Dimensions of collocation in American English

Tony Berber Sardinha (Sao Paulo Catholic University, Brazil)

Ever since the ground-breaking work by John McH. Sinclair in the 1960s (Sinclair, 1966), collocation (the habitual co-occurrence of two words at a short distance from each other) has established itself as one of the mainstays of corpus-based research. Although J. R. Firth (1957/1968) is credited with making collocation "not just as an observable effect of language use, but as an important element of the causes of language patterns" (Barnbrook, Krishnamurthy, & Mason, 2013, p. 36), it is Sinclair who has been considered the 'discoverer' of collocation (Hoey, 2009, p.39), due to the detailed evidence that his analyses of computer corpora provided in support of the existence of collocation in language use. A large number of studies have focused on collocation since then, looking at various aspects of the concept, from the association between lexical patterning and meaning (Moon & Sinclair, 1987), to the actual patterns of collocation for particular words (Stubbs, 2002), among other features. At the same time, research on collocation has generally ignored the relationship between collocation and text varieties (Biber, 2010, p. 245), focusing primarily on patterns of collocation that cut across register differences. An exception is Sinclair, Jones, and Daley (1970/2004), who compared collocations in a science magazine (*New Scientist*) and in conversation and found that the collocations could discriminate between the two registers. The authors concluded that:

> from a linguistic point of view it is interesting to find that 'strength of collocation' provides a useful discriminant between different types of English and it would be interesting to see if the results were so encouraging for two texts which differ very little. (p. 133)

This paper attempts to fill the gap in collocation studies by reporting the results of a multidimensional study on collocations from a register perspective using the 450-million-word Corpus of Contemporary American English (Davies, 2012; see Table 1).

Register	Tokens	
Spoken	90,786,821	20.6%
Magazine	90,780,789	20.6%
Newspaper	87,131,579	19.8%
Academic	86,512,881	19.6%
Fiction	85,907,930	19.5%
Total	441,120,001	100%

Table 1: Composition of COCA 2012, downloadable full-text version

The method for this investigation was inspired by the multidimensional (MD) analysis of register variation, introduced by Biber in the 1980s (Biber, 1988) and subsequently developed by him and colleagues (cf. Berber Sardinha & Veirano Pinto, 2014). The goal of an MD analysis is to determine the dimensions or underlying parameters of variation in the data (see Friginal & Hardy, 2014, for an overview of the method). Several major differences exist between a mainstream MD analysis and the MD analysis carried out here. First, in this investigation, the units upon which the analysis was based were collocations and not texts—more specifically, pairs of words, with one representing a node and the other, a collocate (these nodes and collocates were selected from among the most frequent words in each register in COCA). Second, in this investigation, the measurements

taken for each unit were not text counts, but the log-dice, a word association statistic (Rychly, 2008) that gauged the attraction between the two words. Third, the factor scores were calculated for the collocates of each node rather than the texts in the corpus. Finally, the basis for the interpretation of the factors in this investigation was based primarily (but not solely) on lexical features revealed by their semantic preference (Stubbs, 2007), lexical sets (Sinclair & Jones, 1974/1996), word fields (Lehrer, 1974; Trier, 1931), 'aboutness' (Phillips, 1989; Scott, 2000; Yablo, 2016), topics (Berber Sardinha, 1997), and subject matter (Schütze, 1998), and not primarily on functional / communicative grounds.

A program designed for this project retrieved the most salient node-collocate pairs in each register. The resulting data matrix consisted of 3,511 columns (one for each node) and 23,602 rows (one for each collocate). Each cell in the data matrix contained the log-dice value for the node-collocate pair. The log-dice statistic measures the degree of lexical association between a node and a collocate in a span of four words on either side of the node. The dimensions were determined through a series of factor analyses carried out in SAS University Edition.

Nine dimensions of collocation were identified: 1. Literate discourse; 2. Oral discourse; 3. Objects, people, and actions; 4. Colloquial and informal language use; 5. Organizations and the government; 6. Politics and current affairs; 7. Feelings and emotions; 8. Cooking; and 9. Education. Some of the major collocations that typify each dimension are presented in Table 2.

Dimension	Examples
1. Literate discourse	|issue~n + relate-v|, | factor~n + relate-v|, |seem~v + appropriate-j|, |specific-j + area~n|, |assessment-n + tool~n|
2. Oral discourse	|want~v + know-v|, |people~n + know-v|, |want-v + say~v|, |people~n + think-v|, |think-v + go~v|
3. Objects, people, and actions	|stare-v + window~n|, |stare-v + ceiling~n|, |slide-v + open~j|, |pull-v + trigger~n|, |car~n + pull-v|
4. Colloquial and informal language use	|afraid-j + lose~v|, |mama-n + papa~n|, |mama-n + daddy~n|, |mommy~n + daddy-n|, |glad-j + hear~v|
5. Organizations and the government	|protection~n + agency-n|, |official-n + say~v|, |international-j + monetary~j|, |national-j + association~n|, |district-n + attorney~n|
6. Politics and current affairs	|other~j + politician-n|, |decline~v + interview-v|, |police~n + interview-v|, |deserve-v + credit~n|, |think~v + deserve-v|
7. Feelings and emotions	|feel~v + shame-n|, |feel~v + guilt-n|, |feel~v + rage-n|, |face~n + rage-n|, |feel~v + excitement-n|
8. Cooking	|mix-v + bowl~n|, |mix-v + ingredient~n|, |cup-n + sugar~n|, |add-v + heat~n|, |add-v + onion~n|
9. Education	|student~n + benefit-v|, |rate-v +

	scale~n|, |expose-v + student~n|, |expose-v + child~n|, |educate-v + public~n|

Table 2: Typical collocations in each dimension ('~' indicates a node, and '-' represents a collocate; order of node and collocates reflects order in which they are most often found in COCA).

The greatest mean scores for the registers on each dimension were the following: in dimensions 1 and 9, 'academic'; in dimension 2, 'spoken' and 'fiction'; in dimensions 3, 4 and 7, 'fiction'; in dimension 5, 'newspaper' and 'academic'; in dimension 6, 'spoken' ; and in dimension 8, 'magazines'. Despite these contrasts, the register differences were statistically negligible, due to the irregular distribution of collocation in language. This suggests that each individual dimension is not a reliable predictor of register differences. However, while collocations may not be strong predictor of register categories, it is possible that register categories are strong predictors of collocation. To this end, a discriminant function analysis (DFA) was employed, which used the factor scores of each collocate with each node on each dimension as input and produced discriminant equations that were used to place the collocation in its most likely register, based on its factor scores. A DFA was run in SPSS having as input the dimension scores of each collocation on each dimension as the dependent variable and the register categories as the independent variable. The 'leave one out' option was used in the DFA, and therefore each collocation was excluded from the model used to predict its classification, so as to prevent bias in the classification task. Four samples of different sizes were used—namely, 500, 1,000, 2,000, and 4,000 collocates per register; these consisted of the *n* collocates with the greatest scores per register. The best results were obtained with the 500-word-per-register sample, where the majority of collocations (56.7%) were assigned to their source registers at a rate nearly three times better than chance. The cases of cross-classification (where the collocations from one particular registers were attributed to a different register) were also examined. Magazine was the most cross-classified register, as its collocations were habitually assigned to spoken and newspaper. This suggests that magazine texts have both a 'spoken-like' and a 'newspaper-like' character. Cross-classification was not bi-directional, though. Newspaper collocations were less likely to be wrongly predicted as magazine collocations than the other way around. The least cross-classified register was academic, with 2/3 of its collocation being correctly attributed.

To summarize, this study presents a large-scale study of cross-register variation across American English collocations. Nine general groups of collocation (dimensions) were identified. When these nine dimensions are combined, they can predict the register from which a collocation occurs significantly better than chance. Overall, the results suggest that the use of collocation is sensitive to register constraints. The register differences associated with collocation use shown by this study provide another 'nail in the coffin' in the attempts to describe 'general English' or any other language, as if language were a homogenous whole.

References

Barnbrook, G., Krishnamurthy, R., & Mason, O. M. A. (2013). *Collocation: Applications and Implications*. Basingstoke: Palgrave Macmillan.

Berber Sardinha, T. (1997). *Automatic identification of segments in written texts.* PhD dissertation, AELSU/English Department, University of Liverpool, UK.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2014). *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. Amsterdam/Philadelphia, PA: John Benjamins.

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

Biber, D. (2010). What can a corpus tell us about registers and genres? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (1st ed., pp. 241-254). London ; New York, NY: Routledge.

Davies, M. (2012). Corpus of Contemporary American English. Available at corpus.byu.edu/full-text.

Firth, J. R. (1957/1968). A synopsis of linguistic theory, 1930-55. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth 1952–59* (pp. 168-205). London: Longmans.

Friginal, E., & Hardy, J. A. (2014). Conducting Multi-Dimensional analysis using SPSS. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber* (pp. 298-316). Amsterdam/Philadelphia, PA: John Benjamins.

Hoey, M. (2009). Corpus-driven approaches to grammar: The search for common ground. In R. Schulze & U. Römer (Eds.), *Exploring the Lexis-Grammar Interface* (pp. 34-48). Amsterdam: John Benjamins.

Lehrer, A. (1974). *Semantic Fields and Lexical Structure*. Amsterdam: North-Holland.

Moon, R., & Sinclair, J. M. (1987). The analysis of meaning *Looking up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary* (pp. 86-103). London: Collins.

Phillips, M. (1989). *Lexical Structure of Text*. Birmingham: ELR, University of Birmingham.

Rychly, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008* (pp. 6-9). Brno: Masaryk University.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*(1), 97-123.

Scott, M. (2000). Focusing on the text and its key words. In L. Burnard & A. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (Vol. 2, pp. 103-122). Frankfurt: Peter Lang.

Sinclair, J. M. (1966). Beginning the study of lexis. In C. E. Bazell (Ed.), *In Memory of J R Firth* (pp. 410-430). London: Longman.

Sinclair, J. M., & Jones, S. (1974/1996). English lexical collocations: A study in computational linguistics. In J. A. Foley (Ed.), *J M Sinclair on Lexis and Lexicography* (pp. 22-68). Singapore: UniPress.

Sinclair, J. M., Jones, S., & Daley, R. (1970/2004). *English Lexical Studies: The OSTI Report* (Vol. Ed. by Ramesh Krishnamurthy). London/New York: Continuum.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics, 7*(2), 215-244.

Stubbs, M. (2007). Quantitative data on multi-word sequences in English: the case of the word 'world'. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, Discourse and Corpora* (pp. 163-190). London: Continuum.

Trier, J. (1931). *Der deutsche Wortschatz im Sinnbezirk des Verstandes; die Geschichte eines Sprachlichen feldes*. Heidelberg: C. Winter.

Yablo, S. (2016). *Aboutness*. Princeton, NJ: Princeton University Press.