A register variation perspective on varieties of English

Stella Neumann, RWTH Aachen University
Stefan Evert, FAU Erlangen-Nürnberg

1. Introduction

The notion of variety usually refers broadly to variants of a given language spoken by larger groups of speakers. It typically covers an entire region or political entity as in "Austrian German" or "Brazilian Portuguese". Often, however, the classification also accounts for specific social groups as in "Black South African English" or "African American English". Languages of wider circulation such as Spanish and English often consist of a whole range of varieties reflecting the spread of the respective language. It is specifically this latter type of varieties that calls into question the idea of a given language as some more or less consistent cultural entity. It will intuitively appear plausible to assume that the Spanish spoken in Bolivia is embedded in a different cultural context than the variety spoken in Andalusia. The same holds for varieties of English spoken in cultural contexts as different as the Nigerian, the Indian, the Irish and the New Zealand one. In light of this, Halliday and Matthiessen (2014, 73) characterize English as an "assemblage" of varieties, "subject to dialectal, codal and registerial variation", which will also manifest in variation in lexico-grammar. Following Malinowski (1935), we may assume that the differing cultural contexts will not remain without impact on language use in situational contexts.

Language use in situational context, in turn, finds its reflex in register, that is in "a configuration of meanings that are typically associated with a particular situational configuration" (Halliday and Hasan 1989, 38–39). Given the theoretical link between levels of linguistic description, or strata in systemic functional parlance, meanings are said to be realized by lexico-grammatical and phonological features. By referring to meanings *typically* exchanged or 'at risk' in a given situation (Halliday 1978, 185), this definition of register reflects the potential for construing (similar) meaning in variable ways. In a sense, a situational configuration recognized by a language user will narrow down his/her search space of what meanings are going to be exchanged (including which meanings will have to be inferred) without allowing him/her to predict the exact wordings. Registers as established configurations of meanings in this sense are likely to be given culturally-recognized labels as Biber et al. suggest in their chapter in this volume.

Crucially, initiated language users are able to recognize the situation they are experiencing as an instance of a type of situation, because it bears general similarities with situations they experienced in the past (Gregory 1967). This means that, while the specific configuration of individual interactants at a given time and place is unique, they will respond to and enact more general characteristics of a given type of situation. The recurring character of different situation types and interactants' adapted language behavior explains why the lexico-grammar of any given language is not a homogenous structure where every feature is equally likely to be chosen regardless of the specific context, but rather a probabilistic system (Halliday 1991; see also Taverniers 2019) with situation-specific pre-selections (Matthiessen 1993). These considerations forming the backbone of systemic functional language theory have been corroborated by empirical studies of language use across different registers that have

shown different distributions of linguistic features (see, e.g., Neumann 2014a; outside of SFL, see Biber et al. 1999; and for a more general overview Biber 2019).

Assuming a close relation between cultural and situational context, registers reflecting similar situational configurations should potentially differ across varieties of a language that differ in cultural context. Parliamentary debates, for example, exist across a wide range of cultures in which English plays a central role. However, since the political systems of the respective cultures differ, we might expect differences in the specific meanings exchanged during parliamentary debates. While the more general social activity may thus be comparable, it is likely that more specific aspects such as the exact relationship between members of the parliament and governmental representatives or between different political groupings differs not only due to constitutional differences but also because of different historical developments. As a consequence, parliamentarians can be expected to express partially different meanings in a comparable register.

Cultural differences may of course be so significant that at least for some registers existing in culture A no counterpart can be identified in culture B. The corpus of Nukulaelae Tuvaluan analyzed by Biber (1995), for instance, contains at least one register we might claim not to have a counterpart in English.

Against this conceptual background the present paper reports on a corpus-based exploration of register variation across three varieties of English, applying the multivariate procedure developed by Diwersy et al. (2014) and further worked out by Evert and Neumann (2017). The approach is informed by theoretical concepts developed in the systemic functional tradition of register theory (for a recent summary see Matthiessen 2019). These theoretical concepts specifically enter our derivation of features (see Section 3.2). Our specific goal in this paper is to identify dimensions of register variation between individual texts in three comparable corpora, using the common sampling frame of the corpora – in particular, their text categorization scheme – as a frame of reference. The remainder of the paper is organized as follows. Section 2 reviews previous research on register variation with a specific emphasis on variation across varieties. We introduce the particulars of our corpus and feature extraction in Section 3 and discuss the multivariate analysis in Section 4. The paper is rounded off by some concluding remarks in Section 5.

2. State of the art

Drawing on sources in various linguistic traditions, register linguistics looks back on almost fifty years of scholarly engagement (for a graphical map of traditions, see Matthiessen 2019, 26). Specifically, Halliday and Hasan (1989) worked this out in the systemic functional tradition into a fully-fledged account of text and context with register as the linguistic reflection of situational context. A notion of register embedded in (systemic) functional language theory means that it is conceptualized in a framework with links between basic functions of language, parameters of register and specific lexico-grammatical systems. Moreover, Matthiessen (1993) elaborates the relationship between language theory and probabilities of lexico-grammatical features to be selected in given registers. In a usage-based perspective, this means that some features are described as available and others as “turned off” (Matthiessen 1993, 259) based on their frequency of occurrence in a given situational context. This stochastic thinking underlies Halliday’s characterization of register as “a cluster of associated features having a greater-than-random (or rather, greater than predicted by their unconditioned probabilities) tendency to co-occur” (Halliday 1988, 162). Consequently, SFL-based register analysis is inherently quantitative, as Teich (2013, 417) points out. Despite this and the more general fact that systemic functional language theory emphasizes the centrality of language use and consequently the frequency-based nature of lexico-grammar, only few studies exist which actually put the theoretical claims to a quantitative test.

Besides this functional language-theoretical branch of register linguistics, a second tradition draws on sources in anthropology and sociolinguistics, among others Hymes (1972), Ervin-Tripp (1972), see also Biber and Finegan (1994). It is particularly this line of research which brought about a breakthrough in establishing empirical evidence of systematic language variation based on register, mainly in Biber's seminal work using multivariate analysis (see, among others, Biber 1988; 1995; 2019). Biber introduced a quantitative empirical approach that permitted the identification of dimensions of register variation based on the co-occurrence of linguistic features. This characterization of register variation as a multi-dimensional space exhibits a clear conceptual overlap with the systemic functional approach, which derives claims about multidimensionality from theoretical assumptions.

In some sense, Neumann (2014a) can be seen as a combination of both strands of research, embedding her corpus analysis in the systemic functional framework and specifically using the linguistic indicators of register variation as operationalizations of theory-based register constructs (for a discussion of corpus approaches and SFL-based register theory, see Moore 2020, 37–38). While this allows her to structure the corpus analysis according to sub-dimensions of Halliday's register parameters field, tenor and mode, it results in a very constrained statistical approach, which is essentially univariate. More specifically, she determines the register specificity of features by comparing their frequency distributions in a given register with distributions in a reference corpus, using Student's t-test. This approach has the advantage of yielding a distinctive characterization of individual registers rather than their relative characteristics against the background of a set of other registers. However, its univariate design fails to account for more complex co-occurrence patterns between linguistic features.

As to varieties of English, there is a large body of corpus-based studies of linguistic variation drawing mainly on the International Corpus of English (ICE; Greenbaum 1996a). Usually, such studies concentrate on the comparative distribution of a given vernacular feature (Sand 2004; Hundt, Röthlisberger, and Seoane 2018 to name but two examples). Linguistic variation across varieties is doubtlessly a key research area for sociolinguistics, where the relevance of register is sometimes only acknowledged as a nuisance factor that needs to be controlled for (Szmrecsanyi 2019, 77). Despite the fact that functional variation, i.e. register, is intimately linked to social factors – both user-related, viz. regional and social, and use-related, viz. functional, variation reflect social order, albeit in different perspectives (Halliday 1978, 185) – the interaction between the types of variation has only recently started to receive attention. Szmrecsanyi (2019, 82) lists a number of recent studies the cumulative weight of which he claims to suggest "that register regularly affects the relative frequency with which speakers and writers select particular variants".

A number of studies have also approached the interaction between social/regional and functional variation from the perspective of register variation across varieties of English reporting evidence of variety-specific patterns with the help of Biber's multidimensional approach (MDA). Xiao (2009) applies MDA with an enhanced set of linguistic features to Asian varieties and British English, drawing on the International Corpus of English (ICE). Van Rooy et al. (2010) analyze the East African component of the ICE Corpus in terms of the original MDA dimensions (Biber 1988). Both studies concentrate on the comparison of a subset of geographically (and potentially linguistically) related corpus components. As will be explained below, our selection of components deliberately draws on varieties that differ with respect to several factors. A study that also investigates register variation across varieties of English using MDA is by Kruger and Van Rooy (2018). Their data set compiled for comparing register variation across contact varieties consists of written text categories across native and non-native Englishes from the ICE Corpus as well as translated texts. Given the focus on written text categories, the dimensions yielded by the MDA inevitably differ in the potential influence of spoken text categories. Interestingly, Kruger and Van Rooy (2018, 237) come to the conclusion that the dimensions are generally similar

across varieties. Arguably, this can be explained by their necessary focus on written varieties. The authors point to language users being "drawn to more formal and normative choices" in language contact situations (Kruger and Van Rooy 2018, 237).

In this paper, we take a closer look at register variation in varieties of English by (i) focusing on the variability of individual texts rather than comparing average scores for text categories and varieties, (ii) using the sampling frame of the ICE Corpus as a frame of reference for our multivariate analysis, and (iii) building on a targeted, yet comprehensive set of linguistic features rooted in SFL theory. This type of comparison informed by systemic functional register theory was previously attempted in two small-scale studies by Neumann (2012) and Neumann and Fest (2016), which derived dimensions from register theory and inspected frequencies of features across ICE components without assessing the statistical significance of the observations. Drawing on this experience in working with the ICE Corpus, Neumann (2020) uses the visualization-based approach to multivariate analysis developed by Diwersy, Evert and Neumann (2014). We refer to this approach as geometric multivariate analysis (GMA). GMA is inspired by Biber's MDA, but takes a more geometric perspective emphasizing the visualization of individual texts in multidimensional feature space. Its main object of study are linguistic differences between texts as quantified by Euclidean distances in feature space, in contrast to the feature correlation patterns foregrounded by MDA. As a consequence, GMA uses the orthogonal projections of principal component analysis (PCA) to identify latent dimensions of linguistic variation instead of MDA's factor analysis[i]; GMA dimensions can thus be interpreted as geometric perspectives on the high-dimensional data set of individual texts. PCA is complemented by supervised linear discriminant analysis (LDA), which can bring out subtler patterns of variation by finding a perspective that visually separates pre-determined groups of texts.

Visualization is central to GMA because it shows the position of each individual text in relation to all other texts, i.e. the patterns of (dis)similarities both between text categories and between individual texts. The mathematical properties of the geometric analysis techniques ensure that the visually observed patterns are interpretable as linguistic differences between texts viewed from a specific perspective (for each latent dimension). It is thus possible to appraise not just differences between aggregated groups of data points (in the form of group averages or centroids), but also the variability and specific distribution patterns within each group.

A small set of previous studies have also used sequences of multivariate techniques for related purposes such as document classification, authorship attribution and, to a more limited extent, also for exploring register variation. Karlgren and Cutting (1994) apply LDA for automatic genre classification of texts. Building on their work, Stamatatos, Fakotakis and Kokkinakis (2000) use a larger set of features and extend the application to authorship attribution. Neither paper is concerned with linguistic aspects of register variation. Tambouratzis et al. (2004) use cluster analysis for investigations of register and style in Greek texts. Notably, Argamon et al. (2007) present an approach to stylistic text classification with the help of support vector machines that draws on a theoretical motivation in SFL very similar to our reasoning. However, their study is geared towards the specific purpose of stylistic analysis and consequently only covers a limited feature set insufficient for register analysis. Egbert and Biber (2018) compare unsupervised factor analysis with supervised LDA[ii] for modeling register variation, finding very similar register dimensions but substantial differences in the features associated with each dimension (Egbert and Biber 2018, 258–267).

Neumann (2020) can be seen as a precursor study to this paper. In fact, we build on the same three ICE components and the same feature extraction procedure (see Section 3 for a description of the corpus and the feature extraction pipeline), except that the earlier study uses a higher cut-off point for text length, viz. 500 words instead of 100 words, and fails to remove material labelled as extra-corpus

text by the corpus compilers. The key difference, however, is that Neumann (2020) injects the three ICE components, namely Hong Kong English, Jamaican English and New Zealand English, as supervised information into the LDA, thus foregrounding linguistic variation between the varieties. We use the text categories defining the ICE sampling frame as supervised information – specifically excluding any information on which ICE component a text belongs to – which allows us to study how the varieties differ along dimensions of register variation and relates more closely to previous studies in the MDA framework. Based on her exploration of the LDA dimensions, Neumann (2020) concludes that small, but discernible differences can be observed: texts display variety-specific tendencies which break up clusters formed by the text categories. In terms of register-related variation captured by the unsupervised PCA dimensions, she observes visual evidence for variation between spoken and written text categories. She interprets this as corroboration of Biber's (1988) dimension 1 with the help of a different data set and multivariate procedure, which also sheds light on the gradual overlap between text categories (and varieties). The present paper examines the similarities and differences between text categories in the data set in more detail.

3. Data

3.1. The corpus

The best resource available for investigating register variation across varieties of English is the International Corpus of English (Greenbaum 1996a), the standard corpus resource for research on varieties of English. The corpus covers a large set of different varieties collected as so-called components following a common corpus design and annotation scheme.[iii] Each component has a targeted size of 1 million words across 32 spoken and written text categories. Greenbaum explains that the corpus was designed to capture English as a means of communication between speakers in a country in which English is used "either as a majority first language (for example, Canada and Australia) or an official additional language (for example, India and Nigeria)" (Greenbaum 1996b, 3). It is important to bear in mind that the focus of the corpus is on educated English. As Greenbaum (1996b, 6) writes, the language represented in the corpus is that of adults "who have received formal education through the medium of English to the completion of secondary school". Consequently, the corpus does not cover the full spectrum of different lects in the respective variety.

Using this corpus for research into register variation across varieties of English is not entirely without problems. Collection and mark up of texts are carried out by individual local teams, but following the above mentioned predetermined corpus design. This common design is both an advantage, as it makes the components comparable, and a disadvantage because it potentially obscures the actual variance of language use in the respective country. It should also be noted that the categorization is based on the British cultural context. It is possible that the specific interpretation of what counts as, say, a broadcast interview differs between teams. Moreover, there is also the possibility that one of the predetermined text categories does not exist in another cultural context in the (same) form in which it might (have) exist(ed) in the UK. Register defined as the semantic reflection of a particular situational context which, in turn, is embedded in a given cultural context is therefore not necessarily captured by the pretheoretical ICE text categories. This poses a challenge for a study on register variation across varieties of English. In order to avoid overstating the empirical findings of this paper with respect to register, we therefore use the term 'text category' rather than 'register' when referring to the corpus texts.

Moreover, we also face a dilemma with respect to the large number of text categories, which makes it difficult to separate categories in the visualization (e.g. by showing them in different colors) and to properly assess variation within and between the categories. For example, it can be argued that academic writing across different disciplines exhibits sufficient similarities to warrant grouping it under the general heading of academic writing (rather than the four separate text categories of the ICE

scheme). On the other hand, very broad categories would risk grouping together texts that reflect quite diverse situation types; multivariate analysis would then only be able to capture individual salient features that these texts happen to share rather than the true multidimensional character of register. We might want to question, for instance, whether the categories of administrative writing and skills/hobbies share enough properties to justify their combined analysis as instructional writing (at the coarser level of the ICE category system). To avoid the shortcomings of both options, we use an intermediate set of 20 categories that represent a compromise between the two ICE categorization levels. Table 1 gives an overview of the original ICE category system and a comparison with the 20 text categories used in this paper in the right-most column.

Table 1. Comparison of the classification of text categories according to the ICE initiative and the 20 categories used in this paper

ICE classification			**20 text categories**
Dialogues	private	face-to-face conversations	conversations/phonecalls
		phonecalls	
	public	classroom lessons	classroom lessons
		broadcast discussions	broadcast interactions
		broadcast interviews	
		parliamentary debates	parliamentary debates
		legal cross-examinations	legal cross-examinations
		business transactions	business transactions
Monologues	unscripted	spontaneous commentaries	unscripted monologues
		unscripted speeches	
		demonstrations	demonstrations
		legal presentations	legal presentations
	scripted	broadcast news	scripted monologues
		broadcast talks	
		non-broadcast talks	
Non-printed	student writing	student essays	student writing
		exam scripts	
	letters	social letters	social letters
		business letters	business letters
Printed	academic writing	humanities	academic writing
		social sciences	
		natural sciences	
		technology	
	popular writing	humanities	popular-scientific writing
		social sciences	
		natural sciences	
		technology	
	reportage	press news reports	news reports
	instructional writing	administrative writing	administrative writing
		skills/hobbies	skills and hobbies
	persuasive writing	press editorials	press editorials
	creative writing	novels and short stories	creative writing

To cover a set of varieties likely to reflect cultural differences, we continue to use the set of three ICE components analyzed by Neumann (2020), namely Hong Kong English, Jamaican English and New Zealand English. She motivates the choice by referring to Schneider's (2007) dynamic model of postcolonial Englishes that classifies the development of postcolonial varieties into a foundational phase by colonial expansion, exonormative stabilization through increased contact between colonizers and the indigenous population (phase 2), the emergence of distinctive nativized features (phase 3), codification and endonormative stabilization (phase 4) and, in the last phase, internal differentiation. According to Schneider (2007), Hong Kong English displays exonormative orientation and is in the process of undergoing nativization (phase 3), Jamaican English displays traces of implicit endonormative orientation (phase 4), whereas New Zealand shows signs of dialectal fragmentation and social variation (phase 5). Moreover, they are characterized by different historical developments, different language contact situations and differ as to the learning pathway, with English as L1 for the majority of speakers in New Zealand, as an indigenized L2 variety in Jamaica and as a non-native L2 variety in Hong Kong. We believe that these three components are thus likely to capture sufficient differences required to detect a potential effect of diverging cultural contexts. Note that, for this purpose, it is not particularly relevant whether or not these varieties reflect some kind of standard (in fact, including a standard variety might actually add another dimension of variation that is beyond the scope of this paper). Although the fact that the components have been compiled by different teams is certainly a disadvantage for register studies, it also means that team-specific preferences for selecting texts as representative of a given text category are balanced out to some extent by the preferences of the other teams.

In order to make sure that we target individual texts and not concatenated files consisting of multiple texts, possibly from different language users, ICE Corpus files were split into the individual "sub-texts" where necessary. We also removed material marked as extra-corpus text, which required some manual correction of the corpus annotation.[iv] Some of the resulting texts are very short, leading to inflated sampling variation of quantitative features, so texts containing less than 100 words or 10 sentences were removed from the corpus. The final data set consists of 2,844 texts containing a total of 3,352,389 words. Table 2 summarizes the distribution of texts across the three varieties and written and spoken modes.

Table 2. Distribution of texts across varieties and modes

	HK	JA	NZ
WRITTEN	701	539	431
SPOKEN	412	383	378
TOTAL	1,113	922	809
WORDS	1,184,422	1,045,567	1,122,400

3.2. Feature extraction

Our feature extraction pipeline requires a tokenized corpus annotated with sufficiently detailed part-of-speech tags (POS). For the purposes of this study, we use the POS-tagged versions of the ICE components provided by the ICE initiative.[v] The annotation was performed with the CLAWS tagger using the CLAWS7 tagset (Garside and Smith 1997) and thus provides us with the fine-grained categories needed in order to extract specific lexico-grammatical patterns with high accuracy. Tagging was performed on the original corpus files that include the ICE-specific mark-up (for a detailed discussion see Wong, Cassidy, and Peters 2011). The corpus files suffer from inconsistent use of the mark-up conventions (both within and across components), annotation represented as corpus text and simple mistakes in the mark-up such as missing opening or closing brackets (for a discussion of the different types of issues, see Neumann 2012, 80–82). For our pipeline, the mark-up was converted into well-formed XML

format. In this step, redundant information was removed (such as speaker identification by a separate tag, a common source of mark-up format errors, which is already included in the sentence/utterance ID), as well as any mark-up not needed in the feature extraction. Any remaining mistakes were manually corrected, based on XML validation to find inconsistent and overlapping tags. The result of this revision process is a lean version of the annotated corpus in a "vertical text" format that is compatible with many off-the-shelf NLP and concordancing tools. Finally, the corpus was indexed with the IMS Open Corpus Workbench (CWB; Evert and Hardie 2011) for interactive search and feature extraction.

Derivation of features follows the SFL conceptualization in that for each of the specific sub-dimensions of field, i.e. the aspect of the experiential world negotiated in a type of situation, tenor, i.e. aspects of situation types related to the interactants, their roles and relationships, and mode, i.e. the details of the means of communication and how exactly language is put to work, operationalizations are identified (for a detailed discussion of the relation between parameters, sub-dimensions and linguistic indicators, see Neumann 2014a; 2014b).

Our approach extracts the features thus derived with corpus queries formulated in the CQP query syntax supported by CWB. This strategy has several crucial advantages: (i) queries can be developed, tested and refined interactively using the CQP command-line tool or a Web-based frontend such as CQPweb; (ii) the queries provide a clear formal operationalization of all features and make the corpus analysis reproducible; (iii) the powerful CQP query language – with its macro substitution facility and support for matching against word lists – helps to break down complex queries into manageable units, forming a good compromise between an off-line grammar formalism and interactive queries (see Evert and The CWB Development Team 2020).

The feature extraction pipeline starts with a CQP script that automatically performs all individual queries and obtains per-text counts of the linguistic features, which are collected by a Perl script into a table with one row for each text and one column for each feature count. This table is further processed by an R script that normalizes feature counts to relative frequencies, with an appropriate unit of measurement chosen for each individual feature so that it can be interpreted as an independent choice within an envelope of variation. As Grieve-Smith (2007) explains, this approach reduces spurious correlations between features in multidimensional analysis. For instance, for each text the number of nouns is divided by the number of words, whereas the number of passives is divided by the number of finite verbs as a proxy for clauses (as passive voice is a feature of the clause, not of words). An example also given by Grieve-Smith (2007) is the count of first person pronouns, which is divided by the number of all pronouns rather than all words. This ensures that the features 'pronouns' (per word) and 'first person pronouns' (per pronoun) contribute complementary information to the multivariate analysis; if both counts used the same unit of measurement, they would be trivially correlated (a text that contains many pronouns in general will also contain a large number of first person pronouns).

There are 41 features in total, which are intended to cover a wide range of linguistic traits of texts, thus allowing a comprehensive linguistic characterization of each data point. Originally derived for SFL-based register analysis (see above), they cover information related to the register parameter field, for example, lexical density, nominalizations, neoclassical compounds, attributive adjectives and prepositions (all relative to the number of words), tenor-related information such as interrogatives and imperatives (relative to the number of sentences), titles per number of words and modal verbs per all verbs, and features linked to mode such as specific parts of speech in sentence-initial position as an approximation of theme, i.e. the local point of departure of the message of the clause. At this stage, the feature catalogue does not include verb classes. This represents a limitation of our coverage as verb classes have been linked to semantic differences between registers. The feature extraction scripts

are made available as part of the online supplement to this paper, which also provides access to visualizations, available at http://www.stefan-evert.de/PUB/NeumannEvert2021/.

4. Geometric multivariate analysis

The starting point of our analysis is a text-feature matrix with 2844 rows (texts) and 41 columns (features). All features are standardized to z-scores with zero mean ($\mu = 1$) and unit variance ($\sigma^2 = 1$), then an additional signed log transformation ($f(x) = \mathrm{sgn}(x) \cdot \log(|x| + 1)$) is applied in order to deskew the distributions, which reduces the influence of outlier texts on the multivariate analysis. The data set was checked for collinearities and excessive correlations with the help of a correlation matrix display. The pre-processing steps are detailed in the online supplement.

We perform a supervised LDA across the full data set (with all three varieties of English) in R (R Core Team 2018), using our 20 text categories as the target variable for supervised learning. This creates a perspective on register variation that is aligned with the sampling frame of the ICE Corpus, i.e. a map of the register space that can be interpreted by visualizing the arrangement and spread of text categories in the latent LDA dimensions. The LDA identifies 19 dimensions to separate the 20 text categories with an accuracy of 72.6% (i.e., 72.6% of all texts can be assigned to the correct category based on their coordinates in the 19-dimensional LDA space). Preliminary visual exploration of the scatterplots with text categories indicated by colors and varieties by symbols (see Figure 1) revealed salient and recognizable structures in the first four LDA dimensions that align fairly well with text categories, while higher dimensions show less distinctive patterns. Using only these four dimensions, 60.0% of the texts can still be assigned to the correct text category (compared to a baseline accuracy of 7.5% for random guessing, taking the uneven distribution across categories into account); the four dimensions thus account for a major portion of the registerial variation captured by the ICE text categories.[vi] In an extension of our previous methodology, we perform a PCA-based rotation in the first two dimensions in order to align the visual structure with the dimension axes. This simplifies the discussion of visualizations and the interpretation of the latent dimensions.

4.1. The first four dimensions of the LDA

Figure 1 shows a scatterplot visualization of the first four LDA dimensions, pairing dimension 1 with each of the other three dimensions. Every data point represents a single text from the corpus, split up into written (top row) and spoken (bottom row) texts. Different colors indicate different text categories, with rainbow hues assigned in the order shown in Table 1 so that text categories deemed similar by the corpus designers have similar colors. Keep in mind that LDA dimensions are ordered so that the first dimensions provide the clearest separation of categories, i.e. the highest ratio of between-category to within-category variance. This as well as all other later scatterplot matrices use the same scales and have 1:1 aspect between axes.

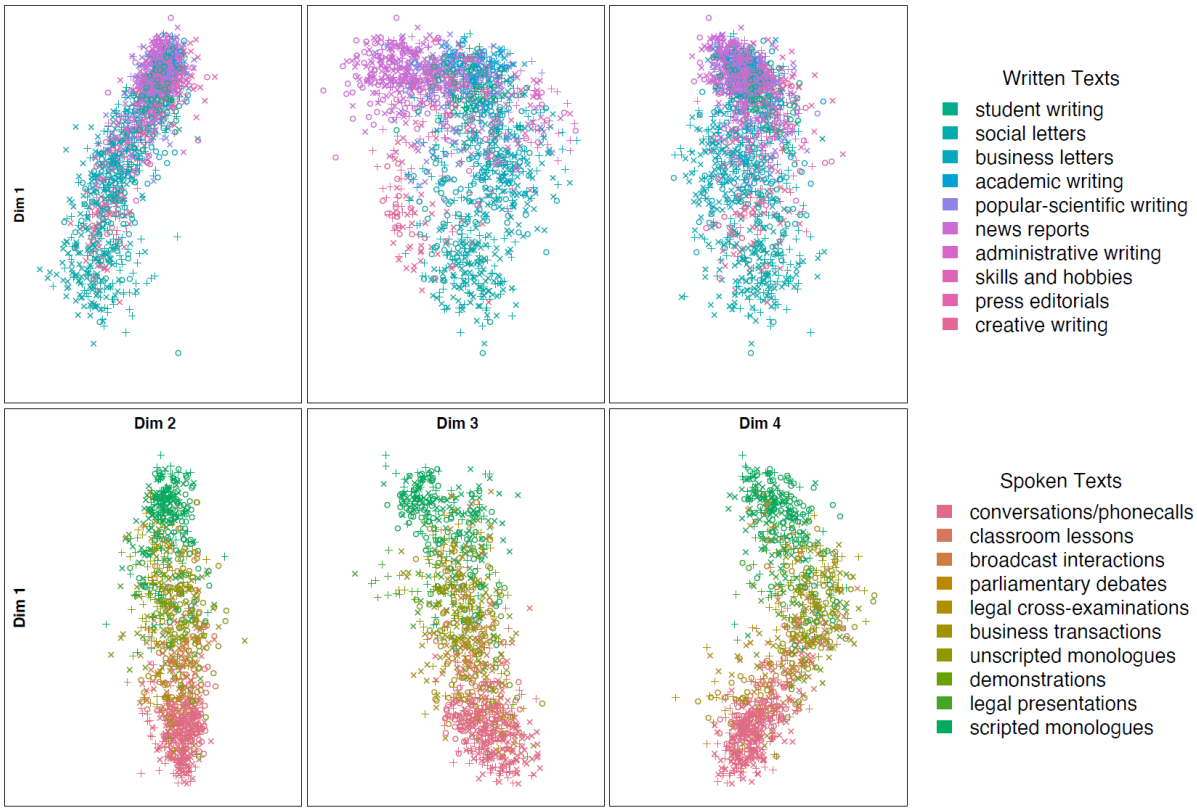

Figure 1. Scatterplots of LDA dimension 1 (vertical axis in both rows) against dimensions 2 (first column), 3 (second column) and 4 (third column) separately for written (first row) and spoken (second row) text categories across all varieties

The combined scatterplot of dimensions 1 and 2 for all texts yields a characteristic "inverted V" shape formed by the written and spoken texts (Figure 1obtained by overlaying the two panels in the left-most column of Figure 1, see also Figure 10). This shape can be inspected in an interactive, zoomable version produced with the web application framework Shiny (Chang et al. 2020) available as part of the online supplement: open the 'Scatterplot Viewer' and choose 'Preset S0' for the full scatterplot. Hover the mouse over the legend on the right-hand side to highlight the corresponding categories in the plot. Dimensions 3 and 4 improve the separation of written and spoken text categories, respectively, but produce a less salient structure than the first two dimensions.

A first general observation we can make based on the visualization of all data points in the scatterplots concerns the large amount of overlap between text categories. Exploring the complete data set in Preset S0 in the Scatterplot Viewer reveals that texts from different text categories frequently cover the same space. This is particularly obvious when selecting confidence ellipses in the viewer. It suggests that registers form a continuous space rather than discrete categories (cf. Biber 1989, 16). In fact, only extreme categories at the opposite ends of a dimension tend not to overlap. In order to make sense of the respective dimensions, the following analysis will foreground such extreme text categories.

Dimension 1 captures a wide continuum of variation with successive areas of overlap between text categories; news reports (*news* for short) are located at the extreme positive end within a cluster of similar text categories and conversations/phonecalls (*conv*) are the most extreme text category at the negative end of the dimension.

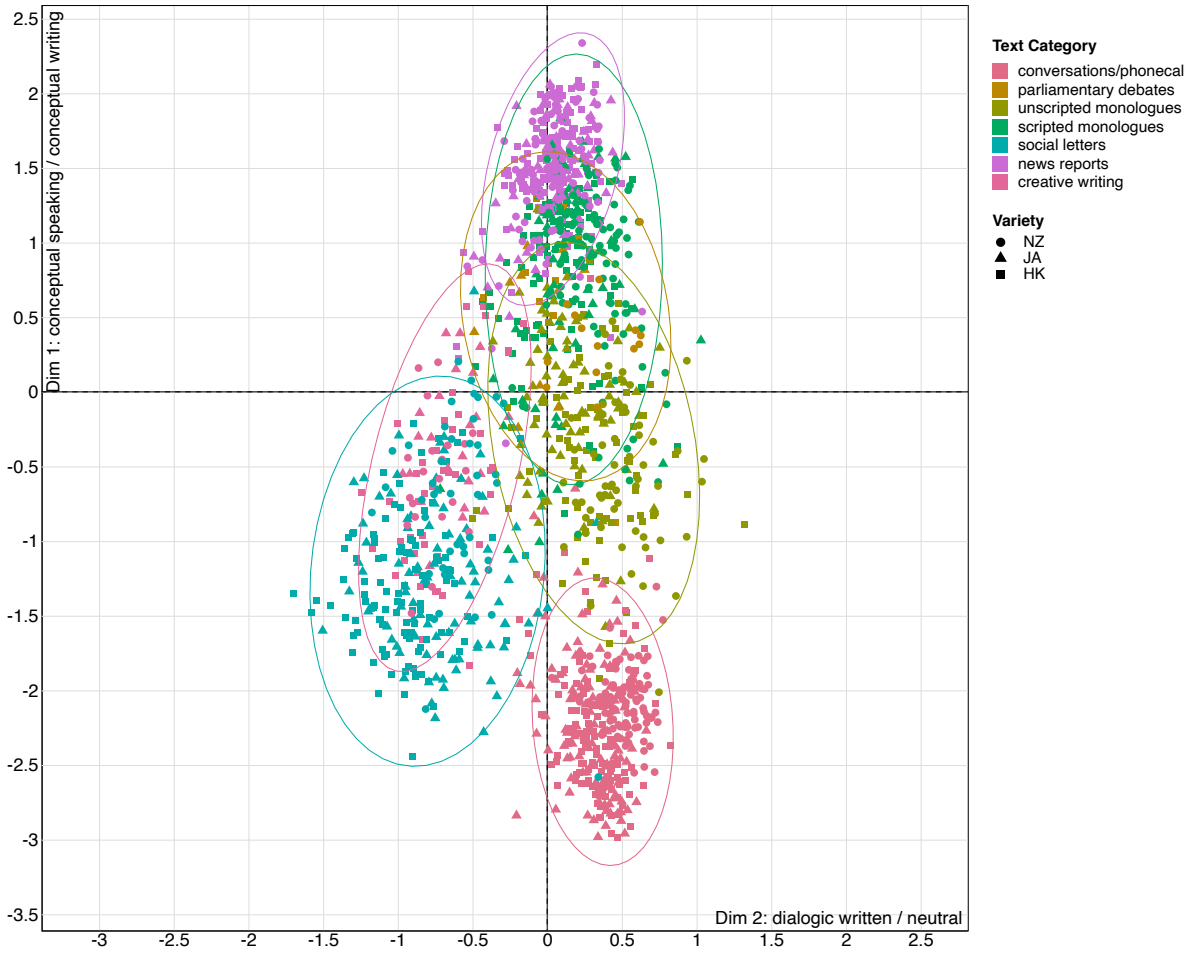

Figure 2. Scatterplot of dimensions 1 (conceptual speaking/conceptual writing; vertical axis) and 2 (dialogic written/neutral; horizontal axis) for the text categories conv, parl, unscr, script, socLet, news *and* creat *across all three varieties*

Despite the fact that these are prototypically written versus spoken text categories, visual inspection of the scatterplot (see Figure 2, available in the Scatterplot Viewer by selecting 'Figure 2' from the list of presets) suggests that we cannot characterize this dimension simply in terms of a contrast between spoken and written mode.[vii] The negative side of the dimension also contains written text categories, namely social letters (*socLet*) and creative writing (*creat*), whereas the positive side also contains spoken categories: The majority of texts classified as scripted monologues (*script*) can be found at values around and above +1.0 and a considerable amount of parliamentary debates (*parl*) and unscripted monologues (*unscr*) are located on this positive side of the dimension.

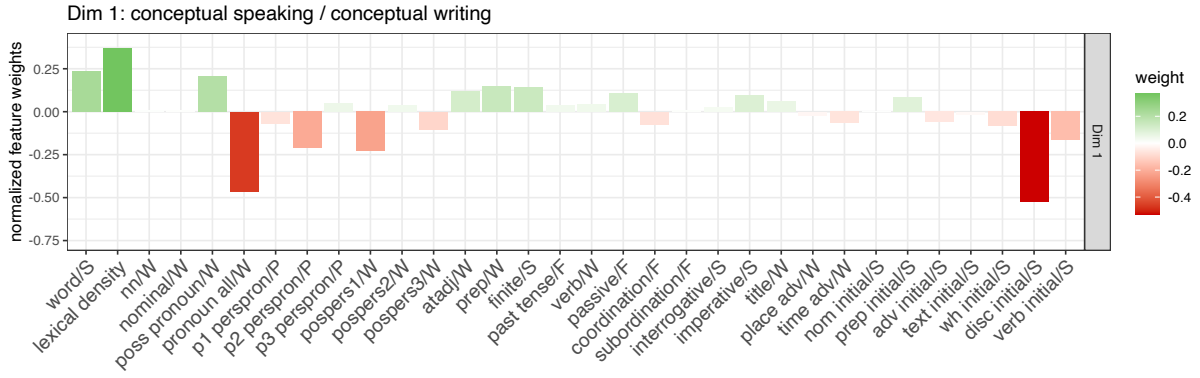

Figure 3. Normalized feature weights for dimension 1 "conceptual speaking/conceptual writing"

Figure 3 visualizes the normalized feature weights for dimension 1. For instance, the log z-score of lexical density is multiplied with +0.368, pushing texts with above-average lexical density towards the positive side of the dimension; the log z-score of pronouns (*pronoun all/W*) is multiplied with –0.465, pushing texts with many pronouns towards the negative side of the dimension. The position of each text on the dimension is the sum of such individual contributions by the features. While it is tempting to assume that conceptually written texts at the positive end of dimension 1 are characterized by long sentences (*word/S*), high lexical density, few pronouns and few sentence-initial discourse markers (*disc initial/S*), such a straightforward interpretation of feature weights is too simplistic, especially for LDA-based dimensions. Therefore, our discussion below takes the individual contributions of features into account, which can be visualized with box-and-whiskers plots for different text categories. These box-plots cannot be included in the paper for lack of space, but the online supplement offers an interactive 'Weights Viewer' along with a density plot indicating the distribution of texts on the dimension for the selected text categories. For example, Preset W1 (reproduced in Figure 4) shows feature contributions for text categories *conv* vs. *news. Word*/S, *lexical density*, *pronoun all/W* and *disc initial/S* strongly push *news* to the positive end and *conv* to the negative end of the dimension, with smaller contributions by features such as *prep/W* and *finite/S*. The vertical extent of the boxes and whiskers indicates variability within the text category, and boxes should always be read in comparison with the overall distribution shown in black (*other*). Keep in mind that negative feature weights (indicated by "(–)" in the label) invert the contribution: *conv* text are pushed towards the negative side by the feature *pronoun all/W* because they contain *more* pronouns than average texts.

The set of features with strong weight can be tentatively interpreted with respect to the absence of a shared production context. Shared context is an environment favorable for the feature carrying the strongest (negative) weight on this dimension, the proportion of pronouns per all words (see Biber et al. 1999, 1042), as it is likely that interactants refer to each other or to things accessible in the shared context with the help of syn-semantic items. This may even include using pronouns to refer to items only textually accessible, as shared context facilitates smooth negotiation of meaning. This is true for *conv* as well as broadcast interactions (*broadc*). In such contexts, discourse markers, particularly those at the beginning of sentences, are a useful means for managing turn-taking. Moreover, a low lexical density may occur in spontaneous interactions in which the elaboration and specification of meaning is avoided (Biber et al. 1999, 1044). Such an interpretation is corroborated by the boxplot (Figure 4 and Preset W1), as discussed above. Other features with similar opposite trends, albeit smaller contributions to the position on this dimension, include sentence length (*words/S*), second person pronouns relative to all pronouns (*p2 perspron/P*), first person possessive and personal pronouns *(pospers1/W*), the number of finite verbs per sentence (*finite/S*), the proportion of passives *(passive/F*), *wh*-elements in sentence-initial position (*wh initial/S*) and prepositions in sentence-initial position (*prep initial/S*).

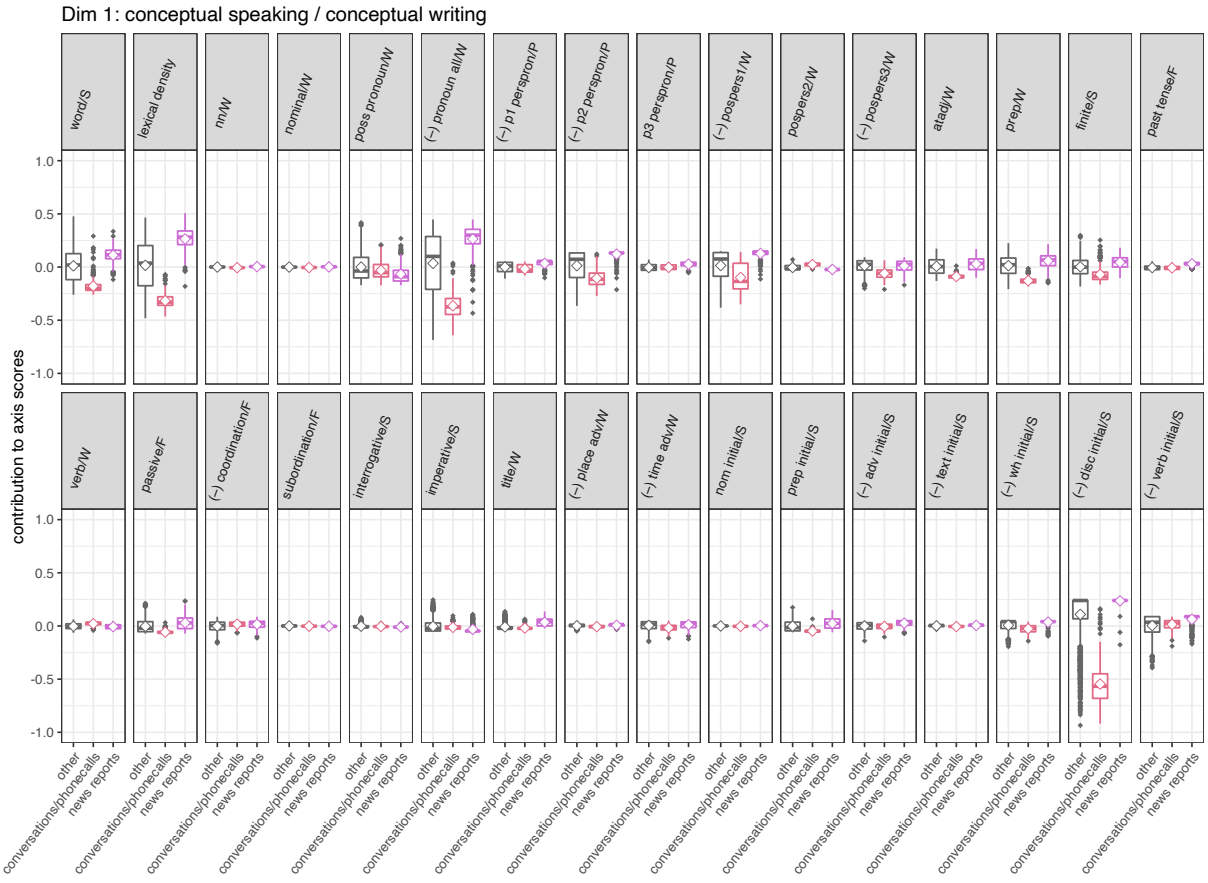

Figure 4. Box-and-whisker plots showing the contribution of each individual feature to the position of texts from the categories conv *and* news *on dimension 1*

Consequently, text categories on the positive side of this dimension are clear examples of non-situated interactions characterized by elaborated reference including features indicating richer lexical information such as attributive adjectives and prepositions (as indirect indicators of prepositional phrases), which receive smaller, but positive weights on this dimension. The text excerpt in (1)[viii] from a press editorial contains complex noun phrases and is generally quite dense. In fact, the first pronoun in this text is token number 168 (not included in the excerpt).

(1) **Home truths**

Even though mortgage rates have fallen to three per cent, the lowest in as many years, and flat prices have more than halved since the peak of 1997, tens of thousands of vacant flats are still looking for buyers.

No wonder the Government is said to be considering slashing the public sector's housing production target by one-third, from 45,700 flats a year to about 32,000. In fact, the original target has long been in effect ditched, and the current move is aimed at reconciling policy with reality.

But a deeper question that policy-makers need to ask is about the future of public housing. Hong Kong's public housing programme was conceived at a time when the territory was much poorer. (ICE-HK:W2E-001#1:1 - ICE-HK:W2E-001#6:1)

This linguistic characterization of the dimension explains why not only *conv* is to be found on the negative side of the dimension but also legal cross-examinations (*crossEx*) and *socLet*, which may, to some extent, reflect the interactivity of conversation. *SocLet* shares a lack of sentence-initial discourse markers (although with high variance) with prototypically written texts at the positive end, but display a

fairly strong contribution of (low) lexical density and a strong contribution of pronouns per words, which are apparently responsible for moving the texts in this category to the negative side. This makes sense, because the writers of social letters may share a considerable amount of context with their addressees, even if this is not a shared physical environment, as illustrated by example (2).

(2) Dear Jane and Phil,
My time with you was tremendous - so good to connect with you again. By the time I arrived back in Wellington, I felt as if I'd had a really good holiday.
Thank you so much for your hospitality - I did enjoy myself. (ICE-NZ:W1B-011#3:1- ICE-NZ:W1B-011#7:1)

Interestingly, Biber (1995, 148) reports generally similar trends for his dimension 1 "involved versus informational production", but only intermediate scores for personal letters and fiction, i.e. the text categories that clearly tend towards the negative side of our dimension 1.

The texts located at the positive side of this dimension are characterized by high lexical density and – still with substantial feature weight – long sentences (see Presets W2 and W2a). The text categories we find at the positive end in addition to *news* are press editorials (*prEdit*), academic writing (*acad*), popular-scientific writing (*popSci*), *script* and *parl*. The four written categories form a tight cluster of texts between +0.5 and +2.0 (see Preset S1). Although clearly located on this positive side of the dimension, the texts in the two spoken categories (*script*, *parl*) are located somewhat closer to zero (see Preset S2). These two spoken categories can be assumed to be based on heavily edited manuscripts which are read to the audience, thus making the texts in this category conceptually written (Koch and Oesterreicher 1985). We might therefore conjecture from this grouping of text categories that these texts are characterized by highly edited, elaborated style.

Summing up, the features mostly contributing to the location of texts on this dimension capture differences in terms of presumed/imagined or real access versus no such access to shared production context of the interaction. Halliday and Hasan (1989, 59) aptly characterize this presumed or imagined access as writing "*as if* I were talking to my friend" (italics in the original). In addition to this, the two most extreme text categories are distinguished by a whole range of features typically connected to written versus spoken mode. The dimension thus reflects what Koch and Oesterreicher (1985) describe as **conceptual speaking versus conceptual writing** (see also the differentiation of medium relationship in Gregory 1967, 189). In this sense, the dimension appears to pick up on some aspects that fall under Halliday's register parameter mode, i.e. the symbolic organization of the text (Halliday and Hasan 1989, 12).

There is much less variance on **dimension 2** (horizontal axis in Figure 2 and in Preset S0), but it is responsible for the characteristic V-shape of the data set. The most extreme text categories along this second dimension are *socLet* at the negative end and *unscr* at the positive end. In fact, spoken texts show very little variance on this dimension (see Preset S3), and spoken variation appears to be primarily captured by dimension 1 (and, to a smaller degree, dimension 4). Focusing on written texts, the four categories located mainly on the negative side of this dimension are, as mentioned before, *socLet*, *creat*, business letters (*busLet*) and skills and hobbies (*skills*).

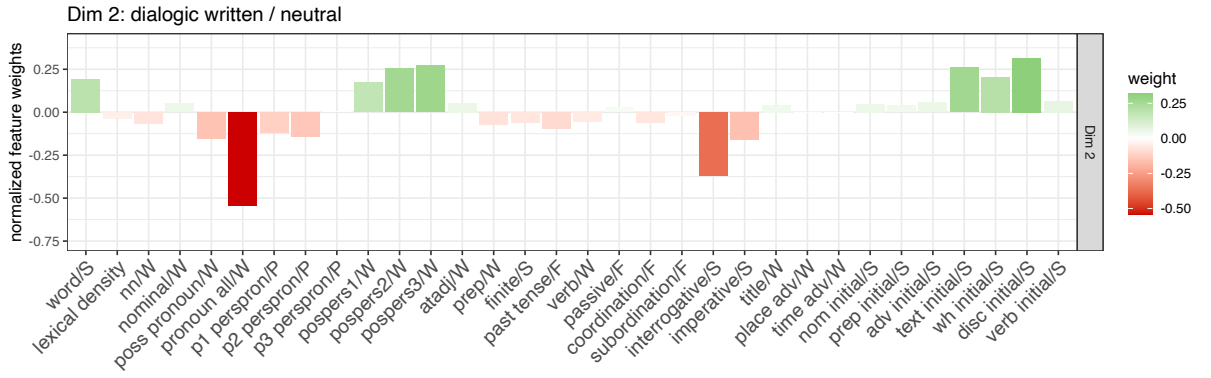

Figure 5. Normalized feature weights for dimension 2 "dialogic written/neutral"

Features with substantial weights on this dimension (see Figure 5) are all pronouns and interrogatives (*interrogative/S*). Additionally, discourse markers in sentence-initial position also have considerable weight (see Preset W3). Since texts deviate more to the negative side of this dimension than to the positive side, our discussion focusses on characteristic features and text categories on this side. We find a highly particular type of texts: written texts with situated, interactive features that we would prototypically expect in spoken mode (many pronouns and interrogatives). This is very compatible with the category of *socLet,* which, despite being produced in writing, can be assumed to be fairly dialogic, something which is otherwise more typical of dialogic texts such as conversations. It is also characteristic of novels and short stories reflecting the dialogic parts of creative writing (for a more detailed discussion of such aspects of literary texts, see Egbert and Mahlberg 2020). At the same time, these texts don't necessarily contain sentence-initial discourse markers, which have been linked to spoken texts above. It is the particular strength of multidimensional analysis (regardless of which exact flavor) to account for such specific aspects. Our approach – and in particular its visual representation of each individual data point – is able to capture the similarities between *socLet* and *creat* with respect to dialogic aspects of written texts, while also revealing their differences along other dimensions (see Presets S4 and W4 for a closer look at the similarities and differences between *socLet, creat, conv* and *unscr*). In short, dimension 2 picks up on (unusual) dialogic features of written texts, which are pertinent in *socLet*, see examples (3) and (4).

(3) Did I tell you she and Dave (boy-friend) are planning to cycle in (Mother takes a deep breath!) Pakistan, Tibet and China next May? And I thought I was adventurous when I was young! (ICE-NZ:W1B-011#65:2 - ICE-NZ:W1B-011#66:2)

(4) I don't want to bore you to death *or* make you think that I'm ABNORMAL!!! I bet Irene has already told you much about our weekly activities/gatherings. We all look forward to Friday every week, coz we can bust loose for a while amidst all the chaos & school work & mid terms! (ICE-HK:W1B-002#18:1 - ICE-HK:W1B-002#20:1)

This characterization also provides us with a possible explanation as to why some of the texts in *unscr* reach the most extreme positions on the positive side of dimension 2. They share one feature with strong weight unusual for prototypically spoken texts, namely a low frequency of pronouns per words. Texts in this category tend to be lectures or speeches which do not necessarily involve dialogue (see example (5)).

(5) Thank you Cedric for those very kind remarks /
As a former student / attached to Irvine Hall between / nineteen sixty-five and nineteen sixty-eight / and who was in fact on campus / during Sir Frank Worrel's tenure / as warden of Irvine Hall / it is my very special pleasure this evening / to been asked to give this the second

lecture in this Sir Frank Worrel Memorial / public lecture series /
As / Cedric I think clearly indicated / Sir Frank Worrel was not only a great cricketer / but he was also a great leader who inspired excellence from all who came in contact with him / (ICE-JA:S2A-023#1:1:A - ICE-JA:S2A-023#3:1:A)

To conclude, dimension 2 captures a particular feature combination in written texts in terms of dialogic character. All features that would otherwise distinguish between spoken and written language use such as frequent nominals or frequent verbal categories are 'muted' by the LDA. This levels out the differences between *conv*, *acad* etc. The dimension therefore has a one-ended character with dialogic written texts extending towards the negative end of the dimension and all other texts being mostly neutral. We might therefore label the dimension as **dialogic written versus neutral** and tentatively link this dimension to tenor-related variation according to Halliday's register parameters. This parameter concerns the kind of relationships that obtain between interactants (Halliday and Hasan 1989, 12).

Dimension 3 captures variation within the group of written text categories with administrative writing (*admin*) as the most extreme category on the positive side and *news* as the most extreme category on the negative side (see Figure 6, reproducible as preset 'Figure 6'). Note that two *news* texts can actually be found within the confidence ellipsis of *admin*. Consequently, not even the most extreme text categories are completely separated along this dimension.

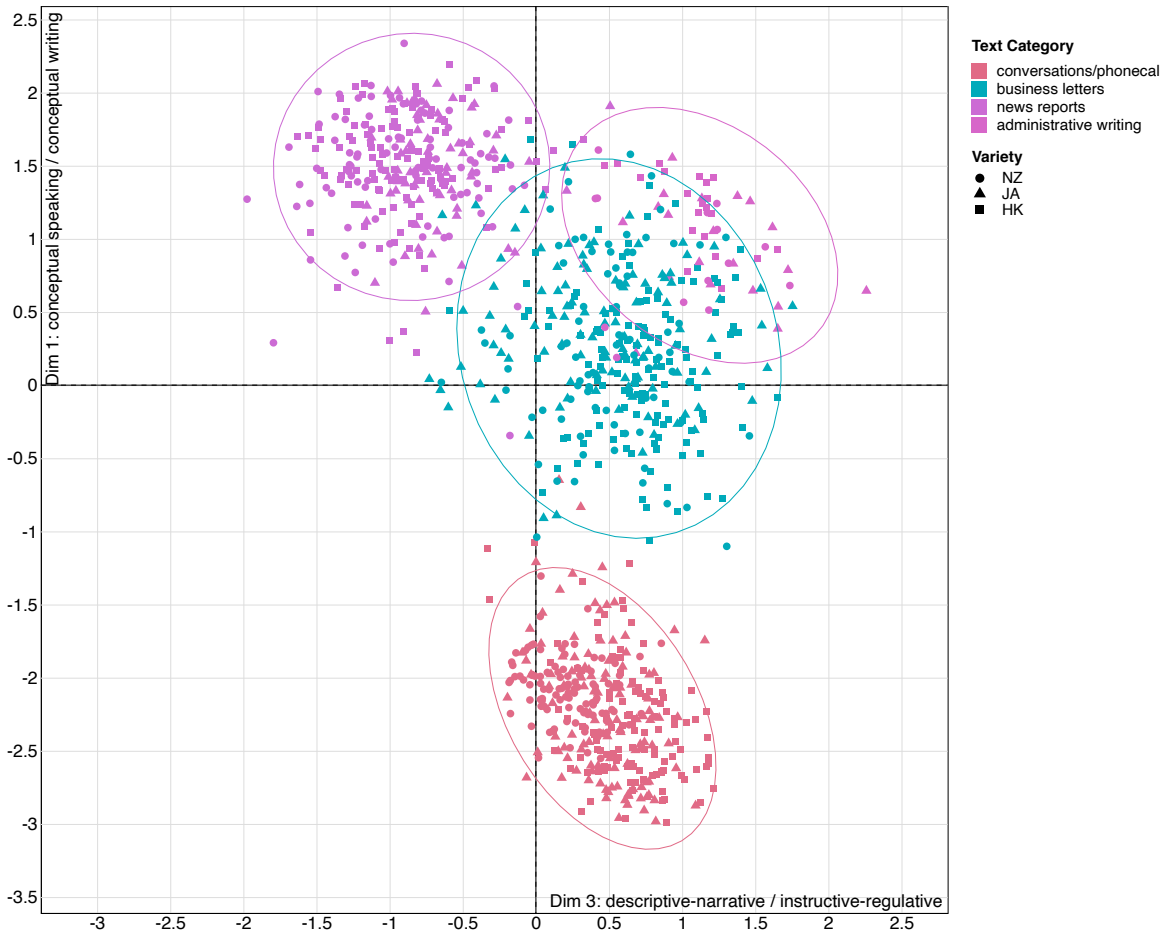

Figure 6. Scatterplot of dimensions 1 (conceptual speaking/conceptual writing; vertical axis) and 3 (descriptive-narrative/instructive-regulative; horizontal axis) for the categories conv, busLet, news, admin *across all three varieties*

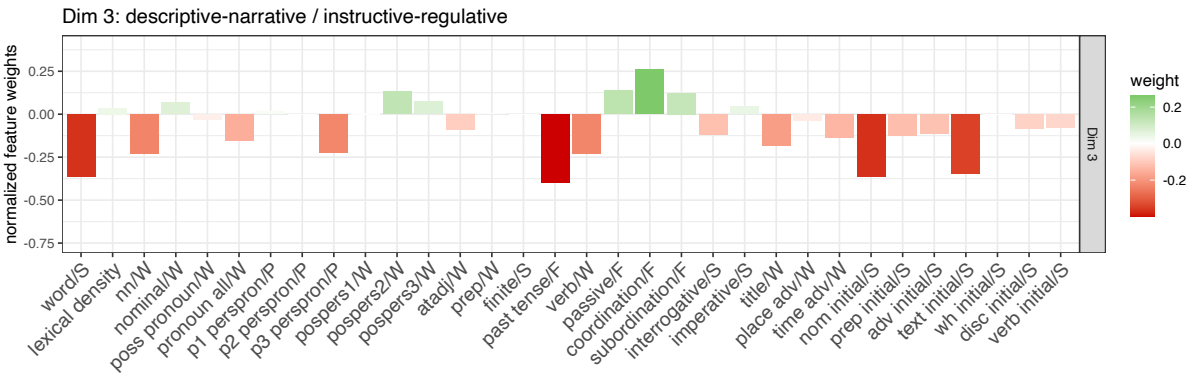

Figure 7. Normalized feature weights for dimension 3 "descriptive-narrative versus instructive-regulative"

As can be seen in Figure 7, features with strongest weight on the positive side of this dimension are the relative absence of past tense verbs, of nominal and textual elements in sentence-initial position and relatively short sentences. Additionally, the occurrence of coordinating conjunctions per finite verbs receives a moderate positive weight (see Preset W5). Especially the preference for present over past tense lends itself to an interpretation of this dimension. Texts on the positive side of the dimension tend to express contents independently of an explicit location in time, which is further corroborated by the relative absence of time adverbs – although this feature contributes very little to the dimension. Intuitively, it does not come as a surprise that *busLet* is the second most extreme category on the positive side. We may assume that *busLet* shares some characteristics with *admin* in terms of a fact-oriented style drawing on reduced verbal routines. In fact, the contribution of a low number of words per sentence is even stronger than for *admin* and the relative absence of past tense verbs receives almost the same weight as for *admin*. Sentence-initial textual elements are similarly infrequent, but the contributions of sentence-initial nominal elements and coordinating conjunctions per finites diverge. To some extent this can be explained by the use of reduced formal features such as letter heads, as illustrated by example (6).

(6) **MEMORANDUM**
TO: Professor Michael Morgan
Head, Language and Linguistics
FROM: Derek Shaw
DATE: November 7, 2000
RE: Project Implementation Proposed Computer writing Centre
We have concluded our preliminary costing and scheduling for implementation of the Proposed Computer Writing Centre in the Department of Language and Linguistics. (ICE-JA:W1B-016#1:1 - ICE-JA:W1B-016#7:1)

News, the category most clearly located on the negative side of this dimension, has some commonalities with *admin* in terms of a higher frequency of nouns, and a relative absence of verbs and sentence-initial textual elements, but differs specifically with respect to three of the features with the strongest weight. *News* has many verbs in the past tense, long sentences and frequent sentence-initial nominal elements (see example (7)). Also the number of coordinating conjunctions per finites differs from *admin*. Arguably, the temporal sequence facilitated by the more frequent use of past tense verbs allows writers to leave more logico-semantic relations between sentences implicit.

(7) Political debates urged
Calls to shift campaigning from 'entertainment' to substance

A CONCERTED appeal has been made for election campaigning to take on a new focus in Jamaica, with meaningful debates among candidates of all contesting political parties. Newly-appointed President and Chief Executive Officer of Crown Eagle Life Insurance Company, Geoffrey Messado, The Press Association of Jamaica (PAJ) and the Christian United Party (CUP) are leading the push for national debates leading up to what is expected to be an early general election. (ICE-JA:W2C-002#73:4 - ICE-JA:W2C-002#76:4)

We may thus tentatively label this dimension in terms of goals language users pursue based mainly on the contribution of past tense verbs with texts describing events in the factual or imaginary past on the negative side of the dimension. Texts on the positive side do not explicitly locate events in time (and space). Furthermore, they have a lower number of words per sentence which, in this case, cannot necessarily be linked to typical features of spoken language (although it probably explains the location of the main part of texts from several spoken categories on the positive side), but rather captures the reduced, formal style of regulative administrative writing. This is further corroborated by the contribution of a low number of verbs per words in *admin*, which suggests that the texts may contain a fairly high share of incomplete sentences as illustrated by the headings in example (8).

(8) 3. PERSONAL GRIEVANCES
3.1 MAIN FEATURES

- A personal grievance may arise where an employee has been dismissed unjustifiably, or has been subject to discrimination, sexual harassment, duress or other unjustifiable action. The remedies available for a proven personal grievance include reinstatement, reimbursement for lost income and compensation. (ICE-NZ:W2D-008#1:1 - ICE-NZ:W2D-008#4:1)

Note that *admin* is not necessarily characterized by many imperatives, the prototypical feature of instructive texts. Although the category displays more variance with respect to this feature with some texts clearly containing a good deal of imperatives, the feature does not display a notable weight (see Preset W6). In fact, texts in the category *skills*, many of which are cooking recipes, contain more imperatives thus reflecting this key aspect of instructive texts more clearly. The majority of texts in this category can also be found on the positive side of dimension 3 (see Preset S5).

To conclude, we might characterize the dimension in terms of rich descriptive, time-bounded versus reduced formal style that might reflect different goals, namely **descriptive-narrative versus instructive-regulative**. This orientation towards the speakers' goals suggests that the dimension can be linked to Halliday's register parameter of field, that is the nature of the social activity in terms of both the acts carried out and their goals (Halliday and Hasan 1989, 56).

As can be gathered from the scatterplot in Figure 8, the most extreme text categories on **dimension 4** are *unscr* and *busLet*, with all unscripted speeches on the positive side of the dimension except for three outliers. *BusLet*, *crossEx* and a considerable number of texts in *conv* are located on the negative side of this dimension (see also Preset S6 including *news*). The majority of all texts in the spoken categories are on the positive side of this dimension, whereas the majority of written texts are on the negative side. Generally, the spoken texts account for most of the variance along this dimension.

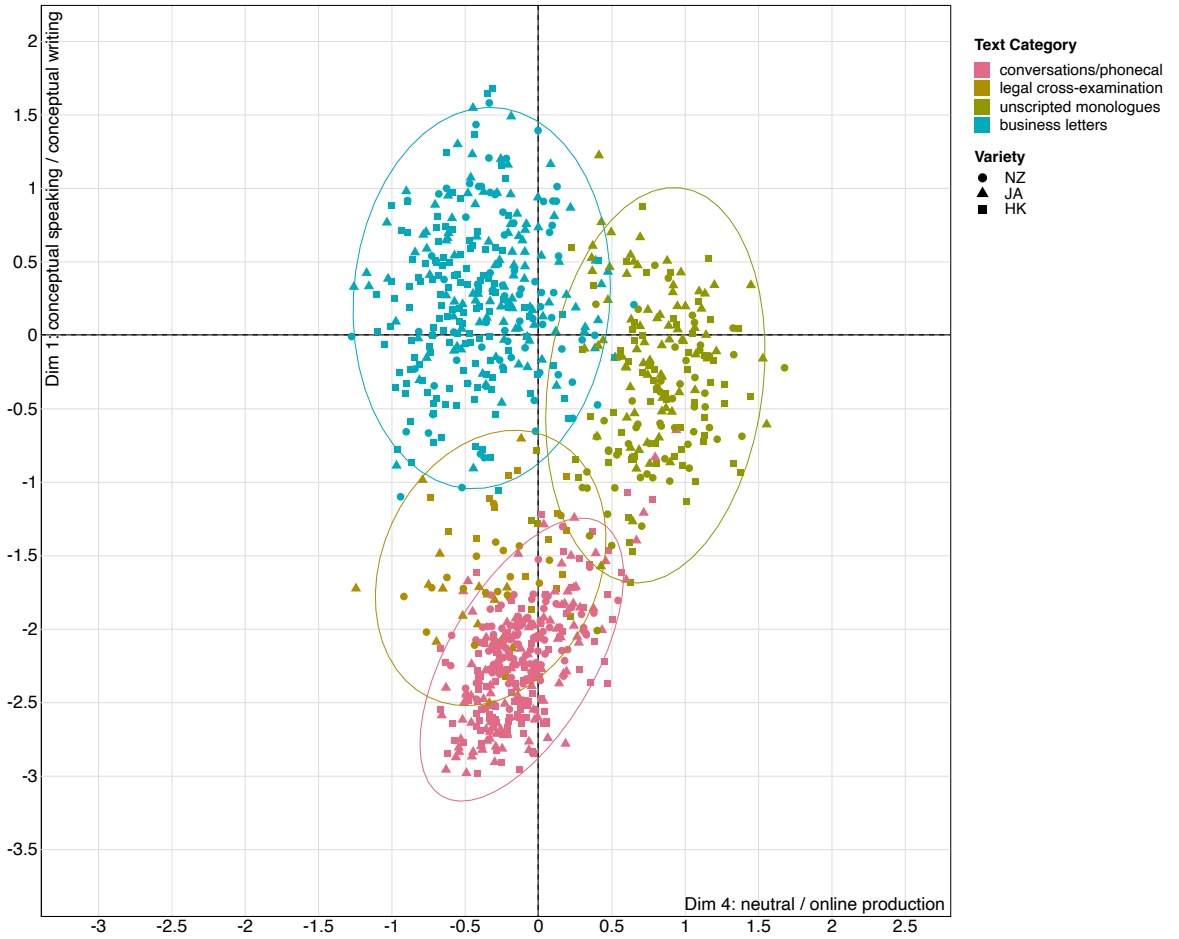

Figure 8. Scatterplot of dimensions 1 (conceptual speaking/conceptual writing; vertical axis) and 4 (…/online production; horizontal axis) for the categories conv, crossEx, unscr, busLet *across all three varieties*

The feature with strongest weight pushing texts to the positive side of the dimension (see Figure 9) is the proportion of finites per sentence, i.e. a proxy of the relative number of clauses and hence an indicator of what Halliday calls grammatical intricacy calculated as the number of ranking clauses in the clause complex (Halliday 2009, 76). This is further corroborated by a strong weight of low lexical density. Interestingly, texts on the positive side of the dimension are also characterized by a relative absence of second person pronouns per words (see Preset W7).

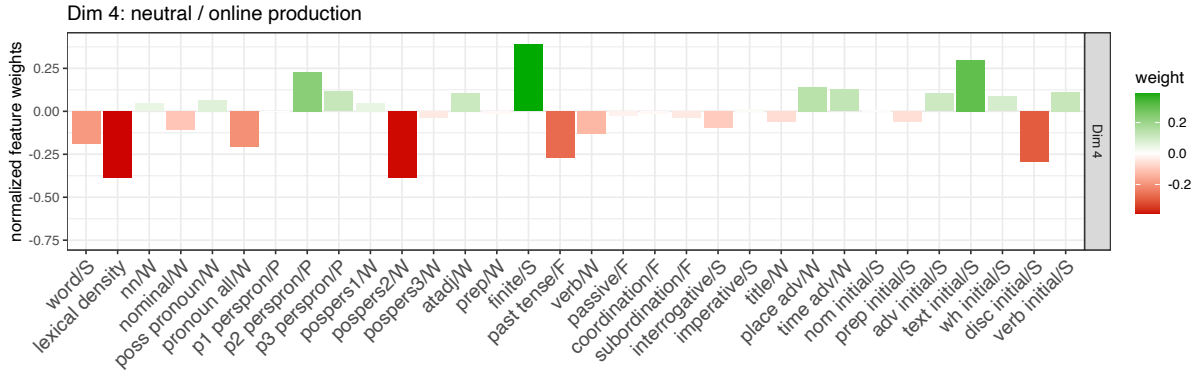

Figure 9. Normalized feature weights for dimension 4 "neutral/online production"

Taking into account the extreme categories on the negative side, this feature makes sense: *crossEx* (as well as *conv*) can be assumed to be characterized by a relatively high frequency of second person pronouns, unlike the speeches contained in the category of *unscr* (see Preset W8). Also, *busLet* – despite its reduced formal characteristics brought out by dimension 3 – can be expected to contain direct forms of address, as illustrated by example (9).

(9) Dear Mr. Pabst,
Thank you very much for your confirmation.
We will run your advertisement (2 columns x 4 cm) into the International Classified Section of our March 24th issue. (ICE-HK:W1B-016#160:7 - ICE-HK:W1B-016#162:7)

The features contributing particular weight to the positive side of this dimension seem to capture the pressure of online production, namely texts fragmented into many clauses and generally low lexical density. Arguably this is particularly obvious in monologic texts such as unscripted speeches, which are not characterized by a high degree of interactivity as illustrated by the pauses in example (5) above. At the same time, sentence-initial textual elements contribute considerable weight. This makes sense again for speeches, which are fragmented into many short clauses because they are produced on the fly and may therefore need to make rhetorical relations explicit.

Closer examination of feature weights reveals an interesting effect for individual text categories. While written texts are mostly located on the negative side of dimension 4, some texts also display features which push them to the positive side. This is most apparent in the category *skills*. Despite being classified as a written category, the majority of texts are located on the positive side of the dimension (see Preset S21). Inspection of feature contributions shows that features with moderate weight such as second person pronouns and the relative absence of past tense verbs (see Preset W26) contribute to the position on the positive side, whereas the features with the highest weight (and in particular those that account for the extreme location of *unscr*) hardly come into play. In the light of this observation, it would obviously be wrong to conclude that *skills* is similar to, but only less extreme than *unscr*. Rather, a different combination of features is responsible for its position on the positive side of the dimension. Focusing exclusively on dimension weights thus obscures the differences between these two text categories.

To sum up, a possible label for this dimension is **neutral versus online production** (Biber et al. 1999, 1066–72). Like dimension 2, this dimension is one-ended, that is, it displays variance in one direction, but is neutral in the other direction as reflected in texts located near zero. We therefore restrict the label to the positive side of the dimension. The combined effect of features reflecting online production pressure (short sentences, low lexical density, many finite clauses linked with textual elements) and avoidance of direct orientation towards the addressee (low frequency of 2nd person pronouns) suggests that dimension 4 can be linked to both Halliday's mode and tenor.

Table 3. Overview of dimensions, influential features and text categories

	Dimension label	**Influential features**	**Text categories representative of the poles**
1	Conceptual speaking/conceptual writing	disc initial/S, pronoun all/W, lexical density, word/S	*conv – socLet – script – news*
2	Dialogic written/neutral	pronoun all/W, interrogative/S, (disc initial/S, pospers3/W)	*socLet – creat –* (*unscr*)
3	Descriptive-narrative/instructive-regulative	past tense/F, nom initial/S, word/S, coordination/F	*news – busLet – admin*

4	Neutral/online production	finite/S, lexical density, text initial/S, (disc initial/S)	(*busLet*) – *unscr*

Table 3 provides a breakdown of the four dimensions discussed in this section along with text categories representative of the end poles of each dimension and the features contributing most to the position of texts at these poles. Note that the features are characteristic only of the text categories mentioned here (features in brackets only for a subset of these categories), and that we cannot deduce a general characterization of the respective dimension from these features. Even individual texts belonging to the text categories mentioned in Table 3 may not necessarily contain the characteristic frequency of an influential feature. Future work involves further examining the exact contribution of features across text categories located in the same region of multidimensional space. As we hope to have shown in the discussion, features contribute in complex ways to the position of individual texts along the dimensions.

4.2. Observations about variation across the three varieties

So far, we have examined overall groupings of texts across the four dimensions. Let us now inspect the variation of text categories across the three varieties included in the analysis, as previous studies suggest that some clusters are influenced by specific constellations in one of the three varieties (see Section 2).

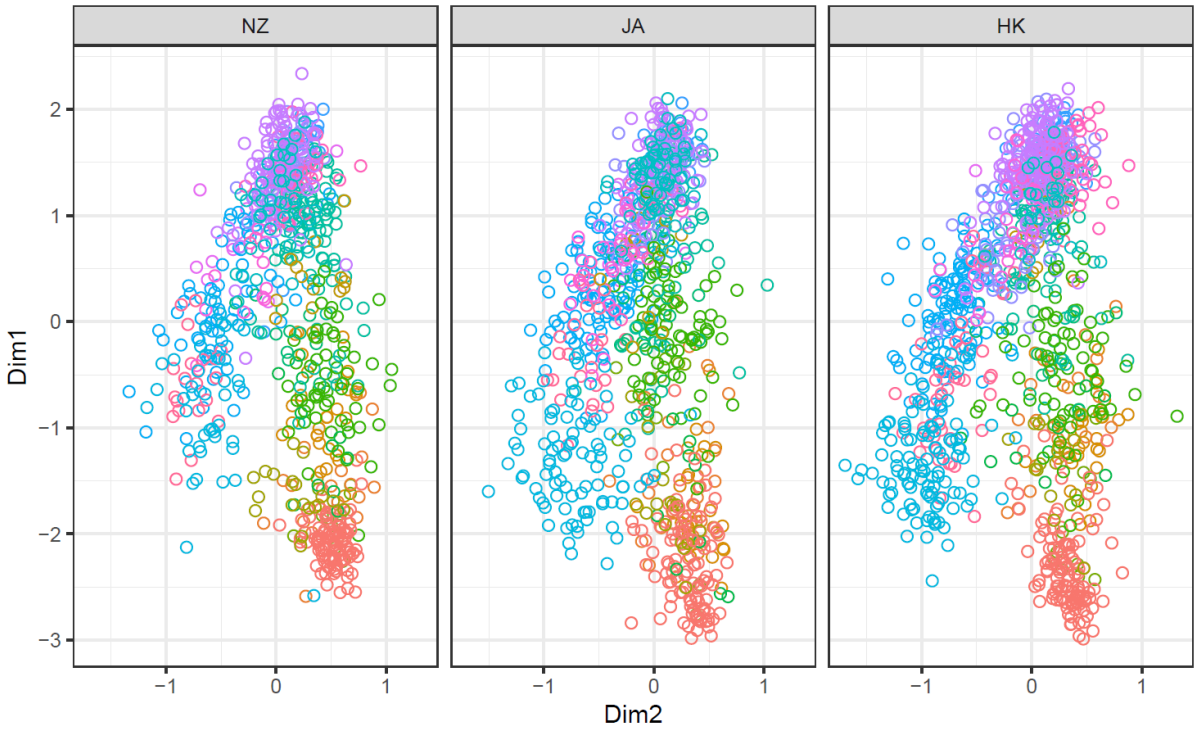

Figure 10. Scatterplots of dimensions 1 (vertical axis) and 2 (horizontal axis) for all text categories by variety

Variance between NZ texts on dimension 1 (vertical axis in Figure 10 and Presets S7–S9) is clearly smaller than between texts in the two other varieties. Compared to the L2 varieties, NZ English also has fewer extreme cases of what we labelled conceptually spoken texts. With the exception of one outlier, texts can be found in a region between –2.6 and +2.0 (see Presets W9–W11). Jamaican and Hong Kong English have clearly more extreme texts at the conceptually spoken end. This observation can be seen as corroboration of Kruger and Van Rooy's (2018, 237) conjecture that spoken language may exhibit more pronounced differences between varieties.

On dimension 2, New Zealand and Jamaican English are similar in variance, with Jamaican texts altogether somewhat shifted to the negative, i.e. marked end of this dimension. Interestingly, Jamaican texts reduce the difference between the marked, i.e. dialogic written and other texts. Therefore, the V-shape of the data set is less clearly visible for texts in this variety, mainly due to the fact that the spoken texts are located closer to zero or even on the negative side of the dimension. This means that the one-endedness of the dimension is even more pronounced in the Jamaican component: either texts are marked as dialogic written or they are entirely neutral. Compared to the other two varieties, the Hong Kong component displays slightly more variance with more extreme cases on the negative side of the dimension (see Presets W12–W14). Dialogic written texts appear to be less unusual in this variety.

It is the New Zealand texts which display a larger variance than the other two varieties on dimension 3 (see Presets S10–S12, W15–W17), with a tendency to allow more extreme positions on the negative, i.e. descriptive-narrative side of the dimension. Jamaican English differs from the other two varieties in that more texts are moved to the positive, i.e. instructive-regulative, end of the dimension. Rather than claiming that Jamaican texts are less descriptive-narrative and more instructive-regulative, we might assume that either the features used for the analysis across all varieties fail to capture these goals as expressed by Jamaican language users optimally or that Jamaican language users do not necessarily use the particular lect represented in the International Corpus of English.

Hong Kong English displays the widest variance on dimension 4. New Zealand English texts appear somewhat more neutral on the negative side and include more extreme texts on the positive side of the dimension with Jamaican texts being located between the texts from the two other varieties and displaying the same variance as NZ English (see Presets S13–S16, W18–W20). Possibly, dimension 4 captures aspects of New Zealand unscripted spoken texts better than of the other two varieties.

On the whole, the three varieties differ in variance across the four dimensions. The New Zealand texts display the smallest and the Hong Kong texts the largest variance overall. Particularly the latter may be interpreted in terms of less established conventions in the individual registers, which therefore leave more room for individual variation. With the Jamaican component figuring somewhere in between, we might tentatively interpret this with respect to the specific language constellation in each of the three regions: Compared to Jamaica and Hong Kong, English simply plays a more important role in New Zealand. In Jamaican English, we seem to observe a more differentiated situation, possibly with some division of labor between the various lects. And lastly, Hong Kong English seems to display more free variation in a linguistic context mainly driven by a *different* language, viz. Cantonese.

4.3. General discussion

Summing up, linear discriminant analysis based on information about 20 text categories (see Section 3.1) provides us with a number of clear indications of differences between groups of texts, but notably also of similarities and overlaps between groups. In fact, registers approximated by the exploration of ICE text categories form a continuous space, an observation corroborating Biber's related claim in the discussion of text types in English (Biber 1989, 41). The range of variation between texts classified as belonging to one text category (and arguably to one register) underlines the individuality of the meanings expressed in each text as reflected in similar, but not identical lexico-grammatical choices. The picture emerging from this analysis is one that characterizes individual text categories in terms of specific linguistic profiles.

An interesting observation is that the analysis brings out clusters of text categories. The cluster on dimension 1 consisting of the five text categories *news*, *acad*, *prEdit*, *popSci* and student writing (*stuEss*; see Preset S17), for instance, could suggest that most texts of the superordinate ICE category 'printed' plus one category under 'non-printed' can simply be grouped together. However, the cluster

is differentiated on dimension 3, which picks up on variation in the written text categories. This is in line with the above claim about register profiles emerging out of specific properties along different dimensions: it is to be expected that registers will display linguistic patterns reflecting similarities along some sub-dimensions of register parameters while differing for other sub-dimensions. This is what we see on dimension 1: as to conceptual writing, the categories in the cluster plausibly don't differ. But they do differ (in interesting ways) along dimension 3 interpreted as capturing field-related variation. Nevertheless, it should also be kept in mind that our query script currently does not include verb classes. It is possible that differences in verb classes would be picked up by the LDA and thus yield an additional dimension that discriminates these written, fact-oriented texts.

A plausibility check of the dispersion of text categories suggests that most categories are fairly compact showing consistent style with respect to the four dimensions, but business transactions (*busiTr)* show conspicuous outliers. On dimension 1, 10 out of 12 Jamaican *busiTr* texts alone are located between –0.2 and –1.4 with the exception of one HK outlier text. The ICE-JA component is thus responsible for the large dispersion of data points (see Presets S19 and S19a). Meta information about Jamaican texts in this category suggests that this is a compilation effect as nine of the above 10 texts appear to be from the same source. Inspection of the dispersion of individual text categories across the four dimensions also yields further interesting patterns, discussion of which is beyond the scope of this paper.

Given the corpus compilation of the International Corpus of English, we have carefully avoided referring to the ICE text categories as registers. What does this kind of analysis then add to the body of knowledge about register variation? In our previous discussion, we linked our interpretation of the dimensions to Hallidayan register parameters. On the whole, we would claim that the text categories reflect aspects of situational variation. Given the tendency of the analysis to pick up on mode-related aspects of the data set, we interpret our findings as corroboration of Biber's (1995) dimension 1 discriminating between involved and informational production (see also Neumann 2020). We should have found a similar dimension with an LDA solely for spoken text categories, but would have lost dimension 2, which separates the "informal" written from spoken texts. Only the third dimension focuses mostly on characteristics pertaining mainly to written texts. Interestingly, unlike the extreme registers on Biber's dimension 1, the extreme text categories on the first dimension of our LDA don't reflect a straightforward distinction between speaking and writing,[ix] as texts produced in either modality can be found towards both ends of the dimension. Rather, our dimension 1 reveals a more intricate aspect of mode, viz. the cross-classification of channel and medium (Halliday and Hasan 1989, 59; Gregory 1967, 188–94). This cross-classification is also observed by Biber (1995, 148) along his dimension 1, but only for texts yielding intermediate scores on this dimension.

Arguably, the analysis suggests that register parameters capturing relevant aspects of the situational context involve some internal hierarchy with variation in mode being a watershed parameter. Furthermore, we would tentatively claim that tenor-related, or – in more traditional linguistic terms – pragmatic aspects contribute more to the variation observed in the data set than field-related aspects. To some extent this may be due to the choice of linguistic features included in the analysis, viz. the limited lexical characterization of the texts, but it also appears plausible to assume that speaking versus writing as well as social relationships between the interactants strongly impact the way we communicate about one and the same field. In addition to the mode-related structure of the data set, the analysis also reveals some interesting commonalities between spoken and written categories. These commonalities further corroborate the multidimensionality of linguistic variation.

Like any exploratory multivariate analysis, the dimensions emerging from our study are also only valid for the linguistic features included in the analysis and of course depend on the precision of the script-based counts. This means that – despite all care that we have given to the accuracy of our analyses –

we expect changes in the location of individual (groups of) texts on the dimensions, in the feature weights and consequently in the interpretation of the dimensions, if more or different features are included in the analysis. This obviously also applies to the composition of the data set. Although the International Corpus of English has proven useful for this type of analysis, it also has some clear limitations. Particularly the selection of texts for each component is a potentially distorting factor: the decision to include particular texts as exemplars of a particular text category may be based on somewhat divergent criteria across components resulting in higher variance between components than in the underlying population of texts. At the same time, compilers of individual components (and actually, more generally, of corpora) may tend to include very similar texts as exemplars of a particular situational context, thus not capturing the real variance of that situation type.

More generally, we are painfully aware of the heavy impact corpus design has on this (or any) kind of exploratory multivariate analysis. The inclusion of specialized text categories such as instruction manuals has been shown to exert a profound effect on the latent dimensions in a multivariate analysis (see Diwersy, Evert and Neumann 2014). This is likely to be true not just for LDA but also for factor analysis and similar techniques used in the multivariate approach and serves as a reminder that findings yielded by exploratory analyses cannot be used as a reliable reference for other data sets.

5. Conclusions

In this paper, we discussed geometric multivariate analysis (GMA) as an approach to the study of register variation across varieties of English. Its geometric nature and use of orthogonal projections enables a meaningful visual inspection of the distribution patterns of individual texts in multidimensional space. The visualization allows us to inspect the position of each text in relation to all other texts. As a consequence, it reveals similarities and more general continuities between groups of texts as well as dimensions that differentiate between such groups of texts. This is in contrast to approaches that focus exclusively on group averages, foregrounding differences between groups. We also discussed a new way of interpreting the contribution of individual linguistic features to the characterization of groups of texts in terms of register.

The approach allowed us to examine register variation mapped onto the sampling frame of the ICE Corpus, that is the register space interpreted by visualizing the arrangement and spread of text categories in the latent LDA dimensions. The analysis also threw into relief variation between the register space reflected in the text categories of the individual ICE components. Although the tentative interpretation of the difference in variance between the Hong Kong, Jamaican and New Zealand texts appears plausible and corroborates findings by previous multivariate studies, the more general question of the interaction between cultural and situational context will still require further investigation. The challenges of compiling a corpus adequately representing the situation types for each respective cultural context while still keeping the overall data set comparable remain. Not just the comparability of ICE components represents a limitation of our study, but also the coverage of features. While we believe that our current feature set captures important aspects of variation, future work will involve further refining the query script in continued alignment with the register-theoretical constructs on which it is based.

The growing body of empirical evidence of situation-dependent patterning of language use leaves no doubt of the relevance of register for language theory. We believe that the inclusion of texts across varieties of English additionally underlines the importance of accounting for more generalized patterns of variation in language theory as well.

Acknowledgments

We would like to thank the two anonymous reviewers as well as Doug Biber for valuable comments which greatly helped to clarify the argument of the paper.

References

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. 'Stylistic Text Classification Using Functional Lexical Features'. *Journal of the American Society for Information Science and Technology* 58 (6): 802–22. https://doi.org/10.1002/asi.20553.

Biber, Douglas. 1986. 'Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings'. *Language* 62 (2): 384–414.

———. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

———. 1989. 'A Typology of English Texts'. *Linguistics* 27: 3–43. https://doi.org/10.1515/ling.1989.27.1.3.

———. 1995. *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

———. 2019. 'Text-Linguistic Approaches to Register Variation'. *Register Studies* 1 (1): 42–75. https://doi.org/10.1075/rs.18007.bib.

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

Biber, Douglas, and Edward Finegan, eds. 1994. *Sociolinguistic Perspectives on Register*. New York/Oxford: Oxford University Press.

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

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson. 2020. *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Diwersy, Sascha, Stefan Evert, and Stella Neumann. 2014. 'A Weakly Supervised Multivariate Approach to the Study of Language Variation'. In *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*, edited by Benedikt Szmrecsanyi and Bernhard Wälchli, 174–204. Linguae & Litterae. Berlin/New York: de Gruyter.

Egbert, Jesse, and Biber, Douglas. 2018. 'Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis'. *Corpus Linguistics and Linguistic Theory* 14 (2): 233–273. https://doi.org/10.1515/cllt-2016-0016

Egbert, Jesse, and Michaela Mahlberg. 2020. 'Fiction – One Register or Two?: Speech and Narration in Novels'. *Register Studies* 2 (1): 72–101. https://doi.org/10.1075/rs.19006.egb.

Ervin-Tripp, Susan. 1972. 'On Sociolinguistic Rules: Alternation and Co-Occurrence'. In *Directions in Sociolinguistics. The Ethnography of Communication*, edited by John J. Gumperz and Dell H. Hymes, 213–50. New York: Holt, Rinehart and Winston.

Evert, Stefan, and Andrew Hardie. 2011. 'Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium'. In *Proceedings of the Corpus Linguistics Conference 2011, University of Birmingham, UK, 20-22 July 2011*. Birmingham, UK: University of Birmingham. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf.

Evert, Stefan, and Stella Neumann. 2017. 'The Impact of Translation Direction on Characteristics of Translated Texts. A Multivariate Analysis for English and German'. In *Empirical Translation Studies. New Theoretical and Methodological Traditions*, edited by Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 47–80. Berlin: de Gruyter.

Evert, Stefan, and The CWB Development Team. 2020. *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial* (version CWB Version 3.5). http://cwb.sourceforge.net/files/CQP_Tutorial/.

Garside, Roger, and Nicholas Smith. 1997. ‘A Hybrid Grammatical Tagger: CLAWS4’. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech, and Anthony McEnery, 102–121. London: Longman. http://ucrel.lancs.ac.uk/papers/HybridTaggerGS97.pdf.

Greenbaum, Sidney, ed. 1996a. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

———. 1996b. ‘Introducing ICE’. In *Comparing English Worldwide. The International Corpus of English*, edited by Sidney Greenbaum, 3–12. Oxford: Clarendon Press.

Gregory, Michael. 1967. ‘Aspects of Varieties Differentiation’. *Journal of Linguistics* 3 (2): 177–98.

Grieve-Smith, Angus B. 2007. The envelope of variation in multidimensional register and genre analyses. In *Corpus Linguistics Beyond the Word*, edited by Eileen Fitzpatrick, volume 60 of Language and Computers, 21–42. Leiden, The Netherlands: Brill | Rodopi.

Halliday, M.A.K. 1978. *Language as Social Semiotic. The Social Interpretation of Language and Meaning*. London: Arnold.

———. 1988. ‘On the Language of Physical Science’. In *Registers of Written English: Situational Factors and Linguistic Features*, edited by Mohsen Ghadessy, 162–177. London: Frances Pinter.

———. 1991. ‘Towards Probabilistic Interpretations’. In *Functional and Systemic Linguistics. Approaches and Uses*, edited by Eija Ventola, 39–61. Berlin, New York: Mouton de Gruyter.

———. 2009. ‘Methods - Techniques - Problems’. In *Continuum Companion to Systemic Functional Linguistics*, edited by M. A. K. Halliday and Jonathan J. Webster, 59–87. London: Continuum.

Halliday, M.A.K., and Ruqaiya Hasan. 1989. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.

Halliday, M.A.K., and Christian M.I.M. Matthiessen. 2014. *Halliday’s Introduction to Functional Grammar*. 4th ed. Abingdon: Routledge.

Hundt, Marianne, Melanie Röthlisberger, and Elena Seoane. 2018. ‘Predicting Voice Alternation across Academic Englishes’. *Corpus Linguistics and Linguistic Theory* 1 (ahead-of-print). https://doi.org/10.1515/cllt-2017-0050.

Hymes, Dell H. 1972. ‘Models of the Interaction of Language and Social Life’. In *Directions in Sociolinguistics. The Ethnography of Communication*, edited by John J. Gumperz and Dell H. Hymes, 35–71. New York: Holt, Rinehart and Winston.

Karlgren, Jussi, and Cutting, Douglass. 1994. ‘Recognizing text genres with simple metrics using discriminant analysis’. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994), Volume 2*, 1071–1075.

Koch, Peter, and Wulf Oesterreicher. 1985. ‘Sprache Der Nähe – Sprache Der Distanz: Mündlichkeit Und Schriftlichkeit Im Spannungsfeld von Sprachtheorie Und Sprachgeschichte’. In *Romanistisches Jahrbuch*, 36:15–43. de Gruyter.

Kruger, Haidee, and 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). http://search.ebscohost.com/login.aspx?direct=true&db=mlf&AN=202016499220&site=ehost-live.

Malinowski, Bronislaw. 1935. *Coral Gardens and Their Magic*. London: Allen & Unwin.

Matthiessen, Christian M.I.M. 1993. ‘Register in the Round: Diversity in a Unified Theory of Register Analysis’. In *Register Analysis. Theory and Practice*, edited by Mohsen Ghadessy, 221–292. London: Pinter.

———. 2019. ‘Register in Systemic Functional Linguistics’. *Register Studies* 1 (1): 10–41. https://doi.org/10.1075/rs.18010.mat.

Moore, Alison Rotha. 2020. ‘Progress and Tensions in Modelling Register as a Semantic Configuration’. *Language, Context and Text* 2 (1): 22–58. https://doi.org/10.1075/langct.00020.moo.

Neumann, Stella. 2012. ‘Applying Register Analysis to Varieties of English’. In *Anglistentag 2011 Freiburg Proceedings*, edited by Monika Fludernik and Benjamin Kohlmann, 75–94. Trier: Wissenschaftlicher Buchverlag Trier.

———. 2014a. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. Berlin: de Gruyter Mouton.
———. 2014b. 'Cross-Linguistic Register Studies: Theoretical and Methodological Considerations'. *Languages in Contrast* 14 (1): 35–57. https://doi.org/10.1075/lic.14.1.03neu.
———. 2020. 'On the Interaction between Register Variation and Regional Varieties in English'. *Language, Context and Text* 2 (1): 121–44. https://doi.org/10.1075/langct.00023.neu.
Neumann, Stella, and Jennifer Fest. 2016. 'Cohesive Devices across Registers and Varieties: The Role of Medium in English'. In *Variational Text Linguistics*, edited by Christoph Schubert and Christina Sanchez-Stockhammer, 195–220. Berlin, Boston: De Gruyter.
R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.
Rooy, Bertus van, Lize Terblanche, Christoph Haase, and Joseph Schmied. 2010. 'Register Differentiation in East African English: A Multidimensional Study'. *English World-Wide* 31 (3): 311–49. https://doi.org/10.1075/eww.31.3.04van.
Sand, Andrea. 2004. 'Shared Morpho-syntactic Features in Contact Varieties of English: Article Use'. *World Englishes* 23 (2): 281–98. https://doi.org/10.1111/j.0883-2919.2004.00352.x.
Schneider, Edgar W. 2007. *Postcolonial English. Varieties around the World*. Cambridge, UK: Cambridge University Press.
Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 2000. 'Automatic Text Categorization in Terms of Genre and Author'. *Computational Linguistics* 26 (4): 471–95. https://doi.org/10.1162/089120100750105920.
Szmrecsanyi, Benedikt. 2019. 'Register in Variationist Linguistics'. *Register Studies* 1 (1): 76–99. https://doi.org/10.1075/rs.18006.szm.
Tambouratzis, George, Stella Markantonatou, Nikolaos Hairetakis, Marina Vassiliou, George Carayannis, and Dimitrios Tambouratzis. 2004. 'Discriminating the Registers and Styles in the Modern Greek Language-Part 1: Diglossia in Stylistic Analysis'. *Literary and Linguistic Computing* 19 (2): 197–220. https://doi.org/10.1093/llc/19.2.197.
Taverniers, Miriam. 2019. 'Semantics'. In *The Cambridge Handbook of Systemic Functional Linguistics*, edited by David Schönthal, Geoff Thompson, Lise Fontaine, and Wendy L. Bowcher, 55–91. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316337936.005.
Teich, Elke. 2013. 'Choices in Analyzing Choice: Methods and Techniques for Register Analysis'. In *Systemic Functional Linguistics: Exploring Choice*, edited by Lise Fontaine, Tom Bartlett, and Gerard O'Grady, 417–31. Cambridge University Press.
Wong, Deanna, Steve Cassidy, and Pam Peters. 2011. 'Updating the ICE Annotation System: Tagging, Parsing and Validation'. *Corpora* 6 (2): 115–44. https://doi.org/10.3366/cor.2011.0009.
Xiao, Richard. 2009. 'Multidimensional Analysis and the Study of World Englishes'. *World Englishes* 28 (4): 421–50. https://doi.org/10.1111/j.1467-971X.2009.01606.x.

[i] Some recent MDA studies have also used PCA instead of factor analysis, though for reasons of computational efficiency rather than a theoretical motivation (see, e.g., Biber and Egbert 2016).
[ii] Egbert and Biber refer to LDA as "canonical correlation analysis", following the terminology of the SAS software they use for their experiments.
[iii] https://www.ice-corpora.uzh.ch/en.html
[iv] A direct comparison of analyses with and without the extra-corpus material, which is not reported here, showed only small differences, indicating that the multivariate approach is robust against the addition of small amounts of extraneous material.

[v] Note that using a part-of-speech tagger trained on a sample from the British National Corpus (Garside and Smith 1997, 115) on corpora of non-standard varieties of English may lead to reduced tagging accuracy – particularly in light of the assumptions made in this paper. To the best of our knowledge, the ICE initiative has not published any analyses of precision and recall. Admittedly, this may impact the accuracy of our counts. While it is conceivable that incorrect tagging might result in spurious variety-specific linguistic patterns, a cursory inspection of the query output suggests that tagging errors tend to manifest mainly in the form of individual wrongly assigned items. There may thus be a tail of single items mis-assigned to a given category, but these mostly random cases should not affect the multivariate analysis in a substantial way except to add some noise to the distribution of text. We did not notice any systematic error patterns that would be likely to skew the latent dimensions.

[vi] Results for an LDA based on the fine-grained set of 32 text categories according to the original ICE scheme are very similar to the results reported in this paper. The interactive viewers provided as part of the online supplement allow readers to switch between the two analyses.

[vii] Density plots capturing the distribution of texts along each dimension can be viewed in the interactive Weights Viewer in the online supplement.

[viii] Formatting of corpus examples follows the original mark-up. For the sake of legibility, detail such as deletions or overlaps are removed. In spoken text, pause information is replaced by slashes. What is labeled as one intonation unit in the corpus (and given an ID) is indicated by line breaks.

[ix] Note that Biber (Biber 1986, 402), too, points out that a simple spoken/written distinction cannot explain the distribution of texts along his dimension 1. Instead, personal involvement and production constraints are identified as explanations.