# Using Phrasal Verbs as an Index to Distinguish Text Genres

**Kyle B. Dempsey, Philip M. McCarthy, Danielle S. McNamara**

Department of Psychology
Institute for Intelligent Systems
University of Memphis
Memphis. TN 38152
{kdempsey, pmmccrth, dsmcnamr} @ memphis.edu)

**Abstract**

Previous studies have shown that text genres can be computationally distinguished by sophisticated computational and statistical methods. The current study adds to the previous body of work by incorporating *phrasal verbs* as a text genre identifier. Results indicate that phrasal verbs significantly distinguish between both the spoken/written and formal/informal dimensions, with considerably less computational expense than previous studies. Phrasal verbs also indicate degree of *spokenness* and *formality* that is significantly similar to previous computationally expensive studies. The study offers useful findings for text-identification research and for materials developers in the field of English as a second language.[1]

## Introduction

Computationally distinguishing spoken registers from written registers has been an ongoing goal of corpus linguistic research (Biber, 1988; Louwerse et al., 2004). For example, Biber (1988) used 67 shallow lexical features to study the variation in spoken versus written registers, and while he was able to report many differences within these registers, he was not able to identify an empirically defined spoken/written dimension. More recently, Louwerse et al. (2004) identified a single dimension of spoken/written registers, but only by using the significantly more sophisticated indices of cohesion and readability made available through a computer system called Coh-Metrix (Graesser et al., 2004).

One linguistic characteristic that neither Biber (1988) nor Louwerse et al. (2004) considered was the incidence of phrasal verbs. Phrasal verbs are verbs plus one or more particles that behave as a syntactic and semantic, and often idiomatic, unit (Rudzka-Ostyn, 2003). Phrasal verbs have been identified as a potentially strong indicator of text genre. For example, Simpson and Mendis (2003) investigated the possibilities of variation in idiom frequency across different types of academic conversation transcriptions. Their study reported that as the particular situation changed, the type and frequency of idioms changed. As idioms are often in the form of phrasal verbs (Rudzka-Ostyn, 2003), this research leads us to hypothesize that differences in frequencies of these verbs across different types of communicative registers may facilitate making those registers more computationally identifiable. Previous research has also supported a view of phrasal verbs being a lexical marker. Biber (1987) marked phrasal verbs as a lexical phenomenon possibly occurring in varying frequency between formal and informal texts. Lastly, Darwin and Gray (1999) reported that phrasal verbs more commonly occur in freshman texts; that is, the papers of students who presumably have least experience with formal written language production. The implications for the current study are that idiomatic phrasal verbs occur in both major modes of communication to varying degrees and, therefore, that they may serve to help researchers better identify text genres.

To conduct our current study, we analyzed phrasal verb occurrences across two corpora: the first being the same corpus used for the two major previous studies (i.e., Biber, 1988; Louwerse et al., 2004), and the second being a different corpus of texts that mirrors the registers of the first corpus while more than doubling the number of texts.

## Purpose of the Study

This study serves three major purposes. First, textbook writers for English as a Second Language (ESL) may be supplied with a better understanding of the prevalence and type of phrasal verbs and idioms. Second, researchers may be better able to identify and sort the register and quality of texts during processes such as text mining and Q&A systems (McCarthy, Briner, et al., 2006). These two fields would benefit from the current study by further narrowing the computer-driven output into more manageable samples. Finally, identifying phrasal verbs as a marker of text genre can be beneficial to current computational tools such as Coh-Metrix (Graesser et al., 2004). While this tool has already proven to be effective in a wide variety of text identification studies (e.g., Louwerse et al., 2004; McCarthy, Briner, et al., 2007; McCar

thy, Lewis, et al., 2006), the addition of phrasal verbs as a marker of text genre can increase the power of Coh-Metrix as a textual analysis tool.

### Predictions

Biber (1988) created an empirical separation of spoken versus written text by using a corpus of spoken text, the London-Lund corpus (LLC), and a corpus of written text, Lancaster-Oslo-Bergen (LOB). Biber's factor analysis focuses on 67 lexical features but was unable to computationally separate the two modes. However, because phrasal verbs are believed to be more common in spoken and informal registers (CITE), we predicted that phrasal verbs frequency would be a sufficient marker to computationally discriminate between spoken and written texts. Specifically, we predicted that phrasal verbs would appear more often in spoken text than in written text.

Biber (1988) identified his fifth dimension, based on degree of abstractness, as a formal/informal distinction. Biber ranked the registers within this dimension from most informal (non-abstract) to most formal (abstract). Since previous studies have suggested that phrasal verbs are more likely to be informal than formal (Biber, 1987; McWhorter, 2001), we predicted a higher frequency of phrasal verbs at the informal end of Biber's fifth dimension.

An equally interesting dichotomy reported by Louwerse et al. (2004) was their first dimension distinction between spoken and written texts. Louwerse and colleagues developed this dimension into a ranking of text registers from most spoken to most written. Since phrasal verbs are more likely contained in spoken text than in written text (Biber, 1987; McWhorter, 2001), we also predicted a higher frequency of phrasal verbs at the spoken end of the Louwerse and colleagues' first dimension. Our predictions, therefore, were that phrasal verb frequencies would correlate significantly with Biber's fifth dimension and Louwerse et al.'s first dimension, as these are the dimensions that most closely reflect the function of phrasal verbs. Such findings would supply evidence that phrasal verbs frequency is indicative of established text dimensions.

In addition, as the Biber fifth dimension has its registers ranked in terms of *degree of informality*, and as the Louwerse and colleagues first dimension ranks their registers in terms of *degree of spokenness*, we further predicted that correlations between these individual rankings and those produced by frequencies of phrasal verbs would be significant. As such, we predicted that not only would phrasal verbs distinguish text genres along the lines of formality and spokenness, but that they would also distinguish text genres in terms of degree of formality and spokenness.

## Experiment 1

To examine whether phrasal verbs predicted spoken/written variation and formal/informal variation, we first analyzed the same corpora as those used in the Biber (1988) and Louwerse et al. (2004) studies (see Table 1). Note, however, that as Biber's study used an unobtainable private collection of personal and professional letters, we used the letters from the Louwerse and colleagues' study.

Table 1: Registers used in Biber (1988) and Louwerse et al. (2004).

| Corpus | Register |
|---|---|
| Lancaster-Olso-Bergen corpus | Press reportage, editorials, press reviews, religion, skills and hobbies, popular lore, biographies, official documents, academic prose, general fiction, mystery fiction, science fiction, adventure fiction, romantic fiction, humor |
| London-Lund corpus | Face-to-face conversation, telephone conversation, public conversations, debates, interviews, broadcast, spontaneous speeches, planned speeches |
| Additional | Personal letters, professional letters |

Because the spoken corpus and written corpus differed in terms of length, only the first 1000 words from each text were considered to control for a text length confound. For a corpus of phrasal verbs, we used the 397 most commonly occurring phrasal verbs, supplied in Hart (1999). To account for grammatical variation, the phrasal verbs were transformed into five forms: *base form* (e.g., *go around*), *3rd person* (e.g., *goes around*), progressive form (e.g., *going around*), 2nd form (e.g., *went around*), and 3rd form (e.g., *gone around*). Only unique transformations were considered.

The frequencies of phrasal verbs were calculated using a Visual Basic program specifically written for this study. The counts were separated by verb form and text. Counts were also normalized before being output for data analysis purposes.

### Results

An ANOVA was performed on the frequency of occurrence of phrasal verbs in the empirical separation of spoken versus written text: LOB (written)/LLC (spoken). The means are shown in Table 2. The frequencies in this analysis and in the remaining analyses tended not to be normally distributed. Therefore, we conducted the Mann-Whitney non-parametric test (i.e., U in Tables) as well as an ANOVA (F in Tables). With the exception of the frequency of 2nd forms, all categories of phrasal forms significantly distinguished spoken from written registers.

Our second question regarded the relevance of phrasal verbs to the formal versus informal distinction made by Biber (1988). Both a Mann-Whitney non-parametric test and an ANOVA were performed on the frequency of occurrence of phrasal verbs in informal versus formal texts as separated by Biber (see Table 3). With the exception of the frequency of 3rd persons, all categories of phrasal forms significantly distinguished formal from informal registers.

Table 2: Occurrence of Phrasal Verbs in Written (W) and Spoken (S) Text (Biber, 1988).

| | W | S | F | U |
|---|---|---|---|---|
| Base | 2.31 (2.25) | 4.58 (3.23) | 76.94** | 12949.5** |
| 3rd Person | 0.27 (0.63) | 0.65 (1.86) | 11.47* | 21377.5* |
| Prog-ressive | 0.59 (0.66) | 1.42 (1.49) | 58.30** | 15379.5** |
| 2nd form | 2.52 (2.41) | 2.91 (2.56) | 2.39 | 21531.0 |
| 3rd form | 0.64 (0.94) | 1.81 (1.94) | 78.63** | 144149.0** |
| Total | 6.35 (4.59) | 11.37 (6.64) | 90.34** | 12482.5** |

Notes: Standard deviations are in parenthesis; * $p < .05$; ** $p < .001$

Table 3: Occurrence of Phrasal Verbs in Formal (Frm) versus Informal (Inf) Text (Biber, 1988).

| | Frm | Inf | F | U |
|---|---|---|---|---|
| Base | 2.24 (1.94) | 3.29 (3.00) | 15.19** | 20194.0* |
| 3rd Person | 0.39 (0.75) | 0.38 (1.27) | 0.01 | 22949.0 |
| Prog-ressive | 0.60 (0.83) | 0.94 (1.25) | 9.12* | 21445.0* |
| 2nd form | 2.17 (1.98) | 2.84 (2.62) | 7.67* | 21397.0* |
| 3rd form | 0.55 (0.83) | 1.16 (1.56) | 20.17** | 19058.5** |
| Total | 5.95 (3.64) | 8.61 (6.27) | 23.10** | 18861.5** |

Note: Standard deviations are in parenthesis; * $p < .05$; ** $p < .001$

Our third question concerned the relevance of phrasal verbs to the spoken versus written distinction made by Louwerse et al. (2004). The means are shown in Table 4. An ANOVA and the Mann-Whitney non-parametric test were performed on the frequency of occurrence of phrasal verbs in spoken versus written texts as separated by Louwerse and colleagues. All categories of phrasal forms significantly distinguished Louwerse and colleagues' distinction of spoken/written registers.

Our final question in Experiment 1 regarded the correlations between the frequency of phrasal verbs and the rankings reported in the Biber (1998) and the Louwerse et al. (2004) studies. The Biber study provided a ranking of the registers along an informality scale ranging from most informal (spontaneous speeches) to most formal (science texts). The Louwerse et al. (2004) study provided a ranking for the registers along a spokenness scale ranging from most spoken (interviews) to most written (professional letters). The correlations were computed between the frequency of occurrence of phrasal verbs and spokenness as postulated by Louwerse et al. (2004). As both Biber and Louwerse and colleagues' studies used factor analyses, the difference in degree between registers is difficult to assess. As such we focus on the rank positions of the registers. Such a test is non-parametric and, therefore, a Spearman correlation was conducted. The results of the correlation were significant ($r = .464$, $p < .001$), indicating a high degree of similarity between the rank order of the register findings in the Louwerse et al. study and the frequencies of phrasal verbs. The Spearman correlation between the frequency of occurrence of phrasal verbs and informality as given by Biber (1988) was also significant ($r = .579$, $p < .001$), indicating that phrasal verb frequencies are also highly indicative of formal/informal differences of register.

Table 4: Occurrence of Phrasal Verbs in Written (W) versus Spoken (S) Text (Louwerse et al., 2004).

| | W | S | F | U |
|---|---|---|---|---|
| Base | 2.28 (2.28) | 4.44 (3.13) | 73.33** | 13708.0** |
| 3rd Person | 0.26 (0.64) | 0.59 (1.78) | 7.96* | 23477.5 |
| Prog-ressive | 0.57 (0.84) | 1.41 (1.46) | 63.16** | 15971.0** |
| 2nd form | 2.43 (2.35) | 3.07 (2.64) | 7.18* | 21393.5* |
| 3rd form | 0.61 (0.89) | 1.77 (1.91) | 83.61** | 15046.5** |
| Total | 6.17 (4.54) | 11.29 (6.43) | 100.47** | 12576.5** |

Note: Standard deviations are in parenthesis; * $p < .05$; ** $p < .001$

## Experiment 2

To confirm the validity of these findings, we conducted the same experiment on a *mirror corpus* of that used in Experiment 1. This mirror corpus replaced the LOB texts with texts of similar length and register from the Brown corpus (Kucera & Francis, 1967) but with the main

difference being number of text used. In total, 600 more texts were included in the mirror corpus. The use of the Brown corpus to serve as a mirror was appropriate for two reasons: First, the LOB corpus was based on the Brown corpus and therefore the two corpora contain the same definition and number of registers. Second, a number of previous studies focusing on extending the findings of Biber (1987, 1988) have been conducted on the Brown corpus as opposed to the LOB corpus (e.g., Karlgren & Cutting, 1994; Kessler, Nunberg, & Schutze, 1997). The LLC texts from the original corpus were similarly replaced by spoken texts from the Wellington Corpus of Spoken English (WCS, Holmes, 1995). As the WSC corpus does not mirror the LLC corpus quite as conveniently as does Brown to LOB, we arranged the divisions of the WSC corpus into mirror registers of the LLC. Personal letters were added to the mirror corpus from a free online source (http://www.openletters.net), and the professional letter mirror corpus was compiled from business letters sent internally at the University of Memphis. As such, the complete mirror corpus matched all registers in the original corpus and more than doubled the number of texts available for analysis.

## Results

To examine whether phrasal verbs predicted spoken/written variation in the new corpus, an ANOVA and a Mann-Whitney were performed on the frequency of occurrence of phrasal verbs in the Brown and Wellington corpora for the empirical separation of spoken versus written texts (Brown/WSC). All categories of phrasal forms significantly distinguished spoken from written registers (see Table 5). This was true even of the 2nd person forms which were not significant in the smaller, original corpus of Experiment 1.

Table 5: Occurrence of Phrasal Verbs in Written (W) versus Spoken (S) Text (Brown/WSC).

| | W | S | F | U |
|---|---|---|---|---|
| Base | 2.33 | 5.36 | 76.94** | 59549.5** |
| | (2.34) | (3.46) | | |
| 3rd | 0.30 | 0.57 | 26.45** | 114531.5** |
| Person | (0.63) | (0.97) | | |
| Prog- | 0.62 | 1.74 | 150.44** | 80877.5** |
| ressive | (0.92) | (1.85) | | |
| 2nd | 2.43 | 3.60 | 42.51** | 95320.0** |
| form | (2.84) | (2.88) | | |
| 3rd | 0.66 | 1.95 | 187.70** | 72446.5** |
| form | (1.01) | (1.88) | | |
| Total | 5.68 | 12.26 | 369.24** | 49771.5** |
| | (4.85) | (6.06) | | |

Note: Standard deviations are in parenthesis; * $p < .05$; ** $p < .001$

Our second question regarded the relevance of phrasal verbs to the formal versus informal distinction made by Biber (1988). An ANOVA and a Mann-Whitney were performed on the frequency of occurrence of phrasal verbs in the Brown and Wellington corpora for informal versus formal texts as separated by Biber. The means are shown in Table 6. With the exception of 3rd person forms, all categories of phrasal forms significantly distinguished formal from informal registers. These results mirror the smaller, original corpus of Experiment 1.

Table 6: Occurrence of Phrasal Verbs in Formal (Frm) versus Informal (Inf) Text (Biber, 1988)

| | Frm | Inf | F | U |
|---|---|---|---|---|
| Base | 1.93 | 4.49 | 132.30** | 55374.0** |
| | (1.87) | (3.45) | | |
| 3rd | 0.36 | 0.46 | 2.64 | 116085.5** |
| Person | (0.63) | (0.88) | | |
| Prog- | 0.44 | 1.43 | 84.08** | 75640.0** |
| ressive | (0.66) | (1.70) | | |
| 2nd | 1.77 | 3.44 | 68.22** | 88864.5** |
| form | (1.80) | (3.10) | | |
| 3rd | 0.60 | 1.55 | 70.00** | 69408.0** |
| form | (1.00) | (1.74) | | |
| Total | 4.51 | 10.49 | 206.21** | 44293.5** |
| | (3.08) | (6.52) | | |

Note: Standard deviations are in parenthesis; * $p < .05$; ** $p < .001$

Table 7: Occurrence of Phrasal Verbs in Written (W) versus Spoken (S) Text (Louwerse et al., 2004).

| | W | S | F | U |
|---|---|---|---|---|
| Base | 2.18 | 5.36 | 305.46** | 53697.5** |
| | (2.20) | (3.44) | | |
| 3rd | 0.31 | 0.55 | 21.58** | 98853.5 |
| Person | (0.64) | (0.96) | | |
| Prog- | 0.57 | 1.74 | 166.71** | 64266.5** |
| ressive | (0.87) | (1.83) | | |
| 2nd | 2.24 | 3.72 | 69.69** | 68421.0** |
| form | (2.64) | (2.99) | | |
| 3rd | 0.62 | 1.94 | 198.65** | 65461.5** |
| form | (0.97) | (1.86) | | |
| Total | 5.30 | 12.31 | 441.36** | 44296.0** |
| | (4.48) | (6.04) | | |

Note: Standard deviations are in parenthesis; * $p < .05$ ** $p < .001$

Our third question concerned the relevance of phrasal verbs to the spoken versus written distinction made by Louwerse et al. (2004). An ANOVA and a Mann-Whitney were performed on the frequency of occurrence of phrasal verbs in the Brown and Wellington corpora for spoken versus written texts as distinguished by the Louwerse et al.

(1988) 1st dimension. The means are shown in Table 7. All categories of phrasal forms significantly distinguished spoken from written registers. These results thus mirror the smaller, original corpus of Experiment 1.

As in Experiment 1, the correlation between the frequency of occurrence of phrasal verbs and the degree of formality as given by Biber (1988) was significant ($r$ = .656, p<.001). Likewise, the correlation between the frequency of occurrence of phrasal verbs in the Brown and Wellington corpora and the degree of spokenness as given by Louwerse et al. (2004) was significant ($r$ = .611, $p$ < .001).

## Discussion

The current study explored the contributions of phrasal verbs as a lexical identifier of 1) the empirical spoken/written distinction as taken from the LOB and LLC corpora; 2) Biber's 5th dimension (informal versus formal); and 3) the Louwerse et al. (2004) 1st dimension (spoken versus written). Also, the current study explored the correlations between the rankings of text registers along ranges of formality (Biber, 1988) and spokenness (Louwerse et al., 2004).

The results of this study suggest that phrasal verbs are highly indicative of Biber's formality/informality dimension, and also of Louwerse et al.'s spoken/written dimension. These results are in line with our hypotheses. The results of Experiment 2 confirm those of Experiment 1 and suggest that the findings are unlikely to be chance and, more importantly, that a phrasal verb discriminator index is reliable across British, American, and New Zealand English.

The results of the ANOVAs suggest that phrasal verbs are able to make significant distinctions between spoken/written texts and informal/formal texts that are highly similar to previous genre identification studies (i.e., Biber, 1988; Louwerse et al., 2004). The correlation results suggest that the distinctions made by phrasal verbs are also highly similar to the ranks of spokenness and formality reported by Louwerse et al. (2004) and Biber (1988) respectively.

The idiomatic nature of phrasal verbs has caused significant difficulty for generations of English language students. Frequently, these students tend to overuse Latinate verb forms such as *descend* instead of the phrasal counterpart *go down* simply because the structure is easier or more familiar (being a cognate). Having a better understanding of where and when phrasal verbs occur will help materials develops to better focus introductions and explanations of phrasal verbs.

Better knowledge of the frequency of occurrence of phrasal verbs also facilitates the fields of text mining and Q&A systems. As phrasal verb frequencies are a good indicator of both spoken/written and formal/informal dimensions, natural language processing researchers can have a better idea of the kind of texts they are retrieving. Phrasal verbs counts may prove particularly valuable to these fields because assessing their frequency is computationally inexpensive.

At a finer grain level, the individual analyses of the verb types indicate that each form may have something unique to offer both genre identification studies and ESL material developers. Specifically, the *spokenness* dimension is not significant in terms of 2nd forms (presumably past tense forms), whereas the *formal* distinction is not significant for 3rd person singular forms (presumably present tense text). The immediate impact of this finding is twofold. First, genre identification research can benefit from distinctions based on phrasal verb forms as well as frequencies which would help direct future tool development. Second, ESL material developers can create study units on phrasal verbs to better match the grammatical forms in which they are more commonly found. Moreover, they could even include more spoken or informal texts as a manner of increasing difficulty of study for ESL students, therefore increasing phrasal verb incidence so as to gauge mastery of the language.

Future research will expand the examination of phrasal verbs to the examination of idiomatic speech and collocations by expanding the search tokens to include common occurring idioms and collocations. Analysis will also examine interactions within the search tokens.

The current study provides evidence that word level analysis can give insight into the text genre. While sophisticated indices such those used in Louwerse et al. (2004) indices certainly provide new and powerful information as to the construction of discourse, more shallow indices such as phrasal verbs clearly also have much to offer textual research. The inclusion for these basic level indices, therefore, can supplement other indices currently available on tools such as Coh-Metrix.

## Acknowledgements

This research was supported by the Institute for Education Sciences (IES R305G020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors would also like to thank Dr. Max Louwerse for his contribution to this paper.

## References

Biber, D. 1987. A textual comparison of British and American writing. *American Speech*, 62, 99-119.

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

Darwin, M., and Gray, L. 1999. Going after the phrasal verb: an alternative approach to classification. *TESOL Quarterly, 33*, 1.

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

Hart, C. W. 1999. *The ultimate phrasal verb book*. Hauppauge, NY: Barron's.

Holmes, J. 1995. The Wellington Corpus of Spoken New Zealand English: A Progress Report. New Zealand English Newsletter: 5-8.

Karlgren, J. and Cutting, D. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of COLING*.

Kessler, B., Nunberg, G., and Schütze, H. 1997. Automatic detection of text genre. In *Proceedings of the 35th ACL/8th EACL*, pp. 32-38.

Kucera, H., and Francis, W. N. 1967. *Computational analysis of present-day English*. Providence, RI: Brown University Press.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., and 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.

McCarthy, P.M., Briner, S.W., Rus, V., and McNamara, D.S. 2007. Textual Signatures: Identifying text-types using Latent Semantic Analysis to measure the cohesion of text structures. In: A. Kao, S. Poteet (Eds.). *Natural Language Processing and Text Mining*. UK: Springer-Verlag.

McCarthy, P.M., Lewis, G.A., Dufty, D.F., and McNamara, D. S. 2006. Analyzing writing styles with Coh-Metrix. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)*, Melbourne, Florida.

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

Rudzka-Ostyn, B. 2003. Word power: Phrasal verbs and compounds: A cognitive approach. Berlin, New York. Mouton de Gruyter.

Simpson, R., and Mendis, D. 2003. A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 37 (Fall) 419-441.