Behav Res (2016) 48:1227–1237
DOI 10.3758/s13428-015-0651-7

The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion

Scott A. Crossley[1] · Kristopher Kyle[2] · Danielle S. McNamara[2]

Published online: 28 September 2015

Abstract This study introduces the Tool for the Automatic Analysis of Cohesion (TAACO), a freely available text analysis tool that is easy to use, works on most operating systems (Windows, Mac, and Linux), is housed on a user's hard drive (rather than having an Internet interface), allows for the batch processing of text files, and incorporates over 150 classic and recently developed indices related to text cohesion. The study validates TAACO by investigating how its indices related to local, global, and overall text cohesion can predict expert judgments of text coherence and essay quality. The findings of this study provide predictive validation of TAACO and support the notion that expert judgments of text coherence and quality are either negatively correlated or not predicted by local and overall text cohesion indices, but are positively predicted by global indices of cohesion. Combined, these findings provide supporting evidence that coherence for expert raters is a property of global cohesion and not of local cohesion, and that expert ratings of text quality are positively related to global cohesion.

Keywords Cohesion · Coherence · Natural language processing · Text difficulty · Writing quality

Cohesion is a crucial element for understanding texts, particularly with challenging texts that present knowledge demands to the reader (Loxterman, Beck, & McKeown, 1994; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996). Hence, measuring cohesion is an important element of discourse-processing research (McNamara et al. 2010a, b). However, freely available natural language processing (NLP) tools that measure linguistic features related to text cohesion are limited. The best-known example is likely Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014), an online tool that measures a number of linguistic features related to lexical sophistication, syntactic complexity, and cohesion. Although Coh-Metrix has been extremely useful and has had a large impact on our understanding of language and discourse, it has several shortcomings with regard to usability and to the facile and broad measurements of its cohesion indices. First, the version made available to the public does not allow for the batch processing of text and is not housed on a user's hard drive (and thus it is dependent on an Internet connection and an external server). Second, the Coh-Metrix cohesion indices generally focus solely on local and overall text cohesion (in contrast to global cohesion), and the publicly available tool includes a limited number of cohesion indices (25, at the time of writing).

This article introduces a new text cohesion analysis tool called the Tool for the Automatic Analysis of Cohesion (TAACO). TAACO is a freely available text analysis tool that is easy to use, works on most operating systems (Windows, Mac, and Linux), is housed on a user's hard drive (rather than having an Internet interface), allows for the batch processing of text files, and incorporates over 150 classic and recently developed indices related to text cohesion. The cohesion indices reported by TAACO evenly focus on local cohesion, global cohesion, and overall text cohesion. Local cohesion refers to cohesion at the sentence level (i.e., cohesion between smaller chunks of text such as noun overlap between sentences or linking sentences through connectives); global cohesion refers to cohesion between larger chunks of text such as paragraphs (e.g., noun overlap between paragraphs in a text); and overall

✉ Scott A. Crossley
sacrossley@gmail.com

[1] Department of Linguistics, Georgia State University, Atlanta, GA, USA

[2] Learning Sciences Institute/Psychology, Arizona State University, Tempe, AZ, USA

text cohesion refers to the incidence of cohesion features in an entire text, but not in comparisons of parts of the text (e.g., lexical diversity, which is calculated as the repetition of words across a text).

In this study, we demonstrate the utility of the cohesion indices provided by TAACO, with a focus on the domain of writing, specifically on persuasive essays. We examine the degree to which the linguistic features of essays related to cohesion predict expert judgments of text organization (i.e., our measure of text coherence) and writing quality. To do this, we collected a corpus of persuasive essays written by college freshmen. The essays were then scored by expert raters in terms of their overall text coherence and writing quality. These expert scores were used as the criteria for assessing the utility of the TAACO indices by examining how the local, global, and overall text cohesion devices reported by TAACO were differentially related to human judgments of text coherence and essay quality. The analyses conducted in this study allow us not only to introduce TAACO and validate the tool (i.e., by testing its predictive validity in assessing human ratings of coherence), but also to further examine the relations between different aspects of cohesion with human ratings of essay quality.

Although it is generally assumed that cohesion is important to writing quality (Collins, 1998; DeVillez, 2003), research on the effects of cohesion on human ratings of essay quality has reported that not all types of cohesion features are uniformly related to essay quality or text coherence. Indeed, the differences among cohesion types across writing and coherence studies were what motivated us to develop TAACO. These differences concern the relations between local, global, and overall text cohesion and human ratings of essay quality and text coherence. For instance, some research has suggested that the use of local cohesion devices may have little to no relation to expert ratings of text quality (McNamara et al. 2013; McNamara et al. 2010a, b), and some cohesion research has even yielded negative correlations to writing quality (Crossley & McNamara, 2010, 2011, 2012; Crossley et al. 2011a, b). In terms of global cohesion and overall text cohesion, the available research is insufficient to assess links between these types of cohesion and expert ratings of text quality, primarily because tools to assess these aspects of cohesion have not been widely available. However, research has shown positive relations between expert raters' judgments of text coherence and text quality (Crossley & McNamara, 2010, 2011) and between expert judgments of text coherence and global, but not local, cohesion cues (Crossley & McNamara, 2010, 2011).

These findings overlap, to a degree, with research on the effects of overall text cohesion on text readability. A number of studies have shown that the benefits of local and overall text cohesion may be limited to low-knowledge readers (McNamara & Kintsch, 1996; McNamara et al., 1996) because these types of cohesion help such readers bridge gaps in their background knowledge. In contrast, for high-knowledge readers, the absence of local cohesion cues in a text can prompt the generation of inferences that connect ideas in the text and to the reader's prior knowledge, thus enhancing text comprehension (McNamara & Kintsch, 1996; McNamara et al., 1996). In the case of high-knowledge readers, it may be that global cohesion cues linking paragraphs also benefit text comprehension. Hence, the judgments of text coherence and writing quality given by expert raters (i.e., high-knowledge readers), such as the ones in this study, may be representative of global cohesion in the text and not of local or overall text cohesion.

Cohesion and coherence

Crucial to the measurement of cohesion is the theoretical distinction between cohesion and coherence. Cohesion generally refers to the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text. Examples of these explicit cues include overlapping words and concepts between sentences. These cues communicate to the reader that the same or similar ideas are being referred to across consecutive sentences. In addition, connectives such as *because*, *therefore*, and *consequently* act as explicit cues that inform the reader that there are relations between ideas and the nature of those relations (Halliday & Hasan, 1976). Thus, most linguistic devices associated with text cohesion are local in nature (i.e., cohesion at the sentence level). However, global cohesion devices are also indicative of text cohesion, although such devices are often more implicit. These devices may include causal relations throughout a text (Graesser, McNamara, Louwerse, & Cai, 2004) and semantic similarity between paragraphs in a text (Foltz, 2007). Less research has been conducted on overall text cohesion features, which are measured across an entire text and are not specific to the sentence or paragraph levels.

Coherence, as compared to cohesion, refers to the understanding that the reader derives from the text (i.e., the coherence of the text in the mind of the reader; McNamara et al., 1996; O'Reilly & McNamara, 2007). A coherent text also matches the expectations of the reader and/or "sticks to the point" (Johns, 1986). This coherence depends on a number of factors, including explicit cohesion cues, implicit cohesion cues (which are more closely linked to text coherence than are explicit cues), and nonlinguistic factors such as prior knowledge and reading skill (McNamara et al., 1996; O'Reilly & McNamara, 2007). Whereas coherence is the quality of the reader's mental representation, comprehension is the observed outcome of that representation. In turn, the effects of the representation depend on the measure used to assess comprehension. For example, measures that rely primarily on surface information from words and sentences (e.g., recognition, multiple choice) depend less on the coherence of the reader's mental representation than do measures that tap into relations between ideas in the text (e.g., inference questions; McNamara & Kintsch, 1996).

Text cohesion and text quality

Another important area of research involves investigating links between cohesion devices in the text and human judgments of text features and text quality. Text features can be measured through judgments of overall text coherence, and text quality can be measured through a holistic score assigned to an essay. In two recent studies, Crossley and McNamara (2010, 2011) examined the degree to which judgments of text coherence were predicted by automated indices of local cohesion reported by the computational tool Coh-Metrix (McNamara et al., 2014). Crossley and McNamara (2010) reported that human judgments of coherence were strongly correlated with human judgments of essay quality (*r*= .80). However, only a few cohesion indices calculated at the local and text levels demonstrated significant correlations with human ratings of coherence (e.g., anaphoric reference, causal cohesion, connectives, and lexical overlap); these indices correlated negatively, and generally with low effect sizes (i.e., *r*< .30). A follow-up study (Crossley & McNamara, 2011) reported similar results; however, global indices of cohesion that calculated overlap between initial, middle, and final paragraphs measured by the Writing Assessment Tool (WAT: Crossley, Roscoe, & McNamara, 2013) were positively correlated at around *r*= .30 with judgments of text coherence. These two studies provide some support for the assumption that, for expert raters, judgments of coherence are not likely a result of local or overall text cohesion devices, but rather of global cohesion devices.

A number of studies have also assessed the degree to which cohesion devices are predictive of writing quality in general. Using Coh-Metrix, McNamara et al. (2010a, b) examined the role that linguistic features (including cohesion devices) play in predicting independent essay quality (i.e., the quality of a writing sample that requires no specific background knowledge). Their results indicated that no local cohesion indices showed significant differences between low- and high-scored essays, and that no indices categorized as being locally cohesive significantly correlated with the essay scores. A follow-up study (Crossley et al. 2011a, b) examined essay scores assigned by expert raters, using linguistic features reported by Coh-Metrix and WAT. The findings indicated that two indices of global cohesion (semantic similarity between initial and middle paragraphs, and semantic similarity between initial and final paragraphs) and one index of overall text cohesion (text givenness) significantly correlated with essay quality. However, the effect sizes for these correlations were small, and only text givenness was included in a regression model that predicted essay quality.

Similar findings have been reported for studies that have focused on predicting second language (L2) writing. For instance, in a study that examined the writing quality of essays produced by Hong Kong high school students, Crossley and McNamara (2012) found that the local and overall text cohesion devices reported by Coh-Metrix, such as content word overlap, positive logical connectives, aspect repetition, and semantic similarity between sentences, were negatively correlated with expert ratings of expert quality. One of the overall text cohesion devices (aspect repetition) was a negative predictor in a regression model that predicted 26% of the variance in the essay scores. In a similar analysis, Guo, Crossley, and McNamara (2013) used Coh-Metrix indices to examine independent essay scores and source-based, or integrated, essay scores (i.e., writing that required the use of reading and/or listening materials as stimuli for composing an essay). Guo et al. reported that local and text indices of cohesion (e.g., aspect repetition, content word overlap, and conditional connectives) were negatively correlated with judgments of essay quality for the independent essays. However, for the integrated essays, which heavily relied on text integration from outside sources, local cohesive indices (e.g., semantic similarity between sentences and noun overlap) were positively correlated with essay quality and were included in regression models that predicted essay scores. Overall, these L2 studies indicate that expert ratings of essay quality either were not predicted by local and overall text cohesion devices or were negatively predicted by those devices. However, local cohesive indices were predictive of L2 writing quality for integrated (source-based) prompts, so that greater cohesion between local elements was related to higher judgments of writing quality.

Method

To validate TAACO and to investigate how local, global, and overall text cohesion indices can be used to assess expert judgments of text coherence and essay quality, we investigated the relations between indices provided by TAACO (outlined in greater detail below) and a corpus of scored student essays. The corpus used for this study comprised a set of independent essays written within a 25-min time frame that were scored by expert raters for essay coherence and overall essay quality.

Corpus

We selected the corpus of essays used in Crossley and McNamara (2011) in order to assess global cohesion. This corpus comprises 313 timed essays written on SAT prompts.

The essays were written by undergraduate freshmen composition students at Mississippi State University. The students were given 25 min to write an essay, during which no outside referencing was allowed. Two SAT prompts were used in the data collection, with students being randomly assigned to either prompt. All of the students were native speakers of English.

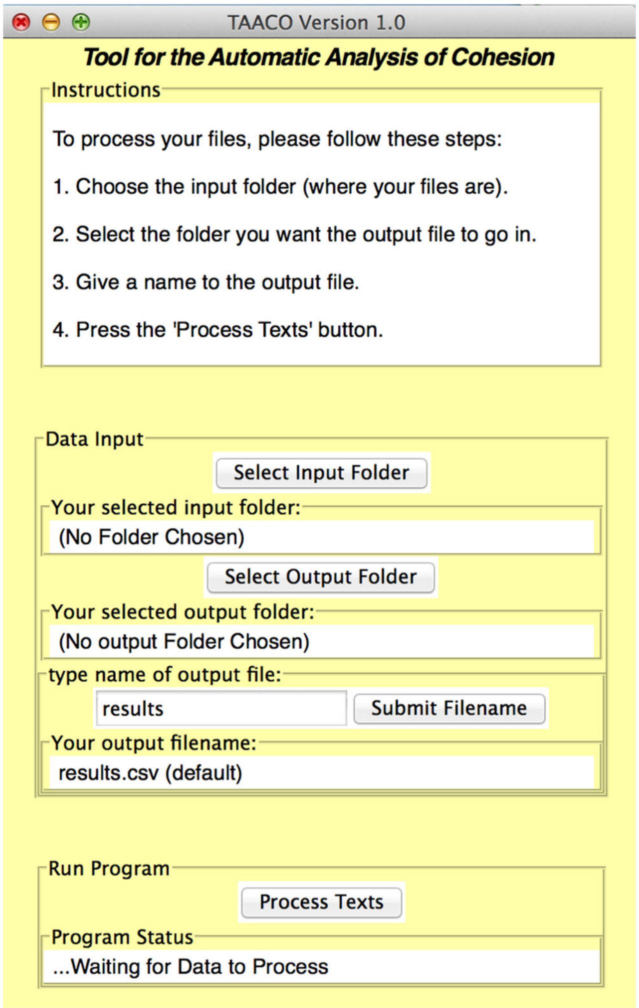
Each essay was read and scored by two trained raters on both overall quality (i.e., a holistic score) and specific textual elements (i.e., analytic scores). Eight raters in total took part. The holistic grading scale was based on a standardized rubric commonly used in assessing SAT[1] essays. The analytic rubric included sections related to the essay purpose, the essay plan, the use of topic sentences, the use of paragraph transitions, essay organization, writer conviction, and grammar and mechanics. Of interest for this analysis was the analytic feature relating to organization (i.e., coherence), which evaluated semantic-based, global cohesion (i.e., that the body paragraphs followed the plan set up in the introduction). Such structural elements promote overall text comprehension through the increase of global cohesion.

The trained raters who evaluated the essays had either master's or doctoral degrees in English, and each rater had at least 3 years experience teaching university-level composition classes. We thus consider these raters to be high-knowledge readers. The raters were informed that the distances between scores were equal. The raters were first trained to use the rubric with 20 practice essays, and after the raters had reached interrater reliabilities of at least r= .50 for the analytic scores and at least r= .70 for the holistic score, the raters then scored the 313 essays independently. After scoring was completed, the differences between raters were calculated. If the difference in ratings on a feature was less than 2 points, an average score was computed for that essay feature. If the difference was greater than 2 points, a third expert rater adjudicated the final rating. The correlations between the raters before adjudication for the holistic score were r= .79, and for the organization score r= .69.

TAACO

TAACO is a freely available text analysis tool that is written in Python, but it is implemented in a way that requires little to no knowledge of programming, since it can be started by double-clicking the TAACO icon. The TAACO interface is an easy to use and intuitive graphical user interface (GUI) that requires the user to select an input folder containing the files of interest (in .txt format). The user then selects an output folder for the output file and enters a name for a .csv file that TAACO will write the results for each text into (the default name is results.csv). The user then selects to process the texts, and a program status box informs the user of how many texts have been processed (see Fig. 1 for the TAACO GUI). Instructions and explanations for using TAACO, and the program itself, are available at www.kristopherkyle.com/taaco.html.

For a number of indices, the tool incorporates a part-of-speech (POS) tagger from the Natural Language Tool Kit (Bird, Klein, & Loper, 2009) and synonym sets from the WordNet lexical database (Miller, 1995). TAACO differs from other automatic tools that assess cohesion (i.e., Coh-Metrix; Graesser et al., 2004; McNamara et al., 2014) in that it reports on a greater number and variety of local, global, and overall text cohesion markers (see Table 1 for an overview). Additionally, TAACO is housed on the user's hard drive, allowing users to work independently of outside servers, which allows for secure processing of sensitive data. TAACO also incorporates part-of-speech (POS) tags and WordNet synonym sets.



TAACO Version 1.0

Tool for the Automatic Analysis of Cohesion

Instructions

To process your files, please follow these steps:

1. Choose the input folder (where your files are).

2. Select the folder you want the output file to go in.

3. Give a name to the output file.

4. Press the 'Process Texts' button.

Data Input

Your selected input folder: (No Folder Chosen)

Your selected output folder: (No output Folder Chosen)

type name of output file: results

Your output filename: results.csv (default)

Run Program

Program Status: ...Waiting for Data to Process

Fig. 1 TAACO interface

[1] The SAT is a college entrance exam commonly administered in the United States. An important component of the exam is a writing section in which test-takers are required to produce an essay based on general knowledge within a 25-min time frame.

Table 1 Cohesion features, categorizations, descriptions, and examples

Feature	Cohesion Type	Description	Example of High Cohesion
Connectives	Local	A number of theoretical and rhetorical lists of connectives	*First*, she was rich and happy.
Givenness	Overall text	Ratio of pronouns to nouns; incidence of demonstratives; definite articles	The man was happy *he* had *that*.
Type Token Ratio (TTR)	Overall text	Word repetition across a text	*The big* dog saw *the big* cat.
Lexical overlap	Both local and global	Overlap between nouns, arguments, stems, content and function words, and POS tags (for both sentences and paragraphs)	The *sun* was bright. The day was *sunny*.
Synonmy overlap	Both local and global	Overlap of synonyms across sentences and paragraphs	The *animal* was small. It was a *cat*.

Lexical overlap TAACO calculates a number of sentence overlap indices that assess local and global cohesion. These indices compute lemma (e.g., the lemma for the words human, humans, humanly, and inhumane is human) overlap between two adjacent sentences and paragraphs and between three adjacent sentences and paragraphs. TAACO calculates average overlap scores across sentences and paragraphs for all lemma overlap, content word lemma overlap, and lemma overlap for POS tags such as nouns, verbs, adjectives, adverbs, and pronouns. TAACO also calculates binary overlap scores for these features, which indicate whether there is any overlap between adjacent sentences or paragraphs. Local cohesion overlap indices have demonstrated positive relations with measures of cohesion in previous studies (McNamara et al. 2010a, b), but generally they demonstrate no significant relations with measures of coherence (Crossley & McNamara, 2010, 2011). Paragraph overlap indices have demonstrated positive relations with measures of text coherence in previous studies (Crossley & McNamara, 2011).

Semantic overlap Using the WordNet database, TAACO calculates overlap between words and sets of word synonyms (synsets) between sentences and between paragraphs. Unlike strict overlap indices, these indices measure overlap between semantically related words (i.e., the synset for jump contains the related words leap, bound, and spring, among others). TAACO calculates semantic overlap between sentences (local cohesion) and paragraphs (global cohesion) for nouns and for verbs. Semantic overlap has demonstrated positive relations with measures of cohesion in previous studies (McNamara et al. 2010a, b), but generally it has demonstrated no significant relations with measures of coherence (Crossley & McNamara, 2010, 2011).

Givenness Givenness is an important element of measuring cohesion and reflects the amount of information that is recoverable from the preceding discourse. To assess givenness, TAACO calculates the incidence of a variety of pronoun types, including first (e.g., I, me, us), second (e.g., you), and third (e.g., he, she, him, them) person pronouns, subject pronouns (i.e., I, you, she, he, but not me, him, and her), and quantity pronouns (e.g., many), under the presumption that pronouns are used when information is given (Crossley, Allen, Kyle, & McNamara, 2014). Following a similar presumption, TAACO calculates the ratio of nouns to pronouns. TAACO also counts the incidence of definite articles (i.e., the) and demonstratives (i.e., this, those, that, and these), under the presumption that definiteness is used for given information. Lastly, TAACO calculates the number and proportion of single content lemmas (e.g., how many lemmas occur only once in a text). Givenness indices have demonstrated positive relations with measures of text coherence in previous studies (Crossley & McNamara, 2011). These indices are calculated at the text level.

Type–token ratio (TTR) TTR measures the repetition of words in the text by dividing the number of individual words (types) by the total number of words (tokens). Thus, it likely taps into the amount of given information in a text. TAACO calculates a number of different TTR indices. These include simple TTR (the ratio of types to tokens), content word TTR (TTR using only content words such as nouns, verbs, adjectives, and adverbs), lemma TTR, and content lemma TTR. In addition to traditional word-based TTR indices, TAACO also calculates TTR for bigrams (i.e., two-word strings) and for trigrams (three-word strings). TTR indices have demonstrated positive relations with measures of cohesion in previous studies (Crossley & McNamara, 2014; McCarthy & Jarvis, 2010), but generally they demonstrate negative relations with measures of text coherence (Crossley & McNamara, 2010; McNamara et al. 2010a, b). TTR indices are calculated at the text level.

Connectives TAACO contains a number of connective indices that measure local cohesion. Many of the connective indices are similar to those found in Coh-Metrix (McNamara et al. 2014) and are theoretically based on two dimensions. The first dimension contrasts positive versus negative connectives, and the second dimension is associated with the particular classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001), such as temporal, additive, and causative connectives. These theoretically based indices have demonstrated negligible or negative correlations with essay quality and essay coherence (Crossley & McNamara, 2010, 2011). A number of new connective indices were also included in TAACO, based on considerations of how connectives operate rhetorically in written texts, as compared to theoretical bases. These connective classes are summarized, with examples, in Table 2. The lists were collected through reference searches and consultation with experts. Some connective indices have demonstrated positive relations with measures of cohesion in previous studies (McNamara et al. 2010a, b), but generally they demonstrate no significant relations with measures of coherence (Crossley & McNamara, 2010, 2011).

Statistical analysis

For the essay analyses, the TAACO indices were the predictor indices, and the human scores (for both coherence and overall essay quality) were the criterion variables. Indices reported by TAACO that lacked normal distributions were removed. The corpus was first divided into training and test sets using a 67/33 split (Witten, Frank, & Hall, 2011). Using the training set, correlations were then calculated to determine whether there was a statistical ($p< .05$) and meaningful (of at least a small effect size, $r> .10$) relation between the TAACO indices and both the human scores for coherence and the human scores for holistic quality. Indices that were highly collinear ($r> .90$) were flagged, and the index with the strongest correlation with human scores was retained while the other indices were removed. The remaining indices were included as predictor variables in a stepwise multiple regression to explain the variance in the human scores of both coherence and overall essay quality. The model from the stepwise regression was then used to predict the variance in the human scores for the essays in the test set. We predicted that the global cohesion indices would positively correlate to the human ratings of coherence and essay quality, and that the local cohesion indices would correlate negatively.

Table 2 Rhetorical connectives reported by TAACO

Class	Example
Coordinating connectives	but, and, or
Semicoordinators	nor, so, yet
Basic coordinators	combined coordinators and semicoordinators
Quasi-coordinators	as well as
Conjunctions	and, but
Disjunctions	or
Simple subordinators	after, until
Complex subordinators	so that
Coordinating conjuncts	however
Addition	further
Contrasts	on the contrary
Sentence linking	Nonetheless
Order	first, finally
Reference	with regard
Reason and purpose	hence
Condition	in case of
Concession	although

Results

Coherence scores

Correlations with human ratings Correlations were conducted between the TAACO indices and the human ratings of essay coherence. Of these variables, 12 demonstrated significant correlations with the human scores for coherence. Of these 12 variables, two demonstrated significant multicollinearity with variables that reported a larger r value (Adjacent overlap binary two paragraphs verb lemma average and Adjacent overlap two paragraphs content words lemma average). These variables were removed from the regression analysis. The correlations for the remaining variables are reported in Table 3.

Regression A stepwise regression analysis using the ten significant indices as the independent variables to predict the human scores of coherence yielded a significant model, $F(3, 197)= 16.55$, $p< .001$, $r= .45$, $R^2= .20$. Three of the TAACO variables were included as significant predictors of the coherence scores. These variables were Adjacent overlap two paragraphs all lemmas average, Adjacent overlap binary two paragraphs adverb lemma average, and Verb synonym sentence lemma overlap. The two global cohesion indices were positive predictors, whereas the local cohesion index (Verb synonym sentence lemma overlap) was a negative predictor.

The model demonstrated that the three variables together explained 20% of the variance in the human scores of coherence for the 197 essays in the training set (see Table 4 for additional information). When the model was applied to the test set, the model yielded $r= .47$, $R^2= .22$, indicating that the three variables together explained 22% of the variance in the human scores of coherence for the 116 essays in the test set and that the model is stable.

Essay scores

Correlations with human ratings Correlations were conducted between the TAACO indices and the human ratings of essay quality. Of these variables, 20 demonstrated

Table 3 Correlations between TAACO indices and human scores of coherence

Index	Cohesion Type	*r*	*p*
Adjacent overlap binary two paragraphs noun lemma average	Global	.42	<.01
Adjacent overlap binary two paragraphs adverb lemma average	Global	.40	<.01
Adjacent overlap two paragraphs all lemmas average	Global	.37	<.01
Adjacent overlap binary two paragraphs pronoun lemma average	Global	.37	<.01
Adjacent overlap binary two paragraphs adjective lemma average	Global	.35	<.01
Verb synonym paragraph lemma overlap	Global	.19	<.01
Noun synonym paragraph lemma overlap	Global	.18	<.01
Verb synonym sentence lemma overlap	Local	–.16	<.01
Incidence of pronouns	Text	.13	<.05
Repeated content word lemmas	Text	.12	<.05

significant correlations with the human scores for essay quality. Two of these variables demonstrated significant multicollinearity with variables that reported a larger *r* value (Adjacent overlap binary two paragraphs argument lemma average and Adjacent overlap two paragraphs content words lemma average). These variables were removed from the regression analysis. The correlations for the remaining variables are reported in Table 5.

Regression A stepwise regression analysis using the 18 significant indices as the independent variables to predict the human scores of coherence yielded a significant model, F(4, 208)= 18.17, $p< .001$, $r= .51$, $R^2= .26$. Four of the TAACO variables were included as significant predictors of the coherence scores. These variables were Adjacent overlap two paragraphs all lemma average, Adjacent overlap binary two paragraphs verb lemma average, Lemma TTR, and Ratio of pronouns to nouns. The two global cohesion indices were positive predictors, whereas the two remaining overall text cohesion indices were negative predictors.

The model demonstrated that the four variables together explained 26% of the variance in the human scores of coherence for the 212 essays in the training set (see Table 6 for additional information). When the model was applied to the test set, the model yielded $r=$ .52, $R^2= .27$, indicating that the four variables together explained 27% of the variance in the human scores of coherence for the 101 essays in the test set and that the model is stable.

Discussion

This article introduces a new tool, TAACO, that automatically analyzes local, global, and overall text cohesion. The findings from this study help to provide predictive validity for the indices reported by TAACO, by demonstrating the potential for the TAACO indices to predict expert judgments of text coherence and judgments of essay quality. Our hope is that the tool will provide researchers in discourse processing, language assessment, education, and cognitive science with access to a greater depth and breadth of linguistic indices related to text cohesion and coherence. The indices in TAACO could be used to study the effects of discourse beyond those tested here (i.e., in text readability studies or with lower-level readers). TAACO indices could also be used by researchers in language assessment and education to develop tests, examine differences in selected texts, and as predictors in automatic essay scoring (AES) systems. As well, TAACO indices could be used by cognitive scientists to develop stimuli for behavioral studies examining language processing. In essence, researchers in any number of fields with an interest in language and discourse structure could use TAACO as a research tool.

Overall, the findings of this study support the notion that expert judgments of text coherence are either negatively

Table 4 Stepwise regression analysis and significance values for TAACO indices predicting essay scores

Entry	Index Added	*r*	Total R^2	*B*	B	*SE*	*t*
Entry 1	Adjacent overlap two paragraphs all lemmas average	.40	.16	0.95	0.23	0.37	2.53^*
Entry 2	Adjacent overlap binary two paragraphs adverb lemma average	.43	.19	0.83	0.23	0.33	2.55^*
Entry 3	Verb synonym sentence lemma overlap	.45	.20	–0.17	–0.14	0.08	$–2.12^*$

B= unstandardized β; B= standardized; *SE*= standard error. The estimated constant term is 2.859. $^*p< .05$

Table 5 Correlations between TAACO indices and human scores of essay quality

Index	Cohesion Type	*r*	*p*
Adjacent overlap two paragraphs all lemmas average	Global	.40	<.01
Adjacent overlap two paragraphs noun lemma average	Global	.37	<.01
Adjacent overlap two paragraphs argument lemma average	Global	.37	<.01
Adjacent overlap binary two paragraphs verb lemma average	Global	.35	<.01
Adjacent overlap binary two paragraphs adverb lemma average	Global	.33	<.01
Adjacent overlap binary two paragraphs adjective lemma average	Global	.31	<.01
All lemma type–token ratio (TTR)	Text	–.29	<.01
Noun synonym paragraph lemma overlap	Global	.26	<.01
Adjacent overlap two paragraphs pronoun lemma average	Global	.24	<.01
Verb synonym paragraph lemma overlap	Global	.22	<.01
Adjacent overlap sentence verb lemma	Local	–.18	<.01
Bigram TTR	Text	–.17	<.01
Content word TTR	Text	–.17	<.01
Ratio of pronouns to nouns	Text	–.15	<.01
Repeated content word lemmas	Text	.14	<.01
Adjacent overlap sentence content words lemma	Local	–.14	<.05
Trigram TTR	Text	–.11	<.05
Verb synonym sentence lemma overlap	Local	–.11	<.05

correlated or not predicted by local cohesion indices. In contrast, and as predicted, expert ratings of coherence are positively correlated and positively predicted by global indices of cohesion calculated at the paragraph level. Similar findings have been reported for expert judgments of writing quality in which overall text cohesion devices are negatively correlated to and negatively predict writing quality, whereas global indices of cohesion positively predict essay quality. Combined, these findings provide supporting evidence that text coherence for high-knowledge readers (in this case, expert raters) is related to global cohesion and that text coherence (in terms of global cohesion) is a positive predictor of essay quality. In contrast, local and overall text cohesion indices are generally not positive predictors or either text coherence or essay quality for high-knowledge expert raters of essays.

Specifically, our first analysis examined relations between local, global, and overall text cohesion indices and expert judgments of text coherence as a means of testing the predictive validity of the TAACO indices. We selected human judgments of text coherence because of links between cohesion and coherence and also because earlier studies had indicated that text coherence was not predicted by indices related to local cohesion, but may be related to indices of global cohesion (Crossley & McNamara, 2010, 2011). Initial correlations strongly supported this notion, in that the majority of all of the TAACO indices that positively correlated with judgments of text coherence were global in nature (i.e., measured overlap between paragraphs and not sentences). The strongest of these indices were noun, verb, and adverb overlap between paragraphs. Local cohesion and overall text cohesion indices generally correlated negatively with judgments of text coherence (e.g., TTR and verb synonym overlap between sentences) or did not demonstrate any significant correlations (e.g., the majority of connective and sentence overlap indices). These findings provide us with two sources of information. First they help support the validity of the TAACO indices in that they replicate previous findings and provide a model of coherence based on global and local indices of cohesion. Second, the results further our understanding of how expert raters develop coherent models of a text based on the repetition of nouns,

Table 6 Stepwise regression analysis and significance values for TAACO indices predicting essay scores

Entry	Index Added	*r*	Total R^2	*B*	B	*SE*	*t*
Entry 1	Adjacent overlap two paragraphs all lemma average	.37	.14	0.02	0.24	0.01	3.46^{**}
Entry 2	Adjacent overlap binary two paragraphs verb lemma average	.42	.18	–17.38	–0.26	4.15	$–4.19^{**}$
Entry 3	Lemma TTR	.46	.21	–4.34	–0.28	1.11	$–3.92^{**}$
Entry 4	Ratio of pronouns to nouns	.51	.26	–1.45	–0.22	0.411	$–3.54^{**}$

B= unstandardized β; B= standardized; *SE*= standard error. The estimated constant term is 5.898. $^{**}p< .001$

verbs, and adverbs across larger and not smaller chunks of text. This is likely a result of expert raters being able to monitor and refer to a variety of propositions across an entire text without the need for repetition of propositions at the sentence level. A regression model using both local and global indices explained 22% of the variance in the human ratings, with two global indices of cohesion as the strongest predictors. The strength of the regression model is lower than that reported by Crossley and McNamara (2011), but Crossley and McNamara included indices related to text structure (length and number of paragraphs), syntactic complexity, and lexical sophistication in their regression analysis, so direct comparisons are not possible in this study. Nonetheless, the analysis presented here provides support for the notion that expert raters do not appear to depend on local cohesion devices to develop a coherent representation of the text, but rather are more influenced by global cohesion.

Our second analysis examined links between the local, global, and overall text cohesion scores and expert ratings of essay quality. This analysis was built on previous studies that indicated that local cohesion devices were either not correlated or negatively correlated with writing quality (Crossley & McNamara, 2012; Crossley et al. 2011a, b; McNamara et al. 2010a, b). A few studies have also indicated the potential for global indices of cohesion to correlate with but not predict writing quality (Crossley & McNamara, 2011; Crossley et al. 2011a, b), but this notion is not fully supported (Crossley & McNamara, 2012). Initial correlations indicated that the strongest correlations were reported for global indices of cohesion, followed by local and overall text cohesion indices. The global cohesion indices all reported positive correlations (i.e., overlap between paragraphs for all lemmas, content lemmas, nouns lemmas, verb lemmas, adverb lemmas, and adjective lemmas), whereas the majority of local and overall text cohesion indices correlated negatively with judgments of writing quality (all TTR measures, sentence overlap indices, and pronoun indices). These results indicate that expert raters judge writing proficiency in a manner similar to how they develop coherent mental representations of text (McNamara & Kintsch, 1996). That is to say, they appear not to rely upon repetition of words at the sentence level, but rather track and process words across larger chunks of texts. From a construction–integration model perspective (Kintsch, 1998), this also indicates that expert raters are able to integrate textual propositions from a more loosely connected network of concepts. A mix of both overall text and global cohesion indices explained 27% of the variance in the human judgments of essay quality.

Comparisons between this model and previous models of essay quality are difficult because automatic essay scoring (AES) models generally sample indices from a variety of linguistic constructs (e.g., text structure, syntactic complexity, discourse components, and lexical sophistication), whereas our model only sampled indices from a single construct: text cohesion. Overall, the analysis helps support the notion that essay quality is negatively related to the use of local cohesion devices but positively related to the use of global cohesion devices, a notion that may help spur the accuracy and validity of future AES systems. Nevertheless, the variance explained by cohesion features alone in our model is on the low end of acceptability when compared with previous research on AES models (Attali & Burstein, 2006; McNamara et al., 2013; Warschauer & Ware, 2006).

The two regression analyses provide support for the notion that high-knowledge readers, such as the expert raters in this study, benefit from texts that exhibit less local and overall text cohesion and greater global cohesion. In previous studies, such a reverse cohesion effect has demonstrated that low-knowledge readers, unlike high-knowledge readers, benefit from local cohesion devices that help the readers bridge gaps in their background knowledge and reading skills, among both native readers (McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996; O'Reilly & McNamara, 2007) and L2 readers (Crossley et al., 2014). In contrast, high-knowledge readers seem to be influenced in their judgments of essay coherence and quality not by local cohesion, but by global cohesion that connects ideas across larger segments of the text. The TAACO indices sampled here support this notion, indicating that cohesion gaps that are resolved at the global level increase coherence for expert raters and, concomitantly, lead the same raters to score the essay higher. This is likely the case because expert raters have greater expertise in writing that affords memory resources that allow them to reference information across larger text segments without losing the meaning of the text (see, e.g., Ericsson & Kintsch, 1995).

Conclusion

This study introduces and demonstrates the use of a new tool, TAACO, which is freely available to researchers. We presume that this tool will facilitate research on the significance of cohesion and coherence in discourse studies, language assessment, education, and cognitive science (among other disciplines). We foresee TAACO being used to examine differences in text readability, grade-level texts, text genres, spoken discourse, writing tasks, and psycholinguistic stimuli, and to develop models of cognitive processing such as those found in AES systems. The study also provides evidence supporting the notion that the different types of cohesion devices are predictive of human evaluations of text coherence and text quality. In general, this evidence indicates that local, global, and overall text cohesion devices are important in different ways for a population of expert readers with high knowledge and experience. An important component of this study was to provide details on how cohesion devices are related to judgments of text coherence and essay quality, which has proven to be difficult to assess successfully using NLP tools. However, the results of this study are promising, with a number of TAACO indices being correlated with essay quality and with expert judgments of text coherence.

Thus, TAACO provides an automated approach for the examination of how cohesion devices at the local, global, and text levels relate to text comprehension and judgments of text quality and coherence. We plan on extending this foundation by developing additional indices of cohesion for inclusion in TAACO. Such indices will examine semantic similarity using latent semantic analysis and latent Dirichlet allocation at the local, global, and text levels. We will use these new indices to examine the effects of cohesion on both low-knowledge and high-knowledge readers in terms of differences in text comprehension and judgments of text coherence. We also plan on expanding our understanding of text coherence by collecting judgments of text coherence from nonexpert raters (i.e., naïve raters) in order to investigate differences in text coherence based on skill and background knowledge.

Author note We thank our expert raters for their assistance in scoring the essays used in this study. We also thank Art Graesser, Carl Cai, Philip McCarthy, and Jianmin Dai for their advice over the years.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*, 3.

Bird, S., Klein, K., & Loper, E. (2009). *Natural language processing with Python*. Beijing, China: O'Reilly.

Collins, J. L. (1998). *Strategies for struggling writers*. New York, NY: Guilford Press.

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using the simple natural language processing tool (SiNLP). *Discourse Processes, 51,* 511–534.

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). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236–1241). Austin, TX: Cognitive Science Society.

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. (2014). Developing component scores from natural language processing tools to assess human ratings of essay quality. In W. Eberle & C. Boonthum-Denecke (Eds.), *Proceedings of the 27th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 381–386). Menlo Park, CA: AAAI Press.

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: AAAI Press.

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011a). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 438–440). New York, NY: Springer.

Crossley, S. A., Weston, J., McLain-Sullivan, S. T., & McNamara, D. S. (2011b). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28,* 282–311.

DeVillez, R. (2003). *Writing: Step by step*. Dubuque, IA: Kendall Hunt.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211–245. doi:10.1037/0033-295X.102.2.211

Foltz, P. W. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 167–184). Mahwah, NJ: Erlbaum.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36,* 193–202. doi:10.3758/BF03195564

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Writing Assessment, 18,* 218–238.

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

Johns, A. (1986). Coherence and academic writing: Some definitions and suggestions for teaching. *TESOL Quarterly, 20,* 247–265.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

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

Loxterman, J. A., Beck, I. L., & McKeown, M. G. (1994). The effects of thinking aloud during reading on students' comprehension of more or less coherent text. *Reading Research Quarterly, 29,* 353–367.

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, 42,* 381–392. doi:10.3758/BRM.42.2.381

McNamara, D. S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55,* 51–62.

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

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45,* 499–515. doi:10.3758/s13428-012-0258-1

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., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22,* 247–288.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43.

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

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), 39–41.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43,* 121–152.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10,* 1–24.

Witten, I. A., Frank, E., & Hall, M. A. (2011). *Data mining*. San Francisco, CA: Elsevier.