Computational Approaches in Writing Assessment

Li Liu, Barley Mak and Tan Jin[1]

[1]Faculty of Education, The Chinese University of Hong Kong
refilane@gmail.com, barleymak@cuhk.edu.hk, tjin@cuhk.edu.hk

Abstract
Since the 1990s, the writing performances of candidates have been analyzed in terms of surface-level measures, comprising the human counts of linguistic features/structures and the intuitive comparisons between texts (see Graesser, McNamara, Louwerse, & Cai, 2004; Reid, 1992). Recently, due to the developments in computational linguistics, corpus linguistics and natural language processing, computational approaches have been widely applied to obtain deeper-level measures in validating the writing assessment (Biber, 1988; Biber, Conrad, & Reppen, 1998; Jurafsky & Martin, 2000; Moore & Wiemer-Hastings, 2003). To this end, this paper reviews certain empirical studies using computational approaches in writing assessment. First, different computational approaches were compared in the context of writing assessment. Second, the use of computational tools in analyzing candidates' writing performances was discussed. Finally, two research agendas were proposed: on one hand, more features are needed to answer the call for better and more reliable predictors. On the other hand, computational tools are suggested to be used together with qualitative evaluations of essays to provide a more accurate and detailed picture of candidates' writing performances.

Keywords
computational approach, writing assessment, candidate performance

1 Introduction
Performance tests have been widely used, taking the place of multiple-choice tests, in the assessment of writing proficiency (see Liu, 2011). Writing performance tests are featured by eliciting candidates' writing performances, resulting in labor-intensive scoring efforts by human raters. Computational tools have therefore been developed for evaluating the quality of essays to mitigate and assist human raters' scoring work. Using the computational tools, comprehensive linguistic features of essays are analyzed—which human raters are not expected to attend to—with systematic and detailed feedback reported. Few attempts, however, have been made to review the computational tools employed in the writing performance tests. To this end, this paper examines the computational approaches adopted in analyzing the writing performances.

2 Computational approaches to writing assessment
In the 1990s, the quality of essays was evaluated in terms of the manual counts of linguistic features/rhetorical structures and/or the intuitive comparison of experts (Reid, 1992). Such evaluation methods, however, were probably influenced by the fallibility of manual counts and the subjectivity of human raters' intuitive judgments. To provide more objective measures, computational tools were thus developed and used. Reid (1992), one of the pioneers, used a computer text-analysis program, the Writer's Workbench (WWB), to identify the syntactic and rhetorical differences of essays among four groups of writers with different language backgrounds. The WWB was subsequently employed to examine the cohesion differences among the four groups of writers, with surface-level cohesion features in the texts computed and analyzed.

Later, Grant and Ginther (2000) employed the computerized tagging system developed by Biber (1988) to analyze L2 learners' essays at different proficiency levels. The computer system tagged and analyzed the linguistic features of the essays at both micro and macro levels, comprising five major categories, namely, *general*, *lexical specificity*, *lexical features*, *grammatical features* and *clause level features*. These five categories were further operationalized as twenty-three features. The results of analyzing these features demonstrated the effectiveness of using the computer tagging system in the writing assessment.

Recently, due to the advances in various disciplines, such as computational linguistics, corpus linguistics and natural language processing, computational approaches have been widely applied to extract and measure the deeper-level features in validating the writing assessment (Biber, 1988; Biber et al., 1998; Jurafsky & Martin, 2000; Moore & Wiemer-Hastings, 2003). These features can be analyzed through a variety of computer-generated measures, including the automated lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters and other measures, which have been developed in the fields of natural language processing, computational linguistics and

cognitive science (Jurafsky & Martin, 2008). Taken together, the computational approaches make it possible that the analyses of many deep-level discourse features can be automated, providing more accurate and detailed information of candidates' writing performances.

A recent synthesis in the field has been achieved by the Coh-Metrix—a computational tool that employs over 600 linguistic indices related to written discourse features (Graesser et al., 2004). The Coh-Metrix has been employed widely to examine the linguistic structures of texts, and to explore textual differences in L2 writing studies (see Crossley & McNamara, 2008). In addition, a number of validation studies have been conducted on Coh-Metrix measures of cohesion (Crossley, Greenfield, & McNamara, 2008) and of lexical indices (Crossley et al., 2009).

Computational tools, however, are not without limitations; many researchers have reported problems with computer analyses (Frase, Faletti, Ginther, & Grant, 1997; Granger, 2002). It is argued that the measures derived from the computational tools are not able to represent the entire lexical, syntactical or rhetorical features within a language (Frase, Faletti, Ginther, & Grant, 1997).

3 Discussion and summary

The current paper has reviewed certain empirical studies employing computational approaches in the context of writing assessment. Future studies are suggested from two aspects. First, the balance among technology, construct relevance and applications in different contexts should be taken into consideration, although a variety of features can be analyzed by the computer. Qualitative evaluations of discourse features such as content and organization can also be employed to provide a more accurate and detailed picture of candidates' writing performances (see Liu, Mak, & Jin, 2012, April). Second, future studies employing computational approaches are needed to identify the features that can characterize candidates' writing performances at different proficiency levels (Jin & Mak, 2012).

References

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

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Crossley, S. A., & McNamara, D. S. (2008). Assessing second language reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, Mc- Carthy, & McNamara (2007). *Language Teaching, 41*, 409–429.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475-493.

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer Analysis of the TOEFL Test of Written English (TOEFL Research Report 64)*, Educational Testing Service, Princeton, NJ.

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.

Grant, L. & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*, 123 – 145.

Jin, T., & Mak, B. (2012). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing*. Advance online publication (21 May 2012). doi: 10.1177/0265532212442637.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Liu, L. (2011). *Towards a definition of second language writing ability: A literature review*. Qualifying Examination of PhD Candidature (PAPER I). The Chinese University of Hong Kong.

Liu, L., Mak, B., & Jin, T. (2012, April). Measuring content quality of writing performance: Putting content in its place. Paper presented at *the 34th Language Testing Research Colloquium* (LTRC 2012), Princeton, New Jersey, USA.

Moore, J. D., & Wiemer-Hastings, P. (2003). Discourse in computational linguistics and artificial intelligence. In A. C. Graesser, M. A. Gernsbacher, and S.R. Goldman (Eds.). *Handbook of discourse processes* (pp.439-486). Mahwah, NJ: Erlbaum.

Reid, J. R. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1,* 79–107.