Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (eds), Multi-Dimensional Analysis: Research Methods and Current Issues, 67-94, London; New York: Bloomsbury Academic

The Multidimensional Analysis Tagger

Andrea Nini

University of Manchester

Abstract

This chapter introduces and describes the Multidimensional Analysis Tagger (MAT), a computer program for the analysis of corpora or single texts using the multi-dimensional model proposed by Biber (1988). The program uses the Stanford Tagger to generate an initial tagged version of the input, which is then used to find and count the original linguistic features used in Biber (1988). The program then plots the text or corpus on to Biber's (1988) dimensions and assigns it a text type as proposed by Biber (1989). Finally, MAT offers a tool to visualize the features of each dimension in the text. The software was tested for reliability by comparing the dimension scores produced by MAT for the LOB corpus against the ones obtained by Biber (1988) in his original analysis. This test shows that MAT can largely replicate Biber's results. The software was also tested on the Brown corpus and the results not only confirm the reliability of MAT in calculating the dimension scores, but also suggest that Biber's (1988) dimensions and text types can be generalized and applied to other data sets. As a further example of a MAT analysis, a study of a corpus of threatening and abusive letters is reported. Although this corpus did not contain the balanced sample of registers required to perform a new multi-dimensional analysis, MAT allowed a text type analysis of the corpus to be performed through a comparison with Biber's (1988; 1989) model of English register variation.

1. Introduction

About thirty years ago, Biber's (1988) *Variation across Speech and Writing* revolutionized our understanding of registers by introducing factor analysis for the extraction of latent

dimensions of variation from patterns of co-occurrence of linguistic features, a methodology later called multi-dimensional analysis. The use of this new methodology also led to a sounder understanding of the most important linguistic and extra-linguistic factors that influence register variation in English. Multi-dimensional analysis was adopted in a large number of other studies on the language used in various registers from academic language (e. g. Biber, 2003; Gray, 2013) to the most recent web registers (Grieve, Biber, and Friginal 2011; Titak and Roberson 2013; Biber and Egbert 2016). The flexibility of multi-dimensional analysis for linguistic research is also demonstrated by its various other applications, such as the study of author styles (Biber and Finegan 1994), sociolects (Biber and Burges 2000), regional variation (Grieve 2014), or diachronic register variation (Biber and Finegan 1989).

Beside the value of multi-dimensional analysis itself, Biber (1988) has also uncovered some very valuable insight on the patterns of variation across the registers of the English language. By extracting the underlying dimensions of variation for a corpus balanced for registers, Biber (1988) was able to propose a set of six dimensions that can account for the linguistic variation in the most important registers of the English language. These six dimensions represent patterns of co-variation of linguistic features and were functionally interpreted according to their constituting features and the registers that they characterized. These original six dimensions are summarized in Table 3.1.

Table 3.1: Short descriptions and summary of the six dimensions of register variation for English found by Biber (1988).

	Description	
Dimension 1 Involved vs. Informational Discourse	Low scores on this Dimension indicate informationally dense discourse, as in the case of academic prose, whereas high scores indicate that the text is affective and interactional, as for conversations.	**Involved production features**: private verbs, that-deletions, contractions, present tenses, second person pronouns, do as pro-verb, analytic negations, demonstrative pronouns, emphatics, first person pronouns, pronoun it, be as main verb, causative subordinations, discourse particles, indefinite pronouns, hedges, amplifiers, sentence relatives, wh- questions, possibility modals, non-phrasal coordinations, wh- clauses, stranded prepositions. **Informational production features**: nouns, average word length, prepositions, type/token ratio, attributive adjectives
Dimension 2 Narrative vs. Non-Narrative Concerns	The higher the score on this Dimension the higher the narrative concern, as in the case of works of fiction.	**Narrative concerns features**: past tenses, third person pronouns, perfect aspects, public verbs, synthetic negations, present participial clauses

Dimension 3 Context-Independent Discourse vs. Context-Dependent Discourse	Low scores on this Dimension indicate dependence on the context as in the case of a sport broadcast, whereas high scores indicate independence from context, as for example in academic prose.	**Context-dependent discourse features**: time adverbials, place adverbials, general adverbs. **Context-independent discourse features**: wh- relative clauses on object position, pied-piping relatives, wh- relative clauses on subject position, phrasal coordinations, nominalizations
Dimension 4 Overt Expression of Persuasion	The higher the score on this Dimension indicate the more the text explicitly marks the author's point of view as well as their assessment of likelihood and/or certainty, as for example in professional letters.	**Overt expression of persuasion features**: infinitives, prediction modals, suasive verbs, conditional subordinations, necessity modals, split auxiliaries
Dimension 5 Abstract vs. Non-Abstract Information	The higher the score on this Dimension the higher the degree of technical and abstract information, as for example in scientific discourse.	**Abstract information features**: conjuncts, agentless passives, past participial clauses, by passives, past participial WHIZ deletion relatives, other adverbial subordinators
Dimension 6 On-Line Informational Elaboration	High scores on this Dimension indicate that the information expressed is produced under certain time constraints, as for example in speeches.	**On-line informational elaboration features**: that clauses as verb complements, demonstratives, that relative clauses on object position, that clauses as adjective complements

The value of these dimensions transcends their role in the English language as subsequent research has also demonstrated a striking cross-linguistic validity for Dimension 1 and Dimension 2 across several languages from different families (Biber 1995; Biber 2014).

In addition to the discovery of the six dimensions, Biber (1989) later introduced the use of cluster analysis to find out the characteristic *text types* of the English language, that is, clusters of texts that are linguistically similar in terms of the six dimensions. Using cluster analysis, this study found out that the same corpus used in Biber (1988) could be divided into eight text types, a summary of which is presented in Table 3.2.

Table 3.2: Short description and summary of the eight text types for English found by Biber (1989).

Text type	Characterizing registers	Dimension profile	Description
Intimate interpersonal interaction	telephone conversations between personal friends	high score on D1, low score on D3, low score on D5, unmarked scores for the other Dimensions	Text type that usually includes interactions that have an interpersonal concern between close acquaintances
Informational interaction	face-to-face interactions, telephone conversations, spontaneous speeches, personal letters	high score on D1, low score on D3, low score on D5, unmarked scores for the other Dimensions	Text type that usually includes personal spoken interactions focused on informational concerns
Scientific exposition	academic prose, official documents	low score on D1, high score on D3, high score on D5, unmarked scores for the other Dimensions	Text type that usually includes informational expositions focused on conveying technical

			information
Learned exposition	official documents, press reviews, academic prose	low score on D1, high score on D3, high score on D5, unmarked scores for the other Dimensions	Text type that usually includes informational expositions focused on conveying information
Imaginative narrative	romance fiction, general fiction, prepared speeches	high score on D2, low score on D3, unmarked scores for the other Dimensions	Text type that usually includes texts with an extreme narrative concern
General narrative Exposition	press reportage, press editorials, biographies, non-sports broadcasts, science fiction	low score on D1, high score on D2, unmarked scores for the other Dimensions	Text type that usually includes texts that use narration to convey information
Situated reportage	sport broadcasts	low score on D3, low score on D4, unmarked scores for the other Dimensions	Text type that usually includes on-line commentaries of events that are in progress
Involved persuasion	spontaneous speeches, professional letters, interviews	high score on D4, unmarked scores for the other Dimensions	Text type that usually includes persuasive and/or argumentative discourse

Besides the pioneering of factor analysis and cluster analysis for the analysis of registers, another achievement of the findings of the two studies above is the elaboration of a model of register variation for the English language that is predictive. Using the results of the multi-dimensional analysis it is possible to determine how a text, corpus, or even register behaves linguistically in comparison to other registers of English. In essence, the multi-dimensional model represents a base-rate knowledge of English that allows the description or

evaluation of other texts or registers.

Despite this potential, the majority of the research on the multi-dimensional analysis of register variation has focused on using factor analysis and cluster analysis on new data sets. Relatively speaking, few studies have used previous multi-dimensional models or the original model itself to describe or evaluate new data. Among these studies, the original multi-dimensional model has been used especially to study the registers of television programs (Quaglio, 2009; Al-Surmi 2012; Berber Sardinha 2014; Berber Sardinha and Veirano Pinto 2017) and written or spoken academic registers (Conrad 1996; Conrad, 2001; Biber et al., 2002).

These studies are evidence that the model can be useful in many applications that involve the comparison of new data to a base-rate knowledge of English registers. For example, the evaluation of similarity of a particular academic text written by a learner of English to the norm for academic registers of English is such an application. Similarly, register variation researchers can use the same model to compare a register to the other registers of English considered in the 1988/1989 model. As opposed to finding new dimensions, which is an endeavor that brings insight in the internal structure of registers, contrasting a data set to a general model of English can be another way to bring to light its register identity.

The application of Biber's original dimensions and text types can also be useful for those interested in looking at variation within small or unstructured corpora. The first multi-dimensional analysis was successful in producing a model that describes English registers because the corpus was carefully sampled by registers and large enough to carry out a statistical analysis. These two pre-requisites are essential to obtain dimensions that can adequately capture register variation. However, depending on the data that one wants to analyze, it might turn out to be impossible to collect a large enough corpus or one that is

internally stratified enough to produce meaningful register dimensions. In such cases, plotting the input corpus onto Biber's model of English can be a reasonable approximation to running a new multi-dimensional analysis. Instead of extracting new dimensions for the register, one can assess how this new register is different or similar to other registers of the English language and in doing so finding out its register identity.

The application of Biber's model is however dependent firstly on an empirical validation of its generalization to new texts and secondly on the development of a tool that can easily allow other researchers to find the location of a new data set in the English multi-dimensional space. The present chapter presents research that assesses both points. Firstly, the chapter introduces the Multidimensional Analysis Tagger (or MAT, freely accessible at https://sites.google.com/site/multidimensionaltagger/), a computer program that facilitates the process of applying the original 1988 model to a new data set. After describing its architecture and validation process, an analysis of the Brown corpus using MAT will be reported to describe to what extent the model can be applied to new texts. Finally, the chapter concludes with a demonstration of the applications of MAT for register analysis.

2. The MAT

2. 1 The architecture of MAT

MAT is a computer program that replicates Biber's (1988) tagger, calculates the dimension scores for each of the dimensions, and then plots the input data onto the multi-dimensional space while also assigning each text to one of the eight text types identified by Biber (1989). This whole process is achievable due to the detailed descriptions of the tagging rules for the linguistic features presented in the appendix of Biber (1988).

After the user has provided an input, MAT returns a tagged version of it using the same 67 linguistic features of Biber (1988). However, MAT does not use the original Biber tagger,

which is not publicly available, and instead uses the Stanford Tagger (Toutanova et al. 2003) for the preliminary tagging of basic parts of speech, such as nouns, adjectives, verbs, or adverbs, followed by Biber's (1988) rules for more complex features, such as sentence relatives, *that* as a demonstrative as opposed to a complementizer, and so forth. Although the original Biber tagger prompted the user with ambiguous cases for certain complex features, MAT does not implement any manual intervention from the user. However, manual intervention on the tagged texts can be performed by a user, if he/she wishes, before the statistical analysis takes place.

Although the tagging rules used by MAT are the same as the original Biber tagger, since the tagging of basic parts of speech is performed by the Stanford Tagger, the tagged files returned by MAT are bound to contain some inconsistencies with the original tagger. While some differences are unavoidable, basic parts of speech attribution generally does not vary greatly across taggers, and thus results should be compatible to the 1988 results. Indeed, the reliability of MAT has been tested and the results are reported in the next section below.

After the input has been tagged, in order to calculate the dimension scores, firstly the occurrences of a feature are counted (with the exception of average word length and type/token ratio), and then their relative frequency per hundred words is calculated. Finally, the standardized scores, or z-scores, for each feature are calculated using the standard formula reported below, where x is the relative frequency of a feature in the user's input, z_x is the resulting z-score of the feature in consideration, μ_B is the mean frequency for that feature in Biber's (1988) corpus, and σ_B is the standard deviation of that feature in Biber's (1988) corpus:

$$z_x = \frac{x - \mu_B}{\sigma_B}$$

MAT applies these formulas and outputs two files, one with the frequencies (per hundred words) and one with the z-scores. As described in Biber (1988), the final dimension scores for

each dimension are calculated by summing or subtracting the z-scores of the dimension features following the features' polarities within a dimension. For example, for Dimension 1, the final score is calculated with the formula:

$$D1 = \left(z_{privateverbs} + z_{thatdeletions} + z_{contractions} + \cdots\right) - \left(z_{nouns} + z_{wordlength} + z_{prepositions} + \cdots\right)$$

In the calculation of dimension scores MAT implements a slight alteration as it includes in each formula only those variables with a mean higher than 1 in Biber (1988: 77). This change has been implemented as the features with a mean lower than 1 are rare features of English—the frequency of which can be highly dependent on sample size. For this reason, if by chance alone one of these rare features is even slightly more common in the user's input than in the original corpus, then the z-scores of these features would be abnormal and thus inflate the dimension scores. The loss of this information does not greatly influence the dimension scores as, given their rarity, these features contribute very little to the dimensions. In addition to this change, MAT also offers the possibility to apply a *z-score correction*, that is, a reduction of all the z-scores of magnitude higher than 5 to 5, in order to avoid unlikely inflated z-scores and dimension scores.

Finally, MAT plots the input data in the multi-dimensional space and assigns a text type to each input text. A graph similar to the graphs displayed in Biber (1988: 172) using means and ranges is produced for the dimensions selected by the user. Using this graph, the user can compare their data against a selection of registers. The program will also print out which register is the most similar to the input data. Another plot is also produced for the text types mirroring Biber's (1989) visualisation. This plot displays the dimensions horizontally and the location of each text type as well as the input data on each dimension vertically. In this way, the user can compare their data to the other text types and assess which text type is the most similar to the input. The most similar text type is assigned using Euclidean distance

from the centroids of the clusters reported in Biber (1989).

After carrying out the analysis, the user can also use MAT to visualize of one or more dimension features in the text. MAT can produce a color-coded file with the selected dimension features, allowing for the qualitative exploration and interpretation of such features.

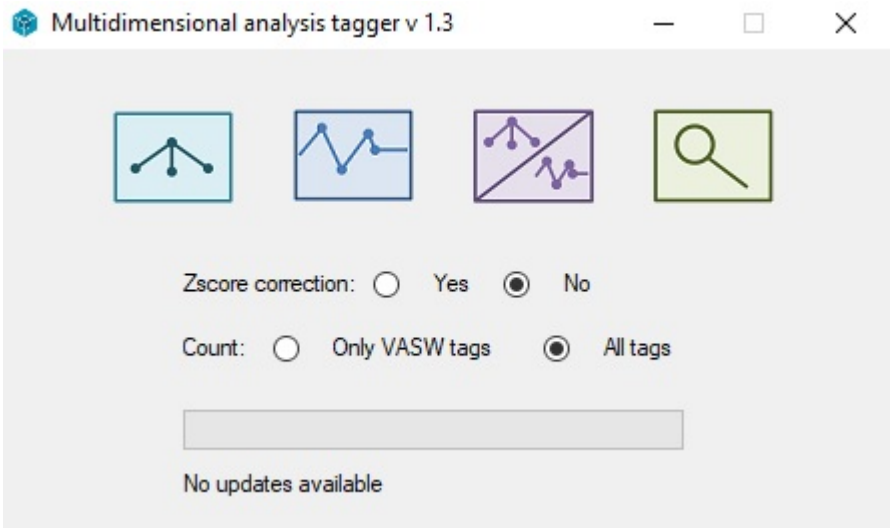

Figure X: Screenshot of MAT interface for Windows.

Figure X shows the interface of MAT for Windows. The third button from the left, *Tag and Analyze*, will take as input one text or a folder of texts, tag it, and then return the input's location in the multi-dimensional model. The two processes can be done in two separate steps, for example if the user wants to manually check the quality of the tagging, by using the first button *Tag* and then the second button *Analyze*. Finally, the final button is the *Inspect* functionality to visualize the dimension features. More information about how to use MAT can be found in the manual, which is also freely downloadable at https://sites.google.com/site/multidimensionaltagger/.

Although great differences are not expected between MAT and the 1988 Biber tagger, the question of whether and to what extent MAT does indeed replicate the original analysis can only be tested if the same data set is analyzed and similar results are found. The

description of such analysis is reported in the section below.

2.2 Testing the reliability of MAT

In order to test whether MAT is reliable a MAT analysis of the original data set used by Biber (1988; 1989) was carried out. The original data set consisted of the LOB corpus (Johansson, Leech, and Goodluck 1978) for the published written material, the London-Lund corpus (Svartvik 1990) for the spoken data, and a small corpus of personal and professional letters collected by Biber.

Despite some efforts, only the LOB corpus could be retrieved, of which only thirteen out of its fifteen registers were found: Press Reportage, Press Editorial, Press Reviews, Religion, Hobbies, Popular Lore, Academic Prose, General Fiction, Mystery Fiction, Science Fiction, Adventure Fiction, Romantic Fiction, and Humour. The test was therefore carried out on this data set.

After running MAT on the thirteen available registers of LOB some differences were observed, but overall Biber's analyses were successfully replicated. The results of the analysis are displayed in Table 3.3, where the first column identifies a register and the following columns list the dimension scores obtained by Biber and the ones returned by MAT. The last column lists and contrasts the distribution of text types for the register using percentages, from the most common to the least common.

Table 3.3: 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.

Registers	D1	D2	D3	D4	D5	D6	Text types
Press reportage MAT	-14.02	0.97	2.81	-0.38	0.52	-0.72	59% General narrative exposition; 39% Learned exposition; 2%

							Involved persuasion; 2% Scientific exposition
Press reportage Biber (1988)	-15.01	0.4	-0.3	-0.7	0.6	-0.9	73% General narrative exposition; 25% Learned exposition; 2% Scientific exposition
Difference	0.99	0.57	3.11	0.32	0.08	0.18	
Press editorials MAT	-8.4	-0.28	4.38	3.3	1.5	0.33	81% General narrative exposition; 7% Involved persuasion; 7% Scientific exposition; 4% Learned exposition
Press editorials Biber (1988)	-10	-0.8	1.9	3.1	0.3	1.5	86% General narrative exposition; 11% Involved persuasion; 4% Learned exposition
Difference	1.6	0.52	2.48	0.2	1.2	1.17	
Press reviews MAT	-12.45	-0.74	5.38	-2.32	0.36	-1.01	53% General narrative exposition; 47% Learned exposition
Press reviews Biber (1988)	-13.9	-1.6	4.3	-2.8	0.8	-1	47% Learned exposition; 47% General narrative exposition; 6% Scientific exposition
Difference	1.45	0.86	1.08	0.48	0.44	0.01	
Religion MAT	-4.26	0.17	4.69	0.85	2.22	1.01	65% General narrative exposition; 29% Involved persuasion; 6% Scientific exposition
Religion Biber (1988)	-7	-0.7	3.7	0.2	1.4	1	59% General narrative exposition; 18% Involved persuasion; 18% Learned exposition; 6% Imaginative narrative
Difference	2.74	0.87	0.99	0.65	0.82	0.01	
Hobbies MAT	-9.42	-2.1	3.15	1.51	2.54	-0.35	34% General narrative exposition; 24% Learned exposition; 24%

							Involved persuasion; 18% Scientific exposition
Hobbies Biber (1988)	-10.1	-2.9	0.3	1.7	1.2	-0.7	43% General narrative exposition; 21% Learned exposition; 21% Involved persuasion; 7% Scientific exposition; 7% Situated reportage
Difference	0.68	0.8	2.85	0.19	1.34	0.35	
Popular lore MAT	-9.58	0.31	3.42	-0.61	1.4	-0.64	36% Learned exposition; 32% General narrative exposition; 20% Involved persuasion; 2% Imaginative narrative; 9% Scientific exposition
Popular lore Biber (1988)	-9.3	-0.1	2.3	-0.3	0.1	-0.8	36% Learned exposition; 36% Involved persuasion; 21% General narrative exposition; 7% Imaginative narrative
Difference	0.28	0.41	1.12	0.31	1.3	0.16	
Academic prose MAT	-12.16	-2.16	5.38	-0.02	5.14	0.23	56% Scientific exposition; 24% Learned exposition; 14% General narrative exposition; 6% Involved persuasion
Academic prose Biber (1988)	-14.09	-2.6	4.2	-0.5	5.5	0.5	44% Scientific exposition; 31% Learned exposition; 17% General narrative exposition; 9% Involved persuasion
Difference	1.93	0.44	1.18	0.48	0.36	0.27	
General fiction MAT	0.35	6.26	0.03	1.79	-0.45	-0.75	55% Imaginative narrative; 31% General narrative exposition; 10% Involved persuasion; 3% Learned

							exposition
General fiction Biber (1988)	-0.8	5.9	-3.1	0.9	-2.5	-1.6	51% Imaginative narrative; 41% General narrative exposition; 3% Informational interaction; 3% Involved persuasion
Difference	1.15	0.36	3.13	0.89	2.05	0.85	
Mystery fiction MAT	0.82	5.76	-0.7	1.55	-0.69	-1.13	67% Imaginative narrative; 29% General narrative exposition; 4% Involved persuasion
Mystery fiction Biber (1988)	-0.2	6	-3.6	-0.7	-2.8	-1.9	70% Imaginative narrative; 23% General narrative exposition; 8% Situated reportage
Difference	1.02	0.24	2.9	2.25	2.11	0.77	
Science fiction MAT	-5.01	6.1	1.08	0.21	-0.54	-0.54	83% General narrative exposition; 17% Imaginative narrative
Science fiction Biber (1988)	-6.1	5.9	-1.4	-0.7	-2.5	-1.6	50% General narrative exposition; 33% Imaginative narrative; 17% Situated reportage
Difference	1.09	0.2	2.48	0.91	1.96	1.06	
Adventure fiction MAT	-0.85	5.89	-1.29	0.19	-0.97	-1.29	69% Imaginative narrative; 24% General narrative exposition; 3% Involved persuasion; 3% Learned exposition
Adventure fiction Biber (1988)	0	5.5	-3.8	-1.2	-2.5	-1.9	70% Imaginative narrative; 31% General narrative exposition
Difference	0.85	0.39	2.51	1.39	1.53	0.61	
Romantic fiction MAT	3.55	6.71	-0.88	2.35	-1.26	-1	79% Imaginative narrative; 17% General narrative exposition; 3% Involved persuasion
Romantic fiction Biber (1988)	4.3	7.2	-4.1	1.8	-3.1	-1.2	92% Imaginative narrative; 8%

							General narrative exposition
Difference	0.75	0.49	3.22	0.55	1.84	0.2	
Humour MAT	-6.19	1.43	1.62	0.43	0.65	-0.56	78% General narrative exposition; 11% Imaginative narrative; 11% Involved persuasion
Humour Biber (1988)	-7.8	0.9	-0.8	-0.3	-0.4	-1.5	89% General narrative exposition; 11% Involved persuasion
Difference	1.61	0.53	2.42	0.73	1.05	0.94	
Mean differences	1.24	0.51	2.27	0.72	1.24	0.51	

In Dimension 1, Involved versus Informational Discourse, the most important of the dimensions in terms of variance explained and universality (Biber 1995), score differences range from 0.28 for Popular Lore to 2.74 for Religion (Mean: 1.24). These differences of the order or one or two points do not affect the identification of the correct location of a new input, as Dimension 1 scores in Biber's study range from roughly -20 to 50.

In Dimension 2, Narrative versus Non-Narrative Concerns, again large differences are not detected, with a range from 0.2 for Science Fiction to 0.87 for Religion (Mean: 0.51). For this dimension, as for all the other dimensions except for the first one, the range of scores presented in Biber (1988) roughly ranges from -5 to 5, with a positive score indicating that a text or register is narrative. Besides the small differences, Table 3 highlights that for all the registers except two, the sign of the scores is the same as in the original study.

Contrary to the two dimensions above, the results for Dimension 3, Context-Independent Discourse versus Context-Dependent Discourse, are not as accurate, with differences ranging from 0.99 for Religion to 3.22 for Romantic Fiction (Mean: 2.27). Such differences are much higher in magnitude than the ones previously observed as the range of

Dimension 3 from Biber (1988) spans from -5 to 5. An average difference of more than 2 points can affect the reliable identification of the location of a text in this space.

As opposed to Dimension 3, the scores for the remaining dimensions are again not largely different from Biber's, ranging, respectively: from 0.19 to 2.25 for Dimension 4, Overt Expression of Persuasion (Mean: 0.72); from 0.08 to 2.11 for Dimension 5, Abstract versus Non-Abstract Information (Mean: 1.24), and from 0.01 to 1.06 for Dimension 6, On-Line Informational Elaboration (Mean: 0.51). Although it could be argued that some differences of the order of magnitude of 2 could be problematic, these are extreme values, as the more modest mean differences reveal.

In terms of the distribution of text types, most of the distributions assigned by MAT are compatible with the ones published in Biber (1989). Despite some differences in the exact percentages, the order of the text types, from most common to least common, is highly compatible and the most common error is the shifting of a particular text type of one rank. Since most text types are unmarked in Dimension 3, the assignation of accurate text types is unaffected by the discrepancies in Dimension 3 scores noted above.

In conclusion, this analysis has found that MAT can replicate Biber's (1988; 1989) analyses as well as assign dimension scores and text types that are reliable. The exception found concerns Dimension 3, were at times moderate differences were observed. Although a careful analysis of these differences was carried out, the cause of the problem could not be identified. A possibility is that the difference lies in the procedure used by the Stanford Tagger to tag basic parts of speech. Qualitative exploration of the z-scores of the Dimension 3 features seems to indicate that the abnormal values are mostly found for the general adverbs z-scores. An abnormal z-score could either indicate difference in adverbs tagging rules between the Stanford Tagger and the 1988 Biber tagger or, alternatively, a transcription error in the mean or standard deviation for adverbs in Biber (1988). Although further investigations

on the issue will be carried out, it is possible to nonetheless conclude that MAT offers a good replication of Biber's (1988; 1989) results and that it can be used to plot new data onto its multi-dimensional space.

3. The reliability of Biber's (1988) original dimensions

After having demonstrated that MAT is successful in replicating Biber's (1988; 1989) results, a question that can be now answered is the extent to which the model itself is reliable. The perfect test for such a question is the application of MAT to a data set that includes as many registers as the ones analyzed in the first study, such as a corpus similar to LOB. Luckily, such a data set is indeed available as the LOB corpus was created as a British replication of the Brown corpus (Francis and Kucera 1979). Since both corpora contain exactly the same registers, in precisely the same categories, text size, and number, the application of MAT to the Brown corpus is an excellent way of testing the reliability of Biber's model for a similar yet new data set. In this section an analysis of MAT on the Brown corpus is reported for the same thirteen registers that were analyzed in the section above, so that a comparison can be carried out both with Biber's (1988; 1989) results and with the results of MAT on the LOB corpus.

MAT results for the Brown corpus show how stable the model is to new data, as well as to the degree of internal consistency of the analysis with MAT. The results are displayed in Table 3.4 in a format similar to the previous analysis of the LOB corpus.

Table 3.4: 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.

Registers	D1	D2	D3	D4	D5	D6	Text types
Press reportage Brown	-17.61	0.09	4.51	-1.55	0.85	-1.11	75% Learned exposition; 20%

							General narrative exposition; 4% Scientific exposition
Press reportage Biber (1988)	-15.01	0.4	-0.3	-0.7	0.6	-0.9	73% General narrative exposition; 25% Learned exposition; 2% Scientific exposition
Difference	2.6	0.31	4.81	0.85	0.25	0.21	
Press editorials Brown	-10.71	-0.59	4.5	1.39	0.63	-0.28	63% General narrative exposition; 7% Involved persuasion; 26% Learned exposition; 4% Scientific exposition
Press editorials Biber (1988)	-10	-0.8	1.9	3.1	0.3	1.5	86% General narrative exposition; 11% Involved persuasion; 4% Learned exposition
Difference	0.71	0.21	2.6	1.71	0.33	1.78	
Press reviews Brown	-13.83	-1.32	5.27	-3.31	0.41	-1.08	59% Learned exposition; 41% General narrative exposition
Press reviews Biber (1988)	-13.9	-1.6	4.3	-2.8	0.8	-1	47% Learned exposition; 47% General narrative exposition; 6% Scientific exposition
Difference	0.07	0.28	0.97	0.51	0.39	0.08	
Religion Brown	-7.17	-0.11	5.1	0.39	2.11	0.49	35% General narrative exposition; 29% Involved persuasion; 24% Learned exposition; 12% Scientific exposition
Religion Biber (1988)	-7	-0.7	3.7	0.2	1.4	1	59% General narrative exposition; 18% Involved persuasion; 18% Learned exposition; 6% Imaginative narrative
Difference	0.17	0.59	1.4	0.19	0.71	0.51	
Hobbies	-12.44	-2.66	4.47	-0.86	1.34	-1.15	50% Learned

Brown							exposition; 36% General narrative exposition; 6% Involved persuasion; 8% Scientific exposition
Hobbies Biber (1988)	-10.1	-2.9	0.3	1.7	1.2	-0.7	43% General narrative exposition; 21% Learned exposition; 21% Involved persuasion; 7% Scientific exposition; 7% Situated reportage
Difference	2.34	0.24	4.17	2.56	0.14	0.45	
Popular lore Brown	-13.3	-0.1	3.9	-1.03	1.38	-0.67	44% Learned exposition; 42% General narrative exposition; 8% Involved persuasion; 6% Scientific exposition
Popular lore Biber (1988)	-9.3	-0.1	2.3	-0.3	0.1	-0.8	36% Learned exposition; 36% Involved persuasion; 21% General narrative exposition; 7% Imaginative narrative
Difference	4	0	1.6	0.73	1.28	0.13	
Academic prose Brown	-13.58	-2.33	5.93	-0.88	4.48	0.01	38% Scientific exposition; 38% Learned exposition; 23% General narrative exposition; 3% Involved persuasion
Academic prose Biber (1988)	-14.09	-2.6	4.2	-0.5	5.5	0.5	44% Scientific exposition; 31% Learned exposition; 17% General narrative exposition; 9% Involved persuasion
Difference	0.51	0.27	1.73	0.38	1.02	0.49	
General fiction Brown	-5.83	5.86	0.19	-0.33	-0.44	-1.22	66% General narrative exposition; 24% Imaginative narrative; 10% Involved

							persuasion
General fiction Biber (1988)	-0.8	5.9	-3.1	0.9	-2.5	-1.6	51% Imaginative narrative; 41% General narrative exposition; 3% Informational interaction; 3% Involved persuasion
Difference	5.03	0.04	3.29	1.23	2.06	0.38	
Mystery fiction Brown	-2.21	5.57	-1.22	0.13	-1.03	-1	46% General narrative exposition; 42% Imaginative narrative; 13% Involved persuasion
Mystery fiction Biber (1988)	-0.2	6	-3.6	-0.7	-2.8	-1.9	70% Imaginative narrative; 23% General narrative exposition; 8% Situated reportage
Difference	2.01	0.43	2.38	0.83	1.77	0.9	
Science fiction Brown	-4.1	4.79	1.3	0.12	0.79	-0.78	50% General narrative exposition; 17% Imaginative narrative; 17% Involved persuasion; 17% Learned exposition
Science fiction Biber (1988)	-6.1	5.9	-1.4	-0.7	-2.5	-1.6	50% General narrative exposition; 33% Imaginative narrative; 17% Situated reportage
Difference	2	1.11	2.7	0.82	3.29	0.82	
Adventure fiction Brown	-6.05	5.88	-0.81	-1.78	-1.05	-1.39	66% General narrative exposition; 31% Imaginative narrative; 3% Learned exposition
Adventure fiction Biber (1988)	0	5.5	-3.8	-1.2	-2.5	-1.9	70% Imaginative narrative; 31% General narrative exposition
Difference	6.05	0.38	2.99	-0.58	1.45	0.51	
Romantic fiction Brown	0.83	6.02	0.41	-0.08	-1.15	-1.08	59% Imaginative narrative; 31% General narrative exposition; 10% Involved persuasion

Romantic fiction Biber (1988)	4.3	7.2	-4.1	1.8	-3.1	-1.2	92% Imaginative narrative; 8% General narrative exposition
Difference	3.47	1.18	4.51	1.88	1.95	0.12	
Humour Brown	-6.76	2.96	2.56	-1.16	0.42	-0.46	67% General narrative exposition; 22% Imaginative narrative; 11% Learned exposition
Humour Biber (1988)	-7.8	0.9	-0.8	-0.3	-0.4	-1.5	89% General narrative exposition; 11% Involved persuasion
Difference	1.04	2.06	3.36	0.86	0.82	1.04	
Mean difference	2.18	0.53	2.73	0.88	1.18	0.57	

With the exception of Press Reportage, MAT assigns text types to the other Press registers with a very similar distribution compared to the original results. The similar distribution is of course a reflection of the similarities in the dimension scores, with the average score differences ranging from 2.6 in Dimension 3 to the impressive 0.07 in Dimension 1 for Press Review. Press Reportage shows some differences from Biber's studies, as Learned Exposition becomes the most common text type for this register as opposed to General Narrative Exposition, which comes second. This difference could be caused by the lower score in Dimension 1 obtained by this register using MAT. Their narrative character is nonetheless correctly identified by MAT through the positive score in Dimension 2. In general, the differences observed could be attributed to differences in the two corpora in topics or themes covered by the reportages.

For the Religion, Hobbies, and Popular Lore registers a generally strong compatibility of results is found, despite the fact that these registers are the most likely to contain different styles and topics compared to the LOB corpus. Religion presents the best results, both in terms of text type distribution and in terms of dimension score differences. Similarly good

results are observed for Popular Lore, with the exception of the Dimension 1 difference score of 4, which is however not too influential given the wide range of Dimension 1. Finally, the worst results for these categories are the ones for Hobbies, which include a difference as high as 4.17 for Dimension 3. As Dimension 3 does not have major weight on the attribution of text types, the impact that this anomaly has on text type distribution is small and leads only to the swap of first and second positions in the ranks.

The fact that Biber's model is reliable can be better observed in those registers in which topics and styles are not expected to vary greatly from corpus to corpus, e.g. Academic Prose. Indeed, for this register an extremely high level of compatibility of results between the Brown and the LOB corpora was found, as shown by the very small differences in dimension scores and a rather impressively similar distribution of text types.

Finally, the last registers discussed are the narrative and Humour, most of which indicate strong compatibility despite the fact that greater variability in styles and topics is expected in narratives across corpora. The narrative character of these registers is well-captured by MAT, as all the registers correctly present positive scores in Dimension 2 and narrative text types in the first ranks. The most common variations in the attribution of text types concern the alternation of first and second positions between the text type General Narrative Exposition and Imaginative Narrative. Both text types characterize narrative texts and the only difference between the two is that Imaginative Narrative tends to be more involved as it is typical of emotional narratives such as romantic novels. Indeed, for Romantic Fiction, both Biber's results and the MAT analysis of the Brown corpus show Imaginative Narrative as the most common text type. For Adventure Fiction and Mystery Fiction, however, MAT assigns General Narrative Exposition as the first text type, as opposed to Imaginative Narrative. These differences aside, the results are largely comparable and show the reliability of the model for narrative texts.

Table 3.5: 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.

Registers	D1	D2	D3	D4	D5	D6	Text types
Press reportage MAT Brown	-17.61	0.09	4.51	-1.55	0.85	-1.11	75% Learned exposition; 20% General narrative exposition; 4% Scientific exposition
Press reportage MAT LOB	-14.02	0.97	2.81	-0.38	0.52	-0.72	59% General narrative exposition; 39% Learned exposition; 2% Involved persuasion; 2% Scientific exposition
Difference	3.59	0.88	1.7	1.17	0.33	0.39	
Press editorials MAT Brown	-10.71	-0.59	4.5	1.39	0.63	-0.28	63% General narrative exposition; 7% Involved persuasion; 26% Learned exposition; 4% Scientific exposition
Press editorials MAT LOB	-8.4	-0.28	4.38	3.3	1.5	0.33	81% General narrative exposition; 7% Involved persuasion; 7% Scientific exposition; 4% Learned exposition
Difference	2.31	0.31	0.12	1.91	0.87	0.61	
Press reviews MAT Brown	-13.83	-1.32	5.27	-3.31	0.41	-1.08	59% Learned exposition; 41% General narrative exposition
Press reviews MAT LOB	-12.45	-0.74	5.38	-2.32	0.36	-1.01	53% General narrative exposition; 47% Learned exposition
Difference	1.38	0.58	0.11	0.99	0.05	0.07	
Religion MAT Brown	-7.17	-0.11	5.1	0.39	2.11	0.49	35% General narrative exposition; 29% Involved persuasion; 24% Learned exposition; 12% Scientific exposition
Religion MAT LOB	-4.26	0.17	4.69	0.85	2.22	1.01	65% General narrative exposition; 29% Involved persuasion; 6% Scientific exposition

Difference	2.91	0.28	0.41	0.46	0.11	0.52	
Hobbies MAT Brown	-12.44	-2.66	4.47	-0.86	1.34	-1.15	50% Learned exposition; 36% General narrative exposition; 6% Involved persuasion; 8% Scientific exposition
Hobbies MAT LOB	-9.42	-2.1	3.15	1.51	2.54	-0.35	34% General narrative exposition; 24% Learned exposition; 24% Involved persuasion; 18% Scientific exposition
Difference	3.02	0.56	1.32	2.37	1.2	0.8	
Popular lore MAT Brown	-13.3	-0.1	3.9	-1.03	1.38	-0.67	44% Learned exposition; 42% General narrative exposition; 8% Involved persuasion; 6% Scientific exposition
Popular lore MAT LOB	-9.58	0.31	3.42	-0.61	1.4	-0.64	36% Learned exposition; 32% General narrative exposition; 20% Involved persuasion; 2% Imaginative narrative; 9% Scientific exposition
Difference	3.72	0.41	0.48	0.42	0.02	0.03	
Academic prose MAT Brown	-13.58	-2.33	5.93	-0.88	4.48	0.01	38% Scientific exposition; 38% Learned exposition; 23% General narrative exposition; 3% Involved persuasion
Academic prose MAT LOB	-12.16	-2.16	5.38	-0.02	5.14	0.23	56% Scientific exposition; 24% Learned exposition; 14% General narrative exposition; 6% Involved persuasion
Difference	1.42	0.17	0.55	0.86	0.66	0.22	

General fiction MAT Brown	-5.83	5.86	0.19	-0.33	-0.44	-1.22	66% General narrative exposition; 24% Imaginative narrative; 10% Involved persuasion
General fiction MAT LOB	0.35	6.26	0.03	1.79	-0.45	-0.75	55% Imaginative narrative; 31% General narrative exposition; 10% Involved persuasion; 3% Learned exposition
Difference	6.18	0.4	0.16	2.12	0.01	0.47	
Mystery fiction MAT Brown	-2.21	5.57	-1.22	0.13	-1.03	-1	46% General narrative exposition; 42% Imaginative narrative; 13% Involved persuasion
Mystery fiction MAT LOB	0.82	5.76	-0.7	1.55	-0.69	-1.13	67% Imaginative narrative; 29% General narrative exposition; 4% Involved persuasion
Difference	3.03	0.19	0.52	1.42	0.34	0.13	
Science fiction MAT Brown	-4.1	4.79	1.3	0.12	0.79	-0.78	50% General narrative exposition; 17% Imaginative narrative; 17% Involved persuasion; 17% Learned exposition
Science fiction MAT LOB	-5.01	6.1	1.08	0.21	-0.54	-0.54	83% General narrative exposition; 17% Imaginative narrative
Difference	0.91	1.31	0.22	0.09	1.33	0.24	
Adventure fiction MAT Brown	-6.05	5.88	-0.81	-1.78	-1.05	-1.39	66% General narrative exposition; 31% Imaginative narrative; 3% Learned exposition
Adventure fiction MAT LOB	-0.85	5.89	-1.29	0.19	-0.97	-1.29	69% Imaginative narrative; 24% General narrative exposition; 3% Involved persuasion; 3% Learned exposition
Difference	5.2	0.01	0.48	1.97	0.08	0.1	

Romantic fiction MAT Brown	0.83	6.02	0.41	-0.08	-1.15	-1.08	59% Imaginative narrative; 31% General narrative exposition; 10% Involved persuasion
Romantic fiction MAT LOB	3.55	6.71	-0.88	2.35	-1.26	-1	79% Imaginative narrative; 17% General narrative exposition; 3% Involved persuasion
Difference	2.72	0.69	1.29	2.43	0.11	0.08	
Humour MAT Brown	-6.76	2.96	2.56	-1.16	0.42	-0.46	67% General narrative exposition; 22% Imaginative narrative; 11% Learned exposition
Humour MAT LOB	-6.19	1.43	1.62	0.43	0.65	-0.56	78% General narrative exposition; 11% Imaginative narrative; 11% Involved persuasion
Difference	0.57	1.53	0.94	1.59	0.23	0.1	
Mean difference	2.84	0.54	0.63	1.33	0.43	0.28	

Table 3.5, shows the internal consistency of MAT through a comparison of the analyses of the LOB and the Brown corpora. The mean differences of the six dimensions displayed in Table 5 are small in magnitude, ranging from the relatively small 2.84 in Dimension 1 to 0.28 in Dimension 6. Text type assignation is also very consistent; the only differences being an inversion of the first and second text types in the ranks.

In conclusion, the results of these analyses suggest that the model proposed by Biber is valid and that MAT is consistent in applying it. Similar scores and text types are returned if the model is applied to a new data set with the same registers tested in Biber's original data set and this is despite the fact that the tagging of basic parts of speech is done with a different tagger. This is an important result, which shows that Biber's model contains valid and stable information about general patterns of register variation in English. That the multi-dimensional

model of English can be used with data sets other than the original one implies that its contribution to linguistics and register analysis goes beyond the introduction of a method of analysis: the model constitutes a base-rate knowledge of linguistic characteristics of registers in the English language that can be used in new applied and theoretical research. As a demonstration of such applications, the next section details the analysis of a register that was not investigated by Biber in 1988 using MAT.

4. Applying Biber's multi-dimensional model of English to a new data set

The last step of this chapter is to demonstrate the usefulness of MAT for register analysis, both for comparing and contrasting new data to other registers of the English language and for performing a multi-dimensional analysis of a corpus that does not meet the its requirements. If a researcher's goal is to study the language of a certain register using the multi-dimensional method, then he/she should have access to a corpus that is large enough and with a sufficient internal stratification of situational parameters to allow for dimensions of register variation to be captured using factor analysis (Biber and Conrad 2009). Unfortunately, depending on the type of data this is not always possible and some data sets might not therefore be analyzed in this way. The example reported in this section concerns a register for which data collection is highly problematic, i.e. the register of malicious forensic texts.

A malicious forensic text, or MFT, is defined as a written piece of communication that is abusive, threatening or defaming and that is used as evidence in a forensic case (Nini 2017). For example, a ransom demand, an abusive or threatening letter, or a slanderous piece of writing that has been part of an investigation or a court trial would all qualify as MFTs. The situational characteristic that MFTs have in common and that links them to each other is the presence of a malicious purpose or speech act, such as a threat. However, texts classified as MFTs can differ in other situational characteristics and thus can be letters, text messages,

notes, and so forth. In the corpus considered, MFTs do tend to share many linguistic and situational characteristics with each other as they tend to have the same situational characteristics of written communication, such as being written with the possibility of being edited, absence of shared time and space between the participants involved, presence of only one recipient with no audience, etc. A full analysis of the situational parameters of the corpus of MFTs considered can be found in Nini (2017).

Although this register has been looked at before, especially at the pragmatic level and in particular for threatening texts (Fraser 1998; Napier and Mardigian 2003; Solan and Tiersma 2005), there seems to be no study on the register identity of these kinds of texts. This gap in the literature could be due to the difficulty of accessing such confidential data. Although forensic linguists working with law enforcement units frequently work with these kinds of texts, even in this field paucity of this type of data is an issue. Another problem of these texts is their relative shortness, as often such texts only contain one hundred or less tokens. Given the data collection problems listed above, finding enough data stratified by situational parameters so that a multi-dimensional register analysis can be carried out is difficult. However, thanks to the application of Biber's model through MAT some of these limitations can be overcome and the register identity of these texts can be investigated.

The MFT corpus here adopted was collected by Nini (2015) and consists of 104 texts, for a total of 39,188 word tokens and an average text length of 357 tokens (min: 103, max: 1610). Almost all the data set was collected using publicly available sources, such as forensic linguistics textbooks (e. g. Olsson, 2003), the FBI Vault repository of texts (https://vault.fbi.gov/), or the web through search engine queries. A smaller section of the corpus was made up of non-public texts made available by forensic linguists who frequently work on real-life cases in the UK and in the US. The corpus was limited to texts that contained at least 100 word tokens, as below this value it is not possible to calculate the

frequency of features reliably (Biber 1993; Biber and Jones 2005).

Although this corpus contains largely non-standard, unedited texts with misspellings, slang, non-standard punctuation or capitalization and so forth, MAT performed well in tagging the corpus. The reliability of MAT for the MFT corpus was tested by hand-checking a random 20% of the MAT tagged files while counting the number of tagging errors and calculating the percentage of correct tags for each text. This test revealed that on average 96% of the tags were correct (min: 87%, max: 100%). After this test, the corpus was used as input for a MAT analysis as described above.

With the application of Biber's model using MAT, it was possible to add MFTs to the register analysis of English presented by Biber (1988; 1989). The present section concerns the dimension scores and text types obtained for the MFTs compared to the remaining dimension scores and text types from Biber (1988; 1989). Table 3.6 gives the mean dimension scores for the MFT register in comparison with two registers that are similar in terms of situational parameters, Personal and Professional Letters.

Table 3.6: Comparison of mean dimension scores and standard deviations for MFTs and Personal and Professional Letters from Biber (1988).

	Personal Letters		MFTs		Professional Letters	
	Mean	SD	Mean	SD	Mean	SD
Dimension 1 Involved vs. Informational Discourse	19.5	5.4	1.5	11.5	-3.9	13.7
Dimension 2 Narrative vs. Non-Narrative Concerns	0.3	1.0	-0.7	3.9	-2.2	3.5
Dimension 3 Context-Independent Discourse vs. Context-Dependent Discourse	-3.6	1.8	2.6	3.9	6.5	4.2
Dimension 4 Overt Expression of Persuasion	1.5	2.6	3.9	5.7	3.5	4.7
Dimension 5 Abstract vs. Non-Abstract Information	-2.8	1.9	0.3	4.0	0.4	2.4
Dimension 6 On-Line Informational Elaboration	-1.4	1.6	0.1	2.6	1.5	3.6

Table 3.6 shows that the mean dimension scores for MFTs are located between the two types of letters, with the exception of Dimension 4, and on average closer to Professional than to Personal Letters.

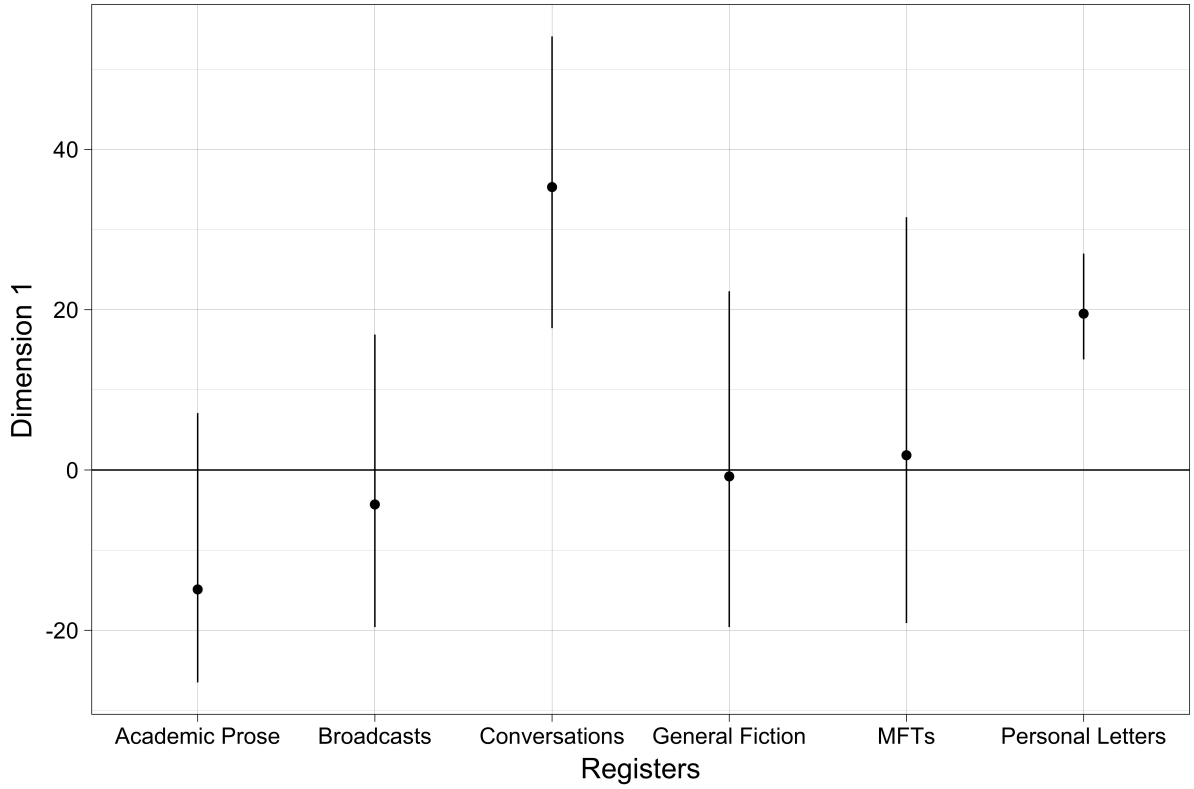

Figure 3.1: Mean scores and ranges for Dimension 1, Involved vs. Informational Discourse, for a selection of Biber's (1988) registers compared to the mean and range for MFTs.

In Dimension 1, Involved versus Informational Discourse, (Figure 3.1), MFTs are rather unmarked texts. With a mean score close to one, the average MFT text contains language that is not characterized by either dimension poles, similarly to that of General Fiction, and rather distant from more involved or informational registers, such as Conversation and Academic Prose, respectively.

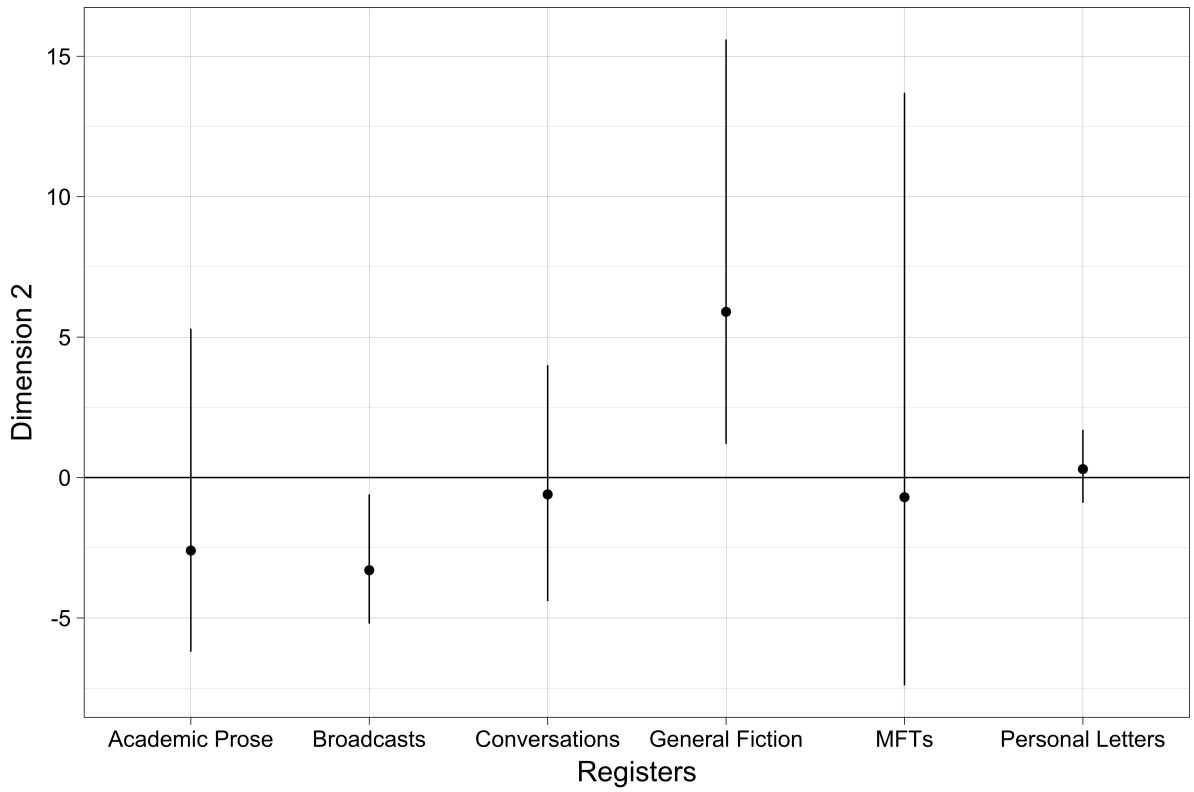

Figure 3.2: Mean scores and ranges for Dimension 2, Narrative vs Non-Narrative Concerns, for a selection of Biber's (1988) registers compared to the mean and range for MFTs.

In Dimension 2, Narrative versus Non-Narrative Concerns, (Figure 3.2), the mean score below zero suggests that MFTs are on average non-narrative texts situated together with other non-narrative registers such as Academic Prose and distant from Fiction registers.

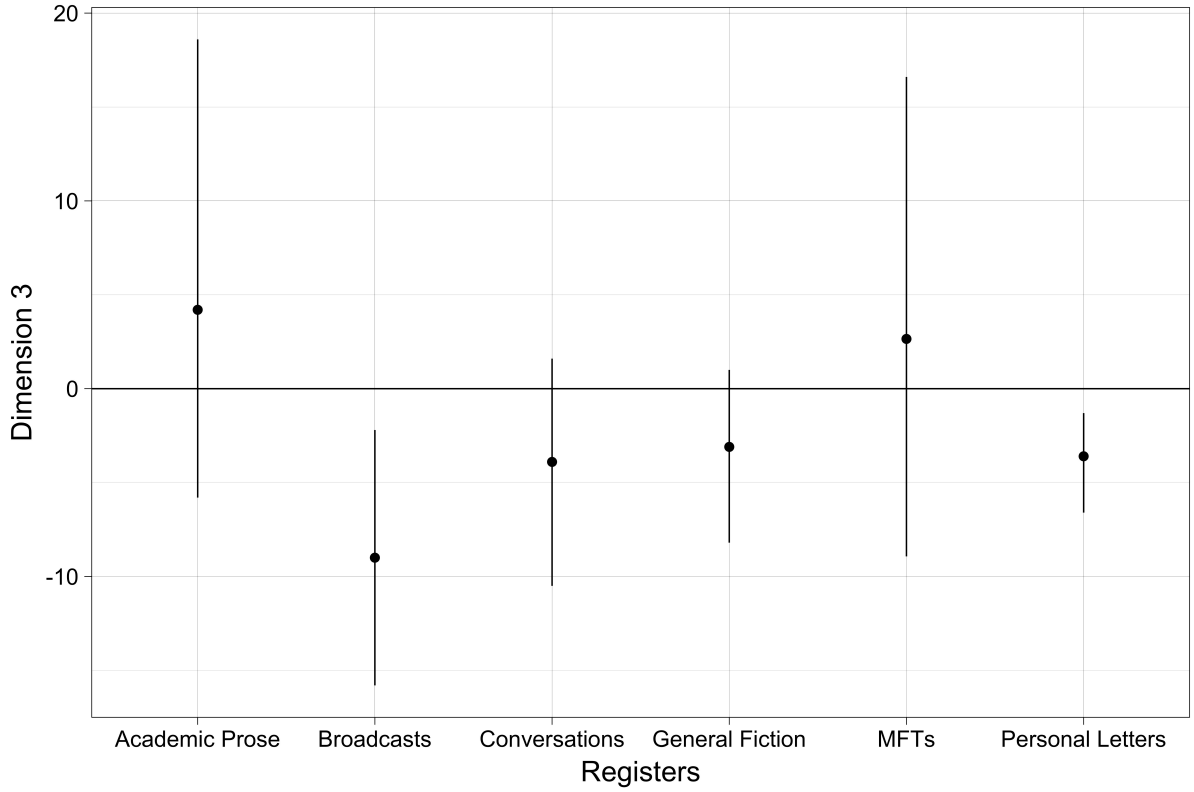

Figure 3.3: Mean scores and ranges for Dimension 3, Context-Independent vs. Context-Dependent Discourse, for a selection of Biber's (1988) registers compared to the mean and range for MFTs.

The analysis of Dimension 3, Context-Independent versus Context-Dependent Discourse (Figure 3.3) reveals that MFTs on average have a tendency towards explicitness of information, just as other written registers, such as Academic Prose and as opposed to the context-dependency of certain spoken registers, such as Conversations or Broadcasts.

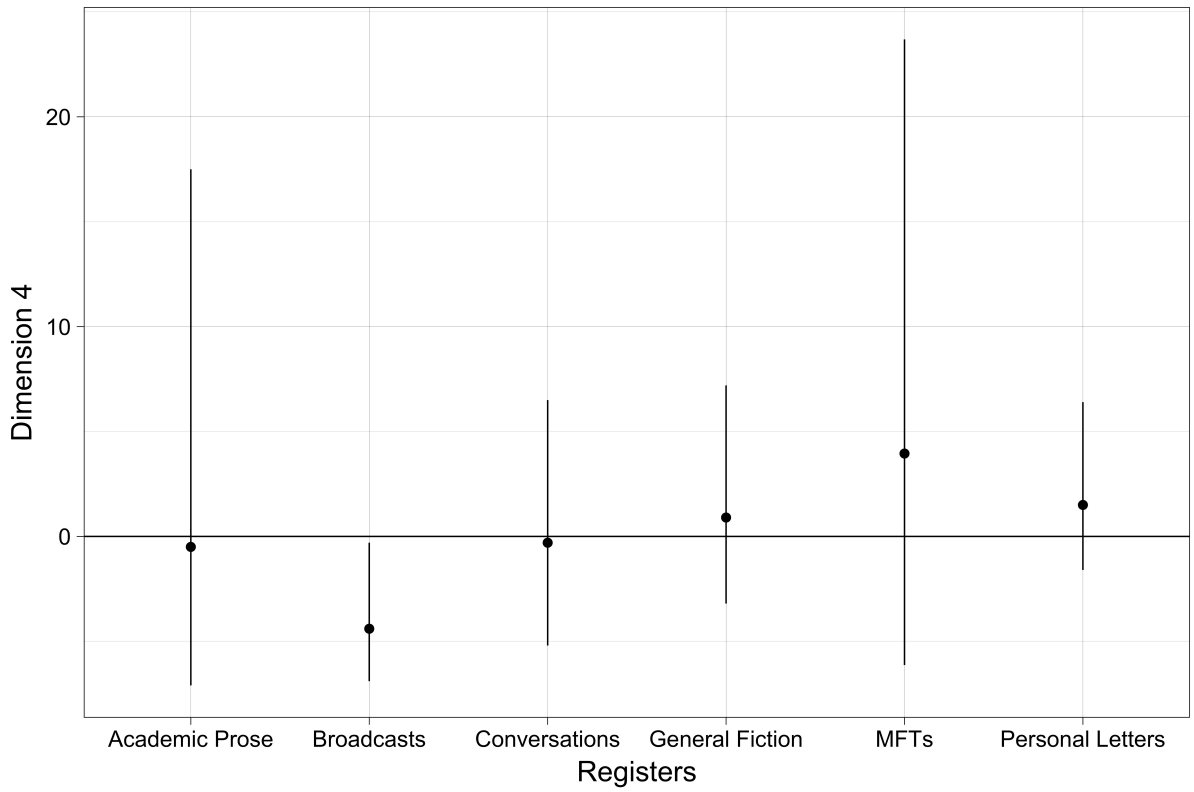

Figure 3.4: Mean scores and ranges for Dimension 4, Overt Expression of Persuasion, for a selection of Biber's (1988) registers compared to the mean and range for MFTs.

Dimension 4, Overt Expression of Persuasion, (Figure 3.4) presents interesting results, as the mean score achieved by the average MFT is far higher than any other register analyzed in Biber's works. In his original study, the highest mean score for this dimension was that of Professional Letters, as this register is the one that most frequently adopts persuasive linguistic means. As this analysis reveals, though, MFTs are far superior in the use of persuasive means than average-scored Professional Letters and the high mean score of MFTs in Dimension 4 could be regarded as the register 'signature'. As Nini (2017) reveals by comparing and contrasting texts with different situational parameters in this same MFT corpus, this high score in Dimension 4 is due to the presence of threatening texts, which are characterized by high levels of persuasion features.

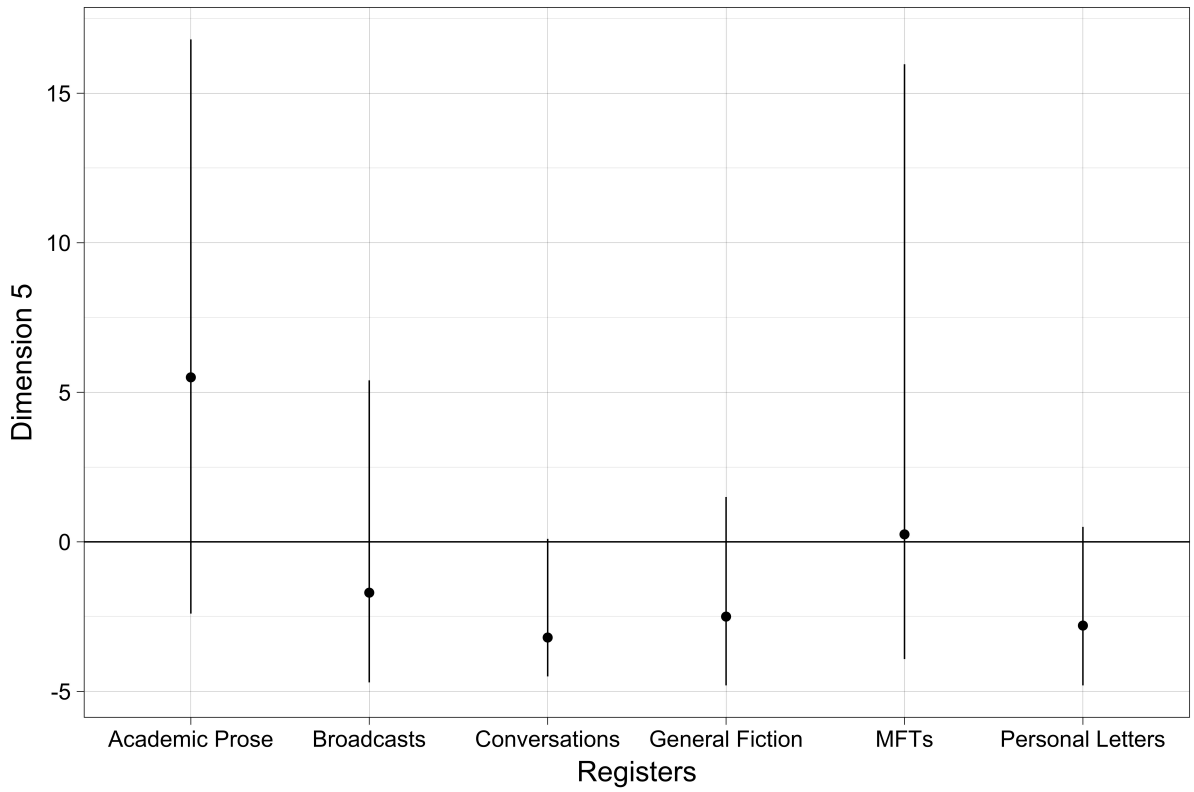

Figure 3.5: Mean scores and ranges for Dimension 5, Abstract vs Non-Abstract Information, for a selection of Biber's (1988) registers compared to the mean and range for MFTs.

In Dimension 5, Abstract versus Non-Abstract Information (Figure 3.5), MFTs are on average unmarked. They are distant from other more abstract written registers, such as Academic Prose and also distant from other registers, such as Conversations, in which abstract discourse is rarely found. As such, for this dimension, MFTs again appear as a kind of less formal and less abstract written register.

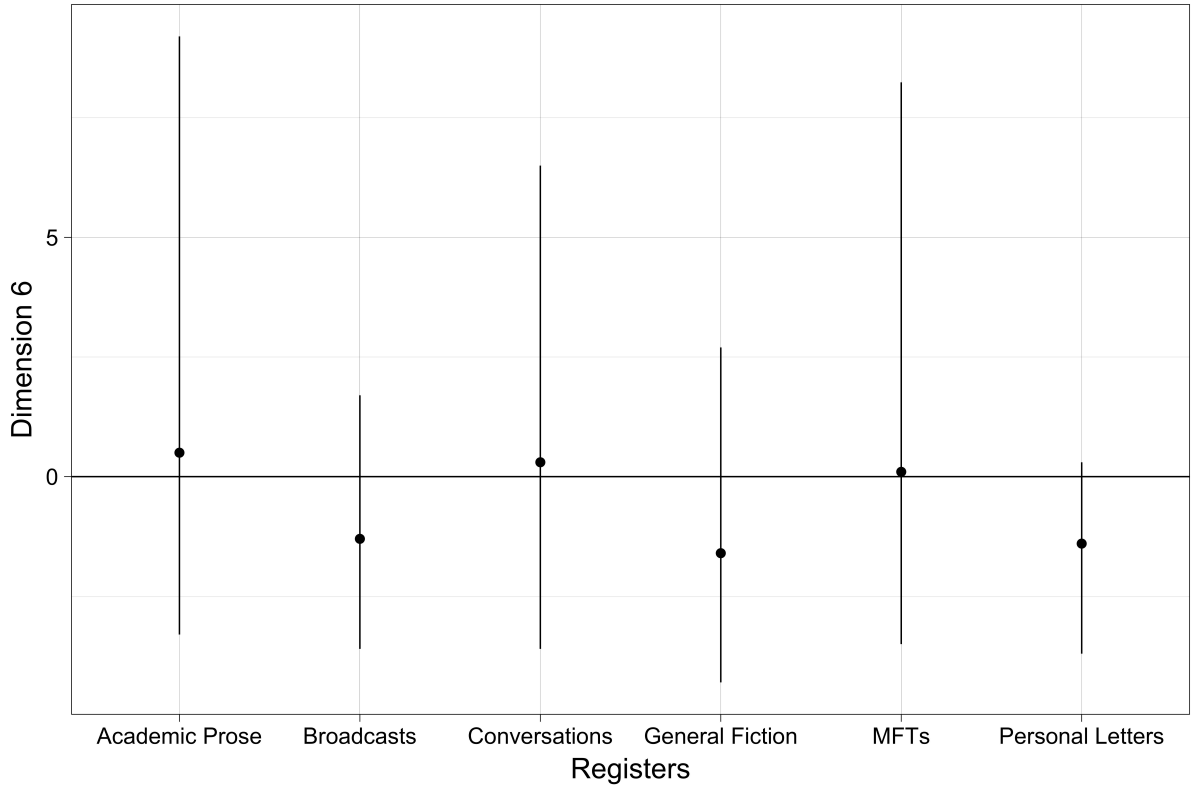

Figure 3.6: Mean scores and ranges for Dimension 6, On-Line Informational Elaboration, for a selection of Biber's (1988) registers compared to the mean and range for MFTs.

Finally, in terms of Dimension 6, On-Line Informational Elaboration (Figure 3.6), MFTs are also unmarked and do not show signs of high degrees of on-line information elaboration.

In terms of text types, MAT reveals that the average text type for MFTs is Involved Persuasion, with 47% of the texts being classified in this category. The high frequency of Involved Persuasion text types is not surprising as this text type is marked by high scores in Dimension 4 and it is the prototypical text type for Professional Letters. The fact that almost 50% of MFTs fall into this text type is therefore a confirmation of their peculiar Dimension 4 scores due to the presence of several texts with threatening content, which often employ modal verbs and other modality features to convey their stance (Gales 2011; Gales 2015; Nini 2017).

Although this analysis is not entirely equivalent to finding out the dimensions of variation for MFTs and/or the text types within this register, there is much value in plotting the register of MFTs against other registers of the English language. The results of this study reveal interesting facts about this relatively unexplored register so that it is possible to get a grasp of its linguistic characteristics even without reading any of the texts and simply by examining the text types these texts are classified as and the scores that these texts have on Biber's dimensions. This type of analysis can also approximate the results of a full multi-dimensional register analysis, as the evaluation of Biber's dimensions can be used to understand that, for example, Dimension 4 is a key dimension for MFTs, or that Dimension 5 and 6 are relatively unimportant. This knowledge is useful by itself, but can also inform future multi-dimensional studies, which in order to find the internal structure of the register, should perhaps focus on modality and other persuasive linguistic features.

In summary, besides concluding that MFTs are similar to Professional Letters, the insight given by this analysis is the individuation of the space of variation for these texts and their role within the ecosystem of the English language. An interested analyst can now predict which features to expect in MFT texts and with which frequency. Such knowledge can empower, for example, forensic linguists who are interested in base-rate knowledge of forensic registers. Similarly, for more theoretical purposes, analyses as the one presented constitute another piece of the puzzle in search for a comprehensive descriptive and predictive framework of the registers of English and for the understanding of the nature of the linguistic features, their extra-linguistic predictive factors, and their history and evolution.

5. Conclusions

The aims of the present chapter were to demonstrate the usefulness of Biber's model of English register variation as well as to show how this can be applied for new data using MAT.

The purpose of the chapter was not only to present a method to do so, but also to encourage new research that exploits the model as a different way to perform multi-dimensional analysis.

Firstly, the application of MAT to the LOB corpus revealed that despite some differences in the tagging of basic parts of speech, MAT can obtain largely the same dimension scores as Biber (1988) and assign text types similarly as Biber (1989). The only questionable results are found in the Dimension 3 scores, although these do not substantially affect the attribution of text types. Besides the reliability of MAT, this test also demonstrates the stability of Biber's original model, which can return the same results even if slightly different taggers are used.

Secondly, the validity of the model was again demonstrated by applying MAT to the Brown corpus, a corpus similar to LOB, for which therefore similar attribution of text types and dimension scores were expected. The large majority of the results prove that Biber's model is not only valid for LOB, but that it is also applicable to new data sets.

Finally, the last step of this chapter was to demonstrate how the model could be applied to new data sets using MAT. Although the corpus of malicious forensic texts used was not large and stratified enough for a full multi-dimensional analysis, the application of Biber's model of English shed light on the register identity of the corpus, the location of malicious forensic texts within the English language, and their linguistic peculiarities. Through this analysis it was revealed that malicious forensic texts are linguistically similar to professional letters and are often characterized by large scores in Dimension 4, the dimension of expression of persuasion.

In addition to the results above, the analyses reported also encourage reflections regarding the importance of using previously generated multi-dimensional models. Since the introduction of multi-dimensional analysis research by Biber (1988), a lot of attention has

been given to the method itself and to its potential for new research. A vast array of studies have applied the multi-dimensional method to registers other than those used in his seminal study and produced exciting results. However, not many studies have exploited the dimensions of variation in English or text types produced by other multi-dimensional analyses. If it is believed that the potential of a multi-dimensional analysis carried out on a representative corpus of a register is to produce a model of said register, which is descriptive and predictive, then it is also advisable that these models become the bricks upon which new research is built. This is particularly the case for those multi-dimensional models that aim at being comprehensive, such as Biber's model of the English language. Research that builds on the foundations of previous multi-dimensional studies is very much welcomed, especially for applied problems. An example of such applications is Crosthwaite's (2016) analysis of a longitudinal corpus of English for Academic Purposes students' writings, in which MAT is used to plot student texts onto Biber's model in order to assess their progress. Such applications reveal the possibilities that past multi-dimensional studies offer and should encourage researchers using this methodology to make their models available and applicable for other researchers.

References

Al-Surmi, Mansoor. 2012. “Authenticity and TV Shows: A Multidimensional Analysis Perspective.” *TESOL Quarterly* 46 (4): 671–694. doi:10.1002/tesq.33.

Berber Sardinha, Tony. 2014. “25 Years Later: Comparing Internet and Pre-Internet Registers.” In *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*, edited by Tony Berber Sardinha and Marcia Veirano Pinto, 81–105. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Berber Sardinha, Tony, and Marcia Veirano Pinto. 2017. “American Television and off-

Screen Registers: A Corpus-Based Comparison." *Corpora* 12 (1): 85–114. doi:10.3366/cor.2017.0110.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

———. 1989. "A Typology of English Texts." *Linguistics* 27 (1): 3–43.

———. 1993. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8 (4): 243–57. doi:10.1093/llc/8.4.243.

———. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge; New York: Cambridge University Press.

———. 2003. "Variation among University Spoken and Written Registers: A New Multi-Dimensional Analysis." *Language and Computers* 46 (1): 47–70.

———. 2014. "Using Multi-Dimensional Analysis to Explore Cross-Linguistic Universals of Register Variation." *Languages in Contrast* 14 (1): 7–34.

Biber, Douglas, and Jena Burges. 2000. "Historical Change in the Language Use of Women and Men: Gender Differences in Dramatic Dialogue." *Journal of English Linguistics* 28 (1): 21–37.

Biber, Douglas, and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge; New York: Cambridge University Press.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. "Speaking and Writing in the University: A Multidimensional Comparison." *TESOL Quarterly* 36 (1): 9–48. doi:10.2307/3588359.

Biber, Douglas, and Jesse Egbert. 2016. "Register Variation on the Searchable Web: A Multi-Dimensional Analysis." *Journal of English Linguistics* 44 (2): 95–137. doi:10.1177/0075424216628955.

Biber, Douglas, and Edward Finegan. 1989. "Drift and the Evolution of English Style: A

History of Three Genres." *Language* 65: 487–517.

———. 1994. "Multi-Dimensional Analyses of Authors' Styles: Some Case Studies from the Eighteenth Century." In *Research in Humanities Computing 3*, edited by D Ross and D Brink, 3–17. Oxford: Oxford University Press.

Biber, Douglas, and James Jones. 2005. "Merging Corpus Linguistic and Discourse Analytic Research Goals: Discourse Units in Biology Research Articles." *Corpus Linguistics and Linguistic Theory* 1 (2): 151–82. doi:10.1515/cllt.2005.1.2.151.

Conrad, Susan. 1996. "Investigating Academic Texts with Corpus-Based Techniques: An Example from Biology." *Linguistics and Education* 8: 299–326.

———. 2001. "Variation among Disciplinary Texts: A Comparison of Textbooks and Journal Articles in Biology and History." In *Variation in English: Multi-Dimensional Studies*, edited by Susan Conrad and Douglas Biber, 94. Harlow: Longman.

Crosthwaite, Peter. 2016. "A Longitudinal Multidimensional Analysis of EAP Writing: Determining EAP Course Effectiveness." *Journal of English for Academic Purposes* 22.

Francis, Winthrop Nelson, and Henry Kucera. 1979. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Brown University.

Fraser, Bruce. 1998. "Threatening Revisited." *Forensic Linguistics* 5 (2): 159–73.

Gales, Tammy. 2011. "Identifying Interpersonal Stance in Threatening Discourse: An Appraisal Analysis." *Discourse Studies* 13: 27–46. doi:10.1177/1461445610387735.

———. 2015. "Threatening Stances: A Corpus Analysis of Realized vs Non-Realized Threats." *Language and Law/Linguagem E Direito* 2 (2): 1–25.

Gray, Bethany. 2013. "More than Discipline: Uncovering Multi-Dimensional Patterns of Variation in Academic Research Articles." *Corpora* 8 (2): 153–81.

Grieve, Jack. 2014. "A Multi-Dimensional Analysis of Regional Variation in American

English." In *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*, edited by Tony Berber Sardinha and Marcia Veirano Pinto, 3–35. Amsterdam: John Benjamins.

Grieve, Jack, Douglas Biber, and Eric Friginal. 2011. "Variation among Blogs: A Multi-Dimensional Analysis." *Genres on the Web* 42: 303–22.

Johansson, Stig, Geoffrey Leech, and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. University of Oslo.

Napier, M, and S Mardigian. 2003. "Threatening Messages: The Essence of Analyzing Communicated Threats." *Public Venue Security*.

Nini, Andrea. 2015. "Authorship Profiling in a Forensic Context." Aston University, UK.

———. 2017. "Register Variation in Malicious Forensic Texts." *International Journal of Speech, Language and the Law* 24 (1). doi:10.1558/ijsll.30173.

Olsson, John. 2003. *Forensic Linguistics: An Introduction to Language, Crime and the Law*. London: Continuum.

Quaglio, Paulo. 2009. *Television Dialogue: The Sitcom Friends vs Natural Conversation*. Amsterdam; Philadelphia: John Benjamins Publishing.

Solan, Lawrence M., and Peter M. Tiersma. 2005. *Speaking of Crime: The Language of Criminal Justice*. Chicago: University of Chicago Press.

Svartvik, Jan. 1990. *The London-Lund Corpus of Spoken English: Description and Research*. Lund, Sweden: Lund University Press.

Titak, Ashley, and Audrey Roberson. 2013. "Dimensions of Web Registers: An Exploratory Multi-Dimensional Comparison." *Corpora* 8 (2): 235–60.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In *Proceedings of*

HLT-NAACL 2003, 252–59.