English Language and Linguistics, 28.2: 371–400.
doi:10.1017/S1360674324000017

An exploratory investigation of functional variation in South Asian online Englishes[1]

MUHAMMAD SHAKIR
University of Münster
(Received 10 July 2023; revised 4 January 2024)

This article conducts an exploratory multidimensional (MD) analysis of four interactive online registers, namely newspaper comments, tweets, web forums and text messages, originating from four South Asian countries (Bangladesh, India, Pakistan and Sri Lanka) and two Inner Circle (Kachru 1985) English-speaking countries (UK and USA). A principal component analysis (PCA) has been performed on the interactive registers using linguistic features tagged by a modified version of the MFTE tagger (Le Foll 2021a). The dimensions resulting from the PCA show that nominal, literate and informational features are generally more common in the South Asian data – which represent varieties belonging to the Outer Circle (Kachru 1985). Additionally, different features are used for expressing persuasion or opinion compared to the two reference varieties.

Keywords: register variation, World Englishes, South Asia, computer-mediated communication, multidimensional analysis

1 Introduction

Registers are situational varieties of language (Biber & Conrad 2019) and one of the most important factors in the study of language variation. In the last three or so decades, the internet has introduced a host of new registers and communicative situations. On the World Wide Web, there are new registers such as blogs and web forums along with older ones like news. On the more recently invented social media websites, there are multiple registers such as tweets, comments etc. Messaging apps like WhatsApp and Telegram enable communication that has certain similarities with spoken conversations (e.g. Jonsson 2015: 290 on ICQ chats and speech). The English language is one of the most dominant languages of the internet, in countries where it is the majority and native language, like the United Kingdom, as well as where it coexists with other indigenous languages, e.g. Bangladesh or India in South Asia. The study of internet-based registers in contexts like South Asia can enlighten us about register and functional variation in these regional varieties of English on the internet, especially if a

[1] I would like to acknowledge that this project has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 452561886.

framework like multidimensional (MD) analysis (e.g. Biber 1988) is used in which both aspects, i.e. the study of regional as well as register variation, can be combined.

The current state-of-the-art in the study of internet-based registers is that most of the MD studies have looked at Inner Circle (IC) varieties of English (Biber & Egbert 2016, 2018; Berber Sardinha 2022). The number of online registers is limited in studies that include Outer Circle (OC) and/or Expanding Circle (EC) (Kachru 1985) varieties (e.g. Bohmann 2019 only uses Twitter data) or their focus is only on one variety (e.g. Shakir & Deuber 2018, 2019 on Pakistani English). The present research article addresses this by presenting an MD analysis of four OC varieties of English from the South Asian region – namely Bangladesh, India (labelled as the regional 'super-central variety' by Mair 2013), Pakistan and Sri Lanka – comprising the following internet-based registers: newspaper comments, tweets, web forums and text messages.[2] For reference I also include the same registers – except text messages – from two main IC varieties (Kachru 1985), i.e. US English – which is described as the 'hyper-central variety' by Mair (2013) – and UK English – which is historically relevant and termed a 'super-central variety' by Mair (2013).

The main aim of the study is to explore register variation in the selected online registers of the above-mentioned four varieties of English in reference to the two IC varieties. More specifically, the first aim is to find linguistic variation in the form of co-occurring linguistic patterns or dimensions in the four interactive CMC registers, i.e. newspaper comments, tweets, web forums and text messages. The second aim is to explore regional variation among corresponding interactive registers in the six countries based on these dimensions. The last aim is to find out if there are any regional differences in South Asian text messages.

The article includes the following sections. I present a brief literature background on World Englishes, MD analysis and MD studies on online registers in the next section. The MD analysis is described along with data, methodology and results in section 3. Section 4 concludes the article.

2 Background

2.1 Register variation, World Englishes and CMC

Register is an important driver of variation. However, register analysis is not equally incorporated in every study of World Englishes. Bohmann (2019: 32–5) – building on Neumann's (2014) categorisation – distinguishes three kinds of studies according to the level of incorporation of register: register monolithic; register controlled; and the ones focusing on register as the main object of inquiry. The first two types consider only single or a few features and apply detailed and contextualised attention to them, while the third type takes a feature-aggregate approach (Bohmann 2019: 34). MD analysis falls under the third category, where the normalised frequencies of a large set of

[2] Text messages could only be collected for the South Asian countries.

functionally relevant lexico-grammatical features are obtained. Then, co-occurring patterns of these features – called 'dimensions' – are discovered using statistical techniques and functionally interpreted in the light of the situational context and linguistic usage of the registers under study (cf. Biber 1988; Biber & Conrad 2019). There have been a few comprehensive MD studies on traditional registers in the context of World Englishes (see section 2.3), but research on online registers has been less common.

In the last few decades, the internet has created new avenues of communication for users. At the same time, this has presented opportunities as well as challenges for researchers in the field of sociolinguistics and World Englishes. At the simplest level, online registers provide a treasure trove of ready-to-use data for small- and large-scale studies (Bohmann 2019: 35). In particular, interactive and informal online registers, which are mediated but increasingly intertwined with the real-life of netizens, supply vernacular data for the study of sociolinguistic phenomena (Heyd & Mair 2014: 249). This is especially important for smaller varieties, e.g. Bangladeshi English, where it is hard to collect data in traditional domains. Since online communication has a global as well as a local aspect (Bohmann 2019: 37), internet-based registers can be used as an alternative to traditional registers. A study of the same set of registers from multiple varieties of English, e.g. those from South Asia, using MD analysis can uncover (communicative) functions associated with English in conjunction with associated features in a holistic and comparative manner across registers and varieties. This can help us to understand the development of these varieties – individually and in the South Asian region as a whole – on the internet. Further, online data can serve as a proxy for offline situations, but with certain caveats. In fact, the use of a multi-register CMC corpus – as opposed to a dataset like GloWbE (Davies & Fuchs 2015) where CMC registers are not as clearly defined (cf. Loureiro-Porto 2017) – can also improve the World Englishes community's understanding of variation in internet registers, which, in turn, can help clarify whether and to what extent different online registers are valid proxies for offline registers.

2.2 Previous relevant MD studies on online registers

In the following, the findings of relevant MD studies on online registers – i.e. comments, tweets, chats and web forums – are discussed.

Ehret & Taboada (2020, 2021) examine Canadian newspaper comments by conducting new MD analyses. They find that newspaper comments are nearer to written registers such as exams, opinions and social letters than spoken registers. In comparison to the online registers of Biber *et al.* (2015), online news comments are informational, argumentative and also include evaluative discourse. They can range from very short to quite long argumentative and encyclopaedic types of texts (Ehret &Taboada, 2021: 8). In contrast, YouTube, ESPN and Yahoo! comments have been found to be involved, personal and opinionated (Titak & Roberson 2013: 255).

Twitter discourse has been included in a number of MD studies. Titak & Roberson (2013) use blogs, emails, newspaper articles, opinion columns, comments (see above) and Twitter/Facebook posts merged in one category. As per the new MD analysis, Twitter and Facebook posts are nominal and informational with very limited narrativity (Titak & Roberson 2013: 254). Friginal *et al.* (2018: 358) use the dimensions identified in Titak & Roberson (2013) to more closely examine Twitter and Facebook posts related to weather, politics, sports and personal topics. Their analysis shows that generally tweets are less personal, less narrative, less involved and less interactive than Facebook posts (Friginal *et al.* 2018: 358). The authors cite length restrictions in Twitter as one of the possible reasons for this type of discourse. Berber Sardinha (2014) in his study comparing online registers to non-internet registers using additive MD analysis – where 'old' dimensions are applied to a new corpus (Berber Sardinha 2014: 81) – based on Biber's (1988) dimensions demonstrates that online registers related to personal communication like Twitter have similarities to spontaneous speeches, interviews and personal letters. Finally, Berber Sardinha's (2022: 665–9) analysis of Twitter along with Facebook and Instagram posts reveals that tweets have a comparatively higher score on dimension 2 'Informal, interactive and speaker-oriented discourse'.

One of the registers most closely related to text messages, i.e. ICQ chats, has been compared with traditional spoken and written registers using Biber's (1988) dimensions of variation (Jonsson 2015). The study's findings indicate that ICQ chats are just like face-to-face and telephonic spoken conversations on dimension 1 – i.e. involved, interactive and affective – but they are comparatively less narrative on dimension 2 'Narrative versus non-narrative production'. Otherwise, ICQ chats are similar to the above-mentioned registers, in that they have situation-dependent reference, no overt expression of persuasion and non-abstract information on dimensions 3, 4 and 5 (Jonsson 2015: 292).

Collot & Belmore (1996) conduct an additive MD analysis using Biber's (1988) dimensions on message boards, an earlier form of web forums. The authors note that message boards are highly involved and relatively situation-dependent due to the shared interest of the participants, despite non-shared time and space. In their opinion, the communicative purposes of giving information and discussing specific issues mean that message boards have a non-narrative and highly persuasive discourse. Lastly, they reason that the similarity between message boards and public interviews could be due to the social roles of addressor, addressee and audience (Collot & Belmore 1996: 26).

2.3 Previous relevant MD analyses on register and regional variation

MD studies focusing on geographical variation use different combinations of traditional registers or traditional and online registers. The two most comprehensive studies have been conducted by Xiao (2009), using registers from the *International Corpus of English* (ICE) – private and public dialogues, scripted and unscripted monologues, unpublished writing such as student essays and personal letters, published writing such

as academic writing, popular writing such as magazines, creative writing, press reportage etc. – and Bohmann (2019), who uses ICE registers along with Twitter discourse. Xiao's (2009) MD analysis shows that Indian English is the most elaborate on dimension 1 'Interactive casual discourse versus informative elaborate discourse'. On dimension 5 'Future projection', Indian English scores the lowest on all registers (Xiao 2009: 447). Bohmann (2019) confirms Xiao's findings, i.e. Indian English is among more informational varieties on dimension 1 'Involved versus informational production', which is true for all mediums, i.e. spoken, written and Twitter (Bohmann 2019: 109). Similar observations have been made by Hussain's (2016) study on an incomplete ICE corpus of Pakistani English. Lastly, Kruger & van Rooy's (2018: 237) analysis of written ICE registers finds that less advanced OC varieties like Indian English generally avoid features related to informality and involvement.

Online registers of OC varieties show similar differences when compared with a native English register. For example, Hardy & Friginal (2012: 159) observe that Philippine English blogs typically deviate 'from personal and involved stylistic features of informal writing such as blogs written by American authors'. Similarly, Shakir & Deuber (2018, 2019) note that Pakistani online registers are informational and restrictive in communicative purposes compared to their US English counterparts.

Overall, whilst there have been comprehensive MD studies on traditional registers across several IC and OC varieties, there has not yet been an examination of multiple CMC registers with a focus on OC varieties from the South Asian region. The present analysis, thus, fills this research gap by conducting an exploratory study by applying MD analysis on four interactive online registers, i.e. newspaper comments, tweets, web forums and text messages, from the six varieties of English (Bangladesh, India, Pakistan, Sri Lanka, UK, USA). Based on previous literature, it is expected that the four South Asian varieties will tend to have formal, literate and informational discourse in these registers.

3 Dimensions of variation in interactive online registers

3.1 Data collection and preprocessing

The data for the present study have been sampled from the SAOnE corpus (Shakir & Deuber 2023). The down-sampling was necessary for better preprocessing, especially checking non-standard spellings (see below). Table 1 presents the corpus structure, brief descriptions of the sources, and the sample included in this analysis. In general, I have sampled 50 files per category, which is 10 per cent of the total of 500 texts in most categories. Text messages are an exception as they have only half the word count (500,000 words) with conversations sometimes larger than 50,000 words. These files were sliced into 2,000-word texts and then 50 files were sampled for each country. The total files included in the study are 50 texts * 6 countries * 4 categories = 1,100. There are no data for text messages from the UK and USA.

Table 1. *Description of corpus categories in the main corpus with texts included in this study (adapted from Shakir & Deuber 2023: 123)*

Register category	Description type	Detail
Comments (CMT)	Sources	Newspaper comments (2 newspapers per country from 4 SA countries + USA & UK).
	Text composition	Comment threads combined in ≈2,000-word files (1 million words per country).
	Current study	50 texts (≈100,000 words) per country using stratified random sampling.
Discussion forums (WBF)	Sources	Discussion threads from public English forums, at least 5 web forums per country on topics like cars, gaming, cricket, defence, technology, study, city, transport, finance etc. from 4 SA countries + USA & UK.
	Text composition	Forum threads combined in ≈2,000-word files (1 million words per country).
	Current study	50 texts (≈100,000 words) per country using stratified random sampling.
Tweets (TWT)	Sources	English tweets originating from 4 SA countries + USA & UK as provided by Twitter API.
	Text composition	1 user/file in ≈2000-word files (1 million words per country).
	Current study	50 texts (≈100,000 words) per country using stratified random sampling.
Text messages (TXM)	Sources	Mostly English group and 1–1 chats on topics like study, work etc. from various messaging platforms (WhatsApp, Facebook Messenger, Telegram, Upwork…) originating from 4 SA countries.
	Text composition	Complete conversations per file, spanning from a few hundred to 100,000 or more words (500,000 words per country).
	Current study	50 texts (≈100,000 words) per country (larger files were sliced into 2,000-word smaller text files and then selected using stratified random sampling).

The preprocessing consisted of three steps. The first step was to check for non-standard spellings to improve the grammatical tagging process in the later stages. For this purpose, I have used the spellchecking module syplls in Python to help the grammatical tagger (Shepelev 2022; Python Core Team 2022). The corrected spellings were checked by the project assistants to reverse any mistakes of the automatic spelling correction. For example, *Ambar* is a proper noun, which was falsely replaced with *Mbar* by the spellchecker. The original word was restored in the review process. The data included in SAOnE were also annotated for indigenous content present in the form of single words, multiword expressions and sometimes full posts (i.e. tweets, comments, web forums, text messages). In the second step multiword phrases and clauses of

indigenous content present in the South Asian data were removed. These expressions would be tagged either as foreign words (FW) or mistagged by the grammatical tagger. Both of these outcomes were not helpful for the present analysis. The third step was hashtag segmentation. Hashtags in tweets can consist of multiple words which can include nouns and adjectives – e.g. *#Trainingprovider* – or even clauses – e.g. '#ILOVENYC' (Berber Sardinha 2022: 659). Following Berber Sardinha, I have used the Python module wordsegment (Jenks 2018) to split hashtags – e.g. *Training provider*, *I LOVE NYC* – to help the grammatical tagger in identifying word classes.[3]

3.2 *Grammatical tagging and MD analysis*

Before describing the actual MD analysis and results, I address three issues related to MD analysis. The first issue is related to the selection of features. Generally, the most popular software used for tagging MD related features is BiberTagger (also used in Biber 1988). The features tagged by BiberTagger are a combination of lexico-grammatical features introduced in the original 1988 study and more detailed semantic subclasses based on Biber *et al.* (1999) and later works such as Biber (2006). Biber's (1988) 67 linguistic features can also be tagged by Multidimensional Analysis Tagger (MAT), which was developed by Nini (2014, 2019) and is now available as an opensource tool. This tool has been modified by Le Foll (2021a) and released as a separate tagger. The main changes introduced by Le Foll are to make the tagging process more reliable and reduce the need for manual correction as much as possible. For this reason, she excludes features that have very low precision and recall. Some examples include present and past participial clauses, present and past participial WHIZ deletion relatives, *that* relative clauses on subject and object positions, *WH* relative clauses on subject and object positions, pied-piping relative clauses, sentence relatives, phrasal and clausal coordination etc. The features listed above have been replaced with a smaller set of features, namely *that* complement clauses, *that* relative clauses, *WH* subordinate clauses, nonfinite *-ing* forms of verb, nonfinite *-ed* forms of verb, coordinating conjunctions etc. Additionally, she includes some semantic verb classes from Biber (2006), e.g. activity verbs, communication verbs, mental verbs etc. (see link for the complete list of features included). I have selected Le Foll's feature set because it has over 90 per cent accuracy (Le Foll 2021b: 34) including on internet-related data, and because manual correction of each feature is a very time consuming task for large number of texts, as for example included here.

I have further expanded Le Foll's (2021a) feature set by including Biber's (2006) and Biber *et al.*'s (1999) features related to semantic classes of nouns, verbs, adjectives and

[3] As pointed out by an anonymous reviewer, hashtags perform a multitude of functions in social media texts and are sometimes embedded in the syntax in such a way that simply segmenting them would not be the solution for each case. The present approach follows Berber Sardinha (2022), who also segments hashtags in social media texts. However, the limitation of this method is also acknowledged here. Certainly, a better solution is needed, e.g. a manual analysis and segmentation of only non-embedded hashtags. This was unfortunately not possible due to the limited time and resources available for this project.

adverbs. Apart from vocabulary items, there are *that*, *to* and *WH* clauses that can be preceded by different types of nouns, adjectives and verbs, for example *that* complement clauses preceded by communication verbs. I have modified Le Foll's script 'Multi-Feature Tagger of English' or MFTE to include these semantic subtypes in the general categories of nouns, verbs, adverbs, adjectives and *that*/*WH*/*to* clauses. The combination of Le Foll's feature set and my modifications enables me to tag for more than one hundred lexico-grammatical features that are very similar to the output of BiberTagger as they are mostly defined and operationalised following Biber's work. Due to space restrictions I cannot list definitions and examples of all these features here. However, I have put online a supplementary document defining each feature (link). The individual features will be mentioned and contextualised as and when they appear in the analysis below.

The second issue is related to the normalisation unit for the absolute frequencies that will later be used as the input table for MD analysis. Traditionally MD studies use a per-thousand-words normalisation approach for all lexico-grammatical features. However, some authors argue that word-based normalisation may not be suitable for every feature. For example, Evert (2022: 27) points out that passive voice is essentially dependent on total number of verbs instead of total number of words in a text. Moreover, word-based normalisation can create confusion between 'covariation of features due to situational variation', which is desired in MD analysis, as opposed to 'covariation due to grammatical structure', which is undesired (Le Foll 2022: 285) – hence prepositions will highly correlate with nouns due to the grammatical constraints of prepositional phrases.

One possible solution for the above is to normalise features, where possible, based on the grammatical feature they depend on mostly – hence prepositions normalised on total number of nouns. Le Foll's (2021b) 'complex normalisation approach' is one such attempt (see footnote).[4] Though it has its own shortcomings, e.g. it is not always clear which feature is the appropriate normalisation baseline for a feature like adjectives. My experiments with both normalisation methods on the present data show that the former (word-based normalisation) results in slightly larger dimensions with some additional features – especially on the first two dimensions –compared to the latter ('complex-normalisation'). Otherwise, the interpretation of the dimensions remains the same due to mostly identical features. I adopt the latter for the present study, as both methods result in conceptually similar dimensions for the present data and the latter, to a certain extent, suppresses grammatical covariation.

[4] Features that are verbal in nature or depend on verbs are normalised based on per hundred finite verbs, for example voice, aspect, tense, complement clauses, semantic subtypes of verbs, conditional conjunctions, contractions, negation etc. Nominal features like compound nouns, nominalisations, attributive adjectives, relative clauses, prepositions, semantic subtypes of nouns, semantic subtypes of adjectives etc. are normalised based on per hundred nouns. Features like discourse markers, total remaining nouns etc. are normalised per 100 words (cf. Le Foll 2021b, 2022: 283–4).

The third issue is related to the method applied to extract the groups of co-occurring lexico-grammatical features or dimensions. Traditionally, exploratory factor analysis (EFA) has been used in MD studies, including in the pioneering study Biber (1988). The researcher needs to decide the number of factors that should be extracted. Additionally, the factors are rotated to create 'simple structure' to help interpret the latent constructs in the data (Bohmann 2019: 89; Le Foll 2022: 289). An oblique rotation like 'promax' is recommended in MD studies as it allows the factors to be correlated just like everything in language is also correlated with each other (Biber 1988: 85). However, orthogonal rotations like 'Varimax', which do not allow correlation among factors, have also been used in MD studies such as Grieve *et al.* (2010: 309).

PCA is a similar dimensionality reduction technique that has also been used in MD studies, e.g. Biber & Egbert (2016). It has been suggested as an alternative to EFA as it can be more objective by eliminating the need to determine the number of dimensions beforehand (Evert 2022: 31). Just like EFA, PCA can also be rotated, as for example Biber & Egbert (2016) do. This, however, reintroduces the issue of determining the number of dimensions to be extracted and the composition of resulting components – the first and second components in a two-component rotated solution will be different from the first two components in a three-component rotated solution (Le Foll 2022: 290). Moreover, the rotated principal components 'no longer successively account for maximum variance' (Rencher 1992: 222). Considering the above and my experiments with various combinations of rotated and unrotated PCA solutions, I have used unrotated PCA for the present analysis as the resulting dimensions are easier to interpret.[5]

The data have been tagged using Stanford Tagger (Toutanova *et al.* 2003) and as the second step by the modified MFTE script as described above.[6] I have utilized psych:: principal() in R (Revelle 2022; R Core Team 2022) for PCA. To handle outliers in a better way, I also used *z* score transformation and then signed-log transformation as recommended by Neumann & Evert (2021: 155). Before the application of PCA, I have removed features having 0 occurrences in 66 per cent or more files as well as features with less than |0.1| correlation with all other features, because such features are most probably less relevant to the data under study. Very high correlations can skew the results (Le Foll 2022: 295) so I have excluded features with very high correlation (>|0.9|) with any other feature. As per Egbert & Staples' (2019: 129) recommendation, features having communalities lower than 0.2 have been removed in this analysis. Kaiser–Meyer–Olkin (KMO) is a measure of sampling adequacy and its value lower than 0.5 means that the data are not appropriate for MD analysis (Egbert & Staples 2019: 129). It can be calculated for individual features as well for the whole data. I checked for individual features having a KMO less than 0.5 and removed them.[7] The

[5] More research, especially in the context of MD studies, is definitely needed to further clarify these issues.

[6] Since my additions to MFTE build on Le Foll's original algorithms, approximately same level of accuracy, i.e. over 90 per cent, is expected.

[7] Only one feature 'present tense verbs' having a KMO lower than 0.5 (0.35) was preserved, as it is linguistically relevant to the present data.

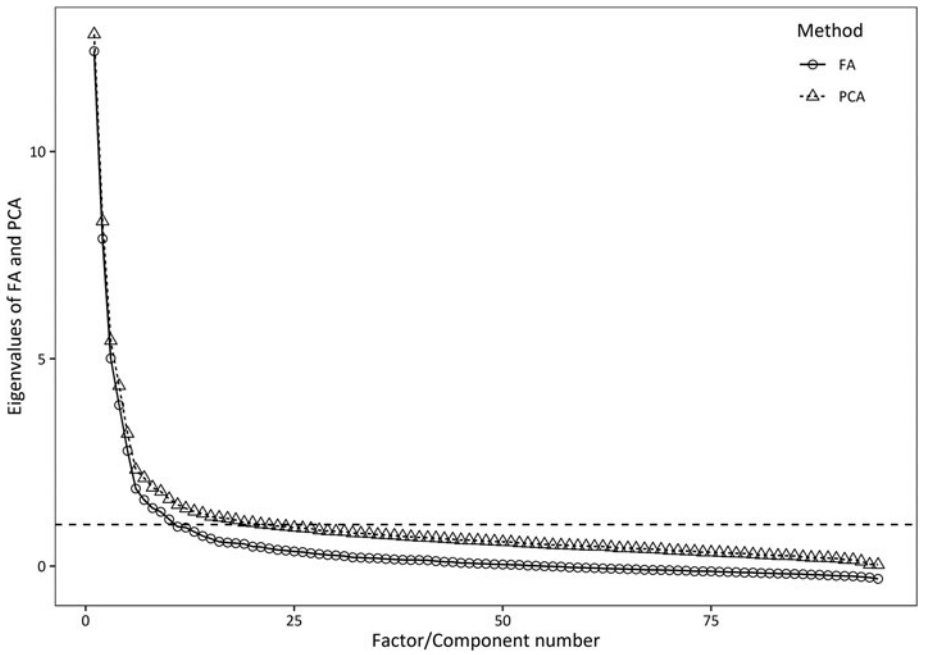

Figure 1. Scree plot of the data (FA and PCA)

final data table has a KMO of 0.85 (=Meritorious) with 80 linguistic features. Five dimensions of variation were retained in the final PCA as the scree plot flattens after five dimensions in both PCA and EFA (see figure 1). The total variance explained by the model is 40 per cent, which is normal for MD studies: e.g. Biber & Egbert (2016: 12) report 42.7 per cent variance accounted for. Dimension scores have been obtained using the regression method, which is the default in the principal() function.

Before presenting the results in the next section, a brief explanation of the statistical methods is given here. Dimension scores do not fulfil the assumptions of classic ANOVA, i.e. the within-group variance is not equal and the distributions are not normal. Therefore, a robust regression model was fitted using MASS::rlm (Venables & Ripley 2002). Then, type III ANOVA tables were generated using car::Anova (Fox & Weisberg 2019). Group comparisons were run on the ANOVA object using the emmeans package at 95 per cent confidence level (Lenth 2023). The value of R^2 gives an indication of the importance of a dimension as it shows how much variance in the dimension scores of the given dimension can be explained by knowing the text categories (Biber & Egbert 2016: 12). The text groups in this case consist of register as well as country-based groupings. The R^2 was generated using the interaction term of country and register in lm() in base R as rlm() estimated higher values.

The figures contain information boxes. Apart from the significance of predictors and R^2 values, the information box shows significant emmeans compared to 'Both' of the IC varieties or 'UK' or 'US'. 'Both:rel. group' indicates that the relevant register in both

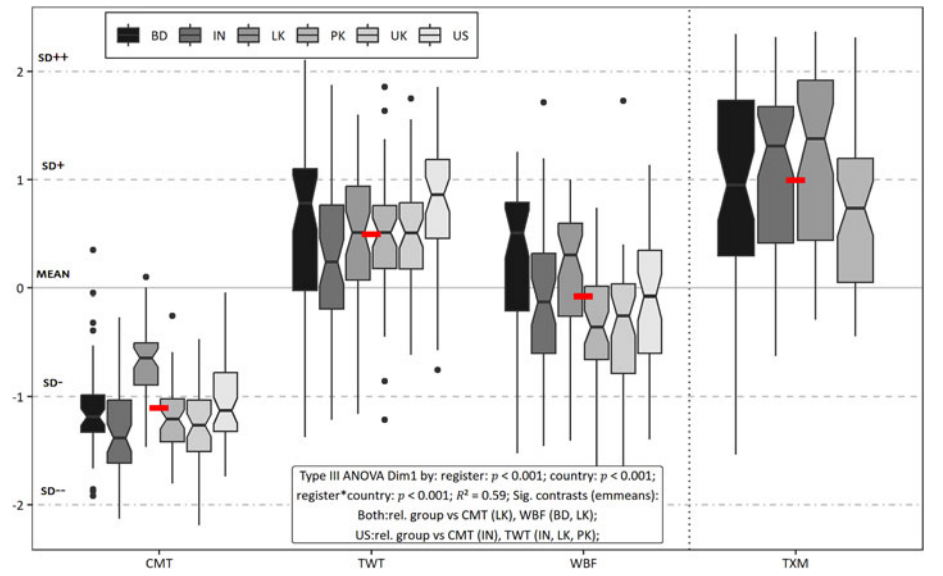

Figure 2. Distribution of registers on Dim1 'Oral personal (top) versus literate impersonal discourse (bottom)'

Note: Country names are two-letter ISO codes: IN is India; BD is Bangladesh; PK is Pakistan; LK is Sri Lanka; UK is United Kingdom; and US is United States of America. Register codes are as follows: CMT for comments; TWT for tweets; WBF for web forums; and TXM for text messages.

IC varieties is significantly different from the registers in South Asian countries shown in brackets. Hence line 3, entry 1 in figure 2 shows that Sri Lankan comments are significantly different from both US and UK comments. Indian comments are only significantly different from US comments (line 4 first entry) and so on. Red dashes show overall register means.

3.3 Dim1 'Oral personal versus literate impersonal discourse'

Dim1 consists of 40 variables, of which 10 occur on the positive side and 30 on the negative side, as table 2 reveals. The positive features are related to involved and informal discourse, e.g. pauses and interjections, contractions, pragmatic markers and politeness markers, as in (1). Non-standard spellings and internet slang are used in informal CMC. There are features related to interactive and personal communication like first- and second-person pronouns and imperatives. The features on the negative side are related to impersonal, nominal and literate discourse, which include prepositions, passives, nominalisations, nouns like *abstract process*, *cognitive* and *place*, non-finite verb *-ed* and *-ing* forms – which may or may not introduce non-finite clauses as in (2), high ratio of content to function words, high type token ratio and longer words. Features like attributive adjectives and determiners modify nouns and make the noun phrase longer and complex. These nouns are most probably not proper

Table 2. *Dimensions and features of MD analysis*[a]

Dim	+/−	Features
Dim1	+	FPUH - Filled pauses and interjections (0.74)
	Oral personal	NSS - Non-standard spellings (0.67)[b]
	vs	FPPAll - All first-person pronouns (0.63)
		SPP2 – Second-person pronouns (0.6)
		DMA - Discourse/pragmatic markers (0.59), e.g. *well*, *right*
		POLITE - Politeness markers (0.56), e.g. *please*
		SLNG - Internet slang like LOL (0.53)
		CONT - Verbal contractions (0.5)
		VIMP - Imperatives (0.49), e.g. *Sit* down.
		TIME - Time references (0.31)
	−	INother - Prepositions having no additional tag (−0.77)
	Literate impersonal discourse	JJATother - Attributive adjectives having no additional tag (−0.77)
		LDE - Lexical density (−0.68), ratio of content to function words
		NOMZ - Nominalisations (−0.64)*
		DT - Determiners (−0.63), e.g. *a*, *an*, *the*…
		PASSAll - All *be* and *get* passives (−0.63)
		TPP3P – Third-person plural pronouns (−0.6)
		CC - Coordinating conjunctions (−0.56)
		THSCother - *that* complement clauses not preceded by a stance adjective or verb (−0.55)
		PrepNSTNC - Prepositions preceded by stance nouns (−0.53)*
		JJTOPIC - Topical adjectives (−0.51)*
		SPLIT - Split auxiliaries and infinitives (−0.51)
		PEAS - Perfect aspect (−0.51)
		NNGRP - Nouns group (−0.49)*
		THRC - *that* relative clauses (−0.48)
		WHSCother - *WH* complement clauses not preceded by a stance verb (−0.44)
		NNCOG - Nouns cognitive (−0.44)*
		MDNE - Necessity modals (−0.42)
		EXIST - Existential or relationship verbs (−0.41)
		NNABSPROC - Nouns abstract and process (−0.4)*
		ELAB - Elaborating conjunctions (−0.39), e.g. *similarly*, *for example*
		OCCUR - Occurrence verbs (−0.38)
		AWL - Average word length (−0.38)
		EX - Existential *there* (−0.38)
		MDWO - modal *would* (−0.37)
		ThVSTNCother - *that* subordinate clauses followed by stance verbs other than communication verbs (−0.37)*
		AMP - Amplifiers (−0.36)
		TTR - Type token ratio (−0.35)

(*Continued*)

Table 2. (*continued*)

Dim	+/−	Features
		NNPLACE - Nouns place (−0.33)*
		VBG - Non-finite verb *-ing* forms (−0.33), e.g. *concerning* your request…
Dim2	+	QUAN - Quantifiers (0.7)
	Oral elaboration	RBother - Adverbs having no additional tag (0.66)
	vs	EMPH - Emphatics (0.57)
		CONT - Verbal contractions (0.54)
		DOAUX - DO auxiliary (0.54)
		DT - Determiners (0.52)
		HDG - Hedges (0.51)
		THATD - Subordinator *that* omission (0.47)
		DEMO - Demonstratives (0.45)
		XX0 - Negation (0.43)
		JJEVAL - Evaluative adjectives (0.38)*
		ToSTNCAll - All *to* infinitive clauses preceded by stance adjectives, nouns and verbs (0.38)*
		RSTNCAll - All adverbs related to stance (0.38)*
		JJSIZE - Size related adjectives (0.38)*
		QUPR - Quantifying pronouns (0.37), e.g. *anybody*, *anyone*, *anything*…
		PIT - pronoun *it* (0.37)
		DMA - Discourse/pragmatic markers (0.37)
		MDWO - modal *would* (0.36)
		YNQU - Yes/no questions (0.36), e.g. *Do* you mind*?*
		SLNG - Internet slang like LOL (0.35)
		COND - Conditional conjunctions (0.35)
		JJREL - Relational adjectives (0.33)*
		NNHUMAN - Nouns human (0.32)*
		NNCOG - Nouns cognitive (0.31)*
		MENTAL - Mental verbs (0.3)
	−	NNother - Nouns not in a semantic class, including proper nouns (−0.71)
	Informational concerns	VBN - Non-finite *-ed* verbforms (−0.6) , e.g. *provided* that…
		AWL - Average word length (−0.59)
		VBG - Non-finite verb *-ing* forms (−0.39)
		VIMP - Imperatives (−0.38)
		ACT - Activity verbs (−0.36)
		CC - Coordinating conjunctions (−0.3)
		TIME - Time references (−0.32)
Dim3	+	NNABSPROC - Nouns abstract and process (0.63)*
	Persuasion- and help-oriented	NNTECH - Nouns technical (0.52)*
		MDWS - *will* and *shall* modals (0.48)
	vs	MDPOSSCAll - All modals of possibility (0.42)
		MDNE - Necessity modals (0.38)
		POLITE - Politeness markers (0.38)
		JJREL - Relational adjectives (0.35)*

(*Continued*)

Table 2. (*continued*)

Dim	+/–	Features
		CAUSE - Facilitation and causative verbs (0.35)
		NOMZ - Nominalisations (0.33)*
		COND - Conditional conjunctions (0.32)
		COMM - Communication verbs (0.32)
		NNHUMAN - Nouns human (0.31)*
		YNQU - Yes/no questions (0.3)
	–	NNother - Nouns not in a semantic class, including
	Informational and	proper nouns (−0.52)
	past-oriented discussions	VBD - Past tense (−0.48)
		RP - Verb particles (−0.45), e.g. I'll look it *up*
		FREQ - Frequency references (−0.41), e.g. *usually*,
		always, *mainly*, *often*…
		EMPH - Emphatics (−0.37)
		CONC - Concessive conjunctions (−0.34)
		CONT - Verbal contractions (−0.31)
		TTR - Type token ratio (−0.3)
Dim4	+	VPRT - Present tense (0.44)
	Focus on humans and mental	NNHUMAN - Nouns human (0.44)*
	processes	AWL - Average word length (0.43)
	vs	BEMA - *be* as main verb (0.42)
		WHQU - Direct WH-questions (0.37), e.g. *What's*
		happening*?*
		JJTOPIC - Topical adjectives (0.33)*
		MENTAL - Mental verbs (0.3)
	–	NCOMP - Noun compounds (−0.5)
	Non-abstract things and	NNCONC - Nouns concrete (−0.49)*
	activities	ACT - Activity verbs (−0.45)
		RP - Verb particles (−0.35)
		NNQUANT - Nouns quantity (−0.34)*
		MDWO - modal *would* (−0.33)
		HDG - Hedges (−0.32)
Dim5	+	JJPRother - Predicative adjectives having no additional
	Addressee involved opinion	tag (0.51)
	and description	VPRT - Present tense (0.5)
	vs	SPP2 – Second-person pronouns (0.35)
		AMP - Amplifiers (0.33)
		BEMA - *be* as main verb (0.32)
		MENTAL - Mental verbs (0.3)
	–	VBD - Past tense (−0.51)
	Narrative tendencies	TPP3S – Third-person singular pronouns (−0.5)

[a]Features different from Biber (1988) have examples in front of them (see also the supplementary feature description file). Features marked with * are my additions based on Biber (2006) or BiberTagger (using Biber *et al.* 1999). Semantic verb classes from Biber (2006) are adopted by Le Foll. Cut-off point is |0.295| (rounded to |0.3|). Dimension 1 was inverted for better interpretation.
[b]Non-standard spellings were calculated separately before replacing them with standard spellings.

nouns, which do not take determiners in English. Moreover, clauses like *that* relatives and *WH* clauses can be used to add further information. Lastly, the presence of third-person plural pronouns indicates that the discourse focuses on third parties. Based on these traits, the first dimension, as in many previous MD studies, can be labelled 'Oral personal versus literate impersonal discourse'.

(1) So **yeah** [**DMA**].
<#>
. **Ah** [**FPUH**].
<#>
Ye ye.
<#>
. That blue colour uniform.
<#>
. When did **you** [**SPP2**] change.
<#>
Yea [**FPUH**] **I** [**FPPAll**] was in Vidura before!! After o/ l..
<#>
Okay [**DMA**] so that makes **you** [**SPP2**] bright.
<#>
. Hyatt.
<#>
. **No** [**DMA**] **I** [**FPPAll**] '**m** [**CONT**] not.
<#>
Yeah [**DMA**] **you** [**SPP2**] are.
<#>
. How many As did **you** [**SPP2**] get for o/ l?
<#>
O/ l results?
<#>
3as.
<#>
Wow [**FPUH**] nice. (Dim1 Oral personal; LK-TXM162)

(2) Start your **own** [**JJATother**] domain **name** [**NNABSPROC**] **and** [**CC**] **hosting** [**VBG**] **company** [**NNGRP**]! **Free** [**JJATother**] billing **and** [**CC**] tech **support** [**NNABSPROC**] 24/7, actually, it 's **the** [**DT**] same **company** [**NNGRP**] **that** [**THRC**] sells godaddy **their** [**TPP3P**] reseller program! You literally have **the** [**DT**] same products **and** [**CC**] **support** [**NNABSPROC**], **they** [**TPP3P**] just answer **the** [**DT**] phone **as** [**INother**] "customer **support** [**NNABSPROC**]". **But** [**CC**] feel safe your reseller program is **managed** [**PASSAll**] **by** [**INother**] **the** [**DT**] same great **support** [**NNABSPROC**], **security** [**NNABSPROC**] **and** [**CC**] **reliability** [**NOMZ**] **as** [**INother**] **the** [**DT**] big dogs! […] (Dim1 Literate impersonal; PK-WBF015.txt)

Scores on Dim1 are the most sensitive to the groups of registers and countries included in these data, as shown by the R^2 value of 0.59 or 59 per cent – hence a good discriminator of these text categories. As figure 2 shows, text messages have the highest score on the positive side and comments are the most literate register. Tweets are second on the oral side, while web forums are slightly literate. The Bangladeshi tweets and web forums have comparatively higher scores in South Asia due to oral features like filled pauses, non-standard spellings (not shown in the example), first- and second-person pronouns, politeness markers etc. See (3) for forum posts where someone is being wished a happy birthday and is answering someone else's post, and (4) for a couple of tweets (fan addressing two famous twitter accounts):

(3) **ahhh** [**FPUH**] **I** [**FPPAll**] see **now** [**TIME**] … Happy Birthday.
<#>
Happy Birthday …….
<#>
Thank [**POLITE**] **you** [**SPP2**] very much to all of **my** [**FPPAll**] buddies. **Your** [**SPP2**] wishes mean a lot to **me** [**FPPAll**]. (3 forum posts from BD-WBF336)

(4) @X **Please** [**FPUH, POLITE**] take care.
<#>
@XxxxxxXxxxxxxxx **I** [**FPPAll**] love this short film !! That'**s** [**CONT**] so inspirational (2 tweets from BD-TWT184)

A comparatively higher presence of the same type of features in the Sri Lankan comments makes them more oral personal compared to the other South Asian comments and the UK and USA. Apart from that, the Indian comments and tweets, and Pakistani and Sri Lankan tweets are significantly more literate in comparison to the relevant categories in UK and USA. Lastly, more South Asian registers have significant differences from their US counterparts (as shown by line 4 in the information box of figure 2) compared to the corresponding UK registers.

3.4 Dim2 'Oral elaboration versus informational concerns'

Dim2 has a total of 33 features with 25 on the positive and eight on the negative side. Some features on the positive side point to an oral and informal discourse, e.g. contractions, auxiliary *do*, *that* omission, demonstratives, discourse markers, internet slang etc. Features like adverbs (general, stance-related, hedges, emphatics), mental verbs like *wonder* and *appreciate*, *to* infinitives preceded by stance words, adjectives (evaluative, size, relational), modal *would* etc. can be used to talk about or express opinions/stance about people and things using human nouns and pronouns (quantifying, *it*). The presence of determiners on the positive side – like the negative side of Dim1 – shows that most of the nouns used here are not proper nouns, hence they require articles and other determiners. For example, a user is expressing their personal opinion about Covid problems in a comment in (6). Thus, the positive side can be labelled 'Oral

elaboration' like Biber & Egbert's (2016) dimension 2. The most important feature on the negative side is general nouns that do not need a determiner, i.e. proper nouns. These long account handles (hence higher average word length) are notable in (7), which is from the account of a Pakistani political worker who tags many party accounts in their tweet. Occasionally such tweets have other features like non-finite *-ed* verbs, non-finite *-ing* verbs, imperatives, or time references (e.g. *today*). The use of time references and imperatives can be seen in (5), which shows an advertisement-related Indian tweet. Hence, the appropriate title for the negative side can be 'Informational concerns'.

(5) **Now** [**TIME**] **Book** [**VIMP**] reliable **and** [**CC**] Safe **Cab** [**NNother**] Service across **Jharsuguda** [**NNother**] at an affordable **price** [**NNother**]. (IN-TWT217)

(6) Since **the** [**DT**] pandemic started, I **do** [**DOAUX**] n'**t** [**XX0, CONT**] **feel** [**MENTAL**] that **much** [**QUAN**] of **a** [**DT**] **need** [**NNCOG**] to exercise (**even** [**RBother**] though I **still** [**RBother**] **do** [**DOAUX**]), **especially** [**RBother**] since school is online, and **nobody** [**QUPR**] **really** [**EMPH**] **sees** [**MENTAL**] **more** [**QUAN**] than your face on **a** [**DT**] screen. **This** [**DEMO**] is **a** [**DT**] **good** [**JJEVAL**] thing for **lots** [**QUAN**] of **people** [**NNHUMAN**] who sometimes **feel** [**MENTAL**] pressured to have **a** [**DT**] "perfect" body, including myself. (Dim2 Oral elaboration; US-CMT333)

(7) Many many returns of the day Happy **birthday** [**NNother**] chairman **sb** [**NNother**] **Allah** [**NNother**] bless you **@XXxxxxxXxxxxxx** [**NNother**] **@XXXXXXXXXXXXXXX** [**NNother**] **@XxxXxxxxx** [**NNother**].

The second dimension is rather less sensitive in discriminating the text groupings in these data as R^2 value based on the interaction of country and register is only 0.26 (26 per cent). The red dash in figure 3 reveals that web forums have the highest overall dimension score compared to other registers. Comments and text messages are also slightly inclined towards the 'oral-elaboration' pole of Dim2. Tweets are mostly informational, which is partly due to the presence of proper nouns in the form of @ mentions. Comments, tweets and web forums also show clear regional trends, with the South Asian countries being significantly less oral-elaborative – i.e. having fewer oral and stance-related features – in comments and web forums, and more informational in tweets compared to both native varieties. Text messages are almost in the middle, with Pakistani text messages inclining towards the informational side probably due to communicative purposes like study/work-related information sharing (see section 3.8). The comparatively lower number of stance-related features in the South Asian comments shows that these comments employ different features for this purpose. The communicative purpose of information sharing appears to be more relevant for South Asian tweets as well as web forums.

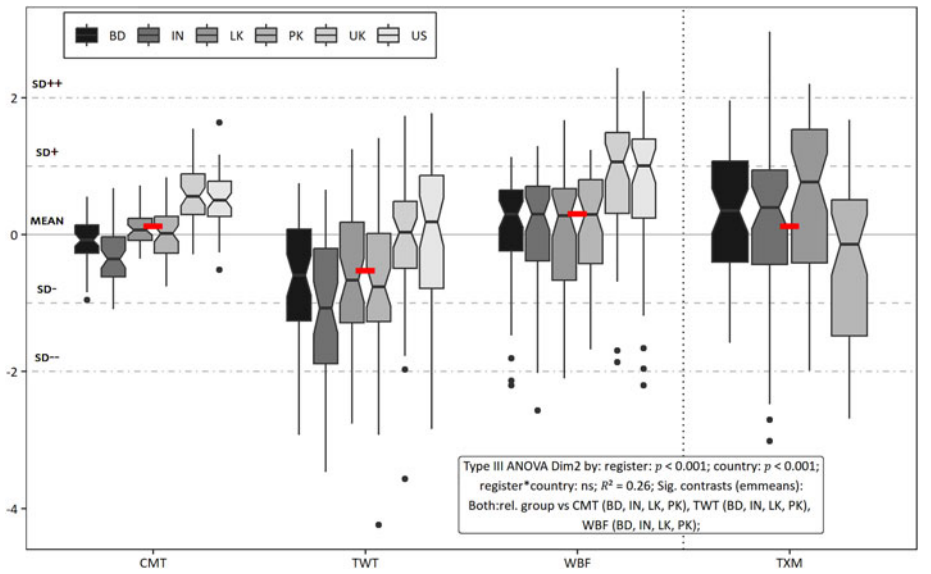

Figure 3. Distribution of registers on Dim2 'Oral elaboration (top) versus informational concerns (bottom)'

3.5 Dim3 'Persuasion- and help-oriented versus informational and past-oriented discussions'

In total 21 linguistic features have loaded on Dim3, divided into 13 and 8 on the positive and negative side respectively. The features on the positive side include vocabulary items like abstract and process nouns, technical nouns and nominalisations, which show that technical and abstract things are under discussion. Some of these features are shared with Biber's (1988) dimension 4 'Overt expression of persuasion', including possibility and necessity modals which refer to actions, i.e. what *could/might* or *can/may* or *will* be done and what *should* or *must* be done. Conditionals and facilitation verbs like *help*, *require*, *force* in the present data indicate that audience is persuaded, e.g. in comments, or explanation/ guidance is provided or demanded – as in the two forum posts in (8). Communication verbs like *threaten*, *encourage*, *ask* etc. facilitate with persuasion or obtaining help, whereas yes/no questions and human nouns can indicate an interactive discourse. The negative side has nominal features like general nouns, including proper nouns and type token ratio. Past tense verbs indicate the presence of narrative elements. There are also some informal features like contractions, verb particles potentially due to phrasal verbs, along with emphatics, frequency-related adverbs, and clause connectors like *still*, *although*, which may suggest some form of discussion, as shown in (9) from Indian comments. Looking at both types of features, this dimension can be labelled 'Persuasion- and help-oriented versus informational and past-oriented discussions'.

(8) I have completed the full course. But, i **could** [**MDPOSSCAll**] not **answer** [**COMM**] the **question** [**NNABSPROC**] no 10. i tried to find the **answer** [**NNABSPROC**] by using logic. but, i **could** [**MDPOSSCAll**] not find the **answer** [**NNABSPROC**]. **Can** [**MDPOSSCAll, YNQU**] you give me the **solution** [**NNTECH**] so that i **can** [**MDPOSSCAll**] learn it?
<#>
You **should** [**MDNE**] be able to get the correct **answer** [**NNABSPROC**], **if** [**COND**] you use countif. (Dim3 Persuasion- and help-oriented discussions; BD-WBF069)

(9) The biggest **tell** [**NNother**]- **tale** [**NNother**] sign of **India** [**NNother**] 's western **neighbour** [**NNother**] 's support for **terrorism** [**NNother**] **took** [**VBD**] place in 2011. In 2011 **mastermind** [**NNother**] of 9/11 **terror** [**NNother**] attack **was** [**VBD**] eliminated in its territory. Immediately thereafter **mastermind** [**NNother**] of 26/11 **terror** [**NNother**] attack along with a **separatist** [**NNother**] from **Kashmir** [**NNother**] valley **held** [**VBD**] a **prayer** [**NNother**] meeting for the eliminated **terror** [**NNother**] mastermind. (Dim3 Informational and past-oriented discussions; IN-CMT380)

As shown in figure 4, the scores on Dim3 are quite sensitive (overall the third most) in discriminating the text categories used in these data, as the variance explained (R^2) based on the interaction of register and country is 0.45 or 45 per cent. Text messages have the highest average scores compared to other registers. The top-scoring texts in this register deal with study-related topics and contain features like abstract and process nouns (*quiz*, *attempt*, *link* etc.), modals (*will*, *need*, *should* etc.) and yes/no questions etc. Comments have the highest and tweets the lowest scores, whereas web forums are in-between on the border. In the first two registers, the South Asian countries are significantly different (i.e. more features closer to the 'persuasion- and help-oriented' pole) from the UK and USA (i.e. more features closer to the 'informational and past-oriented' pole). In terms of forums, Bangladesh, Sri Lanka and Pakistan have significant differences. Also, the spread of the whiskers, especially in Bangladesh and Sri Lanka, shows that there are certain texts where people ask for help on technical topics. Forums related to coding (BD), telecommunication company services (LK) or broadband-related help (PK) appear to be the source of such texts, which have no equivalent in the Indian data included here. Such posts are also less frequent in the UK and USA technology-related forums in these data. The Indian and Pakistani comments also have more past-oriented (i.e. past tense) and general informational features (e.g. general nouns), as shown in (9).

3.6 Dim4 'Focus on humans and mental processes versus non-abstract things and activities'

The fourth dimension consists of 14 features equally divided on both sides. The use of present tense, *be* as the main verb, and human nouns in (10) shows that the focus is on humans. Additionally, mental verbs and *WH* questions indicate that the participants are engaging with each other or the writer by asking questions, e.g. by mentioning other Twitter users as in (11).

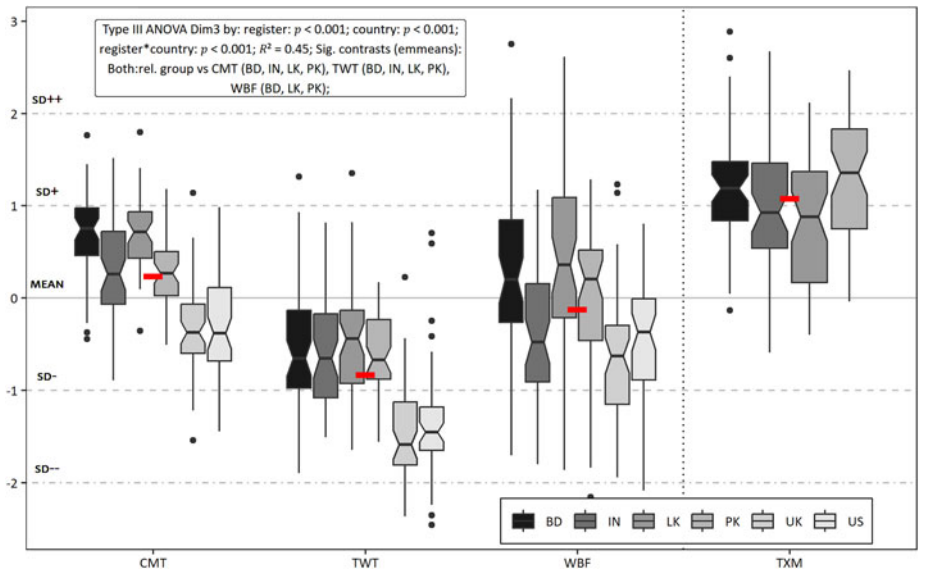

Figure 4. Distribution of registers on Dim3 'Persuasion- and help-oriented (top) versus informational and past-oriented discussions (bottom)'

(10) Bus **drivers** [**NNHUMAN**] **are** [**VPRT, BEMA**] a menace on the road. I **request** [**VPRT**] the **owners** [**NNHUMAN**] to take measures to ensure safety of public on the roads. (Dim4 Focus on humans and mental processes; BD-CMT026)

(11) @xxxxx @XxxxxxxxXXxxxx **where** [**WHQU**] **are** [**VPRT, BEMA**] you from? (LK-TWT276)

Topical adjectives like *economical*, *political*, *human* etc. are used for rather abstract issues, as seen in (12) taken from Bangladeshi comments:

(12) […] This **is** [**VPRT**] n't just about **being** [**BEMA**] a feminist … this **is** [**VPRT, BEMA**] about fighting for the basic **human** [**JJTOPIC**] right to education and liberty. […] (BD-CMT021)

The positive side, thus, can be labelled 'Focus on humans and mental processes'. On the negative side, concrete, quantity and compound nouns refer to non-abstract and technology-related items as observable in (13). Activity verbs and modal verb *would* are associable to users' personal experiences or actions, also exemplified in (13). Hedges and verb particles can be attributed to interactive discourse like web forums. In contrast to the positive pole, 'Focus on non-abstract things and activities' appears to be an appropriate label for the negative side.

(13) […] For **water** [**NNCONC**] **blocks** [**NCOMP, NNCONC**] I '**d** [**MDWO**] like to **use** [**ACT**] **Heat** [**NNQUANT**] Killer but they do n't currently have any plans for RTX3090 **blocks** [**NCOMP, NNCONC**] so I suspect that I might **use** [**ACT**] EK **blocks** [**NCOMP, NNCONC**].

For hardline **tubes** [**NNCONC**] and fittings I 've previously only **used** [**ACT**] […] (Dim4 Focus on non-abstract things and activities; UK-WBF411)

The scores on Dim4 are again reasonably sensitive (overall the second most sensitive) to the text groupings in these data with an R^2 value of 0.54 or 54 per cent, i.e. variance explained if the text categories are known. Comments and tweets comparatively incline to the positive pole 'Focus on humans and mental processes', whereas web forums have more features related to 'Focus on non-abstract things and activities'. Just like the previous dimensions, the South Asian data are significantly more attracted to one pole of the dimension ('Focus on humans and mental processes') in comparison to both IC varieties (or at least one of them) with a reverse trend. This trend is especially notable for the Bangladeshi comments, tweets and web forums. The Indian comments have comparatively lower scores than the other South Asian countries due to the use of concrete and compound nouns; the same is true for the Sri Lankan tweets. The UK web forums have the highest attraction to the negative pole 'Focus on non-abstract things and activities', which also makes them significantly different from most South Asian forums.

The general prominence of South Asian tweets on the positive pole ('Focus on humans and mental processes') may indicate different communicative preferences of these South Asian users. The use of features like human nouns, topical adjectives, *WH* questions and mental verbs in, for example, Pakistan is partly due to the use of Twitter by a large number of political activists (self-observation). The same is somewhat true for the Indian and to a lesser extent the Sri Lankan data. In the Bangladeshi data, many Twitter users are fans of foreign celebrities including those from India and South Korea (Shakir & Deuber 2023: 134). These users tend to interact with other Twitter users and may employ the features listed above. Hence, the difference in communicative purposes of the South Asian netizens versus those from UK and USA appears to be a trend in the South Asian region but with slight differences in each South Asian variety.

3.7 Dim5 'Addressee involved opinion and description'

As shown in table 2, the last dimension has 8 linguistic features with 6 on the positive and only 2 features on the negative side. The features loading on the positive side point to descriptions (predicative adjectives, present tense, *be* as a main verb) along with expression of personal opinion (mental verbs), as shown in (14). The addressee is also involved either by referring to them as in (14), or in many cases through formulaic expressions, e.g. *Thank you*, especially in the public Bangladeshi registers. The negative features (past tense verbs and third-person pronouns) indicate a slight tendency of narration, which is visible in (15) from Sri Lankan comments. As figure 6 exhibits, 0.12 or 12 per cent of the variance in the scores on Dim5 can be explained if the text categories are known, which makes it the least sensitive to the text groupings in the data. The most prominent registers are all from Bangladesh on the positive side and Sri Lankan comments on the negative side. Some of them are significantly different from the native varieties, e.g. Sri Lankan comments and Bangladeshi tweets and forums.

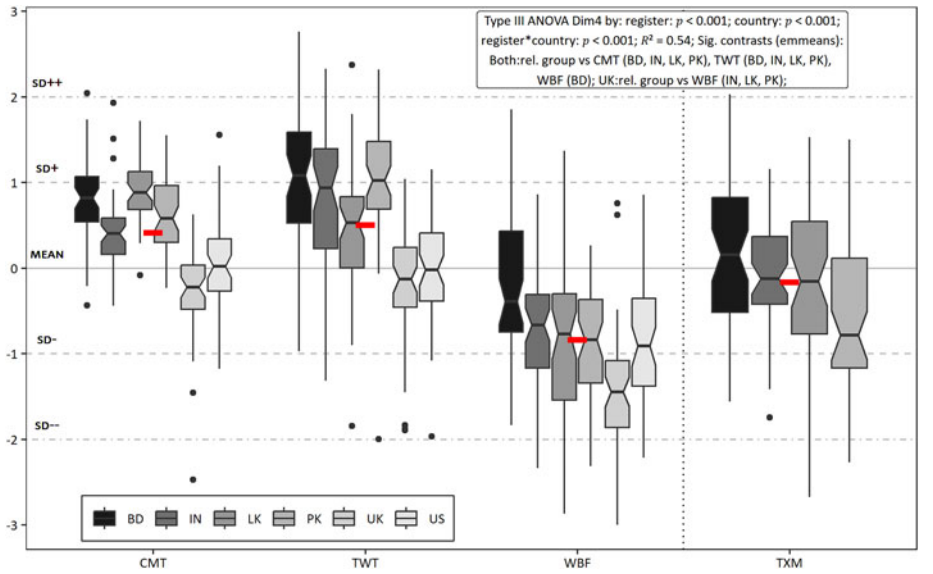

Figure 5. Distribution of registers on Dim4 'Focus on humans and mental processes (top) versus non-abstract things and activities (bottom)'

(14) **Dear** [**JJPRother**] Ms.Aziz, I also **feel** [**VPRT, MENTAL**] like **you** [**SPP2**] whenever I **watch** [**VPRT**] a movie. I **think** [**MENTAL**] it **is** [**VPRT, BEMA**] not **impossible** [**JJPRother**] for anyone to have a sophisticated command on English if he/ she **is** [**VPRT, BEMA**] **sincere** [**JJPRother**] about that. The way **you** [**SPP2**] **have** [**VPRT**] pointed out the obstacles on the way of speaking English fluently, **manifests** [**VPRT**] **your** [**SPP2**] honest endeavour. **You** [**SPP2**] **are** [**VPRT, BEMA**] **sure** [**JJPRother**] to succeed.
I **wish** [**VPRT, MENTAL**] **you** [**SPP2**] all the best. (Dim5 Addressee involved opinion and description; BD-WBF291)

(15) Anyway, **he** [**TPP3S**] **was** [**VBD**] of no use. **He** [**TPP3S**] **was** [**VBD**] always lying to the brim when the nation **knew** [**VBD**] the truth. **He** [**TPP3S**] **was** [**VBD**] a disgrace to the police department. (Dim5 Narrative tendencies; LK-CMT176)

3.8 Regional variation in South Asian text messages

Text messages are only available for South Asia so they have been described separately. As table 3 exhibits, the Pakistani text messages are significantly different from the other three South Asian countries on Dim2, as are the Bangladeshi text messages on Dim5. Pakistani text messages are also more literate impersonal on Dim1. More literate-impersonal and informational discourse entails that sometimes these chats are used for conveying information; e.g. (16) is from a work-related WhatsApp group where pharmaceutical sales representatives exchange information about orders. The use

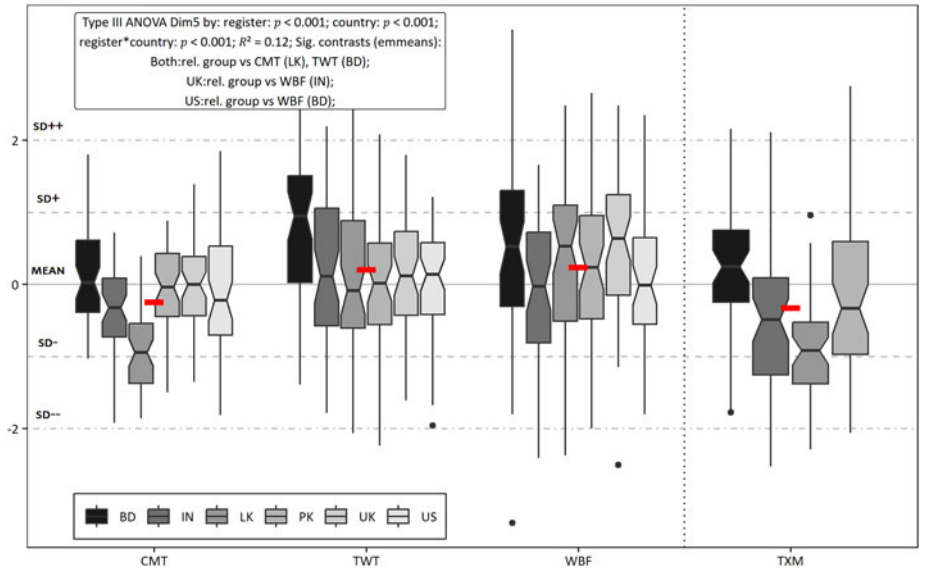

Figure 6. Distribution of registers on Dim5 'Addressee involved opinion and description (top)'

of general nouns including proper nouns, time references, imperatives and activity verbs is visible in the example.

(16) It's **already** [**TIME**] **delivered** [**ACT**] at **FSD** [**NNother**] on 27.02.21 **received** [**VBN, ACT**] by **Imran** [**NNother**]. **Check** [**VIMP, ACT**] with them.
<#>
Respected [**VBN**] **sir** [**NNother**]
Required urgently **stock** [**NNother**] for **Al** [**NNother**] **Aziz** [**NNother**] distribution **sargodha** [**NNother**].
Levopraid [**NNother**] 50 mg [**NNother**] 1000 (Dim2 Informational concerns; PK-TXM010)

The Bangladeshi text messages are significantly more descriptive and addressee inclusive on dimension 5. For example, in (17) participants encourage each other to

Table 3. *Regional variation in South Asian text messages*

Dimension	Type III ANOVA	Significant contrasts (emmeans)
Dim1	country: $p < 0.01$	IN - PK: $p < 0.05$, LK - PK: $p < 0.01$
Dim2	country: $p < 0.01$	BD - PK: $p < 0.05$, IN - PK: $p < 0.05$, LK - PK: $p < 0.001$
Dim3	country: $p < 0.05$	LK - PK: $p < 0.05$
Dim4	country: $p < 0.01$	BD - PK: $p < 0.001$
Dim5	country: $p < 0.001$	BD - IN: $p < 0.001$, BD - LK: $p < 0.001$, BD - PK: $p < 0.05$, LK - PK: $p < 0.01$

speak (or write) English by saying that mistakes are a learning opportunity. These texts are from a few group chats in the Bangladeshi data where participants want to practise English because they are preparing for the IELTS exam.

(17) […] if **you** [**SPP2**] **make** [**VPRT**] mistake no problem other member replace **your** [**SPP2**] mistake. it will **be** [**BEMA**] **very** [**AMP**] **helpful** [**JJPRother**] for everybody.
<#>
However, I **wanted** [**MENTAL**] to say, if I/ we Mistake any sentences or words grammatically, who **knows** [**VPRT, MENTAL**], will refine my or one's Mistake .
<#>
In **bengali** [**JJPRother**]
<#>
when **you** [**SPP2**] **make** [**VPRT**] mistake then **you** [**SPP2**] can **learn** [**MENTAL**] on this mistake. (Dim5 Addressee inclusive opinion and description; BD-TXM096)

3.9 Summary of findings

The resulting dimensions of the MD analysis appear to confirm previous studies (e.g. Biber 1988; Biber & Egbert 2016) in terms of universal dimensions of register variation: oral versus literate discourse (Dim1) and stance/opinion- and persuasion-related features (Dim2 'Oral elaboration' and Dim3 'Persuasion- and help-oriented discussions'). There are narrative elements on the negative sides of Dims 3 and 5 but a full-fledged narrative dimension has not emerged. The South Asian data generally have nominal, literate and formal discourse as shown on the negative ends of Dims 1 and 2. The comments, tweets and web forums show regional variation on Dim2 (less oral-elaborative more informational), Dim3 ('Persuasion- and help-oriented discussions'), and Dim4 ('Focus on humans and abstract processes'). Dim2 and Dim3 show that the devices used to express opinion and persuade others are present in comments as noted previously (e.g. Ehret & Taboada 2021), but the South Asian English comments employ different features (e.g. modal verbs of necessity on Dim3) as opposed to evaluative adjectives and stance *to* clauses (Dim2) in the IC data. The South Asian web forums on Dim3 have different topics than the UK and USA forums, e.g. technical help, which, as discussed above, may be due to the types of forums included in these data. The comparatively high scores of South Asian tweets on Dim4 (positive pole 'Focus on humans and mental processes') are probably due to an abundance of political activist and fan accounts; in the British and American data, in contrast, features like human nouns and topical adjectives are less common.

There are also some individual trends in these South Asian data. The Bangladeshi netizens use features like *WH* questions, topical adjectives, human nouns etc. in comments and web forums more prominently. The communicative purpose of sharing information is noteworthy in the Pakistani text messages. The Indian web forums are informational and past oriented on Dim3 due to the use of past tense and general nouns. Similarly, the Indian comments are literate and narrative as they use more nouns

(general on Dim3; concrete and compound on Dim4), past tense (Dim3 and Dim5) and third-person singular pronouns (Dim5) in comparison to the Bangladeshi and Sri Lankan English comments. The same is partly applicable to the Pakistani comments on Dim3 and Dim4. Finally, the Sri Lankan comments are different from the other South Asian English comments in these data due to oral features like first- and second-person pronouns, discourse markers, etc. (Dim1) as well as narrative features like past tense and third-person pronouns (Dim5).

4 Discussion, limitations and conclusion

An MD analysis has been presented in the previous sections that covers four interactive registers – comments, tweets, web forums and text messages – from six varieties of English, i.e. four OC varieties from South Asia, namely Bangladesh, India, Pakistan and Sri Lanka, and two IC reference varieties, i.e. those from the UK and US. The study results in five dimensions of variation, among which Dim1, Dim3 and Dim4 are the most important dimensions in terms of differentiating regional and register varieties in the present data.

As figure 7 depicts, the data of the four OC varieties form their own cluster distinct from that of the two IC varieties. Based on the present data it appears that the communicative function of conveying and receiving information is important for these OC data as shown by Dim1. Moreover, the functions of persuasion and asking/providing help (Dim3), focus on humans and abstract issues and processes (Dim3, Dim4), and narration (Dim2) are general characteristics of the OC discourse in these data.

Based on previous research, the presence of literate and formal discourse, even in online communication, is unsurprising. The use of abstract nominal features can also be attributed to the formal domains associated with English communication in these countries, e.g. education, business etc. The use of modal verbs to express opinion and persuasion (Dim3) is apparently more straightforward for the OC netizens than other stance-marking devices like adverbs that are abundant in the IC data (Dim2). South Asian Twitter users engage with other users due to the presence of political workers and fans as shown in section 3.6.

Apart from these general trends, there are also differences among the South Asian varieties, as can be seen from the positions of Sri Lanka and Bangladesh in the phylogram. The data from the regional super-central variety, i.e. Indian English, has a number of commonalities with the Pakistani data, for example comments and web forums use more nouns including proper nouns (Dim1, Dim3). The Sri Lankan data are less divergent from the former two compared to the Bangladeshi data. Nonetheless, the informal features in the Sri Lankan comments (Dim1) and oral-elaboration-related features in the text messages (Dim2) make Sri Lankan English dissimilar from other varieties in these data. The Bangladeshi data are the most divergent of all mainly due to higher scores on Dim3–5 in one or more registers.

Based on these results, it is difficult formulate any clear statements about the status or regional power relationships among these varieties, e.g. using Mair's (2013)

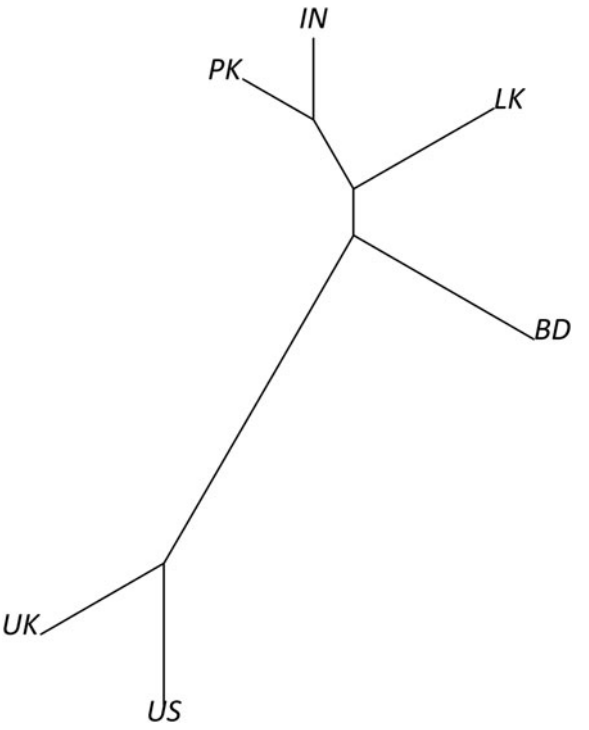

Figure 7. Phylogram of the countries based on mean dimension scores of the interactive registers (excluding text messages) (see Appendix)

classification of varieties of English. Hence, the super-central status of Indian English and the hyper-central status of American English cannot be confirmed or denied, though that is of course a matter of perspective and criteria applied. However, Indian English and Pakistani English, based on these data, appear to be the two well-established varieties that perform very different functions online compared to the two reference varieties. The Sri Lankan and Bangladeshi English data appear to be slightly more unique in certain functions but still similar to the two main South Asian varieties. Bangladeshi English could be considered an EFL (English as a foreign language) variety in view of the restricted uses of English in this country, the scarce availability of text messages and blogs (Shakir & Deuber 2023), and the existence of chat groups on themes like preparing for IELTS, as shown in section 3.8. Overall, the earlier classification by Kachru (1985), i.e. OC varieties, appears to be relevant for the four South Asian countries based on this analysis. The present results further entail that despite the global nature of CMC it is possible to find local patterns of variation (e.g. in the Bangladeshi or Sri Lankan data) as well as regional trends of variation (e.g. the common behaviour of the South Asian comments or tweets), and they indicate connections between online and offline language use.

The data used for the present article have certain limitations. Firstly, the number of texts included, i.e. 1,100 texts, is rather low. Secondly, though SAOnE consists of 4.5 million (or 4 million in the case of UK and USA) words per country, the number of unique web sources in comments (2 newspaper websites per country) and web forums (5–8 forum websites per country) is small. Lastly, text messages are generally limited to formal domains like education and work, with UK and US texts missing. These limitations were unavoidable due to limited time and resources available for data collection, more time and effort needed at the data screening stage to verify countries of origin, and unavailability or inaccessibility of certain types of data (see Shakir & Deuber 2023: 122–8). These limitations entail that these data may not cover the whole spectrum of functional variation in the interactive English CMC of these countries and the resulting dimensions may be only exploratory in nature. Despite these limitations, it certainly becomes clear that register considerations are important and need to be given more attention in World Englishes studies that are based on CMC data and corpora that consist of CMC, e.g. GloWbE (Davies & Fuchs 2015).

To conclude, the study hopes to have shown the advantages of using multi-register CMC data for varieties of English, e.g. Bangladeshi and Pakistani English, where such informal data are not or scarcely available in traditional domains. The use of a register-centric and feature-aggregate approach like MD analysis highlights aspects of these varieties – e.g. the behaviour of Bangladeshi English in these data or the use of English on Twitter to engage with others in South Asia – which are difficult to arrive at using a register-controlled approach focusing on only one or a few features, as is the case with most variationist studies in World Englishes. Thus, MD analysis can provide a useful link between macro-sociolinguistic perspectives and studies of individual language features. In terms of MD analysis tools, an opensource piece of software has been enhanced during the course of this research. The modified version of MFTE has been released (link) to help promote open science practices in the discipline. Lastly, the desirability of more studies using similar approaches with similar or larger datasets has been underlined not only for the varieties investigated here but also other varieties in the Outer and Expanding Circles according to Kachru (1985), with regional aspects as a further point deserving more consideration.

Author's address:

Englisches Seminar
University of Münster
Johannisstr. 12–20
48143 Münster
Germany
muhammad.shakir@uni-muenster.de

References

Berber Sardinha, Tony. 2014. 25 years later: Comparing internet and pre-internet registers. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*, 81–105. Amsterdam and Philadelphia: John Benjamins.

Berber Sardinha, Tony. 2022. Corpus linguistics and the study of social media: A case study using multi-dimensional analysis. In Anne O'Keeffe & Michael J. McCarthy (eds.), *The Routledge handbook of corpus linguistics*, 2nd edn, 656–74. Abingdon: Routledge.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*, 2nd edn. Cambridge: Cambridge University Press.

Biber, Douglas & Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics* 44(2), 95–137.

Biber, Douglas & Jesse Egbert. 2018. *Register variation online*. Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora* 10(1), 11–45.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Bohmann, Axel. 2019. *Variation in English worldwide: Registers and global varieties*. Cambridge: Cambridge University Press.

Collot, Milena & Nancy Belmore. 1996. Electronic language: A new variety of English. In Susan C. Herring (ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, 13–28. Amsterdam and Philadelphia: John Benjamins.

Davies, Mark & Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1), 1–28.

Egbert, Jesse & Shelley Staples. 2019. Doing multidimensional analysis in SPSS, SAS and R. In Tony Berber-Sardinha & Marcia Veirano Pinto (eds.), *Multidimensional analysis research methods and current issues*, 125–44. London: Bloomsbury.

Ehret, Katharina & Maite Taboada. 2020. Are online news comments like face-to-face conversation? A multi-dimensional analysis of an emerging register. *Register Studies* 2(1), 1–36.

Ehret, Katharina & Maite Taboada. 2021. Characterising online news comments: A multi-dimensional cruise through online registers. *Frontiers in Artificial Intelligence* 4, 643770. https://doi.org/10.3389/frai.2021.643770

Evert, Stephanie. 2022. A multivariate approach to linguistic variation. https://stephanie-evert.de/SIGIL/sigil_R/materials/07_multivar_talk.handout.pdf

Fox, John & Sanford Weisberg. 2019. *An R companion to applied regression*. Thousand Oaks, CA: Sage.

Friginal, Eric, Oksana Waugh & Ashley Titak. 2018. Linguistic variation in Facebook and Twitter posts. In Eric Friginal (ed.), *Studies in corpus-based sociolinguistics*, 342–62. New York: Routledge.

Grieve, Jack, Douglas Biber, Eric Friginal & Tatiana Nekrasova. 2010. Variation among blogs: A multi-dimensional analysis. In Alexander Mehler, Serge Sharoff & Marina Santini (eds.), *Genres on the web: Computational models and empirical studies*, 303–22. New York: Springer.

Hardy, Jack A. & Eric Friginal. 2012. Filipino and American online communication and linguistic variation. *World Englishes* 31(2), 143–61.

Heyd, Theresa & Christian Mair. 2014. From vernacular to digital ethnolinguistic repertoire: The case of Nigerian Pidgin. In Véronique Lacoste, Jakob Leimgruber & Thiemo Breyer (eds.), *Indexing authenticity*, 244–68. Berlin: De Gruyter.
Hussain, Zahida. 2016. A multidimensional comparative analysis of Pakistani English. PhD thesis, Govt. College University Faisalabad.
Jenks, Grant. 2018. wordsegment. Python. https://pypi.org/project/wordsegment/
Jonsson, Ewa. 2015. *Conversational writing: A multidimensional study of synchronous and supersynchronous computer-mediated communication*. Frankfurt a.M.: Peter Lang.
Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk & Henry G. Widdowson (eds.), *English in the world – Teaching and learning the language and literature*, 11–30. Cambridge: Cambridge University Press.
Kruger, Haidee & Bertus van Rooy. 2018. Register variation in written contact varieties of English: A multidimensional analysis. *English World-Wide: A Journal of Varieties of English* 39(2), 214–42.
Le Foll, Elen. 2021a. A Multi-Feature Tagger of English (MFTE). Perl. https://github.com/elenlefoll/MultiFeatureTaggerEnglish (accessed 28 January 2023).
Le Foll, Elen. 2021b. Introducing the Multi-Feature Tagger of English (MFTE) version 3.0. https://github.com/elenlefoll/MultiFeatureTaggerEnglish/blob/main/Introducing_the_MFTE_v3.0.pdf. (accessed 26 September 2023).
Le Foll, Elen. 2022. Textbook English a corpus-based analysis of the language of EFL textbooks used in secondary schools in France, Germany and Spain. PhD thesis, University of Osnabrück.
Lenth, Russell V. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans
Loureiro-Porto, Lucía. 2017. ICE vs GloWbE: Big data and corpus compilation. *World Englishes* 36(3), 448–70.
Mair, Christian. 2013. The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34(3), 253–78.
Neumann, Stella. 2014. Cross-linguistic register studies: Theoretical and methodological considerations. *Languages in Contrast* 14(1), 35–57.
Neumann, Stella & Stephanie Evert. 2021. A register variation perspective on varieties of English. In Elena Seoane & Douglas Biber (eds.), *Corpus-based approaches to register variation*, 143–78. Amsterdam: John Benjamins.
Nini, Andrea. 2014. *Multidimensional Analysis Tagger 1.2, Manual*. https://sites.google.com/site/multidimensionaltagger/versions
Nini, Andrea. 2019. The multi-dimensional analysis tagger. In Tony Berber-Sardinha & Marcia Veirano Pinto (eds.), *Multidimensional analysis research methods and current issues*, 67–94. London: Bloomsbury.
Python Core Team. 2022. *Python: A dynamic, open source programming language*. Python Software Foundation. www.python.org
R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org
Rencher, Alvin C. 1992. Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician* 46(3), 217. https://doi.org/10.2307/2685219
Revelle, William. 2022. *psych: Procedures for psychological, psychometric, and personality research*. Evanston, IL: Northwestern University. https://CRAN.R-project.org/package=psych
Shakir, Muhammad & Dagmar Deuber. 2018. A multidimensional study of interactive registers in Pakistani and US English. *World Englishes* 37(4), 607–23. https://doi.org/10.1111/weng.12352
Shakir, Muhammad & Dagmar Deuber. 2019. A multidimensional analysis of Pakistani and U.S. English blogs and columns. *English World-Wide* 40(1), 1–23.

Shakir, Muhammad & Dagmar Deuber. 2023. Compiling a corpus of South Asian online Englishes: A report, some reflections and a pilot study. *ICAME Journal* 47(1), 119–39.
Shepelev, Victor. 2022. spylls. Python. https://pypi.org/project/spylls/ (accessed 29 January 2023).
Titak, Ashley & Audrey Roberson. 2013. Dimensions of web registers: An exploratory multi-dimensional comparison. *Corpora* 8(2), 235–60.
Toutanova, Kristina, Dan Klein, Christopher Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, 252–9.
Venables, W. N. & B. D. Ripley. 2002. *Modern applied statistics with S*, 4th edn. New York: Springer.
Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28(4), 421–50.

Appendix

Table A1. *Mean dimension scores used in figure 7 (excluding text messages)*

Country	PC1	PC2	PC3	PC4	PC5
BD	−0.087	−0.216	0.161	0.545	0.472
IN	−0.393	−0.438	−0.205	0.158	−0.095
LK	−0.021	−0.166	0.231	0.164	−0.19
PK	−0.341	−0.206	−0.134	0.271	0.028
UK	−0.374	0.457	−0.793	−0.642	0.231
US	−0.11	0.397	−0.7	−0.281	−0.015