ICAME Journal • 49(1) • 2025
DOI: 10.2478/icame-2025-0003

sciendo

ICAME Journal

How stable are multivariate findings about register variation across varieties of English? On the replicability of Geometric Multivariate Analysis

Research Article

Florian Frenken[1]*, Stephanie Evert[2], Gerold Schneider[3], Stella Neumann[1]

[1]RWTH Aachen University
[2]Friedrich-Alexander-Universität Erlangen
[3]Universität Zürich

Received 22 January 2025; accepted 3 March 2025

Abstract: Registers reflect the constraints of systematically recurring situational contexts and are therefore embedded in the lingua-culture in which these situations occur. Consequently, when a language – such as English – is used in widely differing cultural contexts, the question arises whether registers in different varieties of the language might not actually reflect cultural differences between similar types of situations. Previous studies have shown that varieties of English fall into different clusters and that informal spoken texts in particular reflect differences between the varieties. With a focus on register variation across varieties of English, Neumann & Evert (2021) suggest that register-related patterns of variation are much more pronounced than differences between varieties. However, they also observe divergence between texts in the same register from different varieties. The generality of both findings is limited, though, because their analysis was based on only three varieties of English. Our paper aims at exploring these questions more thoroughly by drawing on a larger set of nine components of the International Corpus of English (ICE) preprocessed for comparability (Lehmann & Schneider 2012) and by focusing the interpretation on registers that are expected to be more strongly affected by cultural differences. To this end, we extract the same set of 41 lexico-grammatical features from the ICE components as Neumann & Evert (2021), building on the corpus queries made available in their online supplement. In three steps, we first reproduce the geometric multivariate analysis (GMA) of Neumann & Evert (2021) and then replicate it in two increasingly different approaches. These methodological variations allow us to explore to what extent the results of Neumann & Evert (2021) depended on their specific choice of three ICE components and how stable the results of the exploratory analysis with the chosen multivariate approach are.

Keywords: *varieties of English • register variation • reproduction • replication • multivariate analysis*

1 Introduction

The term *register* describes a type of linguistic variation according to situational context where certain patterns of meaning are more likely to be realized by particular configurations of lexico-grammatical features than others (Halliday & Hasan 1989: 38–39). As such, the correlated situations are readily recognized through association and typically given descriptive labels in a culture. Szmrecsanyi & Kortmann (2009) have shown that varieties of English fall into different clusters based on whether specific vernacular features are attested in the Electronic World Atlas of Varieties of English (Kortmann et al. 2020). Given this close relationship between cultural and situational context, it stands to reason that the former should have an impact on the types of registers distinguishable across varieties of a language. For instance, certain formal registers like parliamentary debates may display varying conventions in cultures where social hierarchies are less pronounced or valued differently. Kruger & van Rooy (2018: 237) posit

* *Corresponding author: Florian Frenken, E-mail: florian.frenken@rwth-aachen.de*

that spontaneous spoken language in particular should exhibit significant differences in this respect. Neumann & Evert (2021) investigate this issue with a multivariate analysis of 41 linguistic features across texts from three components of the International Corpus of English (ICE; Greenbaum 1996), namely Hong Kong, Jamaica, and New Zealand.[1] They use geometric multivariate analysis (GMA), which was first introduced as a methodology by Diwersy et al. (2014) and strongly draws on ideas pioneered in Biber's (1988) multidimensional analysis (MDA). The GMA methodology and how it differs from MDA is described in more detail in Section 2.2 and illustrated by the reproduction study in Section 4. GMA and MDA have in common that both identify latent dimensions of linguistic variation represented by linear combinations of features. While MDA studies often focus on interpreting dimensions based on their feature weights, GMA embraces a geometric perspective based on the visualization of individual texts in the latent multivariate space (and, in particular, the distribution of different text categories). The most crucial difference, however, is that MDA only identifies major dimensions of linguistic variation from a purely unsupervised analysis of correlational patterns in the data, whereas GMA includes weakly supervised techniques in order to reveal more fine-grained patterns or to focus on specific aspects of linguistic variation. Neumann & Evert (2021) use this approach to create a four-dimensional multivariate register space characterized by situational contexts according to the ICE text categories. They interpret latent dimensions based on a combination of the geometric distribution of text categories and the feature weights of each dimension, and conclude, for example, that the first dimension separates conceptual speaking from conceptual writing (see Koch & Oesterreicher 1985). Their findings corroborate the assumption of Kruger & van Rooy (2018), as both Jamaica and Hong Kong English display greater variability towards conceptual speaking in this dimension than texts from New Zealand. While Neumann & Evert (2021) plausibly attribute this to the different status that English assumes in these cultures, understanding the precise processes underlying such observations requires further research.

As a first step, the present study aims to validate the observations made by Neumann & Evert (2021) by testing the reproducibility and replicability of their analysis. In doing so, we aim to answer three specific research questions:

RQ 1: *Can we reproduce Neumann & Evert (2021)*? – We test the reproducibility of the analysis on the same three ICE components. We follow exactly the same analysis procedure but obtain the corpus data via the official ICE portal[2] and apply slightly different data preprocessing that aims to be as consistent as possible across components despite design differences between the local compilation teams. We carry out two reproductions: (i) using the original R code from the online supplement of Neumann & Evert (2021) with minimal adaptations to work with our new data set, and (ii) our own code based on the description in the chapter and the original R code, but making use of the GMA implementation in the R package 'gmatools'. We describe only the latter in this paper, but the original R code leads exactly to the same results and is included in our reproduction materials.

RQ 2: *Do their findings replicate on other ICE components*? – We try to replicate the analysis on an extended data set containing six additional ICE components (Canada, Great Britain, Indonesia, Ireland, Philippines, and Singapore). By mapping the additional texts into the existing register space, we can assess whether they corroborate the interaction between regional and register variation observed by Neumann & Evert (2021), or whether the extended visualization might lead to different conclusions.

RQ 3: *How stable is the register space and its interpretation?* – This RQ is addressed by a more challenging form of replication: the multivariate analysis is repeated on the extended data set, resulting in a different register space. If the weakly supervised GMA procedure turns out to be susceptible to peculiarities of the data set used, the new latent dimensions might lead to different interpretations than before and highlight other patterns of variation. We answer the question both by a visual comparison of the two register spaces and by taking a closer look at the interpretation of the resulting latent dimensions.

1 The ICE is well-suited for this investigation because it is stratified for both cultural and situational variation through its organization into components and text categories collected as comparable, standardized samples.

2 https://www.ice-corpora.uzh.ch/.

Our goal in this paper is thus a methodological one. This means that the linguistic discussion of our results takes a backseat compared to the discussion of the reproduction and replication studies. We will, nevertheless, include a close reading of selected corpus texts in Section 5. The remainder of the paper is organized as follows. Section 2 gives some background on the methodological considerations. The data preparation for the reproduction and replication experiments is summarized in Section 3, whereas the actual experiments are described in Section 4. After the above-mentioned close reading in Section 5, Section 6 offers some concluding remarks.

2 Background

2.1 Reproducibility and replication

Conducting and re-conducting experiments – in order to assess different methods, demonstrate the development of an improved algorithm, or to resolve discrepancies between conflicting conclusions – is essential for science. This involves the possibility to reproduce, replicate, and extend experiments carried out by other researchers such as recommended by course books, such as Levshina (2015), Schneider & Lauber (2019), and Winter (2019). However, many practical obstacles remain that hinder successful reproduction or replication of studies in computational and corpus linguistics (Sönning & Werner 2021), for example that methods are not explained in enough detail and the used data is not available so that computational and corpus linguistics are arguably also part of the reproducibility crisis (Baker 2016).

Reproducibility refers to being able to apply the same method to the same data and obtain the same (or at least very similar) results (Barba 2018: 3–4). It implies that analysis procedures and program code are described with sufficient clarity in the original paper and the underlying data sets are readily available. Our RQ 1 addresses this issue for the study of Neumann & Evert (2021). Reproducibility is a first step towards generalizability, whose ultimate objective is to find out if the results and conclusions of a study are valid beyond a particular data set and analysis, i.e., whether they are corroborated by other data samples and study settings. It is typically defined as obtaining the same or almost same results on the same input data (Barba 2018).

The next step towards generalizability is replicability, which refers to using different data. Unfortunately, there are different exact definitions of whether replicability entails using the same methods for a different dataset (the narrow definition, followed, e.g., by Whitaker 2017 and Flanagan 2025) or if it possibly also includes using different methods (Barba 2018). If the narrow definition is used, then applying different (often related) methods to the same data can be referred to as robustness (Nosek et al. 2022; Whitaker 2017; Flanagan 2025). The approach of applying multiple methods and combining multiple perspectives on the same data has become popular under the name of triangulation, which refers to the strategic use of multiple approaches to tackle one research question: “Each approach has its own unrelated assumptions, strengths, and weaknesses. Results that agree across different methodologies are less likely to be artefacts” (Munafò & Smith 2018: 400). The issue of robustness is beyond the scope of the present paper, as it would require a reanalysis of the data set with a different methodological approach.

Strictly speaking, our RQ 1 also encompasses aspects of replicability because we re-create the data set of Neumann & Evert (2021) from the same corpus data, but with a somewhat different preprocessing pipeline (see Section 3). One might thus argue that we are applying the same method to similar (rather than the same) data. The question of replicability is primarily addressed by our RQ 2 and RQ 3, though, where we consider an extended data set with six additional ICE components. RQ 3 represents the most typical replication scenario, while RQ 2 focuses on the generalizability of the observations rather than generalizability of the analytical method.

Complete reproduction materials for our study are provided as an online supplement in the public Github repository at https://github.com/quantor-project/gma-replication/.

2.2 Multidimensional analysis (MDA) and geometric multivariate analysis (GMA)

Previous research on register and variety-specific variation has shown that register-related patterns generally tend to be more pronounced than variety-specific variation (see, e.g., Kruger & van Rooy 2018). This suggests that situational factors play a superordinate role in influencing choices in language production compared to those associated with different cultural contexts. At the same time, however, certain interactions between variety and register have been attested before, which appear as deviations in selected aspects of variation, like elaboration (Xiao 2009: 442–443) or formality (van Rooy et al. 2010). Moreover, it seems that informal spoken registers, in particular, exhibit noticeable differences between different language varieties (see Neumann & Evert 2021). Moreover, the claim that register variation is a multidimensional phenomenon influenced by a range of variables relating to the typical situational context (Halliday & Hasan 1989; Neumann 2014), calling for multivariate analysis to untangle the dimensions of variation, has been corroborated by all quantitative studies of register discussed here.

Multivariate methods are popular in inductive corpus research, where they help to discover 'hidden' (latent) structure in quantitative linguistic data (Desagulier 2020). Most studies in this area rely on unsupervised statistical techniques, in particular a range of closely related latent variable models including factor analysis (FA; used by Biber 1988) and principal component analysis (PCA; used, e.g., by Diwersy et al. 2014; Biber & Egbert 2016). Biber's (1988) seminal work introducing multidimensional analysis (MDA) has deeply influenced research on this topic. Starting from a large set of lexico-grammatical features, FA is applied to detect latent dimensions of variation from correlational patterns among the features. Interpretation of these dimensions is primarily based on the positive or negative factor loadings of the individual features, after applying a threshold to determine which features are to be considered relevant for each dimension. This is usually complemented by inspecting the locations of the registers along a dimension based on the mean scores of all texts of the same register (see, e.g., Biber & Egbert 2016: 107–112), but the distributions of individual texts in the latent space are rarely visualized, even in more recent applications of the method. In addition to questions of the statistical approach, our approach to normalization is also in striking contrast to MDA (see Section 3), which normalizes most features as frequency counts per 1,000 words of text. This results in spurious correlations between features related to the same linguistic units, e.g., different tenses of verbs or different pronouns. We believe that the relation between surface features and the underlying dimensions of variation is reflected more accurately in the multivariate analysis if normalization is based on a unit of measurement adequate for the respective feature: while it makes sense to normalize frequencies of individual parts-of-speech per words, using the same unit of measurement for features related to other linguistic units such as the sentence (e.g., mood choice) makes the frequencies less precise (see also the discussion in Neumann & Evert 2021: 154).

Diwersy et al. (2014) introduced geometric multivariate analysis (GMA) as an alternative to MDA, with a focus on enabling a geometric interpretation of the feature space. Instead of looking at broad correlational patterns, GMA starts from the feature vectors of individual texts as data points and uses Euclidean distance as a measure of their linguistic (dis)similarity. It relies on orthogonal projections to decompose the distance information of the geometric configuration of data points into a low-dimensional focus space and its orthogonal complement, thus obtaining a low-dimensional perspective suitable for visualization. The percentage R^2 of variance (i.e., squared Euclidean distances) captured by the focus space provides a quantitative measure of how faithful the low-dimensional visualization is to the geometric configuration. Accordingly, GMA uses PCA to identify orthogonal latent dimensions that maximize R^2 and thus capture as much of the distance information as possible (in contrast to FA, whose non-orthogonal dimensions are based on correlational patterns rather than geometric structure).

In practice, FA and PCA tend to identify similar latent dimensions (Diwersy et al. 2014: 183) that capture the major patterns of linguistic variation. Both methods thus share two important limitations: (i) the latent dimensions are dominated by major variational patterns and insensitive to more fine-grained differences, for example, between translated and non-translated texts (Diwersy et al. 2014: 187); (ii) the latent dimensions usually fail to disentangle different factors of linguistic variation, for example, register variation vs. the contrast between German and English texts (Evert & Neumann 2017). Both issues are

relevant for the study of Neumann & Evert (2021): register-related variation is much more pronounced than variety-specific differences, which do not become clearly visible in the unsupervised PCA; at the same time, it is mixed with other factors of linguistic variation in the latent dimensions (such as topic effects), which are not the object of study (cf. Figure 3 in Section 4.1).

To address this problem, the GMA approach introduces a weakly supervised intervention in order to guide the latent dimensions towards the desired factor of variation. Neumann & Evert (2021) use the ICE text categories (as a proxy for different registers) for this purpose, resulting in a latent space that separates the categories as clearly as possible while minimizing variability within each category. This is done with the help of linear discriminant analysis (LDA), which is based on the same distributional assumptions as PCA and hence compatible with the geometric perspective. Neumann & Evert (2021) consider the result to be a latent register space that is made interpretable by the visual map of ICE text categories, which enable them to relate different regions of the latent space to familiar registers (cf. Figure 2 in Section 4.1). This is supported by visualizations of the register space in the form of scatterplot matrices (see Section 4.1) and interactive 3d plots, with relevant metadata (in particular ICE text category and language variety) indicated through colors, plot symbols, or animation. An R software package 'gmatools' supporting the entire GMA analysis procedure has recently become available (Evert 2025). We use this software package for our reproduction and replication study.

Application of the supervised LDA method is perhaps the most crucial difference between the GMA and MDA approaches, combined with the strong focus on visualization and geometric interpretation in GMA. LDA is often claimed to achieve better performance as a preprocessing step in text classification tasks than PCA and related unsupervised techniques (e.g., Torkkola 2001) because it maximizes class separability rather than reflecting the maximal differences in the data. Due to its supervised nature, it can be given supplementary information about which features in the data are relevant for the respective classification. For example, completely unsupervised stylistic research is often misled by differences in topic, content, or editorial conventions, thus missing far more subtle stylistic features such as passive structures (Hundt et al. 2016). While PCA and LDA are often seen as alternatives, some authors have pointed out that performing a PCA first and then applying LDA only to the resulting space with already reduced dimensions can be beneficial (Krzanowski et al. 1995). Such systems have been used for image recognition, particularly for face recognition (Belhumeur et al. 1997), but rarely for text analysis yet. In many areas of content analysis, combining supervised and unsupervised methods is useful, for example, in structural topic models (Roberts et al. 2019). We believe that the combination of LDA and PCA has considerable potential. In GMA, it is usually applied in a different way than in text classification: a small number of LDA dimensions are complemented by further orthogonal PCA dimensions to capture factors of language variation that have been backgrounded by the LDA focus space. For very high-dimensional feature spaces, prior application of PCA may be useful to avoid overfitting of the LDA (similar to partial least-squares regression, cf. Abdi 2010).

3 Data preparation

The corpus used in this study consists of the following nine components from the ICE: New Zealand (NZ), Jamaica (JAM), Hong Kong (HK), India (IND), Philippines (PHI), Singapore (SIN), Canada (CAN), Ireland (IRE) and Great Britain (GB). These components cover a good range of Inner and Outer Circle varieties and vary in the developmental phase according to Schneider's (2007) dynamic model of postcolonial Englishes. All ICE components are compiled on the basis of a common design (Greenbaum 1996), which stratifies texts into a range of text categories at various levels of granularity. Like Neumann & Evert (2021), we use the 20 mid-level categories, which should be specific enough to capture multiple dimensions of variation across texts yet not so specific that we run the risk of missing out on abstractions language users might make across groups of texts.

The nine components are used in the cleaned-up version by Lehmann & Schneider (2012), which ensures consistent and comparable markup. Neumann & Evert (2021: 153) report their own, different

procedure for cleaning up the three components they analyze; hence, the markup of these components will diverge slightly from the version we use for this study. As we are interested in the characterization of texts in their respective specific situational context (and to ensure comparability with Neumann & Evert 2021), individual texts ('sub-texts' in ICE initiative parlance) are extracted from the corpus files. This ensures that the specific situational features influencing each language user's linguistic choices are adequately covered by our analysis. The resulting number of texts is given in Table 1. In total, our data set contains 7,930 texts.[3]

Table 1. Overview of the number of texts per component across the 20 mid-level text categories of the International Corpus of English.

	NZ	JAM	HK	IND	PHI	SIN	CAN	IRE	GB
conversations/ phonecalls	100	114	116	111	101	116	130	103	137
classroom lessons	20	22	22	20	20	17	28	20	22
broadcast interactions	30	33	36	32	31	36	69	30	34
parliamentary debates	14	10	9	12	10	10	15	10	11
legal cross-examinations	18	15	16	10	9	11	13	10	12
business transactions	10	12	10	10	15	11	12	11	13
unscripted monologues	58	72	71	51	51	61	60	53	98
demonstrations	11	11	12	13	29	13	11	11	17
legal presentations	11	17	13	10	20	10	14	11	12
scripted monologues	106	60	105	64	82	102	145	59	79
student writing	44	66	24	73	47	40	27	24	41
social letters	43	93	107	25	57	67	57	50	67
business letters	79	111	133	16	112	104	73	104	96
academic writing	50	69	110	48	77	58	83	83	42
popular-scientific writing	46	43	53	42	42	45	40	40	40
news reports	95	67	117	57	84	59	96	95	79
administrative writing	14	20	21	11	12	47	9	15	11
skills and hobbies	13	24	27	23	26	23	24	33	10
press editorials	31	22	63	23	40	35	37	35	36
creative writing	21	23	45	21	20	22	21	20	20
total	**814**	**904**	**1110**	**672**	**885**	**887**	**964**	**817**	**877**

As in previous work, textual segments marked as extra-corpus material by the corpus compilers (such as turns by the interviewers in elicited tasks) were removed, even though their inclusion seems to have no significant impact on the distribution of linguistic features across the whole text (Neumann & Evert 2021: 152). The components are used in the officially distributed version annotated for part of speech using the CLAWS tagger (Garside & Smith 1997), whose finer granularity compared to other taggers facilitates the extraction of more complex grammatical features that are relevant for register differences. The feature extraction is designed as a reproducible corpus query pipeline for a total of 41 features that cover the three main dimensions of contextual variation according to systemic functional register theory, namely field, tenor, and mode of discourse (see Halliday & Matthiessen 2014: 32–36). Features include,

3 The overall number of texts diverges slightly from Neumann & Evert (2021): NZ: 809; JAM: 922; HK: 1,113, This difference is due to differences in cleaning up the markup resulting in different text lengths, and then cutting off texts with less than 100 words or 10 sentences.

for example, counts for different mood choices such as imperatives, which (together with accompanying features like present tense and reduced pronoun usage) are expected to be more frequent in instructive or regulatory writing, implying tenor-related differences in terms of personal authority. The extraction pipeline has been implemented in the form of a script using the CQP query processor of Corpus Workbench (Evert & Hardie 2011), followed by post-processing in Perl and R. The CQP queries are adopted from Neumann & Evert (2021), with a correction made to the query for passives and a slight improvement to attributive adjectives, which are now counted separately in coordinated pre-modifiers. In addition, the query for adverbs of time was expanded by a short word list. A comprehensive evaluation of the feature extraction pipeline, including more substantial improvements, is currently underway. In the post-processing steps, absolute frequency counts obtained by the query script are normalized to relative frequencies with respect to sensible units of measurement (e.g., passives as a proportion of all finite verbs), so as to represent them statistically as individual choices within an envelope of variation. For this step and further pre-processing of the data set of feature vectors, we use the R scripts from the online supplement of Neumann & Evert (2021) with minimal modifications. Note that – like Neumann & Evert (2021) – we use the count of the respective unit of measurement per text as the denominator, thus yielding straightforward proportions.

We also follow Neumann & Evert (2021) in discarding all texts shorter than 100 words or consisting of fewer than 10 sentences. Feature counts are sparse and unreliable in such short texts, which has two effects: (i) the resulting feature distributions are often skewed and do not meet the normality assumptions of PCA and LDA; and (ii) both PCA and LDA are known to be susceptible to outlier data points from short texts. This filtering step affects certain parts of the ICE more than others, ranging from 5% of texts in the Indian component to almost 14% for Canadian texts. This is presumably a side effect of differences in the corpus compilation related to padding files with individual texts such that the desired length of 2,000 tokens is reached. In a similar manner, certain text categories like letters also contain disproportionately more texts below the threshold due to the nature of such texts. We consider this acceptable for the purposes of the present study (and for the purpose of replicating the previous analysis). However, in future work, we plan to develop an improved version of GMA that can be applied to texts of any length. Again following Neumann & Evert (2021), features are standardized to z-scores, which ensures that all features make the same contribution to (squared) Euclidean distances, followed by a signed logarithmic transformation in order to deskew the distributions. One difference from the original study is that our z-scores are based on the means and standard deviations across all nine ICE components rather than only across Hong Kong, Jamaica, and New Zealand. The resulting values are very similar to the means and standard deviations for the original three components, so this should have little effect on the multivariate analysis. The distributions of our log-transformed z-scores shown by the box plots in Figure 1 are strikingly similar to those obtained by Neumann & Evert (2021).[4] We refer to our full data set as ICE9, and to the subset comprising only Hong Kong, Jamaica, and New Zealand as ICE3 (but using the same standardization as ICE9).

4 Results

4.1 Reproduction

In this section, we attempt to reproduce Neumann & Evert (2021) using the recreated ICE3 dataset. We follow their analysis procedure and visualizations as closely as possible, starting with a supervised LDA based on the same 20 mid-level text categories. Neumann & Evert (2021: 155) use the first four LDA dimensions as their focus space, arguing that a supervised classifier[5] can assign texts to the correct category with 60% accuracy within this focus space (compared to 72.6% with all 19 LDA dimensions) so that categories are indeed well separated and reducing the register space to four dimensions loses

4 See https://www.evert.de/PUB/NeumannEvert2021/data/prepare_data.html.
5 SVM classifier with RBF kernel and 5-fold cross-validation.

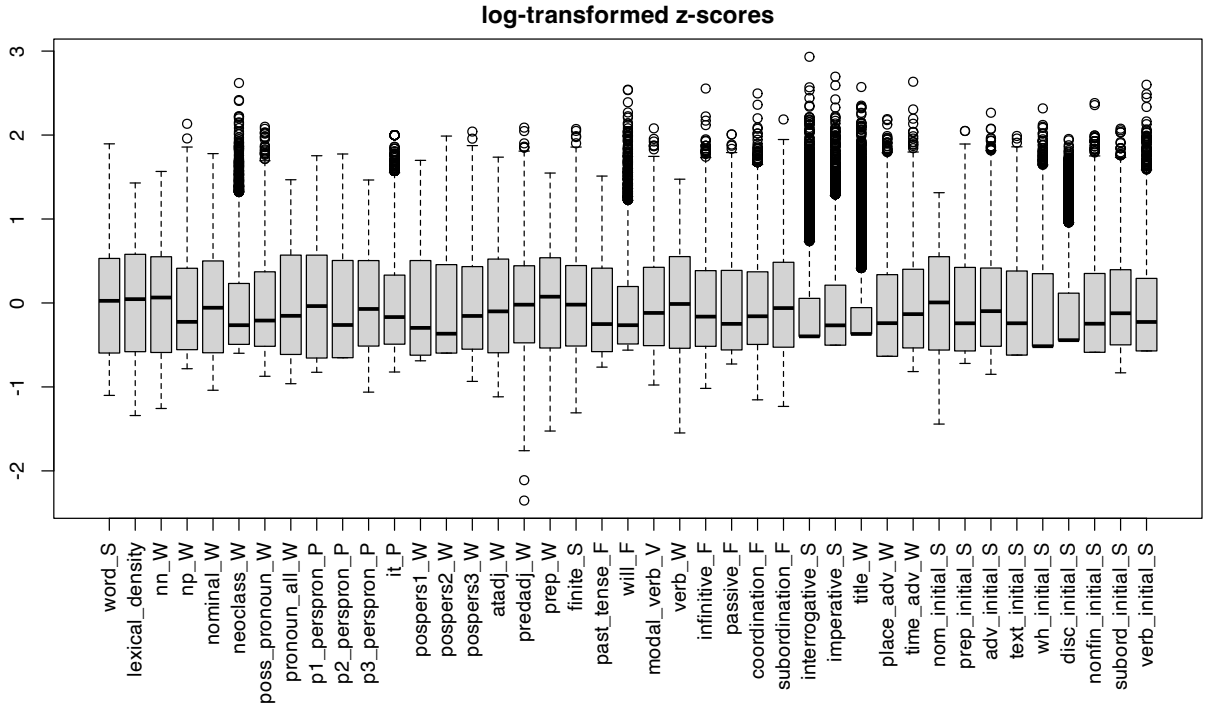

Figure 1. Distribution of all features in ICE9 as log-transformed z-scores.

comparatively little structural information. For the purpose of reproduction, we also use the first four LDA dimensions, even though they account only for R^2 = 17.87% of the total variance. This observation underlines how LDA specifically captures linguistic variation related to differences between the register-related ICE text categories rather than merely accounting for the largest proportion of overall variation. We also follow Neumann & Evert (2021) in applying a PCA-based rotation[6] to the first two dimensions in order to align the geometric structure better with the dimension axes. We then visualize the orthogonal projection of the ICE3 data set into the resulting focus space in the form of a reduced scatterplot matrix matching the original study (Figure 2). The panels of this plot are scatterplots of the first LDA dimension against the other three dimensions, respectively. Text categories are indicated by color, using a rainbow scale that visually groups similar text categories. For better readability, we show separate plots for written texts (first row) and spoken texts (second row). Keep in mind that, due to its geometric approach, GMA works with orthonormal basis vectors of the focus space rather than the original LDA discriminants (which are usually oblique rather than orthogonal). For simplicity, we nonetheless refer to these basis vectors as 'LDA dimensions'.

The visualization immediately reveals a clearly structured register space that places different text categories in different regions and often arranges closely related categories next to each other. It is particularly remarkable how the LDA – rather than maximally separating individual categories in an arbitrary manner – has lined up written and spoken categories in two overlapping banana-like shapes (separated in two rows by mode in Figure 2), recreating the gradients of the rainbow color scale. This shows that our focus space indeed captures more general regularities underlying the ICE text categories. GMA considers the analysis weakly supervised because no information about connections and similarities between the 20 text categories is provided to the LDA (e.g., via the hierarchical structure of the category system, cf. Neumann & Evert 2021: 151). A visual comparison with Neumann & Evert

6 This operation is similar to the "varimax" rotation used in factor analysis (see, e.g., Merenda 1997). Unlike FA rotations, GMA only allows true geometric rotations that preserve distances and angles within the focus space.

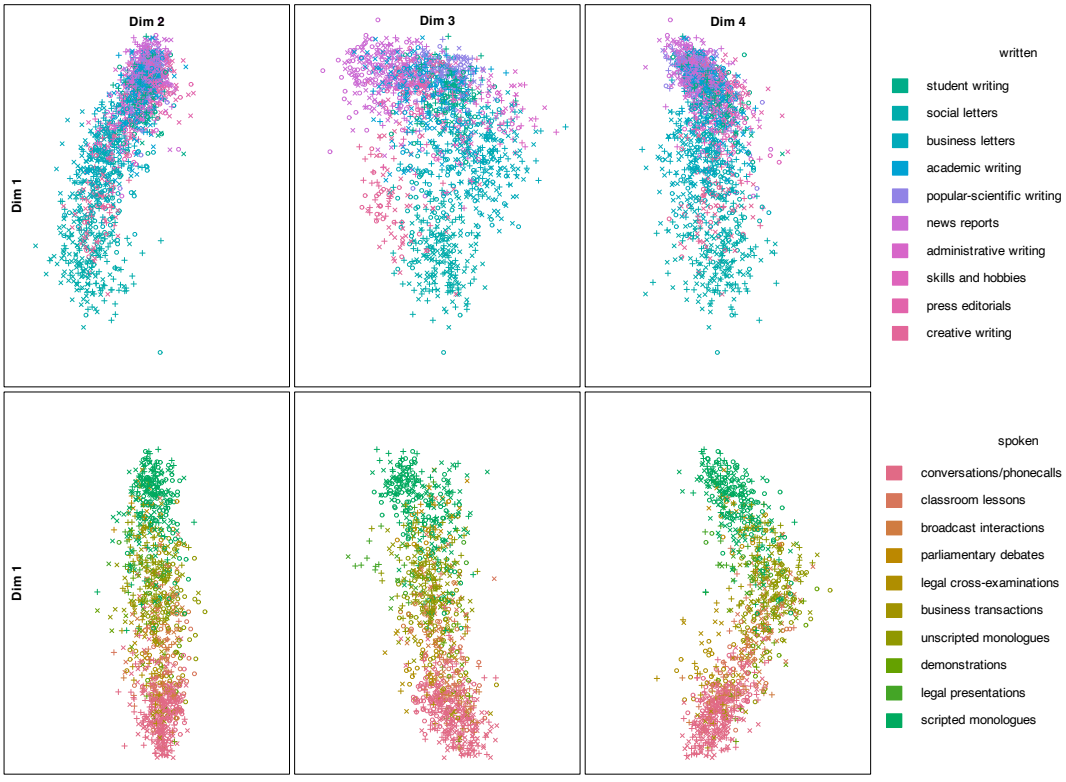

Figure 2. Partial scatterplot matrix of the four LDA focus dimensions for ICE3 separated by mode (first row: written texts; second row: spoken texts).

(2021: Figure 1) shows virtually indistinguishable shapes (even down to many of the outlier data points), indicating that we have successfully reproduced their original analysis despite small differences in our data preparation.

We have already pointed out above that orthogonal projection into the LDA focus space captures only R^2 = 17.87% of the (squared) distance information from the original data set. This suggests that only a comparatively small portion of the overall linguistic variation is necessary to model register-related groupings, which the LDA foregrounds against the much more prevalent variation between individual texts. This is in stark contrast to the first four dimensions of an unsupervised PCA, which account for R^2 = 51.25% of the variance, nearly three times as much as the LDA focus space. Figure 3 shows an analogous visualization of these PCA dimensions, which clearly capture some aspects of register variation but show less structure and more variance, arguably picking up on specific factors that potentially only apply to individual texts. PCA dimension 1 is correlated with our first focus dimension but also captures substantial amounts of variation within each text category. Dimension 2 also helps to separate certain text categories and might be related to the third focus dimension, but again there is a large amount of within-category variability. Dimensions 3 and 4 appear to focus on other factors of variation that are not directly related to text categories. Overall, PCA fails to provide a clear-cut visual map of the ICE text categories and hence produces a latent space that is much harder to interpret.

The next question to ask is whether our reproduction would have led us to the same interpretation of the LDA dimensions as Neumann & Evert (2021), who label them as *conceptual speaking vs. conceptual writing* (LDA dim 1), *dialogic written vs. neutral*[7] (LDA dim 2), *descriptive-narrative vs.*

7 'Neutral' is Neumann & Evert's (2021: 164) notion for one-ended dimensions along which individual data points only reach values sufficiently removed from zero to identify notable linguistic traits towards one pole, meaning that all other data points do not display specific linguistic features on this dimension.

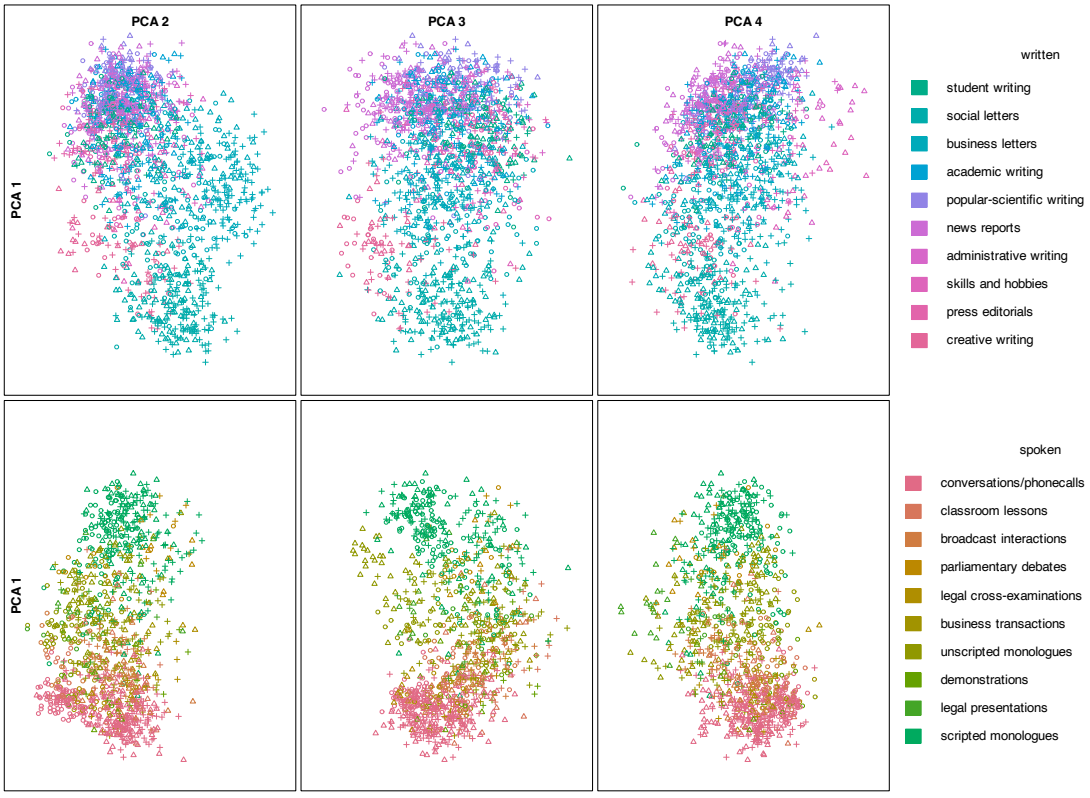

Figure 3. Partial scatterplot matrix of the first four PCA dimensions for ICE3 separated by mode (first row: written texts; second row: spoken texts).

instructive-regulative (LDA dim 3), and *neutral vs. online production* (LDA dim 4). Their interpretation is based on the visual reference system created by the positions of different text categories within the focus space, combined with the feature weights of the LDA dimensions, which are also Biber's (1988) main entry point for interpretation. We have already seen that the visual map of text categories is almost identical to the original study, guiding the interpretation of the dimensions in the same direction. Figure 4 visualizes the feature weights for all four LDA dimensions, with green bars above the neutral line indicating positive weights and red bars below the neutral line negative weights. Features that do not contribute substantially to any of the four dimensions (with a weight ≥ 0.1) have been omitted for clarity. Comparison with Neumann & Evert (2021: Figure 3, 5, 7, and 9) shows very similar patterns of the feature weights, though some minor differences for individual features might change details of the interpretation. Neumann & Evert (2021) complement their interpretation by looking at the contribution of different features to the discriminant scores of text categories, which Evert & Neumann (2017) insist on to avoid misinterpretation of feature weights. The numerous comparisons of different categories for each LDA dimensions are only made possible by an interactive Web app, so we do not include this step in our reproduction experiment. Considering the striking similarities between Neumann & Evert (2021) and our reproduction, both in terms of textual groupings as well as feature weights, our analysis clearly aligns with their original interpretation, which we therefore do not repeat here.

Concerning regional variation, Figure 5 shows separate scatterplots for texts from the New Zealand, Jamaica, and Hong Kong components of the ICE, focusing on the first two LDA dimensions. It corresponds to Neumann & Evert (2021: Figure 10), but separates written and spoken texts for additional clarity. No strong patterns of variety-specific register variation emerge, but some smaller differences correspond well to the observations made in the original study. In particular, both Jamaica and Hong Kong show more variability along dimension 1 (*conceptual speaking vs. conceptual writing*)

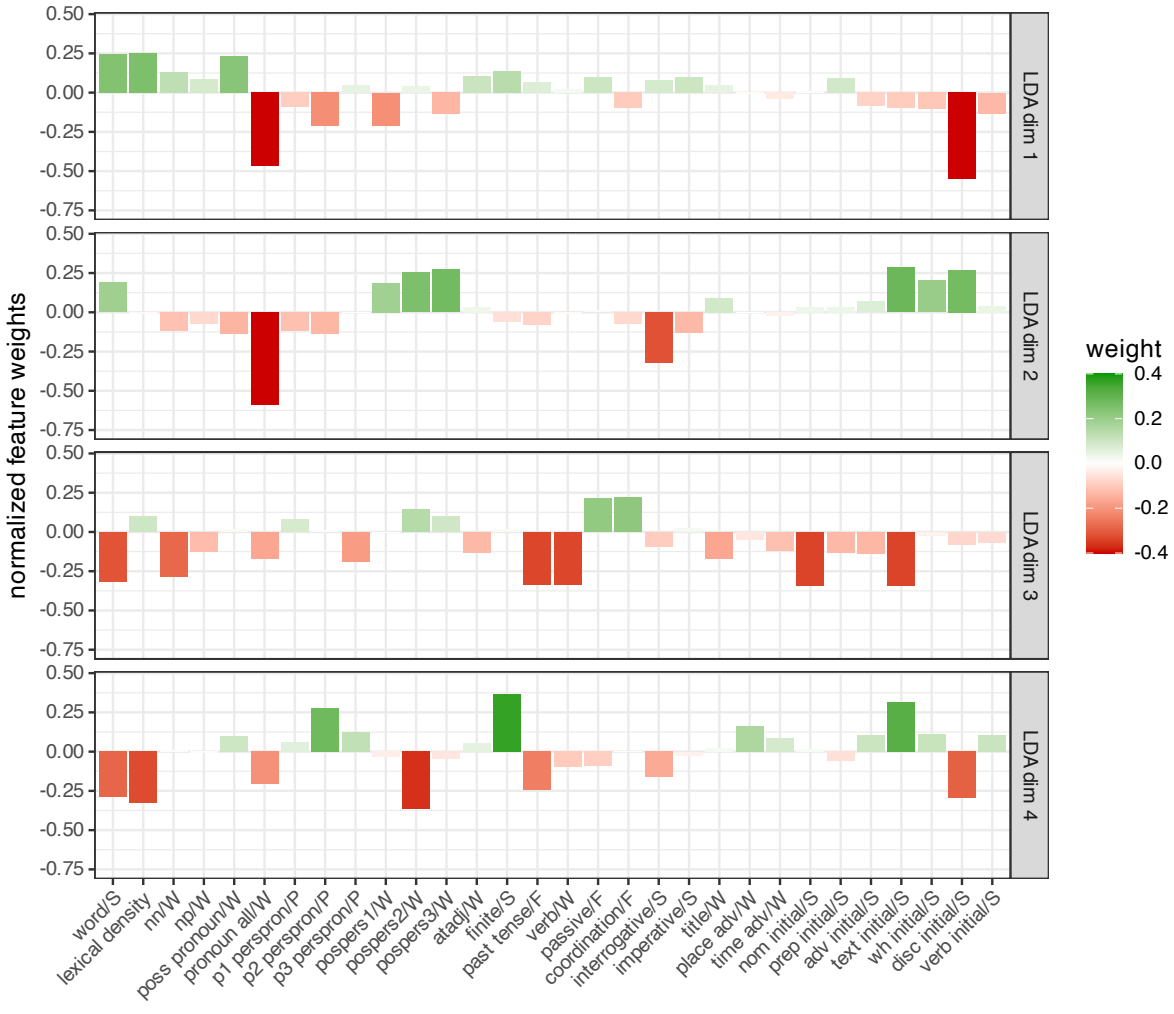

Figure 4. Feature weights of the four LDA focus dimensions for ICE3.

than New Zealand, reproducing a key finding of the first study that supported the conjecture of Kruger & van Rooy (2018). In dimension 2 (*dialogic written vs. neutral*), all texts from the Jamaica component are shifted slightly to the *dialogic written* side compared to the same text categories in the New Zealand component; this suggests a general difference between the varieties that is not related to individual registers.

Summing up, we can give a fully positive answer to RQ 1. We have been able to reproduce the analysis and findings of Neumann & Evert (2021) very well, even with the use of slightly different preprocessing steps during data preparation.

4.2 Replication 1: Adding data points

In this section, we attempt to replicate the results of Neumann & Evert (2021) on our extended ICE9 data set. In order to address RQ 2, we keep the four-dimensional focus space from Section 4.1 (which we have confirmed to match the focus space of the original study) and project all texts from the additional six ICE components into this subspace. In other words, the visual map and interpretation of the dimensions remain the same, but we test whether other varieties of English confirm the prior observations or show entirely different patterns of register-specific variation.

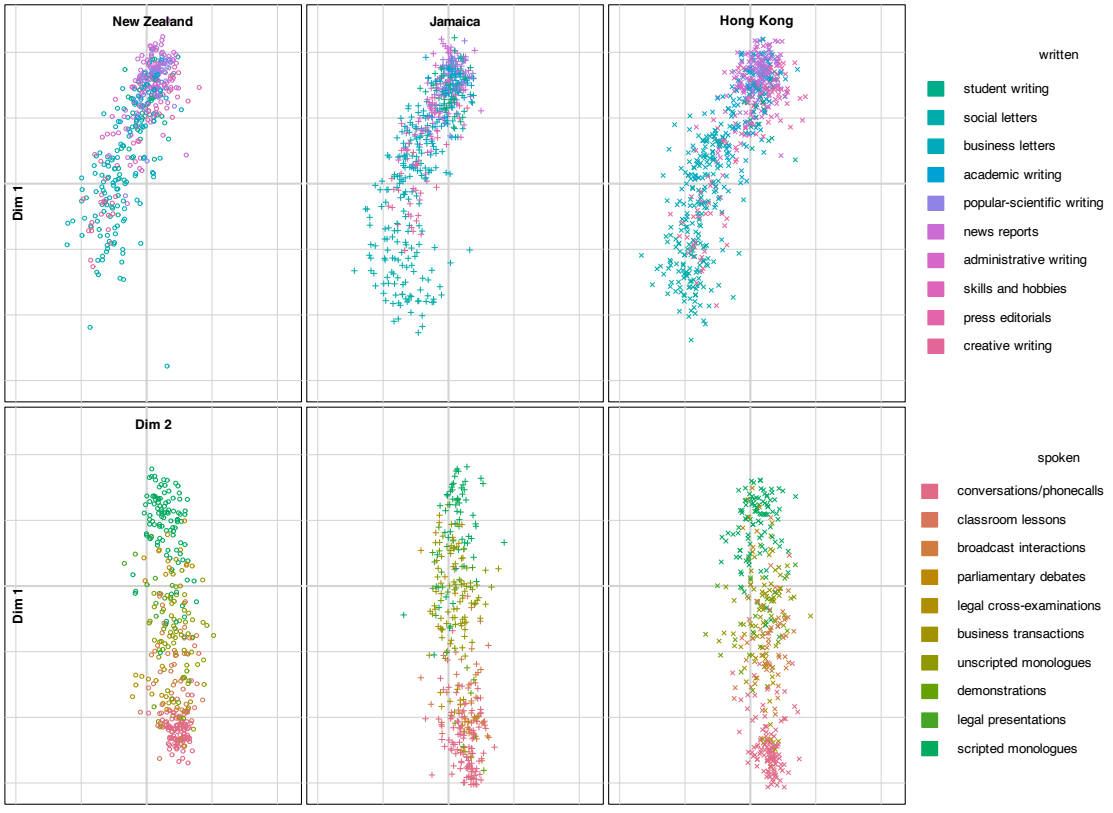

Figure 5. Scatterplots of the first two LDA focus dimensions for ICE3 separated by variety (columns) and mode (rows; first row: written texts; second row: spoken texts).

Figure 6 shows scatterplots corresponding to Figure 5 for ICE components Canada, Ireland, and Great Britain; Figure 7 shows scatterplots for India, the Philippines, and Singapore. Overall, the distributions of data points look very similar to the original study, and no surprising new patterns emerge. For all six additional varieties of English, the variability of spoken texts along dimension 1 (*conceptual speaking vs. conceptual writing*) is similar to Jamaica and Hong Kong (if anything, Great Britain and Ireland are spread particularly far), showing that New Zealand is in fact the outlier with reduced variability. In their brief discussion of the comparison between the three varieties, Neumann & Evert (2021: 171) implied that the New Zealand variety was the reference which the two Outer Circle varieties moved away from. This implication thus needs to be reviewed against the background of more Inner *and* Outer Circle varieties.

Written texts, on the other hand, extend further towards conceptual speaking in Jamaica and Hong Kong than for most other varieties. In the second dimension (*dialogic written vs. neutral*), the overall slight shift to the left observed for Jamaica and Hong Kong also holds for the three Asian varieties (India, Philippines, Singapore).

The answer to our RQ 2 is thus partly positive. Overall, the lack of any substantial register-related differences between the language varieties is confirmed and the original analysis based on only three ICE components did not give a misleading impression. However, the additional six components open a new perspective on the small differences observed in dimensions 1 and 2. The tentative conclusions of Neumann & Evert (2021) cannot be fully confirmed and may need to be modified.

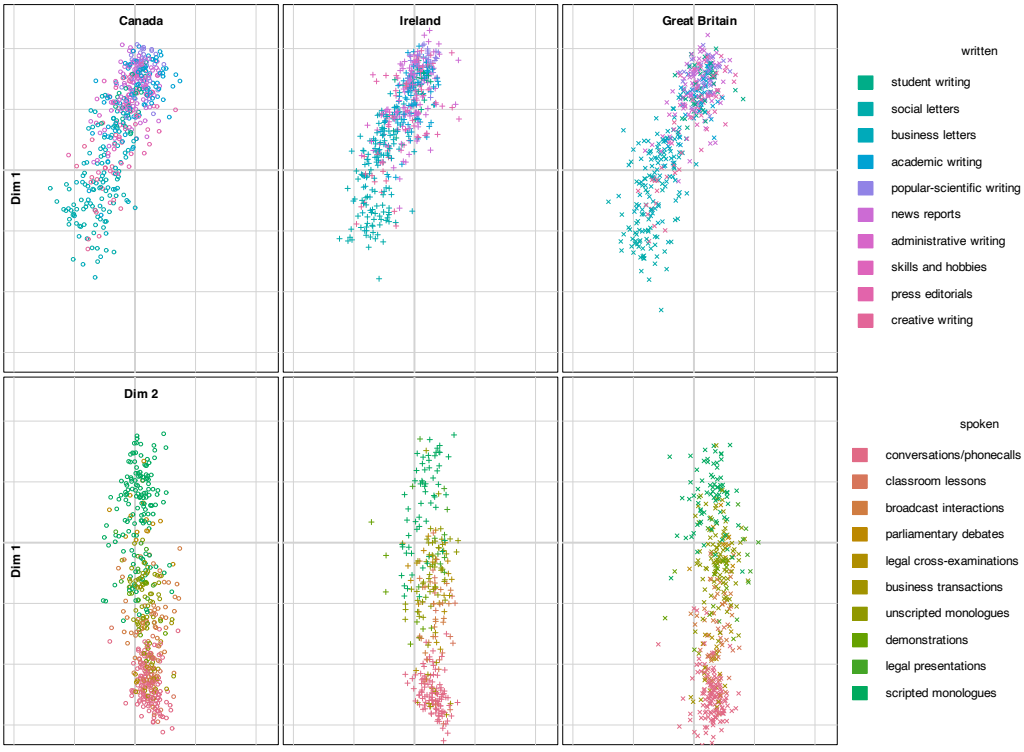

Figure 6. Scatterplots of the first two LDA focus dimensions for CAN, IRE and GB separated by variety (columns) and mode (rows; first row: written texts; second row: spoken texts).

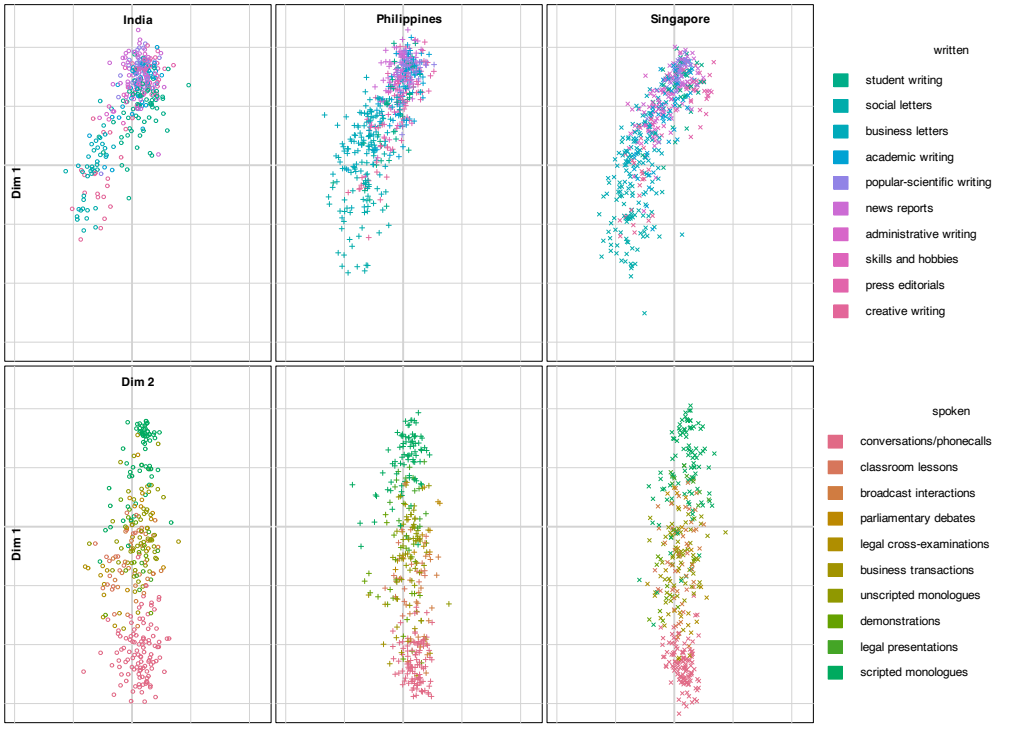

Figure 7. Scatterplots of the first two LDA focus dimensions for IND, PHI and SIN separated by variety (columns) and mode (rows; first row: written texts; second row: spoken texts).

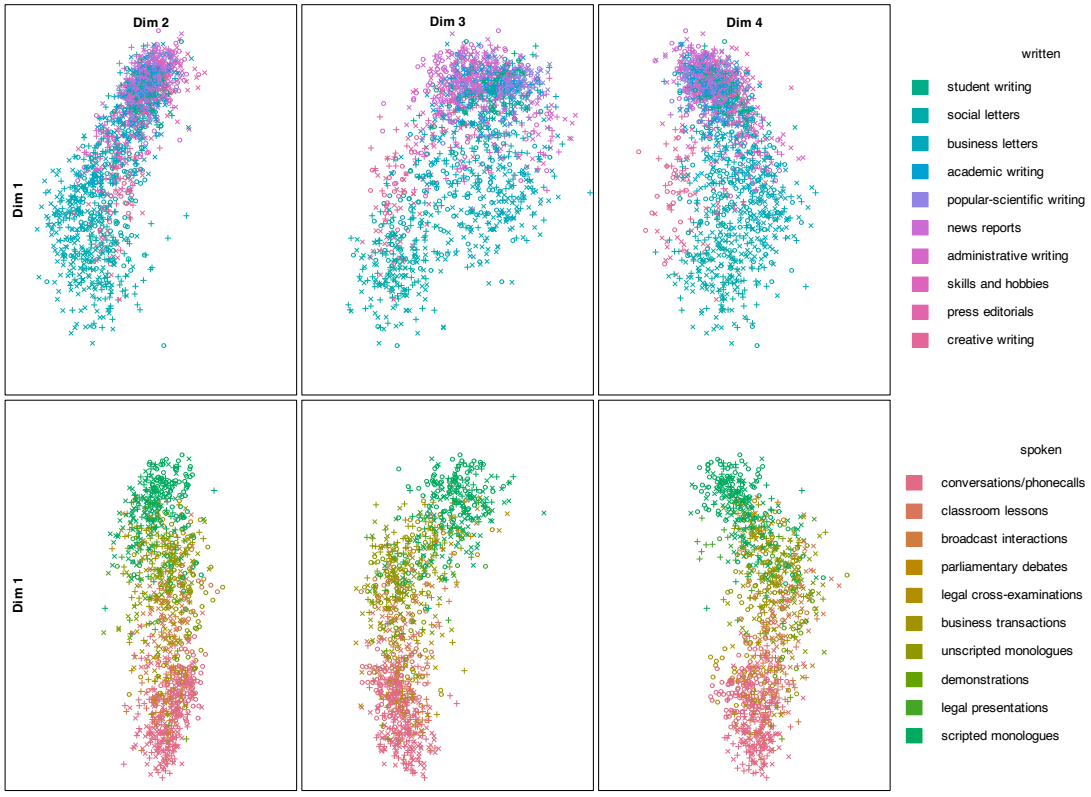

Figure 8. Partial scatterplot matrix of the first four LDA focus dimensions for the ICE9 data set, showing only ICE3 texts separated by mode (rows; first row: written texts; second row: spoken texts).

4.3 Replication 2: A new focus space

We finally turn to RQ 3 and a more substantial replication, now by creating a new focus space from an LDA across all nine ICE components. Our procedure follows the original study as closely as possible. In particular, we also use four LDA dimensions and perform a PCA rotation in the first two. This replication tests to what extent the register space of Neumann & Evert (2021) is a product of their choice of varieties; ideally, LDA on another set of ICE components should still result in a very similar space. We assess similarity of the focus spaces visually by comparing scatterplots of the ICE3 and ICE9 focus space. Figure 8 shows only ICE3 varieties in order to facilitate direct comparison with Figure 2.

The first two dimensions appear quite similar to those of the ICE3 focus space, with slightly increased variability along dimension 2 (*dialogic written vs. neutral*), which is easily explained by additional variation brought in by the six additional components. However, the visual impression of dimension 3 (*descriptive-narrative vs. instructive-regulative*) is markedly different (and to a lesser extent also of dimension 4). The bend to the right in LDA dimension 4 (*neutral vs. online production*) now also affects the written texts to a certain extent, while a complementary curvature can be seen in LDA dimension 3 (towards the top right of the two center panels). The feature weights of the new focus space, shown in Figure 9, also suggest at least a broadly similar interpretation for LDA dimensions 1 and 2, but point into an entirely different direction for dimensions 3 and 4.

It would thus seem at first glance that the first two dimensions of the register space are fairly stable, but further dimensions strongly depend on the language varieties included. The ‘gmatools’ package offers a quantitative measure of the similarity of two focus spaces, which can be interpreted as the (fractional) number of shared dimensions between the two subspaces.

Here, we obtain a similarity score of 3.714 out of 4 dimensions, which suggests that the LDA on ICE9 is less different from the original LDA than the visual impression has led us to believe. Subspace

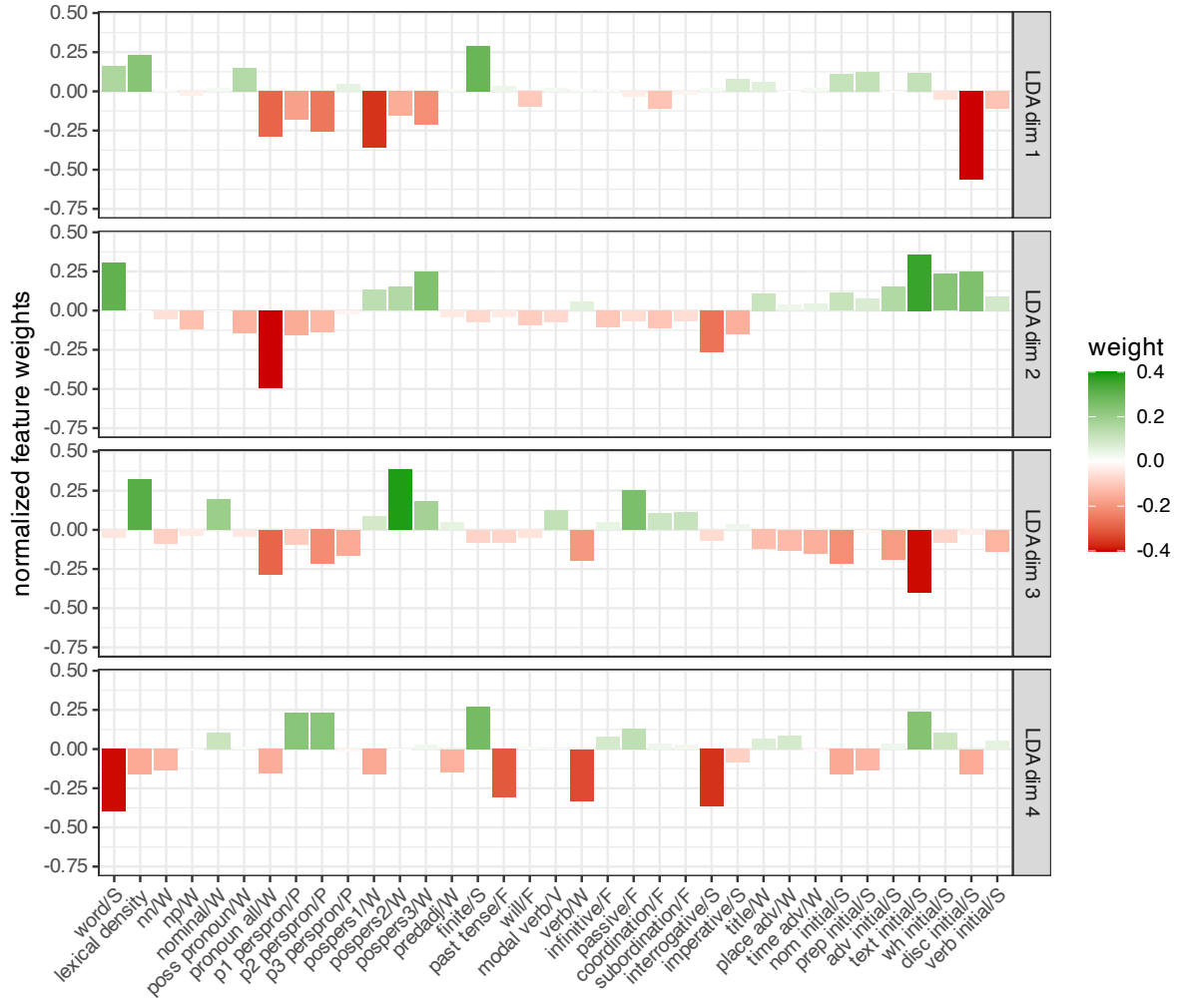

Figure 9. Feature weights of the LDA focus dimensions for ICE9.

Table 2. Subspace similarity scores for the ICE9 vs. ICE3 focus spaces.

Aligned basis vector	Similarity	Angle (degrees)
#1	0.9918	7.77
#2	0.9597	16.31
#3	0.8896	27.18
#4	0.8742	29.04

similarity is computed by aligning the basis vectors of the two subspaces and then measuring the angles between corresponding basis vectors, as shown in Table 2 (with the total score obtained as the sum of cosine similarities).

After the alignment, two dimensions of the focus space are fairly close to the original analysis, while the other two are oblique at an angle of less than 30 degrees. This should still ensure that we see comparable patterns in all four dimensions. A possible explanation is that the two focus spaces are indeed reasonably similar but that the LDA dimensions are rotated in the ICE9 space. The 'gmatools' package also offers functionality to perform a rotation of the ICE9 focus space that matches the ICE3 dimensions as closely as possible.

Figure 10 shows a scatterplot of the ICE3 components in the ICE9 focus space after the rotation, which now bears striking similarity to Figure 2. Feature weights for dimensions 3 and 4, shown in Figure 11, also have become much more similar to the corresponding ICE3 feature weights (though there are

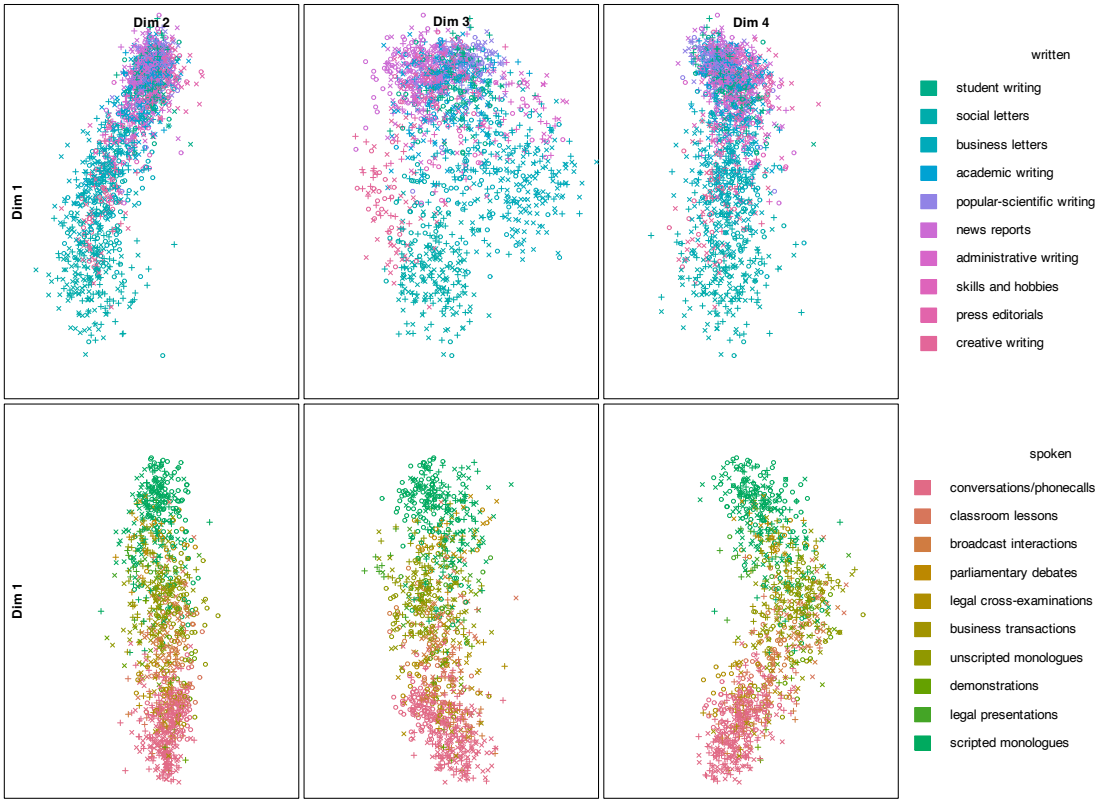

Figure 10. Partial scatterplot matrix of the first four LDA focus dimensions for ICE9 after alignment with the ICE3 focus space.

still noticeable differences in all four dimensions), leading to an interpretation that is at least broadly in line with Neumann & Evert's (2021) dimensions.

Finally, a look at differences between the nine language varieties in the new matching perspective (omitted for lack of space, but available in the online supplement) shows patterns that are very similar to those in Figure 5, Figure 6, and Figure 7. The main findings of the study are thus seen to replicate across different versions of the LDA register space.

To sum up, the answer to RQ 3 is more complicated than for the other research questions. On the one hand, our second replication study has to be considered successful because (i) the new focus space has a very similar visual map of ICE text categories because (ii) the interpretation of its dimensions will be reasonably consistent with the original study, and because (iii) it leads to the same findings concerning register-related differences between the language varieties. On the other hand, our first visual impression was markedly different and may well have guided our discussion and interpretation in another direction. A successful replication was only possible after the automatic alignment of dimensions. This shows that even with advanced visualizations, understanding multidimensional spaces can be challenging and suitable rotations of the dimensions might be needed in order for the quantitatively measurable similarity of different focus spaces to become visible.

5 Discussion

Our aim in this paper has been a methodological one, namely, to examine reproducibility and replication of inductive multivariate analysis of linguistic variation in a corpus. The previous section answered the

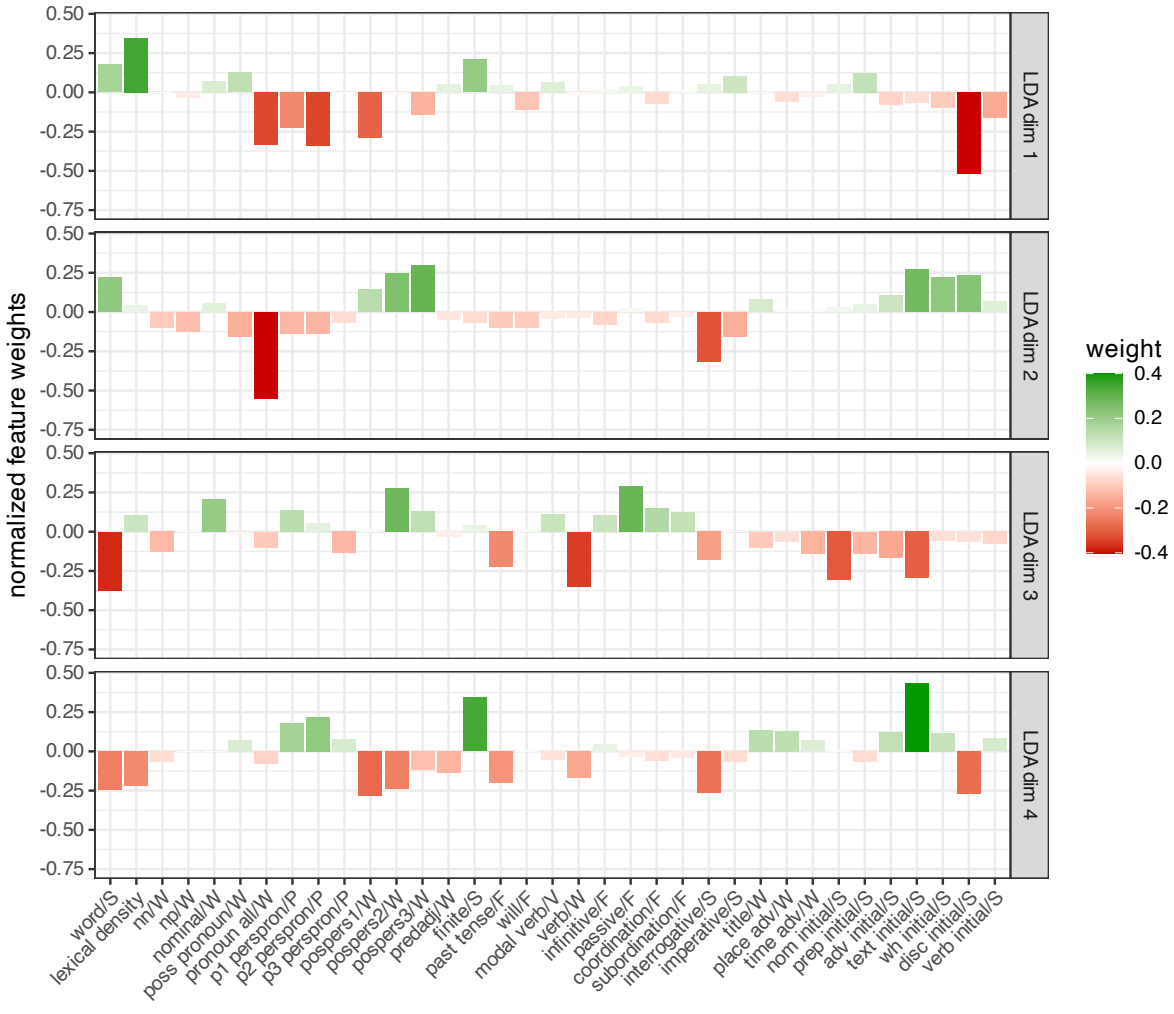

Figure 11. Feature weights of the LDA focus dimension for ICE9 after alignment with the ICE3 focus space.

respective research questions. What does this mean for the usability of such a quantitative approach – and more specifically for geometric multivariate analysis? Given the overall similarities between the previous GMA on three ICE components by Neumann & Evert (2021) and the three increasingly different analyses demonstrated here, we can conclude that the results are fairly stable – including the result that register differences are much stronger than differences between varieties. Using slightly different preprocessing steps with the same data does not significantly affect the results of a GMA analysis both in terms of visualization and interpretation. Although this is unsurprising given that Neumann & Evert (2021) reported that even including or removing extra-corpus text present in the ICE corpus files did not significantly change their results, our further experiments conducted in this paper give us additional confidence in the strength of the method. The analysis does not appear to be very sensitive to the influence of additions to the data set – at least if its overall structure is as consistent as is the case for the different ICE components, which are compiled on the basis of a common corpus design. The stability of the focus space is demonstrated in particular by our second replication experiment (after applying suitable rotations to align it with the ICE3 focus space). While it is conceivable that varieties differ in their register structure, this is not what we see. We believe that, despite some observable differences in patterns across varieties, the overall effect of register, which is also responsible for the main visual structure in our data, is so strong that it creates the same statistical regularities for clusters of texts

across all varieties (see Section 1) – even in the face of slight alterations of the texts (such as noise in the form of extra-corpus material) or addition of further varieties.

Complex models like the one discussed here are often difficult to interpret. In order to verify our approach and findings, close reading of individual examples is an important and indispensable step. It can also help to shed light on the linguistic patterns underlying the statistical patterns and correlations identified by the GMA procedure. For our brief linguistic interpretation, we focus on three spoken and two written text categories that are clearly located in the conceptual speaking range of LDA dim 1 (i.e., negative coordinate values). This is because we suspect the most striking variation in this area, as stated in the introduction. Specifically, we zoom in on conversations/phonecalls, broadcast interactions, and legal cross-examination as representatives for spoken text categories, and social letters and creative writing as representatives for written ones. We use the full ICE9 data set and the ICE9 LDA dimensions after they have been rotated to give a good visual match to ICE3. We look at documents at the extreme ends of the dimensions, and at texts that appear in unexpected places in the space. The texts at the extreme ends are particularly striking examples, as we illustrate with texts from the first dimension now.

We follow Neumann & Evert (2021) in labeling dimension 1 as expressing the difference between conceptual speaking and conceptual writing. In order to gain a more intuitive grasp of the underlying linguistic contrast, we can look at texts located at the extremes of this dimension. At the conceptual speaking side, there are six texts with dimension scores below –3.2, all of which are transcripts of phone calls. For the close reading analysis, we pick text icegb_s1a-095_2 from the British ICE component, an excerpt of which is shown as example (1):

(1)	icegb_s1a-095_2_A	126	Hallo
	icegb_s1a-095_2_B	127	Hallo
	icegb_s1a-095_2_B	128	How are you
	icegb_s1a-095_2_A	129	I'm very well thank you
	icegb_s1a-095_2_B	130	Jolly good
	icegb_s1a-095_2_B	131	How's work
	icegb_s1a-095_2_A	132	Uh oh not too bad
	icegb_s1a-095_2_A	133	not too bad
	icegb_s1a-095_2_B	134	Yeah
	icegb_s1a-095_2_B	135	Still all there
	icegb_s1a-095_2_B	136	Nobody else been thrown out
	icegb_s1a-095_2_A	137	Uh not yet no
	icegb_s1a-095_2_B	138	Not yet
	icegb_s1a-095_2_B	139	Is it in the offing or
	icegb_s1a-095_2_A	140	Well
	icegb_s1a-095_2_A	141	Could well be I think
	icegb_s1a-095_2_B	142	Yeah
	icegb_s1a-095_2_A	143	Uhm yeah
	icegb_s1a-095_2_A	144	Oh sorry the Gi Giles
	icegb_s1a-095_2_A	145	Giles's uh whatever they are lamb something or others are
	icegb_s1a-095_2_A	147	Sorry combined effort on the cooking going on there
	icegb_s1a-095_2_B	148	yeah
	icegb_s1a-095_2_A	149	Uh
	icegb_s1a-095_2_B	150	Yes so uh anything else interesting happening
	icegb_s1a-095_2_A	151	Uh well I don't quite know how to answer that one
	icegb_s1a-095_2_B	152	I see
	icegb_s1a-095_2_B	153	I see

We can see characteristically spoken language features, such as short sentence length, frequent use of personal pronouns (especially in the first and second person) and low lexical density. Though

our feature set does not include other stylometric features (such as (low) vocabulary richness and word frequency) or specific features of spoken language (such as hesitation markers, false starts, etc.), their prominence in this short excerpt is also striking. Notable is also the frequent use of discourse markers in sentence-initial position, which are characteristic of more casual (especially face-to-face) interactions due to the shared production context where they are commonly used to manage turn-taking. Given the high frequency of syn-semantic features such as pronouns and the low frequency of auto-semantic features such as nouns, the exchange appears focused on re-affirming the relationship and almost void of referential information. As such, this text is extremely prolific in its use of spoken language features – more so even than most other spoken texts. These observations can be linked to the quantitative analysis by measuring the contribution of individual features to the dimension score of text icegb_s1a-095_2, i.e., which features "push" the text towards the negative side of the dimension (for the general role of feature contributions in the interpretation of LDA dimensions, see Neumann & Evert 2021: 158–160). For example, we find that discourse markers in sentence-initial position alone contribute –0.76 to the dimension score. They are followed by the contributions of first and second person pronouns (–0.75 across four features), lexical density (–0.40), pronouns (–0.28), finites (as a proxy of clauses) per sentence (–0.21), and sentence length (–0.18). Together, these eight features already push the text to position –2.58 on dimension 1 (*conceptual speaking vs. conceptual writing*), well in the conceptual speaking end.

Let us now also look at the conceptual writing side of the dimension. In order to make a direct and meaningful comparison, it might be more interesting to look at spoken texts with positive dimension scores rather than at the most extreme written text categories. The highest scores of spoken texts in our data set are close to +2.0 (compared to the overall maximum of +2.56 for written texts). We select the only text from ICE-GB among the top ranks (icegb_s2a-033_1), which happens to be an unscripted speech (whereas other extreme texts are scripted broadcast news and talks). Its dimension score of +1.88 can be traced to contributions from low frequencies of personal pronouns (+0.60 from four individual features) and pronouns in general (+0.29), few sentence-initial discourse markers (+0.23), high lexical density (+0.22), relatively few sentence-initial verbs (+0.09), and many attributive adjectives (+0.08), as well as long sentences (+0.08).[8] Together, these features push the text to position +1.60 on dimension 1. It has to be noted that there are also some features with negative contributions that are more characteristic of spoken texts, but their total contribution is only –0.33. These characteristics are clearly visible in the first 4 turns of the text given as example (2).

(2)	icegb_s2a-033_1_A	1	The term clinical trial can be applied to any type of planned experiment which involves patients and is des designed to determine the most appropriate treatment for future patients with a given medical condition
	icegb_s2a-033_1_A	2	Historically medicine has advanced in a haphazard and in inefficient way
	icegb_s2a-033_1_A	3	The first clinical trial wasn't run until nineteen forty-eight and it's only in recent years that it's become widely recognised that properly conducted clinical trials which follow the principles of scientific experimentation provide the only reliable basis for evaluating the efficacy and safety of new treatments
	icegb_s2a-033_1_A	4	The essential characteristic of a clinical trial is that the results based on a limited sample of patients are used to make inferences about treatment for the general population of patients

In a similar vein, let us also consider texts belonging to the written text categories located on the negative side of the dimension, which captures conceptual speaking. In particular, these are texts from the categories of creative writing and social letters, which are produced in the written mode

8 Pronouns in general and attributive adjectives have particularly extreme values (among the top 3% and 1%, respectively) even when compared to written texts.

but linguistically more akin to spoken texts, hence their location near the truly spoken texts. In other words, unlike the previous text categories, these texts are clearly interactions which do not involve co-presence of the participants, but they still retain features of more conversational style in some respects. In particular, we would usually not expect to see many sentence-initial discourse markers in social letters, but lexical density should still be low and accompanied by more personal pronouns compared to other text categories like academic writing. This is because the interactants will presumably still share a significant amount of communicative background, despite not being physically part of the same situation. A similar observation can be made for creative writing, which may lack common ground between author and reader but often make up for this with the heavy use of dialogic segments that diegetically present a shared situation for the characters in the narrative. As such, they are likewise typically characterized by a frequent use of pronouns as well as interrogatives as a device for moving the story forward.

However, both text categories show considerable variability and extend quite far into the conceptual writing side of dimension 1. Example (3) contains an excerpt from the fifth most extreme data point in the two categories, a text from the Canadian component (icecan_w2f-013_1). Its dimension score of +1.38 places this text firmly in the conceptual writing range.

(3) icecan_w2f-013_1 My father's peripatetic piety includes not only the missions but other churches and tabernacles, some far afield, in Elizabeth, in West New York, on Staten Island, all of which require hours of buses and ferries and buses again, several churches in Manhattan which are reached by subway, a place in Park Slope where a renegade group from Elim calling themselves The Latter Rain anoint one another apostles and prophets and speak ecstatically of last things.

This single sentence is very long, abundant in rare words, and it features complex syntactic structures – as such it can be seen as a text that is more "written" than most written texts. Contrary to what one might think at first sight, the positive score does not result from a small number of such fairly extreme features outweighing negative contributions from many other features that are more characteristic of conceptual speaking. In fact, the total negative contribution for this text is only –0.24 compared to a total positive contribution of +1.62. The impression from the small excerpt above is partially confirmed insofar as syntactic complexity (finite verbs per sentence, +0.14), sentence length (+0.13), and lexical density (+0.06) do make noticeable contributions. For the entire text, they are overshadowed by personal pronouns (total contribution +0.42) and sentence-initial discourse markers (+0.23), though.

In order to obtain a thorough understanding of the dimensions, it is also of interest to explore a text category that appears in an unexpected place in the register space, as it may shed more light on what registers may have in common on one dimension, while being unrelated on another. We observed that some texts from the spoken mode appear on the dialogic written side of dimension 2 (*dialogic written vs. neutral*), where written texts would be expected. These texts are legal cross-examinations and are extremely dialogic in nature, consisting almost exclusively of question-answer pairs, which are in line with the contribution of interrogatives to the position of texts on this dimension. At the same time, however, they are more technical and denser than what spoken language often is, so they also generally contain more nouns, leading to a higher lexical density. Example (4) is an excerpt from a legal cross examination from the Indian component with a position far on the side of dialogic written texts.

(4)	iceind_s1b-062_1_a	Mr Angale
	iceind_s1b-062_1_b	Yes
	iceind_s1b-062_1_a	in the year nineteen eighty-one what business you were doing ?
	iceind_s1b-062_1_a	Were you in service or business ?
	iceind_s1b-062_1_b	No I was in business
	iceind_s1b-062_1_a	What business were you doing in eighty-one ?
	iceind_s1b-062_1_b	That time I was an artist that those days and I was running my shows in Bombay
	iceind_s1b-062_1_a	What shows ?
	iceind_s1b-062_1_b	variety entertainment programme shows

iceind_s1b-062_1_a	I see not any business you were not in business I suppose
iceind_s1b-062_1_b	No not that time
iceind_s1b-062_1_a	I was not in business in the year nineteen eighty-one
iceind_s1b-062_1_c	However I was doing varieties of entertainment
iceind_s1b-062_1_a	Varieties entertainment programme

What we hope to have shown in this close reading of examples is the overall variance in the data set. As shown throughout by the scatterplots, each text category (and arguably register) is characterized by a great deal of variation. That is why we follow Neumann & Evert (2021) in not reporting mean values for the text categories. And that is also what our discussion of specimens of the different text categories has demonstrated.

6 Conclusion

To conclude, we find that both our reproduction and replication efforts for the results of Neumann & Evert (2021) were largely successful. As such, our work not only further corroborates fundamental assumptions about register-based linguistic variation, but also underlines that quantitative approaches become generally quite reliable with increasing amounts of data (as demonstrated by our two replication experiments). By introducing additional data (roughly twice the amount of the original data set) into an already existing LDA register space we showed that the overall structure of its categories generalizes across language varieties, although the extra data cast a different light on some patterns of regional variation. The LDA focus space obtained from a multivariate analysis of the full extended data set was also very similar to the original register space. While immediately obvious for the first two dimensions, the full extent of the similarities only became apparent after a rotation that aligned the dimensions of the new focus space with the original space. Without the previous study, the analysis might have led to different interpretations and conclusions for the third and fourth dimension. This, then, suggests that the LDA register space is robust and stable per se, but it is challenging to grasp its four-dimensional geometric structure intuitively, which can possibly mislead the linguistic interpretation. We demonstrated that the tools provided by the GMA approach, in particular weakly-supervised analysis and its visualization, are an excellent starting point for reproducible research, but they might need to be extended with more sophisticated automatic and manual rotation algorithms in order to uncover the most interpretable perspective and dimensions.

Funding information

The authors gratefully acknowledge funding by the German Research Council (DFG) and the Swiss National Fund (SNF) for the WEAVE project *QuanTOR – Quantitative Analysis of Textual Organisation across Registers* (DFG project no. 528467412).

Author contributions

All authors contributed to the conceptualization of the study. Data curation was performed by FF, data analysis was performed and written up by SE, close reading was described by GS and FF. All authors contributed to writing the manuscript.

References

Abdi, Hervé. 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics* 2 (1). 97–106. https://doi.org/10.1002/wics.51.

Baker, Monya. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533 (7604). 452–454. https://doi.org/10.1038/533452a.

Barba, Lorena A. 2018. Terminologies for reproducible research. *arXiv*. https://doi.org/10.48550/arXiv.1802.03311.

Belhumeur, Peter Neil, João Pedro Hespanha & David Jay Kriegman. 1997. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (7). 711–720. https://doi.org/10.1109/34.598228.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024.

Biber, Douglas & 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.

Desagulier, Guillaume. 2020. Multivariate exploratory approaches. In Magali Paquot & Stefan Thomas Gries (eds.), *A practical handbook of corpus linguistics*, 435–469. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_19.

Diwersy, Sascha, Stefan Evert & Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis*, 174–204. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110317558.174.

Evert, Stefan & Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In Gert de Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies*, 47–80. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110459586-003.

Evert, Stephanie. 2025. gmatools: Geometric multivariate analysis in R. R package version 0.1. https://github.com/schtepf/GMA/.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf.

Flanagan, Joseph. 2025. Reproducibility, replicability, robustness, and generalizability in corpus linguistics. *International Journal of Corpus Linguistics*. Online first. https://doi.org/10.1075/ijcl.24113.fla.

Garside, Roger & Nicolas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In *Corpus annotation: Linguistic information from computer text corpora*, 102–121. London: Longman.

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

Halliday, M.A.K. & Ruqaiya Hasan. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, M.A.K. & Christian M.I.M. Matthiessen. 2014. *Halliday's introduction to functional grammar*. 4th edition. London: Routledge. https://doi.org/10.4324/9780203431269.

Hundt, Marianne, Gerold Schneider & Elena Seoane. 2016. The use of the *be*-passive in academic Englishes: Local versus global usage in an international language. *Corpora* 11 (1). 29–61. https://doi.org/10.3366/cor.2016.0084.

Koch, Peter & Wulf Oesterreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36 (1). 15–43. https://doi.org/10.1515/9783110244922.15.

Kortmann, Bernd, Kerstin Lunkenheimer & Katharina Ehret. 2020. cldf-datasets/ewave: The Electronic World Atlas of Varieties of English. Zenodo. https://doi.org/10.5281/ZENODO.3712132.

Kruger, Haidee & Bertus van Rooy. 2018. Register variation in written contact varieties of English: A multidimensional analysis. *English World-Wide. A Journal of Varieties of English* 39 (2). 214–242. https://doi.org/10.1075/eww.00011.kru.

Krzanowski, Wojtek J., Philip Jonathan, W. V. McCarthy & Mervyn R. Thomas. 1995. Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Applied Statistics* 44 (1). 101–115. https://doi.org/10.2307/2986198.

Lehmann, Hans Martin & Gerold Schneider. 2012. BNC Dependency Bank 1.0. In Signe Oksefjell, Jarle Ebeling & Hilde Hasselgård (eds.), *Aspects of corpus linguistics: Compilation, annotation, analysis*. Helsinki: Research Unit for Variation, Contacts, and Change in English, online. https://www.zora.uzh.ch/id/eprint/68479/.

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: Benjamins. https://doi.org/10.1075/z.195.

Merenda, Peter Francis 1997. A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development* 30 (3). 156–164. https://doi.org/10.1080/07481756.1997.12068936.

Munafò, Marcus Robert & George Davey Smith. 2018. Robust research needs many lines of evidence. *Nature* 553 (7689). 399–401. https://doi.org/10.1038/d41586-018-01023-3.

Neumann, Stella. 2014. *Contrastive register variation: A quantitative approach to the comparison of English and German* (Trends in Linguistics. Studies and Monographs 251). Berlin: De Gruyter Mouton.

Neumann, Stella. 2020. On the interaction between register variation and regional varieties in English. *Language, Context and Text. The Social Semiotics Forum* 2 (1). 121–144. https://doi.org/10.1075/langct.00023.neu.

Neumann, Stella & Stefan Evert. 2021. A register variation perspective on varieties of English. In Elena Seoane & Douglas Biber (eds.), *Corpus-based approaches to register variation*, 143–178. Amsterdam: Benjamins. https://doi.org/10.1075/scl.103.06neu.

Nosek, Brian Arthur, Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B. Nuijten, Julia M. Rohrer, Felipe Romero, Anne M. Scheel, Laura D. Scherer, Felix D. Schönbrodt & Simine Vazirel. 2022. Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology* 73 (1). 719–748. https://doi.org/10.1146/annurev-psych-020821-114157.

Roberts, Margaret E., Brandon M. Stewart & Dustin Tingley. 2019. stm: An R package for structural topic models. *Journal of Statistical Software* 91 (2). https://doi.org/10.18637/jss.v091.i02.

Schneider, Edgar W. 2007. *Postcolonial English. Varieties around the world*. Cambridge, UK: Cambridge University Press.

Schneider, Gerold & Max Lauber. 2019. *Statistics for linguists: A patient, slow-paced introduction to statistics and to the programming language R*. Zurich: Digitale Lehre und Forschung UZH. https://doi.org/10.5167/UZH-183632.

Sönning, Lukas & Valentiin Werner. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59 (5). 1179–1206. https://doi.org/10.1515/ling-2019-0045.

Szmrecsanyi, Benedikt & Bernd Kortmann. 2009. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua* 119 (11). 1643–1663. https://doi.org/10.1016/j.lingua.2007.09.016.

Torkkola, Kari. 2001. Linear discriminant analysis in document classification. In *IEEE ICDM workshop on text mining*, vol. 29. https://www.cse.unr.edu/~buchmann/doc_class/papers/LDA_doc_classification.pdf

Van Rooy, Bertus, Lize Terblanche, Christoph Haase & Josef J. Schmied. 2010. Register differentiation in East African English: A multidimensional study. *English World-Wide: A Journal of Varieties of English* 31 (3). 311–349. https://doi.org/10.1075/eww.31.3.04van.

Whitaker, Kristie. 2017. *Publishing a reproducible paper* [Conference presentation]. Open science in practice summer school, Lausanne, Switzerland. https://doi.org/10.6084/m9.figshare.5440621.v2

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York: Routledge. https://doi.org/10.4324/9781315165547.

Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28 (4). 421–450. https://doi.org/10.1111/j.1467-971X.2009.01606.x.