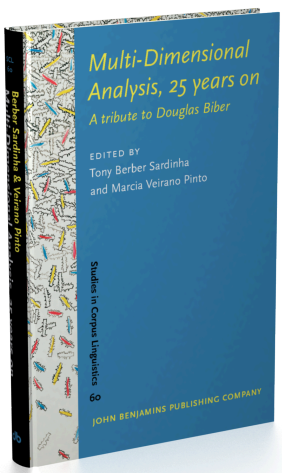
Studies in Corpus Linguistics 60

Opening. Multi-Dimensional analysis

A personal history

Douglas Biber | Northern Arizona University, USA

https://doi.org/10.1075/scl.60.005bib



Pages xxix–xxxviii of
Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber
Edited by Tony Berber Sardinha and Marcia Veirano Pinto
[**Studies in Corpus Linguistics**, 60] 2014. xxxviii, 328 pp.

Multi-Dimensional analysis

A personal history

Douglas Biber
Northern Arizona University, USA

This essay takes a personal perspective on the history of Multi-Dimensional (MD) Analysis, including motivations for the first MD studies, the influence of previous seminal research studies, and the influence of faculty colleagues at the University of Southern California. The essay also provides a short personal narrative on the development of the MD approach relative to the author's own background and interests. A brief survey of MD studies over the intervening decades is then followed by discussion of prospects, with discussion of what has been accomplished using this approach, and what remains to be done.

When I first began to study the linguistic similarities and differences between speech and writing, I never imagined that it would result in Multi-Dimensional (MD) analysis. In fact, I had no awareness of corpora at all. In the pilot study for my dissertation, I actually spent considerable time counting the occurrence of grammatical features in texts by hand! It was only later that I came to realize that the analysis of corpora provided an ideal research approach for investigating this issue.

I was extremely lucky in this enterprise to be in the right place at the right time. As an undergraduate, I had developed a strong background in science (with a degree in geophysics from Penn State University), including two courses in Fortran computer programming, and some research experience working on the computer modeling of earthquake fault zones in southern California. However, I did not really build on that experience after graduation. Rather, I spent time drawing seismic maps as a geophysicist; then went back to graduate school in theoretical linguistics; then supervised a Somali adult literacy program in northeast Kenya; and eventually ended up in the Ph.D. program in linguistics at the University of Southern California, where I initially focused my research efforts mostly on phonology and historical linguistics. I gradually shifted my interests to issues in sociolinguistics, focusing especially on spoken versus written discourse. But two mentors at USC had a major influence on me during this period, resulting in the development of the MD approach.

First, Ed Purcell helped me to develop the technical skills needed for MD analysis. Ed taught me both statistical analysis as well as advanced computer programming skills. Through Ed's courses, I learned how to carry out univariate and multivariate statistical analyses, with extensive discussion of how those techniques could be applied to linguistic research questions. And my development in computer programming skills occurred mostly as on-the-job training, when Ed hired me to work in a computer lab on campus. We worked on translating acoustic analysis software from Fortran to EDL (a computer language used on IBM Series/1 minicomputers), and in the process, I learned how to write software for linguistic analysis. That job led to a full-time position as a programmer in the university computing center, which placed me in the ideal position for working on the MD analysis for my dissertation in the evenings.

A second Ed – Ed Finegan – was central to my development as a corpus linguist, and as a researcher and writer in general. Ed was my dissertation chair and completely supportive of my general ideas to compare spoken and written discourse. But then one day in 1983, Ed told me that he had read an article about an electronic collection of texts (the Brown Corpus). I had never heard of a 'corpus' before, so didn't really know what it could do for me. But Ed suggested that I could apply my programming skills to corpus analysis, radically changing the methodology that I had intended to apply in my dissertation research on spoken and written discourse. Ed helped me obtain university funding to purchase the Brown Corpus, LOB Corpus, and London-Lund Corpus, providing the foundation for the first MD analyses.

In addition to the two Eds, there were several published papers that especially influenced my early work on developing the MD approach. First, there were theoretical discussions by linguists like Ervin-Tripp (1972), Hymes (1974), and Brown and Fraser (1979) who emphasized the importance of linguistic co-occurrence for the analysis of differences among registers (or 'speech styles'). So, for example, Brown and Fraser (1979, p. 38–39) argued that it can be 'misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the co-occurrence of sets of markers'. Chafe (1982) applied this concept to the comparison of speech and writing, proposing two parameters of linguistic variation: 'integration/fragmentation' and 'detachment/involvement'. Each of these parameters was composed of a set of related linguistic features. For example, the 'integration/fragmentation' parameter was composed of features like nominalizations, participles, and attributive adjectives versus clause co-ordination. Chafe identified these sets of linguistic features on an intuitive basis, but the notion that features work together as co-occurring sets was clearly evident in his work.

Of course, the distinctive methodological innovation of MD analysis was the application of factor analysis to empirically identify sets of linguistic features that

tend to co-occur in texts. This innovation had its roots in Carroll (1960) – a truly amazing study for its time, although I'm not sure I fully appreciated that fact in the early 1980s. Although the paper provides essentially no information on the methods for the linguistic analysis, we can only assume that it was done entirely by hand: counting the occurrence of 39 linguistic variables in 150 text passages (each 300 words in length). These counts were then subjected to a statistical factor analysis, carried out with 'the aid of high-speed electronic computing machines' (Carroll 1960, p. 288) – presumably an early version of a mainframe computer. Regardless of the methodological details, the resulting analysis identified six major 'vectors of prose style'. Each of these vectors was composed of subjective, perceptual variables co-occurring with objective, linguistic variables. Conceptually and methodologically, these vectors are very similar to the 'dimensions' in MD analysis. This seems to have been Carroll's only foray into the domain of linguistic stylistics (he was much more interested in language testing, human cognition, and psychometrics). However, the 1960 paper must have had a huge influence on my own thinking, helping me to realize that statistical factor analysis could be used to empirically identify the linguistic co-occurrence patterns that linguists had been positing on theoretical grounds.

So, with that background, I started in 1983 to develop a computer program – the first version of my grammatical tagger – to analyze lexico-grammatical characteristics in corpora. That version of the tagger (written in PL/1, a computer language that ran on IBM mainframes) was entirely rule-based, relying to a large extent on the grammatical descriptions found in Quirk, Greenbaum, Leech, and Svartvik (1972). Using this software, I tagged the LOB and London-Lund corpora, and then wrote another program to count the occurrence of 41 linguistic features in the texts of the corpora. I analyzed that data with factor analysis, providing the foundation for the first MD studies of speech and writing in English: my dissertation in 1984 (Biber 1984), and journal articles published in 1985 and 1986 (Biber 1985, 1986).

After I graduated with my Ph.D. in 1984, I accepted an Assistant Professor position at USC. I obtained a book contract with CUP to further pursue this area of research. For that project, I completely rewrote and extended the scope of my grammatical tagger, and re-ran the factor analysis on that expanded set of 67 linguistic features. That work resulted in my 1988 Cambridge book (Biber 1988), the study that most people recognize as the first MD analysis.

During my time at USC, I worked with colleagues (especially Ed Finegan) and Ph.D. students who were interested in applying the MD approach to the analysis of other languages and/or specialized discourse domains in English. This was one of the most stimulating periods of my academic life, with daily discussions about language variation and change, and new methods for capturing those patterns.

Ed and I talked almost daily about these issues, including long discussions during some great hikes in the San Gabriel Mountains. Those discussions resulted in several projects and published papers on historical register variation, including a 1989 article (Biber & Finegan 1989) published in *Language*, and a National Science Foundation grant to construct and analyze the ARCHER corpus.

Three of the Ph.D. students that I had the opportunity to work with during this period were interested in applying MD analysis to the study of other languages: Niko Besnier working on Nukulaelae Tuvaluan; YongJin Kim on Korean; and Mohamed Hared on Somali. Because I had spent three years in NE Kenya, I also knew Somali. Having Mohamed as a Ph.D. student at USC offered a unique opportunity for collaboration. As a result, I was able to obtain a National Science Foundation grant to carry out a major MD analysis of synchronic and diachronic register variation in Somali. This entailed fieldwork in Somalia (to construct the corpora), computational work to develop a grammatical tagger for Somali, tagging and tag-editing the Somali corpora, and then the synchronic and historical MD analyses. During this time, I also had the chance to work with Jeff Connor-Linton and Dwight Atkinson, who both applied the 1988 MD analysis to specialized discourse domains in English.

Since that time, the MD approach has been applied to many specialized discourse domains in English, as well as many other languages. For English, those studies include investigations of: elementary school spoken and written registers (Reppen 1994, 2001), job interviews (White 1994), 18th c. speech-based and written registers (Biber 2001), university spoken and written registers (Biber 2006), Google text types (Biber & Kurjian 2007), moves in science research articles (Kanoksilapatham 2007; Biber & Jones 2005), conversational text types (Biber 2008), call center discourse (Friginal 2009), World English spoken and written registers (Xiao 2009), written legal registers (Goźdź-Roszkowski 2011), blogs (Grieve, Biber, Friginal, & Nekrasova 2011), academic research articles across disciplines (Gray 2011), 19th c. fictional novels (Egbert 2012), and ESL spoken and written exam responses (Biber & Gray 2013). In addition, numerous studies have applied the 1988 dimensions of variation to study the relations among English registers in more specialized discourse domains (see, e.g. the papers in Conrad & Biber 2001).

Cross-linguistically, the approach has been applied to analyze register variation in an equally extensive set of languages, including Nukulaelae Tuvaluan (Besnier 1988), Somali (Biber & Hared 1992, 1994; cf. Biber 1995), Korean (Kim & Biber 1994), Taiwanese (Jang 1998), Spanish (Biber, Davies, Jones, & Tracy-Ventura 2006; Parodi 2007; Asención-Delaney & Collentine 2011; Asención-Delaney, in this volume), Czech (Kodytek 2008), Bagdani (Purvis 2008), and Brazilian Portuguese (Berber Sardinha, Kauffmann, & Acunzo, in this volume).

These MD studies of register variation have uncovered both surprising similarities and notable differences in the underlying dimensions of variation. Each of these MD analyses has identified dimensions that are specialized to a discourse domain or language, reflecting the particular communicative priorities of that language/culture or domain of use. For example, the MD analysis of Somali identified a dimension interpreted as 'Distanced, directive interaction', represented by optative clauses, 1st and 2nd person pronouns, directional pre-verbal particles, and other case particles. Only one register is especially marked for the frequent use of these co-occurring features in Somali: personal letters. This dimension reflects the specialized inventory of grammatical devices in Somali combined with the particular communicative priorities of personal letters in Somali, which are typically interactive as well as explicitly directive.

From both theoretical and methodological perspectives, it is not surprising that each MD analysis would uncover specialized dimensions that are peculiar to a given language and/or discourse domain. After all, each of these studies differs with respect to the set of linguistic features included in the analysis, and the set of registers represented in the corpus for analysis. Given those differences, it would be reasonable to expect that the parameters of variation that emerge from each analysis would be fundamentally different.

Given that background, it would be much more surprising to discover dimensions of variation that occurred across languages and discourse domains. However, two such dimensions have emerged in nearly all of these MD studies, making them candidates for universal parameters of register variation: a dimension associated with 'oral' versus 'literate' discourse, and a dimension associated with narrative discourse (see also Biber, to appear). The robustness of narrative dimensions across languages and discourse domains indicates that this rhetorical mode is basic to human communication, whether in speech or in writing. But the most surprising finding is the oral/literate opposition, which emerges as the very first dimension in nearly all MD studies.

In MD studies based on general corpora of spoken and written registers, this oral/literate dimension clearly distinguishes between speech and writing. However, MD studies of specialized discourse domains show that this is not a simple opposition between the spoken and written modes. In fact, this dimension emerges consistently in studies focused exclusively on spoken registers, as well as studies focused on written registers.

In terms of communicative purpose, the 'oral' registers characterized by this dimension focus on personal concerns, interpersonal interactions, and the expression of stance. These registers are usually produced in real time, with little or no opportunity for planning, revising, or editing. In contrast, 'literate' registers focus on the presentation of propositional information, with little overt acknowledgement of

the audience or the personal feelings of the speaker/writer. These registers usually allow for extensive planning and even editing and revising of the discourse.

Linguistically, this first dimension opposes two discourse styles: an 'oral' style that relies on pronouns, verbs, and adverbs, versus a 'literate' style that relies on nouns and nominal modifiers. The oral style relies on clauses to construct discourse – including a dense use of dependent clauses. In contrast, the complexity of the literate style is phrasal. This finding, replicated across languages, is especially surprising, because it runs counter to assumptions about syntactic complexity held by many linguists. But it is perhaps the most important and robust finding to emerge cross-linguistically from MD studies: spoken registers (and 'oral' written registers) rely on clausal discourse styles, including a dense use of dependent clauses; written registers (and 'literate' spoken registers) rely on phrasal discourse styles, especially the dense use of phrasal modifiers embedded in noun phrases (see also Biber & Gray 2011, Biber, Gray, & Poonpon 2011).

In sum, the patterns of variation observed across MD studies provide considerable empirical evidence to support the possibility of universals of register variation. One major need for future research is analysis of additional languages, to confirm the generalizability of these basic dimensions.

The converse focus – describing the specialized dimensions that emerge from each MD analysis – requires perhaps even more attention in future research. We need to better understand the underlying functional bases of these specialized dimensions and identify possible generalizable patterns across languages and discourse domains.

In many cases, these specialized dimensions reflect the particular communicative purposes and other situational characteristics of specialized registers found in the target discourse domain (e.g. differences between Introductions versus Methods sections of science research articles). Similarly, analyses of some languages/cultures will include specialized registers (like maneapa speeches in Nukulaelae Tuvaluan) not generally found in other languages/cultures, and it is likely that the MD analysis of those languages will uncover specialized dimensions associated with those registers. In other cases, specialized dimensions reflect the linguistic resources that are available in the language. For example, the 'Spoken irrealis discourse' dimension in the Spanish analysis reflects the existence of verb inflections for subjunctives and conditionals in that language. Similarly, the 'Honorification' dimension in Korean reflects the existence of honorific forms in that language.

But some of the apparent differences relating to these specialized dimensions across languages reflect the representativeness of the corpus, rather than genuine characteristics of the language/culture. Given the resources that are available on the Web, it is presently possible to construct a corpus that represents a much wider range of registers than what was considered feasible even two decades ago. And

as a result, MD analyses of these languages have identified specialized dimensions that reflect the communicative characteristics of the specialized registers included in the corpus. For example, the 2012 MD study of Brazilian Portuguese by Berber Sardinha, Kauffmann & Acunzo is based on an especially large and comprehensive corpus, and for that reason, it was able to identify specialized dimensions such as 'Evaluative discourse' (defined primarily by *que*-clause constructions and other kinds of stance devices; and distinguishing horoscopes and political speeches from other registers), and 'Procedural discourse' (defined primarily by present subjunctive verbs, imperative verbs, and subject pronoun-drop; and distinguishing recipes from most other registers).

Thus, one important methodological issue here concerns the corpus: How can we determine the extent to which a corpus represents the range of register variation in a language? I have been interested in this methodological issue since the early 1990s (see Biber 1990, 1993), and most recent textbooks on corpus linguistics also address the importance of this issue. Corpus size (how many texts; how many words) is one important consideration in this regard; but corpus composition is equally important, especially the extent to which we have represented the full range of register variation in a language.

These issues affect all quantitative corpus-based research – not just MD analyses. Research that disregards register differences leads to incomplete descriptions, and in some cases, inaccurate conclusions (see Biber 2012). Thus, there is a need in all corpus-based studies of language use to develop better methods for evaluating the register-representativeness of the corpus itself.

MD studies of specialized discourse domains have usually been exemplary in this regard, beginning with a situational description of the domain of use, followed by careful methods for sampling texts and sub-registers from across that domain. This ideal has also motivated the corpus design and construction utilized in MD analyses of cultures with a restricted range of written registers (e.g. Nukulaelae Tuvaluan, Bagdani, and even Somali, with its short history of written registers). In these cases, it has been feasible to sample texts representing essentially the full range of available registers in the culture/language.

However, that ideal has been more difficult to achieve for the analysis of languages/cultures with a long history of literacy. The corpora utilized for general MD studies of spoken and written register variation in a language have always attempted to include samples from across the spectrum of registers. But in most cases, it has not been feasible to include the full set of specialized registers in the corpus for these studies.

The Berber Sardinha, Kauffmann & Acunzo (2012) study of Brazilian Portuguese indicates that this situation is changing, in large part due to the resources of the Web. As a result, one important area of future research will be cross-linguistic

comparisons of MD analyses based on matched corpora, representing both the full range of general as well as more specialized registers in each culture/language.

There is strong evidence from MD studies to date that the robust, 'universal', dimensions of variation will emerge from the analysis of almost any corpus. These include the ubiquitous 'oral/literate' dimension, dimensions associated with narration, and dimensions associated with the expression of stance. But future analyses of the more specialized dimensions of variation across languages can be enhanced by paying more attention to the corpus designs, ensuring that corpora across languages are all representing the complete range of register variation available to sampling.

References

Asención, Y., & Collentine, J. (2011). A Multidimensional Analysis of a Written L2 Spanish Corpus. *Applied Linguistics, 32*, 299–322.

Berber Sardinha, T., Kauffmann, C., & Acunzo, C.M. (2012). Register variation in Brazilian Portuguese. Talk. Northern Arizona University.

Besnier, N. (1988). The linguistic relationships of spoken and written Nukulaelae registers. *Language, 64*, 707–736.

Biber, D. (1984). *A model of textual relations within the written and spoken modes*. (Unpublished doctoral dissertation). University of Southern California, Los Angeles, CA.

Biber, D. (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics, 23*, 33760.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, *62*, 384414.

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

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, *5*, 257–269.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 243–257.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic perspective*. Cambridge: CUP.

Biber, D. (2001). Dimensions of variation among eighteenth-century speech-based and written registers. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional studies* (pp. 200–214). London: Longman.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (2008). Corpus-based analyses of discourse: Dimensions of variation in conversation. In V. Bhatia, J. Flowerdew, & R. Jones (Eds.), *Advances in discourse studies* (pp. 100–114). London: Routledge.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory, 8*, 9–37.

Biber, D. (to appear). Using multi-dimensional analysis to explore cross – linguistic universals of register variation. *Languages in Contrast*.

Biber, D., & Finegan, E. (1989). Drift and the evolution of English style: A history of three genres. *Language, 65*, 487–515.

Biber, D., & Hared, M. (1992). Dimensions of register variation in Somali. *Language Variation and Change, 4*, 41–75.

Biber, D., & Hared, M. (1994). Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 182–216). Oxford: OUP.

Biber, D., & Jones, J.K. (2005). Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles. *Corpus Linguistics and Linguistic Theory, 1*,151–182.

Biber, D., Davies, M., Jones, J.K., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A Multi-Dimensional analysis. *Corpora, 1*, 7–38. DOI: 10.3366/cor.2006.1.1.1

Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: A Multi-Dimensional analysis. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 109–132). Amsterdam: Rodopi.

Biber, D., & Gray, B. (2011). Grammar emerging in the noun phrase: The influence of written language use. *English Language and Linguistics, 15*, 223–250.

Biber, D., Gray, B., Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*, 5–35.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking responses on the TOEFL iBT. Princeton, NJ: Educational Testing Service.

Brown, P., & Fraser, C. (1979). Speech as a marker of situation. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 33–62). Cambridge: CUP.

Carroll, J. (1960). Vectors of prose style. In T.A. Sebeok (Ed.), *Style in language* (pp. 283–292). Cambridge: CUP.

Chafe, W.L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35–54). Norwood, NJ: Ablex.

Conrad, S., & Biber, D. (Eds.) (2001). *Multi-Dimensional studies of register variation in English*. London: Longman.

Egbert, J. (2012). Style in nineteenth century fiction: A Multi-Dimensional analysis. *Scientific Study of Literature, 2*, 167–198. DOI: 10.1075/ssol.2.2.01egb

Ervin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In J. Gumperz & D. Hymes (Eds.), *Directions in Sociolinguistics* (pp. 213–250). New York, NY: Holt.

Friginal, E. (2009). *The language of outsourced call centers*. Amsterdam: John Benjamins.

Goźdź-Roszkowski, S. (2011). *Patterns of linguistic variation in American legal English*. Frankfurt: Peter Lang.

Gray, B. (2011). *Exploring academic writing through corpus linguistics: When discipline tells only part of the story*. (Unpublished Doctoral Dissertation). Northern Arizona University, Flagstaff, AZ.

Grieve, J., Biber, D., Friginal, E., & Nekrasova, T. (2011). Variation among blogs: A Multi-Dimensional analysis. In A. Mehler, S. Sharoff, & M. Santini (Eds.), *Genres on the web: Computational models and empirical studies* (pp. 303–322). London: Springer.

Hymes, D. (1974). *Foundations in Sociolinguistics*. Philadelphia, PA: University of Pennsylvania Press.

Jang, S.-C. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study*. (Unpublished doctoral dissertation). University of Hawaii, Manoa, HI.

Kanoksilapatham, B. (2007). Rhetorical moves in biochemistry research articles. In D. Biber, U. Connor, & T. A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure* (pp. 73–120). Amsterdam: John Benjamins.

Kim, Y.J., & Biber, D. (1994). A corpus-based analysis of register variation in Korean. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 157–181). Oxford: OUP.

Kodytek, V. (2008). *On the replicability of the Biber model: The case of Czech*. Unpublished manuscript.

Parodi, G. (2007). Variation across registers in Spanish. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 11–53). London: Continuum.

Purvis, T.M. (2008). *A linguistic and discursive analysis of register variation in Dagbani*. (Unpublished doctoral dissertation). Indiana University, Bloomington, IN.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *A Grammar of Contemporary English*. London: Longman.

Reppen, R. (1994). *Variation in elementary student writing*. (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, AZ.

Reppen, R. (2001). Register variation in student and adult speech. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional studies* (pp. 187–199). London: Longman.

White, M. (1994). *Language in job interviews: Differences relating to success and socioeconomic variables*. (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, AZ.

Xiao, R. (2009). Multi-Dimensional analysis and the study of World Englishes. *World Englishes, 28*(4), 421–450. DOI: 10.1111/j.1467-971X.2009.01606.x