Multi-Dimensional Analysis

Research Methods and Current Issues

Edited by Tony Berber Sardinha and Marcia Veirano Pinto
Multi-Dimensional Analysis
Also available from Bloomsbury

Research Methods in Applied Linguistics, Brian Paltridge
Corpus Linguistics and Linguistically Annotated Corpora, Sandra Kuebler
Experimental Research Methods in Language Learning, Aek Phakiti
Multi-Dimensional Analysis

Research Methods and Current Issues

Edited by
Tony Berber Sardinha and
Marcia Veirano Pinto
To Marilisa and Julia

Tony

To Walter, Otto, and Rafael

Marcia
## Contents

List of Illustrations ix
List of Contributors xv
Acknowledgments xvi

Introduction  *Tony Berber Sardinha* and *Marcia Veirano Pinto* 1

### Part One  Understanding the Principles: Origins of the Method, Corpus Design, and Annotation

1  Multi-Dimensional Analysis: A Historical Synopsis  *Douglas Biber* 11
2  Corpus Design and Representativeness  *Jesse Egbert* 27
3  Tagging and Counting Linguistic Features for Multi-Dimensional Analysis  *Bethany Gray* 43
4  The Multi-Dimensional Analysis Tagger  *Andrea Nini* 67

### Part Two  Conducting an MD Analysis: Quantitative and Qualitative Aspects

5  Multivariate Statistics Commonly Used in Multi-Dimensional Analysis  *Pascual Cantos-Gomez* 97
6  Doing Multi-Dimensional Analysis in SPSS, SAS, and R  *Jesse Egbert* and *Shelley Staples* 125
7  From Factors to Dimensions: Interpreting Linguistic Co-occurrence Patterns  *Eric Friginal* and *Jack A. Hardy* 145
8  Adding Registers to a Previous Multi-Dimensional Analysis  *Tony Berber Sardinha*, *Marcia Veirano Pinto*, *Cristina Mayer*, *Maria Carolina Zuppardi*, and *Carlos Henrique Kauffmann* 165

### Part Three  Exploring the Method

9  Examining Lexical and Cohesion Differences in Discipline-Specific Writing Using Multi-Dimensional Analysis  *Scott A. Crossley*, *Kristopher Kyle*, and *Ute Römer* 189
10 Using Discriminate Function Analysis in Multi-Dimensional Analysis  Marcia Veirano Pinto  217
11 Using Multi-Dimensional Analysis to Detect Representations of National Cultures  Tony Berber Sardinha  231

Index  259
List of Illustrations

Figures

Figure 2.1 Rates of occurrence (per million words) for verbs in COCA 33
Figure 3.1 Example interface for an interactive tag editing program (FixTag) 61
Figure 4.1 Screenshot of MAT interface for Windows 73
Figure 4.2 Mean scores and ranges for Dimension 1, Involved versus Informational Discourse, for a selection of Biber’s (1988) registers compared to the mean and range for MFTs 89
Figure 4.3 Mean scores and ranges for Dimension 2, Narrative versus Non-narrative Concerns, for a selection of Biber’s (1988) registers compared to the mean and range for MFTs 89
Figure 4.4 Mean scores and ranges for Dimension 3, Situation-Dependent versus Explicit Reference, for a selection of Biber’s (1988) registers compared to the mean and range for MFTs 90
Figure 4.5 Mean scores and ranges for Dimension 4, Overt Expression of Persuasion, for a selection of Biber’s (1988) registers compared to the mean and range for MFTs 90
Figure 4.6 Mean scores and ranges for Dimension 5, Abstract versus Nonabstract Information, for a selection of Biber’s (1988) registers compared to the mean and range for MFTs 91
Figure 4.7 Mean scores and ranges for Dimension 6, On-Line Informational Elaboration, for a selection of Biber’s (1988) registers compared to the mean and range for MFTs 92
Figure 5.1 Plot correlating factors and genres (adapted from Biber 1988, 18) 102
Figure 5.2 Mean factor scores and related genres (adapted from Biber and Finegan 1986, 32) 103
Figure 5.3 Rotated solution graph 105
Figure 5.4 Dendrogram of nine Russian near synonyms meaning “to try” (adapted from Gries 2009, 22) 108
List of Illustrations

Figure 5.5  Cluster analysis of fifteen text types based on one hundred common words (adapted from Nishina 2007, 31)  109
Figure 5.6  Dendrogram  110
Figure 5.7  Data display  111
Figure 5.8  Data display of nonhierarchical clustering  112
Figure 6.1  SPSS results containing the KMO statistic  134
Figure 6.2  SPSS scree plot  135
Figure 6.3  SPSS output containing variance explained by the factors in the model  135
Figure 6.4  SPSS output for complete factor matrix (only first five features displayed)  136
Figure 6.5  SAS results containing the KMO statistic  137
Figure 6.6  SAS scree plot  137
Figure 6.7  SAS output containing eigenvalues for each factor  138
Figure 6.8  SAS output for complete factor matrix (only first five features displayed)  138
Figure 6.9  R results containing KMO statistic  138
Figure 6.10  R scree plot  139
Figure 6.11  R output containing variance accounted for by each factor  139
Figure 6.12  R output for complete factor matrix (only first five features displayed)  139
Figure 7.1  Comparison of average factor scores for Dimension 1: Involved production (+) versus Information production (−). Adapted from Biber (1988, 128)  152
Figure 7.2  Group comparison of caller versus agent texts from Friginal’s Dimension 2  156
Figure 8.1  Calculation of a factor score in an Excel spreadsheet  178
Figure 8.2  Vertical plot of Web registers added to Biber’s (1988) Dimension 3, Explicit versus Situation-Dependent reference  180
Figure 8.3  Horizontal bar plot of Web registers added to Biber’s (1988) Dimension 3, Explicit versus Situation-Dependent reference  181
Figure 9.1  Dimension 1 loadings: Ease of function words  199
Figure 9.2  Dimension 2 loadings: Text simplicity  201
Figure 9.3  Dimension 3 loadings: Content word frequency  203
List of Illustrations

Figure 9.4  Dimension 4 loadings: Word overlap 205
Figure 9.5  Dimension 5 loadings: Function word repetition 206
Figure 10.1 SPSS main discriminant analysis screen 220
Figure 10.2 SPSS discriminant analysis: Statistics screen 220
Figure 10.3 SPSS discriminant analysis: Classification screen 221
Figure 10.4 SPSS discriminant analysis: Save screen 221
Figure 11.1 Scree plot for American 237
Figure 11.2 Scree plot for Brazilian 238
Figure 11.3 Sequential mean decade scores for Dimension 1, American 240
Figure 11.4 Sequential mean decade scores for Dimension 2, American 242
Figure 11.5 Sequential mean decade scores for Dimension 3, American 243
Figure 11.6 Sequential mean decade scores for Dimension 4, American 245
Figure 11.7 Sequential mean decade scores for Dimension 5, American 246
Figure 11.8 Sequential mean decade scores for Dimension 1, Brazilian 249
Figure 11.9 Sequential mean decade scores for Dimension 2, Brazilian 250
Figure 11.10 Sequential mean decade scores for Dimension 3, Brazilian 252
Figure 11.11 Sequential mean decade scores for Dimension 4, Brazilian 253
Figure 11.12 Comparison between Dimension 2 and 4, Brazilian 253
Figure 11.13 Sequential mean decade scores for Dimension 5, Brazilian 254

Tables

Table 1.1  MD studies of specialized discourse domains in English 18
Table 1.2  MD studies based on other linguistic characteristics 19
Table 3.1  Contingency table for calculating precision and recall for POS tags (adapted from Manning et al. 2008) 57
Table 3.2  Contingency table for calculating precision and recall for passives (data based on Gray 2015) 57
Table 4.1  Short descriptions and summary of the six dimensions of register variation for English found by Biber (1988) 68
Table 4.2  Short description and summary of the eight text types for English found by Biber (1989) 69
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.3</td>
<td>Comparison of dimension scores and distribution of text types between Biber’s (1988, 1989) analysis of the LOB corpus and a MAT analysis of the same corpus</td>
<td>74</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Comparison of dimension scores and distribution of text types between Biber’s (1988, 1989) analysis of the LOB corpus and a MAT analysis of the Brown corpus</td>
<td>79</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Comparison of dimension scores and distribution of text types between the analysis of the LOB corpus and the analysis of the Brown corpus with MAT</td>
<td>84</td>
</tr>
<tr>
<td>Table 4.6</td>
<td>Comparison of mean dimension scores and standard deviations for MFTs and personal and professional letters from Biber (1988)</td>
<td>88</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Factors and factor loadings (adapted from Biber and Finegan 1986, 28)</td>
<td>101</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Rotated sums of squared loadings</td>
<td>104</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Rotated component matrix with suppressed loadings &lt;.4</td>
<td>105</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>Example of a distance matrix</td>
<td>109</td>
</tr>
<tr>
<td>Table 5.5</td>
<td>Agglomeration schedule</td>
<td>111</td>
</tr>
<tr>
<td>Table 5.6</td>
<td>Cluster centers</td>
<td>112</td>
</tr>
<tr>
<td>Table 5.7</td>
<td>Cluster membership (two clusters)</td>
<td>114</td>
</tr>
<tr>
<td>Table 5.8</td>
<td>Cross-validation classification results</td>
<td>116</td>
</tr>
<tr>
<td>Table 5.9</td>
<td>American movie dimensions of variation (Berber Sardinha and Veirano Pinto 2015, 82)</td>
<td>117</td>
</tr>
<tr>
<td>Table 5.10</td>
<td>Classification results by genres (Berber Sardinha and Veirano Pinto 2015, 92)</td>
<td>118</td>
</tr>
<tr>
<td>Table 5.11</td>
<td>Variables that do not pass the tolerance test</td>
<td>119</td>
</tr>
<tr>
<td>Table 5.12</td>
<td>Eigenvalues</td>
<td>119</td>
</tr>
<tr>
<td>Table 5.13</td>
<td>Wilks’s lambda test</td>
<td>120</td>
</tr>
<tr>
<td>Table 5.14</td>
<td>Scores on different teaching methods</td>
<td>120</td>
</tr>
<tr>
<td>Table 5.15</td>
<td>ANOVA results</td>
<td>121</td>
</tr>
<tr>
<td>Table 5.16</td>
<td>Duncan’s multiple test</td>
<td>121</td>
</tr>
<tr>
<td>Table 5.17</td>
<td>Test of homogeneity of variance</td>
<td>122</td>
</tr>
<tr>
<td>Table 6.1</td>
<td>Summary statistics for the meta-analysis of 23 MD analyses</td>
<td>132</td>
</tr>
<tr>
<td>Table 6.2</td>
<td>SAS syntax and line-by-line descriptions</td>
<td>137</td>
</tr>
<tr>
<td>Table 6.3</td>
<td>R syntax and line-by-line descriptions</td>
<td>138</td>
</tr>
</tbody>
</table>
List of Illustrations

Table 7.1  Biber's (1988) co-occurring features of LOB and London-Lund  150
Table 7.2  Linguistic features of Friginal's Factor 2  154
Table 7.3  Linguistic features of Hardy and Römer's (2013) Factor 1, later named “Involved, academic narrative production versus descriptive, informational production”  160
Table 7.4  Linguistic features of Hardy and Römer's (2013) Factor 3, later named “Situation-Dependent, Non-Procedural evaluation versus Procedural discourse”  161
Table 8.1  Linguistic features in Biber (1988) Dimension 2  171
Table 8.2  Spreadsheet with the formula for the calculation of normed frequency counts  172
Table 8.3  Output from the Biber TagCount program  173
Table 8.4  Partial spreadsheet with normed frequency counts of Web registers  175
Table 8.5  Spreadsheet exemplifying the calculation of standardized counts  176
Table 8.6  Dimension 3: Explicit versus Situation-Dependent Reference (Biber 1988)  177
Table 8.7  Spreadsheet with the calculation of the mean dimension scores for the register encyclopedia for Dimension 3 (Biber 1988)  179
Table 8.8  Design of the Web corpus  179
Table 9.1  Details of the MICUSP sub-corpora used in this study  192
Table 9.2  Distribution of papers across paper types in MICUSP sub-corpora  193
Table 9.3  First ten eigenvalues from the principal component analysis  197
Table 9.4  Factor 1 loadings: Ease of function words  197
Table 9.5  Factor 2 loadings: Text simplicity  198
Table 9.6  Factor 3 loadings: Content word frequency  198
Table 9.7  Factor 4 loadings: Word overlap  198
Table 9.8  Factor 5 loadings: Function word repetition  198
Table 9.9  Descriptive statistics for weighted component 1  200
Table 9.10 Classification results for Dimension 1: Ease of function words  200
Table 9.11 Descriptive statistics for weighted component 2  202
Table 9.12 Classification results for Dimension 2: Text simplicity  204
Table 9.13 Descriptive statistics for weighted component 3  204
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.14</td>
<td>Classification results for Dimension 3: Content word frequency</td>
<td>204</td>
</tr>
<tr>
<td>9.15</td>
<td>Classification results for Dimension 4: Word overlap</td>
<td>206</td>
</tr>
<tr>
<td>9.16</td>
<td>Classification results for Dimension 5: Function word repetition</td>
<td>207</td>
</tr>
<tr>
<td>10.1</td>
<td>Spreadsheet with corpus data</td>
<td>219</td>
</tr>
<tr>
<td>10.2</td>
<td>Classification function coefficients</td>
<td>222</td>
</tr>
<tr>
<td>10.3</td>
<td>Classification results&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>225</td>
</tr>
<tr>
<td>11.1</td>
<td>Factorial pattern for American</td>
<td>239</td>
</tr>
<tr>
<td>11.2</td>
<td>ANOVAs and R&lt;sup&gt;2&lt;/sup&gt;, American</td>
<td>247</td>
</tr>
<tr>
<td>11.3</td>
<td>Factorial pattern for Brazilian</td>
<td>248</td>
</tr>
<tr>
<td>11.4</td>
<td>ANOVAs and R&lt;sup&gt;2&lt;/sup&gt;, Brazilian</td>
<td>256</td>
</tr>
</tbody>
</table>
List of Contributors

Berber Sardinha, Tony
São Paulo Catholic University, Brazil

Biber, Douglas
Northern Arizona University, USA

Cantos-Gómez, Pascual
University of Murcia, Spain

Crossley, Scott A.
Georgia State University, USA

Egbert, Jesse
Northern Arizona University, USA

Friginal, Eric
Georgia State University, USA

Gray, Bethany
Iowa State University, USA

Hardy, Jack A.
Emory College of Arts and Sciences, USA

Kauffmann, Carlos Henrique
São Paulo Catholic University, Brazil

Kyle, Kristopher
Georgia State University, USA

Mayer, Cristina
São Paulo Catholic University, Brazil

Nini, Andrea
Aston University, UK

Römer, Ute
Georgia State University, USA

Staples, Shelley
Purdue University, USA

Veirano Pinto, Marcia
São Paulo Federal University, Brazil

Zuppardi, Maria Carolina
São Paulo Catholic University, Brazil
Acknowledgments

Tony Berber Sardinha would like to thank both CNPq, Brasília, Brazil (research grants #303710/2013-6 and #306994/2017-8), and PIpeq, PUCSP, for funding his research.
Introduction

Tony Berber Sardinha and Marcia Veirano Pinto

Multi-Dimensional (MD) Analysis is a corpus linguistics approach developed by Douglas Biber in the mid-1980s for describing the patterns of variation in written and spoken registers in English. In the research tradition of MD Analysis, registers are “a cover term for any language variety defined by its situational characteristics, including speaker’s purpose, the relationship between speaker and hearer, and the production circumstances” (Biber 2009, 823), which can be defined in different degrees of specificity. Such conceptualization makes register the key element in the approach, a concept that captures the systematic nature of choices constrained by situational factors. Over the years, what was initially developed for modeling register variation in English was applied to a whole range of different languages (English, Somali, Korean, Spanish, Portuguese, etc.) and discourse domains (academia, movies, television, music, services, airlines, etc.). The evidence provided by MD studies of variation strongly suggests that register variation is the norm, rather than the exception, in language use (Berber Sardinha 2017; Biber 2012; Herrmann and Berber Sardinha 2015); as such, analysts should “begin a research study with the hypothesis that such register differences exist” (Biber 2012, 34) and be suspicious of accounts that see language as a unified whole. Language use varies a great deal from context to context and is therefore inherently extravagant. MD Analysis has consistently shown that this extravagance is patterned rather than chaotic—that is, large-scale systematic patterns of variation can be detected and described by means of factor analysis, which enables the detection of latent (i.e., not directly observed) variables from the observed variables, that is, the counts of linguistic characteristics; the latent variables are the dimensions of variation.

The patterns of register variation are grounded on the notion of linguistic co-occurrence (Ervin-Tripp 1972; Hymes 1974; Brown and Fraser 1979). For example, one of the main differences between natural conversation and academic prose is the presence of a set of co-occurring linguistic features, such as first and second person pronouns, contractions, *wh* questions, and the pronoun *it* in the former and nouns, prepositions, and attributive adjectives in the latter. The multivariate statistic procedure of factor analysis allows for the identification of such sets of co-occurring linguistic features, which define continuous spaces of variation, where registers are compared in relation to the extent of whether or not they share a range of linguistic features. These spaces of variation are then analyzed qualitatively—the assumption is that the co-occurrence of linguistic features shares situational, social, and cognitive functions—and the dimensions of a corpus emerge. Thus, dimensions of variation comprise both linguistic and functional content. The linguistic content of a dimension is the set of
linguistic features that co-occur with a high frequency in texts, whereas the functional content comprises the underlying situational, social, and cognitive functions. The dimensions enable registers to be explored as continuous constructs that are more or less linguistically similar to one another. The degree of similarity and difference across registers is determined by frequent and infrequent salient co-occurrence patterns within functional parameters of variation.

This book covers the main concepts, principles, procedures, statistics, and innovations associated with the MD Analysis framework. It is divided into three main parts. The chapters in Part 1 present the origins of the approach and discuss the principles of corpus design and annotation, which lay the groundwork for the analysis. The chapters in Part 2 examine the core phases in an MD study—namely, the statistical analyses, the factorial extraction, and the interpretation of the factors as dimensions; this section also presents the additive type of MD Analysis, which does not require a factorial extraction, as it “sits on top” of a previous “full” MD Analysis. Part 3 of the book introduces extensions to the MD approach, which explore novel applications, such as the analysis of cohesion, lexis, and discursive representations, as well as the application of a dimensional model in text classification.

Part 1 of the book consists of four chapters written by Biber, Egbert, Gray, and Nini. Biber’s chapter offers a historical panorama of MD Analysis, contextualizing the author’s personal history and chronicling when and how the basic ideas underlying the method came about. The author discusses his realization of the need for the introduction of computers in linguistic analysis and how his skills as a computer programmer helped him make the most of the technology available at the time in his early career as a linguist. The chapter provides a bibliographic history of the ideas that shaped the development of MD Analysis, including crucial studies that noted the importance of linguistic patterns of co-occurrence in differentiating text varieties (such as Ferguson 1959; Carroll 1960; Bernstein 1970; Longacre 1976; Irvine 1979; Ochs et al. 1979; Chafe 1982). The chapter presents an overview of MD studies looking at a wide range of specialized discourse domains (e.g., television, web, university, pop songs, conversational registers, world Englishes, regional dialects, nineteenth-century fictional style), including methodological extensions of the approach designed to meet such goals as the identification of text types (i.e., groups of texts that have similar MD characteristics), the automatic classification of texts into register categories based on MD scores, and the development of MD models that draw on linguistic characteristics not normally explored in MD studies (e.g., metaphors, bigrams, discourse markers, cohesion and coherence, collocation). The author points out that, despite the wide range of data and applications of MD Analysis, many studies converge toward the identification of “universal” dimensions distinguishing oral/literate and narrative/non-narrative parameters of variation.

Egbert’s chapter is devoted to a crucial element in corpus linguistic research in general and MD studies in particular—namely, the design of representative corpora. The author presents the state of the art in corpus representativeness and goes on to lay out the main principles that should guide corpus design, such as the careful consideration of the characteristics underlying the population sampled by the corpus, the independence between balance and representativeness, the relationship between corpus design and research design, and the considerations involved in determining
the size of the corpus. Egbert ends his chapter by arguing that it is the responsibility of corpus creators, researchers, and consumers of corpus-based publications to ensure accuracy in the description of language use and variation.

Gray’s chapter looks at the major issues involved in corpus annotation for MD Analysis. Annotating (tagging) a corpus for linguistic characteristics is a requirement for MD Analysis because MD studies use counts of these annotated features as data. Therefore, ensuring that the corpus has been accurately annotated and that the tagged features have been reliably tabulated is an absolute necessity for MD Analysis. The author discusses the major stages involved in the annotation and counting phases of an MD Analysis project—namely, using automatic taggers such as the Biber tagger and CLAWS for annotating the linguistic features, verifying the degree of accuracy in the annotation, identifying errors in and fixing the tagging of particular linguistic features, and counting the linguistic features in such a way that the counts are compatible with MD Analysis.

Nini closes Part 1 by presenting the Multidimensional Analysis Tagger (MAT), an integrated corpus analysis tool designed to help users analyze English corpora from an MD perspective. The tool utilizes the freely available Stanford tagger to emulate the tags used by the Biber tagger, which has restricted access. The author shows how MAT annotates the texts, finds and counts the features used in Biber (1988), plots the registers or text onto the six dimensions of variation included in the original study, assigns text types (Biber 1989), and enables the researcher to visualize the features belonging to each of the dimensions in each text. The tool is intended to provide a user-friendly environment for new and experienced MD analysts to conduct MD research on corpora of English texts.

Part 2 consists of four chapters as well, written by Cantos-Gomez; Egbert and Staples; Friginal and Hardy; and Berber Sardinha, Veirano Pinto, Mayer, Zuppardi, and Kauffmann. Cantos-Gomez presents an introduction to the multivariate statistics commonly used in MD Analysis: factor analysis, cluster analysis, discriminant function analysis (DFA), and analysis of variance. The author describes the common characteristics across all of these different multivariate statistics, highlighting their usefulness in tackling important problems in data analysis. The chapter details the steps involved in carrying out a multivariate statistical analysis in SPSS, a popular statistical package normally used in MD Analysis. The chapter comments on the tables and graphs reporting the results of the multivariate procedures, helping readers and analysts understand and focus on the most relevant points in the output.

Egbert and Staples focus on the single most important multivariate procedure involved in MD Analysis: factor analysis. Because of its central role in the method, factor analysis requires a detailed examination. The chapter includes instructions on how to perform the procedure using different statistical packages (SPSS, SAS, and R) and how to make the most appropriate decisions involved in successfully conducting a factor analysis of MD data. Although these different programs all provide the means for running a factor analysis, they differ considerably in how the procedure is actually conducted. The authors help analysts navigate the various options in the different programs by providing click-by-click instructions as well as the programming code needed to complete a factorial extraction with the different programs.
The chapter by Friginal and Hardy addresses the last major phase of a complete MD Analysis—namely, the interpretation of the factors as dimensions (i.e., the transition from statistical results to actual functional communicative parameters of linguistic variation). The interpretation of factors is a complex endeavor that must take into consideration the functions performed by the set of the features loading on the (poles of) the factors, which requires both linguistic expertise and familiarity with the situational contexts of the registers. The authors illustrate the decision-making process involved in interpreting the factors and devising appropriate labels for the dimensions with examples from Biber (1988) as well as from studies that looked at dimensions of variation in cross-cultural business communication and in the Michigan Corpus of Upper-level Student Papers (MICUSP).

The final chapter in Part 2, written by Berber Sardinha, Veirano Pinto, Mayer, Zuppardi, and Kauffmann, is devoted to additive MD analyses, a different kind of MD Analysis than the one focused on in the previous chapters. Unlike a “full” MD Analysis, an additive analysis does not require the extraction of factors; instead, it enables researchers to apply existing dimensions of variation (e.g., Biber 1988) to “new” registers. Adding registers to previous dimensions basically consists of calculating the dimension scores for a new corpus of texts using the mean and standard deviations from the previous reference study. Software programs exist that calculate these dimension scores automatically, simplifying the process of conducting additive analyses. Because it does not require conducting a factor extraction, additive analyses remove one of the major barriers for researchers who are not familiar with multivariate statistical techniques. In addition, an additive analysis is technically less demanding, making it a suitable entry point into the MD framework. Yet additive analyses are not less informative than full analyses. In this chapter, the authors review the major studies reporting additive analyses and show how they provide important insights into register variation, both synchronically and diachronically.

Part 3 of the book introduces innovations and extensions to the basic MD Analysis framework; it includes chapters by Crossley, Kyle, and Römer; Veirano Pinto; and Berber Sardinha. Crossley, Kyle, and Römer examine differences in discipline-specific student writing as found in the MICUSP using an MD Analysis approach. The features that inform their MD Analysis are reported by two natural language processing tools: Tool for the Automatic Analysis of Lexical Sophistication (TAALES) and Tool for the Automatic Analysis of Cohesion (TAACO). These tools consist of lexical and cohesive features of the texts, respectively. The chapter investigates student writing in science, engineering, social sciences, and social studies. The goal of the study is to better understand differences within and between student writing genres, to better understand the expectations of advanced student writing, and to better inform pedagogical practices at the university level.

Veirano Pinto discusses the use of DFA as a means to validate the dimensional model produced by an MD Analysis. More specifically, the author carried out a text classification task on a corpus of texts (movie dialogues) scored on a set of dimensions (identified by Veirano Pinto 2014), and measured the degree of success in the classification task. As it turns out, the texts were more frequently classified into their current register categories than expected by chance, which suggested that the
Introduction

5
dimensional model was robust enough to predict the register categories from the
dimension scores. The chapter provides a step-by-step guide to running a DFA in SPSS
using dimension scores, and Veirano Pinto argues that a DFA complements an MD
Analysis by providing further verification (beyond ANOVAs and the coefficient of
determination [$R^2$]) of the strength of the findings.

Berber Sardinha closes the book by introducing a lexical type of MD Analysis as a
means to identify representations of national identity. The chapter reports on a study
that analyzed the most salient two-word sequences beginning with either “American” or
“Brazilian,” as found in the Google Books N-Gram database. These items were entered
in individual factor analyses (one for “American” and the other for “Brazilian”), and the
sets of words loading on each factor were interpreted as discursive constructions that
both shape and reflect the public understanding of the cultures of the United States and
Brazil. The results suggest that the factors captured some of the major representations
of the United States and Brazil from the 1800s to the 2000s. The chapter is an example
of how MD Analysis can benefit from using lexical data as well as how it can be applied
discourse analysis.

We hope this book motivates readers to take up MD Analysis as a method for the
study variation in language use. As a major exponent of quantitative corpus linguistics,
the method requires a good deal of technical expertise, especially with respect to the
use of computer software and the application of multivariate statistical techniques.
Although the methodological requirements are adequately presented in the MD
literature, the level of detail is not normally sufficient to enable most linguists to carry
out an MD study with competence (an exception is Friginal and Hardy 2014, which
is a guide to MD Analysis in SPSS). The chapters herein cover the most important
technical elements of the method, providing step-by-step guidance, and we hope that
readers find these instructions clear and helpful. At the same time, we also hope that
the book inspires analysts to explore further avenues of research with MD Analysis;
the chapters in Part 3 of the book point out some possibilities, but many others surely
exist that are worth exploring. The MD framework offers a flexible method that can be
adapted to take into account different facets of language use, beyond register variation,
for many different languages. Extensions of the basic method provide exciting new
possibilities of research on language use. We believe that the MD approach has made
extraordinary contributions to corpus-based studies of language use over the course of
its thirty years of existence and, as such, it is one of the major developments not only
in corpus linguistics, but also in linguistic science. We are certain the MD approach
will continue to be a reference for quality language studies that have a profound impact
on how we understand variation in language use, and we hope this book will provide
useful guidance for more analysts to engage with the method.

Notes

1 Because of the central place occupied by the concept of register in the framework, it is
important to explore some of the basic similarities and differences between “register”
and its closely associated term “genre.” Both terms are frequently used in linguistics to
refer to a text variety defined in terms of its context of use, which exhibits particular linguistic patterns (cf. Biber and Conrad 2009, 21). In the MD tradition, however, the term “register” has been preferred to “genre,” although “genre” has been used in place of “register,” especially in the early literature (e.g., Biber 1988, 1989), as in Biber (1989, 5), where genres are defined as “text categories readily distinguished by mature speakers of a language.” However, in particular fields, the two terms have specific meanings; therefore, care must be taken to distinguish them. In Bakhtinian discourse analysis, for example, genres are “relatively stable types” of utterances that characterize particular spheres of activity (Bakhtin 1986, 60). In Systemic Functional Linguistics, in turn, genres are characterized as “staged, goal oriented social processes” and, as such, are “defined as a recurrent configuration of meanings” that “enact social practices of a given culture” (Martin and Rose 2008, 6). Register, on the other hand, is a configuration of the register variables of tenor (role relationships), field (action taking place), and mode (part language is playing) (Martin and Rose 2008, 11). Consequently, in systemics, genre and register vary independently: genre variation is the variation at the level of culture, whereas register variation is the variation at the level of situation. In addition to these theoretical distinctions, methodological distinctions exist between the terms, with each term identifying a particular analytical perspective. According to Biber and Conrad (2009, 16), a genre perspective would generally entail a focus on complete texts, specialized expressions, and rhetorical organization; a register perspective, in turn, focuses on either whole texts or text samples, any lexico-grammatical features, and the frequent and pervasive linguistic characteristics found in the texts.

The single most important statistical procedure used in MD Analysis is factor analysis, without which it would not be possible to determine the factors underlying the corpus and, consequently, the dimensions of variation, which are the end result of the interpretation of the factors. In an MD study, the factors are groupings of linguistic characteristics that co-occur in the texts. Factor scores are computed for each text based on the occurrence of the linguistic variables loading on the factor; as a result, each text is scored on each factor—that is, a number is attached to the text that represents the incidence of the group of co-occurring features. A high score means a high incidence of the set of linguistic variables associated with the factor, whereas a low score means a low incidence of such features. Following the factor extraction and scoring, an analysis of variation (ANOVA) is typically carried out on the dimension scores across the corpus components (e.g., registers) in order to gauge the amount of variation captured by each factor. As a matter of course, the output of an ANOVA produced by the major statistical packages will include two major statistics often used in MD analyses: the F score and the coefficient of determination. The F score indicates whether the variation in the data is statistically significant across the corpus components. The coefficient of determination (R²) specifies the amount of variation explained by these components. In other words, whereas the F scores tell the analyst whether the variation picking up the dimensions is significant, the R² indicates the strength of the variation in relation to the corpus components. Both statistical measures are important in MD analyses, as they provide information on whether there are statistical grounds to assume that variation actually exists in the data among the designated categories (e.g., registers) and the extent to which these categories are reliable predictors of the variation. Two additional statistical procedures may be used in conjunction with an MD Analysis: cluster analysis and discriminant function analysis. A cluster analysis may be carried out on the dimension scores in order to determine the text types present in the corpus. A text type, in the Biberean tradition, is a “grouping of texts that are similar in their linguistic form” (Biber 1989, 13). The groupings of texts that give rise to a text type
are determined based on their dimension scores; texts with similar dimension scores are placed together in a cluster "such that the texts within each cluster are maximally similar to each other in the exploitation of the textual dimensions, while each cluster is maximally distinct from the others" (Biber 1989, 13). Unlike a factor analysis, in which all texts are scored in each factor/dimension (hence, the name Multi-Dimensional), in a cluster analysis each text is designated to a single cluster. However, similar to a factor analysis, where each factor is assumed to embody a dimension, each cluster is assumed to correspond to a different text type; therefore, the job of the analyst is similar in both a factor analysis and a cluster analysis: to interpret the statistical output in order to detect the underlying principles motivating the existence of the groups. Finally, a statistical procedure that is less often employed in MD Analysis is discriminant function analysis. Discriminant analysis is used to discriminate among observations by classifying them into a number of discrete categories. Following an MD Analysis, a discriminant analysis can be run using the dimension scores and the corpus component categories (e.g., register) as a means to verify the extent to which the texts are classified according to their actual categories. The accuracy of the classification will give an indication of the reliability of the Multi-Dimensional model of variation. For instance, if more texts are assigned to their actual categories than expected by chance using the dimension scores as predictor variables (e.g., more texts from register "A" are assigned to classification group “A” than to other groups), it is possible to suggest that the dimensions are reliable predictors of the categories. As mentioned, in a typical MD Analysis, the relationship between the corpus group categories and the dimension scores is normally tested through ANOVAs. A discriminant analysis is not meant to replace an ANOVA, but rather to add an extra layer of methodological rigor to the analysis by means of a blind classification task. See Cantos-Gomez’s chapter for an introduction to these four major multivariate procedures as well as for guidance on how to carry out these procedures in SPSS. Egbert and Staples’s chapter provides instructions on how to conduct a factor analysis in three major statistical packages (SPSS, SAS, and R). See Nini’s chapter for a discussion of English text types and Veirano Pinto’s chapter for a presentation on the use of discriminant analysis as a means to check the strength of an MD model.

Gray describes the basic algorithm used by the Biber tagger, which involves using dictionaries (word lists) to look up the words being tagged in a text. These dictionaries are not exhaustive; therefore, if a word is not found, a process of guessing the grammatical category of the word is put in place. This algorithm is not a unique feature of the Biber tagger. All taggers employ some form of predefined look-up list of words, and all of them are prepared to make an informed guess on any words that are not part of their dictionaries. In addition to grammatical features, the tagger also encodes semantic information in the annotated output. Various types of semantic tag complements are employed, ranging from the type of modal verb (possibility, necessity, or prediction) to the semantic grouping of nouns, verbs, adjectives, and adverbs. The semantic classification of nouns includes eight categories, such as animate (e.g., teacher, child), cognitive (e.g., fact, knowledge), and concrete (e.g., rain, modem). The semantic classes of verbs comprise seven categories, like activity (e.g., smile, open), communication (e.g., suggest, declare), and mental (e.g., know, think). There are six categories of adjectives, including size (e.g., big, high), time (e.g., new, young), and color (e.g., white, red), and two categories of adverbs (i.e., time and place). The attribution of a semantic domain is based on predefined lists of words for each semantic category. For a more detailed description of the semantic categories encoded by the tagger as well as for an overview of the words in each semantic class, see Biber (2006, 244–50).
References


Part One

Understanding the Principles: Origins of the Method, Corpus Design, and Annotation
Introduction

For many years, researchers have been interested in variation in language use across different situations: the description of “registers” and “register variation” (see, for example, Biber and Conrad 2009). Studies of registers have come from a wide range of disciplines and subdisciplines, including functional/sociolinguistics, applied linguistics, corpus/computational linguistics, composition/rhetoric studies, and communication research. (See the surveys of previous research in Biber and Atkinson 1994; Barbieri 2009.)

Many studies have described the situational and linguistic characteristics of a particular register, such as conversations, radio broadcasts, university lectures, grant proposals, PhD dissertations, and so on. However, the study of register can also be approached from a comparative perspective, investigating the patterns of variation among the range of registers in a discourse domain. Register variation is inherent in human language: a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar associated with different registers, reflecting the situational characteristics of those registers. The ubiquitous nature of register variation was noted early on by a number of scholars, for example:

Each language community has its own system of registers ... corresponding to the range of activities in which its members normally engage. (Ure 1982, 5)

Register variation, in which language structure varies in accordance with the occasions of use, is all-pervasive in human language. (Ferguson 1983, 154)

No human being talks the same way all the time. ... At the very least, a variety of registers and styles is used and encountered. (Hymes 1984, 44)

However, despite its fundamental importance, there were few comprehensive linguistic analyses of register variation before the 1980s. This disregard was due mostly to methodological difficulties: until that time, it was simply not feasible to analyze the