Detecting Manipulated Second Language Reading Texts

Scott A. Crossley*, Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, Max M. Louwerse, Danielle S. McNamara

*Department of English, Mississippi State University, Mississippi State, MS 39762 (scr544@msstate.edu)
Department of Psychology, University of Memphis, Memphis. TN 38152
(pmccarthy, glewis, d.dufty, mlouwerse, d.mcnamara @mail.psyc.memphis.edu)

An important debate in the field of second language learning concerns assumed differences between *authentic* and *simplified* text. Texts are *simplified* through techniques such as increasing high frequency vocabulary, controlling for connective and abstract language, revising complex syntax, and increasing redundancy within text to reduce cognitive processing load (Simenson 1987).

Crossley (2006) argued that moderate, shallow, textual changes can significantly affect discourse structures and have the potential to affect discourse processing. Crossley also demonstrated that teachers of English could reliably distinguish these two text-types. In this study, we build on this research by evaluating the use of the computational tool, *Coh-Metrix* (Graesser et al., 2004) as a means to replicate human ability to distinguish the two text types.

For this analysis, we constructed a corpus from both authentic and manipulated second language reading texts..In total, 224 texts used for second language instruction were excerpted from 11 intermediate reading textbooks. The dataset was divided into a training set (n=113 texts) and a test set (n=111 texts). Based on Crossley (2006), we selected five predictors from Coh-Metrix: *logical connectors*, *textual abstractness/ambiguity*, *syntactic complexity*, *lexical co-reference*, and *word information*. The final three predictors were selected to closely match those used by humans when successfully discriminating between authentic and manipulated texts (Crossley, 2006).

Results and Discussion

An initial analysis of variance conducted on the training set suggested a number of significant differences between the two text types. Authentic texts contained more logical connectors ($F(1,112)=27.57$, $p<.001$), had greater abstractness in the form of higher verb hypernymy scores ($F(1,112)=4.21$, $p<.05$), and also demonstrated greater syntactic complexity ($F(1,112)=10.69$, $p<.001$). Manipulated texts, on the other hand, contained greater co-reference ($F(1,112)=7.53$, $p<.05$) and had lower lexical age of acquisition scores ($F(1,112) = 4.48$, $p<.05$).

We also conducted a discriminant function analysis on the training set using *text-type* (authentic or manipulated) as the dependent variable. The derived algorithm when applied to the entire dataset correctly allocated 156 of the 224 texts, an average accuracy rate of 70% (N=224, $\chi 2=33.55$, $p<.001$). Using the test data set only, the accuracy of the analysis remained high, with 67 of the 110 texts correctly allocated: an average accuracy rate of 60% (N=111, $\chi 2= 4.55$, $p<.05$).

The findings indicate the degree and type of differences between manipulated and authentic texts. The results also suggest that computational tools such as *Coh-Metrix* can be used to distinguish groups of highly similar text types comparably to humans.

Future studies will analyze how artificially modifying texts according to a few simple pedagogical principles may cause unintended consequences for the overall structure of the discourse and potentially affect how the text is processed, comprehended, and understood by second language readers.

Acknowledgments

This research was supported by the Institute for Education Sciences (IES R3056020018-02).

References

Crossley, S. A. (2006). *A computational approach to assessing second language reading texts*. Unpublished doctoral dissertation. University of Memphis.

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

Simensen, A. M. (1987). Adapted readers: How are they adapted? *Reading in a Foreign Language*, *4*, 41-57.