spurious – we would be hard-pressed to explain why similar pictures of diachronic change emerge from different corpora, since on that assumption we would expect the results from different corpora to differ randomly. But the pictures emerging from studies of the Family and studies of ARCHER clearly *are* similar. Thus, a more cautious interpretation of Millar's data is called for. We might posit, perhaps, that the use of a sampling frame such as that of the Brown Corpus somehow evens out the large year-to-year differences observable in a single-source corpus such as *TIME*. But it remains to be worked out in detail exactly how this would come about. In the meantime, we would argue that in extending the Brown Family to the twenty-first century, snapshot periods shorter than three decades may be prudent, where practical. In fact, exactly this approach has been taken in the construction of British English 2006 (Baker 2009), a match for LOB that samples 2005–2007 – a lapse of *fifteen* years from FLOB, not thirty.

As we noted, Biber's (2004) study of modality and other markers of stance over time was also a study of registers. *Registers* are groupings of texts defined by external factors – that is, social or situational features of the medium they use, the context in which they were produced, or the purpose of their creation. It is in fact his approach to variation across registers which forms the core of Biber's research methods, and it is to this approach to the study of language that we now turn.

5.4 The multi-dimensional approach to variation

5.4.1 An overview of the MD method

The multi-dimensional (MD)[7] approach to studying textual variation is associated primarily with Douglas Biber. It was first introduced in a study (Biber 1986) which aimed to explain certain puzzling findings in earlier work on variation between speech and writing, and between different registers of each. Biber (1986: 386) argues that work in this field had produced contradictory results because of differences in the *linguistic features* taken into account when contrasting two or more registers. Biber suggested an innovative, much more comprehensive approach, which looks at the use of a very large range of features of language in different registers and uses statistical techniques to weave them together into a more complicated and subtle picture of how registers differ from one another.

This more sophisticated approach looks at a *list* of sixty-seven linguistic features, in contrast to earlier studies which typically focused on one feature or a much smaller group of features (hence it can be labelled a 'multi-feature' approach). Biber developed this list of features on the basis of a survey of the previous literature on distinguishing spoken and written discourse. The next step in the MD method is to measure the frequency of each of the features within

a corpus sampled from a heterogeneous set of registers. Biber's (1988) study of Standard English used samples drawn largely from LOB, covering all fifteen sections of the Brown sampling frame, but also adding various spoken texts from the London-Lund Corpus (see section 4.2) as well as samples of personal and professional letter-writing. Subsequent MD studies (discussed below) have employed corpora of comparable size and heterogeneity.

A statistical analysis is then applied to (normalised) frequencies of the many linguistic features. The purpose of this statistical procedure – called a *factor analysis* – is to cluster together linguistic features which tend to vary with one another.[8] This is summarised by Biber *et al*. (1998: 278) as follows:

> In a factor analysis, the correlations between a large number of variables (i.e. the linguistic features) are identified, and the variables that are distributed in similar ways are grouped together. Each group of variables is a factor – which is then interpreted functionally as a dimension of variation . . .

So, for instance, when Biber (1986, 1988) applied this analysis, it emerged that texts which contain (relatively) many past-tense verbs *also* tend to contain (relatively) many third person pronouns. So these two features are grouped together. The factor analysis continues in this way until it has reduced the very large list of linguistic features to a much smaller number of *factors* that describe the variation among the texts in the dataset. The key to the MD approach, and the reason why it is 'multi-dimensional', is that these factors are interpreted as *dimensions* – that is, independent scales along which a text can vary (and as there are many factors, the approach is multi-dimensional). For instance, we know that a text might be formal or informal. We also know that a text might concern concrete subject matter or abstract subject matter. However, there is no necessary relationship between these two parameters of the register. A text on abstract matters could equally well be formal or informal, as could a text on concrete matters. More subtly, there is also the possibility of *gradience* – rather than a binary opposition – between concrete and abstract, formal and informal. So formality and concreteness of subject matter are independent, and we can imagine them as the *x* and *y* axes on a two-dimensional graph. Variation in abstractness might affect the 'horizontal' position of a text on the graph, variation in formality the 'vertical' position – that is, two independent sliding scales.

The visual metaphor of two dimensions for two independent forms of register variation is an appealing and powerful one. However, it is problematic because choosing abstractness and formality as our two dimensions is essentially arbitrary. Why only two types of variation? Why not three or four? Where do you stop adding dimensions? Which dimensions of variation in situation, context and purpose are actually significant for studying variation in *language*? Biber's MD method offers a way around these questions. The factors which emerge from the statistical analysis discussed above must, by virtue of the way they have been calculated from linguistic feature frequency statistics, necessarily represent aspects of variation that *are* significant for studying language. So Biber argues

that the factors can be considered as *dimensions* of register variation. On the basis of the linguistic features grouped together on each dimension, Biber proposes a functional interpretation for that dimension. The dimensions thus link together the functional requirements of a particular register with particular linguistic features that are favoured by those functional requirements.

This is quite hard to understand in the abstract, so let us take a particular example. Factor 2 in Biber's statistical results puts together the following linguistic features (Biber 1988: 102, 108–9):

- *high frequency of*: past-tense verbs, third person pronouns, perfect aspect verbs, public verbs, synthetic negation, present participial clauses;
- *low frequency of*: present-tense verbs, attributive adjectives.

Biber argues that features such as the past tense (and, thus, relative lack of present tense) and third person pronouns relate to the function of narrative discourse: namely, the relation of *past events* with *specific participants*. Meanwhile, a high frequency of attributive adjectives is associated with elaborate noun phrases, a feature of non-narrative discourse, and so the opposite feature – *low* frequency of attributive adjectives – becomes associated with narrative. Biber thus allots to this dimension the functional label 'Narrative versus Non-Narrative Concerns'.

Biber (1988) proposes six such dimensions from seven statistical factors, one of the factors being statistically too weak to be safely interpreted as a dimension. An earlier version of his analysis of Standard English (Biber 1986) found three dimensions; in later work (e.g. Biber 1989), he tends to focus on only the five strongest factors, to which he assigns the following functional labels:

- Dimension 1: 'Involved versus Informational Production'
- Dimension 2: 'Narrative versus Non-Narrative Concerns'
- Dimension 3: 'Explicit versus Situation-Dependent Reference'
- Dimension 4: 'Overt Expression of Persuasion'
- Dimension 5: 'Abstract versus Non-Abstract Information'

Biber does not stop at identifying these dimensions but uses the raw statistics for the linguistic features to produce an overall 'score' for each register on each dimension. As noted above, these vary incrementally, so on each dimension there are registers with high scores, registers with low scores and registers with middling scores. Furthermore, because the dimensions are substantially independent, the relative ordering of the registers may be completely different from dimension to dimension. The registers that score low on Dimension 2, for instance, include telephone conversations, professional letters, academic prose and official documents. The registers that score high include all kinds of fiction, biography and spontaneous speeches (Biber 1988: 136). This precise spread is not observed on any other dimension. An illustration of the spread of registers on Dimension 2 is given as Figure 5.1.

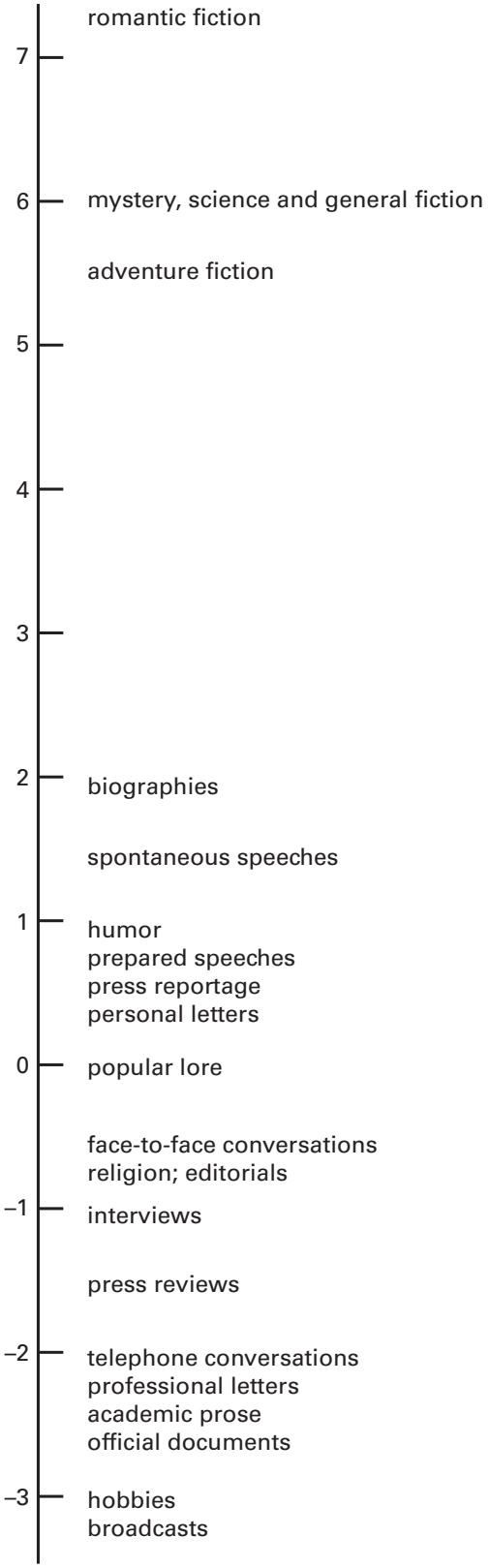

Figure 5.1 *Biber's Dimension 2,* Narrative versus Non-Narrative Concerns*; narrative texts have high scores, non-narrative texts have low scores. Reproduced from Biber (1988).*

Biber (1986) argues, critically, that no one dimension equates to a straightforward distinction between speech and writing, and that all the dimensions are needed to characterise the differences between the registers in his dataset. We can see this from the functional labels on the dimensions. Prototypical conversation is involved rather than informational – but this is not true of *all* spoken registers: consider broadcast speech such as, say, a television news report. Similarly, prototypical conversation tends to make use of situation-dependent reference rather than explicitly spelling out what is being discussed – but some spoken registers do *not* do this, such as prepared and spontaneous speeches. In fact, the oral–literate distinction is represented to some degree by three different dimensions – 1, 3 and 5. Equally, on Dimension 2 (see Figure 5.1), there are both spoken and written registers that score very highly, as well as both spoken and written registers that score very low.

It is in this way that Biber uses the results of his MD approach to English to explain the previous contradictory findings of research in the 1960s, 1970s and early 1980s. Because only one or a few linguistic features were being investigated to map out the difference between speech and writing, earlier researchers could not identify multiple dimensions of variation. Thus, Biber's initial work on the MD model resulted in the most detailed model of register variation yet proposed, providing a notable demonstration of the power of radically corpus-based, computational and statistical approaches to problems in text and language analysis.

5.4.2 Applications of the MD approach

Since developing the MD approach, a clear strand of Biber's research has focused on extending the method or applying it to new areas of investigation. The pattern of register variation developed in Biber (1988) has proven a useful starting point for a number of diverse analyses. For example, Biber and Finegan (1989) undertake a diachronic analysis within the MD framework of Biber's (1988) dimensions. Using a dataset of fiction, essays and personal letters from the past four hundred years, Biber and Finegan examine the evolution of style within these three registers, using the three dimensions linked to the oral–literate distinction as a basis for comparison. A key methodological point is that Biber and Finegan do not repeat the process that produced the dimensions on this new data – as that would not allow comparison between datasets. Rather, the dimensions established on the basis of Biber's (1988) corpus are treated as given, and the historical genres are positioned on those dimensions. Biber and Finegan find major changes over time relative to each of the three dimensions; furthermore, the patterns for each register on each of the three dimensions are generally alike: '17th- and 18th-century texts tend to be moderately or extremely literate, with a transition towards more oral styles in the 19th century and the development of a distinctly oral characterisation in the modern period . . . across the four centuries

all genres have tended towards more involved, more situated, and less abstract styles' (Biber and Finegan 1989: 507). They convincingly connect this finding to the social history of literacy and style across this period. In contrast to Biber's (2004) work on modals over time, and indeed to most diachronic corpus analysis, Biber and Finegan's emphasis is very clearly on the development of the registers, rather than of any particular linguistic feature or features.

In another direction, Biber (1989) applies the MD approach to the study of text types. Text types are distinct from registers in a crucial respect: while a register is a group of texts defined on the basis of language-external features (i.e. context, medium, purpose), a text type is a group of texts defined on the basis of linguistic similarity, with no necessary implication that they are from similar contexts of use. Biber again uses the dimensions previously established, this time as a means of measuring the distance of each text in his data from each other. On the basis of these distances – relying on all five dimensions, and thus on a very large proportion of his underlying list of linguistic features – he uses a *clustering* procedure (rather than a factor analysis on this occasion) to group texts into eight distinct text types.[9] These text clusters form groups in some ways similar to the registers. However, Biber argues (1989: 39) that this approach identifies finer distinctions among text types than were previously considered to exist; for instance, Biber identifies two different text types with characteristics of narrative, whereas earlier studies had often described 'narrative' as a unitary text type. Of course, given that narrativity is a *dimension* in Biber's model, it is not surprising that a high position on that dimension should characterise more than one text type.

Perhaps the most surprising application of the MD approach has been as a tool for cross-linguistic analysis. Biber (1995a) presents the results of a lengthy programme of collaborative research applying MD methods to corpora of Somali, Korean and Nukulaelae Tuvaluan – three languages genetically and geographically quite separate from one another and from English. In view of this, it is interesting that, when the dimensions that emerge for each language are compared, the overall picture is one of *similarity* (Biber 1995a: 278). All three of these languages, like English, have multiple dimensions that relate to the oral–literate distinction; all lack a single, clear speech-versus-writing distinction. Furthermore, 'each language has dimensions that mark personal stance [. . . and . . .] a dimension marking narrative versus non-narrative discourse' (Biber 1995a: 237). These findings are potentially highly significant, for functional and typological linguistics as well as text linguistics and corpus linguistics, in that they may point the way towards universals of the *functions to which language is put* – which, in turn, affect linguistic choices in ways captured by the different dimensions.[10] More recently, the MD method has been applied to Spanish (Biber *et al*. 2006), with the same kinds of pattern emerging.

Biber's most recent work within the MD framework has focused on language as it is used in the context of universities. The ultimate goal of this undertaking is to assist, especially, non-native speakers in English-speaking higher educational

settings (Biber 2006: 2). Using a tailor-made corpus of 2.7 million words, and a list of 129 linguistic features, Biber undertakes a full MD analysis of the different university registers (textbooks, academic speech in different contexts and so on). In this case, then, the prior results of Biber (1988) did not form part of the analysis. Biber (2006: 181–2) argues that this procedure is appropriate when approaching a new discourse domain, where that domain contains many registers. Another way to think about this might be that the registers represented within the university domain only overlap marginally with the domain of general English examined by Biber (1988) (for instance, the register of academic writing is present in both). It cannot therefore be assumed that the same dimensions will differentiate registers in the new domain. Some functional factors that were relevant to general English may not apply; likewise, university-specific functional factors may emerge. In fact, three out of the four dimensions that emerge for the university registers have parallels to a dimension identified for general English (Biber 2006: 211).

The MD approach is not the only methodology applied in Biber's study of university language. Vocabulary usage, variation of particular grammatical features and the expression of stance by a variety of linguistic devices all form a part of his analysis. In every case, the variation of these features *across separate registers* in the university domain is the key focus of Biber's analysis. Another major element of the analysis is *lexical bundles* – that is, highly frequent multi-word sequences such as *in the light of*.[11] The notion of lexical bundles was first used in Biber *et al*.'s (1999; Chapter 13) corpus-based reference grammar of English (see also Biber and Conrad 1999).

Methodologically and technically, 'lexical bundles' are simply n-grams – recurring sequences of *n* words. However, the term has come to be associated with Biber and colleagues' particular approach to the interpretation of n-grams, rooted in their descriptive goal of capturing the overall characteristics of registers (individually or in contrast to one another). This approach, most notably, concentrates on the 'structural and functional' interpretation of the lexical bundles (Biber 2006: 172). As with the dimensions in the MD method, the aim is to explain *functionally* the frequencies of particular bundles across registers. Biber also notes that '[t]he patterns of use for lexical bundles are strikingly different from those found for traditional lexico-grammatical features'. For example, consider the lexical bundles identified by Biber *et al*. (2004: 384, 389–90) as having the function in classroom teaching of expressing *personal epistemic stance*: these include *I don't know if*, *I don't know how*, *you know what I*, *I thought it was* and others. Although they have various grammatical structures, these all work to express uncertainty in some way – and share the fixity of form that results in their identification as lexical bundles.

In sum, then, Biber's applications of his MD model have contributed significantly to many different subdisciplines of linguistics. However, the approach has not been more widely adopted. In a generally positive review of Biber (1995a), Kilgarriff (1995: 613) notes that:

> [t]he MD research program has been proceeding for over a decade, but as yet the methodology has only been used by Biber and a small group of collaborators. This could be because other readers have not been impressed by the work, or it could just be that the methodology is technically difficult and time-consuming to implement. I suspect the latter. Each stage of the methodology involves a new set of obstacles and skills, and an MD analysis is not lightly undertaken.

Regrettably, this is still true, more or less, at the time of writing – sixteen years on from Kilgarriff's review and twenty-five years on from the initial appearance of Biber's (1986) paper in the journal *Language*. Other methodologies developed by Biber and his colleagues have been picked up and exploited to full advantage by other researchers. For example, Culpeper and Kytö (2002) investigate the use of lexical bundles in dialogue text in Early Modern English. Lexical bundles, however, are computationally and statistically much more straightforward than the factor analysis that underlies an MD analysis. It is almost certainly this complexity that has inhibited the widespread uptake of what appears to be a useful technique. It has been argued by Tribble (1999) and by Xiao and McEnery (2005) that the methodological complexity of the MD analysis can be reduced by using a keywords analysis to achieve much the same effect; likewise Crossley and Louwerse (2007) show that an MD analysis is possible using only bigram frequencies as the input features. However, that aside, the lack of uptake of MD methods by a wider community supports, we would argue, a point we made in section 2.5.4: it is imperative that corpus tools accessible to the non-technical user should continue to develop and to incorporate more and more specialised and complex procedures, up to and beyond the level of complexity of an MD analysis. Unless this occurs, and procedures like MD analysis become as easy for the user to run as a straightforward concordance, without technical training linguists will never be able to access and benefit from the full gamut of corpus methodologies.

5.4.3 Criticisms of Biber's MD methodology

While we have discussed briefly a practical impediment to the uptake of the MD approach, Biber's approach to language variation has also been subject to more fundamental criticism. Not all such criticisms have been entirely well founded, however. For example, Watson (1994) attempts to apply Biber's MD model to the analysis of one postmodern author's novels, a purpose quite different from the model's original use, and on the basis of problems encountered during this effort argues for a set of deficiencies in the MD approach per se. The publication of this paper sparked a debate in the pages of the journal *Text* (Biber 1995b; Watson 1995) in which Biber refutes Watson's criticisms comprehensively. This debate is in itself of limited significance, but it serves to underline the critical importance of careful methodological awareness in the application of corpus linguistic techniques.

Three criticisms of the MD approach, however, deserve some further consideration. All are broadly methodological. The first – and least serious – is the nature of the dataset on which Biber's study of spoken and written English was based (and which thus was also the foundation of several of Biber's later investigations). It is rather small and consists of short samples of longer texts (like LOB, from which it was largely drawn). In fact, we might say that Biber's MD studies, especially those of languages other than English, epitomise the small, carefully designed sample corpus approach pioneered by UCL and Lancaster (see section 1.4.3), and thus the drawbacks of that kind of dataset necessarily apply in full to the MD framework. However, in response to criticisms along these lines, Biber has done important empirical work assessing the impact of corpus representativeness (Biber 1990, 1993). For example, Biber (1990) demonstrates that re-running his MD analyses on small subsets of his corpus produces very similar results to the original analysis. He argues that this supports the contention that the corpus is sufficiently representative.

A more important criticism concerns the replicability of Biber's results. While much of Biber's own further research supports his original results for English, Doyle (2005: 4) points out that others have had difficulty replicating Biber's findings independently, due in part to the unavailability of relevant software and datasets. Furthermore, there are reports in the literature of data suggesting that some of Biber's dimensions may not be statistically replicable in other general English datasets (Lee 2001). If we were to take a devil's advocate approach to Biber's (2006) MD analysis of university language, which found similar but not identical dimensions to Biber (1988), we might argue that this actually demonstrates that the dimensions that emerge are relative to the spectrum of particular texts under analysis at any given time. Hence, no strong claim can be made for the validity of dimensions beyond the corpus they are calculated for. This may be a legitimate argument as far as the *precise identity* of dimensions is concerned, but the full body of Biber's (cross-domain and cross-linguistic) MD research provides powerful support for consistency in the *general outline* of observable dimensions. Replicability remains, however, something of a concern for the MD framework.

Finally, some criticism has been directed at Biber's approach to the choice of linguistic features to use in order to generate the corpus statistics that drive the factor analysis calculation. Altenberg (1989: 171–3) points out that the choice of features is actually one of the keys to the results of a factor analysis, and that some of Biber's (1988) features are questionable on one ground or another. For example, the choice of features is limited to what can be searched for in a part-of-speech-tagged, but not parsed, corpus. Developing this criticism, Ball (1994) discusses in detail some of the problems that can arise in searching corpora for grammatical patterns, and argues that studies like Biber's, based on multiple features whose identification by a corpus search is not unproblematic, are premature. A different kind of problem emerges with regard to some other features. Altenberg (1989: 172) points out that preposition phrases (which constitute just one of Biber's sixty-seven features) are functionally heterogeneous, as they can either postmodify a