Matsumoto, Yuji/Kitauchi, Akira/Yamashita, Tatsuo/Hirano, Yoshitaka/Matsuda, Hiroshi/Takaoka, Kazuma/Asahara, Masayuki (2000), *Morphological Analysis System ChaSen Version 2.2.1 Manual.* NIST Technical Report.

McEnery, Tony/Wilson, Andrew (2001), *Corpus Linguistics.* 2nd edition. Edinburgh: Edinburgh University Press.

Miller, George (1957), Some Effects of Intermittent Silence. In: *American Journal of Psychology* 52, 311−314.

Möbius, Bernd (2003), Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis. In: *International Journal of Speech Technology* 6, 57−71.

Muller, Charles (1977), *Principes et méthodes de statistique lexicale*. Paris: Hachette.

Naranan, Sundaresan/Balasubrahmanyan, Vriddhachalam K. (1998), Models for Power Law Relations in Linguistics and Information Science. In: *Journal of Quantitative Linguistics* 5, 35−61.

Oakes, Michael (1998), *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Sampson, Geoffrey (2002), Review of Harald Baayen: Word Frequency Distributions. In: *Computational Linguistics* 28, 565−569.

Simon, Herbert A. (1955), On a Class of Skew Distribution Functions. In: *Biometrika* 42, 425−440.

Sinclair, John (2005), Corpus and Text: Basic Principles. In: Wynne, Martin (ed.), *Guide to Good Practice in Developing Linguistic Corpora*. Oxford: Oxford Books, 1−6. Available from http://ahds.ac.uk/litlangling/linguistics/index.html.

Ueyama, Motoko/Baroni, Marco (2006), Automated Construction and Evaluation of a Japanese Web-based Reference Corpus. In: *Proceedings of Corpus Linguistics 2005*, available from http://www.corpus.bham.ac.uk/PCLC.

Tweedie, Fiona/Baayen, R. Harald (1998), How Variable May a Constant Be? Measures of Lexical Richness in Perspective. In: *Computers and the Humanities* 32, 323−352.

Witten, Ian/Moffat, Alistair/Bell, Timothy (1999), *Managing Gigabytes.* 2nd edition. San Francisco: Morgan Kaufmann.

Zipf, George Kingsley (1949), *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.

Zipf, George Kingsley (1965), *The Psycho-biology of Language*. Cambridge MA: MIT Press.

Marco Baroni, Trento (Italy)

38. Multi-dimensional approaches

1. Studying register and register variation

For many years, researchers have studied the language used in different situations: the description of *registers*. *Register* is used here as a cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances.

In many cases, registers are named varieties within a culture, such as novels, letters, memos, editorials, sermons, and lectures. However, registers can be defined at any level of generality, and more specialized registers may not have widely used names. For example, 'academic prose' is a very general register, while 'methodology sections in experimental psychology articles' is a much more highly specified register.

Although registers are defined in situational terms, they can also be compared with respect to their linguistic characteristics: the study of *register variation*. Register variation is inherent in human language: a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar reflecting a range of situational factors. The ubiquitous nature of register variation has been noted by a number of scholars, for example:

> "register variation, in which language structure varies in accordance with the occasions of use, is all-pervasive in human language" (Ferguson 1983, 154)
>
> "no human being talks the same way all the time ... At the very least, a variety of registers and styles is used and encountered." (Hymes 1984, 44)

However, despite the fundamental importance of register variation, there have been few comprehensive analyses of the register differences in a language. This gap is due mostly to methodological difficulties: until recently, it has been unfeasible to analyze the full range of texts, registers, and linguistic characteristics required for comprehensive analyses of register variation. With the availability of large on-line text corpora and computational analytical tools, such analyses have become possible. Multi-dimensional (MD) analysis – the focus of the present article – is a corpus-based research approach developed for the comprehensive analysis of register variation.

2. Theoretical background

In a few cases, registers can be distinguished by the presence of distinctive *register markers*: linguistic features restricted to a single register. For example, Ferguson (1983) describes how the grammatical routine known as 'the count', as in *the count is two and one*, is a distinctive register marker of baseball game broadcasts. In most cases, though, register differences are realized through the relative presence or absence of *register features* – core lexical and grammatical features – rather than by the presence of a few distinctive register markers. Register features are found to some extent in almost all texts and registers, but there are often large differences in their relative distributions across registers. In fact, many registers are distinguished only by a particularly frequent or infrequent occurrence of a set of register features.

Register analyses of these core linguistic features are necessarily quantitative, to determine the relative distribution of linguistic features. Further, such analyses require a comparative approach. That is, it is only by quantitative comparison to a range of other registers that we are able to determine whether a given frequency of occurrence is notably common or rare. A quantitative comparative approach allows us to treat register as a continuous construct: texts are situated within a continuous space of linguistic variation, enabling analysis of the ways in which registers are more or less different with respect to the full range of core linguistic features.

It turns out, though, that the relative distribution of common linguistic features, considered individually, cannot reliably distinguish among registers. There are simply too many different linguistic characteristics to consider, and individual features often have idiosyncratic distributions. However, when analyses are based on the *co-occurrence* and *alternation* patterns for groups of linguistic features, important differences across registers are revealed.

The theoretical importance of linguistic co-occurrence was recognized well before corpus-based methods were developed to analyze these patterns. For example, Ervin-Tripp (1972) identified 'speech styles' as varieties that are defined by a shared set of co-occurring linguistic features, while Brown/Fraser (1979, 38−39) observed that it can be "misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the co-occurrence of sets of markers".

The Multi-Dimensional (MD) approach was developed as a corpus-based methodology to analyze the linguistic co-occurrence patterns associated with register variation. The following section provides a conceptual overview of the approach, while section 4 summarizes the methodological techniques used for MD analyses.

3. Conceptual introduction to the multi-dimensional approach

MD analysis was developed as a corpus-based methodological approach to: (1) identify the salient linguistic co-occurrence patterns in a language, in empirical/quantitative terms; and (2) compare registers in the linguistic space defined by those co-occurrence patterns. The approach was first used in Biber (1985, 1986) and then developed more fully in Biber (1988).

The notion of linguistic co-occurrence has been given formal status in the MD approach, in that different co-occurrence patterns are analyzed as underlying *dimensions* of variation. The co-occurrence patterns comprising each dimension are identified quantitatively. That is, based on the actual distributions of linguistic features in a large corpus of texts, statistical techniques (specifically factor analysis) are used to identify the sets of linguistic features that frequently co-occur in texts. The methods used to identify these co-occurrence patterns are described in section 4.

Qualitative analysis is required to interpret the functions associated with each set of co-occurring linguistic features. The dimensions of variation have both linguistic and functional content. The linguistic content of a dimension comprises a group of linguistic features (e. g., nominalizations, prepositional phrases, attributive adjectives) that co-occur with a high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted in terms of the situational, social, and cognitive functions most widely shared by the linguistic features. That is, linguistic features co-occur in texts because they reflect shared functions.

A simple example is the way in which first and second person pronouns, direct questions, and imperatives are all related to interactiveness. Contractions, false starts, and generalized content words (e. g., *thing*) are all related to the constraints imposed by real-time production. The functional bases of other co-occurrence patterns are less transparent, so that careful qualitative analyses of particular texts are required to help interpret the underlying functions.

In sum, the salient characteristics of the MD approach are:

- The research goal of the approach is to describe the general patterns of variation among registers, considering a comprehensive set of linguistic features and the range of registers in the target domain of use.
- The unit of analysis in the approach is each text, rather than individual linguistic constructions.
- The importance of variationist and comparative perspectives is assumed by the approach. That is, different kinds of text differ linguistically and functionally, so that analysis of a single text variety cannot adequately represent a discourse domain. From a quantitative point of view, it is not possible to determine which linguistic distributions are noteworthy without comparison to other text varieties.
- The approach is explicitly multi-dimensional. That is, it is assumed that multiple parameters of variation will operate in any discourse domain.
- The approach is empirical and quantitative. Analyses are based on normed frequency counts of linguistic features, describing the relative distributions of features across the texts in a corpus. The linguistic co-occurrence patterns that define each dimension are identified empirically using multivariate statistical techniques (see also article 40).
- The approach synthesizes quantitative and qualitative/functional methodological techniques. That is, the statistical analyses are interpreted in functional terms, to determine the underlying communicative functions associated with each distributional pattern. The approach is based on the assumption that statistical co-occurrence patterns reflect underlying shared communicative functions.

4. Methodology in the multi-dimensional approach

A complete multi-dimensional analysis follows eight methodological steps:

1. An appropriate corpus is designed based on previous research and analysis (cf. article 9). Texts are collected, transcribed (in the case of spoken texts), and input into the computer. (In many cases, pre-existing corpora can be used; cf. article 20.)
2. Research is conducted to identify the linguistic features to be included in the analysis, together with functional associations of the linguistic features.
3. Computer programs are developed for automated grammatical analysis, to identify – or ‘tag’ – all relevant linguistic features in texts (cf. articles 23, 24).
4. The entire corpus of texts is tagged automatically by computer, and all texts are edited interactively to ensure that the linguistic features are accurately identified.
5. Additional computer programs are developed and run to compute normed frequency counts of each linguistic feature in each text of the corpus.
6. The co-occurrence patterns among linguistic features are identified through a factor analysis of the frequency counts (cf. article 40).

7. The ‘factors’ from the factor analysis are interpreted functionally as underlying dimensions of variation.
8. Dimension scores for each text are computed; the mean dimension scores for each register are then compared to analyze the salient linguistic similarities and differences among registers.

In practice, there are two different types of MD study: those that carry out a full MD analysis and those that apply previously identified dimensions to new areas of research. Methodologically, the two types differ in whether or not they include steps 6 and 7. Full MD studies, such as the original MD studies (Biber 1985, 1986, 1988), identify and interpret underlying dimensions of register variation and then use those dimensions to characterize registers; they thus include all eight methodological steps. However, many MD studies use the dimensions identified in Biber (1988) to describe and compare additional registers. These studies omit steps 6 and 7; since they use the previously-identified dimensions, such studies do not require a separate factor analysis.

4.1. Linguistic features and tagging the corpus

One of the first tasks in a MD study is to identify the linguistic features to be used in the analysis. The goal here is to be as inclusive as possible, identifying all linguistic features that might have functional associations, including lexical classes, grammatical categories, and syntactic constructions. Thus, any feature associated with particular communicative functions, or used to differing extents in different text varieties, is included. Occurrences of these features are counted in each text of the corpus, providing the basis for all subsequent statistical analyses.

The identification of functionally important linguistic features for the 1988 study of register variation was relatively easy, due to the large body of previous research studies on speech and writing. That study was based on 67 linguistic features, taken from 16 major grammatical and functional categories:

1) tense and aspect markers
2) place and time adverbials
3) pronouns and pro-verbs
4) questions
5) nominal forms
6) passives
7) stative forms
8) subordination features
9) prepositional phrases, adjectives, and adverbs
10) lexical specificity
11) lexical classes
12) modals
13) specialized verb classes
14) reduced forms and discontinuous structures
15) coordination
16) negation

These features are identified in the texts of a corpus by using an automatic 'tagger' (see Biber 1988, appendix II).

The current version of the tagger has both probabilistic and rule-based components, and uses multiple large-scale dictionaries. The tagger has been developed with three primary considerations: achieving high accuracy levels; robustness across texts from different registers (with different processing options for 'oral' and 'literate' texts); and identification of a large set of linguistic characteristics (e. g., distinguishing simple past tense, perfect aspect, passive voice, and postnominal modifier for past participle verbs; identifying the gap position for *WH*-relative clauses; identifying several different kinds of complement clause, and the existence of *that*-complementizer deletion). In recent years, several linguistic distinctions have been added to the tagger as part of the analyses carried out for the *Longman Grammar of Spoken and Written English* (Biber et al. 1999); these include many lexico-grammatical features, such as mental verbs controlling *that*-clauses, or verbs of effort controlling *to*-clauses. To ensure accurate tagging, problematic linguistic features are corrected interactively using a grammar checker (see also Biber/Conrad/Reppen 1998, methodology boxes 4 and 5).

Once texts have been tagged and interactively tag-edited, other programs are used to calculate the 'normed' (or 'normalized') rate of occurrence of linguistic features in each text (e. g., the number of nouns per 1,000 words; see Biber/Conrad/Reppen 1998, methodology box 6). (Some linguistic features have non-linear distributions and so must be adjusted in other ways; see article 37.) These normed counts provide the basis for the factor analysis, described in the following section.

4.2. Factor analysis to identify the 'dimensions' of variation

As described above, co-occurrence patterns are central to MD analyses: each dimension represents a different set of co-occurring linguistic features. These co-occurrence patterns are identified quantitatively, using a statistical technique known as factor analysis; each set of co-occurring features is referred to as a *factor*. In a factor analysis, a large number of original variables – in this case the linguistic features – are reduced to a small set of derived, underlying variables: the factors. In the 1988 MD analysis, the 67 linguistic features were reduced to 7 factors.

Each linguistic feature has some relation to each factor, and the strength of that relation is represented by *factor loadings*. (The factor loading is essentially a correlation, representing the amount of variance that a feature has in common with the total pool of shared variance accounted for by a factor.)

The factor loadings for the 1988 MD analysis of spoken and written registers are given in Table 38.1. Factor loadings can range from 0.0, which shows the absence of any relationship, to 1.0 (positive or negative), which shows a perfect correlation. The factor loading indicates the extent to which a linguistic feature is representative of the dimension underlying a factor; the size of the loading reflects the strength of the co-occurrence relationship between the feature in question and the total grouping of co-occurring features represented by the factor.

As Table 38.1 shows, each linguistic feature has a loading on each factor. However, when interpreting a factor, only features with salient or important loadings are consid-

Tab. 38.1: Factor loadings in the factor analysis of register variation in English

LING FEATURE	FACT1	FACT2	FACT3	FACT4	FACT5	FACT6	FACT7
Past tense	−0.083	0.895	0.002	−0.249	−0.049	−0.052	0.021
Perfects	0.051	0.480	0.049	−0.016	−0.101	0.146	0.143
Present tense	0.864	−0.467	−0.008	0.229	−0.006	0.011	0.011
Place adverbs	−0.417	−0.060	−0.492	−0.094	−0.067	−0.018	−0.023
Time adverbs	−0.199	−0.062	−0.604	−0.020	−0.290	0.116	−0.046
1st pers. pro.	0.744	0.088	0.025	0.026	−0.089	0.008	−0.098
2nd pers. pro.	0.860	−0.043	−0.018	0.016	0.007	−0.168	−0.064
3rd pers. pro.	−0.053	0.727	−0.074	−0.018	−0.167	−0.076	0.138
Pronoun *it*	0.706	−0.021	−0.038	−0.034	−0.038	0.022	0.060
Dem. pronouns	0.756	−0.166	−0.001	−0.108	0.004	0.306	−0.077
Proform *any*	0.618	0.046	0.011	0.085	−0.094	−0.085	−0.032
Proform *do*	0.821	0.004	0.071	0.049	−0.057	−0.077	−0.056
Wh questions	0.523	−0.024	0.117	−0.111	−0.032	0.036	−0.094
Nominaliz.	−0.272	−0.237	0.357	0.179	0.277	0.129	−0.019
-ing nouns	−0.252	−0.127	0.216	0.177	0.087	−0.052	0.052
Other nouns	−0.799	−0.280	−0.091	−0.045	−0.294	−0.076	−0.213
Agentless pasv.	−0.388	−0.145	0.109	0.060	0.430	0.063	−0.057
By pasv.	−0.256	−0.189	0.065	−0.124	0.413	−0.089	−0.045
Stative *be*	0.713	0.056	0.075	0.008	0.014	0.292	0.180
Existential *there*	0.262	0.108	0.113	−0.124	−0.004	0.318	0.017
That verb clause	0.045	0.228	0.125	0.265	0.053	0.558	−0.122
That adj clause	−0.124	0.066	−0.080	0.123	0.171	0.360	0.183
Wh clause	0.467	0.143	0.221	0.032	−0.050	−0.044	−0.027
Infinitive	−0.071	0.059	0.085	0.760	−0.274	−0.005	−0.074
Advl clause *-ing*	−0.211	0.392	−0.142	−0.076	0.268	−0.217	0.121
Advl clause *-ed*	−0.025	−0.154	0.029	−0.050	0.415	−0.142	−0.059
Whiz *-ed*	−0.382	−0.336	−0.071	−0.137	0.395	−0.128	−0.103
Whiz *-ing*	−0.325	−0.114	0.080	−0.169	0.212	−0.070	−0.093
That rel. subj.	0.051	−0.036	0.021	0.019	−0.058	0.184	0.033
That rel. obj.	−0.047	0.053	0.201	0.223	−0.125	0.457	−0.065
Wh- rel. subj.	−0.087	−0.067	0.453	−0.027	−0.174	0.228	0.047
Wh- rel. obj.	−0.072	0.049	0.627	−0.060	−0.083	0.302	0.165
Wh- rel. pied pip.	−0.029	0.026	0.606	−0.144	0.046	0.280	0.192
Sentence rel.	0.550	−0.086	0.152	−0.118	−0.025	0.048	−0.041

Tab. 38.1: (continued)

LING FEATURE	FACT1	FACT2	FACT3	FACT4	FACT5	FACT6	FACT7
Advl. cl. − reason	0.661	−0.080	0.110	0.023	−0.061	0.078	−0.076
Advl. cl. − conc.	0.006	0.092	0.100	−0.071	0.010	−0.056	0.300
Advl. cl. − cond.	0.319	−0.076	−0.206	0.466	0.120	0.103	−0.007
Advl. cl. − other	−0.109	0.051	−0.018	0.008	0.388	0.102	0.109
Prepositions	−0.540	−0.251	0.185	−0.185	0.234	0.145	−0.008
Attributive adj.	−0.474	−0.412	0.176	−0.055	−0.038	−0.064	0.299
Predicative adj.	0.187	0.076	−0.089	0.248	0.311	−0.012	0.210
Adverbs	0.416	−0.001	−0.458	−0.020	−0.156	0.053	0.314
Typetoken ratio	−0.537	0.058	0.002	−0.005	−0.311	−0.228	0.219
Word length	−0.575	−0.314	0.270	−0.009	0.023	0.028	0.081
Conjuncts	−0.141	−0.160	0.064	0.108	0.481	0.180	0.217
Downtoners	−0.084	−0.008	0.021	−0.080	0.066	0.113	0.325
Hedges	0.582	−0.156	−0.051	−0.087	−0.022	−0.145	0.096
Amplifiers	0.563	−0.156	−0.028	−0.124	−0.124	0.225	−0.018
Emphatics	0.739	−0.216	0.015	−0.027	−0.188	−0.087	0.210
Disc. particles	0.663	−0.218	−0.128	−0.029	−0.096	0.165	−0.140
Demonstratives	0.040	−0.062	0.113	0.010	0.132	0.478	0.153
Pos. modals	0.501	−0.123	0.044	0.367	0.122	−0.022	0.115
Nec. modals	−0.007	−0.107	−0.015	0.458	0.102	0.135	0.042
Pred. modals	0.047	−0.056	−0.054	0.535	−0.072	0.063	−0.184
Public verbs	0.098	0.431	0.163	0.135	−0.030	0.046	−0.279
Private verbs	0.962	0.160	0.179	−0.054	0.084	−0.049	0.106
Suasive verbs	−0.240	−0.035	−0.017	0.486	0.051	0.016	−0.237
Seem/appear	0.054	0.128	0.160	−0.010	0.015	0.045	0.348
Contractions	0.902	−0.100	−0.141	−0.138	−0.002	−0.057	−0.032
That deletions	0.909	0.036	0.098	−0.059	−0.005	−0.178	−0.081
Stranded preps	0.426	0.007	−0.124	−0.210	0.023	0.340	−0.100
Split infinitives	----------- DROPPED -----------						
Split auxiliaries	−0.195	0.040	0.012	0.437	0.043	0.120	0.239
Phrasal coord.	-0.253	−0.091	0.355	−0.066	−0.046	−0.324	0.126
Clausal coord.	0.476	0.041	−0.052	−0.161	−0.139	0.218	−0.125
Synthetic neg.	−0.232	0.402	0.046	0.133	−0.057	0.176	0.110
Analytic neg.	0.778	0.149	0.017	0.125	0.019	0.001	0.037

ered. In the 1988 analysis, features with loadings smaller than .35 were considered not important in the interpretation of a factor. Positive or negative sign does not influence the importance of a loading; for example, nouns, with a loading of −.799, have a larger weight on Factor 1 than first person pronouns, with a loading of .744.

Rather than reflecting importance, positive and negative sign identify two groupings of features that occur in a complementary pattern as part of the same factor. That is, when the features with positive loadings occur together frequently in a text, the features with negative loadings are markedly less frequent in that text, and vice versa. (For more technical information about the factor analysis, see Biber 1995, chapter 5.)

4.2.1. Interpretation of factors as dimensions of variation

Factor interpretations depend on the assumption that linguistic co-occurrence patterns reflect underlying communicative functions. That is, linguistic features occur together in texts because they serve related communicative functions. The interpretation of a factor is based on (1) analysis of the communicative function(s) most widely shared by the set of co-occurring features, and (2) analysis of the similarities and differences among registers with respect to the factor.

For example, Table 38.2 lists the features with salient loadings on Factor 1 in the 1988 MD analysis (i. e., the features with loadings greater than 0.35). In the interpretation of a factor, it is important to consider the likely reasons for the complementary distribution between positive and negative feature sets as well as the reasons for the co-occurrence patterns within those sets.

On Factor 1, the interpretation of the negative features is relatively straightforward. Nouns, word length, prepositional phrases, type/token ratio, and attributive adjectives all have negative loadings larger than |.45|, and none of these features has a larger loading on another factor. These features reflect an informational focus, a careful integration of information in a text, and precise lexical choice. Text sample 1 illustrates these co-occurring linguistic characteristics in an academic article:

> *Text sample 1: Technical academic prose*
>
> Apart from these very general group related aspects, there are also individual aspects that need to be considered. Empirical data show that similar processes can be guided quite differently by users with different views on the purpose of the communication.

This text sample is typical of written expository prose in its dense integration of information: frequent nouns and long words, with most nouns being modified by attributive adjectives or prepositional phrases (e. g., *general group related aspects*, *individual aspects*, *empirical data*, *similar processes*, *users with different views on the purpose of the communication*).

The set of features with positive loadings on Factor 1 is more complex, although all of these features have been associated with interpersonal interaction, a focus on personal stance, and real-time production circumstances. For example, first and second person pronouns, *WH*-questions, emphatics, amplifiers, and sentence relatives can all be interpreted as reflecting interpersonal interaction and the involved expression of personal stance (feelings and attitudes). Other positive features are associated with the constraints

Tab. 38.2: Factor 1 features and loadings in the 1988 MD analysis of register variation

Dimension 1: Involved vs. Informational Production	
Positive features:	
private verbs	.96
that deletion	.91
contractions	.90
present tense verbs	.86
2nd person pronouns	.86
do as pro-verb	.82
analytic negation	.78
demonstrative pronouns	.76
general emphatics	.74
first person pronouns	.74
pronoun *it*	.71
be as main verb	.71
causative subordination	.66
discourse particles	.66
indefinite pronouns	.62
general hedges	.58
amplifiers	.56
sentence relatives	.55
wh questions	.52
possibility modals	.50
non-phrasal coordination	.48
wh clauses	.47
final prepositions	.43
(adverbs	.42)
Negative features:	
nouns	−.80
word length	−.58
prepositions	−.54
type/token ratio	−.54
attributive adjs.	−.47
(place adverbials	−.42)
(agentless passives	−.39)
(past participial postnominal clauses	−.38)

of real time production, resulting in a reduced surface form, a generalized or uncertain presentation of information, and a generally 'fragmented' production of text; these include *that*-deletions, contractions, pro-verb DO, the pronominal forms, and final (stranded) prepositions. Text sample 2 illustrates the use of positive Dimension 1 features in a formal conversation (an interview) from the London-Lund Corpus:

Text sample 2: Interview

B: come in . come in - - ah good morning
A: good morning
B: you're Mrs Finney
A: yes I am
B: how are you - my names Hart and this is Mr Mortlake
C: how are you

A: how do you do .
B: won't you sit down
A: thank you - -
B: mm well you are proposing . taking on . quite something Mrs Finney aren't you
A: yes I am
B: mm
A: I should like to anyhow
B: you know what you'd be going into
A: yes I do

Overall, Factor 1 seems to represent a dimension marking interactional, stance-focused, and generalized content (the features with positive loadings on Table 38.2) versus high informational density and precise word choice (the features with negative loadings). Two separate communicative parameters seem to be represented here: the primary purpose of the writer/speaker (involved versus informational), and the production circumstances (those restricted by real-time constraints versus those enabling careful editing possibilities). Reflecting both of these parameters, the interpretive label 'Involved versus Informational Production' was proposed for the dimension underlying this factor.

4.2.2. Computing dimension scores

The second major step in interpreting a dimension is to consider the similarities and differences among registers with respect to the set of co-occurring linguistic features. To achieve this, *dimension scores* are computed for each text, and then texts and registers are compared with respect to those scores. Dimension scores (or *factor scores*) are computed by summing the individual scores of the features with salient loadings on a dimension. In the 1988 MD study, only features with loadings greater than |.35| on a factor were considered important enough to be used in the computation of dimension scores. For example, the Dimension 1 score for each text was computed by adding together the frequencies of private verbs, *that*-deletions, contractions, present tense verbs, etc. − the features with positive loadings on Factor 1 (from Table 38.2) − and then subtracting the frequencies of nouns, word length, prepositions, etc. − the features with negative loadings.

All individual linguistic variables are standardized to a mean of 0.0 and a standard deviation of 1.0 before the dimension scores are computed. This process converts feature scores to a single scale representing standard deviation units, so that all linguistic features have the same range of variation and therefore equivalent weights in the computation of dimension scores (see Biber 1988, 93−97).

Once a dimension score is computed for each text, the mean dimension score for each register can be computed. Plots of these mean dimension scores allow linguistic characterization of any given register, comparison of the relations between any two registers, and a fuller functional interpretation of the underlying dimension. Standard statistical procedures (such as ANOVA) can be used to further analyze the statistical significance of differences among the mean dimension scores.

For example Figure 38.1 plots the mean dimension scores of registers along Dimension 1. The registers with large positive values (such as face-to-face and telephone conversations), have high frequencies of present tense verbs, private verbs, first and second

```
     | TELEPHONE CONVERSATIONS
     |
 35  + FACE-TO-FACE CONVERSATIONS
     |
     |
     |
 30  +
     |
     |
     |
 25  +
     |
     |
     |
 20  + Personal letters
     | PUBLIC CONVERSATIONS, SPONTANEOUS SPEECHES
     | INTERVIEWS
     |
 15  +
     |
     |
     |
 10  +
     |
     |
     |
  5  +
     | Romance fiction
     | PREPARED SPEECHES
     |
  0  + Mystery and adventure fiction
     | General fiction
     | Professional letters
     | BROADCASTS
 −5  +
     | Science fiction
     | Religion
     | Humor
−10  + Popular lore, editorials, hobbies
     |
     | Biographies
     | Press reviews
−15  + Academic prose, Press reportage
     |
     | Official documents
```

Fig. 38.1: Mean scores of registers along Dimension 1: Involved vs. Informational Production. Underlining denotes written registers; capitalization denotes spoken registers. (Adapted from Figure 7.1 in Biber 1988) ($F = 111.9$, $p < .0001$, $r^2 = 84.3\%$)

person pronouns, contractions, etc. − the features with salient positive weights on Dimension 1. At the same time, registers with large positive values have markedly low frequencies of nouns, prepositional phrases, long words, etc. − the features with salient negative weights on Dimension 1. Registers with large negative values (such as academic prose, press reportage and official documents) have the opposite linguistic characteristics: very high frequencies of nouns, prepositional phrases, etc., plus low frequencies of private verbs, contractions, etc.

The relations among registers shown in Figure 38.1 confirm the interpretation of Dimension 1 as distinguishing among texts along a continuum of involved versus informational production. At the positive extreme, conversations are highly interactive and involved, with the language produced under real-time circumstances. Registers such as public conversations (interviews and panel discussions) are intermediate: they have a relatively informational purpose, but participants interact with one another and are still constrained by real time production. Finally, at the negative extreme, registers such as academic prose are non-interactive but highly informational in purpose, produced under controlled circumstances that permit extensive revision and editing.

The statistics given for F, p, and r^2 at the bottom of Figure 38.1 report the results of statistical tests. The F and p values give the results of an ANOVA, which tests whether there are statistically significant differences among the registers with respect to their mean Dimension 1 scores. The value for r^2 is a direct measure of strength or importance. The r^2 value measures the percentage of the variance among dimension scores that can be predicted by knowing the register categories (in this case, 84.3 %).

The sections above have provided an overview of the methodological techniques used in MD analysis, illustrating the steps involved in the interpretation of a factor through consideration of Dimension 1 from the 1988 MD analysis. In section 5 below, I summarize the other major dimensions of variation from the 1988 study. Several publications provide fuller discussion of the relevant methodological issues, including: designing a representative corpus, the stability of linguistic feature counts, methodological details of factor analysis and computing factor scores, the stability of factor solutions, and the general goals, strengths, and weaknesses of MD analysis (see especially Biber 1988, chapters 4–5; 1995, chapter 5; 1990, 1993a, 1993b; Biber et al. 2003).

5. Summary of the 1988 MD analysis of English registers

The first major MD analysis (Biber 1988) was undertaken to investigate the relationship between spoken and written language in English. Most previous studies had been based on the assumption that speech and writing could be approached as a simple dichotomy; so, for example, a comparison of conversations and student essays was sometimes interpreted as representing general differences between speech and writing. In contrast, MD analysis is based on the assumption that all registers have distinctive linguistic characteristics (associated with their defining situational characteristics). Thus, the 1988 MD study of speech and writing set out to describe the relations among the full range of spoken registers and the full range of written registers – and to then compare speech and writing within the context of a comprehensive analysis of register variation.

For example, Figure 38.1 shows that there is a large range of variation among spoken registers with respect to the linguistic features that comprise Dimension 1 (‘Involved versus Informational Production’). Conversation has extremely large positive Dimension 1 scores; spontaneous speeches and interviews have moderately large positive scores; while prepared speeches and broadcasts have scores around 0.0 (reflecting a balance of positive and negative linguistic features on this dimension). The written registers simi-

Tab. 38.3: Linguistic features on Dimensions 2−5 from the 1988 MD analysis

DIMENSION 2: Narrative vs. Non-narrative Discourse	
Positive features:	
past tense verbs	.90
third person pronouns	.73
perfect aspect verbs	.48
public verbs	.43
synthetic negation	.40
present participial clauses	.39
Negative features:	
(present tense verbs	−.47)
(attributive adjs.	−.41)
DIMENSION 3: Situation-dependent vs. Elaborated Reference	
Positive features:	
time adverbials	.60
place adverbials	.49
adverbs	.46
Negative features:	
Wh relative clauses on object positions	−.63
pied piping constructions	−.61
Wh relative clauses on subject positions	−.45
phrasal coordination	−.36
nominalizations	−.36
DIMENSION 4: Overt Expression of Argumentation	
Positive features:	
infinitives	.76
prediction modals	.54
suasive verbs	.49
conditional subordination	.47
necessity modals	.46
split auxiliaries	.44
(possibility modals	.37)
[No negative features]	
DIMENSION 5: Abstract versus Non-abstract Style	
Positive features:	
conjuncts	−.48
agentless passives	−.43
past participial adverbial clauses	−.42
BY-passives	−.41
past participial postnominal clauses	−.40
other adverbial subordinators	−.39
[No negative features]	

larly show an extensive range of variation along Dimension 1. Expository informational registers, like official documents and academic prose, have very large negative scores; the fiction registers have scores around 0.0; while personal letters have a relatively large positive score.

This distribution shows that no single register can be taken as representative of the spoken or written mode. At the extremes, written informational prose is dramatically different from spoken conversation with respect to Dimension 1 scores. But written personal letters are relatively similar to spoken conversation, while spoken prepared speeches share some Dimension 1 characteristics with written fictional registers. Taken together, these Dimension 1 patterns indicate that there is extensive overlap between the spoken and written modes in these linguistic characteristics, while the extremes of each mode (i. e., conversation versus informational prose) are sharply distinguished from one another.

The overall comparison of speech and writing resulting from the 1988 MD analysis is actually even more complex, because six separate dimensions of variation were identified, and each of these defines a different set of relations among spoken and written registers. Table 38.3 displays the features and their loadings for Dimensions 2−5. (Dimension 6 has few salient linguistic features and is not considered here; see Biber 1988, 113−114, 154−160.) The name of each factor summarizes its interpretation. Dimension 1 − Involved vs. Informational Production − has been described above; the other dimensions are described and exemplified below.

5.1. Dimension 2: Narrative vs. Non-narrative Concerns

Dimension 2 is entitled Narrative vs. Non-narrative Concerns. The features with positive weights − past tense verbs, third-person pronouns, perfect aspect verbs, public verbs, synthetic negation and present participial clauses − are associated with past time narration. Past tense and perfect aspect verbs are used to describe past events, while the third-person pronouns refer to participants in the events. Public verbs (e. g., *say*, *tell*, *declare*) are used to express communication acts. Present participial clauses are typically used to add description and imagery to the narration. No features have strong negative loadings on this dimension (compared to their loadings on other dimensions); therefore, the dimension is a continuum reflecting the use of narrative features versus absence of those features.

Text sample 3 from romance fiction illustrates many of the features associated with Narrative Concerns. Particularly noticeable in this extract are the past tense verbs, third person pronouns (*he* and *his*), public verbs (particularly *said*), and the present participial clause which adds a descriptive detail to the action (*waving the manager away*).

> *Text sample 3: Romance fiction*
>
> But Mike Deegan was boiling mad now. When the inning was over he cursed the Anniston catcher all the way into the dugout ...
>
> The Anniston manager came right up to the dugout in front of Mike. His face was flushed. ‘Deegan,’ the manager said, his voice pitched low, quivering.
>
> ‘That was a rotten thing to do.’
>
> ‘For God’s sake,’ Mike said, waving the manager away, ‘Stop it, will you? Tell your guys not to block the plate!’

The distribution of registers along Dimension 2, shown in Figure 38.2, further supports its interpretation as Narrative vs. Non-narrative Concerns. All types of fiction have

```
NARRATIVE
    | Romance fiction
  7 +
    |
    |
    |
  6 + Mystery, Science, and General fiction
    |
    | Adventure fiction
    |
  5 +
    |
    |
    |
  4 +
    |
    |
    |
  3 +
    |
    |
    |
  2 + Biographies
    |
    | SPONTANEOUS SPEECHES
    |
  1 + Humor
    | PREPARED SPEECHES
    | Press reportage
    | Personal letters
  0 + Popular lore
    |
    | FACE-TO-FACE CONVERSATIONS
    | Religion; Editorials
 −1 + PUBLIC CONVERSATIONS
    |
    | Press reviews
    |
 −2 + TELEPHONE CONVERSATIONS
    | Professional letters
    | Academic prose
    | Official documents
 −3 + Hobbies
    | BROADCASTS
    |
NON-NARRATIVE
```

Fig. 38.2: Mean scores for registers along Dimension 2: Narrative versus Non-narrative Discourse. (F = 32.3, p < .0001, r^2 = 60.8 %)

markedly high positive scores, reflecting their emphasis on narrating events. In contrast, registers which are typically more concerned with events currently in progress (e. g., broadcasts) or with building arguments rather than narrating (e. g., academic prose) have negative scores on this dimension. Finally, some registers have scores around 0.0, reflecting a mix of narrative and other features. For example, face-to-face conversation will often switch back and forth between narration of past events and discussion of current interactions.

5.2. Dimension 3: Elaborated vs. Situation-dependent Reference

Dimension 3 is labeled 'Elaborated vs. Situation-dependent Reference'. The majority of positive features on this dimension are relative clause constructions − *WH*-relative clauses on object position, *WH*-relative clauses on subject position, and 'pied piping' constructions. These features explicitly identify referents or provide elaboration about referents.

In contrast, the negative features on this dimension are commonly used to refer to places and times outside of the text itself, in either the real world or an imaginary world created by the text. Place and time adverbials are used for temporal and locative reference (e. g., *earlier*, *soon*; *there*, *behind*). The other adverbs can have a wider range of functions, such as descriptions of manner.

Dimension 3 thus represents a continuum between texts that have elaborated, explicit reference, versus reference that is more dependent on the situational context. Figure 38.3 displays the distribution of registers along Dimension 3. Those with large positive scores − official documents, professional letters, academic prose, and press reviews − frequently use *WH*-relative clauses, along with phrasal coordinators and nominalizations (and a lack of time and place adverbials). Those with large negative scores − broadcasts and telephone conversations − rely more heavily on time and place adverbials and other adverbs in order to situate the discourse.

The two contrasting poles of Dimension 3 are exemplified by text samples 4 and 5. Text sample 4 is a short extract from an official document and illustrates the use of *WH*-relative clauses (*321 of whom were approved*, *230 of whom were approved*, *who were awarded* ...) to elaborate noun referents. Phrasal coordination is also used.

Text sample 4: Official document

During the past year 347 candidates were examined by the Surgical Section, 321 of whom were approved, and 352 were examined by the Dental Section, 230 of whom were approved, making a total of 230 candidates who were awarded the Licence in Dental Surgery.

Text sample 5 comes from a radio broadcast of a soccer match. In contrast to the official document, time adverbials (*now*) and place adverbials (e. g., *just below us, here, forward*) are used to refer directly to the physical situation of the broadcast.

Text sample 5: Sports broadcast

and from the foot of Hemsley − the ball into touch − just below us here [...] a strike forward − but of course now turned − by manager O'Farrell [...] quickly taken by Brian Kydd − Kydd now to number seven

```
ELABORATED
    | Official documents
  7 +
    | Professional letters
    |
  6 +
    |
    |
  5 +
    |
    | Press reviews; Academic prose
  4 +
    | Religion
    |
  3 +
    |
    | Popular lore
  2 +
    | Editorials; Biographies
    | SPONTANEOUS SPEECHES
  1 +
    |
    | PREPARED SPEECHES; Hobbies
  0 +
    | Press reportage; INTERVIEWS
    | Humor
  1 +
    | Science fiction
    |
 −2 +
    |
    |
 −3 +
    | General fiction
    | Personal letters; Mystery & Adventure fiction
 −4 + FACE-TO-FACE CONVERSATIONS; Romance fiction
    |
    |
 −5 +
    | TELEPHONE CONVERSATIONS
//
 −9 +
    |
    | BROADCASTS
SITUATION-DEPENDENT
```

Fig. 38.3: Mean scores for registers along Dimension 3: Situation-Dependent versus Elaborated Reference. ($F = 31.9$, $p < .0001$, $r^2 = 60.5\%$)

5.3. Dimension 4: Overt Expression of Persuasion

Like Dimension 2, Dimension 4 has only features with positive weights: infinitives, prediction modals (e. g., *will*, *would*), suasive verbs (e. g., *agree*, *ask*, *insist*, *recommend*), conditional subordination, necessity modals (e. g., *ought*, *should*), split auxiliaries, and possibility modals (e. g., *might*, *may*).

```
OVERTLY ARGUMENTATIVE
     |
     | Professional letters
     |
   3 + Editorials
     |
     |
     |
   2 + Romance fiction
     | Hobbies
     | Personal letters
     |
   1 + PUBLIC CONVERSATIONS; General fiction
     |
     | PREPARED SPEECHES; TELEPHONE CONVERSATIONS
     | SPONTANEOUS SPEECHES; Religion
   0 + Official documents
     | FACE-TO-FACE CONVERSATIONS; Popular lore; Humor
     | Academic prose
     | Biographies; Mystery & Science fiction; Press reportage
  −1 +
     | Adventure fiction
     |
     |
  −2 +
     |
     |
     | Press reviews
  −3 +
     |
     |
     |
  −4 +
     |
     | BROADCASTS
     |
NOT OVERTLY ARGUMENTATIVE
```

Fig. 38.4: Mean scores for registers along Dimension 4: Overt Expression of Argumentation. (F = 4.2, $p < .0001$, $r^2 = 16.9\%$)

This dimension has been interpreted as reflecting overt persuasion or argumentation, as exemplified in Text sample 6 from a professional letter:

> *Text sample 6: Professional letter*
>
> Furthermore, it really would be inappropriate for me to put words in your mouth. In short, you should really take the format of the resolution and put in your own thoughts [...] The association is already sampling opinion on a number of other matters and it may be possible to add this one. If it is not possible to add your concern this year, it would certainly be possible to add it next year.

Typical of texts with a large positive score on Dimension 4, this professional letter uses prediction modals to show what will be possible in the future (*it would be possible to add it next year*) or to discuss hypothetical situations (*it really would be inappropriate ...*).

Necessity modals express obligation for the addressee (*you should really* ...), and possibility modals convey the likelihood of certain events (*it may be possible* ...). Conditional subordination sets limits on the circumstances under which other actions or events may occur (*If it is not possible to add your concern this year* ...). The sample also illustrates the common use of infinitives as complements controlled by adjectives that encode the writer's attitude or stance (*inappropriate to put words in your mouth*, *possible to add this one*). Taken together, these features function to structure an argument, identify alternatives, present the author's stance about ideas, and directly encourage certain thinking or action on the part of others.

The distribution of registers along this dimension (Figure 38.4) shows that professional letters and editorials have a high frequency of these features, while press reviews and broadcasts have a relative absence of these features. Many registers are unmarked for this dimension, and thus cluster around 0 in Figure 38.4.

In MD work subsequent to 1988, Dimension 4 has been referred to both as 'Overt Expression of Persuasion' and 'Overt Expression of Argumentation'. Either 'persuasion' or 'argumentation' can characterize the use of these features.

5.4. Dimension 5: Abstract vs. Non-abstract Style

Dimension 5, like Dimensions 2 and 4, has only features with positive loadings. These features include conjuncts (e. g., *thus*, *however*), agentless passives, passives with *by*-phrases, past participial (passive) adverbial clauses, and past participial (passive) postnominal clauses (also called past participial WHIZ deletions). Most of these structures are passives, and are used to present information with little or no emphasis on the agent, as in this extract from an engineering report:

> *Text sample 7: Engineering report*
>
> Eventually however fatigue cracks were noticed in the roots of two of the blades and it was suspected that the lack of freedom in the drag hinges was the possible cause.
>
> Later, after new blades had been fitted, it was thought better to run with drag hinges free and so reduce root stresses, experience having shown that the possibility of resonance was small [...] This question of blade fatigue is more fully discussed in the appendix.

This short extract contains many passive constructions. Agents of the actions are not mentioned; instead, inanimate referents are the focus of the discourse (e. g., *fatigue cracks were noticed*, *the question of blade fatigue is more fully discussed*). Two sentences use non-referential *it* as subject (*it was suspected*, *it was thought*), further eliminating mention of the animate agent. In other texts of this type, noun phrases are also often modified with past participial passive modifiers (e. g., *the exhaust air volume required by the 6-ft. x 4-ft. grid*).

The distribution of registers along this dimension (Figure 38.5) shows that academic prose and official documents are particularly marked in their use of these features. Thus, the register distribution reinforces the interpretation that this style of discourse is typically used with abstract or technical information. Conjuncts and adverbial subordinators co-occur with the passive forms to mark the logical relationships among clauses.

```
ABSTRACT
  6 +
    |
    | Academic prose
    |
  5 +
    | Official documents
    |
    |
  4 +
    |
    |
    |
  3 +
    |
    |
    |
  2 +
    |
    | Religion
    | Hobbies
  1 +
    | Press reviews
    | Press reportage
    | Professional letters; Editorials
  0 + Popular lore
    |
    | Humor; Biographies
    |
 −1 +
    |
    |
    | BROADCASTS
 −2 + PREPARED SPEECHES; INTERVIEWS
    |
    | General, Science, and Adventure fiction; SPONTANEOUS SPEECHES
    | Mystery fiction; Personal letters
 −3 +
    | Romance fiction; FACE-TO-FACE CONVERSATIONS
    |
    | TELEPHONE CONVERSATIONS
 −4 +
    |
NON-ABSTRACT
```

Fig. 38.5: Mean scores for registers along Dimension 5: Abstract versus Non-abstract Style. (F = 28.8, $p < .0001$, $r^2 = 58.0\%$)

In contrast, conversation and fiction have large negative scores, indicating an absence of these features. As text samples 2 and 3 above illustrate, the subjects of sentences in conversation and fiction are often actors, and passive constructions tend to be rare. Thus, this dimension marks a continuum of impersonal, abstract style versus a more personal, non-abstract style. (In MD studies subsequent to 1988, this dimension has also been referred to as ‘Impersonal vs. Non-impersonal Style’).

5.5. Overall patterns of register variation in the 1988 MD analysis

The 1988 MD analysis showed that English registers vary along several underlying dimensions associated with different functional considerations, including: interactiveness, involvement and personal stance, production circumstances, informational density, informational elaboration, narrative purposes, situated reference, persuasiveness or argumentation, and impersonal presentation of information.

Two of these dimensions have no systematic relationship to speech and writing (Dimension 2: Narrative Discourse; and Dimension 4: Argumentation). However, the other three dimensions identify sharp distinctions between 'oral' and 'literate' registers, where the term 'oral' is used to refer to stereotypical speech (i. e. conversation), and the term 'literate' is used to refer to stereotypical writing (i. e. academic prose or other kinds of formal, informational prose). On Dimension 1, conversation is at one extreme, marked as extremely involved and restricted by real time production circumstances; academic prose is at the other extreme, marked as extremely informational and carefully crafted and edited. On Dimension 3, conversation is at one extreme, marked as extremely situated in reference; academic prose is at the other extreme, marked as extremely elaborated in reference. On Dimension 5, conversation is at one extreme, marked by the absence of passive constructions; academic prose is at the other extreme, marked by impersonal styles of presentation.

Thus, the spoken and written modes can be exploited in extreme ways, resulting in register characterizations not found in the other mode. There are genuine differences in the production circumstances of speech and writing, and these differences provide the potential for styles of expression in writing that are not (normally) feasible in speech. In addition, spoken registers rarely adopt the extreme informational communicative focus of written expository registers. (Even classroom teaching is much more interactive and involved in purpose than typical written expository registers; see Biber et al. 2002). Thus, written academic prose and official documents are extremely 'informational' (Dimension 1), 'elaborated in reference' (Dimension 3), and 'impersonal' (Dimension 5) − extreme linguistic characterizations not found in any spoken register.

However, despite the existence of these oral/literate dimensions, no dimension identifies an absolute distinction between speech and writing. On Dimension 1, written registers can be 'involved' (e. g., personal letters), while spoken registers can be moderately informational (e. g., prepared speeches). And on Dimensions 3 and 5, written registers like fiction and personal letters are similar to conversation in being 'situated' and not 'impersonal'.

5.6. Validation of the 1988 MD analysis

Several studies have attempted to replicate the factor analysis from the 1988 MD study, assessing the stability of that analysis and the validity of the resulting factors. Biber (1990) compares the factor structure of analyses run on various sub-corpora, finding that the factor structure is generally stable as long as the same range of variation is maintained in the target corpus. Biber (1992) uses confirmatory factor analysis to compare the goodness-of-fit for several different factorial models; de Mönnink/Brom/Oostdijk (2003) attempt to replicate the 1988 MD structure on the texts in the ICE-GB corpus; and Lee (2000) analyzes a 4-million-word sample from the BNC, testing the influence of statistical parameters (e. g., factor extraction methods and rotation methods) and alterations to the input data (the linguistic variables and the design parameters of the corpus) on the resulting dimensions of variation.

6. Types of MD study

In practice, there are two different types of MD study: those that carry out a full MD analysis and those that apply the 1988 dimensions to new areas of research. Methodologically, the two types differ in whether or not they include a new factor analysis (steps 6 and 7 in section 4 above). Full MD studies, such as the original MD studies (Biber 1986, 1988), carry out a factor analysis to identify and interpret underlying dimensions of register variation; they thus cover all eight methodological steps. Over the years, there have been several subsequent studies that have undertaken full MD analyses of this type, identifying the underlying dimensions that operate in particular discourse domains of English and other languages (see section 8 below). However, many MD studies apply the dimensions identified in Biber (1988) to describe and compare additional registers. These studies omit steps 6 and 7; since they use the previously-identified dimensions, such studies do not require a separate factor analysis (see section 7 below).

The decision to conduct a new, complete MD analysis or to apply the 1988 dimensions depends on the research issues that are being investigated, because the two approaches will give different perspectives on register variation. Using the established dimensions allows researchers to compare new registers or specialized sub-registers to a wide range of spoken and written registers in English (i. e., the basis of the 1988 study). Other research, however, seeks to explore a particular domain and determine the dimensions of variation for that domain. Section 7 below surveys previous studies that have applied the 1988 dimensions to a range of research questions, while section 8 surveys studies that have undertaken full MD analyses. Then, section 9 briefly surveys studies that have undertaken MD analyses of languages other than English.

7. Application of the 1988 dimensions to other discourse domains

7.1. Specialized registers and discourse domains

Many studies have applied the 1988 dimensions of variation to study the linguistic characteristics of more specialized registers and discourse domains. Biber et al. (2002) describe the patterns of variation among a range of spoken and written university registers, including academic registers (e. g., lectures, office hours, study groups, textbooks, research articles) as well as non-academic articles (e. g., service encounters, course syllabi, institutional writing). However, other MD studies have usually focused on variation among written registers. For example, Biber (1987) compares the MD characteristics of matched written registers in British English and American English, showing how the AmE registers tend to be both more ‘involved’ and less ‘elaborated’ than the corresponding BrE registers. Many MD studies have focused on academic writing. For example, Conrad (1996, 2001) compared research articles, textbooks, and student writing in two academic disciplines: biology and history. Biber/Finegan (1994b) document the systematic patterns of variation among the sections of medical research articles (i. e., introduction, methods, results, and discussion). MD studies have also been used to investigate more specialized written registers. For example, Connor/Upton (2003) describe the characteristics of direct mail letters, while Connor/Upton (2004) investigate non-profit grant proposals, comparing the different 'move' types used in proposals (e. g., 'benefits' versus 'territory'). Two other studies have focused on author style: Connor-Linton (2001) compares the styles of authors writing about the possibility of nuclear war, while Biber/Finegan (1994a) compares the styles of 18th century authors writing essays and fiction.

Several other MD studies have focused on registers in earlier historical periods, tracking the patterns of change across time. For example, Biber/Finegan (1989, 1997) document the patterns of change for a range of written and speech-based registers, documenting a general 'drift' for popular written registers to become more 'oral', while specialist written registers have evolved to become more 'literate' (with respect to Dimensions 1, 3, and 5). Atkinson (1992, 1996, 1999) provides detailed descriptions of historical change for two written registers: medical research articles and scientific research articles. In these studies, Atkinson tracks the patterns of change with respect to the 1988 dimensions, and then provides detailed interpretation of those patterns relative to the socio-historical contexts of the target registers. Finally, Geisler (2002, 2003) focuses on the patterns of variation among 19th century written registers.

Fewer MD studies have focused exclusively on spoken registers. Helt (2001) compares the characteristics of BrE and AmE conversational registers, identifying parallel differences to those found for written registers (in Biber 1987; e. g., with the AmE registers being consistently more 'involved' than the corresponding BrE registers). Connor-Linton/Shohamy (2001) study variation with oral proficiency interviews (used for ESL assessment), showing how the language produced by students for different elicitation tasks varies systematically across dimensions. Csomay (2002) focuses specifically on academic lectures, comparing their multi-dimensional profiles across levels of instruction and interactivity.

Three MD studies have focused on dramatic dialogue. Quaglio (2004) describes how the conversations in the TV show *Friends* are very similar in their MD profiles to natural face-to-face conversations. The other two studies have added the social parameter of gender, showing how men and women talk in systematically different ways in dramatic dialogue. Rey (2001) tracks general changes in the dialogues of *Star Trek* episodes from 1966−1993, focusing especially on how women and men have been portrayed differently over time. For example, women in *Star Trek* used extremely 'involved' (Dimension 1) speech styles in the 1960's but shifted to a more moderate 'involved' style in the most recent shows. In contrast, men used only moderately involved speech styles in the earliest shows, but they actually shifted to become slightly more involved than women in the most recent shows. Biber/Burges (2000) discuss similar kinds of patterns on a larger historical scale, considering the language of women and men as portrayed by female and male authors, in dramatic discourse over the past three centuries. This study found that the gender of the addressee was an important consideration, with both men and women using more 'involved' styles when speaking to a woman than when speaking to a man. Female and male authors use similarly involved styles for female speakers, but they differ in their portrayals of male speakers: modern female authors tend to portray male speakers using 'involved' styles (especially when speaking to a woman), while male authors tend to portray male speakers using less involved styles.

7.2. Automatic identification of register and 'text type' studies

Several studies use multi-dimensional approaches to automatically determine the register or genre category of unknown texts. Biber (1993b) uses the 1988 dimensions as predictors in a discriminant analysis, while studies like Karlgren/Cutting (1994), Beaudouin et al. (2001), and Folch et al. (2000) adopt similar approaches testing the predictive power of parameters that incorporate only linguistic features that are computationally easy to identify (see also Kessler/Nunberg/Schütze 1997; Louwerse et al. 2004). de Mönnink/Brom/Oostdijk (2003) compare the predictive power of factors based on the Biber (1988) feature set with factors derived from a larger set of word class tags. In addition, several studies have used multi-dimensional methodologies for studies of stylometry, to investigate authorship attribution (see article 50) or to automatically identify other demographic characteristics of authors (e. g., Palander-Collin 1999; Koppel/Argamon/Shimoni 2002).

A complementary perspective on textual variation is to identify and interpret the text categories that are *linguistically* well defined, referred to as *text types*. Text type distinctions have no necessary relation to register distinctions. Rather, text types are defined such that the texts within each type are maximally similar in their linguistic characteristics, regardless of their situational/register characteristics. However, because linguistic features have strong functional associations, text types can be interpreted in functional terms.

In the MD approach, text types are identified quantitatively using cluster analysis, with the dimensions of variation as predictors. Cluster analysis groups texts into 'clusters' on the basis of shared multi-dimensional (linguistic) characteristics: the conversations grouped in a cluster are maximally similar linguistically, while the different clusters are maximally distinguished.

Biber (1989, 1995) describes the text types in a general corpus of spoken and written English texts. Text types and registers represent complementary ways of dissecting the textual space of a language. Text types and registers are similar in that both can be described in linguistic and in situational/functional terms. However, the two constructs differ in their primary bases: registers are defined in terms of their situational characteristics, while text types are defined linguistically. Thus, a single text type can include texts from several different registers. For example, the 'involved persuasion' text type in the 1989 study is defined primarily by extremely large positive scores on Dimension 4 ('Overt Expression of Persuasion'); all texts grouped into this text type share these linguistic characteristics, even though they come from 14 different registers (including interviews, spontaneous speeches, academic prose, professional letters, and personal letters). Similarly, texts from a single register can be distributed across multiple text types. For example, academic prose texts are distributed across four text types in the 1989 study: 'scientific exposition', 'learned exposition', 'general reported exposition', and 'involved persuasion'. This analytical approach has also been used to identify text types in more restricted discourse domains (see section 8 below) and in other languages (Somali; see section 9 below).

8. Other MD analyses of English registers

As noted in 6 above, several studies have undertaken new MD analyses to determine the dimensions of variation operating in a particular discourse domain (i. e., including Steps 6 and 7 listed in section 3 above). An early study of this type is Grabe (1987), who investigated the dimensions of variation for written expository registers. Connor-Linton (1989) identified dimensions of variation in an extremely restricted discourse domain, comparing the conversational styles of Soviet and American participants in 'Space-bridge' interactions on TV talk shows. Meurman-Solin (1993) identified dimensions of variation in a corpus of early Scottish prose texts (1450−1700). White (1994) investigated the language of job interviews, comparing the language of interviewers and interviewees, while Reppen (2001) compared the MD characteristics of elementary school spoken and written registers. de Mönnink/Brom/Oostdijk (2003) analyze the dimensions in the ICE-GB corpus, comparing the multi-dimensional models derived from three different feature sets (the Biber 1988 set of 67 features; a set of 129 word class tags; and a set of 103 sentence structures).

Biber (1992) adopted a somewhat different approach, using confirmatory factor analysis to compare the 'goodness-of-fit' of several different multi-dimensional models for discourse complexity features in English. More recently, Biber (2001) identifies six dimensions of variation for 18th century written and speech-based registers, while Biber (2006) identifies the dimensions operating among university spoken and written registers.

Other recent studies carry out a complete MD analysis coupled with text type analysis (see section 7.2. above). For example, Biber (to appear) uses factor analysis to identify the dimensions of variation operating in a corpus of conversational texts, and then uses cluster analysis to identify the conversation text types that are well-defined in terms of those dimensions. Other studies go a step further: they first segment texts into topically coherent discourse segments. These discourse segments are then used as the 'texts' in a factor analysis and cluster analysis, to identify the discourse unit types that are well defined linguistically (see Csomay 2004; Biber/Jones 2005; Biber/Connor/Upton 2007).

It is interesting to compare the kinds of dimensions identified in these studies. Given that each of these studies is based on a different corpus of texts, representing a different discourse domain, it is reasonable to expect that they would each identify a unique set of dimensions. This expectation is reinforced by the fact that the more recent studies have included additional linguistic features not used in earlier MD studies (e. g., semantic classes of nouns and verbs). However, despite these differences in design and research focus, there are certain striking similarities in the set of dimensions identified by these studies.

Most importantly, in nearly all of these studies, the first dimension identified by the factor analysis is associated with an informational focus versus a personal focus (personal involvement/stance, interactivity, and/or real time production features). Table 38.4 summarizes the major features that define Dimension 1 in each of these studies.

It is perhaps not surprising that Dimension 1 in the original 1988 MD analysis was strongly associated with an informational versus (inter)personal focus, given that the corpus in that study ranged from spoken conversational texts to written expository texts. For the same reason, it is somewhat predictable that a similar dimension would have emerged from the 2001 study of 18th century written and speech-based registers, and the 2006 study of university spoken and written registers (although the corpus studied

Tab. 38.4: Comparison of Dimension 1 across multi-dimensional studies of English subsequent to Biber (1988)

Study	Corpus	Linguistic features defining the dimension
White 1994	job interviews	long words, nouns, nominalizations, prepositions, WH questions, 2nd person pronouns versus 1st person pronouns, contractions, adverbs, discourse particles, emphatics, etc.
Reppen 1994	elementary school registers	nouns, long words, nominalizations, passives, attributive adjs., prepositions versus initial *and*, time adverbials, 3rd person pronouns
Biber 2001	18th c. written and speech-based registers	prepositions, passives, nouns, long words, past tense verbs versus 1st and 2nd person pronouns, present tense, possibility and prediction modals, *that*-deletion, mental verbs, emphatics
Biber 2006	university spoken and written registers	nominalizations, long words, nouns, prepositions, abstract nouns, attributive adjectives, passives, stance noun + *to*-clause, etc. versus contractions, demonstrative pronouns, *it*, 1st person pronouns, present tense, time advs, *that*-omission, WH-questions, etc.
Biber, to appear	conversations	long words, nominalizations, prepositions, abstract nouns, relative clauses, attributive adjs. Versus contractions, 1st and 2nd person pronouns, activity verbs

for each of those two studies was more specialized than the general corpus in the 1988 study). However, it was completely unexpected that a similar oral/literate dimension − realized by essentially the same set of co-occurring linguistic features − would be fundamentally important in highly restricted discourse domains, including studies of job interviews, elementary school registers, and conversations.

A second parameter found in most MD analyses corresponds to narrative discourse, reflected by the co-occurrence of features like past tense, 3rd person pronouns, perfect aspect, and communication verbs (see, e. g., the Biber (2006) study of university registers; Biber 2001 on 18th century registers; and the Biber (to appear) study of conversation text types). In some studies, a similar narrative dimension emerged with additional special characteristics. For example, in Reppen's (2001) study of elementary school registers, 'narrative' features like past tense, perfect aspect, and communication verbs co-occurred with once-occurring words and a high type/token ratio; in this corpus, history textbooks rely on a specialized and diverse vocabulary to narrate past events. In the job interview corpus (White 1994), the narrative dimension reflected a fundamental opposition between personal/specific past events and experiences (past tense verbs co-occurring with 1st person singular pronouns) versus general practice and expectations (present tense verbs co-occurring with 1st person plural pronouns).

At the same time, most of these studies have identified some dimensions that are unique to the particular discourse domain. For example, the factor analysis in Reppen (2001) identified a dimension of 'Other-directed idea justification' in elementary student registers. The features on this dimension include 2nd person pronouns, conditional clauses, and prediction modals; these features commonly co-occur in certain kinds of student writing (e. g., *If you wanted to watch TV a lot you would not get very much done*).

The factor analysis in Biber's (2006) study of university spoken and written registers identified four dimensions. Two of these are similar linguistically and functionally to dimensions found in other MD studies: Dimension 1: 'Oral vs. Literate Discourse'; and Dimension 3: 'Narrative Orientation'. However, the other two dimensions are specialized to the university discourse domain: Dimension 2 is interpreted as 'Procedural vs. Content-focused Discourse'. The co-occurring 'procedural' features include modals, causative verbs, 2nd person pronouns, and verbs of desire + *to*-clause; these features are especially common in classroom management talk, course syllabi, and other institutional writing. The complementary 'content-focused' features include rare nouns, rare adjectives, and simple occurrence verbs; these co-occurring features are typical of textbooks, and especially common in natural science textbooks. Dimension 4, interpreted as 'Academic stance', consists of features like stance adverbials (factual, attitudinal, likelihood) and stance nouns + *that*-clause; classroom teaching and classroom management talk is especially marked on this dimension.

A final example comes from Biber's (to appear) MD analysis of conversational text types, which identified a dimension of 'stance-focused versus context-focused discourse'. Stance focused conversational texts were marked by the co-occurrence of *that*-deletions, mental verbs, factual verb + *that*-clause, likelihood verb + *that*-clause, likelihood adverbs, etc. In contrast, context-focused texts had high frequencies of nouns and *WH*-questions, used to inquire about past events or future plans. The text type analysis identified different sets of conversations characterized by one or the other of these two extremes.

In sum, studies that have incorporated complete MD analyses of English registers (i. e., including a new factor analysis) have uncovered both surprising similarities and notable differences in the underlying dimensions of variation. Two parameters seem to be fundamentally important, regardless of the discourse domain: a dimension associated with informational focus versus (inter)personal focus, and a dimension associated with narrative discourse. At the same time, these MD studies have uncovered dimensions particular to the communicative functions and priorities of each different domain of use. The following section shows that similar patterns have emerged from MD studies of languages other than English.

9. MD analyses of other languages

In addition to the MD studies of English surveyed in sections 5−8 above, there have been several MD studies of other languages. A few of these have attempted to apply the 1988 dimensions for English to describe register variation in other languages (see, e. g., the study on Spanish carried out by Lux/Grabe 1991). However, most MD studies of other languages have recognized the need to analyze the range of linguistic devices actually found in the target language, and to carry out independent factor analyses to identify the ways in which features actually co-occur in that language. For example, Sáiz (1999) built a corpus of parallel English-Spanish expository texts and then carried out a separate MD analysis for each language, comparing the dimensions across languages. Two more recent studies have also reported on MD analyses of register variation in Spanish: Parodi (2005), and Biber et al. (2006).

Four non-western languages have been studied to date: Besnier's (1988) analysis of Nukulaelae Tuvaluan; Kim's (Kim/Biber 1994) analysis of Korean; Biber/Hared's (1992) analysis of Somali; and Jang's (1998) study of Taiwanese. Taken together, these studies provide the first comprehensive investigations of register variation in non-western languages.

Biber (1995) synthesizes these studies to investigate the extent to which the underlying dimensions of variation and the relations among registers are configured in similar ways across languages. These languages show striking similarities in their basic patterns of register variation, as reflected by:

- the co-occurring linguistic features that define the dimensions of variation in each language;
- the functional considerations represented by those dimensions; and
- the linguistic/functional relations among analogous registers.

For example, similar to the full MD analyses of English, these MD studies have all identified dimensions associated with informational versus (inter)personal purposes, and with narrative discourse.

At the same time, each of these MD analyses have identified dimensions that are unique to a language, reflecting the particular communicative priorities of that language and culture. For example, the MD analysis of Somali identified a dimension interpreted as 'Distanced, directive interaction', represented by optative clauses, 1st and 2nd person pronouns, directional pre-verbal particles, and other case particles. Only one register is especially marked for the frequent use of these co-occurring features in Somali: personal letters. This dimension reflects the particular communicative priorities of personal letters in Somali, which are typically interactive as well as explicitly directive.

The cross-linguistic comparisons further show that languages as diverse as English and Somali have undergone similar patterns of historical evolution following the introduction of written registers. For example, specialist written registers in both languages have evolved over time to styles with an increasingly dense use of noun phrase modification. Historical shifts in the use of dependent clauses is also surprising: in both languages, certain types of clausal embedding – especially complement clauses – turn out to be associated with spoken registers rather than written registers.

There are important possible confounding influences that must be considered when interpreting cross-linguistic MD comparisons. One consideration has to do with corpus design: do the corpora include the same range of spoken and written registers?

A second possible confounding influence for cross-linguistic comparisons is that each of these languages has a different inventory of structural devices and distinctions. The analytical goal in each case has been to include the full range of structural/functional distinctions found in the target language. However, the multi-dimensional patterns for each language reflect a complex interaction between the available structural resources and the register distinctions that are systematically marked by those resources. For example, the existence of subjunctive mood verbs in Spanish provides the linguistic resources for a dimension associated with irrealis discourse. Similarly, the existence of two past tenses in Spanish provides the structural resources for a specialized dimension associated with informational reports of past events. In the Korean MD analysis, personal stance features are grouped on one dimension, while features of honorification and self-humbling are grouped on a separate dimension.

The existence of structural distinctions does not necessarily entail the existence of systematic register differences, but previous MD analyses show that languages/cultures have often evolved to take advantage of these linguistic resources. However, these analyses have further shown that the ways in which a language/culture exploits such structural resources are not always what we would have anticipated. For example in Korean, the co-occurring features associated with the 'stance' dimension (e. g., emphatics, hedges, other epistemic and attitudinal features) are especially common in the (inter)personal registers, including all conversations and personal letters. In contrast, the features associated with the honorific/self-humbling dimension are especially common only in the *public* spoken registers, such as public interviews and public speeches. Both dimensions are generally related to the expression of stance. However, the MD analysis shows that they are exploited in different ways for specific cultural purposes.

These patterns illustrate the general finding that structural resources come to be exploited in particular (often unanticipated) ways in particular cultures. Some linguistic features are distributed widely across different languages, and they are exploited in very similar − possibly universal − ways to distinguish among registers across cultures. For example, features like 1st and 2nd person pronouns, questions, reduced/contracted forms, and simple hedging or emphatic stance features are found in many languages, and the MD analyses carried out to date indicate that these features tend to co-occur cross-linguistically associated with conversation and other (inter)personal spoken registers. Similarly, nouns, adjectives, and various kinds of nominal modifiers are found in many languages, and they tend to co-occur cross-linguistically associated with formal expository writing. In contrast, other linguistic resources are more specialized, occurring in comparatively few languages, and these resources have come to be exploited for more specialized and more distinctive dimensions of register variation.

10. Conclusion

The present article has briefly introduced the goals and methodologies of MD analysis, and surveyed the major studies undertaken with this approach. Two general patterns are especially noteworthy from these MD studies: 1) the extent to which similar dimensions of variation operate within different languages and within specific discourse domains; and 2) the specialized dimensions that are peculiar to a particular language or discourse domain. The first kind of finding relates to the possibility of universals of register variation. For example, based on prior MD studies, we could make the strong hypothesis that the texts in any language will vary systematically along at least two dimensions: one associated with an informational versus (inter)personal focus, and one associated with narrative versus non-narrative discourse. The second kind of finding relates to the distinctiveness of every language and every discourse domain, reflecting the unique communicative priorities and situational circumstances of the language/domain. These differences are systematically reflected in the dimensions of variation that exist in the language/domain, and in the relations among registers defined by each dimension. In fact, a comparison of the multi-dimensional profiles of different languages might indicate the different communicative priorities of those languages, because it reflects the way in which each one allocates linguistic resources for functional purposes (see Biber 1995, especially 264−270).

Obviously, much further research of this kind is required to confirm the existence of universal patterns of register variation, and to investigate the range of more specialized dimensions found in various languages and domains. The MD studies carried out to date, however, indicate the importance of this research approach, and the feasibility of achieving these goals through further studies of this type.

11. Literature

Atkinson, D. (1992), The Evolution of Medical Research Writing from 1735 to 1985: The Case of the Edinburgh Medical Journal. In: *Applied Linguistics* 13, 337−374.

Atkinson, D. (1996), The Philosophical Transactions of the Royal Society of London, 1675−1975: A Sociohistorical Discourse Analysis. In: *Language in Society* 25, 333−371.

Atkinson, D. (1999), *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675−1975*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Beaudouin, V./Fleury, S./Habert, B./Illouz, G./Licoppe, C./Pasquier, M. (2001), Typweb: Decrire la toile pour mieux comprendre les parcours. In: *Colloque international sur les usages et les services des télécommunications (CIUST01)*. Paris, France. Available at: http://www.cavi.univparis3.fr/ilpga/ilpga/sfleury/typweb.htm.

Besnier, N. (1988), The Linguistic Relationships of Spoken and Written Nukulaelae Registers. In: *Language* 64, 707−736.

Biber, D. (1985), Investigating Macroscopic Textual Variation through Multi-feature / Multi-dimensional Analyses. In: *Linguistics* 23, 337-360.

Biber, D. (1986), Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. In: *Language* 62, 384-414.

Biber, D. (1987), A Textual Comparison of British and American Writing. In: *American Speech* 62, 99−119.

Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1989), A Typology of English Texts. In: *Linguistics* 27, 3−43.

Biber, D. (1990), Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. In: *Literary and Linguistic Computing* 5, 257−269.

Biber, D. (1992), On the Complexity of Discourse Complexity: A Multidimensional Analysis. In: *Discourse Processes* 15, 133−163. Reprinted in: Conrad/Biber 2001, 215−240.

Biber, D. (1993a), Representativeness in Corpus Design. In: *Literary and Linguistic Computing* 8, 1−15.

Biber, D. (1993b), Using Register-diversified Corpora for General Language Studies. In: *Computational Linguistics* 19, 219−241.

Biber, D. (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D. (2001), Dimensions of Variation among 18th Century Registers. In: Diller, H.-J./Görlach, M. (eds.), *Towards a History of English as a History of Genres*. Heidelberg: C. Winter, 89−110. Reprinted in: Conrad/Biber 2001, 200−214.

Biber, D. (2006), *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D. (to appear), Corpus-based Analyses of Discourse: Dimensions of Variation in Conversation. In: Jones, R./Bhatia, V./Flowerdew J. (eds.), *Advances in Discourse Studies*. Routledge.

Biber, D./Burges, J. (2000), Historical Change in the Language Use of Women and Men: Gender Differences in Dramatic Dialogue. In: *Journal of English Linguistics* 28, 21−37. Reprinted in: Conrad/Biber 2001, 157−170.

Biber, D./Connor, U./Upton, T. (2007), *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.

Biber, D./Conrad, S./Reppen, R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D./Conrad, S./Reppen, R./Byrd, P./Helt, M. (2002), Speaking and Writing in the University: A Multi-dimensional Comparison. In: *TESOL Quarterly* 36, 9−48.

Biber, D./Conrad, S./Reppen, R./Byrd, P./Helt, M. (2003), Strengths and Goals of Multi-dimensional Analysis: A Response to Ghadessy. In: *TESOL Quarterly* 37, 151−155.

Biber, D./Davies, M./Jones, J. K./Tracy-Ventura, N. (2006), Spoken and Written Register Variation in Spanish: A Multi-dimensional Analysis. In: *Corpora* 1, 7−38.

Biber, D./Finegan, E. (1989), Drift and the Evolution of English Style: A History of Three Genres. In: *Language* 65, 487−515.

Biber, D./Finegan, E. (1994a), Multi-dimensional Analyses of Authors' Styles: Some Case Studies from the Eighteenth Century. In: Ross, D./Brink, D. (eds.), *Research in Humanities Computing*, Vol. III. Oxford: Oxford University Press, 3−17.

Biber, D./Finegan, E. (1994b), Intra-textual Variation within Medical Research Articles. In: Oostdijk, N./de Haan, P. (eds.), *Corpus-based Research into Language*. Amsterdam: Rodopi, 201−222. Reprinted in: Conrad/Biber 2001, 108−123.

Biber, D./Finegan, E. (1997), Diachronic Relations among Speech-based and Written Registers in English. In: Nevalainen, T./Kahlas-Tarkka, L. (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique, 253−276. Reprinted in: Conrad/Biber 2001, 66−83.

Biber, D./Hared, M. (1992), Dimensions of Register Variation in Somali. In: *Language Variation and Change* 4, 41−75.

Biber, D./Johansson, S./Leech, G./Conrad, S./Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Biber, D./Jones, J. K. (2005), Merging Corpus Linguistic and Discourse Analytic Research Goals: Discourse Units in Biology Research Articles. In: *Corpus Linguistics and Linguistic Theory* 1, 151−182.

Brown, P./Fraser, C. (1979), Speech as a Marker of Situation. In: Scherer, K. R./Giles, H. (eds.), *Social Markers in Speech*. Cambridge: Cambridge University Press, 33−62.

Connor, U./Upton, T. (2003), Linguistic Dimensions of Direct Mail Letters. In: Meyer, C./Leistyna, P. (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 71−86.

Connor, U./Upton, T. (2004), The Genre of Grant Proposals: A Corpus Linguistic Analysis. In: Connor, U./Upton, T. (eds.), *Discourse in the Professions*. Amsterdam: John Benjamins, 235−256.

Connor-Linton, J. (1989), Crosstalk: A Multi-feature Analysis of Soviet-American Spacebridges. PhD dissertation, University of Southern California.

Connor-Linton, J. (2001), Authors' Style and World-view: A Comparison of Texts about Nuclear Arms Policy. In: Conrad/Biber 2001, 84−93.

Connor-Linton, J./Shohamy, E. (2001), Register Variation, Oral Proficiency Sampling, and the Promise of Multi-dimensional Analysis. In: Conrad/Biber 2001, 124−137.

Conrad, S. (1996), Investigating Academic Texts with Corpus-based Techniques: An Example from Biology. In: *Linguistics and Education* 8, 299−326.

Conrad, S. (2001), Variation among Disciplinary Texts: A Comparison of Textbooks and Journal Articles in Biology and History. In: Conrad/Biber 2001, 94−107.

Conrad, S./Biber, D. (eds.) (2001), *Variation in English: Multi-dimensional Studies*. Harlow/London: Pearson Education.

Csomay, E. (2002), Variation in Academic Lectures: Interactivity and Level of Instruction. In: Reppen, R./Fitzmaurice, S. M./Biber, D. (eds.), *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins, 203−224.

Csomay, E. (2004), Linguistic Variation within University Classroom Talk: A Corpus-based Perspective. In: *Linguistics and Education* 15, 243−274.

de Mönnink, I./Brom, N./Oostdijk, N. (2003), Using the MF/MD Method for Automatic Text Classification. In: Granger, S./Petch-Tyson, S. (eds.), *Extending the Scope of Corpus-based Research: New Applications, New Challenges*. Amsterdam: Rodopi, 15−26.

Ervin-Tripp, S. (1972), On Sociolinguistic Rules: Alternation and Co-occurrence. In: Gumperz, J. J./Hymes, D. (eds.), *Directions in Sociolinguistics*. New York: Holt, 213−250.

Ferguson, C. A. (1983), Sports Announcer Talk: Syntactic Aspects of Register Variation. In: *Language in Society* 12, 153-172.

Folch, H./Heiden, S./Habert, B./Fleury, S./Illouz, G./Lafon, P./Nioche, J./Prévost, S. (2000), Typtex: Inductive Typological Text Classification by Multivariate Statistical Analysis for NLP Systems Tuning/Evaluation. In: *Proceedings of the Second Language Resources and Evaluation Conference*. Athens, Greece, 141−148.

Geisler, C. (2002), Investigating Register Variation in Nineteenth-century English: A Multi-dimensional Comparison. In: Reppen, R./Fitzmaurice, S. M./Biber, D. (eds.), *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins, 249−271.

Geisler, C. (2003), Gender-based Variation in Nineteenth-century English Letter Writing. In: Meyer, C./Leistyna, P. (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 87−106.

Gorsuch, Richard L. (1983), *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Grabe, W. (1987), Contrastive Rhetoric and Text-type Research. In: Connor, U./Kaplan, R. B. (eds.), *Writing across Languages: Analysis of L2 Text*. Reading, MA: Addison-Wesley, 115−138.

Helt, M. (2001), A Multi-dimensional Comparison of British and American Spoken English. In: Conrad/Biber 2001, 171−184.

Hymes, D. (1984), Sociolinguistics: Stability and Consolidation. In: *International Journal of the Sociology of Language* 45, 39-45.

Jang, S-C. (1998), Dimensions of Spoken and Written Taiwanese: A Corpus-based Register Study. PhD dissertation, University of Hawaii.

Karlgren, J./Cutting, D. (1994), Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: *Proceedings of Coling 94*. Kyoto, Japan, 1071−1075. Reprinted in: Karlgren, J. (2000), *Stylistic Experiments for Informational Retrieval*, chapter 7. PhD dissertation, Stockholm University.

Kessler, B./Nunberg, G./Schütze, H. (1997), Automatic Detection of Text Genre. In: *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 32−38.

Kim, Y./Biber, D. (1994), A Corpus-based Analysis of Register Variation in Korean. In: Biber, D./ Finegan, E. (eds.), *Sociolinguistic Perspectives on Register*. New York: Oxford University Press, 157−181.

Koppel, M./Argamon, S./Shimoni, A. R. (2002), Automatically Categorizing Written Texts by Author Gender. In: *Literary and Linguistic Computing* 17, 401−412.

Lee, D. (2000), Modelling Variation in Spoken and Written English: The Multi-dimensional Approach Revisited. PhD dissertation, Lancaster University.

Louwerse, M./McCarthy, P. M./McNamara, D. S./Graesser, A. (2004), Variation in Language and Cohesion across Written and Spoken Registers. In: Forbus, K./Gentner, D./Regier, T. (eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum, 843−848.

Lux, P./Grabe, W. (1991), Multivariate Approaches to Contrastive Rhetoric. In: *Lenguas Modernas* 18, 133−160.

Meurman-Solin, A. (1993), *Variation and Change in Early Scottish Prose*. (Dissertationes Humanarum Litterarum 65.) Helsinki: Suomalainen Tiedeakatemia.

Palander-Collin, M. (1999), Male and Female Styles in 17th Century Correspondence. In: *Language Variation and Change* 11, 123−141.

Parodi, G. (2005), Lingüística de corpus y análisis multidimensional: Exploración de la variación en el corpus PUCV-2003. In: Parodi, G. (ed.), *Discurso Especializado e Instituciones Formadoras*. Valparaíso, Chile: Ediciones Universitarias de Valparaíso, 83−126.

Quaglio, P. M. (2004), The Language of NBCs Friends: A Comparison with Face-to-face Conversation. PhD dissertation, Northern Arizona University.

Reppen, R. (1994), Variation in Elementary Student Language: A Multi-dimensional Perspective. Unpublished PhD dissertation, Northern Arizona University, Flagstaff, AZ.

Reppen, R. (2001), Register Variation in Student and Adult Speech and Writing. In: Conrad/Biber 2001, 187−199.

Rey, J. M. (2001), Changing Gender Roles in Popular Culture: Dialogue in the *Star Trek* episodes from 1966 to 1993. In: Conrad/Biber 2001, 138−156.

Sáiz, M. (1999), A Cross-linguistic Corpus-based Analysis of Linguistic Variation. Manchester: PhD dissertation, UMIST.

White, M. (1994), Language in Job Interviews: Differences Relating to Success and Socioeconomic Variables. PhD dissertation, Northern Arizona University.

Douglas Biber, Flagstaff, AZ (USA)

39. Machine learning

1. Introduction

In corpus linguistics the computer can play various roles as assistant to the corpus linguist, alleviating some of the linguist's tasks. This article describes methods which can automatically learn to assign linguistic annotations to digital corpora, on the basis of example annotations presented to them in a training phase. If successful, these methods can generate annotation layers automatically at superhuman speeds; alternatively, they can be integrated into the human annotation process as automatic pre-annotators that do at least part of the work, or post-annotators that search for inconsistencies in the annotations. These methods come from machine learning, a subfield of artificial intelligence.

Research in machine learning focuses on computer programs that learn from experience. Learning is a manifold concept, and is usually indirectly quantified as the measurable improvement of a learner on some task after learning has taken place (Mitchell 1997). The role of machine learning this article focuses on is that of a means to automate knowledge discovery and linguistic modeling on the basis of annotated corpus material, where the annotation is the target of learning. Annotation layers usually represent some abstraction over the text of the corpus, ranging from word-level abstractions (e. g. part-of-speech tags) to text-level abstractions (e. g. topical text categorization).