Identifying Text Genres Using Phrasal Verbs

Kyle B. Dempsey, Philip M. McCarthy, and Danielle S. McNamara
Department of Psychology
Memphis. TN 38152

{kdempsey, pmccarthy, d.mcnamara} @mail.psyc.memphis.edu)

Understanding the textual distinction between spokenness-informality and writtenness-formality serves many purposes. It can facilitate text mining, improve parser accuracy, offer better appraisals of student writing, and may also facilitate better interpretations of experimental data. Previous studies of such textual variation (e.g., Biber, 1988, Louwerse et al., 2004) have failed to produce a simple and effective method for computationally distinguishing these text types. Indeed, Biber (1988) using 67 lexical features could not determine any spoken/written dimension and Louwerse et al. (2004) using over 200 textual indices could not identify a formal/informal dimension.

In this study, we tested the hypothesis that *phrasal verbs* could distinguish such text-types because the presence of this verb construction is often claimed to be indicative of both spoken and less formal discourse (e.g., McWhorter, 2001).

To test our hypotheses, we used Coh-Metrix (Graesser et al., 2004) to calculate the incidences of various phrasal verbs forms across two corpora: the Biber Corpus (Biber, 1988; Louwerse et al., 2004); and a larger, yet structurally identical second corpus. For the spoken/written distinction, we used texts identified in the Louwerse et al. first dimension (LSWD). For the formal/informal distinction, we used tests identified in the Biber fifth dimension (BFID).

Results and Discussion

We conducted a series of ANOVAs on the incidence of phrasal verbs across both text distinctions of both corpora. We also examined correlations between the incidence of phrasal verbs and the degrees of spokenness and informality for each corpus. Overall, we found a significant difference in the incidence of phrasal verbs in both the LSWD texts $F(1,480) = 100.469$, $MSE=27.188$, $p<.001$, and the BFID texts, $F(1,480) = 23.103$, $MSE=31.369$, $p<.001$. We also found a significant correlation between the rank ordering of texts by incidence of phrasal verbs and the order of degree of spokenness in the LSWD texts ($r=.464$, $p< .001$), as well as the incidence of phrasal verbs and the degree of informality in BFID texts ($r=.579$, $p<.001$). The results suggest that phrasal verbs are significant markers for distinguishing differences in both spoken/written and formal/informal distinctions.

In a second experiment, we performed the same analyses on a larger *mirror corpus* of texts, containing 1028 texts as compared to 482 texts in the Biber corpus. Overall, we found a significant difference in the incidence of phrasal verbs LSWD texts, $F(1,1026) = 441.359$, $MSE=28.616$, $p<.001$, and the BFID texts $F(1,1026) = 206.210$, $MSE=34.077$, $p<.001$. We also found a significant correlation between the rank ordering of texts by incidence of phrasal verbs and the order of degree of spokenness in the LSWD texts ($r=.611$, $p< .001$), as well as the incidence of phrasal verbs and the degree of informality in BFID texts ($r=.656$, $p<.001$). The results supported the findings from Experiment 1 and suggest that phrasal verbs are significant markers for identifying spokennness and informality in texts.

Our study suggests that phrasal verbs offer an efficacious and computationally inexpensive approach to identifying the degree of textual *spokenness* and *informality*. Such an index serves to benefit research in both textual mining and text analysis tools. A better understanding of textual composition serves the learning community by increasing the accuracy of textual appraisals, facilitating better feedback to researchers, students, and authors alike.

Acknowledgements

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

References

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

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

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.

McWhorter, J.H. (2001). The power of Babel: A natural history of language. Times Books: Henry Holt and Company, New York.