Writing Analytics for Epistemic Features of Student Writing

Simon Knight, University of Technology Sydney, simon.knight@uts.edu.au
Laura Allen, Arizona State University, laurakallen@asu.edu
Karen Littleton, Open University, k.s.littleton@open.ac.uk
Bart Rienties, Open University, bart.rienties@open.ac.uk
Dirk Tempelaar, Maastricht University, d.tempelaar@maastrichtuniversity.nl

Abstract: Literacy, encompassing the ability to produce written outputs from the reading of multiple sources, is a key learning goal. Selecting information, and evaluating and integrating claims from potentially competing documents is a complex literacy task. Prior research exploring differing behaviours and their association to constructs such as epistemic cognition has used 'multiple document processing' (MDP) tasks. Using this model, 270 paired participants, wrote a review of a document. Reports were assessed using a rubric associated with features of complex literacy behaviours. This paper focuses on the conceptual and empirical associations between those rubric-marks and textual features of the reports on a set of natural language processing (NLP) indicators. Findings indicate the potential of NLP indicators for providing feedback regarding the writing of such outputs, demonstrating clear relationships both across rubric facets and between rubric facets and specific NLP indicators.

Keywords: epistemic cognition, literacy education, writing analytics, learning analytics

Introduction

Literacy, including the abilities to comprehend rich multimedia, and effectively communicate through written texts, is key to learning, and full participation in society across age ranges (OECD, 2013; OECD & Statistics Canada, 2010). Indeed, the 2009-2015 PISA definition of reading indicates that: "Reading literacy is understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2013, p. 9). Thus Rouet (2006) suggests that literacy in the context of rich-multimedia environments (such as the internet) involves the skills of: *integration* of prior knowledge and across documents (including competing claims); *sourcing* of features that identify the provenance, genre, etc. of the information; and *corroboration* to check information across multiple sources.

Multiple document comprehension

One class of research into this area of literacy has explored multiple document processing, the ability to read, comprehend and integrate information from across sources, (see, for examples, Bråten, 2008; Bråten, Britt, Strømsø, & Rouet, 2011; Ferguson, 2014; Foltz, Britt, & Perfetti, 1996; S. R. Goldman et al., 2011; Hastings, Hughes, Magliano, Goldman, & Lawless, 2012; Kobayashi, 2014; Rouet & Britt, 2011), with some research specifically viewing these behaviours through the lens of epistemic cognition – characterised as beliefs about the certainty, simplicity, source, and justification of knowledge (see, for examples, Bråten, 2008; Bråten et al., 2011; Ferguson, 2014). Such tasks provide a context in which to explore the ways in which different sources are treated and drawn on in subsequent tests of knowledge or argumentation; the language participants use in these subsequent contexts may relate to the particular documents implicitly or explicitly drawn on. In this vein, in recent work on literacy and epistemic cognition (Anmarkrud, Bråten, & Strømsø, 2014; Bråten, Braasch, Strømsø, & Ferguson, 2014) students were asked to produce written outputs, which were then scored for:

1. Presence of explicit or implicit sourcing (i.e. explicit reference to the source, or indirect reference such as "one article spoke of [specific detail]" but without direct use of source information);
2. References to trustworthiness of the source or information from that source (coding separately for negative and positive evaluations;
3. Finally, whether connections were made between content-source trustworthiness (for example, whether content was trusted more because of the properties of the document from which it was sourced)

That research found that, approximately half of sourcing references were explicit (with the other half implicit) and students did not make reference to the full list of sources (approximately 3 of 6 references). In other multiple document processing research, Goldman, Lawless, Pellegrino and Gomez (2012) identified three clusters of students from their written outputs: *satisficers*, who selected few sources; *selectors* who selected many sources but did not connect them; and *synthesisers* who selected sources and integrated them.

Developing language technologies

Given these prior research findings, the development of multiple document processing tasks provides opportunity to explore relationships among psychological constructs, outputs, and learning process data (Knight & Littleton, 2015). Emerging language technologies raise potential for such research into relationships between features in output texts and score-descriptors on rubric facets grounded in theorized constructs such as epistemic cognition. For example, analysis of the written outputs for: rhetorical moves that are indicative of claims making, evaluation, and connecting (or synthesis) (see, for example, de Waard, Buitelaar, & Eigner, 2009; Groza, Handschuh, & Bordea, 2010; Simsek, Buckingham Shum, Sandor, De Liddo, & Ferguson, 2013); text cohesion (McNamara, Louwerse, McCarthy, & Graesser, 2010); and topic coverage and integration (see, for example, Hastings et al., 2012).

In the last of these studies (Hastings et al., 2012), students were asked to use three texts with relatively little semantic overlap to answer the inquiry question "In 1830 there were 100 people living in Chicago. By 1930, there were three million. Why did so many people move to Chicago?" They compared three methods to match source material to student writing outputs: Pattern matching approaches (i.e. looking for common text-strings); latent semantic analysis (LSA) to compare semantic-content at a sentence level across student outputs and assigned texts; and machine learning (using support vector machines) assigning student sentences to topic-classes assigned by human-raters. They found that LSA performed best in identifying explicit use of the assigned texts, while pattern-matching approaches were superior for detecting intra and inter-textual inferences (which could be characterised as synthesis or integration of information).

The varied approaches to text analysis are of potential interest in developing approaches to a rich understanding of literacy and writing encompassing not only topic coverage, but also the ways information from multiple sources are integrated (synthesis), which sources information is drawn from (source diversity), and markers of evaluation and contrast (source quality and evaluation). The potential of language technologies, then, is to connect particular types or styles of language to epistemic characterisations; further work to connect computational outputs to human interpretable scores or feedback would then be required. Throughout this prior work the use of key-content, implicit and explicit citation, evaluation of those citations and (separately) their content, and the synthesis of information are foregrounded.

Current study

In the research reported in this paper, a conceptual alignment is drawn between the key literacy considerations of topic and source coverage, synthesis, and sourcing (including evaluation and justification), and a set of language-technology indicators. Examples of the kinds of language of interest to these features of written text are given, with preliminary results indicating relationships between outcomes on a rubric, and language technology indicators. The paper thus addresses the research aim to produce alignment between epistemic features in writing based on multiple document processing, and automated textual analysis of features that align with those key epistemic considerations.

Methods

Participants and ethics

270 students at Maastricht University (Netherlands) enrolled in a 1st year business and economics course took part in the study from which this data is drawn. Participants worked in pairs (n = 135), assigned by the researcher in the class environment. The research followed British Psychological Society ethics guidelines (British Psychological Society, 2014), with all students consenting to take part in the research and informed of their right to withdraw at any time. As part of this consent, students agreed to the Terms of Use of the Coagmento tool (described below), including that data obtained could be used for research purposes. Data is shared among the collaborators such that the Open University team holds all data, with other collaborators having access to subsets of data (anonymized for non-Maastricht researchers). The research was conducted on university PCs, and the browser cache was cleared between sessions to ensure no personal data (or active logins) was exposed either between participants, or to the researcher.

Materials and procedure

Participants completed the tasks in a computer lab using the Firefox browser (Mozilla, 2014) along with a customised add-on called Coagmento (Shah, 2010, 2014). Coagmento is a collaborative information-seeking (CIS) tool designed to support people in collaborating on shared information tasks. It includes a chat tool, and an

integrated 'etherpad' environment – a shared document editor, such that collaborators can write together in real time and share 'snips' (copied text) via the browser add-on tool.

In this research, participants were given a short warm-up task to familiarise themselves with the respective tools and their paired collaborator, following which they were requested to write collaboratively using the etherpad, to create a report of the "best support claims" around the risks of a substance (a herbicide called glyphosate). This topic selection follows the work described in the introduction; topics with conflicting perspectives and a variety of source-qualities were sought to foreground participant's commitments to varying source-content qualities, and abilities to integrate a variety of perspectives. The topic thus:

1. Provides a focussed topical research area which can be studied in isolation, within a 1 hour session;
2. Is not a topic that was high profile or/and a large scale controversy (such as climate change, or genetically modified crops, both of which receive large amounts of press coverage);
3. Has a variety of source-types and qualities regarding it, from varying perspectives.

For the task, a set of eleven documents was collated from materials that discussed a herbicide, the safety of which has been questioned in terms of human health implications and agricultural risks. A simplified document-model (building on Rouet's work 2006, Chapter 3) is given in Figure 1, depicting the three key themes identified in this document set (the presence of glyphosate markers in human urine; the risks to human health of glyphosate; and the agricultural risks of glyphosate use), the document stance (broadly negative – orange; critique/broadly positive – green; largely neutral/scientific – blue) and the relations among them (+ - support; - - critique; note document 11 relates to 2 primary themes). Documents were presented via a webpage, with titles and short 'snippets' given (mirroring a search engine results page presentation of website titles and excerpts).

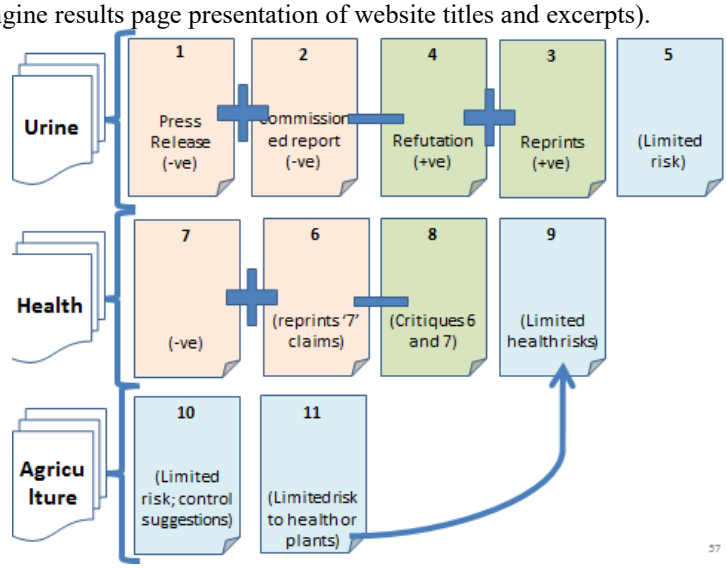

Figure 1. Simplified MDP Document Module.

For each document, the original source (HTML or PDF) was saved and formatted for presentation to students (to ensure it would load without scripts, etc., and would load in html without need to use a PDF viewer or other external reader). The documents were also cleaned, to remove extraneous detail and to reduce them to core claims around glyphosate. Only the abstracts of 5, 7 and 9 were given, while 10 was reduced to the abstract and first section of the introduction; 11 was reproduced in an abridged form, it was also the most comprehensive document in terms of coverage of potential risks.

Note in particular that the set of documents compiled is rather complex. For example, the author of 7 is criticised in 8. Ostensibly, 7 is more trustworthy because it is in a peer review journal and republished by Reuters (6), while 8 is a blog. However, the critique provided in 8 (and the evidence referred to) is strong and the source features of the blog (also based at MIT) are also strong. Furthermore, the author of 7 has been criticised for publishing in an area they are not an expert in, (including praising the discredited Andrew Wakefield on autism), and while the journal is peer reviewed, it is primarily a physics journal not a health-sciences one. We also see in documents '3' and '4' a reprint on a trade website (Farmers Weekly; 4) of an independent critique (3) – something students might identify as raising concerns of bias in '4', although the content is identical. Thus the selection of documents provides a set of conflicting sources, of varying quality, with a range of sub-topics present. As such, the topic and selected documents provide good source material for probing students' abilities to extract, integrate and evaluate information from across sources.

Participant outputs were assessed by the first author using a rubric. The rubric, based on the particular task design, and the MDP work described above, consisted of:

1. Topic coverage – The text covers a range of different topics and relates them to the question (the risks of the substance)
2. Range of sources coverage – The text uses a range of sources
3. Quality of sources – The text evaluates the quality of sources cited
4. Synthesis of information – The text synthesizes information from across sources

In this rubric, a score of '1' indicates content coverage, '2' the sourcing of that content, '3' evaluation of the source features and content, and '4' the ways intertextual ties that are identified and made in the text. Each rubric was assessed on a 1-3 scale. The same rubric was used for a parallel task based on different materials, for which a second rater assessed the student outputs, with acceptable reliability (>.8 Cohen's Kappa on all indices except the synthesis score which had a .58 Kappa).

Analysis

Quantitative analysis

Quantitative analysis was conducted to assess the variance in rubric scores for the written outputs, intending to explore whether or not epistemic differences in writing are captured by the task design. Following a conceptual alignment (highlighted in the next section) between the rubric and quantifiable textual indicators in the written outputs, correlation analysis was conducted to explore these relationships empirically. The potential of this approach is to design models (for example, using regression analyses) to align the rubric facets with particular textual features that may be identified automatically using natural language technologies.

Qualitative analysis

Qualitative analysis was conducted to ground the conceptual alignment drawn between the rubric facets and anticipated textual features. The texts were analysed with regard to their epistemic and textual properties, with key features identified. Across the rubric facets variations in outcome were characterized by, for example:

1. **Synthesis:** Use of lists and extracts from individual articles, versus integrated text organized thematically and drawing from across multiple sources
2. **Topic Coverage:** A sparse use of topically salient keywords, or/and a focus on individual subtopics rather than drawing from the full range of themes and their keywords
3. **Source Diversity:** A focus on sourcing information from 'one best' article, versus the discussion and integration of claims from multiple sources
4. **Source quality:** Uncritical citation of claims, even where claims disagreed, versus identification and connection of disagreements and a critical balancing of claims based on source features

Natural language processing analysis

To assess the linguistic properties of the students' writing, we utilized the *Tool for the Automatic Analysis of COhesion* (TAACO). TAACO is an automated text analysis tool that calculates 150 classic and recently developed indices related to both the local and global cohesion of a text, in contrast to other tools that either do not assess cohesion or focus solely on local cohesion (e.g. Coh-Metrix, McNamara, Graesser, McCarthy, & Cai, 2014). An additional strength of the tool is that it incorporates part-of-speech (POS) tags and synonym sets. The POS tags in TAACO are identified using the POS tagger developed as part of the Natural Language Tool Kit (Bird, Klein, & Loper, 2009), and the synonym sets are taken from the WordNet lexical database (Miller, 1995). In the sections below, we provide descriptions of the categories of TAACO indices most relevant to the current paper (i.e., *basic indices*, *sentence overlap*, *paragraph overlap*, and *connectives*). For more specific information on all of the indices provided by TAACO, see Crossley, Kyle, and McNamara (2015).

Basic indices. TAACO provides indices related to basic information about a text, such as the number of words (i.e., tokens), number of word types (i.e., unique words), and the type-token ratio (TTR). TTR calculates word repetition by dividing the total number of words in a text (tokens) by the number of individual words (types). Therefore, this index describes the amount of given information in a particular 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) and function word TTR (TTR using only function words such as pronouns, preposition, and determiners. TTR indices have demonstrated positive relations with measures of cohesion in previous studies (Crossley & McNamara, 2014; McCarthy & Jarvis, 2010), but

generally demonstrate negative relations with measures of text coherence (Crossley & McNamara, 2010; McNamara, Crossley, & McCarthy, 2010).

TTR indices may account for both local and global characteristics of cohesion. Because they are measured at the level of the overall text, it is difficult to determine whether word repetition is occurring between sentences or larger portions of the text. In the current study, we calculated the *total number of words* and the *total number of word types*. Additionally, we used the basic *type token ratio* index to provide basic information about the lexical diversity in the students' texts.

Sentence overlap. The TAACO tool provides multiple sentence overlap indices to assess *local* text cohesion. These indices calculate lemma overlap between two adjacent sentences and among three adjacent sentences. TAACO also provides average overlap scores across the text for lemma overlap, content word lemma overlap, and lemma overlap for POS tags, such as nouns, verbs, adjectives, adverbs, and pronouns. Finally, TAACO calculates binary overlap scores for all features; these scores indicate if there is or isn't (i.e., 1 or 0) any overlap between adjacent sentences. Overall, overlap indices tend to be positively related to measures of cohesion (see, e.g., McNamara, Louwerse, et al., 2010); however, they typically are unrelated to measures of text *coherence* (Crossley & McNamara, 2010, 2011).

In the current study, we assessed local cohesion in students' texts using three measures of sentence overlap: *adjacent sentence overlap (all words)*, *adjacent sentence overlap (content words)*, and *adjacent sentence overlap (function words).* These three indices all provide information about the local cohesion established between sentences in a text. We included the content and function word indices to provide more specific information about the type of local cohesion found (or not) in a given text. Content word overlap would indicate that similar topics are being discussed in adjacent sentences, whereas function word overlap would be more indicative of similar rhetorical information and sentence structures.

Connectives. The TAACO tool also includes a number of connective indices to measure local cohesion. A number of these indices are based on indices found in the Coh-Metrix tool (McNamara et al., 2014) and can be theoretically described according to two dimensions. The first connective dimension differentiates between positive and negative connectives, whereas the second dimension relates to previously defined categories of cohesion (Halliday & Hasan, 2014; Louwerse, 2001), such as temporal, additive, and causative connectives. Previous research has found negative or no correlation between these indices and measures of writing quality and coherence (Crossley & McNamara, 2010, p. 20, 2011).

TAACO includes multiple new connective indices that are based on the rhetorical purposes of connectives (see Crossley et al., 2015 for more thorough descriptions). Some of these indices have demonstrated positive relations with measures of cohesion in previous studies (McNamara, Louwerse, et al., 2010), but typically do not significantly relate to measures of coherence (Crossley & McNamara, 2010, 2011).

We analyzed three connective indices in the current study: *basic connectives*, *sentence linking connectives,* and *reason and purpose connectives*. These three indices all provide information about the local cohesion established between sentences in a text. However, they differ from basic sentence overlap because they describe how links are being established at the sentence level. We included the basic connectives index to provide an overall measure of the connectives present in a students' text. Additionally, we included the sentence linking (e.g., *nonetheless*) and reason and purpose (e.g., *hence*) connectives to indicate local cohesion that is being established for specific rhetorical purposes.

Paragraph overlap. Finally, in addition to the local cohesion indices, TAACO calculates multiple paragraph overlap indices to assess *global* cohesion. These indices include lemma overlap between two adjacent paragraphs and among three adjacent paragraphs. These indices are based on the same features as the sentence overlap indices (i.e., average and binary lemma overlap, content word lemma overlap, and lemma overlap for POS tags). Previous research suggests that paragraph overlap indices are positively related to text coherence (Crossley & McNamara, 2011).

We assessed the global cohesion of students' texts using three measures of paragraph overlap: *adjacent paragraph overlap (all lemmas)*, *adjacent paragraph overlap (content lemmas)*, and *adjacent paragraph overlap (function lemmas).* These three indices provide information about the global cohesion established between paragraphs in a text. Similar to the sentence overlap indices, we included the content and function word indices to provide more specific information about the type of global cohesion found (or not) in a given text. Additionally, we investigated overlap among *lemmas*, as opposed to explicit words because we were interested in the degree to which paragraphs overlapped in their general meaning (i.e., global overlap/cohesion), as opposed to specific words.

Findings

Rubric scores were correlated with the TAACO indicators as indicated in Table 1. The significant relationships between rubric facets and TAACO indicators vary across rubric facets, indicating that along with some general relationships, there were distinct textual features associated with each facet of the rubric design, as discussed further below.

Table 1: TAACO Indicators

	§1: Basic Indices		
Rubric Feature	Number of Words	Number of Types	Type/Token Ratio
Synthesis	-.117*	-.177**	-.071
Topic Coverage	.447**	.487**	-.271**
Source Diversity	.395**	.456**	-.220**
Source Quality	.137*	.089	-.243**
Rubric Feature	**§2: Sentence Overlap Indices**		
	Adjacent Overlap (All)	Adjacent Overlap (Content)	Adjacent Overlap (Function)
Synthesis	.260**	.231**	.059
Topic Coverage	-.161**	-.134*	-.029
Source Diversity	-.120*	-.117*	.011
Source Quality	.164**	.170**	-.002
Rubric Feature	**§3: Connectives Indices**		
	Basic Connectives	Sentence Linking	Reason and Purpose
Synthesis	.133*	.251**	.147*
Topic Coverage	-.137*	-.164**	-.089
Source Diversity	.038	-.161**	-.168**
Source Quality	-.043	-.003	.118*
Rubric Feature	**§4: Paragraph Overlap Indices**		
	Adjacent Overlap (All)	Adjacent Overlap (Content)	Adjacent Overlap (Function)
Synthesis	.035	.113	-.027
Topic Coverage	.119*	.101	.126*
Source Diversity	.150*	.110	.173**
Source Quality	.130*	.187**	.080

Synthesis. Across relationships to the synthesis facet, there were a number of significant relationships, which can be aligned with the theorized account above. Longer texts tended to include less synthesis (§1), with more sentence-level (but not paragraph-level) overlap (§2 & 4) and associations to sentence-level and purposeful connectives (§3), indicating higher levels of local cohesion establishing links between connected ideas through use of overlapping words (between sentences) and explicit links (connectives), but perhaps thematic shifts between paragraphs (with no relationship to paragraph level indices).

Topic coverage. Across relationships to the topic coverage facet we see strong relationships (§1) to number of words and types but negative relationship to type/token ratio indicating that lexical diversity, or information given – rather than number of words per se – is related to this facet. The facet is negatively related to local cohesion (§2) and connectives (§3), indicating that higher topic scores perhaps tended to involve more 'listing' of claims from sources, with less integration of those claims on a local level (a feature observed in the scoring exercise) although there is some paragraph level (§4) cohesion perhaps indicating general themes threaded through participant outputs.

Source diversity. Across relationships to the source diversity facet we see similar relationships to those indicated in topic coverage. In both cases, a level of global cohesion is indicated at the rhetorical level; that is, they are maintaining similarities in the way that they are talking, even where talking about different content. However, differences can be seen in task approach to developing cohesion such that those scoring high on topic coverage tended to jump between listed ideas, while source diversity scores might be more related to establishing clusters of claims for which multiple sources were cited (and logical connectives used to link these).

Source quality. Across relationships to the source quality facet we again see relationships to lexical diversity (or information given) in the type/token ratio (§1), and (as in synthesis) a relationship to sentence overlap (§2) indicating that local cohesion was being built (suggesting local argumentation focused on specific topics). We also see associations to paragraph overlap (§4) indicating that those who evaluated tended to build a cohesive

argument through their text, making purposeful connections (§3) between sentences. Thus while connectives generally are associated with synthesis (as above), evaluation (in source quality writing) is associated only with purposeful connectives.

Conclusions and implications

Being able to read, select information from, critique, and synthesis from multiple – oft competing – documents is an important skill, which often manifests in the writing of reports. Understanding the textual features of such written outputs is important for developing techniques to support writing tasks, including the potential of automated or semi-automated feedback to students through the use of NLP technologies. The potential, then, is for the assessment of written texts along an epistemic rubric, in which particular textual moves or features are associated with particular epistemic stances or cognitions. This work has given a preliminary demonstration of the potential conceptual and empirical alignment between such features. The potentials drawn highlight some significant relationships, although their sizes are generally small, but does not discuss the range of non-significant relationships or relationships between textual features that mediate rubric scores; further work should analyse this issue. We demonstrate that different cohesion categories varied in their relation to scores on an epistemically-aligned rubric, in line with previous work finding that variability in cohesion can be indicative of differences in cognitive processes (Allen, McNamara, & McCrudden, 2015). Overall, this work demonstrates the importance of investigating the fine-grain properties of students' writing. Further work should examine these differences using deeper analyses and study designs to probe varying cohesion types at a fine grain.

References

Allen, L. K., McNamara, D. S., & McCrudden, M. T. (2015). Change your Mind: Investigating the Effects of Self-Explanation in the Resolution of Misconceptions. Presented at the Society for Text & Discourse, Minneapolis, Minnesota, USA.

Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences*, *30*, 64–76. http://doi.org/10.1016/j.lindif.2013.01.007

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

Bråten, I. (2008). Personal Epistemology, Understanding of Multiple Texts, and Learning Within Internet Technologies. In M. S. Khine (Ed.), *Knowing, Knowledge and Beliefs* (pp. 351–376). Dordrecht: Springer Netherlands.

Bråten, I., Braasch, J. L. G., Strømsø, H. I., & Ferguson, L. E. (2014). Establishing Trustworthiness when Students Read Multiple Documents Containing Conflicting Scientific Evidence. *Reading Psychology*, *0*(0), 1–35. http://doi.org/10.1080/02702711.2013.864362

Bråten, I., Britt, M. A., Strømsø, H. I., & Rouet, J.-F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. *Educational Psychologist*, *46*(1), 48–70. http://doi.org/10.1080/00461520.2011.538647

British Psychological Society. (2014). Code of Human Research Ethics. British Psychological Society.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 1–11.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In *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 *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236–1241).

Crossley, S. A., & McNamara, D. S. (2014). Developing Component Scores from Natural Language Processing Tools to Assess Human Ratings of Essay Quality. In *The Twenty-Seventh International Flairs Conference*.

de Waard, A., Buitelaar, P., & Eigner, T. (2009). Identifying the epistemic value of discourse segments in biology texts. In *Proceedings of the Eighth International Conference on Computational Semantics* (pp. 351–354). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ferguson, L. E. (2014). Epistemic Beliefs and Their Relation to Multiple-Text Comprehension: A Norwegian Program of Research. *Scandinavian Journal of Educational Research*, *0*(0), 1–22. http://doi.org/10.1080/00313831.2014.971863

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. W. Cottrell (Ed.), *Proceedings of the 18th Annual Cognitive Science Conference* (pp. 110–115). Lawrence Erlbaum, NJ.

Goldman, S. R., Lawless, K., Pellegrino, J. W., & Gomez, K. (2012). A Technology for Assessing Multiple Source Comprehension: An Essential Skill of the 21st Century. In M. Mayrath (Ed.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Information Age Publishing (IAP).

Goldman, S. R., Ozuru, Y., Braasch, J. L. G., Manning, F. H., Lawless, K. A., Gomez, K. W., & Slanovits, M. (2011). Literacies for learning: A multiple source comprehension illustration. In N. Stein L. & S. Raudenbush (Eds.), *Developmental science goes to school: Implications for policy and practice* (pp. 30–44). Abingdon, Oxon: Routledge.

Groza, T., Handschuh, S., & Bordea, G. (2010). Towards automatic extraction of epistemic items from scientific publications. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (pp. 1341–1348). New York, NY, USA: ACM. http://doi.org/10.1145/1774088.1774377

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in english*. Routledge.

Hastings, P., Hughes, S., Magliano, J. P., Goldman, S. R., & Lawless, K. (2012). Assessing the use of multiple sources in student essays. *Behavior Research Methods*, *44*(3), 622–633. http://doi.org/10.3758/s13428-012-0214-0

Knight, S., & Littleton, K. (2015). Developing a multiple-document-processing performance assessment for epistemic literacy. Presented at the The 5th International Learning Analytics & Knowledge Conference (LAK15): Scaling Up: Big Data to Big Impact, Poughkeepsie, NY, USA.

Kobayashi, K. (2014). Students' consideration of source information during the reading of multiple texts and its effect on intertextual conflict resolution. *Instructional Science*, *42*(2), 183–205.

Louwerse, M. (2001). An analytic and cognitive parametrization of coherence relations. *Cognitive Linguistics*, *12*(3), 291–316.

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*(2), 381–392.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication*, *27*(1), 57–86. http://doi.org/10.1177/0741088309351547

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

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, *47*(4), 292–330. http://doi.org/10.1080/01638530902959943

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

Mozilla. (2014). *Firefox*. Retrieved from https://www.mozilla.org/en-US/firefox/new/

OECD. (2013). PISA 2015: Draft reading literacy framework. OECD Publishing.

OECD, & Statistics Canada. (2010). Literacy in the Information Age - Final Report of the International Adult Literacy Survey. OECD.

Rouet, J.-F. (2006). *The Skills of Document Use: From Text Comprehension to Web-Based Learning* (First Edition edition). Mahwah, NJ: Routledge.

Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Information Age Publishing (IAP).

Shah, C. (2010). Coagmento-a collaborative information seeking, synthesis and sense-making framework. *Integrated Demo at CSCW*, 6–11.

Shah, C. (2014). *Coagmento*. Retrieved from http://coagmento.org/

Simsek, D., Buckingham Shum, S., Sandor, A., De Liddo, A., & Ferguson, R. (2013). XIP Dashboard: visual analytics from automated rhetorical parsing of scientific metadiscourse. Presented at the 1st International Workshop on Discourse-Centric Learning Analytics, Leuven, Belgium.

Acknowledgments

This work was conducted as part of the first author's PhD at the Open University, UK. We thank participating students at Maastricht University, and Dr Dirk Tempelaar there for his collaboration on the work. Our thanks also to Dr Chirag Shah and Matthew Mitsui at Rutgers for their collaboration on the use of Coagmento, and study design.