Behav Res (2013) 45:499–515
DOI 10.3758/s13428-012-0258-1

Natural language processing in an intelligent writing strategy tutoring system

Danielle S. McNamara · Scott A. Crossley · Rod Roscoe

Published online: 6 October 2012

Abstract The Writing Pal is an intelligent tutoring system that provides writing strategy training. A large part of its artificial intelligence resides in the natural language processing algorithms to assess essay quality and guide feedback to students. Because writing is often highly nuanced and subjective, the development of these algorithms must consider a broad array of linguistic, rhetorical, and contextual features. This study assesses the potential for computational indices to predict human ratings of essay quality. Past studies have demonstrated that linguistic indices related to lexical diversity, word frequency, and syntactic complexity are significant predictors of human judgments of essay quality but that indices of cohesion are not. The present study extends prior work by including a larger data sample and an expanded set of indices to assess new lexical, syntactic, cohesion, rhetorical, and reading ease indices. Three models were assessed. The model reported by McNamara, Crossley, and McCarthy (Written Communication 27:57-86, 2010) including three indices of lexical diversity, word frequency, and syntactic complexity accounted for only 6 % of the variance in the larger data set. A regression model including the full set of indices examined in prior studies of writing predicted 38 % of the variance in human scores of essay quality with 91 % adjacent accuracy (i.e., within 1 point). A regression model that also included new indices related to rhetoric and cohesion predicted 44 % of the variance with 94 % adjacent accuracy. The new indices increased accuracy but, more importantly, afford the means to provide more meaningful feedback in the context of a writing tutoring system.

Keywords Intelligent tutoring systems · Natural language processing · Corpus linguistics · Computational linguistics · Writing pedagogy · Automated essay scoring

D. S. McNamara (✉) · R. Roscoe
Learning Sciences Institute, Arizona State University,
Phoenix, AZ, USA
e-mail: dsmcnamara1@gmail.com

S. A. Crossley
Department of Applied Linguistics/ESL, Georgia State University,
Atlanta, GA, USA

Introduction

The Writing Pal is an intelligent tutoring system (ITS) that provides high school and college students with training on the use of strategies to improve writing quality and, more specifically, on how to write essays (McNamara et al., 2012). We developed this system because of the importance of writing to student education and achievement and because of the lack of available tutoring systems that focus on providing students with instruction on writing strategies. In the Writing Pal, students are provided with lessons on strategies to help them more effectively and efficiently enact the various phases of writing, such as generating and organizing ideas before writing (i.e., freewriting and planning strategies), drafting an essay (i.e., strategies for building the introduction, body, and conclusion), and revising the essay (i.e., strategies for reviewing the essay goals, improving cohesion, and paraphrasing). Each lesson includes practice in the form of mini-games. Students can also practice the strategies by writing prompt-based essays in the Essay Writing Module. An important criterion of an ITS addressing writing instruction is that it must be able to assess students' written work and provide meaningful formative feedback. What makes such a tutoring system *intelligent* is its ability to convincingly "grade" students' essays and return valid, formative feedback that students can apply to improve their writing proficiency. Thus, the creation of the Writing Pal necessitated the development of sophisticated natural

language processing (NLP) algorithms. These algorithms are used to drive interactions within the practice games, to assess the quality of writing, and to guide feedback in the Essay Writing Module.

NLP is a means of creating intelligence for many ITSs, particularly those systems that address ill-defined areas such as writing or that interact with the user via dialogue (e.g., iSTART, McNamara, Levinstein, & Boonthum, 2004; AutoTutor, Graesser et al., 2004). This contrasts with ITSs that address well-defined domains (e.g., algebra, geometry, vocabulary) wherein the concepts and evaluation criteria are tractable and constrained. Within ITSs that accept natural language as input (e.g., essays, verbal explanations of text, problems, or scientific processes), students' responses are open-ended and potentially ambiguous. When a user enters natural language into a system and expects useful feedback or a reasonable response, NLP is used to interpret that input. Indeed, NLP algorithms provide a key source of the perceived intelligence of the Writing Pal.

Figure 1 illustrates the relationship between the user who inputs natural language into an automated system and the algorithms that drive the subsequent response or feedback to the user. NLP algorithms are developed on the basis of principles of artificial intelligence and generally follow the approach of either simulating or imitating human processes. When the objective is to simulate cognitive processes, the variables or features that are used to create the algorithm are guided and constrained theoretically. The overarching goal in this case is often to assess theoretical perspectives about a domain. When the algorithms are situated within the objective of creating a system that mimics human performance (i.e., imitation), the variables or features may be guided by theory but may also have no a priori theoretical connection to the underlying cognitive processes. For example, the linguistic or textual features used to mimic the scoring of essays may not necessarily be the same features that influenced the human (raters') scoring processes. Thus, such features may provide insight into human processes, but not necessarily. This is the case for the Writing Pal and for most, if not all, automated essay scoring algorithms developed to mimic human scoring of essays.

NLP algorithms, in turn, drive the feedback or response to the user (see Fig. 1). This feedback may be primarily guided by the algorithm itself (i.e., the features and variables that make up the algorithm). We label this a *computational approach* because the basis of the feedback emerges from the computation. For example, if the algorithm included grammatical errors, the feedback to the user would use that feature to drive feedback on those errors by instructing the writer to double-check their grammar. Feedback can also be informed through *pedagogical theory*, wherein the features within the algorithm are interpreted in light of a particular pedagogical objective. For example, if the algorithm included the number of words, feedback to the user might suggest strategies for further "elaboration of ideas," rather than merely suggesting that the writer "add more words." That is, features measured by the algorithm may be interpreted as indicators of pedagogical concepts. Whether the feedback is computationally or pedagogically guided, it drives interactions with the user and, theoretically, influences the user's next input or next steps in the system. The degree to which the response is believable, appropriate, well worded, useful, and efficacious determines the *intelligence* of the system.

The construction of effective, intelligent NLP algorithms to interpret writers' input and, subsequently, inform better feedback systems has been one of the major hurdles in Writing Pal development (McNamara et al., 2012). For example, McNamara, Crossley, and McCarthy (2010) used Coh-Metrix to investigate the role of cohesive devices and linguistic sophistication in explaining human ratings of essay quality. Coh-Metrix provides an assortment of indices on the characteristics of words, sentences, and discourse (Graesser & McNamara, 2011; McNamara & Graesser, 2012). Coh-Metrix analyzes text on several dimensions of cohesion including coreferential cohesion, causal cohesion, density of connectives, temporal cohesion, spatial cohesion, and latent semantic analysis (LSA). Coh-Metrix incorporates lexical sophistication indices such as psycholinguistic information about words (concreteness, imagability, meaningfulness, and familiarity scores from the MRC Psycholinguistic Database), semantic word features (polysemy and hypernymy values from WordNet), word frequency indices (CELEX database), and lexical diversity. Coh-Metrix also provides indices related to part-of-speech tagging and syntactic complexity. The primary objective of Coh-Metrix is to provide indices that are potentially related to text difficulty, particularly text cohesion (McNamara, Louwerse, McCarthy, & Graesser, 2010).

McNamara et al. (2010) assessed whether the Coh-Metrix indices successfully distinguished between high- and low-quality essays using a corpus of 120 (untimed, persuasive) college freshman essays scored by human raters using a holistic SAT scoring rubric. A discriminant function analysis (DFA) correctly classified 67 % of the essays as high or low proficiency using three Coh-Metrix indices related to *lexical diversity* (i.e., *MTLD*), *word frequency* (i.e., CELEX logarithm frequency), and *syntactic complexity* (i.e., number of words before the main verb). A stepwise regression analysis using the essay ratings as the dependent variable and Coh-Metrix indices from the DFA as the predictor variables showed that the three indices explained 22 % of the variance in human judgments of essay quality. Overall, the study indicated that human judgments of essay quality were best predicted at the linguistic level by indices related to lexical sophistication (i.e., word frequency and lexical diversity) and syntactic complexity. However, the

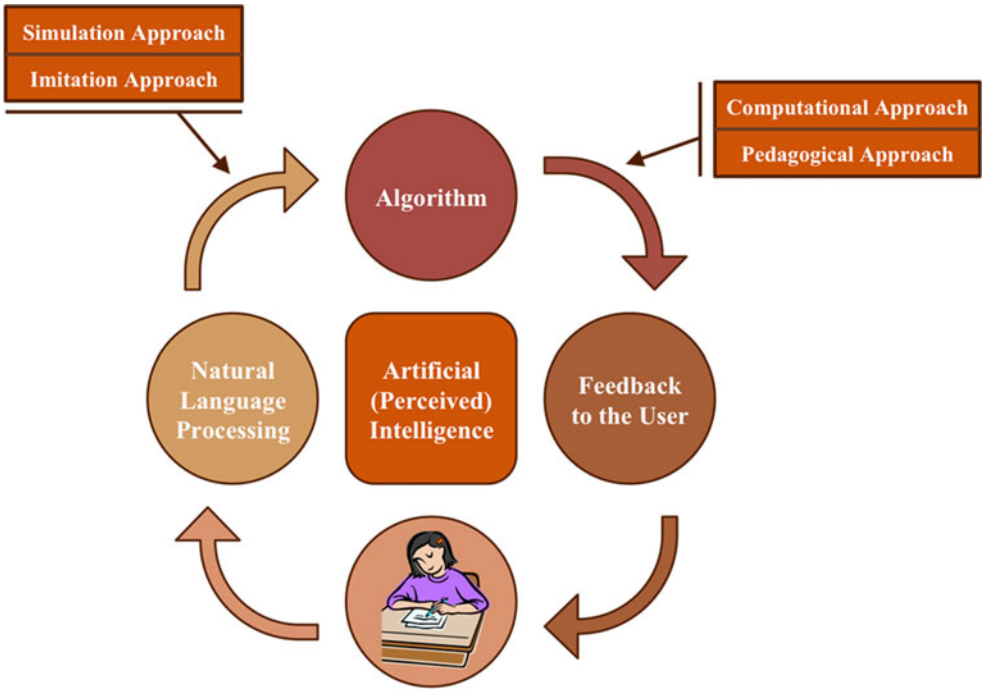

Fig. 1 Cycle between natural language processing and feedback in intelligent tutoring systems to produce *intelligence*

analysis found that no indices of coreference or connectives were significantly correlated with essay scores.

Crossley and McNamara (2010) sought to clarify the importance of text cohesion in writing quality by examining both text features and human judgments of text quality. They examined the degree to which analytical rubric scores of essay quality (e.g., essay cohesion, essay coherence, essay structure, strength of thesis, conclusion type) predicted holistic essays scores. Human judgments of text coherence were the most informative predictor of human judgments of essay quality, explaining 65 % of the variance.

Crossley and McNamara (2010) also examined links between the cohesive devices reported by Coh-Metrix (e.g., semantic coreference, causal cohesion, spatial cohesion, temporal cohesion, connectives and logical operators, anaphoric resolution, word overlap) and human judgments of coherence. They found that few cohesion indices showed significant correlations with the human ratings. Those that were correlated showed a negative relation. Thus, human ratings of coherence were important indicators of holistic evaluations of essay proficiency; however, how human raters construct a coherent mental representation did not correlate positively with the cohesive devices provided by Coh-Metrix.

The results of these studies indicate that writing quality is related to the words and the syntax contained in a text, but not to the cohesive features of the text (although human judgments of text coherence were the most highly predictive features of writing quality). On the surface, this might suggest that feedback from the Writing Pal should focus on these levels of students' writing. Indeed, many automated essay scoring systems provide detailed feedback on lower-level essay features, such as syntax, grammar, spelling, and other characteristics of words and sentences (e.g., Kellogg, Whiteford, & Quinlan, 2010; Shermis & Burstein, 2003). However, a recent meta-analysis of writing interventions conducted by Graham and Perin (2007) indicated that feedback at some lower levels such as grammar and spelling is ineffective. These types of interventions showed an average negative (deleterious) effect size of −.32. By contrast, the most effective interventions were those that provided students with instructions on how to use strategies for various stages of writing such as planning, drafting, editing, and summarizing (Cohen's d = .82). Across the studies reviewed in their meta-analysis, their results indicate that interventions should focus on writing strategies and that writing feedback should seek to help students improve the structure and rhetorical quality of the essay, rather than improving the grammar and spelling within an essay. This makes intuitive sense, especially in the context of very weak essays. When an essay is poor quality with respect to multiple features, it does little good to repair only the grammar and spelling. For example, if only a third of the essay has been written or if the essay is poorly structured and disorganized, it will be more productive to suggest strategies for elaboration or planning than to correct the student's spelling. The student needs to be provided with feedback at the levels that will lead to a more substantive essay. Certainly, grammar and spelling contribute to clearer writing, but providing feedback or instruction at that level does not help the writer to produce higher *quality* essays.

Given the pedagogical objectives of the Writing Pal, algorithms focused primarily on lower levels of writing,

such as lexical sophistication and syntactic complexity, are unlikely to inform feedback on writing at the higher level of rhetorical writing strategies. Hence, our goal in this study was to go beyond the traditional Coh-Metrix developed to assess text difficulty and consider a broader array of indices potentially related to writing quality. These include indices of text difficulty available at the time of the McNamara et al. (2010) study (i.e., *old Coh-Metrix indices*), as well as text difficulty indices that are either newly developed or had not been previously included in analyses of writing (i.e., *new Coh-Metrix indices*). We also included indices developed specifically for the purpose of analyzing writing, which we refer to as *new writing indices*. All of these indices are described in the following section.

Computational indices

Traditional Coh-Metrix indices

Coh-Metrix provides descriptive information about text (e.g., number of words) and linguistic features of text at the level of the word, sentence (i.e., syntax), and intersentenial relationships (i.e., cohesion). Coh-Metrix provides nearly 1,000 linguistic indices about text, of which we included a subset of indices that have been adapted previously in our writing studies. The indices we selected from Coh-Metrix are described briefly below. For a full description of these indices, please see Graesser et al. (2004), and McNamara and Graesser (2012).

Descriptive indices

Coh-Metrix provides a variety of indices that describe the basic properties and structure of a text, such as the number of words, the number of paragraphs, the average length of words, and the average length of sentences.

Lexical indices

Hypernymy

Hypernymy describes the specificity or abstractness of a word. For example, consider the words *car, vehicle*, and *machine*: *Car* is more specific than *vehicle*, which is in turn more specific than *machine*. In other words, *vehicle* is a hypernym (i.e., a more abstract term) for *car*, and *machine* is a hypernym for both *car* and *vehicle*.

To assess hypernymy, Coh-Metrix uses the WordNet database (Fellbaum, 1998). WordNet is a computational lexical database containing over 170,000 English nouns, verbs, adjectives, and adverbs, which have been annotated by experts on various linguistic and psychological features. The words are organized in lexical networks based on connections between related lexical concepts, and each word is located on a hierarchical scale allowing for the measurement of the number of subordinate words below and superordinate words above the target word. Less specific words are assigned a lower value, and thus a lower value equates to less specific word use. Coh-Metrix calculates a mean hypernymy rating across words in the text; thus, a lower score reflects an overall use of less specific words, while a higher value reflects an overall use of more specific words.

Polysemy

Polysemy refers to the number of senses or core meanings of a word and is indicative of text ambiguity. For example, the word *bat* has at least two senses, one referring to an object used to play baseball and the other referring to a flying mammal. Texts that include more polysemous words are less precise, because the words may be understood in different ways. Coh-Metrix measures word polysemy via WordNet and calculates an average polysemy value for content words in a text. A higher value indicates greater polysemy.

Lexical diversity

Lexical diversity (LD) refers to the variety of words used in a text. LD indices generally measure the number of types (i.e., unique words occurring in the text) by tokens (i.e., all instances of words). When the number of word types is equal to the total number of tokens, all of the words are different. By contrast, lexical diversity is lower when more words are used multiple times across the text. Traditional indices of lexical diversity are highly correlated with text length, so Coh-Metrix also reports more sophisticated LD indices, including *MTLD* (McCarthy & Jarvis, 2010) and D (Malvern, Richards, Chipere, & Durán, 2004). Lexical diversity measures relate to the number of words a writer knows.

Word frequency

Word frequency refers to how often particular words occur in the English language and is an important indicator of lexical knowledge. The presence of more uncommon words in a text suggests that the writer possesses a larger vocabulary. The indices reported by Coh-Metrix are obtained from CELEX (Baayen, Piepenbrock, & Gulikers, 1995), a 17.9 million word corpus. Coh-Metrix reports a mean frequency score across words.

Familiarity

Word familiarity refers to how familiar or easily recognized a word seems to a typical adult. For example, the words

table, smile, and *dog* have a higher average familiarity as compared with the words *cortex, dogma,* and *wigwam*. Sentences that contain more familiar words are processed more quickly. Word familiarity ratings are provided via the MRC Psycholinguistic Database (Coltheart, 1981), which provides ratings for several thousands of words along several psychological dimensions. Coh-Metrix reports a mean familiarity score across words in a text. Importantly, more familiar words are not necessarily more frequent. For example, the words *eat* and *while* are equally frequent in language, but the word *eat* is more familiar.

Concreteness

Word concreteness describes the extent to which a word can be understood in terms of concrete sensory experiences (e.g., sight, sound, and touch) rather than an abstract or philosophical meaning. For example, words like *box* or *doctor* that reference objects, materials, or people are more concrete than abstract concepts or ideas like *truth* or *justice*. Concreteness ratings are provided by the MRC Psycholinguistic Database (Coltheart, 1981), and Coh-Metrix calculates the average concreteness rating for nouns in a text.

Imagability

Word imagability refers to how easily one can construct a mental image of a word in one's mind. High-imagery words include terms like *bride* or *hammer*, whereas words like *dogma* or *quantum* are much less imagable. These ratings are provided by the MRC Psycholinguistic Database (Coltheart, 1981), and Coh-Metrix provides the average ratings for nouns in a text.

Meaningfulness

Meaningful words have a greater depth of meaning as given by a high semantic association with other words. For example, the word *people* is semantically related to many more words than is a term such as *abbess.* The MRC Psycholinguistic Database (Wilson, 1988) provides meaningfulness ratings from a corpus developed by Toglia and Battig (1978). Coh-Metrix provides the average ratings for text content words.

Syntactic indices

Syntactic complexity

Sentences that contain a higher number of words before the main verb, a higher number of high-level constituents (sentences and embedded sentence constituents) per word in the sentence, and more modifiers per noun phrase are more syntactically complex and more difficult to process and comprehend (Perfetti, Landi, & Oakhill, 2005). Coh-Metrix calculates the average number of these constructions across sentences in the text.

Syntactic similarity

Syntactic similarity refers to the uniformity and consistency of syntactic constructions in the text at the clause, phrase, and word level. More uniform syntactic constructions result in less complex syntax that is easier for the reader to process (Crossley, Greenfield, & McNamara, 2008). Coh-Metrix calculates the mean level of consistency of syntax at different levels of the text.

Cohesion indices

A primary purpose of Coh-Metrix is to provide measures of text cohesion. The following cohesion measures are validated and described in greater detail in McNamara et al. (2010).

Lexical overlap

Lexical overlap refers to the extent to which words and phrases overlap across sentences and text, thus making a text more cohesive and facilitating text comprehension (Kintsch & van Dijk, 1978). Coh-Metrix considers four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap.

Semantic overlap

Semantic overlap refers to the extent to which phrases overlap semantically across sentences and text. Coh-Metrix measures semantic overlap using LSA, a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts. LSA cosines represent semantic similarity between the words in sentences and paragraphs, an important indicator of cohesion (Landauer, McNamara, Dennis, & Kintsch, 2007).

LSA given/new

Given information has been presented earlier in a discourse. Processing given information can be easier because it is recoverable from the preceding discourse (Chafe, 1975; Halliday, 1967). Coh-Metrix calculates text givenness using perpendicular and parallel LSA vectors (Hempelmann et al., 2005). This is referred to as LSA given/new.

Causal cohesion

Causal cohesion depends on causal relations between events and actions, which helps to create relationships between clauses (Pearson, 1974–1975). Causal cohesion is measured in Coh-Metrix by calculating the ratio of causal verbs (e.g., *kill, break*) to causal particles (e.g., *because*, *by, due to*). The causal verb count is based on the number of main causal verbs identified through WordNet (Fellbaum, 1998).

Connectives

Connective phrases, such as *moreover* or *on the other hand*, make the relationships among clauses and sentences more explicit, and play an important role in the creation of cohesive links between ideas (Longo, 1994). Coh-Metrix assesses the incidence of connectives on two dimensions. The first dimension contrasts positive versus negative connectives, whereas the second dimension is associated with particular classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001). These connectives are associated with positive additive (*also, moreover*), negative additive (*however, but*), positive temporal (*after, before*), negative temporal (*until*), and causal (*because, so*) measures.

Logical operators

Logical operators make the logical flow and relations between ideas explicit and include terms such as *or*, *and*, *not*, and *if–then*. Such terms have been shown to relate directly to the density and abstractness of a text (Costerman & Fayol, 1997). Coh-Metrix assesses the incidence of these terms, combinations of terms, and their common variants.

Anaphoric reference

Anaphoric reference refers to the presence of pronouns that must be resolved by inferring the noun to which they refer from a previous sentence. Anaphoric reference is an important indicator of text cohesion (Halliday & Hasan, 1976). Coh-Metrix measures anaphoric links between sentences by comparing pronouns with previous noun references.

Spatial cohesion

Spatial cohesion helps construct the situational model of a text (Zwaan, Langston, & Graesser, 1995) by developing a spatial representation. According to Herskovits (1998), there are two kinds of spatial information: location information and motion information. Coh-Metrix uses a list of particles provided by Herskovits to capture these two aspects of spatiality. For example, *beside, upon, here,* and *there* indicate location spatiality, whereas the prepositions *into* and *through* indicate motion spatiality. Coh-Metrix also extends Herskovits's theory by assuming that motion spatiality is represented by motion verbs (move, go, run) in WordNet (Fellbaum, 1998) and that location spatiality is represented by location nouns (*place, region*) in WordNet. Coh-Metrix estimates spatial cohesion by tracking the relative frequency of these spatial signals in text (Dufty, Graesser, Lightman, Crossley, & McNamara, 2006).

Temporal cohesion

Temporal cohesion refers to the use of consistent temporal references, such as maintaining the same temporal tense (e.g., past, present, or future) throughout a section of text. Temporal cohesion is also an important element of situational knowledge. Temporal cohesion is measured in Coh-Metrix in three ways: aspect repetition (e.g., progressive and perfect verb forms), tense repetition (e.g., present and past tense), and the combination of aspect and tense repetition.

New Coh-Metrix indices

We selected a variety of Coh-Metrix indices that have been used in previous studies of text analysis but have not been included within our studies of writing quality.

Lexical indices

Lexical categories

Many words can be assigned to multiple syntactic categories. For example, the word *bank* can be a noun (*river bank*) or a verb (*don't bank on it*). Coh-Metrix uses the Charniak parser to calculate incidence scores for all of the part-of-speech tags reported by the Penn Tree Bank Tag Set (Marcus, Santorini, & Marcinkiewicz, 1993). In Coh-Metrix, each word is assigned a lexical category, and these categories are segregated into content words (e.g., nouns, verbs, adjectives, adverbs) and function words (e.g., prepositions, determiners, pronouns). Coh-Metrix assigns only one part-of-speech category to each word on the basis of its syntactic context. Coh-Metrix then computes the *relative frequency* of each word category by counting the number of instances of the category per 1,000 words of text, called *incidence scores*. These indices, which generally relate to grammatical properties of the text, have not been previously investigated in analyses of writing quality. They include measures of adjectives and adverb types (e.g., comparative, superlative), noun types (e.g., singular, plural, proper), personal pronouns, determiners, and verb types (e.g., verb base form, gerunds, past participle, third-person singular).

Syntactic indices

Syntactic categories

Similar to lexical categories for words, many clauses and phrases can also be assigned to particular syntactic categories. For example, phrasal components can include the incidence of noun, verb, and prepositional phrases. Clausal components can include declarative sentences and the number of embedded sentences (s-bars). Coh-Metrix uses the Charniak parser to calculate incidence scores for a variety of syntactic categories and the phrase and clause level. Like lexical categories, these indices have not been investigated in previous Coh-Metrix studies of writing.

Reading ease

Recent research on text readability has led to the development of *component scores* that reflect the ease of processing a text, which were added to Coh-Metrix and current analyses. Graesser, McNamara, and Kulikowich (2011) conducted a principal components analysis including 54 Coh-Metrix indices on 37,520 texts in the TASA (Touchstone Applied Science Associates) corpus. The results showed that eight components accounted for a substantial 67.3 % of the variance of the variability among texts. These eight components are provided in Coh-Metrix both in the form of *Z*-scores and percentile scores (with higher scores indicating greater ease of the text). The eight components described briefly below are described in greater detail in Graesser et al. (2011).

Narrativity

Narrative text tells a story, with characters, events, places, and things that are familiar to the reader. Narrative is closely affiliated with everyday oral conversation. This component is affiliated with word familiarity, world knowledge, and oral language. Nonnarrative texts on less familiar topics lie at the opposite end of the continuum.

Syntactic simplicity

This component reflects the degree to which the sentences in the text contain fewer words and use familiar syntactic structures. At the opposite end of the continuum are texts that contain sentences with more words and use complex, unfamiliar syntactic structures.

Word concreteness

Texts that contain content words that are concrete, meaningful, and evoke mental images are easier to process and understand. Abstract words represent concepts that are difficult to represent visually. Texts that contain more abstract words are more challenging to understand.

Referential cohesion

A text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader. Low-cohesion text is typically more difficult to process because there are fewer connections that tie the ideas together for the reader.

Deep (situation model) cohesion

This dimension reflects the degree to which the text contains causal and intentional connectives when there are causal and logical relationships within the text. These connectives help the reader to form a more coherent and deeper understanding of the causal events, processes, and actions in the text. When a text contains many relationships but does not contain those connectives, the reader must infer the relationships between the ideas in the text. If the text is high in cohesion, those relationships and global cohesion are more explicit.

Verb cohesion

This dimension reflects the degree to which there are overlapping verbs in the text. When there are repeated verbs, the text likely includes a more coherent event structure that will facilitate and enhance comprehension. This dimension is likely to be more relevant for texts intended for younger readers and for narrative texts (McNamara, Graesser, & Louwerse, 2012).

Connectivity

This dimension reflects the degree to which the text contains explicit adversative, additive, and comparative connectives to express relations in the text. This score reflects the number of logical relations in the text that are explicitly conveyed.

Temporality

Texts that contain more cues about temporality and that have more consistent temporality (i.e., tense, aspect) are easier to process and understand. In addition, temporal cohesion contributes to the reader's situation model level understanding of the events in the text.

New writing indices

The Coh-Metrix team has developed a variety of new indices specifically to assess the quality of persuasive essays in

the Writing Pal system. Many of these indices have not been reported in previous studies. Therefore, we discuss them in greater detail below.

Lexical indices

Basic lexical types

For the purpose of analyzing writing, we developed two new indices that measure the basic properties of text. The first index was the number of lexical types in the text (i.e., *total types*). This variable represented the number of different words included in the essays (as opposed to tokens). The measure, total types, is thus indicative of the variation in word usage in the essay. The second was the number of content words contained in the text. Content words include verbs (e.g., *act*, *run*), nouns (e.g., *chair*, *person*), adverbs (e.g., *slowly*, *carefully*), and adjectives (e.g., *red*, *pretty*), as opposed to function words (e.g., *the*, *a*, *this*, *that*, *what*).

Lexical sophistication indices

Given the importance of lexical sophistication in predicting human judgments of essay quality (e.g., McNamara et al., 2010), we developed new indices of lexical sophistication that incorporated the Academic Word List (570 words commonly found in academic writing; Coxhead, 2000), a list of vague words (e.g., *whatever*, *people*, *stuff*, *thing*), and the total number of word types found in the text. We expected writing quality to be positively related to academic writing and negatively related to vague words.

Cohesion indices

Global cohesion indices

While a large number of cohesion indices are included in Coh-Metrix, we nonetheless developed new indices in consideration of the findings of Crossley and McNamara (2010) that text coherence as judged by expert raters was the most predictive analytical feature of essay quality. Moreover, they found that the analytical judgments of text coherence were not positively correlated with the cohesion indices reported by Coh-Metrix. Our new indices of cohesion are intended to capture elements of coherence that may be specific to essay writing. Specifically, these new indices calculate keyword and LSA comparisons, respectively, to assess lexical and semantic overlap between specific paragraphs in the essays (initial to middle paragraphs, middle paragraphs to final paragraph, and initial paragraph to final paragraph), with the understanding that lexical and semantic links between paragraphs will help to develop text coherence in the mental representation of the reader.

Contextual cohesion indices

Because Coh-Metrix focuses on linguistic features, it does not consider cohesion that may be driven by contextual factors. Here, we consider semantic characteristics of the essay such as the degree of overlap between the prompt and the essay. For example, a student may be prompted to write on a certain topic, such as the difference between heroes and celebrities. The contextual cohesion indices we developed assess lexical and semantic overlap between a prompt the essay. These indices included LSA comparisons between the prompt and the essay and keyword comparisons between the prompt and the essay. Thus, keyword comparisons capture lexical overlap between the prompt and essay, and LSA captures semantic overlap between the two. Such indices provide a means to assess whether or not the essay produced by the writer is contextually relevant to the prompt.

We also compute the number of key words and key types used for each essay using a reference corpus that is specific for individual prompts. The reference corpus consists of at least 30 essays written for an individual prompt. From this corpus, we extract the key words that are common to the corpus on the basis of the frequency distribution across the essays (i.e., through a measure of entropy). This list of key words is then used to calculate the incidence of key words and key types (i.e., word token count and a word type count) found in an individual essay. Such a measure assesses how well the writer is producing words that are contextually relevant to the prompt.

Rhetorical indices

Rhetorical strategies are used to persuade the reader. There are a number of strategies, such as the use of exemplification, convincing arguments, description, narrations, and so on. Hence, we created semantic categories related to or proxies for various rhetorical strategies as found in Quirk, Greenbaum, Leech, and Svartvik (1985). These include indirect pronouns (*all*, *none*, *some*), amplifiers and emphatics (*extremely*, *definitely*), downtoners (*slightly*, *somewhat*, *almost*), and exemplification (*for instance*, *namely*).

We also developed n-gram indices for words and phrases common in high-quality introduction, body, and conclusion paragraphs taken from a corpus of argumentative essays written by freshmen college students and scored by trained human raters. These indices differ from the other rhetorical indices we developed because they are domain specific. For these n-gram indices, we compared a corpus of high-quality paragraph types (e.g., introductions) with a corpus of low-quality paragraphs of the same type. We used WordSmith (Scott, 1996) to identify n-grams that were unique to high-quality paragraph types, as compared with low-quality paragraph types, on the basis of keyness (i.e., an n-gram that

occurs more often than would be expected by chance, in comparison with the reference corpus). The key n-grams were then categorized on the basis of rhetorical features. For instance, introductory paragraphs contained n-grams related to reported speech (i.e., *said*), contrast (i.e., *but some*), strength of argument (i.e., *we see*), and outside reference (i.e., *a person*). Body paragraphs contained n-grams related to providing examples (i.e., *addition to*), pronouns, conditionals (i.e., *if an*), and contrast (i.e., *while the*). Concluding paragraphs contained concluding statements (i.e., *in conclusion*), statements of fact (i.e., *it is*), negation, conditionals, modals, opinion (i.e., *I think*), and reason (i.e., *because*).

Method

The goal of this study is to investigate the roles of linguistic, cohesive, and rhetorical features in persuasive essays that predict essay scores assigned by human raters. We also investigate the added value of considering text difficulty indices such as lexical, syntactic, cohesion, rhetorical, and reading ease indices previously unexamined in writing research. In addition, we developed new writing indices that examined rhetorical devices, global cohesion, contextual cohesion, and additional elements of lexical sophistication. For this study, we used a methodology similar to that in McNamara et al. (2010), except that we collected a larger essay corpus that better reflected high-stakes testing conditions. To investigate the Coh-Metrix indices on the larger corpus, we conducted three studies. We first tested the regression model reported by McNamara and colleagues on the larger corpus. We next conducted a second analysis using only traditional (i.e., old) Coh-Metrix indices. These are indices designed to assess text difficulty, but not writing quality in particular. This analysis provided a baseline from which to compare the success of the new indices. Last, we conducted a regression analysis using traditional Coh-Metrix indices combined with the newly developed Coh-Metrix and Writing indices to assess both their predictive strength and their value in providing meaningful feedback in the context of the Writing Pal.

Corpus collection

We collected 313 timed (25-min) essays written by 313 college freshmen at the Mississippi State University (i.e., the MSU timed corpus; Crossley, Roscoe, McNamara, & Graesser, 2011). All essays were written in response to two Scholastic Aptitude Test (SAT) writing prompts. The prompts did not require specific domain knowledge and were intended to relate to a variety of ideas. This corpus differed from the corpus analyzed in McNamara et al. (2010) in that the essays were timed and the prompts were general knowledge. We chose to use timed essays primarily because these types of essays better reflected the conditions under which students usually complete prompt-based essays, such as the SAT essay, and because timed prompt-based essays are primarily the target of the Writing Pal. Hence, the results of the current algorithm are more likely to be accurate in the context of the Writing Pal.

Essay evaluation

Eight expert raters with at least 4 years of experience teaching freshman composition courses at a large university rated the quality of the 313 essays in the corpus. Two raters evaluated each essay on the basis of a commonly used standardized SAT rubric. The rubric generated a holistic quality rating with a minimum score of 1 and a maximum of 6. Raters were informed that the distance between each score was equal. The raters were first trained to use the rubric with 20 similar essays taken from another corpus. Pearson correlations were conducted between all possible pairs of rater responses. The resulting eight correlations were averaged to provide a mean correlation between the raters. This correlation was then weighted on the basis of the number of raters (Hatch & Lazaraton, 1991). Once the correlations within the raters reached a threshold of $r = .70$ ($p < .001$), the raters were considered trained. After the first round of training, all ratings for the holistic scores correlated at .896. The final interrater reliability for all essays in the corpus was $r > .75$. We used the mean score between the raters as the final value for the quality of each essay unless the differences between the two raters was ≥ 2, in which case a third expert rater adjudicated the score.

Statistical analysis

For each of our three analyses, two statistical analyses were conducted. The first analysis assessed the strength of our selected indices in predicting the human scores of the MSU timed corpus using a regression analysis. The second analysis examined how accurately the scores produced by the regression model matched the human scores. For the regression analysis, we first conducted Pearson correlations between the Coh-Metrix indices and the human scores assigned to each essay. After correcting for multicollinearity (i.e., eliminating predictor variables with correlations $\geq .70$), these variables were then used to predict the human scores using a linear regression model. This model was then tested using tenfold cross-validation, in which the data (in this case the 313 essays) are split into 10 subsets. Nine of these subsets are used to develop a regression model that is then tested on the left-out subset. This process is repeated 10 times, so that all data are used to both train and test the

model. Such an approach allows for the calculation of predictability for the variables in an independent corpus. We selected a tenfold cross-validation approach because numerous experiments have shown it to be the best choice for deriving an accurate estimate (Lecocke & Hess, 2006; Molinaro, Simon, & Pfeiffer, 2005; Witten & Frank, 2005).

Our second statistical analysis assessed two types of accuracy with the human scores: exact accuracy and adjacent accuracy (i.e., within 1 point). Exact accuracy examines how accurate the regression model is in terms of assigning the same score to the essay as did the human raters. Adjacent accuracy examines the accuracy of the regression model in assigning a score to the essay that is either exactly the same or adjacent to that assigned by the human raters. For this analysis, we rounded the score derived from the regression up or down to the closest whole number. Thus, if the model assigned a score of 3.56 (rounded to a score of 4) to an essay that was rated by humans as a 4, the exact accuracy would be 1, and the adjacent accuracy would be 1. If the model assigned the same essay a score of 3.2 (rounded to a score of 3) the exact accuracy would be 0, and the adjacent accuracy would be 1. If the model assigned the essay a score of 2.0 (or 6.0), the exact accuracy and the adjacent accuracy would be 0. We also calculated the chi-square and weighted and unweighted Cohen's kappa for the predicted versus actual classifications.

Results

McNamara et al. (2010) model

Regression model

We used the regression model reported by McNamara et al. (2010) on the 313 essays in the data. The model yielded r = .247, r^2 = .061. The results from this model extended to the larger data set demonstrate that the combination of the three variables reported by McNamara, Crossley, and McCarthy accounts for only 6 % of the variance in the human evaluations of essay quality.

Exact and adjacent matches

We used the scores derived from the regression model to assess the exact and adjacent accuracy of the regression scores when compared with the human-assigned scores. This is a standard method employed by researchers and developers who assess the reliability of essay scoring rubrics and automated scoring algorithms because a score that is only 1 score off (i.e., adjacent accuracy) is more acceptable than a score that is off by 2 or more points (Attali & Burstein, 2006; Dikli, 2006; Rudner, Garcia, & Welch, 2006; Shermis, Burstein, Higgins, & Zechner, 2010). The regression model produced exact matches between the predicted essay scores and the human scores for 90 of the 313 essays (29 % exact accuracy). The model produced adjacent matches for 228 of the 313 essays (73 % adjacent accuracy). The measure of agreement between the actual score and the predicted score produced a weighted Cohen's kappa for the adjacent matches was 0.143, demonstrating a poor agreement.

The confusion matrix for this analysis provided in Table 1 provides the alignment between the predicted scores based on the regression equation and the human scores. This matrix further illustrates the poor performance of the model. Perfect performance would be reflected by high frequencies along the diagonal, indicating that the predicted score was the same as the actual human score. However, in this case, the predicted scores are not well aligned with the actual scores.

Traditional Coh-Metrix indices

Pearson correlations training set

We selected the traditional (i.e., old) Coh-Metrix indices that demonstrated the highest Pearson correlation when compared with the human essay scores and that did not demonstrate multicollinearity with one another. Multicollinearity was established if the variables correlated ≥ .70. The highest correlated variables were then retained. The 10 selected variables along with their r values and p values are presented in Table 2, sorted by the strength of the correlation.

Among these indices, we observe results similar to those reported in other data sets. Essay quality is positively correlated with essay length (i.e., number of words), syntactic complexity (high-level constituents per word), lexical specificity and imageability (noun hypernymy, word imageability), and lexical diversity (D). Essay quality is also negatively correlated with cohesion indices related to content word overlap and spatial cohesion, along with lexical simplification (word frequency and word meaningfulness). Unlike past data sets, there is a positive correlation with LSA given/new, indicating that better essays have some sense of semantic

Table 1 Predicted human score: McNamara et al. (2010a) model

Actual human score	Predicted human score					
	1	2	3	4	5	6
1	2	0	1	1	0	0
2	11	10	13	5	3	4
3	10	33	45	26	10	0
4	10	16	20	24	8	0
5	2	7	16	25	9	1
6	0	0	0	0	1	0

Table 2 Correlation between traditional Coh-Metrix indices and essay scores

Index	Type	*r*	*p*
Number of words	Descriptive	.517	<.001
Word frequency content words	Lexical	−.343	<.001
Noun hypernymy	Lexical	.291	<.001
Lexical diversity D	Lexical	−.232	<.001
Word imageability content words	Lexical	.189	<.010
High-level constituents per word	Syntactic	−.184	<.050
LSA given/new	Cohesion	.175	<.050
Word meaningfulness	Lexical	−.137	<.050
Spatial cohesion	Cohesion	−.118	<.050
Content word overlap	Cohesion	−.120	<.050

cohesion that is picked up by this index. In sum, among the Coh-Metrix indices, there are few surprises.

Multiple regression

A linear regression analysis was conducted including the 10 variables. These 10 variables were first regressed onto the human raters' score for the 313 essays in the corpus and were checked for outliers and multicollinearity (i.e., through Tolerance checks, VIF values, and correlations). No outliers or multicollinearity was found between variables. The linear regression yielded a significant model, $F(6, 306) = 36.282$, $p < .001$, $r = .645$, $r^2 = .416$ (see Table 3 for details). Six variables were significant predictors: *number of words*, *word frequency*, *LSA given/new*, *noun hypernymy*, *word imageability*, and *content word overlap*. The results from the linear regression demonstrate that the combination of the six variables accounts for 42 % of the variance in the human evaluations of essay quality.

To validate the model developed from the initial regression, we used tenfold cross-validation modeling. The model produced an estimated value for each writing sample in the test set. We then conducted a Pearson correlation between the estimated scores and actual scores. We used this correlation along with its r^2 to evaluate the strength of the model using cross-validation. The model for the tenfold cross-validation set yielded $r = .614$, $r^2 = .377$. Thus, the combination of the six variables accounted for 38 % of the variance in a cross-validated set.

Table 3 Regression analysis results for timed Mississippi State University (MSU) corpus using traditional Coh-Metrix variables

Entry	Variable added	*R*	R^2	*B*	B	*SE*
1	Number of words	.517	.267	0.004	0.483	0.000
2	Word frequency content words	.607	.368	−1.810	−0.300	0.361
3	LSA given/new	.621	.386	4.331	0.189	1.275
4	Noun hypernymy	.630	.397	0.300	0.134	0.111
5	Word imageability content words	.638	.407	−0.007	−0.127	0.003
6	Content word overlap	.645	.416	−3.017	−0.126	1.410

Constant = 6.211

Exact and adjacent matches

We used the scores derived from the tenfold cross-validated regression to assess the exact and adjacent accuracy of the regression scores when compared with the human-assigned scores. The regression model produced exact matches between the predicted essay scores and the human scores for 133 out of the 313 essays (32 % exact accuracy). The model produced adjacent matches for 284 of the 313 essays (91 % adjacent accuracy). The reported weighted Cohen's kappa for the adjacent matches was 0.293, demonstrating a fair agreement.

A confusion matrix for this analysis is provided in Table 4. The matrix illustrates an improvement over the McNamara, Crossley, and McCarthy (2010) model, particularly in terms of adjacent matches. That is, when the predicted score is incorrect, the matrix shows that the errors tend to be centered around the actual score (i.e., within 1 point). Nonetheless, the performance is poor to fair.

Coh-Metrix and writing indices

Pearson correlations training set

We selected the traditional and new Coh-Metrix indices that demonstrated the highest Pearson correlation when compared with the human essay scores and that did not demonstrate multicollinearity with one another. Among these variables, 40 showed correlations at $p < .05$. Table 5 presents the top 26 variables that were significantly correlated at $p < .001$, sorted by the strength of the correlation.

Among the significant correlations, three quarters comprise the new indices (starred and labeled in Table 5). Among the traditional Coh-Metrix indices, we observe

Table 4 Predicted human score: Traditional Coh-Metrix indices

Actual human score	Predicted human score					
	1	2	3	4	5	6
1	0	2	2	0	0	0
2	2	14	24	6	0	0
3	0	11	79	34	0	0
4	0	2	40	35	1	0
5	0	0	19	36	5	0
6	0	0	0	0	1	0

Table 5 Correlations: New and traditional indices to essay scores

Index	Old/New	Type	*r*	*p*
Total types[a]	W-New	Lexical	.526	.000
Academic words[a]	W-New	Lexical	.427	.000
Key types[a]	W-New	Lexical	.362	.000
Word frequency content words	Old	Lexical	−.343	.000
Body paragraphs n-grams[a]	W-New	Rhetorical	.326	.000
LSA introduction to middle paragraphs[a]	W-New	Cohesion	.323	.000
Incidence of declarative sentences[a]	C-New	Syntactic	−.294	.000
Amplifiers and emphatics[a]	W-New	Rhetorical	.293	.000
Noun hypernymy	Old	Lexical	.291	.000
Indirect pronouns[a]	W-New	Rhetorical	.280	0.000
Lexical diversity D	Old	Lexical	.232	.000
Incidence of verb phrases[a]	C-New	Syntactic	−0.232	.000
Narrativity score[a]	C-New	Ease	−.222	0.000
Incidence of S-bars[a]	C-New	Syntactic	−.199	0.000
Word imageability content words	Old	Lexical	.189	.001
Verb cohesion[a]	C-New	Ease	−0.186	.001
Exemplification[a]	W-New	Rhetorical	.183	.001
Incidence of prepositional phrases[a]	C-New	Syntactic	.177	.002
Conclusion paragraph n-grams[a]	W-New	Rhetorical	.176	.002
LSA given/new	Old	Cohesion	.175	.002
Downtoners[a]	W-New	Rhetorical	.174	.002
LSA essay to prompt[a]	W-New	Cohesion	.169	.003
Modifiers per noun phrase	Old	Syntactic	.166	.003
Vague nouns[a]	W-New	Rhetorical	.165	.003
Incidence of verb base forms[a]	C-New	Lexical	−.162	.004
Keyword initial to final paragraph[a]	W-New	Cohesion	.162	.004

[a] = new indices; Old = traditional Coh-Metrix indices; C-New = new Coh-Metrix indices; W-New = new writing indices; Bolded indices were retained in the regression analysis

results similar to those reported in the second analysis (with the exception of content word overlap, which was removed because of multicollinearity with the LSA given/new index). Of the new writing indices, a few tapped constructs similar to those assessed by Coh-Metrix. The lexical index, total types, is related to the number of words in the text or the length of the text. Academic words are signals for lexical sophistication, and vague words are signals for the lack of lexical sophistication. In addition, several of the rhetorical indices showed positive correlations with essay quality, including the use of amplifiers (*extremely*, *definitely*), indirect pronouns (*all*, *none*, *some*), downtowners (*slightly*, *somewhat*, *almost*), and exemplification (*for instance*, *namely*). Narrativity also showed a significant negative correlation, with essay score demonstrating that essays with more narrativity and less information were scored lower. In addition, the correlations confirm that higher quality essays have greater semantic overlap between the initial paragraph and the body paragraphs and share more keywords between the introduction and the conclusion paragraphs. Additionally, two indices of relevance were positively correlated with essay score: LSA cosines between the essay and the prompts and the number of key types in the essay. Lastly, a variety of part-of-speech and syntactic categories demonstrated significant correlations with essay scores. These indices indicate that grammatically and syntactically less complex essays were scored lower.

Multiple regression

A linear regression analysis was conducted with the 40 new and traditional Coh-Metrix variables. These 40 variables were first regressed onto the raters' score for the 313 essays in the corpus and were checked for outliers and multicollinearity. The linear regression yielded a significant model, $F(8, 299) = 35.453$, $p < .001$, $r = .698$, $r^2 = .473$ (see Table 6 for details). Eight variables were significant predictors: *total types*, *LSA given/new*, *narrativity reading ease score*, *noun hypernymy*, *LSA essay to prompt*, *conclusion paragraph n-grams*, *body paragraph n-grams*, *and word frequency.* The results from the linear regression demonstrate that the combination of these eight variables accounts for 47 % of the variance in the human evaluations of essay quality.

To validate the model developed from the initial regression, we used tenfold cross-validation modeling. The model produced an estimated value for each writing sample in the test set. We then conducted a Pearson correlation between the estimated scores and actual scores. We used this

Table 6 Regression analysis results for timed Mississippi State University (MSU) corpus using traditional and new Coh-Metrix indices and new writing indices

Entry	Variable added	*R*	R^2	*B*	B	*SE*
1	Total types[a]	.531	.282	0.011	0.473	0.001
2	LSA given/new	.593	.351	5.887	0.253	1.018
3	Narrativity score[a]	.636	0.404	−0.213	−0.126	0.082
4	Noun hypernymy	.654	.428	0.265	0.119	0.104
5	LSA essay to prompt[a]	.672	.451	1.872	0.168	0.479
6	Conclusion paragraph n-grams[a]	.684	.467	0.016	0.113	0.006
7	Body paragraphs n-grams[a]	.692	.478	0.005	0.106	0.002
8	Word frequency content words	.698	.487	−0.737	−0.121	0.335

Constant = −1.16

[a] New indices

correlation along with its r^2 to evaluate the strength of the model on an independent data set. The model for the tenfold cross-validation set yielded $r = .675$, $r^2 = .456$. Thus, the combination of the six variables accounted for 46 % of the variance in a cross-validated set.

Exact and adjacent matches

We used the scores derived from the tenfold cross-validated regression to assess the exact and adjacent accuracy of the regression scores when compared with the human-assigned scores. The regression model produced exact matches between the predicted essay scores and the human scores for 139 out of the 313 of the essays (44 % exact accuracy). The model produced adjacent matches for 294 of the 313 essays (94 % adjacent accuracy). The reported weighted Cohen's kappa for the adjacent matches was 0.401, demonstrating a moderate agreement.

A confusion matrix for this analysis is provided in Table 7. This matrix reflects an increase in exact matches and the stronger adjacent agreement using this model. The predicted scores tend to be within 1 point of the actual score. Nonetheless, performance remains moderate, with a fair number of misclassifications. This level of performance is partially due to the number of categories that are being predicted, which renders the classification task more difficult.

Discussion

The intelligence of a tutoring system for writing instruction, or any ITS that must assess and respond to open-ended student responses, is grounded in the natural language algorithms that process those responses. Recent advances in disciplines such as computational linguistics, discourse processing, and information retrieval have made it possible to computationally investigate textual features that impact judgments of essay quality. Together, these advances enable accurate, detailed, and automated analyses of surface and deep-level factors of lexical sophistication, syntactic complexity, contextual relevance, rhetorical features, and various levels of cohesion.

Table 7 Predicted human score: Coh-Metrix indices and new writing indices

Actual human score	Predicted human score					
	1	2	3	4	5	6
1	0	3	1	0	0	0
2	2	20	21	3	0	0
3	0	12	78	34	0	0
4	0	2	40	33	3	0
5	0	0	13	39	8	0
6	0	0	0	0	1	0

In the present study, we extended our prior work by including a larger data sample and an expanded set of linguistic, cohesive, and rhetorical features to assess the added value of considering new Coh-Metrix and writing indices. We computed three regression models. First, we assessed the fit of the model reported by McNamara et al. (2010) including the three indices of lexical diversity, word frequency, and syntactic complexity. Whereas their model accounted for 22 % of the variance in their data set, it accounted for only 6 % of the variance in the human evaluations of essay quality when applied to the present data set. The reduced accuracy of their model with the present data set may be attributable to any number of factors. However, the most salient difference between the studies is that the essays in the previous data set were untimed (take home) essays, whereas the essays in the present data set were timed (25 min) essays. Whether or not an essay is timed will potentially affect both its content and the length of the essays. In the case of the essays used in McNamara and colleagues, the assignment called for 750 word essays, and consequently, most of the essays approximated 750 words. Thus, the model reported by McNamara, Crossley and McCarthy did not include number of words (or word types), because length was not a source of variation between essays. However, the length of an essay is often a strong predictor of untimed essay quality. These findings provide some impetus for future research to explore the differences between timed and untimed essays.

The purpose of our second analysis was to examine the predictive value of the full set of traditional Coh-Metrix indices used in prior studies of writing. The findings from this study demonstrated that a combination of six computational indices including the number of words in the essay, word frequency, LSA given/new, noun hypernymy, word imagability, and content word overlap accounted for 38 % of the variance in human scores of essay quality with 91 % adjacent accuracy (i.e., within 1 point). This analysis provides a baseline and indicates which of the Coh-Metrix linguistic indices account for the most variance in the present essay corpus. The high-quality essays were longer, with more sophisticated words, and also had more specific, imageable words, indicating the potential importance of providing grounded examples in argumentative essays. Cohesion as measured by content word overlap was negatively related to essay quality, as found in previous studies (Crossley & McNamara, 2010, 2011). However, LSA given/new was positively related to essay quality. The latter two results indicate that although higher quality essays did not have greater overlap between sentences, they did have some sense of global, semantic cohesion. Across the essay, there was a greater proportion of information that had already been provided in the essay (given) than of information that was new.

The third regression model was computed to examine the added value of including new indices potentially more related to writing. Our goal was to account for a greater amount of variance in the essays, but also to include more indices with potential links to writing strategies and writing strategy feedback. The algorithms produced by using traditional text difficulty indices in Coh-Metrix provide some pointers to feedback for the writer. But they do not provide feedback on the use of the types of writing strategies that have been shown to have the largest effects on writing ability (Graham & Perin, 2007). Hence, we included a variety of new indices, including indices related to rhetoric and cohesion. We assumed that rhetorical cues were potentially important to human raters' scoring of essay quality because the use of these cues is a signature of text quality. These indices have not been included in Coh-Metrix because it was constructed primarily to provide indices of text difficulty, not text quality.

The findings from the third regression analysis demonstrated that a combination of eight computational indices including *the number of different words (total types)*, *LSA given/new*, *narrativity reading ease score*, *noun hypernymy*, *LSA essay to prompt*, *conclusion paragraph n-grams*, *body paragraph n-grams*, and *word frequency* accounted for 44 % of the variance, with 94 % adjacent accuracy. The new indices increase accuracy but, more important, afford the means to provide more meaningful feedback in the context of a writing tutoring system.

What do these indices tell us about essay quality? First, longer essays with more sophisticated vocabulary were judged higher in quality. In terms of strategy interventions, students can be taught strategies to help them to generate more text (e.g., strategies to facilitate freewriting, planning, drafting, and elaboration), and such strategies improve their essay quality. However, improving students' knowledge and skill at the levels of word knowledge requires time and deliberate practice.

In contrast to past studies, we also found that higher quality essays displayed higher global and contextual cohesion. This was manifested in the form of more given information and greater semantic (LSA) overlap between the prompt and the essay. Essays judged higher in quality maintained stronger links to previously supplied information (i.e., higher LSA given/new) and better maintained the topic of the prompt across the entire essay. These results are important because prior studies suggested that local cohesion played little role in human essay ratings (McNamara et al., 2010), while human judgments of an essay's coherence are strongly related to overall judgments of quality (Crossley & McNamara, 2010). The results of this study indirectly point toward more successful measures of experts' ratings of essay coherence. In turn, within the Writing Pal, students are taught multiple strategies for building and maintaining a common thread of ideas throughout the essay and for addressing the topic of the prompt. The results of the present study can be applied to carefully guide where and how students receive feedback on cohesion-building and revision strategies, or when students are directed to review cohesion building lessons and practice the strategies via mini-games.

Three additional indices contributed significantly to the model: *narrativity reading ease score*, *conclusion paragraph n-grams*, *and body paragraph n-grams.* We consider these three indices to provide signatures of the presence of rhetorical cues in the essays. The reading ease index, narrativity, provides a measure of the difficulty of a text, because narrative texts that are low in narrativity contain less familiar words and generally cover more unfamiliar topics (Graesser et al., 2011). This index indicates that the better essays included more information or content, fewer pronouns, and fewer stories about events. Thus, as expected, the better essays included more features characteristic of informational than did narrative texts.

The higher quality essays also included more phrasal constructions (n-grams) typical of higher quality essay bodies and conclusions. For example, in the bodies of the essays, writers were more likely to include examples (i.e., *addition to*) and make contrasts between ideas (i.e., *while the*). In the conclusion paragraphs, the writers were more likely to include concluding statements (i.e., *in conclusion*), statements of fact (i.e., *it is*), and reasons (i.e., *because*). In the Writing Pal, students are taught strategies for drafting and improving body and conclusion paragraphs that map onto these findings. For instance, students are taught how to identify and edit evidence in the body of the essay that is overly speculative (i.e., too many hypothetical claims) rather than fact based and objective. Similarly, students are taught to write conclusions that succinctly summarize major arguments without presenting additional or new evidence.

Referring back to Fig. 1, one aspect of algorithm development, particularly in the context of natural language, is the degree to which simulation and imitation are objectives. The approach that we have adopted is imitation in the sense that our goal is to mimic the human essay scores, but not necessarily to simulate the underlying processes in either the scoring process or the processes engaged in writing the essay. As such, the relationship to essay quality is explored for a relatively large number of indices, and those that are most highly correlated are included in the algorithm. This approach is appropriate when the goal is not to assess or compare theories of writing but, rather, to develop a system that mimics human intelligence artificially. This approach contrasts with a more theoretically driven approach where the goal is to simulate behavior. In the latter case, the indices included in the model would be based solely on a theoretical model or framework.

In this study, the set of indices are also motivated theoretically. First, the indices in Coh-Metrix were developed on the basis of the theoretical assumption that texts can be understood in terms of levels of comprehension, including surface, textbase, and situation model levels (Graesser & McNamara, 2011). Thus, in order to assess text readability, linguistic indices must be able to capture text properties relevant to each level and relations among each level. Second, another guiding theoretical assumption was that judgments of writing quality involve levels beyond that of text comprehension. Specifically, judgments of writing are assumed to be affected by rhetorical and cohesion cues in the text. This assumption was confirmed in both the pattern of correlations and the regression results. Thus, the resulting algorithm is informed by strong theoretical principles related to text and writing quality and by statistical AI methods for extracting specific indices that are meaningfully predictive of writing quality.

Indeed, one major hurdle we have faced in the development of the Writing Pal has been the development and improvement of algorithms to improve the interpretation of writers' input such that we can, in turn, develop better feedback systems (Roscoe, Kugler, Crossley, Weston, & McNamara, 2012; Roscoe et al., 2011). The new indices we have explored in this study provide some insight into human judgments of essay quality. For example, a common fault in student essays is the lack of a clear concluding paragraph. Our model reported a positive relationship between essay quality and the incidence of conclusion n-grams (e.g., concluding phrases, conditionals, and modals). Thus, our new n-gram-based measure seemed able to detect the presence of a key rhetorical element important in judgments of essay quality. Additionally, lower quality essays include more personal narratives, suggesting that weaker writers relied more on writer-based prose than reader-based prose (e.g., Flowers, 1979).

How might the latter indices inform the delivery of formative feedback for developing writers in the Writing Pal? Two illustrative examples can be considered. If a clear conclusion is not detected by our index, students could be provided with feedback that reminds them of the role and importance of conclusions, as well as provide them strategies for authoring an effective conclusion. Indeed, this instruction is provided in the Conclusion Building module of the Writing Pal, and our new automated indices may allow us to determine whether and when to remind students of helpful mnemonics or direct them to the Conclusion Building module for further practice. Our results also indicated that higher narrativity was related to lower essay scores. Greater narrativity indicates that the writer relied on personal perspectives or anecdotes to communicate and argue their main ideas. Our new indices can detect essays that are overly narrative and inform writers that they may need to develop evidence that appeals to a broader audience. Students could also be directed to study the Planning or Body Building modules and practice games, which discuss how to select, organize, and present one's arguments in an essay.

The results demonstrate that the expansion of the indices assessed in our analyses, such as new rhetorical and contextual indices, contributed positively to the predictive power of the resulting algorithms. A number of these new measures were correlated with human ratings of essay quality and may be worth further exploration as the algorithm development process continues. Additionally, as research on automated linguistic analysis continues to advance, so does our ability to detect and understand the textual features that contribute to effective writing. In turn, this empowers us to teach developing writers how to harness such knowledge to further their academic and professional goals, both via traditional feedback given by teachers and by automated feedback and strategies taught by intelligent tutoring systems such as the Writing Pal. Within the Writing Pal project, we continue to explore not only the intervention we are building, the Writing Pal lessons, but also the algorithms and the feedback generated on the basis of those algorithms. As we do so, we continue to learn more about the writing process and how to improve both the algorithms and the feedback that is provided to the student. As such, our present and future work continues to build upon and expand our understanding of writing, NLP, and intelligent tutoring.

Acknowledgments This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We are thankful to the members of the Writing Pal project who have contributed feedback to various aspects of this study and other studies that have led to this study. We are particularly thankful to Zhiqiang Cai and Art Graesser. We also thank Russell Brandon, Laura Varner, and Jen Weston, as well as Brad Campbell, Daniel White, Steve Chrestman, Michael Kardos, Becky Hagenston, LaToya Bogards, Ashley Leonard, and Marty Price, who scored the essays in this study.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4,* 1–30.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX*. Philadelphia, PA: Linguistic Data Consortium.

Chafe, W. L. (1975). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and topic* (pp. 26–55). New York: Academic.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497–505.

Costerman, J., & Fayol, M. (1997). *Processing interclausal relationships: Studies in production and comprehension of text*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 2,* 213–238.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42,* 475–493.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984–989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *IJCEELL, 21,* 170–191.

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*. Palm Springs, FL: AAAI Press.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5,* 1–35.

Dufty, D. F., Graesser, A. C., Lightman, E., Crossley, S. A., & McNamara, D. S. (July, 2006). *An algorithm for detecting spatial cohesion in text*. Presentation at the 16th Annual Meeting of the Society for Text and Discourse, Minneapolis, MN.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Flowers, L. (1979). Writer-based prose: A cognitive basis for problems in writing. *College English, 41,* 19–37.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. M., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers, 36,* 180–193.

Graesser, A., & McNamara, D. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 2,* 371–398.

Graesser, A., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223–234.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445–476.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English. *Journal of Linguistics, 3,* 199–244.

Halliday, M. A., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Newbury House.

Hempelmann, C. F., Dufty, D., McCarthy, P. M., Graesser, A. C., Cai, Z., & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 941–946). Mahwah, NJ: Erlbaum.

Herskovits, A. (1998). Schematization. In P. Olivier & K. P. Gapp (Eds.), *Representation and processing of spatial expressions* (pp. 149–162). Mahwah, NJ: Lawrence Erlbaum Associates.

Kellogg, R., Whiteford, A., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42,* 173–196.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363–394.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Hand-book of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum.

Lecocke, M., & Hess, K. (2006). An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. *Cancer Informatics, 2*, 313–327.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics, 12,* 291–315.

Longo, B. (1994). Current research in technical communication: The role of metadiscourse in persuasion. *Technical Communication, 41,* 348–352.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19,* 313–330.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, NH: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 381–392.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57–86.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers, 36,* 222–233.

McNamara, D., & Graesser, A. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Lanham, MD: R & L Education.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P. M., & Graesser, A. C. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298–311). Hershey, PA: IGI Global.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47,* 292–330.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics, 21*, 3301–3307.

Pearson, D. (1974–1975). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations. *Reading Research Quarterly*, *10*, 155–192.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford, England: Blackwell.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. A. (1985). *Comprehensive grammar of the English language*. London: Longman.

Roscoe, R. D., Kugler, D., Crossley, S. A., Weston, J. L., & McNamara, D. S. (2012). Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 466–471). Menlo Park, CA: The AAAI Press.

Roscoe, R. D., Varner, L. K., Cai, Z., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2011). Internal usability testing of automated essay feedback in an intelligent writing tutor. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 543–548). Menlo Park, CA: AAAI Press.

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment, 4,* 1–21.

Scott, M. (1996). *WordSmith Tools*. Oxford: Oxford University Press.

Shermis, M., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd ed.). Oxford, UK: Elsevier.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioral Research Methods, Instruments, and Computers, 20,* 6–11.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6,* 292–297.