World Englishes, Vol. 28, No. 4, pp. 421–450, 2009. 0883-2919

Multidimensional analysis and the study of world Englishes

RICHARD XIAO*

ABSTRACT: The multidimensional analysis (MDA) approach was originally developed by Biber to compare written and spoken registers in English, but it has since been applied extensively in synchronic and diachronic analysis of registers in English as well as non-Western languages. While the *WordSmith*-style keyword analysis represents a quick and simple means of evaluating a genre against the MDA dimensions, the keyword approach nevertheless provides a less comprehensive contrast of genres and may not work for more fine-grained types of genre analysis. However, Biber's MDA model has so far been confined largely to grammatical categories, though Biber and others have started to incorporate semantic categories of some word classes in the model. In this paper, we will seek to enhance the MDA framework with semantic components and also introduce this enhanced MDA model, for the first time, in the research of world Englishes by exploring variation across twelve registers and five varieties of English in the International Corpus of English (ICE), which are annotated with grammatical and semantic categories.

INTRODUCTION

World Englishes refer to 'localized forms of English' used around the world (Bolton 2005: 69). They have attracted great scholarly interest over the past three decades. For example, in addition to dozens of authored and edited books in this area (e.g. Bauer 2002; Crystal 2003; Jenkins 2003; Kirkpatrick 2002, 2007; Hickey 2004; Trudgill 2004; Kachru, Kachru and Nelson 2006; Deterding 2007; Schneider 2007; Fishman 2008; Kachru and Smith 2008; Mesthrie and Bhatt 2008), a number of special issues of this journal have been devoted to varieties of English used in Hong Kong (19/3), South Africa (21/1), China (21/2), South America (22/2), the Philippines (23/1), and Russia (24/4). World Englishes have been approached broadly from socio-cultural as well as linguistic perspectives. The former approach has focused on the elaboration of theories and models of development, spread, classification and interaction of new Englishes (e.g. Kachru 1992, 2008; Yano 2001; Berns 2005; Schneider 2007; Kachru and Smith 2009; Michieka 2009; Seidlhofer 2009), while the latter is largely concerned with linguistic descriptions of world Englishes by exploring linguistic features in selected English varieties at various levels including phonetics and phonology, morphology, syntax and discourse.

For example, Gisborne (2000) studies relative clauses in Hong Kong English; Kasanga (2006) discusses requests in a South African variety of English; van Rooy (2006) describes the extension of the progressive aspect in Black South African English; Ulrike (2007) examines the influence of first language on final consonant clusters in English in Singapore and Nigeria; Bao and Hong (2006) explore register variation in Singapore English; Lim (2007) provides a comprehensive account of discourse particles in colloquial Singapore English; Mukherjee and Hoffmann (2006) describe verb-complement constructions in Indian English while Hoffmann and Mukherjee (2007) compare ditransitive verbs in Indian English and British English. In addition, there are a number of studies that compare

*Edge Hill University, Ormskirk, Lancashire, L39 4QP, United Kingdom. E-mail: Richard.Xiao@edgehill.ac.uk

patterns of language use in a range of English varieties. For example, Kachru (2003) compares expressions of definite reference in English as used in India, Nigeria, Singapore and the USA; Sand (2004) investigates the article use in contact varieties; Nelson (2006) is concerned with the 'common core' of lexis in six varieties of English; Mair (2007) explores the collocational and cultural profiles of varieties of English around the world; Collins (2009) explores the distribution patterns of a set of modals and quasi-modals in nine varieties of English; and finally, Kirkpatrick (2007) presents a book-length account of socio-cultural and historical backgrounds of world Englishes as well as their linguistic features.

Previous linguistically-oriented studies of world Englishes, such as those cited above, have typically been concerned with a few opportunely selected linguistic features separately. The present study will take a different approach, namely the multidimensional analysis (MDA) approach which looks at a large number of linguistic features in a large amount of text simultaneously. The MDA approach to register analysis was originally developed to compare written and spoken registers in English (Biber 1988). It has since been used extensively in a wide range of research areas of language variation including, for example (see Xiao and McEnery 2005: 63):

- synchronic analyses of specific registers and genres and author styles;
- diachronic studies describing the evolution of registers;
- register studies of non-Western languages and contrastive analyses;
- research of university English and materials development; and
- move analysis and study of discourse structure.

In addition, MDA has also been applied in addressing corpus design issues and the definitional issues of register/genres and text types. More recently, Biber, Conrad, Reppen, Byrd and Helt (2002) discuss the implications of MDA for teaching materials development. Two edited volumes published recently (Conrad and Biber 2001; Reppen, Fitzmaurice and Biber 2002) demonstrate the growing interest in MDA.

While the *WordSmith*-style keyword analysis (Scott 1996, 2004, 2008) represents a quick and simple means of evaluating a genre against the MDA dimensions, the keyword approach nevertheless provides a less comprehensive contrast of genres and may not work for more fine-grained types of genre analysis (Xiao and McEnery 2005). However, Biber's MDA model has so far been confined largely to grammatical categories, though Biber, Connor and Upton (2007) have started to incorporate semantic categories of some word classes in the model.

This paper aims to enhance the MDA framework with semantic components, and also to introduce this enhanced MDA model, for the first time, in the research of world Englishes. The study is based on five existing components of the International Corpus of English (ICE) annotated grammatically and semantically using the Wmatrix corpus comparison tool (Rayson 2003, 2008), which incorporates the CLAWS part-of-speech tagger and the USAS (UCREL semantic annotation system) semantic tagger. We will seek to answer the two research questions: (1) How can Biber's MDA framework be enhanced by combining grammatical and semantic analysis? (2) In what way, if any, are spoken and written registers in the five varieties of English different or similar along the dimensions established in this study? This study hopes to contribute to corpus research theoretically by enhancing the MDA model, and to the study of world Englishes methodologically as well by introducing the enhanced MDA analytical framework.

In the sections that follow, the data and the linguistic features covered will be introduced and then factor analysis that establishes the dimensions considered in this study will be discussed. The main parts of this study will use these dimensions to compare the twelve registers and five world varieties of English covered in the ICE corpora. The final section concludes the study by summarising the paper's findings and exploring directions for future research.

THE CORPORA AND LINGUISTIC FEATURES

The International Corpus of English (ICE) is specifically designed for the synchronic study of world Englishes, which aims to create twenty corpora of one million words each, with each composed of written and spoken English produced during 1990–1994 in countries or regions in which English is a first or official language. As the primary aim of the ICE is to facilitate comparative studies of English varieties worldwide, each component follows a common corpus design, comprising of five hundred 2,000-word texts sampled from a wide range of spoken (60%) and written (40%) genres, as shown in Table 1 (see Nelson 1996: 29–30).

This study is based on five corpora, namely, five ICE components for Great Britain (ICE-GB), Hong Kong (ICE-HK), India (ICE-IN), the Philippines (ICE-PH), and Singapore (ICE-SG). While uniform criteria for data collection and markup style have been applied for all ICE corpora, different levels of linguistic annotation have been undertaken for different components. For example, the ICE-GB corpus is part-of-speech tagged and syntactically parsed, but all other ICE components used in this study are raw corpora without annotation. Consequently, these corpora were further processed to allow us to undertake a factor analysis. As the first step, we removed all annotations and tags in the original version of the five corpora. The raw text corpora were then retagged using the same tool, namely, Wmatrix which provides a web interface combining the CLAWS part-of-speech tagger and the USAS semantic tagger. The results of grammatical and semantic analysis were downloaded to a local computer and programs were written in PERL (practical extraction and retrieval language) for further processing to render them searchable with the *WordSmith Tools* and our customised program scripts that extract

Table 1. Design of the ICE corpora

Code	Register	Number of samples
S1A	Spoken-dialogue-private	100
S1B	Spoken-dialogue-public	80
S2A	Spoken-monologue-unscripted	70
S2B	Spoken-monologue-scripted	50
W1A	Written-non-printed-student writing	20
W1B	Written-non-printed-letters	30
W2A	Written-printed-academic	40
W2B	Written-printed-popular	40
W2C	Written-printed-reportage	20
W2D	Written-printed-instructional	20
W2E	Written-printed-persuasive	10
W2F	Written-printed-creative	20

the frequencies of the required 141 linguistic features from each of the 2,500 corpus samples.

The Wmatrix system applies the CLAWS C7 tagset for part-of-speech tagging, which consists of 135 tags (in addition to punctuations), much more fine-grained than the C5 tagset (61 tags) applied on the British National Corpus (BNC).[1] The USAS semantic tagger applies a tagset composed of a hierarchical structure with 21 major discourse fields, which are expanded into more than 232 category labels (see Archer, Wilson and Rayson 2002 for illustrations and examples). When a corpus is tagged grammatically and semantically with such details, it is very convenient to extract linguistic features required in this study with the help of part-of-speech and semantic tags, using standardised software tools such as *WordSmith* and our own programs.

Biber (1988) used 67 functionally-related linguistic features, which are largely confined to grammatical categories. However, while grammatical features are undoubtedly of prominence in language variation study, semantic analysis is clearly also indispensable given that grammatical features are closely associated with meaning. Indeed, Biber et al. (2007) have started to incorporate semantic categories of some word classes in their new model. The present study takes advantage of the very detailed grammatical and semantic analysis produced by Wmatrix, and uses combinations of the two types of linguistic annotation where appropriate, to extract 141 linguistic features that are functionally related and relevant to language variation research. The list combines linguistic features from Biber (1988) and those in MDA studies published in recent years as well as some grammatical and semantic categories annotated by the Wmatrix system. Some examples of these linguistic features are given in the sections introducing the enhanced MDA model and exploring variation across registers and world varieties, while further illustrations with more examples of Wmatrix semantic categories can be found in Archer et al. (2002).[2] Examples of more traditional grammatical features are also available in Biber (1988: 211–245).

A. Nouns

1. Nominalisation (nouns with suffixes such as *-tion*, *-ment*, *-ness* in singular and plural forms), 2. other nouns.

Semantic categories of nouns: 3. common nouns, 4. locative nouns, 5. numeral nouns, 6. temporal nouns, 7. nouns of evaluation, 8. nouns of classification, 9. nouns of comparison, 10. animate nouns, 11. nouns for people, 12. nouns for group/affiliation, 13. nouns for substance/material, 14. nouns for objects, 15. general/abstract terms, 16. nouns of communications, 17. nouns of speech acts, 18. nouns of social actions/states/processes, 19. proper nouns, 20. nouns of mental objects, 21. technical terms.

B. Verbs

22. DO as pro-verb, 23. BE as main verb, 24. existential structure.

Tense and aspect markers: 25. past tense verbs, 26. non-past tense verbs, 27. perfect aspect verbs.

Passives: 28. agentless passives, 29. *by*-passives.

Modals: 30. possibility/permission/ability modals, 31. necessity/obligation modals, 32. predictive/volitional modals.

Semantic categories of verbs: 33. verbs for modification/change, 34. causative verbs, 35. verbs for comparison, 36. general/abstract verbs relating to being/existing, 37. verbs of classification, 38. verbs of evaluation, 39. verbs of movement/activity, 40. verbs of communications, 41. verbs of speech acts, 42. verbs of social actions/states/processes, 43. verbs of remaining/stationary/inactivity, 44. cognitive verbs, 45. sensory verbs, 46. suasive verbs, 47. public verbs, 48. private verbs.

C. Pronouns

49. first person pronouns, 50. second person pronouns, 51. third person pronouns, 52. pronoun IT, 53. possessive pronouns, 54. nominal possessive pronouns, 55. reflexive pronouns, 56. indefinite pronouns, 57. demonstrative pronouns (*this, that, these* or *those* followed by a noun), 58. demonstratives (*this, that, these* or *those* used alone without a noun).

D. Adjectives

59. attributive adjectives, 60. predicative adjectives.

Semantic categories of adjectives: 61. adjectives of evaluation, 62. adjectives of comparison, 63. adjectives of importance, 64. adjectives of ease/difficulty, 65. adjectives of general appearance/physical attributes, 66. adjectives of judgement/appearance, 67. adjectives of colour, 68. adjectives of shape, 69. adjectives of texture.

E. Adverbs

70. general adverbs, 71. time adverbs, 72. place adverbs, 73. degree adverbs, 74. prepositional adverbs/particles, 75. adverbs introducing appositions, 76. exclusiviser/particulariser adverbs.

F. Prepositions

77. prepositions, 78. final prepositions.

G. Subordination

79. causative subordinator (*because*), 80. conditional subordinators, 81. other subordinators.

H. Co-ordination

82. phrasal co-ordination, 83. non-phrasal co-ordination.

I. WH-questions and clauses

84. WH-questions, 85. WH-nominal clauses.

J. Nominal post-modifying clauses

86. sentence relatives, 87. THAT relative clauses, 88. WH relative clauses, 89. pied piping constructions, 90. past participial WHIZ deletion.

K. THAT complement clauses

91. THAT clauses as verb complements, 92. THAT clauses as noun complements, 93. THAT clauses as adjective complements.

L. Infinitive clauses

94. infinitive clauses, 95. infinitive clauses as verb complements, 96. infinitive clauses as adjective complements.

M. Participial clauses

97. past participial clauses, 98. present participial clauses.

N. Reduced forms and dispreferred structures

99. contractions, 100. split auxiliaries, 101. split infinitives, 102. THAT deletion.

O. Lexical and structural complexity

103. standardised type-token ratio, 104. mean word length, 105. mean sentence length.

P. Quantifiers

106. numeral, 107. measurement, 108. quantity, 109. frequency/recurrence rate.

Q. Time

110. past, 111. present/simultaneous, 112. future, 113. momentary, 114. period, 115. beginning, 116. ending, 117. old/mature, 118. new/young, 119. early, 120. late.

R. Degrees

121. intensifiers for non-specific degree, 122. maximisers, 123. boosters, 124. approximators, 125. compromisers, 126. diminishers, 127. minimisers, 128. comparative/superlative degrees.

S. Negation

129. analytic negation, 130. synthetic negation.

T. Power relationship

131. power/organising, 132. respect, 133. competition, 134. permission.

U. Definite

135. positive, 136. negative.

V. Helping/hindrance

137. helping, 138. hindrance

X. Linear order

139. expressions relating to linear movement, order, sequencing, etc.

Y. Seem/appear

140. expressions relating to impression/appearance.

Z. Discourse bin

141. discourse markers and emphatic communication terms.

The frequencies of all of these linguistic features were extracted from our corpora. As a frequency profile of these features is required for each of the 2,500 samples, a series of PERL scripts were designed to extract a number of features at one go. But statistics for some linguistic features (e.g. standardised type-token ratio, average word and sentence length) were readily available using the Wordlist function of the *WordSmith Tools*. As our corpus files may vary in size – though they are all around 2,000 words – the raw frequencies were normalised to a common base of 1,000 words. The profiles of normalised frequencies, together with information pertaining to file ID, register and world variety were saved as an Excel spreadsheet for use in factor analysis.

THE ENHANCED MDA MODEL

The key to the MDA approach is factor analysis, which is used extensively in social sciences to identify clusters of variables, thus, reducing a large number of variables to a manageable set of underlying factors or dimensions. Factor analysis is a common data reduction method available in many standard statistics packages such as the Statistical Package for the Social Sciences (SPSS). In this study, the SPSS (Release 14.0) was used in the factor analysis to establish the dimensions or factorial structures underlying the 141 linguistic features.

In this study, Principal Axis Factoring in SPSS is used to extract factors, which is essentially the same extraction method used in Biber (1988). In SPSS the default setting is to retain all factors with eigenvalues greater than 1.0. However, many of the factors extracted in this way had loadings with weights less than 0.30, the minimum loading weight considered significant (cf. Costello and Osborne 2005: 4), which is also the cut-off point in Biber (1988). There were other problems associated with over extraction, for example, too many cross loadings, and weak and unstable factors because of few significant loadings. In this context, the 'best choice' recommended by Costello and Osborne (2005: 3) is the scree plot, which displays eigenvalues graphically, as shown in Figure 1.

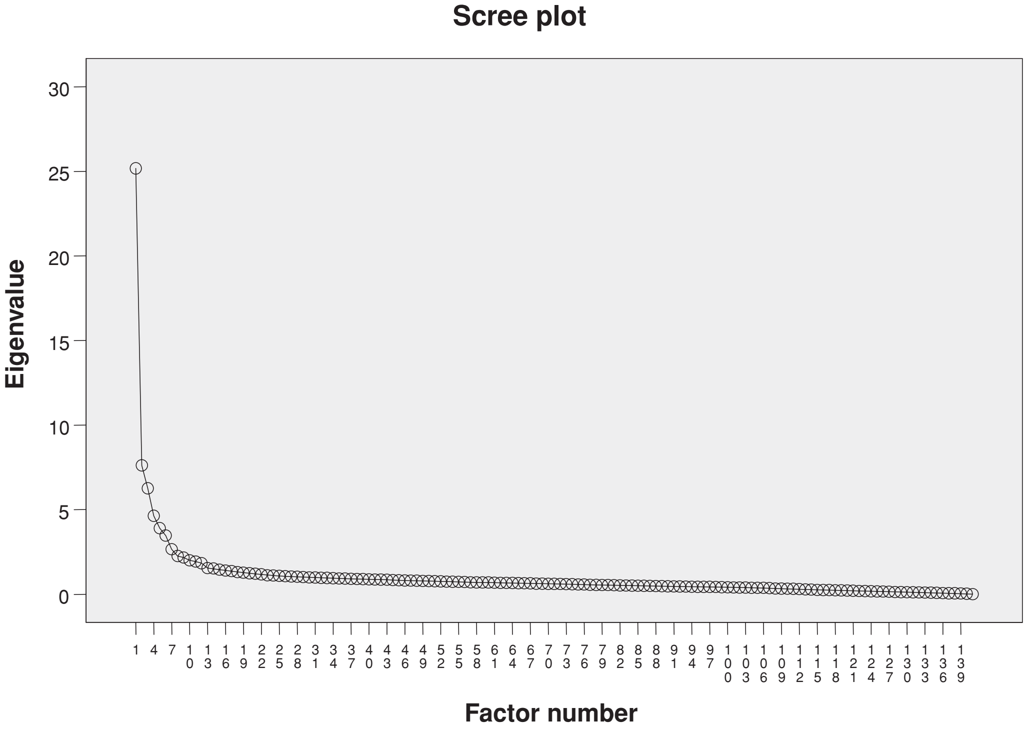

Figure 1. The scree plot

As can be seen in Figure 1, there are natural break points in the curve of eigenvalues. The number of data points *above* the break (i.e. not including the breaking point) is supposed to the number of factors to retain. A number of natural breaks can be noted in Figure 1, the most obvious of which are between the first eight data points, and between data points 12 and 13. While overextraction has some drawbacks as noted earlier, underextraction is also undesirable. On the one hand, there is a loss of information in underextraction as too many linguistic features will be excluded from final analysis; on the other hand, under extraction can produce a "confused picture" (Biber 1988: 88) of linguistic features when factors are collapsed, thus, making the interpretation more difficult. Biber (1988) extracted seven factors on the basis of 481 samples (*c*. 960,000 words) of spoken and written registers in British English, but the last factor has only one significant loading, which is considered weak. As a result, we decided to follow the recommendations of Costello and Osborne (2005: 3) in running multiple factor analyses by manually setting the number of required eigenvalues as 6 (the a priori number of factors in Biber 1988), and 7–12 as indicated in the scree plot. After a comparison of the seven factorial structures based on 6–12 factors in terms of the number of significant loadings (above 0.30) on each factor, cross loadings, and the ease of interpretation of the extracted factors, we decided to establish a nine-factor factorial structure on the basis of 2,500 texts totalling five million words.

The nine factors and their loadings after rotation are shown in Tables 2–10.[3] The tables also give their mean frequencies and standard deviations, which are required to compute their factor scores. Please note that the loadings/weights enclosed in brackets in these tables are not included in computing the factor score of a particular factor as they are loaded more strongly on another factor. One loading was used to compute the factor score of only one factor – in cases of cross loadings, the factor on which it has the greatest weight. Also note that all loadings with absolute values less than 0.30 are discarded as insignificant.[4]

As can be seen in Table 2, linguistic features positively loaded on Factor 1 are associated either with involved and interactive discourse type, or with a less elaborate style, which actually usually co-occur in a discourse. Linguistic features of the former category include examples such as private verbs (e.g. *believe*, *decide*, *feel*, *forget*, *guess*, *hope*, *know*, *mean*, *realise*, *see*, *understand*), discourse bin (i.e. discourse markers such as *ah* and *bravo*) and emphatic communication terms such as *after all* and *believe it or not*, and various types of pronouns (especially first and second person pronouns), frequent use of general/abstract verbs relating to being/existing (of which various word forms of verb *be* are most common, especially *be* as main verb), *do* as pro-verb, cognitive verbs (e.g. *believe*, *know*, *think*), sensory verbs (e.g. *hear*, *listen*, *look*, *see*), verbs of communications (e.g. *call*, *mean*, *phone*, *ring*, *write*), general adverbs such as *very*, exclusiviser/particulariser adverbs (e.g. *just*, *only*, *especially*, *simply*, *utterly*), boosters (e.g. *considerably*, *highly*, *like hell*), and compromisers (i.e. downtoners like *quite*, *pretty*, *rather*, *relatively*, *some way*). This kind of interactive discourse typically uses non-past tense and involves frequent use of *wh*-questions and *wh*-clauses, and causative subordinate clauses introduced by *because*. The less elaborate style of this kind of interactive discourse is signalled by frequent use of contracted forms, analytic negation (i.e. *not* negation), *that*-deletion, non-phrasal co-ordination, and predicative adjectives. In contrast, the linguistic features negatively loaded on this factor are either informationally heavy devices, or markers of a more elaborate style. Linguistic features of the first category include examples such as various types of nouns (especially nominalisation as a grammatical metaphor that reduces a process into a noun, see Halliday and Matthiessen 2004), attributive use of adjectives, adjectives of importance, prepositional phrases, and sentence relatives.[5] Informative discourse tends to use longer words and sentences and have a greater lexical density, as indicated by average word length and sentence length, and standardised type-token ratio. Such informationally dense discourse also tends to use general/abstract terms, terms depicting power/authority, influence, and organisation/administration, and terms depicting commencement. It is more elaborate in style, as indicated by its frequent use of phrasal co-ordination, passives, past participial post-nominal clauses (e.g. *a work written in nineteen thirteen*) and other types of past participial clauses. Causative verbs and other lexical devices indicating helping/hindrance are commonly used in this type of carefully planned discourse to reduce clause complexes of causative relationship to single clauses, in contrast to the frequent use of *because* in less elaborate interactive discourse. It is clear that this first dimension in our factorial structure is concerned with ‘interactive casual discourse vs. informative elaborate discourse’.

Table 3 shows the linguistic features loaded on Factor 2. They are all positive loadings. The most prominent features loaded on this factor are *that*-clauses as complements on nouns, verbs and adjectives. For pragmatic reasons of data extraction, *that*-complement clauses with a zero complementiser are omitted. *That*-clauses as noun complements are typically used in appositive constructions for informational elaboration (e.g. *Sometime later there was news that they were being held by the Iraqis but no more.* ICE-GB: S2B), while *that*-clauses as complements of verbs and adjectives tend to make evaluation

Table 2. Factor 1: Interactive casual discourse vs. informative elaborate discourse

Loading	Weight	Mean	Std. dev.
Positive			
Private verbs	0.90	14.14	9.88
Non-past tense	0.87	28.51	17.52
Discourse bin	0.86	30.00	31.65
Contraction	0.84	12.88	14.64
General/abstract verbs relating to being/existing	0.84	28.67	11.29
2nd person pronouns	0.84	12.53	13.45
1st person pronouns	0.83	21.21	18.64
Analytic negation	0.81	8.70	5.92
Cognitive verbs	0.78	9.28	5.65
Pronoun IT	0.75	11.22	7.43
That-deletion	0.74	2.97	3.11
BE as main verb	0.74	19.54	7.29
General adverbs	0.73	38.66	14.67
Causative subordinator	0.67	1.80	1.90
Indefinite pronouns	0.65	3.30	2.74
3rd person pronouns	0.64	25.30	15.54
Boosters	0.58	6.46	4.17
WH nominal clauses	0.58	0.50	0.68
Exclusiviser/particulariser adverbs	0.56	3.03	2.19
WH questions	0.52	0.83	1.26
Non-phrasal co-ordination	0.47	1.49	1.84
DO as pro-verb	0.45	1.13	1.18
Predicative adjectives	0.43	4.67	2.30
Compromisers	0.41	0.75	1.07
Sensory verbs	0.38	3.04	2.47
Demonstrative pronouns	0.36	6.15	4.17
Verbs of communications	0.34	2.34	2.06
(Degree adverbs	0.55)		
(Time adverbs	0.37)		
(Conditional subordinators	0.37)		
Negative			
Prepositions	−0.86	92.16	25.76
Other nouns	−0.83	196.96	51.40
Common nouns	−0.80	174.11	48.40
Attributive adjectives	−0.79	37.42	18.09
Standardised type-token ratio	−0.75	36.75	6.82
Average word length	−0.72	4.45	0.51
Nominalisation	−0.65	21.50	15.17
Phrasal co-ordination	−0.63	5.80	4.29
Mean sentence length	−0.63	17.76	8.74
Past participial WHIZ deletion	−0.59	1.56	1.55
Agentless passives	−0.56	7.15	5.08
By-passives	−0.55	1.09	1.11
General/abstract terms	−0.54	28.67	11.29
Power relationship: Power/organising	−0.53	6.39	5.98

(*Continued*)

Table 2. *Continued*

Loading	Weight	Mean	Std. dev.
Proper nouns	−0.48	28.98	23.18
Helping/hindrance: Helping	−0.48	3.49	3.48
Nouns for group/affiliation	−0.44	3.14	3.74
Adjectives of importance	−0.39	1.88	1.79
Nouns of social actions/states/processes	−0.37	1.96	2.14
Past participial clauses	−0.36	0.28	0.50
Sentence relatives	−0.36	0.27	0.58
Time: Beginning	−0.35	4.57	2.66
Helping/hindrance: Hindrance	−0.34	0.75	1.25
Causative verbs	−0.31	1.68	1.66
(Measurement	−0.38)		
(Numeral	−0.31)		

Table 3. Factor 2: Elaborative online evaluation

Loading	Weight	Mean	Std. dev.
THAT clauses as noun complements	0.73	2.66	2.16
THAT relative clauses	0.71	1.83	1.60
Demonstratives	0.59	6.60	3.49
WH relative clauses	0.55	2.58	2.09
THAT clauses as verb complements	0.49	3.10	2.24
Existential structure	0.42	2.43	1.77
Pied piping constructions	0.35	0.53	0.71
Nouns of evaluation	0.33	2.65	3.18
Nouns of speech acts	0.33	5.51	4.62
THAT clauses as adjective complements	0.30	0.36	0.51
(General/abstract verbs relating to being/existing	0.31)		
(BE as main verb	0.38)		
(No negative loading)			

(e.g. … *and history has proved that we forget at our peril.* ICE-HK: W2B; *I think it is possible that people do discriminate about subconsciously for things like like dress.* ICE-HK: S1A). Existential sentences are also linked to informational elaboration (cf. Biber 1988: 228). In addition to these complement clauses, *that*- and *wh*-relative clauses, pied piping (a special type of *wh*-relative clause), and demonstratives are all used to make the reference more specific. These linguistic features co-occur with nouns of evaluation and nouns of speech acts in discourse that focuses on elaborative evaluation in real-time speech production.[6] Hence Factor 2 in the model is labelled 'elaborative online evaluation'.

The linguistic features loaded on Factor 3 are given in Table 4. The most prominent feature is prepositional adverbs or particles in phrasal verbs. According to Biber, Johansson, Leech, Conrad and Finegan (1999: 409), approximately 75% of total phrasal verbs are activity verbs, which are very common in fiction and spoken English. It is therefore hardly surprising that phrasal verbs and verbs of movement and activity co-exist in the same

Table 4. Factor 3: Presentational concern

Loading	Weight	Mean	Std. dev.
Positive			
Prep adverbs/particles	0.79	5.85	4.17
Time: Present/simultaneous	0.65	4.93	3.46
Time adverbs	0.60	6.10	4.36
Place adverbs	0.60	3.96	3.33
Verbs of movement/activity	0.58	6.07	4.09
Linear order	0.44	6.88	4.14
Time: Ending	0.36	1.63	1.50
Diminishers	0.35	1.00	1.40
Negative			
Adjectives of comparison	−0.32	3.78	2.61

context. Linguistic features associated with time are prominent on this factor, for example, expressions related to present/simultaneous time (e.g. *at the moment*, *at the same time*, *at this point*, *by now*, *so far*, *meanwhile*, *now*), time adverbs, and expressions indicating ending in time (e.g. *abolition*, *come to an end*, *cancel*, *complete*, *end*, *finish*, *no longer*, *stop*). In addition to time expressions, other positive loadings on this factor include adverbs of place and degree (i.e. diminishers as a subset of downtoners such as *a bit*, *a little*, *partly*, *slightly*, *somewhat*, *to some extent*), and expressions of linear order (e.g. *afterwards*, *at first*, *before*, *earlier*, *eventually*, *finally*, *in the end*, *following*, *last*, *next*, *secondly*, *then*). These linguistic features play an important role in presentation (e.g. *And it's Louis who has taken up the lead from Doncaster at the moment*. ICE-GB: S2A). Hence, we propose to label this factor as 'presentational concern'. As comparison is not a focus in presentation discourse, it is unsurprising that adjectives of comparison are listed as a loading with a negative weight.

As can be seen in seen Table 5, only one positive loading is included for Factor 4, namely, nouns for people (i.e. terms relating to males or females). Two other linguistic features

Table 5. Factor 4: Human vs. object description

Loading	Weight	Mean	Std. dev.
Positive			
Nouns for people	0.32	5.83	5.33
(Power relationship: Power/organising	0.34)		
(Nouns for group/affiliation	0.32)		
Negative			
Nouns for object	−0.54	5.08	5.44
Measurement	−0.51	7.94	6.36
Nouns for substance/material	−0.44	3.00	7.58
Adjectives of general appearance/physical attributes	−0.41	0.45	0.82
Adjectives of shape	−0.32	0.33	0.88
Adjectives of colour	−0.31	0.77	1.85
(Possibility/permission/ability modals	−0.34)		

Table 6. Factor 5: Future projection

Loading	Weight	Mean	Std. dev.
Positive			
Infinitive clauses	0.75	13.87	5.08
Infinitive clauses as verb complements	0.60	6.50	2.90
Definite: Positive	0.59	10.64	4.89
Predictive/volitional modals	0.56	6.58	4.23
Time: Future	0.49	6.84	5.12
Possibility/permission/ability modals	0.48	5.77	3.35
Conditional subordinators	0.42	2.89	2.44
Infinitive clauses as adjective complements	0.37	1.10	0.93
Split auxiliaries	0.33	3.54	1.84
Necessity/obligation modals	0.32	1.87	1.87
(No negative loading)			

that are cross-loaded on this factor also relate to people (terms for power/organising, and nouns for group/affiliation), but they are not included in this factor. In contrast, negative loadings on this factor are associated with concrete descriptions of objects, including nouns for object, and substance/material, as well as adjectives describing appearance/physical attributes, shape, and colour. As such, this factor might be labelled as ‘human vs. object description’.

Table 6 shows the linguistic features loaded on Factor 5, which are all positive loadings. Of these, infinitive clauses, including those as complements on verbs and adjectives, are the prominent features. As infinitive clauses typically refer to what will happen in the future, they go naturally with future time expressions. The same can be said of the three kinds of modals loaded on this factor, namely, predicative/volitional modals (*will*, *would*, *shall*), possibility/permission/ability modals (*can*, *may*, *might*, *could*), and necessity/obligation modals (*ought*, *must*, *should*), all of which can express an epistemic meaning. Split auxiliaries co-occur with modal verbs because these auxiliaries are usually modals (cf. Biber 1988: 111). Conditional subordinators (e.g. *if*, *unless*, *as long as*) are clearly concerned with probability in the future. Positive expressions of definiteness (e.g. *achievable*, *by all means*, *can*, *certain*) relate to modality (e.g. *possible*, *necessary*, *certain*); they are in fact largely modals *per se*. All of these suggest that Factor 5 is concerned with ‘future projection’.

Table 7 indicates that linguistic features positively loaded on Factor 6 largely relate to possessive and reflexive pronouns as well as expressions of judgement and impression (e.g. adjectives of judgement/appearance such as *attractive*, *beautiful*, *gorgeous*, *impressive*, *lovely*, *nice*, *untidy*, and ‘seem/appear’ expressions such as *apparently*, *appear*, *look*, *seem*, *show*) through sensory experiences (though sensory verbs are not included in the factorial structure of this factor), especially of what happened in the past. According to Smith (2001: 11), possessive and reflexive pronouns are linguistic features contributing to subjectivity (see also Brinton 1995). Consequently, we label Factor 6 as ‘subjective impression and judgement’. The negative loading on this factor (i.e. numeral nouns like *hundred*, *thousand*, and *millions*) indicates that such subjective judgements are impressionistic and do not have a quantifying focus.

Table 7. Factor 6: Subjective impression and judgement

Loading	Weight	Mean	Std. dev.
Positive			
Possessive pronouns	0.65	11.80	7.22
Present participial clauses	0.49	0.18	0.45
Reflexive pronouns	0.486	1.04	1.14
Adjectives of judgement/appearance	0.33	1.06	1.34
Seem/appear	0.31	0.88	0.98
(Past tense	0.42)		
(Standardised type-token ratio	0.34)		
(Sensory verbs	0.31)		
Negative			
Numeral nouns	−0.36	1.30	2.24
(Numeral	−0.40)		

Table 8. Factor 7: Lack of temporal/locative focus

Loading	Weight	Mean	Std. dev.
(No positive loading)			
Negative			
Time: Period	−0.77	9.07	6.40
Temporal nouns	−0.56	7.24	4.48
Time: Momentary	−0.47	2.26	3.41
Numeral	−0.44	14.74	12.50
Nominal possessive pronouns	−0.41	0.25	0.79
Locative nouns	−0.32	0.25	0.73

As can be seen in Table 8, Factor 7 has no positive loadings.[7] The negative loadings indicate that this factor is primarily associated with avoidance of expressions of time (both temporal period and instantaneous moment), location, as well as nominal possessive cases of personal pronouns (e.g. *mine*, *yours* and *theirs*). Temporal nouns such as *day*, *hour*, *week* and *year* and numerals are also related to temporal periods. The avoidance of these linguistic features appears to suggest that Factor 7 lacks a temporal/locative focus, and it is thus labelled as such.

Table 9 shows that all positive loadings on Factor 8 relate to two related concerns. One is degree, as indicated by linguistic features such as general degree adverbs (e.g. *quite*, *very*), approximators (e.g. *approximately*, *more or less*), and maximisers (*completely*, *most of all*, *on the whole*). Boosters and compromisers are also expressions of degree, though these two cross loadings are not included in this factor. The other concern is quantity, as indicated by expressions of quantity (numerals and other quantifiers) and measurement, the latter of which is not included in this factor. Hence, Factor 8 can be labelled as ‘concern with degree and quantity’. This factor does not have a communicative focus, as signalled by the avoidance of nouns of communications (e.g. *agenda*, *document*, *leaflet*, *letter*, *message*, *telephone*).

Table 9. Factor 8: Concern with degree and quantity

Loading	Weight	Mean	Std. dev.
Positive			
Degree adverbs	0.57	6.78	3.79
Comparative/superlative degrees	0.49	4.67	2.84
Approximators	0.42	1.53	1.47
Quantity	0.41	23.35	8.59
Maximisers	0.33	1.29	1.16
(Boosters	0.43)		
(Measurement	0.35)		
(Compromisers	0.31)		
Negative			
Nouns of communications	−0.33	3.54	5.18

Table 10. Factor 9: Concern with reported speech

Loading	Weight	Mean	Std. dev.
Positive			
Public verbs	0.81	4.75	3.82
Verbs of speech acts	0.76	9.84	5.65
Past tense	0.48	21.85	18.42
Suasive verbs	0.37	1.69	1.47
(*That*-deletion	0.39)		
(No negative loading)			

Table 10 gives the loadings on Factor 9, the last factor in the model. Biber (1995: 152) observes that public verbs are used to report direct and indirect speech acts. Unsurprisingly, they tend to co-occur with verbs of speech acts. In fact, there is considerable overlap between the two categories: public verbs are “speech act verbs indicating indirect statements” (Quirk, Greenbaum, Leech and Svartvik 1985: 1180). Suasive verbs function as “mandative” and express “a directive to or intention for change” (Hinkel 2002: 105). They also overlap with public verbs (Quirk et al. 1985: 1182). Past tense is not uncommon in reporting what other people have said before, while *that*-deletion habitually occurs in reported clauses (e.g. *Mr Poon said it was too early to say whether he would run for Legco next year through the Election Committee.* ICE:HK: W2C). As such, Factor 9 is labelled as ‘concern with reported speech’.

Having established the factors and discussed how they are interpreted, it is now appropriate to introduce how the factor scores are computed. The factor score (κ) of a feature in a text can be computed using the following formula:

$$\kappa = \frac{F - \mu}{\sigma} \qquad (1)$$

In Equation (1), F is the normalised frequency (i.e. frequency per 1,000 words in this study) of the feature in the text, σ stands for standard deviation, and μ is the mean frequency of the feature in the whole group of texts (i.e. registers in a language variety).

The factor score of a feature in a group of text ($\bar{\omega}$) is the mean score of the feature in the group, that is, the sum of factor scores of the feature in each text of the group divided by the number of texts (N) in that group even if some files may not contain such a feature:

$$\bar{\omega} = \frac{\sum \kappa}{N} \tag{2}$$

The dimension score of a group of texts can be obtained by adding together the mean factor scores of all features with positive weights on a factor and then subtracting the mean factor scores of all features with negative weights on the same factor (see Tables 2–10):

$$\omega = \sum \bar{\omega} \tag{3}$$

The factor scores of the linguistic features loaded on the nine factors can be computed on the basis of their normalised frequencies in each text and their mean frequencies and standard deviations given in Tables 2–10 on the basis of Equation (1), while Equations (2) and (3) can be used to compute the factor score of a text and a register respectively. Since the three formulae would involve hundreds of thousands of time-consuming calculations, computer programs were written to perform these tasks in batches. The results are presented and discussed in the following two sections, which explore variation across registers and world varieties respectively.

REGISTER VARIATION

This section discusses variation in the 12 ICE registers in the whole dataset along the nine dimensions established in the previous section. Note that world varieties are combined in this investigation of register variation in this section.

Figure 2 shows the mean factor scores of the 12 registers on Factor 1 ('interactive casual discourse vs. informative elaborate discourse'). As can be seen, private conversation (direct and telephone conversations) is the most interactive and casual while academic writing is the most informative and elaborate. Spoken registers are generally more interactive and less elaborate than written registers with the exceptions of creative writing (novels and stories),

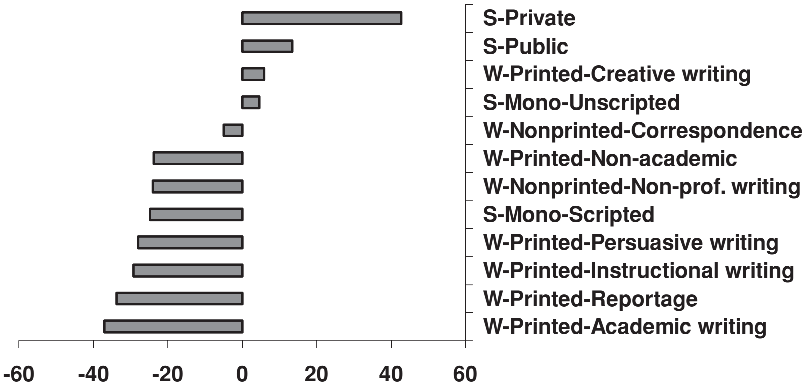

Figure 2. Factor 1: 'Interactive casual discourse vs. informative elaborate discourse' (F = 775.86, $p < 0.001$, R^2 = 77.4%).

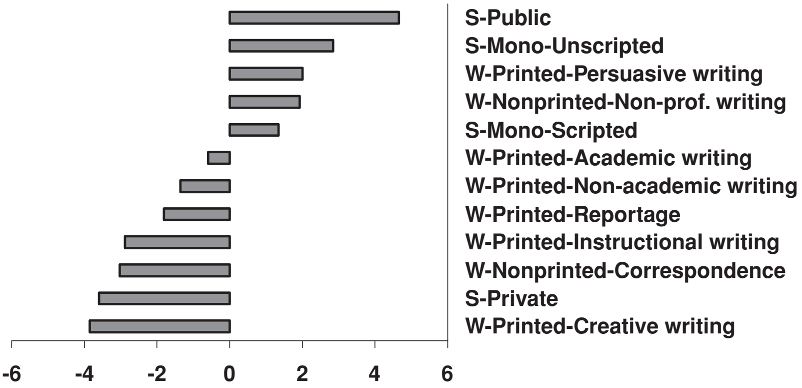

Figure 3. Factor 2: 'Elaborative online evaluation'
(F = 102.20, $p < 0.001$, $R^2 = 31.1\%$).

which are more akin to spoken registers (possibly because of the heavy use of fictional dialogues), and scripted monologue (broadcast news, broadcast talks and non-broadcast talks), which is actually written to be spoken. The ANOVA test was used to test for the significance of differences among these registers along each factor (see the next section for results of the two-way ANOVA test which takes account of register and world variety as co-variables). The test results are given in the brackets in Figures 2–10. A probability value $p < 0.05$ is considered as statistically significant. As can be seen, the F score for Factor 1 is 775.86, with $p < 0.001$, which means that the difference is highly significant. R^2 measures the percentage of variance in factor scores (77.4% in Factor 1) that can be accounted for directly by merely looking at the register label.

As can be seen in Figure 3, which compares the 12 registers along Factor 2, public speech (e.g. class lessons, broadcast discussions, parliamentary debates, legal cross-examinations and business transactions) has the most prominent focus on elaborative online evaluation; unscripted monologue also involves a high level of elaborative online evaluation. While persuasive writing may involve elaborative evaluation, such evaluation is not made online and thus, not restricted by real-time production. Evaluation is not a concern in creative writing. While private conversation can involve online evaluation, it is nevertheless least elaborative even if the evaluation is made online.

Figure 4 compares the registers in terms of presentational concern (Factor 3). As can be seen, unscripted monologue (e.g. demonstrations, presentations and commentaries) has a presentational concern in the ICE corpora. Creative writing is also presentational in the sense that it unfolds a story step by step before the reader. In contrast, presentation is not a concern in academic writing, non-professionals' writing (i.e. student essays and exam scripts), and instructional writing (e.g. administrative writing, skills and hobbies).

Figure 5 shows that along Factor 4, private conversation is most likely to have a focus on people. Correspondence (i.e. social letters and business letters) also involves human description. In contrast, instructional writing tends to give concrete descriptions of objects. Academic prose and non-academic writing can also be concrete when an object or substance is described, for example, by using measurement words and adjectives describing shape and colour.

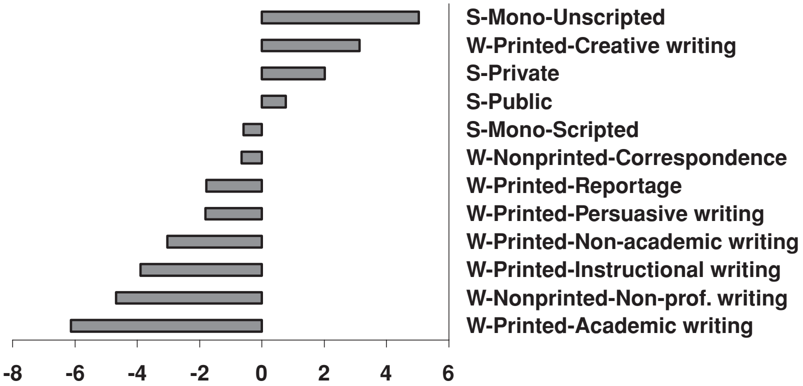

Figure 4. Factor 3: 'Presentational concern'
(F = 134.50, $p < 0.001$, $R^2 = 37.3\%$).

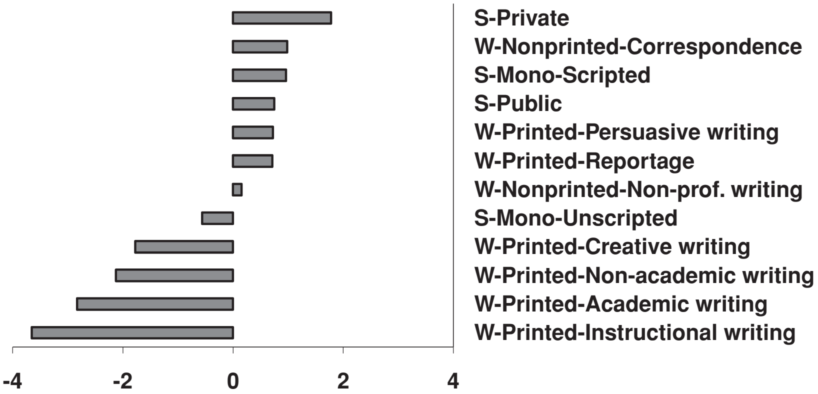

Figure 5. Factor 4: 'Human vs. object description'
(F = 44.03, $p < 0.001$, $R^2 = 16.3\%$).

It can be seen in Figure 6 that along Factor 5, persuasive writing such as press editorials has the most prominent focus on future projection, which is unsurprising given that persuasion is concerned with people's future attitudes and actions. Correspondence (i.e. social and business letters) and public speech (e.g. class lessons, broadcast discussions, parliamentary debates, legal cross-examinations and business transactions) also involve future projection to varying extents. In contrast, academic writing, which is concerned with timeless truths, is least concerned with future projection. Neither is future projection obtrusive in students' essays and exam scripts.

Figure 7 shows the distribution of twelve registers along Factor 6 'Subjective impression and judgement'. As can be seen, the factor score for creative writing is by far greater than any other register, whereas instructional writing, private conversation, and student essays demonstrate the lowest scores in this dimension. Creative writing scores high on this factor because of its frequent use of possessive and reflexive pronouns, as well as adjectives of judgement/appearance (e.g. *She stared at me with her radiant smile*. ICE-HK: W2F).

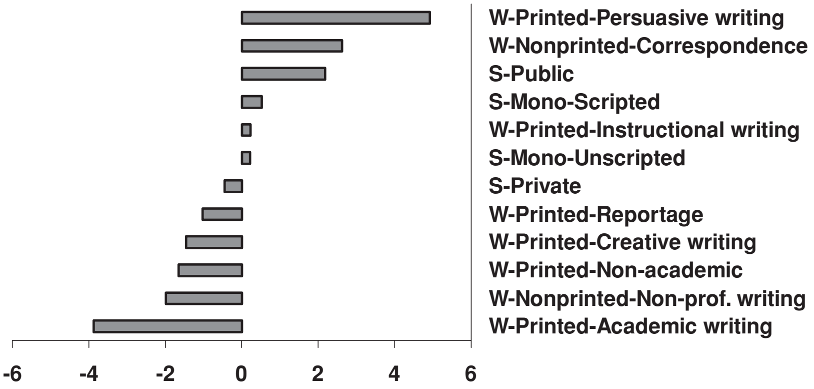

Figure 6. Factor 5: 'Future projection'
(F = 28.10, $p < 0.001$, $R^2 = 11.1\%$).

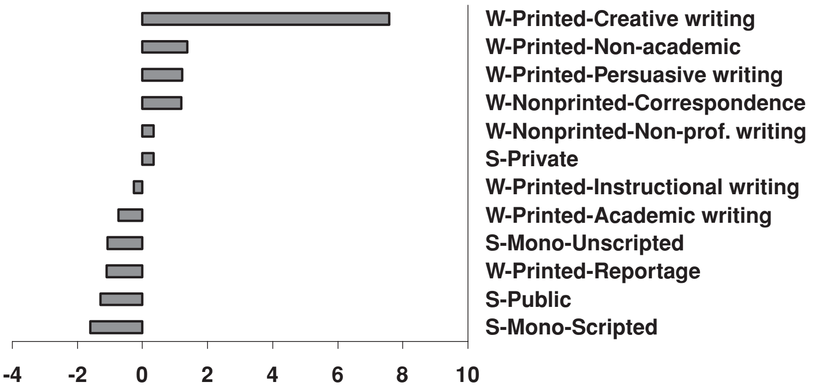

Figure 7. Factor 6: 'Subjective impression and judgement'
(F = 126.22, $p < 0.001$, $R^2 = 35.8\%$).

Brinton (1995) finds that reflexive pronouns, which are generally non-anaphoric in literary works, are strategically deployed to "represent the consciousness of narrated characters from their own point of view" (Finegan 1995:11).

This finding is also supported by the 100-million-word British National Corpus (BNC), where reflexive pronouns are most frequent in the fiction category, with a normalised frequency of 97.87 instances per million words, which nearly doubles that of the next most frequent category, namely, non-academic prose and biography (50.17 instances per million words). Instructional writing, private conversation and student essays, in contrast, show low scores because they do not have a focus on subjective impression and judgement. It is of interest to note that academic prose and non-academic writing (i.e. popular writing) in the same domains (humanities, social sciences, natural sciences, and technology) demonstrate different propensities in this dimension. While academic writing tends to avoid the tone of subjective impression and judgement as far as possible, non-academic writing does not appear to have this tendency.

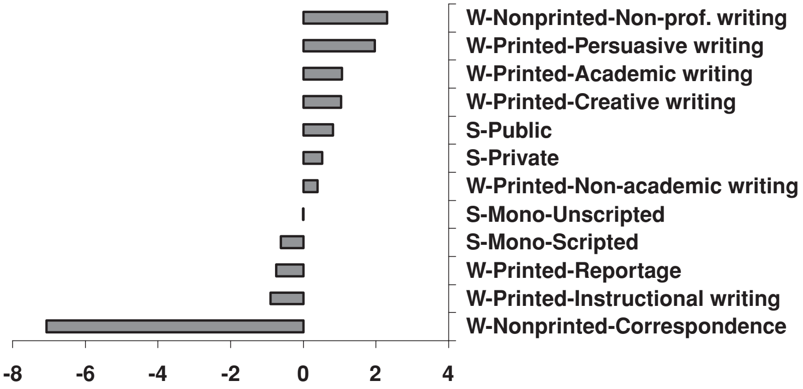

Figure 8. Factor 7: 'Lack of temporal/locative focus' ($F = 89.55$, $p < 0.001$, $R^2 = 28.4\%$).

It can be seen in Figure 8 that student essays and persuasive writing do not have a temporal/locative focus. They are not concerned with concepts such as *when*, *how long*, and *where*, whereas such specific information is of vital importance in correspondence (social and business letters).

Figure 9 indicates that along Factor 8 non-academic popular writing has the greatest concern of degree and quantity. Persuasive writing also displays a high propensity for expressions of degree and quantity. In contrast, such expressions tend to be avoided in instructional writing (e.g. administrative documents) and correspondence.

Finally, Figure 10 shows the distribution of various registers along Factor 9. It can be seen that news reportage is the register which has the greatest concern with reported speech (both direct and indirect speech). In news stories, it is not uncommon to find paragraphs after paragraphs of reported speech. Reported speech is also very common in creative writing like novels and stories. In contrast, instructional writing and academic

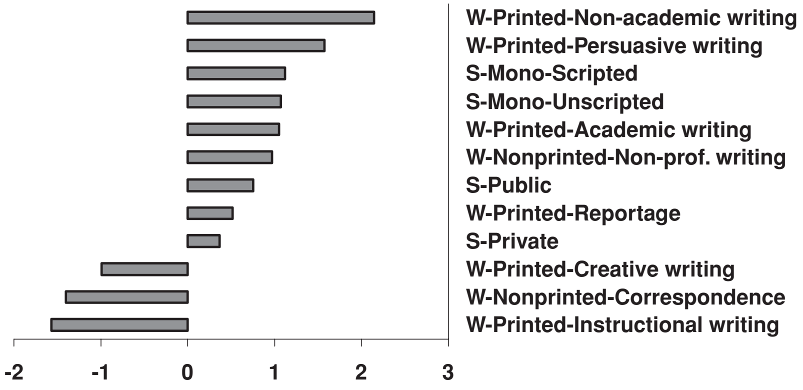

Figure 9. Factor 8: 'Concern with degree and quantity' ($F = 19.33$, $p < 0.001$, $R^2 = 7.9\%$).

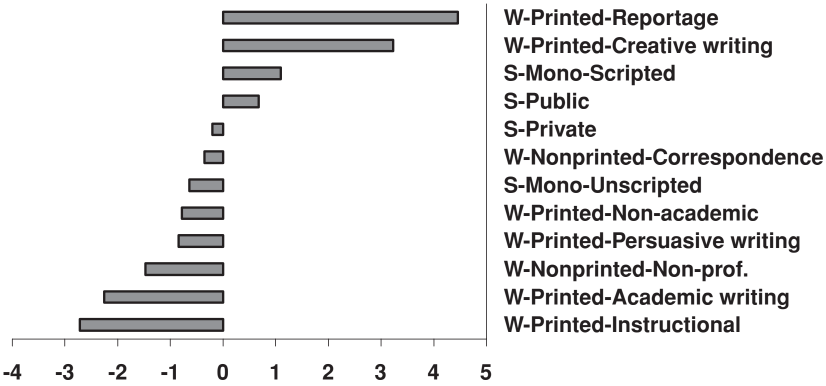

Figure 10. Factor 9: ‘Concern with reported speech’ (F = 80.02, $p < 0.001$, $R^2 = 26.1\%$).

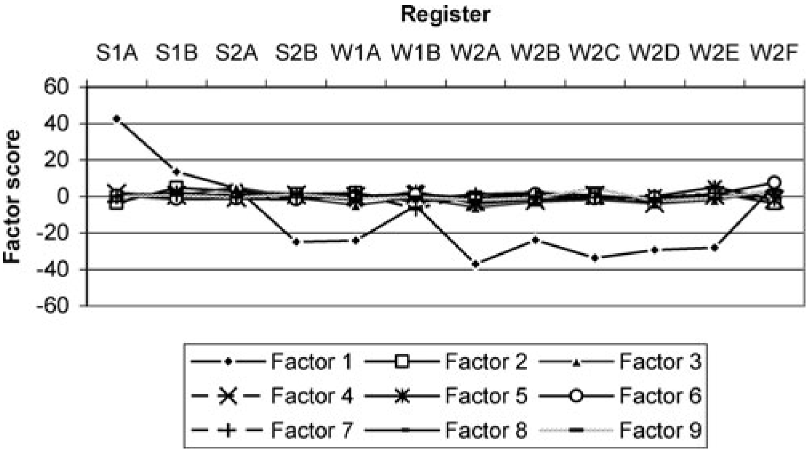

Figure 11. Contrasting 12 registers along nine factors

prose appear at the other extreme of the continuum, which do not have a concern with reported speech.

Of the nine factors established in this study, Factor 1 is the strongest dimension along which the 12 registers demonstrate the sharpest contrasts, as illustrated in Figure 11. This is hardly surprising given that this factor concerns the distinction between interactive casual discourse and informative elaborate discourse, which is a fundamental aspect of variation across registers.

VARIATION ACROSS FIVE VARIETIES OF ENGLISH

We have so far compared the twelve ICE registers along nine factors. While the results of such a comparison have demonstrated the robustness of the model established in this study, they do not show any potential similarities and differences between the five varieties of English under consideration, which will be explored in this section.

Table 11. Results of two-way ANOVA tests for Factors 1–9

Factor	Variety		Register		Variety × register	
	F score	Sig. level	F score	Sig. level	F score	Sig. level
Factor 1	19.37	<0.001	859.20	<0.001	3.20	<0.001
Factor 2	10.11	<0.001	106.22	<0.001	1.42	0.035
Factor 3	7.58	<0.001	139.98	<0.001	2.21	<0.001
Factor 4	4.51	0.001	44.56	<0.001	1.13	0.254
Factor 5	48.79	<0.001	32.23	<0.001	4.24	<0.001
Factor 6	14.33	<0.001	132.18	<0.001	1.99	<0.001
Factor 7	2.70	0.029	92.66	<0.001	2.76	<0.001
Factor 8	17.66	<0.001	20.72	<0.001	2.68	<0.001
Factor 9	5.23	<0.001	84.23	<0.001	3.87	<0.001

As language may vary across registers even more considerably than across language varieties (cf. Biber 1995: 278), many register-based subtleties can be blurred if we compare the five varieties of English on the basis of combined registers. Hence, we decided to compare world varieties along 12 registers factor by factor. However, due to lack of space, the first five dimensions have been chosen, which are also the strongest factors in the enhanced model.

Before the register-based variation across world varieties was studied, statistical tests to determine whether such variation is statistically significant along each of the nine dimensions were run. As there are two independent variables (i.e. world variety and register), two-way ANOVA tests are appropriate, the results of which are shown in Table 11. As can be seen, both register and variety demonstrate significant main effects in all dimensions. The variety by register interaction is also significant in all dimensions except Factor 4, where the variety by register interaction is not significant ($p = 0.254$).

Now let us first consider Factor 1, 'interactive casual discourse vs. informative elaborate discourse'. Figure 12 shows that factor scores of 12 registers in the five varieties of English along this dimension. It can be seen the Indian English displays the lowest score for

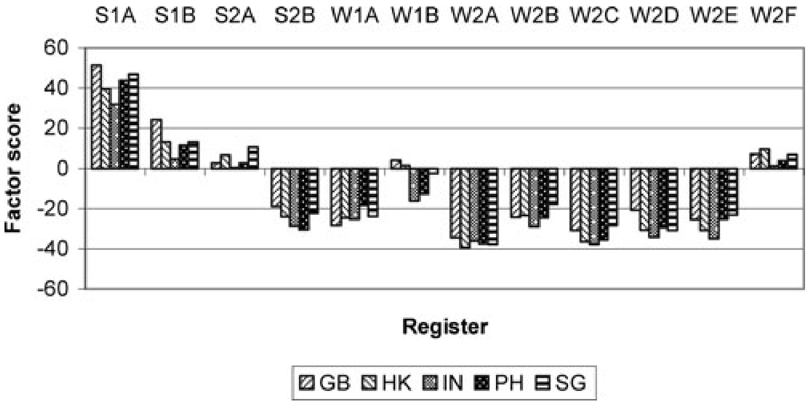

Figure 12. Variation across language varieties along Factor 1

Factor 1 in nearly all registers, meaning that it is less interactive but more elaborate. If "elaborateness" is interpreted as indirectness and clumsiness, we must agree with Sanyal (2007), who describes in the blurb of the book Indian English as "clumsy Victorian English [that] hangs like a dead albatross around each educated Indian's neck". Indian English suffers from "flatulent orotundity, a form of high-flown language that tries to impress but instead obscures" (Cutts 2007: 2). This is partly a legacy of the Raj and the East India Company, and partly a result of influences of native Indian languages, which give primacy to nouns rather than verbs. Consequently, Indian English is characterised with a "nouny" style. Nouns of various categories, and relatedly prepositions, are all negative loadings on Factor 1. In contrast, modern British English appears to be most interactive and least elaborate in registers such as private and public conversations (S1A and S1B), and instructional writing (W2D). A sharp contrast between Indian English and British English is self-evident in excerpts (1) and (2):

> (1) I deem it to be privileged to communicate to you, on behalf of the Executive Committee of Maharashtra State Commerce Conference their decision to felicitate the seniormost Commerce & Management Educators in appreciation of their contribution to the field of Commerce & Management Education and also their patronage in their field, and that the Executive Committee, with all humility has suggested your name for the rare kind of felicitation (ICE-IN: W1B).
>
> (2) I am writing to you personally, on behalf of the University College London branch of the Association of University Teachers, as I understand you are a member of the Council of Queen Mary and Westfield College (ICE-GB: W1B).

The three varieties of English as used in Southeast Asia (Hong Kong, Singapore and the Philippines) are very similar along Factor 1, lying between British English and Indian English. In some registers, for example, academic writing (W2A) and creative writing (W2F), the differences are hardly distinguishable. This similarity can be accounted for by the fact that these Asian varieties of English either share common background languages (e.g. Chinese in Singapore and Hong Kong) or influence each other through language contact (e.g. the influence of Philippine English on Hong Kong English because of the large number of Filipino domestic workers in Hong Kong) (cf. Hickey 2004: 514–515).

Figure 13 gives the factor scores of 12 registers along Factor 2, 'elaborative online evaluation'. It can be seen that British English has a generally higher score for this dimension than other world varieties, as in academic writing (W2A), correspondence (W1B) and scripted monologue (S2B), where non-native varieties of English in this study are strikingly similar. Excerpt (3), taken from unscripted monologue (S2A) of British English, illustrates the frequent use of *that*-clause for online information elaboration.

> (3) The arguments that the great mass of the ancient Egyptian population could not read or write is partly based upon our general perception of the nature and structure of the society apart from the point that there is no positive evidence that anyone of clearly commoner status ever wrote anything (ICE-GB: S2A).

The differences in creative writing (W2F) and private conversation (S1A) are less marked because these registers either do not have an evaluative concern or are not produced online under real-time constraints. While it is not immediately clear why British English tends to show a higher score in most registers in this dimension, which would be better explained

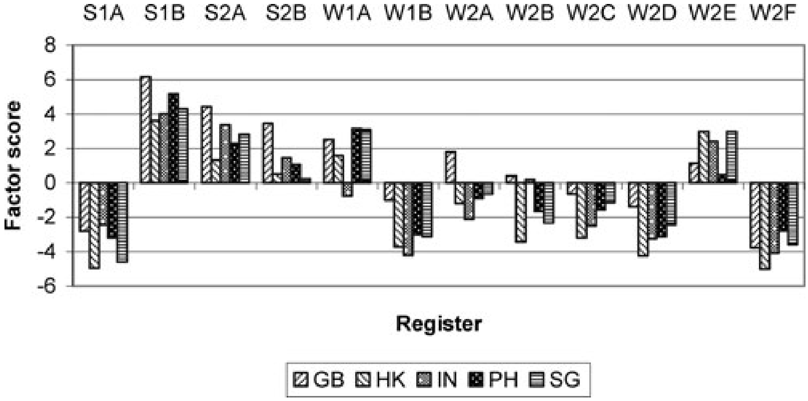

Figure 13. Variation across language varieties along Factor 2

by using methods other than the corpus approach including socio-cultural and language acquisition research, it might be speculated that online elaboration is a dimension that distinguishes between native and non-native varieties of English, because this dimension involves producing elaborate discourse under real-time conditions.

It can be seen in Figure 14 that the five varieties of English do not differ much along Factor 3 in a number of registers, either because those registers have a presentational concern in all world varieties (e.g. creative writing W2F), or because they do not (e.g. academic writing W2A, student essays W1A). However, the overall difference between the five world varieties along this dimension is statistically significant because of sharp contrasts in registers such as private conversation (S1A) and correspondence (W1B). It is of interest to note that in the register of private conversation (S1A), Hong Kong English has a negative score, which is in sharp contrast with the other varieties. A closer examination of individual features loaded on this factor indicates that prepositional adverbs/particles (i.e. phrasal verbs), namely, the feature with the greatest weight (see Table 4), are significantly less frequent in this register in Hong Kong English than in the other four varieties, with a

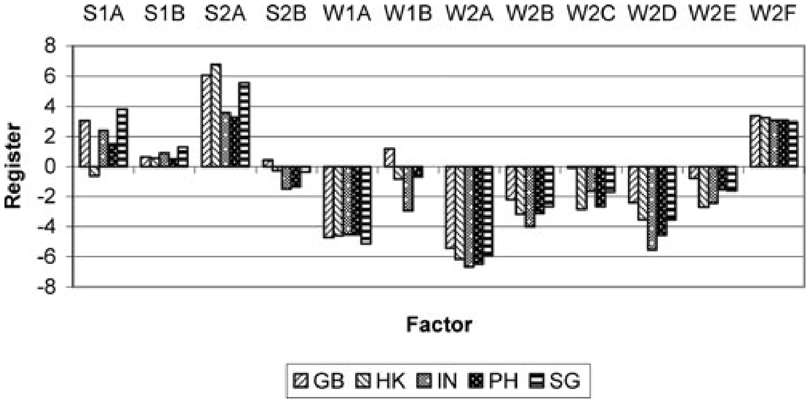

Figure 14. Variation across language varieties along Factor 3

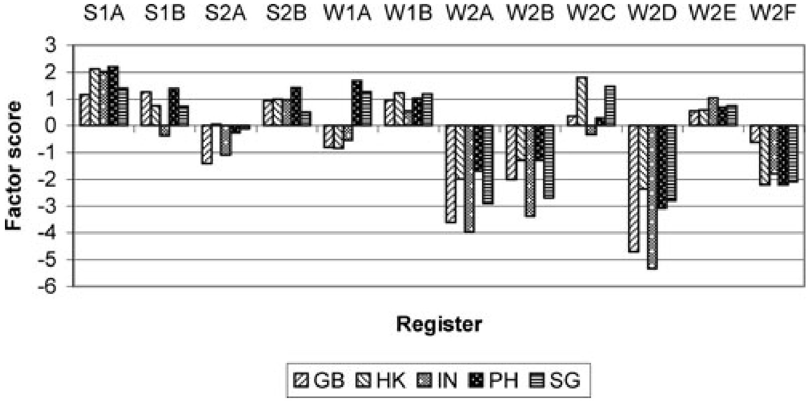

Figure 15. Variation across language varieties along Factor 4

normalised frequency less than half of that in British English (486.8 vs. 1021.45 instances per million words respectively). Bolton and Nelson (2002: 260) highlight phrasal verbs as one of the three categories of features of potential interest in Hong Kong English. According to Hung (2000), the vocabulary of many Hong Kong English speakers does not incorporate many phrasal verbs. One possible explanation Hung suggests is that phrasal verbs are largely unpredictable and therefore have to be learned individually. Similarly, in the register of correspondence (W1B), Indian English displays a much lower dimension score than other varieties, which results from exceptionally lower frequencies of adverbs of time/place and expressions for present/simultaneous time. The lack of specificity of temporal/locative information in Indian English may have a cultural root because India, with a high-context culture, has a concept of time which is different from Western cultures. As a result, Indian English is least concerned with presentation in correspondence (W1B), instructional writing (W2D), and unscripted monologue (S2A), in contrast with British English, which demonstrates a greater propensity for presentational concern, most noticeably in news reportage (W2C) and instructional writing (W2D).

Figure 15 shows variation along Factor 4, ‘human vs. object description’. As can be seen, while there are a number of registers which are very similar in the five world varieties along this dimension (e.g. W2E – persuasive writing, W1B – correspondence, as well as S2B – unscripted monologue), it is also of interest to note that Indian English and British English are similar in a greater range of registers including, in addition to the above mentioned, non-printed writing (W1A, W1B), academic writing (W2A), news reportage (W2C) and instructional writing (W2D). Creative writing in world varieties other than British English is strikingly similar, possibly reflecting a potential boundary between native and non-native varieties. In addition, varieties of English used in Hong Kong and Singapore demonstrate great similarity along this dimension in unscripted monologue (S2A), news reportage (W2C) and instructional writing (W2D). Finally, if we look at the whole picture, it appears that Indian English tends to be least concerned with human description while giving concrete descriptions of objects, as illustrated in excerpt (4), taken from instructional writing in Indian English:

> (4) However, some of the visual characteristics, by which one can diagnose the affected soils in the field itself, are given below: Prominently white fluffy salt encrustation on the surface during dry period, when

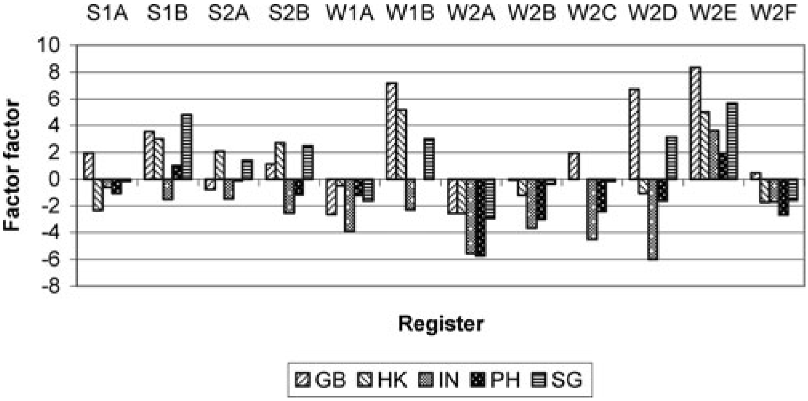

Figure 16. Variation across language varieties along Factor 5

> the net movement of soil moisture is upward. The salts dissolved in the soil water move to the surface, where they are left as a crust when the water evaporates (ICE-IN: W2D).

Figure 16 indicates that there is considerable variation along Factor 5, which is concerned with future projection. British English has the highest score in all printed written registers (W2A–W2F), non-printed correspondence (W1B) and private conversation (S1A). In contrast, Indian English shows the lowest score in nearly all registers (cf. Berglund 1997), with the exceptions of private conversation (S1A), persuasive writing (W2E) and creative writing (W2F), where differences among non-native varieties are less marked. The cross-register distribution patterns of English used in Hong Kong and Singapore appear to be closer to British English, whereas Philippine English is more akin to Indian English along this dimension. Such interesting results, which certainly merit further research, can probably be explained from the socio-cultural perspective. As Shastri (1988: 18) points out, "[m]aybe the Indian mind is not given to thinking much in terms of the future". On the other hand, Hong Kong and Singapore varieties are similar in this dimension possibly because they are both influenced by British English as former British colonies while they also share a common Chinese cultural background.

This section compared five varieties of English along the first five dimensions in the MDA model. This study has also attempted to provide some explanations for the similarities and differences observed in its corpora. The final section that follows will conclude this study by summarising the findings and exploring directions for future research.

CONCLUSIONS

In answer to the two research questions in the Introduction section, this paper has demonstrated that semantic categories can indeed be used to enhance the multidimensional analysis (MDA) approach established in Biber (1988); the paper has also introduced the new enhanced model in the study of world Englishes by exploring language variation across ICE registers and world varieties. The MDA approach is undoubtedly a very powerful analytical framework for language variation study, but Biber's (1988) model is largely confined to grammatical categories. The present study has sought to incorporate in the MDA approach the results of CLAWS and USAS, two powerful corpus annotation systems developed at Lancaster University for grammatical and semantic analyses. A total of 141 linguistic features, both grammatical and semantic, were used at the initial stage of modelling, while the final model includes 109 linguistic features, with others dropped because they were either infrequent in the data, or they overlapped to a great extent with some more generally defined features. The 109 linguistic features are (positively or negatively) loaded on nine factors, all of which are very strong and stable, with at least four loadings with weights above 0.30. These dimensions are: (1) interactive casual discourse vs. informative elaborate discourse; (2) elaborative online evaluation; (3) presentational concern; (4) human vs. object description; (5) future projection; (6) subjective impression and judgement; (7) lack of temporal/locative focus; (8) concern with degree and quantity; and (9) concern with reported speech.

This new model is based on five million words of spoken and written data sampled from 12 registers representing five world varieties covered in the International Corpus of English: Great Britain, Hong Kong, India, the Philippines, and Singapore. A multidimensional analysis of the 12 registers has demonstrated that the new model is very robust. On the other hand, a comparative study of the five world varieties has painted a complex picture, demonstrating both similarities and differences. This is so because variations in language use involve regional varieties as well as variants in different registers and along different dimensions (i.e. factors).

For example, in the first dimension, modern British English appears to be most interactive while Indian English is most elaborate; Southeast Asian varieties of English as used in Hong Kong, Singapore and the Philippines are very similar. In dimension 2, British English displays a higher propensity for elaborative online evaluation while the four non-native world varieties are quite similar. In terms of the presentational concern in dimension 3, the five world varieties are similar in registers such as creative writing and academic prose; British English demonstrates a greater propensity for presentational concern, but Hong Kong English is least concerned with presentation in private conversation while Indian English displays a much lower dimension score in correspondence. In dimension 4 concerning human vs. object description, British English and Indian English are closer to each other than other world varieties, while the four non-native varieties also display great similarities in creative writing. Along dimension 5 for future projection, British English has the highest score whereas Indian English shows the lowest score in nearly all registers; Hong Kong English and Singapore English are closer to British English whereas Philippine English is more similar to Indian English. While further research is certainly required to be based on resources other than corpora to provide an explanation of these interesting similarities and differences, it can be speculated that language status (i.e. native vs. non-native), language contact, and cultural background are among the relevant contributing elements.

There are a number of directions which we think can be fruitfully explored in future research. First, the present study has been based on data from one English variety from the Inner Circle and four from the Outer Circle, the latter of which are all used in Asia. Clearly, future research will benefit from inclusion of other native English varieties, for example, the New Zealand and US components of the ICE corpus; it will also benefit from a wider and more balanced coverage of geographical regions for Outer Circle world varieties, for example, ICE-East Africa (which needs adjustments to make it comparable to other components) and ICE-Jamaica. Second, the model has intentionally excluded many socio-culturally relevant semantic categories to make this pilot study manageable. It will be desirable for such semantic categories to be included on a larger project in the future, because socio-cultural concepts and semantic domains are obviously relevant to studies of world Englishes as indicated earlier in this paper.

Finally, the corpus-based approach taken in this study has also defined its limitation. A corpus-based study is necessarily more descriptive than explanatory. The enhanced MDA model established in this study is only possible when a large amount of text and a large number of linguistic features are examined at the same time. Unsurprisingly, it is robust enough to account for a large amount of attested data and can provide interesting insights in variation across registers and world varieties. Nevertheless, while a corpus can show some interesting findings, it will not explain what has been found in the corpus. Hence, future research will greatly benefit from combining corpora and more traditional resources in socio-cultural and historical research in an attempt to provide an adequately descriptive and sufficiently explanatory multidimensional comparative account of world Englishes.

ACKNOWLEDGEMENTS

We are obliged to Doug Biber for his timely advice and very insightful comments and to Jiajin Xu, Andrew Hardie and Anthony Grant for their constructive suggestions. Special thanks go to the audience at the Fourth Inter-Varietal Applied Corpus Studies (IVACS) Conference at the University of Limerick on 13–14 June 2008, for their useful comments.

NOTES

1. See http://ucrel.lancs.ac.uk/claws7tags.html for more details of the CLAWS C7 tagset.
2. Note that some of the semantic categories in the USAS semantic tagset which are more socio-culturally related are excluded in this pilot study to make the initial set of linguistic features manageable and appropriate for a pilot project of this size.
3. Factor rotation helps "to simplify and clarify the data structure" (Costello and Osborne 2005: 3). As a result, the rotated factorial structure is easier to interpret.
4. A loading of 0.32 has been cited as the threshold value for statistical significance, which roughly corresponds to 10% overlapping variance with the other loadings in a particular factor (cf. Costello and Osborne 2005: 4).
5. Sentence relatives are all non-restrictive relative clauses, which have no qualifying function but serve to provide extra information (e.g. *Activity modules, which are scheduled by a kernel, require access interfaces in ports.* ICE-GB: W2A). According to Biber (personal communication), sentence relatives "are restricted almost entirely to speech". It is then surprising to find that this feature should be a negative loading on Factor 1, albeit with a relatively small weight. It would be less surprising, however, if we have a look at the distribution pattern of sentence relatives in the 100-million-word British National Corpus (BNC), where sentence relatives are nearly twice as frequent in writing as in speech (53.66 and 27.19 instances per million words in the written and spoken BNC respectively). If sentence relatives are viewed as a spoken feature, they are more likely to be found in context-governed and written-to-be-spoken registers (39.83 and 32.85 instances per million words), where they are still less common than in written books and periodicals (with a normalised frequency of 54.49) and miscellaneous writings (with a normalised frequency of 48.41). Sentence relatives are relatively uncommon in typical spoken English such as direct conversation (with a normalised frequency of 8.74).
6. In earlier MDA studies, linguistic features such as *that*-clauses as noun complements and pied piping constructions are usually associated with written registers. It seems surprising to find that three spoken registers are among those displaying a positive score on this factor (see the 'Register Variation' section). They are public speech (broadcast discussions, parliamentary debates, legal cross-examinations, etc.) and scripted and unscripted monologues. However, since these written-to-be-spoken registers or prepared speeches are very different from typical spoken registers like direct conversation, which is at the other end of the continuum in this dimension, it is unsurprising to find some written features in such pseudo-spoken registers (e.g. *There is still a possibility that the hot cars may still fall under the hands of Cabinet members and other government officials.* ICE-PH: S2B).
7. The polarity of this factor is reversed to facilitate comparison across factors. This transformation will not affect the results as the positive vs. negative sign only indicates polarity but not strength of association (cf. Biber 1995: 408).

REFERENCES

Archer, Dawn, Wilson, Andrew, and Rayson, Paul (2002) *Introduction to the USAS category system*. UCREL Working paper, Lancaster University. Accessed on 28 April 2008 at http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf.

Bao, Zhiming, and Hong, Huaqing (2006) Diglossia and register variation in Singapore English. *World Englishes 25*, 105–14.

Bauer, Laurie (2002) *An Introduction to International Varieties of English*. Edinburgh: Edinburgh University Press.

Berglund, Ylva (1997) Future in present-day English: Corpus-based evidence on the rivalry of expressions. *ICAME Journal 21*, 7–20.

Berns, Margie (2005) Expanding on the expanding circle: where do WE go from here? *World Englishes 24*, 85–93.

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

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

Biber, Douglas, Connor, Ulla and Upton, Thomas A. (2007) *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.

Biber, Douglas, Conrad, Susan, Reppen, Randi, Byrd, Pat, and Helt, Marie (2002) Speaking and writing in the university: a multidimensional comparison. *TESOL Quarterly 36*, 9–48.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, and Finegan, Edward (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Bolton, Kingsley (2005) Where WE stands: approaches, issues, and debate in world Englishes. *World Englishes 24*, 69–83.

Bolton, Kingsley & Nelson, Gerald (2002) Analysing Hong Kong English: sample texts from the International Corpus of English. In Kingsley Bolton (ed.), *Hong Kong English: Autonomy and Creativity* (pp. 241–64). Hong Kong: Hong Kong University Press.

Brinton, Laurel (1995) Non-anaphoric reflexives in free indirect style: expressing the subjectivity of the non-speaker. In Dieter Stein and Susan Wright (eds.), *Subjectivity and Subjectivisation* (pp. 173–94). Cambridge: Cambridge University Press.

Collins, Peter (2009) Modals and quasi-modals in world Englishes. *World Englishes 28*, 281–92.

Conrad, Susan & Biber, Douglas (eds.) (2001) *Variation in English: Multidimensional Studies*. Cambridge: Cambridge University Press.

Costello, Anna B., and Osborne, Jason W. (2005) Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation 10*, 1–9.

Crystal, David (2003) *English as a Global Language*, 2nd edn. Cambridge: Cambridge University Press.

Cutts, Martin (2007) Foreword. In Jyoti Sanyal, *Indlish – the Book for Every English-speaking Indian*. New Delhi: Viva Books.

Deterding, David (2007) *Singapore English*. Edinburgh: Edinburgh University Press.

Finegan, Edward (1995) Subjectivity and subjectivisation: an introduction. In Dieter Stein and Susan Wright (eds.), *Subjectivity and Subjectivisation* (pp. 1–15). Cambridge: Cambridge University Press.

Fishman, Joshua A. (2008) *Standards and Norms in the English Language*. Berlin: Mouton de Gruyter.

Gisborne, Nikolas (2000) Relative clauses in Hong Kong English. *World Englishes 19*, 357–71.

Halliday, Michael A. K., and Matthiessen, Christian M. I. M. (2004) *An Introduction to Functional Grammar*. London: Arnold.

Hickey, Raymond (2004) Englishes in Asia and Africa: origin and structure. In Raymond Hickey (ed.), *Legacies of Colonial English: Studies in Transported Dialects* (pp. 503–35). Cambridge: Cambridge University Press.

Hinkel, Eli (2002) *Second Language Writers' Text: Linguistic and Rhetorical Features*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hoffmann, Sebastian & Mukherjee, Joybrato (2007) Ditransitive verbs in Indian English and British English: a corpus-linguistic study. *Arbeiten aus Anglistik und Amerikanistik 32*, 5–24.

Hung, Tony (2000) Interlanguage analysis and remedial grammar teaching. In Hong Kong Baptist University (ed.), *Papers in Applied Language Studies* (Vol. 5, pp. 155–68). Hong Kong: Hong Kong Baptist University.

Jenkins, Jennifer (2003) *World Englishes: A Resource Book for Students*. London: Routledge.

Kachru, Braj B. (1992) Teaching world Englishes. In Braj B. Kachru (ed.), *The Other Tongue, English across Cultures*, 2nd edn. (pp. 354–65). Urbana, IL: University Illinois Press.

Kachru, Braj B. (2008) The first step: The Smith paradigm for intelligibility in world Englishes. *World Englishes 27*, 293–6.

Kachru, Braj B., Kachru, Yamuna, and Nelson, Cecil L. (2006) *The Handbook of World Englishes*. Oxford: Blackwell.

Kachru, Yamuna (2003) On definite reference in world Englishes. *World Englishes 22*, 497–510.

Kachru, Yamuna, and Smith, Larry E. (2008) *Cultures, Contexts, and World Englishes*. London: Routledge.

Kachru, Yamuna, and Smith, Larry E. (2009) The Karmic cycle of world Englishes: Some futuristic constructs. *World Englishes 28*, 1–14.

Kasanga, Luanga A. (2006) Requests in a South African variety of English. *World Englishes 25*, 65–89.

Kirkpatrick, Andy (2002) *Englishes in Asia: Communication, Identity, Power and Education*. Melbourne: Language Australia Ltd.

Kirkpatrick, Andy (2007) *World Englishes*. Cambridge: Cambridge University Press.

Lim, Lisa (2007) Mergers and acquisitions: on the ages and origins of Singapore English particles. *World Englishes 26*, 446–73.

Mair, Christian (2007) Varieties of English around the world: collocational and cultural profiles. *Topics in English Linguistics 54*, 437–70.

Mesthrie, Rajend, and Bhatt, Rakesh M. (2008) *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Michieka, Martha M. (2009) Expanding Circles within the Outer Circle: the rural Kisii in Kenya. *World Englishes 28*, 352–64.

Mukherjee, Joybrato, and Hoffmann, Sebastian (2006) Describing verb-complementational profiles of New Englishes: a pilot study of Indian English. *English World-Wide 27*, 147–73.

Nelson, Gerald (1996) The design of the corpus. In Sidney Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English* (pp. 27–35). Oxford: Clarendon Press.

Nelson, Gerald (2006) The core and periphery of world Englishes: a corpus-based exploration. *World Englishes 25*, 115–129.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, and Svartvik, Jan (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Rayson, Paul (2003) Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD, Lancaster University.

Rayson, Paul (2008) *Wmatrix: A Web-based Corpus Processing Environment*. Computing Department, Lancaster University. Software available at http://ucrel.lancs.ac.uk/wmatrix/

Reppen, Randi, Fitzmaurice, Susan M., and Biber, Douglas (eds.) (2002) *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Sand, Andrea (2004) Shared morpho-syntactic features in contact varieties of English: article use. *World Englishes 23*, 281–98.

Sanyal, Jyoti (2007) *Indlish – the Book for Every English-Speaking Indian*. New Delhi: Viva Books.

Schneider, Edgar W. (2007) *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.

Scott, Mike (1996, 2004, 2008) *The WordSmith Tools* [versions 3.0, 4.0, 5.0]. Oxford: Oxford University Press.

Seidlhofer, Barbara (2009) Common ground and different realities: world Englishes and English as a lingua franca. *World Englishes 28*, 236–45.

Shastri, S. V. (1988) The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal 12*, 15–26.

Smith, Carlota (2001) *Accounting for Point of View and Subjectivity*. Germanistisk Institutt SPRIK report 4. Oslo: University of Oslo.

Trudgill, Peter (2004) *New-dialect Formation: The Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.

Ulrike, Gut (2007) First language influence and final consonant clusters in the new Englishes of Singapore and Nigeria. *World Englishes 26*, 346–59.

van Rooy, Bertus (2006) The extension of the progressive aspect in Black South African English. *World Englishes 25*, 37–64.

Xiao, Richard Z., and McEnery, Anthony M. (2005) Two approaches to genre analysis: three genres in modern American English. *Journal of English Linguistics 33*, 62–82.

Yano, Yasukata (2001) World Englishes in 2000 and beyond. *World Englishes 20*, 119–132.

(Received 24 June 2008.)