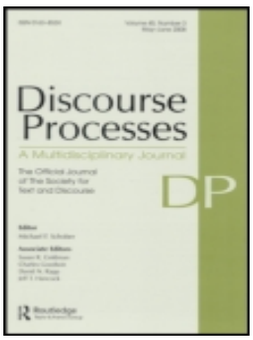

Discourse Processes

ISSN: 0163-853X (Print) 1532-6950 (Online) Journal homepage: http://www.tandfonline.com/loi/hdsp20

Analyzing Discourse Processing Using a Simple Natural Language Processing Tool

Scott A. Crossley, Laura K. Allen, Kristopher Kyle & Danielle S. McNamara

To cite this article: Scott A. Crossley, Laura K. Allen, Kristopher Kyle & Danielle S. McNamara (2014) Analyzing Discourse Processing Using a Simple Natural Language Processing Tool, Discourse Processes, 51:5-6, 511-534, DOI: 10.1080/0163853X.2014.910723

To link to this article: http://dx.doi.org/10.1080/0163853X.2014.910723

Accepted author version posted online: 09 Apr 2014.
Published online: 09 Apr 2014.

Discourse Processes, 51:511–534, 2014

ISSN: 0163-853X print/1532-6950 online
DOI: 10.1080/0163853X.2014.910723

Routledge
Taylor & Francis Group

Analyzing Discourse Processing Using a Simple Natural Language Processing Tool

Scott A. Crossley
Department of Applied Linguistics/ESL
Georgia State University

Laura K. Allen
Department of Psychology, Learning Sciences Institute
Arizona State University

Kristopher Kyle
Department of Applied Linguistics/ESL
Georgia State University

Danielle S. McNamara
Department of Psychology, Learning Sciences Institute
Arizona State University

Natural language processing (NLP) provides a powerful approach for discourse processing researchers. However, there remains a notable degree of hesitation by some researchers to consider using NLP, at least on their own. The purpose of this article is to introduce and make available a *simple* NLP (SiNLP) tool. The overarching goal of the article is to proliferate the use of NLP in discourse processing research. The article also provides an instantiation and empirical evaluation of the linguistic features measured by SiNLP to demonstrate their strength in investigating constructs of interest to the discourse processing community. Although relatively simple, the results of this analysis reveal that the

Correspondence concerning this article should be addressed to Scott A. Crossley, Department of Applied Linguistics/ESL, 34 Peachtree Street, Suite 1200, One Park Tower Building, Georgia State University, Atlanta, GA 30303, USA. E-mail: scrossley@gsu.edu

tool is quite powerful, performing on par with a sophisticated text analysis tool, Coh-Metrix, on a common discourse processing task (i.e., predicting essay scores). Such a tool could prove useful to researchers interested in investigating features of language that affect discourse production and comprehension.

INTRODUCTION

Natural language processing (NLP), or computational linguistics, provides a powerful research approach for discourse processing researchers. In the last decade, NLP has opened research paths that were previously only dreamed of and, in the process, eliminated the need to laboriously tag words, sentences, and texts to painstakingly calculate simple statistics, such as word frequency or readability. NLP approaches have led to an exponential growth in the availability of automated tools, allowing researchers to glean just about any aspect of text, language, or discourse imaginable. We use these tools to explore and better understand language, to test theoretical assumptions, to reinforce experimental studies, and to support natural language dialogue. Indeed, thanks to NLP, we now have a new breed of intelligent tutoring systems that hold conversations with users and provide adaptive feedback on a wide range of content including short answers, long answers, explanations, and essays (Graesser & McNamara, 2012a, 2012b; Graesser, McNamara, & VanLehn, 2005; Roscoe, Varner, Crossley, & McNamara, 2013).

However, although we have made great strides, NLP remains elusive to many. There remains a notable degree of hesitation by some discourse researchers to consider using NLP, at least on their own. NLP seems beyond their own skill sets—an unattainable ability possessed by only a few. Admittedly, developing NLP tools from which to automatically compute linguistic features can be a challenging, time-consuming, and expensive endeavor (McKevitt, Partridge, & Wilks, 1992; Kogut & Holmes, 2001). However, it need not be.

The purpose of this article is to provide discourse processing researchers (and any other brave, linguistically inclined neophytes) with a *simple* NLP (SiNLP) tool. This tool is easy to install and has a user-friendly graphical user interface; SiNLP will rapidly process text (in batches) and is easily expandable so researchers can add their own lexical, rhetorical, semantic, and grammatical categories. The greatest strength of the tool is its simplicity. However, although the tool is simple, it exhibits the ability to measure complex discourse constructs using surface-level linguistic features. Our objective is to introduce and make available this tool to researchers with the overarching goal of proliferating the use of NLP in discourse processing research. It is our hope that this proliferation will help to further our understanding of text and discourse.

SiNLP is written in the Python programming language. We selected Python because it is free, runs on most platforms, allows linguistically useful tasks to be easily accomplished using relatively short scripts, is structured so that small scripts can be combined to create complicated programs, and, as programming languages go, is relatively logical and easy to use. We do not have much space to dedicate to the Python language. Hence, we refer readers to Zelle (2004) and Bird, Klein, and Loper (2009) for further details. In the Method section, we also provide additional information on SiNLP, such as how it computes linguistic features.

We also provide an instantiation and empirical evaluation of the variables provided by SiNLP to demonstrate their strength in investigating constructs of interest to the discourse processing community. For the evaluation, we select a corpus of short essays and aim to predict human judgments of essay quality. We use SiNLP to calculate a small set of linguistic features, which are regressed onto expert ratings of essay quality. We then compare the outcome of using SiNLP to that of Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014), a state of the art NLP tool.

Discourse Processing

Discourse processing researchers are generally interested in examining the processes that underlie the comprehension and production of naturalistic language, such as that found in textbooks, personal narratives, lectures, conversations, and novels. The primary purpose of investigating the linguistic properties found in a text is that such language can provide cues that highlight aspects of the text that listeners and readers should pay attention to and remember (e.g., linguistic features related to coherence that help steer discourse memory construction, Gernsbacher, 1990; Givon, 1992). These cues can range from single words (e.g., connectives) that establish relations among concepts (Sanders & Noordman, 2000) to textual events that establish intentions of the characters and the goals and purpose presented in the text (Zwaan, Langston, & Graesser, 1995).

Accordingly, linguistic processing is a critical component of comprehension at multiple levels of the text. At the surface levels, individuals process the basic lexical and syntactic features of a text and begin to encode the language for its basic meaning (such as understanding specific idea units; Kintsch & van Dijk, 1978). Evidence for the importance of the lexicon in comprehension can be seen in studies that demonstrate that higher frequency words are recognized (Kirsner, 1994) and named (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Forster & Chambers, 1973; Frederiksen & Kroll, 1976) more rapidly than lower frequency words. Additionally, texts with more frequent words are read more

quickly and better comprehended, because frequent words are more easily decoded (Crossley, Greenfield, & McNamara, 2008; Chall & Dale, 1995). Syntactic structure is also related to successful text processing and comprehension, with simpler syntax affording the integration of decoded words into larger syntactic structures, which are necessary for meaning construction (Just & Carpenter, 1987; Rayner & Pollatsek, 1994).

Beyond simple word- and sentence-level features, researchers have examined the broad discourse processes that explain comprehension of an entire text. Investigations of such processes can be used to examine how readers and listeners construct meaning from large segments of texts by developing coherent text representations. Thus, although surface level linguistic features of a text can explain text comprehension, the connections developed among these surface level elements are likely stronger determinants of comprehension (Sparks & Rapp, 2010). A variety of different linguistic cues is available to listeners that operate at the level of discourse to help build coherent text representations. These include linguistic elements that help establish explicit relational information, such as connectives and logical operators (Crossley & McNamara, 2010, 2011; Sanders & Noordman, 2000), features related to anaphoric resolution that help highlight important text elements (Dell, McKoon, & Ratcliff, 1983), and syntactic and semantic features that can distinguish given from new information (Haviland & Clark, 1974; Hempelmann et al., 2005).

Natural Language Processing

NLP involves the automatic extraction of linguistic features such as those discussed above from a text using a computer programming language (Jurafsky & Martin, 2008). In general, NLP focuses on using computers to understand, process, and manipulate natural language text to achieve a variety of objectives. The principle aim of NLP is to gather information on how humans understand and use language through the development of computer programs intended to process and understand language in a manner similar to humans (Crossley, 2013).

NLP techniques have been used in a variety of contexts to understand both discourse comprehension and processing. NLP techniques can be used alone to address discourse processing or in combination with other investigative techniques. For instance, Lintean, Rus, and Azevedo (2012) developed an automatic method for detecting student mental models in the intelligent tutoring system, MetaTutor. In this study, students interacted with the MetaTutor system and generated paragraphs as part of a "prior knowledge activation" activity. These paragraphs were hand coded by humans based on the level of students' understanding of the topic material. NLP techniques and machine learning approaches were then combined to predict these human judgments from the information provided in the paragraph. The results of the study suggest that

NLP techniques can serve as a form of stealth assessment to provide critical information about students' comprehension of complex topics. More recently, Klein and Badia (2014) examined whether creative processing of information could be statistically modeled using NLP techniques. Specifically, they used NLP techniques and a web-based corpus to build a frequency-based method for solving the Remote Associates Test (a test that determines a person's creative potential). The results of this analysis revealed that the NLP techniques outperformed humans on Remote Associates Test items. The findings from the study provide information about the domain generality of the creative process, as well as the role of convergent and divergent thinking for creativity. Together, these studies demonstrate that computational techniques can provide valuable insight into discourse related tasks such as text comprehension and language processing.

A variety of NLP tools has been developed for English that are freely available (or for a fee) and require few to no computer programming skills. The Linguistic Inquiry Word Count (LIWC; http://www.liwc.net) tool developed by Pennebaker, Booth, and Francis (2007) is one such tool. LIWC calculates the percentage of words in a text that are in particular linguistic and psychological categories. Example categories include punctuations (e.g., comma, period), parts of speech (e.g., pronouns, past tense), psychological constructs (e.g., causations, sadness), and personal constructs (e.g., work, religion). LIWC counts the number of words that belong to each word category and provides a proportion score that divides the number of words in the category by the total number of words in the text. Another approach to NLP, Coh-Metrix (http://cohmetrix.memphis.edu), measures textual attributes on a broad profile of language, cohesion, and conceptual characteristics. The system integrates various tools including lexicons (i.e., word lists), pattern classifiers, part-of-speech taggers (Brill, 1995), syntactic parsers (Charniak, 2000), Latent Semantic Analysis (Landauer, McNamara, Dennis, & Kintsch, 2007), and a variety of other components developed in the field of computational linguistics.

Both of these tools can be extremely powerful, capturing a wide range of psychological and linguistic attributes of text such as writing quality (e.g., Crossley & McNamara, 2012; Varner, Roscoe, & McNamara, 2013), text readability (Crossley et al., 2008), and psychological states found in texts (Pennebaker, 2011). Nonetheless, both tools have some limitations. LIWC can be an attractive choice because it provides a wide range of psychologically motivated classes of words. As a cautionary note, however, the user should not take the word classes at face value (i.e., users should carefully examine the list of words contributing to a category). In addition, many of the word classes are populated with only a few relatively uncommon words, leading to non-normal distributions for word counts in shorter texts (i.e., many of the word classes will report zero incidence scores on smaller texts). One positive attribute of LIWC is that it is expandable (i.e., the advanced user can add word categories). However,

the user must pay a small fee to use LWIC. By contrast, Coh-Metrix is provided for no charge (currently). On the negative side, Coh-Metrix is computationally heavy and, thus, slow in processing texts. In addition, the online tool does not allow batch processing, requiring the user to enter texts individually. Finally, and perhaps most importantly, Coh-Metrix is not extendible and, thus, does not allow users to create new linguistic indices to assess text features that may be important to their specific research questions.

Focus of This Study: Essay Quality

Our objective in this article is to introduce a freely available, extendable NLP tool (SiNLP) that can be used to address a wide variety of linguistic questions. In addition, we compare SiNLP to Coh-Metrix in a common discourse processing task that involved text processing, comprehension, and evaluation: the prediction of essay quality by human raters. One motivation for selecting this task is because predicting writing quality may appear more elusive, in the sense that using automated tools to predict writing quality may provide less reliable results when compared with predicting more objective text characteristics such as text readability and difficulty.

A second motivation for our focus on writing is related to our own interests: we have conducted a variety of studies to develop algorithms that predict various aspects of writing quality and to better understand the process of writing (Crossley & McNamara, 2010, 2011; McNamara, Crossley, & McCarthy, 2010; McNamara et al., 2013). McNamara et al. (2010), for example, found that human ratings of essay quality were strongly related to sophisticated language use, such as greater lexical diversity, syntactic complexity, and use of infrequent words. We have also examined relationships between linguistic features of essays to differences between teachers' essay ratings and students' self-assessments of their own writing (Varner et al., 2013). This study found that students' assessments of their own writing were less systematically related to text features and more strongly related to the use of sophisticated vocabulary. Overall, these studies have demonstrated the utility of using NLP tools to explore various aspects of writing and hence provide a useful construct to assess the reliability of SiNLP in an authentic discourse processing task.

METHOD

Our goal is to demonstrate the use of SiNLP to examine discourse processes. Specifically, we use SiNLP to predict and model human judgments of essay quality using linguistic features contained in essays. We then test this tool against a state of the art NLP tool (Coh-Metrix) to compare

differences between the tools and to examine potential benefits of a simple approach to NLP.

Corpus

The target corpus comprises 126 timed (25-minute) essays composed by 126 11th grade students from the Metropolitan District of Columbia area. All essays were written within the Writing Pal, which provides writing strategy instruction to high school and entering college students (McNamara et al., 2012). Essay writing is an essential component of Writing Pal. The system allows students to compose essays and then provides holistic scores and automated, formative feedback based on natural language input. All essays were written in response to a single Scholastic Aptitude Test writing prompt that centered on the benefits of either competition or cooperation. The prompt did not require specific domain knowledge and was intended to relate to a variety of ideas. 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 Scholastic Aptitude Test essay, and because timed prompt-based essays are the primary target of Writing Pal.

Essay Evaluation

Two expert raters with at least 4 years of experience teaching freshman composition courses at a large university rated the quality of the 126 essays in the corpus using a standardized Scholastic Aptitude Test rubric that assesses writing quality (for the rubric, see http://sat.collegeboard.org/scores/sat-essay-scoring-guide). The rubric has been validated in a number of studies (for an overview, see Korbin, Patterson, Shaw, Mattern, & Barbuti, 2008). The rubric generated a holistic quality rating with a minimum score of 1 and a maximum of 6. According to the rubric, higher quality essays are linguistically distinguishable from lower quality essays in that they demonstrate clearer coherence; exhibit more skillful use of language; use more varied, accurate, and apt vocabulary; and contain more meaningful variety in sentence structure. Conceptually, higher quality essays develop better points of views, use better examples, and demonstrate stronger critical thinking. 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 the two raters' responses. Once the correlations between 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 above $r = .70$. 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.

Research Instruments

This section first describes the indices extracted using Coh-Metrix. It then provides a description of SiNLP and the indices calculated using the SiNLP code.

Coh-Metrix

For this analysis, we selected a number of Coh-Metrix indices that have successfully predicted human-rated essay quality in previous studies (e.g., Crossley & McNamara, 2013; Crossley et al., 2013; McNamara et al., 2010, 2013). These indices relate to the number of words, the number of paragraphs, the number of sentences, the number of word types, word frequency, incidence of determiners and demonstratives, incidence of pronouns, lexical diversity, incidence of conjuncts, incidence of connectives, incidence of negations, incidence of modals, and syntactic complexity. These are discussed briefly below in reference to the larger linguistic and discourse features they measure.

Text structure. Coh-Metrix measures a number of text structures, including number of words, number of sentences, and number of paragraphs. Although relatively simple to compute, these indices are extremely powerful and relate to measures of fluency and the development of more sophisticated discourse structure (Dean, 2013; Kellogg, 1988; McCutchen, 1996, 2000).

Vocabulary. Coh-Metrix computes a number of indices related to vocabulary knowledge and use. The two we selected for this study are the number of unique words used (i.e., word types, which relate to vocabulary breadth) and word frequency. Coh-Metrix calculates word frequency using the CELEX database (Baayen et al., 1995), which consists of word frequencies computed from a 17.9-million-word corpus. Word frequency indices measure how often particular words occur in the English language and are important indicators of lexical knowledge and essay quality (McNamara et al., 2010).

Givenness. Given information is information that was previously available in the text and thus not new information. Although Coh-Metrix has an index of givenness that is calculated using perpendicular and parallel Latent Semantic Analysis vectors, we opted for simpler indices of givenness computed by Coh-Metrix' syntactic parser. These indices include part of speech tags for determiners (*a*, *an*, and *the*) and demonstratives (*this, that, these*, and *those*), both of which indicate given information in a text. In terms of discourse comprehension, given information is easier to process than new information (Chafe, 1975; Halliday, 1967).

Anaphor use. Coh-Metrix also uses a syntactic parser to compute the number of pronouns contained in a text. Coh-Metrix provides an overall count for pronouns and individual counts for first-, second-, and third-person pronouns. These indices can be used as a proxy for anaphoric use, because their instances presume there is a previous anaphoric referent. The use of anaphors relates to a text's coherence (Halliday & Hasan, 1976).

Lexical diversity. Lexical diversity indices measure the ratio of types (i.e., unique words occurring in the text) by tokens (i.e., all instances of words) where higher numbers (from 0 to 1) indicate greater lexical diversity (Templin, 1957). Lexical diversity is at a maximum when all words in a text are different or the number of word types is equal to the total number of words (tokens). In that case, the text is likely to be either very low in cohesion (i.e., low word overlap) or very short. By contrast, lexical diversity is lower (and cohesion is higher) when words tend to be used multiple times across a text, providing greater lexical overlap. Coh-Metrix provides several sophisticated lexical diversity indices that control for text length, including *D* (Malvern, Richards, Chipere, & Durán, 2004), which we use in this study.

Connectives. Coh-Metrix calculates the incidence score for connectives in a text through counts of positive (*also*, *moreover*), negative (*however*, *but*), causal (*because*, *so*), contrastive (*although*, *whereas*), additive (*moreover*, *and*), logical (*or*, *and*), and temporal (*first*, *until*) connectives (Halliday & Hasan, 1976; Louwerse, 2001). Coh-Metrix also uses the syntactic parser to calculate the density of negations and contains a conjunct count based on Biber (1988). Connectives are argued to be an important indicator of text coherence (Halliday & Hasan, 1976; Louwerse, 2001).

Future. Coh-Metrix can assess the temporality of the text using a number of features. For this study, we selected an index of future time, which is computed by Coh-Metrix through a modal count retrieved from the syntactic parser. Tense can be an important indicator of rhetorical stance and situational cohesion (Duran, McCarthy, Graesser, & McNamara, 2007).

Syntactic complexity. Coh-Metrix calculates a number of indices that measure syntactic complexity though a parser. For this study, we selected the mean number of words before the main verb, which calculates how long it takes for a reader to arrive at the predicate of the sentence. Indices of syntactic complexity are important indicators of writing quality (McNamara et al., 2010).

SiNLP Tool

We developed a simple NLP tool in Python that automatically computes features similar to those selected by Coh-Metrix but without the need for heavy computation or the use of outside databases. The key element of SiNLP is its simplicity and the power and extendibility found within this simplicity. The tool broadly focuses on indices that count surface level textual and lexical features that have been shown to be good predictors of discourse processing. Although these indices directly measure linguistic features, they only indirectly measure constructs of theoretical interest in discourse processing (i.e., coherence, syntactic complexity, lexical access). Thus, each of these features serves as a proxy for more complex features of language processing and comprehension (e.g., see McNamara et al., 2014). In the sections below, we briefly discuss how to use SiNLP and then discuss the SiNLP features and how they are computed.

To operate the tool, the user should first create a folder that contains the files to be analyzed with the SiNLP tool. These files must be in plain text (.txt) format and must contain no extra hard returns after the last paragraph (this will result in the tool counting an extra paragraph in the text). After this folder has been created, the user should download the SiNLP tool from soletlab.com (under the Tools button). The tool is available in packages for Mac OS, Windows OS, and Linux (Ubuntu). After clicking on the tool icon, a graphical user interface will open (Figure 1).

The graphical user interface first asks users to select a list dictionary. List dictionaries are a key element of SiNLP. These custom lists allow users to create their own custom word lists to analyze texts specific to their research questions. SiNLP list dictionaries may be created in any spreadsheet software that allows for files to be saved as tab-delimited texts (e.g., Excel). SiNLP lists are saved as rows, with the first item in the row as the list name, and all other items in the row as list items. SiNLP dictionaries can include up to 20 lists (rows). Any lists beyond the first 20 are ignored. If users have more than 20 lists, they will be required to create multiple list dictionaries. A SiNLP list can include any combination of single words and/or n-grams, and limited wildcard use is also allowed. For example, wildcard "*" can be added to the beginning and/or end of any list item. Adding this wildcard will match all characters of the list item that precedes or follows the wildcard and then will match any other characters until a word-break is found. For example, if a list included the entry "process*," this list entry would match "process," "processes," and "processing" (and any other words that start with "process"). If the list included the entry "*n't," this list entry would match any contract negations such as "don't," "wouldn't," and "couldn't." The SiNLP package comes with a default list dictionary that includes the categories *determiners*, *demonstratives*, *all pronouns*, *first-person pronouns*, *second-person pronouns*, *third-person pronouns*, *conjuncts*, *connectives*, *negation*, and *future*. These categories are discussed in greater detail below.

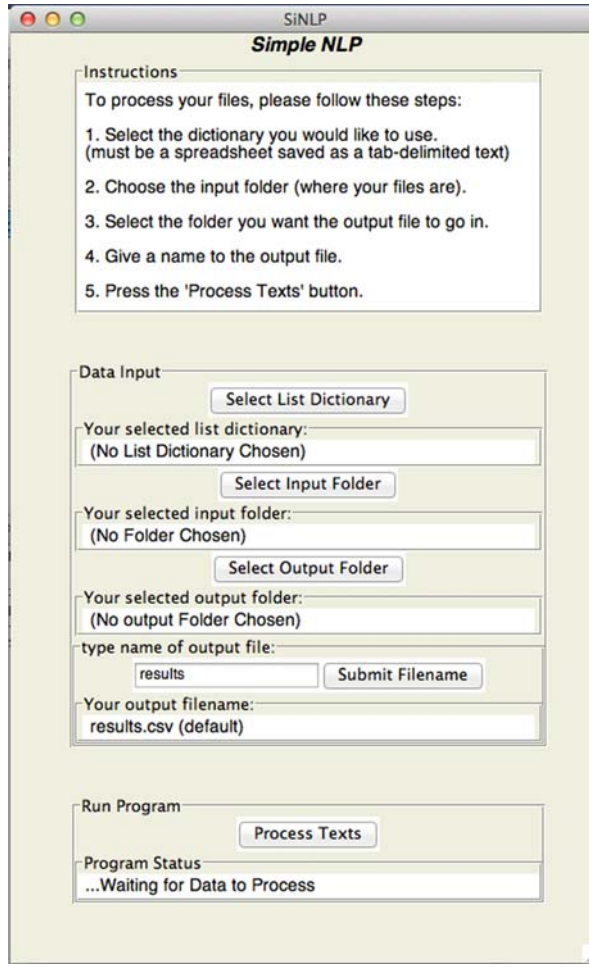

FIGURE 1 Screenshot of SiNLP graphical user interface.

After selecting a list dictionary, users select the input (text) files through a file-dialog prompt, choose where they want the output file to be saved, and click a button to process the texts. While processing, a status field displays the name of the file being processed in real time, giving the user an indication of how quickly the program is working. After processing the texts, SiNLP saves a comma-separated values file (.csv) that can be opened in any spreadsheet software program. The output includes the filename for each text file analyzed, the number of words, the number of word types, a type token ratio score, the number of letters per word, the

number of paragraphs, the number of sentences, the number of words per sentence, and a normalized count of the instances of each custom list in each text file (i.e., the instance of the words in the list divided by the number of words in the text). This format can be easily uploaded to statistical packages such as SPSS or freely available machine learning packages such as the Waikato Environment for Knowledge Analysis, or WEKA. All categories reported by SiNLP and the manner in which they are calculated in Python are discussed below.

Text structure. Like Coh-Metrix, SiNLP measures text structures including number of words, number of sentences, and number of paragraphs. SiNLP calculates the number of words by first assigning a simple count function to a variable (a count variable). This count function is set to 0. SiNLP then tokenizes the text by splitting the text based on whitespace characters. Thus, the sentence *John saw the cat* would be split at each instance of a white space leading to the four-word list [John, saw, the, cat]. SiNLP then runs a loop through each word in the list and adds a count of 1 to the count variable. Word count indices tap into discourse concepts such as fluency, propositional density, and lexical accessibility (Kintsch & Keenan, 1973; McNamara et al., 2013). SiNLP counts the number of paragraphs in a similar fashion to number of words but computes paragraphs based on the occurrence of new lines in the text. Paragraph indices provide evidence of text structuring.

SiNLP computes the number of sentences by using a regular expression to find all the instances of "*;*," "*?,*" "*!,*" and "." and counts their occurrence. SiNLP also uses regular expressions to "not count" those instances of punctuation that occur between capital letters (i.e., it will not count punctuations in acronyms like "U. S.A."). The occurrence of these punctuation markers is then assigned to a variable, which provides a count for the number of sentences in the text. Like word count indices, sentence count indices can tap into fluency and propositional density.

Vocabulary. SiNLP counts the number of unique tokens in a text to provide a measure of vocabulary use and computes a proxy measure for word frequency (i.e., number of letters per word). For the number of types, SiNLP counts how many unique items there are in a list. Such an index can provide information about fluency, vocabulary breadth, and lexical accessibility (Crossley & McNamara, 2013). For the word frequency proxy, SiNLP measures the number of letters per word with the presumption that less frequent words are longer and more frequent words are shorter. Such an index can provide a measure of lexical richness (Flesch, 1948). After counting the number of sentence ending punctuation characters and the number of words, SiNLP deletes all the punctuation from the text (so that punctuation is not counted as a letter). For each word in the list of words, the number of characters are counted. The total number of characters is then divided by the number of words to provide a count of the number of letters per word.

Givenness. SiNLP assesses givenness by computing the incidence of determiners and demonstratives in a text. A greater use of determiners and demonstratives should equate to more givenness in a text, which is an indicator of text cohesion (Gundel, Hedberg, & Zacharski, 1993). Two variables are first assigned a list that contains the English determiners and demonstratives, respectively. SiNLP iterates through all the words in the text and then establishes if the determiners and the demonstratives in the lists are in the word list. For each determiner and demonstrative in the word list, a count variable is increased by one.

Anaphor use. SiNLP assesses the use of anaphors in a text by counting the incidence of all pronouns and, separately, the occurrence of first-, second-, and third-person pronouns. The programming is exactly like that for counting determiners, except the lists are different. Anaphor use can also indicate text cohesion (Halliday & Hasan, 1976).

Lexical diversity. SiNLP computes lexical diversity by dividing the number of types (i.e., unique words in the text) by the number of tokens (i.e., all words in the text) providing a type-token ratio, which relates to word overlap and can be used as a measure of text cohesion (McCarthy & Jarvis, 2010).

Connectives and conjuncts. SiNLP counts the number of connectives and the number of conjuncts in a text using lists in the exact manner used to calculate determiners, demonstratives, and pronouns. The list of connectives contains coordinating connectives that can be used to join independent clauses. Conjuncts in this list serve to connect sentences with previous parts of the text and are based on work by Biber (1988). Both connectives and conjuncts are strong indicators of text cohesion (Halliday & Hasan, 1976).

Future. SiNLP assesses the use of future words by using a list that contains both future words (e.g., could, may) and a wildcard expression to count the incidence of words that end in *'ll*. The use of future time words can indicate temporality and syntactic complexity (i.e., the use of modal verbs increase verb complexity and changes verb inflection; Tomasello, 1992).

Syntactic complexity. SiNLP includes two additional measures that map onto syntactic complexity. The first is the number of words per sentence (calculated using previous counts of words and sentences), which is a proxy for the number of embedded sentences and adjuncts per sentence (Flesch, 1948). The second is the number of negations, which require syntactic transformations that increase the syntactic complexity of a sentence (Chomsky, 1957). The

number of negations is calculated exactly the same as incidence of future terms, but the wildcard expression searches for *n't* instead of *'ll* and the list of words related to negations is different (e.g., none, no).

RESULTS

Statistical Analysis

For each tool, we first conducted correlations between the indices and the human judgments of essay quality. Second, we checked for multi-collinearity among the indices. Finally, we then conducted a regression analysis using 10-fold cross-validation techniques to predict the essay quality of the 126 essays in our corpus using the indices that demonstrate at least a small effect size with essay quality and do not exhibit multi-collinearity with other indices. In 10-fold cross-validation, the data (in this case the 126 essays) were split into 10 subsets. Nine of these subsets were used to develop a regression model that was then tested on the left-out subset. This process was repeated 10 times so all data were 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 10-fold 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).

Coh-Metrix Analysis

Pearson Correlations

Correlations between the 17 indices calculated from Coh-Metrix and the human ratings of essay quality demonstrated 8 significant correlations and 13 correlations that showed at least a small effect size ($r > .010$; Cohen, 1988) (Table 1). Checks for multi-collinearity showed that two indices (*number of word types* and *number of words*) were correlated at above $r > .90$. Because *number of word types* showed a stronger correlation with the dependent variable, it was retained and *number of words* was removed from the subsequent regression analysis.

Regression Analysis

We included all indices that demonstrated at least a small effect size as independent variables in the regression analysis. These variables were used to predict the human scores for the 126 essays using 10-fold cross-validation techniques. The linear regression reported $r = .548$, $r^2 = .300$. Five variables were included in the model: *number of word types*, *CELEX frequency*, *part of speech*, *second-person pronouns*, *density of negations*, and *number of sentences*. The model from the regression is as follows:

TABLE 1
Correlations Between Essay Scores and Coh-Metrix Indices

Index	*r*	*p*
Number of word types	.561	< .001
Number of sentences	.530	< .001
Number of words	.517	< .001
Number of paragraphs	.388	< .001
Incidence of conjuncts	.256	< .01
Density of negations	.224	< .05
Part of speech: second-person pronouns	− .208	< .05
Lexical diversity *D*	.197	< .05
Part of speech: pronouns	− .154	> .05
CELEX word frequency	− .147	> .05
Part of speech: modals	− .118	> .05
Number of words before the main verb	.111	> .05
Part of speech: determiners	.108	> .05
Part of speech: demonstratives	.099	> .05
Incidence of connectives	− .067	> .05
Part of speech: third-person pronouns	.052	> .05
Part of speech: first-person pronouns	.004	> .05

$$\begin{aligned}\text{Predicted essay score} = {} & 2.270 + (.005 \times \textit{number of word types}) \\ & + (-.377 \times \textit{CELEX frequency}) \\ & + (-.006 \times \textit{part of speech: second-person pronouns}) \\ & + (.009 \times \textit{density of negations}) \\ & + (.009 \times \textit{number of sentences})\end{aligned}$$

Exact and Adjacent Matches

We used the scores derived from the 10-fold cross-validated regression to assess the exact and adjacent accuracy of the regression scores when compared to the human-assigned scores. The regression model produced exact matches between the predicted essay scores and the human scores for 77 of the 126 of the essays (61% exact accuracy). The model produced exact and adjacent matches for 125 of the 126 essays (99% adjacent accuracy).

SiNLP Analysis

Pearson Correlations

Correlations between the 17 indices calculated from the SiNLP tool and the human scores demonstrated 7 significant correlations and 12 correlations that showed at least a small effect size ($r > .010$; Cohen, 1988) (Table 2). Checks for

TABLE 2
Correlations Between Essay Scores and SiNLP Indices

Index	*r*	*p*
Number of word types	.573	<.001
Number of sentences	.531	<.001
Number of words	.517	<.001
Number of paragraphs	.388	<.001
Letters per word	.231	<.01
Incidence of negations	.226	<.05
Incidence of second-person pronouns	−.210	<.05
Type-token ratio	−.164	>.05
Number of words per sentence	−.149	>.05
Number of pronouns	−.146	>.05
Incidence of demonstratives	.144	>.05
Incidence of future words	−.139	>.05
Incidence of conjuncts	.085	>.05
Incidence of determiners	.081	>.05
Incidence of third-person pronouns	.076	>.05
Incidence of connectives	.055	>.05
Incidence of first-person pronouns	.007	>.05

multi-collinearity showed that two indices (*number of word types* and *number of words*) were correlated at above $r > .90$. Because *number of word types* showed a stronger correlation with the dependent variable, it was retained and *number of words* was removed from the subsequent regression analysis.

Regression Analysis

We included all indices that demonstrated at least a small effect size as independent variables in the regression analysis. These variables were used to predict the human scores for the 126 essays using 10-fold cross-validation techniques. The linear regression reported $r = .563, r^2 = .316$. Four variables were included in the model: *number of word types*, *letters per word*, *number of paragraphs*, and *incidence of negations*. The model from the regression is as follows:

$$\begin{aligned}\text{Predicted essay score} = &-.569 + (0.008 \times \textit{number of word types}) \\ &+ (.439 \times \textit{letters per word}) \\ &+ (.008 \times \textit{number of paragraphs}) \\ &+ (.008 \times \textit{incidence of negations})\end{aligned}$$

Exact and Adjacent Matches

We used the scores derived from the 10-fold 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 83 of the 126 of the essays (66% exact accuracy). The model produced exact and adjacent matches for 126 of the 126 essays (100% adjacent accuracy).

DISCUSSION

In this article we present a simple NLP tool (SiNLP) that can be used by discourse processing researchers to investigate text comprehension and production processes. The tool is available in packages that run on Mac, Windows, and Linux operating systems and is easy to operate, fast, capable of batch processing, and extendible. Although relatively simple, the tool is also quite powerful, performing on par with the Coh-Metrix tool in predicting essay scores for the corpus of essays used in this study. Such a tool could prove useful to researchers interested in investigating features of language that affect the production and comprehension of language.

In terms of efficiency, SiNLP performed as well as, and potentially better than, Coh-Metrix in its prediction of essay scores for the corpus in this study. An obvious caveat is that we selected only a limited number of indices from Coh-Metrix, which were directly comparable with the indices computed by SiNLP. In all likelihood, an analysis using the full breadth of indices reported by Coh-Metrix would provide equivalent results to those provided by SiNLP (although previous studies using Coh-Metrix on different corpora have not reported such high accuracies; McNamara et al., 2013). Nevertheless, this study has demonstrated compatibility between the two tools as well as the efficacy of simple automated indices to assess writing quality. The model derived from the SiNLP indices provided exact matches for 66% of the essays and adjacent matches for 100% of the essays. These percentages were slightly higher than the model based on the Coh-Metrix indices (61% and 99%, respectively), and both models were in line with, or slightly higher than, exact and adjacent models reported by established automated essay scoring systems, such as e-rater (Attali & Burstein, 2006; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012), IntelliMetric (Rudner, Garcia, & Welch, 2005, 2006), Coh-Metrix, and the Writing Pal (McNamara et al., 2013).

As in previous research, vocabulary breadth (as measured by the number of types in both SiNLP and Coh-Metrix), word frequency (as measured in SiNLP by letters per word and in Coh-Metrix by CELEX), and essay structure (as measured by number of paragraphs in SiNLP and number of sentences in Coh-Metrix) were strong predictors of essay quality, with higher quality essays containing more word types, more infrequent words, and more paragraphs/sentences. These indices indicate that higher rated essays are more difficult to process. Further, the

linguistic complexity that drives this text processing difficulty is at least partially responsible for the scores assigned to the essay. Although these indices may not overlap with domain knowledge, cultural and background knowledge, and variation in rhetorical purposes (all additional key elements of essay quality), they are potentially indicative of a writer's ability to quickly and easily produce complex text, freeing up cognitive resources that can be used to address rhetorical and conceptual concerns in the text (Deane, 2013).

For both SiNLP and Coh-Metrix models, the number of negations was also a positive predictor of essay quality, indicating that greater use of negative terms is a strong predictor of essay quality. The use of second-person pronouns was also a negative predictor of essay quality in the Coh-Metrix model, indicating that addressing the reader through pronominal references leads to lower essay scores.

Our writing quality analysis provided evidence for the usability and validity of the SiNLP tool in reference to discourse processing tasks. Unlike Coh-Metrix (and other NLP tools), SiNLP has other benefits in terms of efficiency, ease of use, and extendibility. From an efficiency standpoint, SiNLP is extremely fast and computationally light. It could take well over a day (and most likely much longer) to calculate linguistic features for the 126 essays sampled in this study using the online Coh-Metrix version. On the other hand, SiNLP can process the same texts in less than a minute (although SiNLP provides many fewer indices). Additionally, because the tool is housed on the user's computer hard drive, the user does not need to depend on an Internet connection or servers and does not have to queue in line behind other users. Another important feature of SiNLP is that it allows for batch processing, so users can process thousands of texts concurrently without the need to return to a website and enter texts one by one (as with Coh-Metrix). Finally, and perhaps most importantly, SiNLP allows users to add in linguistic categories that are of interest to them to extend the tool to meet their personal research needs.

It is the extendibility of the tool that is likely its most powerful feature. By design, the tool currently reports on a limited number of linguistic features. However, researchers have the capacity to adapt the tool to suit their own needs, especially in circumstances in which the variable of interest is text based. For instance, researchers interested in investigating how text information can help construct a coherent situation model may want to develop new categories for casuality, spatiality, temporality, and intentionality. In other cases, researchers may be interested in assessing differences in stimuli for psycholinguistic studies, text recall data, or transcribed conversational data. In each case, the user would only need to create a new dictionary containing the selected category terms in a spreadsheet program and then save the dictionary into a tab-delimited.txt file. The user would then load this dictionary using the SiNLP program and process the texts of interest.

As examples, indices for causality could be based on counts of causal verbs, causal particles, or a combination of both. Indices of spatiality could be based on

locational prepositions and common spatial terms (e.g., terms related to size and shape of objects). Temporal indices could be based on the use of common present and past tense verbs and intentionality indices could be informed through the use of words related to certainty and tentativeness. In addition to situation model indices, some researchers may be interested in accounting for the role of emotions in discourse processing (cf. D'Mello & Graesser, 2012) under the premise that texts and conversations often create emotional responses on the part of the reader/listener. Thus, the investigation and detection of emotional language use can provide additional information about the properties of discourse that affect language processing (Sparks & Rapp, 2010). In this case, a tool such as SiNLP could be extended to included affective categories developed by the researcher. These categories could be simple categories of positive and negative emotional words or more specific based on the research question asked. In essence, the value of SiNLP is that the user can create as many semantic, grammatical, and rhetorical categories that they can theoretically conceive.

Similarly, consider the discourse area of text readability. How might researchers in text readability use SiNLP to investigate readability from a strictly NLP approach? For instance, let's say that a researcher's objective is to develop indices to predict readability scores on a series of passages (e.g., the Bormuth text passages; Bormuth, 1969). The researcher, following the hypothesis that text readability mainly comprises word decoding, syntactic parsing, and meaning construction (Just & Carpenter, 1987; Rayner & Pollatsek, 1994), may first develop a series of lists related to decoding (in addition to using the SiNLP indices *type counts* and *number of letters per word*). These lists, most likely derived from corpus analyses, may include the most frequent content words in English, a list of function words (frequent and infrequent), academic word lists (Coxhead, 2000), or even academic formula lists (Simpson-Vlach & Ellis, 2010). Next, the researcher may develop lists that are related to syntactic complexity (in addition to using the SiNLP calculation for *number of words in a sentence*). These lists would likely revolve around syntactic transformations that make a sentence more difficult to process such as incidence of negations, wh-words, clausal subordinators, or clausal coordinators. Next, the researcher would develop lists that are indicative of meaning construction (i.e., how idea units are linked). Such lists might include the incidence of connectives, conjuncts, or situational model features such as frequent present and past verbs, spatial and movement prepositions, or causal particles. Finally, the researcher would use SiNLP to collect incidence scores for these categories and use machine learning techniques to investigate how predictive these scores are of text processing. Although the process seems simple, the researcher would (optimally) rely on theoretical and empirical evidence to populate the lists and would ensure that the lists provide sufficient lexical coverage of the assessed texts to avoid non-normal distributions for the dictionary counts (a problem commonly found in some LIWC dictionaries). In sum, a tool such as

SiNLP affords researchers the opportunity to explore theoretically and empirically driven questions concerning text and discourse.

Notably, SiNLP can also be used to analyze text that is written in languages other than English. One of the most common questions regarding Coh-Metrix (other than why it is not working) is whether it can be used to analyze other languages. Although Coh-Metrix tools have been developed to analyze some other languages (e.g., Portuguese and Chinese), the online Coh-Metrix tool is limited to the English language. Moreover, building a complete Coh-Metrix for other languages including all of its components is a challenging, time-consuming, and expensive endeavor. In comparison, SiNLP can be easily modified for other languages simply by replacing the predefined lists with appropriate defined lists in another language. The other linguistic features reported by SiNLP (e.g., number of words, number of types, sentence length, etc.) would provide correct counts for languages other than English as well. It is imperative that NLP tools are available to investigate cross-language differences and explore linguistic and discourse questions in multiple languages. SiNLP provides such a tool.

However, caution should be used when interpreting results from SiNLP analyses. Although the tool potentially provide indices that enhance our understanding of text and discourse processing and comprehension, most indices do not address these constructs directly. Rather, any indices provided by a tool such as SiNLP map onto linguistic constructs indirectly as proxies. For instance, the number of words in a sentence is a proxy for measuring syntactic complexity, as is the number of conjuncts in a text for measuring text cohesion (see McNamara et al., 2014). Both indices map onto the linguistic constructs of interest, but do not provide a complete picture of the constructs' intricacy.

Overall, SiNLP can provide discourse researchers with an additional tool with which to identify and examine the mental representations and processes involved in the production and comprehension of language. Because SiNLP can provide textual information at the discourse level, we are hopeful that its use will forward discourse processing studies by providing a more reliable and fine-grained analysis of the linguistic properties that affect the text comprehension and production processes.

ACKNOWLEDGMENTS

We thank Mark Johnson for providing an early draft of his book, *Essential Python for Corpus Linguistics*, to be used in an NLP class at Georgia State University. The class and the book acted as an inspiration for the development of SiNLP. We also thank Mary Sellers for her design of the SiNLP icon.

FUNDING

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.

REFERENCES

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2.0. *Journal of Technology, Learning and Assessment*, *4*, 1–31.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media. Retrieved from http://nltk.org/book

Bormuth, J. R. (1969). *Development of readability analyses* (Final Report, Project No. 7-0052, Contract No. 1, OEC-3-7-070052-0326). Washington, DC: U.S. Office of Education.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, *21*, 543–566.

Chafe, W. (1975). Givenness, contrastiveness, definiteness, subjects, topics, and point of view in subject and topic. In C. Li (Ed.), *Subject and topic* (pp. 25–55). New York, NY: Academic Press.

Chall, J., & Dale, E. (1995). *Readability revisited: The new DaleChall Readability*. Cambridge, MA: Brookline Books.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the First Conference on North American Chapter of the Association for Computational Linguistics* (pp. 132–139). San Francisco, CA: Morgan Kaufmann.

Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands/Paris, France: Mouton.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, *46*, 256–271.

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. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21*, 170–191.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, *35*, 115–135.

Crossley, S. A., & McNamara, D. S. (2013). Text analysis tools for spoken response grading. *Language Learning & Technology*, *17*, 171–192.

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 208–213). Menlo Park, CA: The AAAI Press.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7–24.

Dell, G. S., McKoon, G., & Ratcliff, R. (1983). The activation of antecedent information during the processing of anaphoric reference in reading. *Journal of Verbal Learning and Verbal Learning and Verbal Behavior*, *22*, 121–132.

D'Mello, S. K., & Graesser, A. C. (2012). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, *5*, 304–317.

Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods*, *39*, 212–223.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635.

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 361–379.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Givon, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, *30*, 5–55.

Graesser, A. C., & McNamara, D. S. (2012a). Automated analysis of essays and open-ended verbal responses. In H. Cooper, P. Camic, R. Gonzalez, D. Long, & A. Panter (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (pp. 307–325). Washington, DC: American Psychological Association.

Graesser, A. C., & McNamara, D. S. (2012b). Reading instruction: Technology-based supports for classroom instruction. In C. Dede & J. Richards (Eds.), *Digital teaching platforms: Customizing classroom learning for each student* (pp. 71–87). New York, NY: Teachers College Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, *36*, 193–202.

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist*, *40*, 225–234.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307.

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

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

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*, 512–521.

Hempelmann, C. F., Dufty, D., McCarthy, P., 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. Bara (Ed.), *Proceedings of the 27th Annual Meetings of the Cognitive Science Society* (pp. 941–946). Mahwah, NJ: Erlbaum.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing*. Englewood, NJ: Prentice Hall.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston, MA: Allyn & Bacon.

Kellogg, R. T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 355–365.

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, *5*, 257–274.

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

Kirsner, K. (1994). Implicit processes in second language learning. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283–312). San Diego, CA: Academic Press.

Klein, A., & Badia, T. (2014). The usual and the unusual: Solving remote associates test tasks using simple statistical natural language processing based on language use. *Journal of Creative Behavior*. Advance online publication. doi:10.1002/jocb.57

Kogut, P., & Holmes, W. (2001). AeroDAML: Applying information extraction to generate DAML annotations from web pages. In S. Handschuh, R. Dieng, & S. Stabb (Eds.), *Proceedings of the K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation*. Aachen, Germany: CEUR-WS.

Korbin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average*. New York, NY: The College Board.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: 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.

Lintean, M., Rus, V., & Azevedo, R. (2012). Automatic detection of student mental models based on natural language student input during metacognitive skill training. *International Journal of Artificial Intelligence in Education*, *21*, 169–190.

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

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

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*, 299–325.

McCutchen, D. (2000). Knowledge processing and working memory: Implications for a theory of writing. *Educational Psychologist*, *35*, 13–23.

McKevitt, P., Partridge, D., & Wilks, Y. (1992). Approaches to natural language discourse processing. *Artificial Intelligence Review*, *6*, 333–364.

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

McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499–515.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188–205). Hershey, PA: IGI Global.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., Dai, J., … Graesser, A. C. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298–311). Hershey, PA: IGI Global.

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

Pennebaker, J. W. (2011). *The secret life of pronouns*. New York, NY: Bloomsbury Press.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin, TX: LIWC.net.

Ramineni, C., Trapani, C. S., Williamson, D. M. W., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the TOEFL® independent and integrated prompts* (ETS Research Report No. RR-12-06). Princeton, NJ: ETS.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology*, *8*(4), 362–381.

Rudner, L., Garcia, V., & Welch, C. (2005). *An evaluation of Intellimetric™ essay scoring system using responses to GMAT® AWA prompts*. McLean, VA: GMAC.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric™ essay scoring system. *Journal of Technology, Learning, and Assessment*, *4*(4), 1–22.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*, 37–60.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics (Oxford)*, *31*, 487–512.

Sparks, J. R., & Rapp, D. N. (2010). Discourse processing—Examining our everyday language experiences. *WIREs Cognitive Science*, *1*, 371–381.

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis, MN: University of Minnesota Press.

Tomasello, M. (1992). *First verbs: A case study of early lexical development*. Cambridge, UK: Cambridge University Press.

Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research*, *5*, 35–59.

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

Zelle, J. M. (2004). *Python programming: An introduction to computer science*. Wilsonville, OR: Franklin Beedle.

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.