Register variation in school EFL textbooks

Elen Le Foll
Osnabrück University

This study applies additive Multi-Dimensional Analysis (MDA) (Biber 1988) to explore the linguistic characteristics of 'school English' or 'textbook English'. It seeks to find out how text registers commonly featured in English as a Foreign Language (EFL) textbooks differ from comparable registers found outside the EFL classroom. To this end, a Textbook English Corpus (TEC) of 43 coursebooks used in European schools is mobilised. The texts from six textbook register subcorpora and three target language corpora are mapped onto Biber's (1998) 'Involved vs. Informational' dimension of General English. Register accounts for 63% of the variance in these dimension scores in the TEC. Additional factors such as textbook level, series and country of publication/use only play a marginal role in mediating textbook register variation. Textbook dialogues score considerably lower than the Spoken BNC2014, whereas Textbook Fiction scores closest to its corresponding reference Youth Fiction Corpus. Pedagogical and methodological implications are discussed.

Keywords: coursebooks, language teaching materials, multidimensional analysis (MDA), textbook English

1. Background

1.1 School English as a Foreign Language (EFL) textbooks

Although no reliable data on textbook usage is available, it would appear that virtually all lower secondary EFL classrooms in Europe are equipped with textbooks. In most cases, they are the *de facto* interpretation of the curriculum and their tables of contents dictate the syllabus (cf. Vellenga 2004). At lower secondary level, few additional materials are used; hence, textbooks can be assumed to be the main source of language input for at least the first four to five years of EFL learning at secondary school (Usó-Juan & Martínez-Flor 2010: 424). If, as postulated by usage-based approaches, language learning is driven by frequency and frequency distributions of exemplars within constructions (cf. Ellis & Collins 2009), under-

https://doi.org/10.1075/rs.20009.lef | Published online: 23 November 2021
Register Studies 3:2 (2021), pp. 207–246. ISSN 2542-9477 | E-ISSN 2542-9485

standing what characterises the type of language that learners are exposed to via their textbooks is crucial to understanding learner language development.

Though corpus-based textbook analysis can be traced back to the pioneering work of Dieter Mindt in the 1980s, secondary school (as opposed to university-level) EFL textbook language remains a relatively understudied area. As for the transfer of corpus linguistic insights into EFL textbooks, the much-awaited breakthrough has yet to materialise (cf. Römer 2006). Although some textbooks authors and publishers have started to make use of corpora, the rise in the number of corpus-informed pedagogical publications appears to primarily apply to learner dictionaries, grammars, English for Special Purposes (ESP) and English for Academic Purposes (EAP) textbooks (Meunier & Gouverneur 2009: 180–181). With few exceptions (e.g., Cambridge University Press), general EFL textbooks, especially those designed for national primary and secondary school markets, remain largely unaffected by these developments (personal communication with French and German publishers).

The school EFL textbooks examined in the present study are designed to provide sufficient materials for a whole school year's worth of (in most cases compulsory) English lessons at lower secondary school in France, Germany and Spain, where communicative approaches to foreign language teaching are favoured. They thus explain and provide exercises for grammar and vocabulary, as well as include tasks designed to develop reading, writing, speaking, listening and mediation skills. It can be assumed that the majority of the texts featured in these textbooks have been (co-)written by the authors of the textbooks since only very few texts, mostly from the fiction register, are clearly labelled as extracts or simplified versions of original texts (e.g., novels, newspaper articles).

1.2 Textbook English studies

If learners are to be equipped with the necessary skills to deal with real-world communicative situations in English, it is crucial that they be exposed to the kinds of language patterns that they will later on encounter outside the classroom. However, previous corpus-based Textbook English studies have shown how a range of linguistic features are frequently (mis-)represented in ESL/EFL textbooks as compared to various interpretations of what is often termed "authentic", "natural" or "real" English (for an overview of textbook research from 1990 to 2009, see Meunier & Gouverneur 2009). In the past, the reference corpora of choice for such comparisons have often been well-known general British or American English corpora such as the British National Corpus 1994 (e.g., Chujo 2004; Gabrielatos 2013). Whilst these are carefully sampled, balanced corpora, they tend to have a strong focus on edited, professionally-written language, which may not correspond to secondary school EFL learners' target language (Le Foll 2020).

The range of language features studied in previous Textbook English research ranges from the use of individual words (e.g., Conrad 2004 on the preposition *though*) and phraseological patterns (e.g., Gouverneur 2008 on high-frequency verbs), to tenses and aspects (e.g., Barbieri & Eckhardt 2007 on reported speech; Römer 2004 on modals), and has more rarely ventured into the study of pragmatics (e.g., hedging in ESP/EAP textbooks, Hyland 1994) and spoken grammar (Gilmore 2004). However, they have only ever focused on one or at most a handful of individual features. Taken together, these studies provide valuable insights into "the kind of synthetic English" (Römer 2004: 185) that pupils are exposed to via their school textbooks. However, three crucial aspects have commonly been neglected in past endeavours to study the language of EFL/ESL textbooks.

First, interactions between the frequencies of individual linguistic features have generally not been considered. Usage-based approaches to language acquisition, however, claim that the co-occurrence information that learners perceive in language input "is stored as points in a multi-dimensional space at coordinates, and that speakers process this stored linguistic information in ways that allow them to identify (under certain conditions and defined by various types of frequency occurrences) abstract linguistic patterns" (Rautionaho & Deshors 2018: 229). Thus, whilst some influential studies have helped us understand how EFL/ESL learners can be misled by their textbooks to make unidiomatic use of specific linguistic features (e.g., progressive aspect, Römer 2005), only a multivariate approach can paint the full picture as to how "Textbook English" – as a whole – differs from the English that language learners will later encounter outside the classroom.

The second frequently neglected aspect concerns potential register differences between the various types of texts typically featured in school foreign language textbooks. It has long been established that situational characteristics of texts are a major driver of functional linguistic variation (cf., e.g., Biber 2012; Gray & Egbert 2019). Given that school EFL textbooks may feature, for example, extracts of a short story, a dialogue, instructions, and exercises on any double page, Textbook English cannot be meaningfully examined without taking a register-based approach. Up until now, however, register variation within EFL textbooks has largely been ignored (however, see Miller 2011 with respect to university-level ESL textbooks). In the few cases where register has been taken into consideration, the focus has almost exclusively been on the representations of spoken language, e.g., Mindt (1987, 1995) and Römer (2005) who compared the dialogues of secondary school EFL textbooks to corpora of spoken and pseudo-spoken native speaker English. However, to the author's best knowledge, other textbook registers, such as fiction, instructions, or informative texts, have yet to be explored in EFL textbooks.

Finally, previous quantitative corpus-based studies of textbook language have usually been undertaken at the corpus level (rather than at coursebook volume, chapter, unit or individual text level) and have therefore not been able to take the potential impact of the varying proficiency levels of the textbooks or any potential idiosyncrasies of textbook authors, editors or publishers into consideration.

1.3 A multivariate exploration of textbook English

Consequently, this study aims to explore the specificities of Textbook English by:

a. accounting for a broad range of lexical, grammatical and semantic features,
b. taking account of potential register differences within textbooks, and
c. using statistical methods that can model for the potential effects and interactions of textbook register, series and proficiency levels.

To do so, Biber's MDA framework is applied to the study of register variation in school EFL textbooks. In his pioneering study, Biber (1988) elaborated a robust model of language variation in written and spoken English along six dimensions (cf. Nini 2014, 2019 for an empirical validation of its generalisation to new texts using the Brown Corpus). The MDA framework uses factor analysis to reduce the co-occurrence patterns of a large set of lexico-grammatical features to a parsimonious set of latent factors, which are functionally interpreted (cf. Biber 1988; Conrad & Biber 2001/2013; Berber Sardinha & Biber 2014). Biber's (1988) model of general written and spoken English was elaborated on the basis of the co-occurrence patterns of 67 (largely automatically tagged) linguistic features observed in a large corpus covering a broad range of registers, including face-to-face conversation, press, official documents, letters, etc.

Post-1988, two approaches to register variation studies applying MDA have emerged. The first compares one or more new or more specialised registers relative to the dimensions of an earlier analysis of registers, most commonly Biber's (1988) model; this is referred to as additive MDA (cf. Berber Sardinha et al. 2019). The second consists of conducting a new, full MDA for an entire (new) set of registers (cf. Friginal & Hardy 2014; Egbert & Staples 2019). Additive MDAs bear the advantage of requiring considerably smaller datasets. Indeed, when conducting a full MDA, large and internally well-stratified corpora are essential to be able to extract meaningful dimensions. Where obtaining such data is not feasible, "plotting the input corpus onto Biber's model of English can be a reasonable approximation to running a new [MDA]" (Nini 2019: 70).

Despite this potential, relatively few studies have applied Biber's or other subsequent MDA-derived models to describe or evaluate new registers and/or varieties of English (Berber Sardinha et al. 2019). To date, two registers have been the focus of most additive MDAs: television language (cf. Quaglio 2009; Al-Surmi 2012; Forchini 2012; Berber Sardinha & Veirano Pinto 2017) and academic English (cf. Conrad 1996, 2001/2013; Biber et al. 2002, 2004) – all of which relied on Biber's 1988 model as their baseline.

Conrad (1996, 2001/2013) applied Biber's (1988) model to research articles and university-level textbooks. She compared dimension scores for the two registers (research articles/textbooks) and two disciplines (ecology/history). On Biber's first dimension, all disciplinary texts clustered at the negative, informational end of the scale, thus pointing to overall high informational density. However, fine-grained analyses of the many features that were entered in the MDA revealed notable differences between the two academic registers: the research articles featured more nouns, prepositions, attributive adjectives, and longer words, thus conveying information that is more densely packed than the textbooks that, by contrast, tended to feature more linguistically redundant explanations and examples.

In addition to textbook evaluation, additive MDA may also be used in the development of pedagogical materials: Zuppardo (2013) applied the method to compare the language of aircraft manuals to Biber's (1988) model. The results revealed the salient linguistic features of this specialised register. These can be used by teachers and textbook authors to develop ESP/EAP materials.

Whilst Conrad (1996, 2001/2013) and Zuppardo (2013) have demonstrated the potential of additive MDA in textbook language research, this method has yet to be applied to secondary school EFL textbooks. In fact, as mentioned above, EFL textbook studies, so far, have largely been univariate and, with few exceptions, have mostly ignored potential register-based linguistic variation within textbooks. It is quite probable that the sheer complexity of carrying out an MDA may have hitherto been prohibitive to applying Biber's (1988) model to applied research questions such as register variation in school EFL textbooks (cf. Nini 2019: 92). This study will therefore also investigate the potential of using a freely available and all-in-one programme (Nini 2019), which automatically tags, counts, and computes dimension scores for the first five of Biber's (1988) dimensions, for the analysis of register variation in secondary school EFL textbooks.

1.4 Aims and research questions

This paper aims to overcome some of the limitations of past EFL textbook studies by applying MDA to explore linguistic variation within school EFL textbooks and thus provide a more comprehensive view of the defining characteristics of Textbook English. This paper therefore seeks to tackle the following research questions:

RQ1. What is the extent of the linguistic variation across the major registers of Textbook English? Do some textbook series show significantly more or less register-based variation? Do the proficiency levels of textbooks significantly interact with register-based variation?

RQ2. To what extent do Textbook English registers differ from situationally-similar, naturally-occurring registers? Are any significant differences observed between different textbook series and/or the proficiency level of individual textbook volumes?

RQ3. What are some of the defining linguistic features that characterise Textbook English registers as compared to situationally-similar target language registers?

In addition, the strengths and limitations of applying additive MDA to the investigation of Textbook English using readily available software are considered and discussed.

2. Data and methods

2.1 Corpus design and data collection

2.1.1 *Textbook English corpus (TEC)*

The data explored in this paper is part of the Textbook English Corpus (TEC) (Le Foll in preparation). The TEC is made up of all the texts printed in 43 EFL coursebooks used in secondary schools in France, Germany and Spain, as well as the transcripts of the accompanying audio and video materials (see Table 1). Nine best-selling textbook series from eight major publishers are represented. Each series corresponds to the first four or five years of English instruction at secondary school level.

To be able to compare pedagogical materials used in different educational systems, each textbook was labelled for proficiency level on a universal scale of A to E (see Table 1): level A textbooks correspond to the first year of EFL instruction at secondary level, in other words, beginner level to roughly A1 on the CEFR scale (European Council 2004), whilst level E corresponds to the fifth year (CEFR B1–B2). French textbook series only cover the first four years of secondary school (which take place at *Collèges*), which is why, whenever possible, a textbook from the same publisher corresponding to the fifth year of instruction (the first year of

Table 1. Composition of the textbook English corpus (TEC)

Country of use	Publisher	Textbook series	Volume	Level	Publication date
France	Bordas	Hi There	6ème	A	2012
			5ème	B	2013
			4ème	C	2014
			3ème	D	2015
		New Mission	2nde	E	2014
	Nathan	Join the Team	6ème	A	2010
			5ème	B	2011
			4ème	C	2012
			3ème	D	2013
		New Bridges	2nde	E	2010
	Le Livre Scolaire	Piece of Cake	6ème	A	2017
			5ème	B	
			4ème	C	
			3ème	D	
Germany	Klett	Green Line	1	A	2006
			2	B	
			3	C	2007
			4	D	2008
			5	E	2009
	Klett	New Green Line	1	A	2014
			2	B	2015
			3	C	2016
			4	D	2017
			5	E	2018
	Cornelsen	Access	1	A	2013
			2	B	2014
			3	C	2015
			4	D	2016
			5	E	2017

Table 1. *(continued)*

Country of use	Publisher	Textbook series	Volume	Level	Publication date
Spain	Richmond	Achievers	A1+	A	2015
			A2	B	
			B1	C	
			B1+	D	
			B2	E	
	Cambridge University Press	English in Mind	Starter	A	2010
			1	B	
			2	C	
			3	D	2011
			4	E	
	Oxford University Press	Solutions	Elementary	A	2014
			Pre-Intermediate	B	2016
			Intermediate	C	2017
			Intermediate Plus	D	2017

Note.
For the full bibliographic metadata see http://doi.org/10.5281/zenodo.4922819

Lycée) was added. At the time of corpus compilation, Le Livre Scolaire did not produce any textbooks for *Lycées*.

Each of the 43 textbook volumes were digitalised and manually subdivided into text units, where one exercise, reading passage, or transcript corresponds to one text unit. At the same time, these texts were also coded for eight major textbook registers: Conversation, Informative texts, Fiction, Personal Correspondence (letters, diary entries, social media posts, and e-mails), Instructional texts (instructions and explanations), Poetry (songs and poems), Other texts (timetables, shopping lists, etc.) and Words & Phrases (e.g., contextless words and sentences from exercises). The categories Other texts and Words & Phrases were not analysed in the context of this paper. Example texts of the six textbook registers examined here can be found in the Appendix.

The coding was carried out by the author and a student research assistant. The coding scheme was developed following a cyclical categorisation process and was tested by having both coders blind-annotate three full textbook volumes and comparing the results. Inter-rater agreement rate was found to be satisfactorily high (96.65%). The only notable difficulty consisted in distinguishing between individual sentences and isolated words/phrases; hence these two categories were merged into one in the final annotation scheme. The use of custom macros activated using keyboard shortcuts considerably facilitated the XML annotation process and reduced the potential for inattention errors (Le Foll 2020).

The majority of textbook texts are too short for normalised linguistic feature counts to be meaningful. Linguists attempting to apply MDA to social media texts face a similar problem. To solve this issue in their multi-dimensional analysis of Twitter data, Clarke & Grieve (2017: 2) opted for binary feature frequencies (i.e., whether a feature is present or absent within a tweet) rather than relative frequencies. If, as Clarke & Grieve did, one considers a single tweet (as opposed to a thread of tweets) as a single text, this approach is very sensible because single tweets have, by corpus linguistic standards, a very small maximum character limit (currently 280 characters) and as a result, relative frequencies would largely depend on tweet length. The case of textbook texts, however, is much more complex: whilst many textbook texts are as short as a tweet (e.g., brief instructions, short rhymes), countless others run well over 1,000 words (e.g., short stories, newspaper articles). Indeed, defining text units in school EFL textbooks is a particularly challenging task. Numerous possibilities arise (cf. Le Foll 2020). Up until now, entire textbook volumes have often been conceived as single texts. However, as highlighted in Section 1.3, such an approach entirely ignores the variety of text registers encountered within a single textbook volume. A second approach might consider all the texts of one register found within a chapter or unit of a textbook volume to constitute one text. In some cases, this may be justified because texts within a textbook unit will often be thematically related and may therefore form a coherent whole; however, this will depend on the textbook series and is not always consistent across an entire textbook series, either.

In addition to the problem of defining text units, the great variety of text lengths encountered in school EFL textbooks must also be considered. Whilst there is no standard minimum text length for MDA studies, in order to carry out an additive MDA based on Biber's 1988 model, the type/token ratio variable must be calculated on the basis of the first 400 words of any text (Biber 1988: 238–239). It has long been established that type/token ratios must be calculated on the basis of text samples of equal text length as this lexical diversity measure is highly sensitive to text length (e.g., Brezina 2018: 58). Consequently, texts shorter than 400 words could not be included in the present analysis.

In light of both the great variety of text lengths encountered in school EFL textbooks and the fact that the majority are under 400 words, shorter texts within each textbook volume and register were collated into longer text files. This means that, for example, a number of short, consecutive instructional texts from any one textbook volume were combined until a total word count of at least 400 words was reached. This was done sequentially within each textbook volume so that short files from within a chapter/unit or across directly adjacent chapters/units are grouped together. Hence, the collated text files also correspond to the progression that the learners are expected to make. This resulted in the exclusion of Poetry texts from thirteen volumes, Fiction texts from seven volumes, and Informative texts from two volumes because the texts of these registers did not total to at least 400 words. Following these data preparation steps, 1,949 textbook text files were created (thereafter collectively referred to as the TEC, see Table 2).

Table 2. Textbook English Corpus (TEC) texts entered in the additive MDA

Textbook register	Number of texts	Number of words
Conversation	529	407,591
Fiction	285	205,072
Informative texts	363	265,224
Instructional texts	647	499,324
Personal Correspondence	88	58,534
Poetry	37	22,358
Total	**1,949**	**1,458,103**

2.2 Target language reference corpora

In answering RQ2 and RQ3, this paper focuses on three major textbook registers: Conversation, Fiction and Informative texts by comparing these three subcorpora of the TEC with reference corpora of situationally-similar target language registers. This section briefly outlines the composition of these reference corpora.

2.2.1 *Spoken BNC2014*

The Textbook Conversation subcorpus is compared to the Spoken BNC2014, an 11.4-million-word corpus of 1,251 orthographically transcribed conversations among L1 speakers in the U.K. (Love et al. 2017). The Spoken BNC2014 is rich in metadata and has been manually anonymised; however, for the purposes of this study, all mark-ups have been eliminated and anonymising tags replaced with placeholders of the corresponding word class (e.g., all anonymised place names have been replaced by *IVYBRIDGE*).

2.2.2 *Youth fiction corpus (YFC)*

The Fiction subcorpus of the TEC was compared to the Youth Fiction Corpus (YFC), which consists of 300 novels targeted at teenagers and young adults (Le Foll in preparation). This is a better match for the narrative texts featured in school EFL textbooks than the fiction included in Biber's 1988 corpus, both in terms of target readership and publication dates. For the present study, four random samples of approximately 5,000 words were extracted from each of the 300 books in the corpus (splitting was performed at sentence boundaries, hence the slightly varying word counts), except for three short stories, which were only sampled once each in full. With a total of 1,191 YFC texts, this procedure resulted in a number of texts comparable to that of the Spoken BNC2014.

2.2.3 *Informative texts for teens corpus (ITTC)*

The Informative Texts for Teens Corpus (ITTC) was built by originally retrieving over 10,000 texts from 14 popular web domains of news and information specially targeted at English-speaking teenagers. Care was taken to include a broad range of topics including current affairs, science, technology, history, and entertainment (Le Foll in preparation). Of these, 4,895 text files were under 400 words and were thus discarded for the MDA. Following a stratified sampling approach, 100 texts from each web domain were then randomly selected. This number was chosen to approximately match the number of texts in the other two reference corpora. Fewer than 100 texts longer than 400 words were retrieved from two domains; for these, the full domain datasets were retained. The final selection thus consisted of 1,414 text files (see Table 3).

2.3 Comparative additive MDA

For reasons of space, this paper focuses on register variation in secondary school EFL textbooks along Biber's (1988) first 'Involved vs. Informational Production dimension. With its 23 features that contribute to higher dimension scores (positive loadings) and six to negative scores (negative loadings) (see Table 8), this dimension is the most powerful predictor of register variation in Biber's corpus of general English. It accounts for 84% of the variation in Dimension 1 scores (Biber 1988: 126–127). It has since proven to be a stable and robust baseline for additive MDAs across a wide range of domains (cf. Egbert & Mahlberg 2020: 82). Furthermore, this dimension's 'involved/oral/verbal' vs. 'informational/literate/nominal' opposition has, for a range of languages and domains, almost universally emerged as the strongest and most stable predictor of variation in full MDAs post-1988 (Biber 2014).

Table 3. Informative texts of teen corpus (ITTC) texts entered in MDA

Domain name	Number of texts	Number of words
bbc.co.uk/history	100	74,722
dogonews.com	100	60,762
ducksters.com	100	67,894
encyclopedia.kids.net.au	100	74,566
factmonster.com	100	60,395
historyforkids.net	100	71,955
quatr.us	100	62,254
revisionworld.com (GCSE only)	100	74,301
sciencekids.co.nz	100	57,097
sciencenewsforstudents.org	100	82,258
teen.wng.org	85	45,515
teenkidsnews.com	100	81,765
teenvogue.com	100	82,117
tweentribune.com	29	26,166
whyfiles.org	100	85,492
Total	**1,414**	**1,007,259**

2.3.1 *Tagging and counting linguistics features*

To conduct an additive MDA using Biber's (1988) original model as "a base-rate knowledge of English" (Nini 2019: 70), it is necessary to tag and count exactly the same 67 features used in Biber's original study. This was achieved using the Multidimensional Analysis Tagger (hereafter MAT; Nini 2014, 2019): a freely available programme that aims to replicate the original Biber Tagger. It tags all 67 lexical, grammatical and semantic features using the regular expressions described in Biber (1988: 211–245), and normalises all feature frequencies to the number of occurrences per 100 words. The validity and reliability of the MAT as compared to the Biber Tagger has been demonstrated in Nini (cf. 2019: 92).

2.3.2 *Computing the mean dimension scores for the new registers*

To compute dimension scores, normalised counts must be standardised to avoid frequent features from having a disproportionate influence on the model. In an additive MDA, however, *z*-scores are not calculated on the basis of the features' means and standard deviations from the corpora under study, but rather from the original corpus from which the baseline model was derived. Consequently, texts whose normalised count for any one variable is equal to the variable mean in Biber's corpus (1988: 77) have a *z*-score of 0. Positive *z*-scores indicate that a feature occurs more frequently than on average across Biber's corpus, whilst negative *z*-scores indicate below average normalised counts.

Finally, to compute the dimension scores of the new texts, the *z*-scores of the features with positive loadings are added and those with negative scores are subtracted. The standardisation step and the computing of the dimension scores were also performed using the MAT.

2.3.3 *Computing dimension scores for additional reference corpora*

In theory, conducting an additive MDA makes it possible to compare "new" registers to Biber's "old" general English registers without resorting to any additional reference corpora. However, in this study, three target language reference corpora are also mapped onto Biber's first dimension for comparison with the registers of the TEC. Both theoretical and methodological reasons justify this additional step.

First, although the registers included in Biber's 1988 model undoubtedly provide useful comparison points for EFL textbook registers, any differences observed, say between Biber's fiction registers and the fiction featured in EFL textbooks, could potentially be due to different target readerships. Indeed, the fiction subcorpora of the Lancaster-Oslo-Bergen Corpus of British English (LOB) corpus, on which Biber based his original analysis, predominantly contain samples from literature aimed at an adult readership, rather than secondary school students. Further, the corpora from which Biber's model was derived consist of texts published in 1961 (LOB; Johansson, Leech, & Goodluck 1978) and spoken material recorded between 1953 and 1987 (London-Lund; Svartvik & Quirk 1980). Modern EFL textbooks, however, can reasonably be expected to reflect more recent language change, especially in the conversation register.

Second, whilst Nini (2014, 2019) demonstrated the overall reliability of the MAT, his analyses pointed to minor differences in some feature counts as compared to the original Biber Tagger. Needless to say, results of dimension score comparisons are more likely to be valid if the exact same method is used to tag and count the features of any corpora to be compared.

Consequently, this additive MDA compares register variation across six Textbook English registers, and additionally compares their Dimension 1 scores to three target language corpora.

2.3.4 *Comparing dimension scores*

To compare different registers on any one dimension, the mean dimension scores of all the texts in any one register can be compared to each other. Such comparisons have typically been tested and quantified using ANOVAs and coefficients of determination (e.g., Biber 1988: 95; Biber et al. 2004: 64; Gray 2015: 216; Berber Sardinha & Veirano Pinto 2019: 6), or with nonparametric Kruskal Wallis ANOVAs (Muhammad 2020). More recently, the use of predictive Discriminant Function Analysis (DFA) as a post-hoc analysis has been proposed to verify the robustness of dimensions as predictors of register (e.g., Crossley, Allen, & McNamara 2014; Crossley, Kyle, & Römer 2019; Veirano Pinto 2019). However, a crucial assumption of both ANOVAs and DFAs is that the data points be independent of each other (cf. Gries 2015; Winter 2019: chaps 14–15; on the consequences of using DFA on non-independent data, cf. Mundry & Sommer 2007). However, in the context of the present additive MDA, and, indeed, in many corpus linguistic studies, this assumption is not met. In the present study, each textbook series has largely been written by the same group of authors. They are thus not truly independent. Similarly, the YFC and the ITTC consist of several samples from any one book or web domain (see 2.2.2–2.2.3).

As a result, linear mixed effects models were computed using the R package *lme4* (Bates et al. 2015). First, register variation within Textbook English is modelled on Biber's Dimension 1. To estimate the relationships between textbook register and Dimension 1 scores, a model was fitted with a random effect structure consisting of by-series varying intercepts and by-series varying slopes for each register to account for the non-independence of texts from within one textbook series. Dimension 1 scores are the outcome variable. Textbook register and textbook level are modelled as fixed level predictors. In addition, their two-way interaction term is also fitted, since it can be hypothesised that, as the proficiency of learners increases, the dimension scores of textbook texts within a register may move closer to their target language equivalents. For instance, upper-intermediate fictional texts from textbooks may be more like teenage/young adult fiction than a short story printed in a beginner textbook. If this were true, we would expect Dimension 1 scores for some registers to increase as learners are expected to become more proficient, whilst they may decrease for others.

To compare the Dimension 1 scores of Textbook Conversation, Fiction and Informative texts with the three corresponding target language reference corpora, a second mixed effect model was computed. In this model, the random effect structure consists of varying by-source intercepts and slopes, where 'source' corresponds to a factor variable with nine textbook levels corresponding to each textbook series for the TEC corpus, 300 book levels for the YFC, 14 web domain levels for the ITTC, and one level for the Spoken BNC2014. These levels have been chosen as the best-available proxies to capture the variation inherent to each (group of) author(s)/editor(s). The fixed effects are corpus type (Textbook vs. Target Language Reference), register (Conversation, Fiction and Informative texts) and their two-way interactions.

For data sparsity reasons, a subset of the data that excluded the textbook register Poetry was used for all statistical modelling since several textbook volumes do not include any poems or songs longer than 400 words that could therefore be entered in the MDA (see 2.1.1).

Model diagnostic plots were inspected to check the assumptions of linearity, homogeneity of variance, and the normal distribution of residuals of the model (i.e., the differences between the observed and fitted values).

In the model summaries, the CI ranges reported are 95% confidence intervals. The R^2-values reported summarise the predictive power of the fixed effects only (R^2_{marginal}) and of both fixed and random effects ($R^2_{\text{conditional}}$) and were computed using the R package *sjPlot* (Lüdecke 2020). The estimators of relative contrast effects between each register under study were calculated using the default parameters of the *emmeans* package (Lenth 2020). *P*-value adjustment followed the Tukey method (confidence level = 0.95).

3. Results and discussion

Section 3.1 explores intra-textbook linguistic variation by comparing six Textbook English registers on Biber's Dimension 1 (RQ1). Large within-register dispersions are further examined and examples of salient features that contribute to strikingly low or high scores are discussed in context. This is followed, in Section 3.2, by a more fine-grained comparison of three key textbook registers (Conversation, Informative texts and Fiction) to three comparable target language corpora (see Section 2.2) with the aim of investigating the extent to which textbook registers differ from similar registers encountered outside the classroom (RQ2). The results of this comparative additive MDA provide answers to RQ3 which seeks to pinpoint the linguistic features which most contribute to these differences. Limitations of the method are discussed throughout the results and summarised in the concluding discussion (see Section 4).

3.1 Variation across textbook English registers

As expected, among the six textbook registers, Textbook Conversation scores highest on Biber's first dimension ($\bar{x} = 15.75$, $SD = 7.89$), followed by Personal Correspondence ($\bar{x} = 9.62$, $SD = 6.81$) and Fiction ($\bar{x} = 5.03$, $SD = 8.29$). The lowest scores are found in the Informative ($\bar{x} = -5.26$, $SD = 7.53$) and Instructional ($\bar{x} = -4.69$, $SD = 4.60$) registers.

As illustrated in Figure 1, textbook register is clearly a strong predictor of Dimension 1 scores among textbook texts. A simple model featuring only register

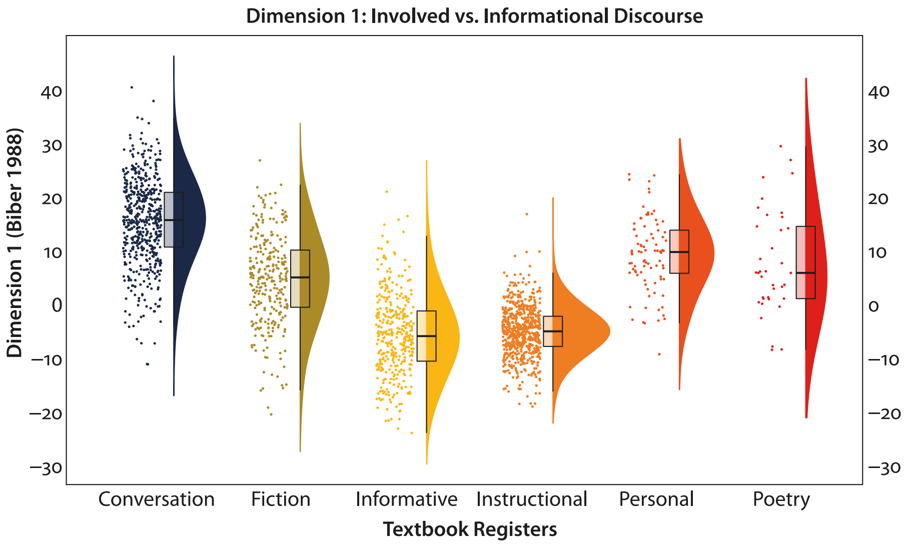

Figure 1. The six registers of the textbook English corpus (TEC) on Biber's (1988) Dimension 1 (Le Foll 2021. Zenodo. Retrieved 7 May 2021. http://doi.org/10.5281/zenodo.4732286)

as a fixed effect and by-series varying intercepts already accounts for some 63% of the variance in Dimension 1 scores ($R^2_{marginal} \approx 0.63$, $R^2_{conditional} \approx 0.66$). Although model comparisons revealed that the proficiency level of textbooks is also a significant predictor of Dimension 1 scores ($\chi^2(4) = 52.27$, $p < 0.001$, as compared to the baseline model), its predicting power is very weak ($R^2_{marginal} \approx 0.03$, $R^2_{conditional} \approx 0.08$). We can thus conclude that text register within textbooks is a much stronger driver of linguistic variation than the proficiency levels the textbooks are designed for.

The full model for intra-textbook variation along Dimension 1 is summarised in Table 4. It is a fairly good predictor of Dimension 1 scores with a predictive power of 65% with fixed predictors only, and 71% with both fixed and random effects. Figure 2 presents a visualisation of the model summarised in Table 4. In addition to providing a visualisation of the model fit, Figure 2 also serves as a reminder of the categories for which there is only sparse or no data: e.g., there are few Personal Correspondence texts, the textbook series *Piece of Cake* (POC) and *Solutions* only go as far as Level D, and some series feature very few or no Fiction texts at certain levels (see 2.1.1).

With Textbook Conversation scoring highest and Textbook Informative at the bottom of the scale, the distribution of scores on this first dimension echoes Biber's original as well as subsequent additive MDAs. The results indicate that

Table 4. Summary of the model for intra-textbook variation along Biber's (1988) Dimension 1: Dim1 ~ (Register|Series) + Register + Level + Level: Register

Predictors	**Estimates**	**CI**	***p*-value**
(Intercept)	16.41	14.35 – 18.47	**< 0.001**
Register [Fiction]	−6.53	−9.06−−4.00	**< 0.001**
Register [Informative]	−20.02	−23.23−−16.80	**< 0.001**
Register [Instructional]	−21.21	−23.70−−18.73	**< 0.001**
Register [Personal]	−5.87	−9.69−−2.05	**0.003**
Level [B]	0.30	−1.38 – 1.98	0.723
Level [C]	−0.41	−2.08 – 1.26	0.631
Level [D]	−2.23	−3.92−−0.54	0.010
Level [E]	−1.68	−3.61 – 0.25	0.088
Register [Fiction] * Level [B]	−3.92	−7.02−−0.83	0.013
Register [Informative] * Level [B]	−0.86	−4.02 – 2.29	0.592
Register [Instructional] * Level [B]	0.59	−1.72 – 2.90	0.618
Register [Personal] * Level [B]	−2.09	−6.68 – 2.49	0.371
Register [Fiction] * Level [C]	−4.33	−7.44−−1.21	**0.006**
Register [Informative] * Level [C]	−0.45	−3.47 – 2.57	0.771
Register [Instructional] * Level [C]	1.36	−0.92 – 3.65	0.242
Register [Personal] * Level [C]	−0.81	−5.47 – 3.84	0.732
Register [Fiction] * Level [D]	−6.54	−9.57−−3.52	**< 0.001**
Register [Informative] * Level [D]	−1.10	−4.11 – 1.91	0.472
Register [Instructional] * Level [D]	2.25	−0.05 – 4.55	0.055
Register [Personal] * Level [D]	0.06	−4.82 – 4.94	0.982
Register [Fiction] * Level [E]	−5.90	−9.03−−2.76	**< 0.001**
Register [Informative] * Level [E]	−3.28	−6.51−−0.06	0.046
Register [Instructional] * Level [E]	−0.07	−2.62 – 2.48	0.957
Register [Personal] * Level [E]	0.25	−4.67 – 5.17	0.919
	Random Effects		
σ^2	38.31		
$\tau_{00\ \text{Series}}$	6.07		
$\tau_{11\ \text{Series.RegisterFiction}}$	1.03		
$\tau_{11\ \text{Series.RegisterInformative}}$	9.07		
$\tau_{11\ \text{Series.RegisterInstructional}}$	7.19		
$\tau_{11\ \text{Series.RegisterPersonal}}$	3.20		
ICC	0.17		
N_{Series}	9		
Observations	1,912		
R^2_{marginal} and $R^2_{\text{conditional}}$		0.645 / 0.706	

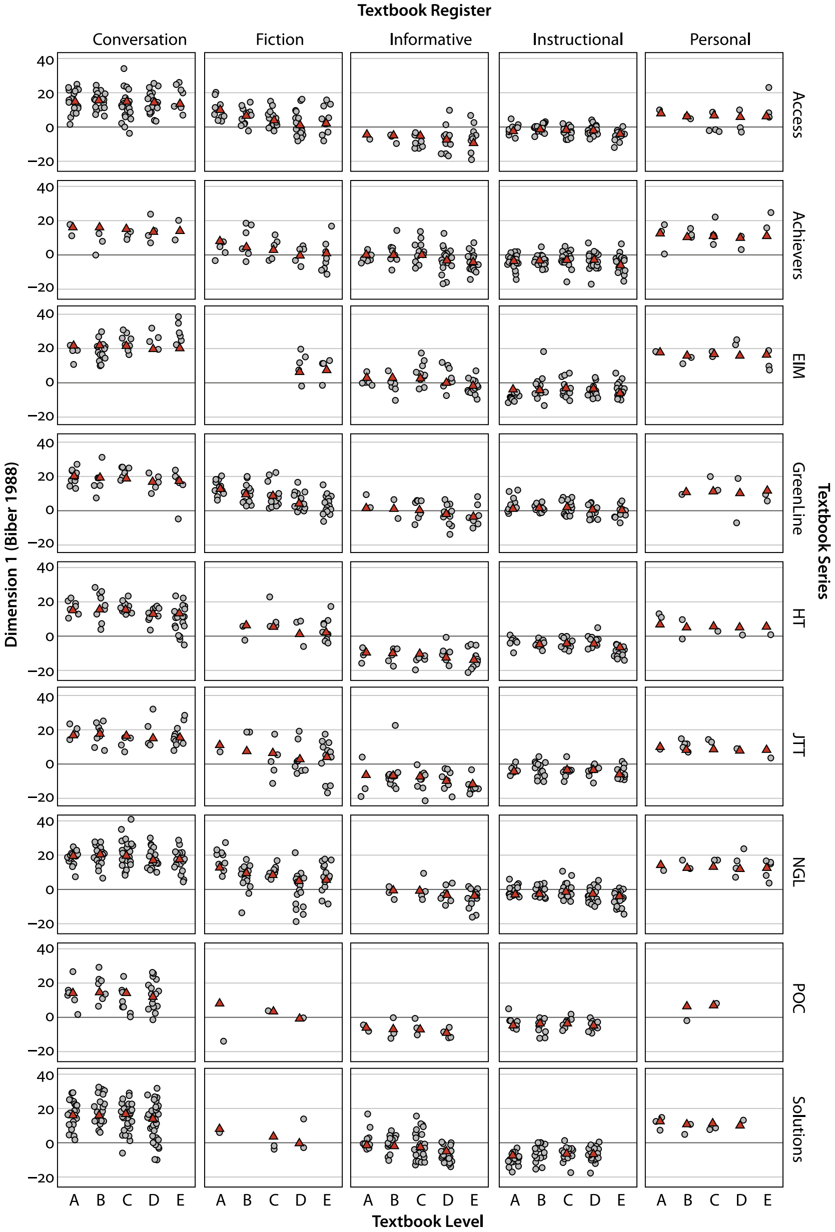

Figure 2. Observed (grey circles) and predicted (red triangle) Dimension 1 scores across textbook register, level and series. Predicted values as computed by the model summarised in Table 4 (Le Foll 2021. Zenodo. Retrieved on 7 May 2021. http://doi.org/10.5281/zenodo.4732323)

textbook authors do make different, register-based linguistic choices when crafting the texts of secondary school EFL textbooks. Indeed, Table 5 shows that the register means for Dimension 1 are all significantly different from each other ($p < .001$), except for the Informative-Instructional, Conversation-Personal Correspondence, and Fiction-Personal Correspondence contrasts (as illustrated in Figure 2, the latter two are likely due to the fact that there are relatively fewer Personal Correspondence texts in the TEC). Thus, these results confirm the need to examine textbook language under the lens of register. Indeed, textbook register appears to have a much larger impact on the choice and frequencies of linguistic features of the texts featured in textbooks than the proficiency level of the textbook, or the linguistic idiosyncrasies of its authors (as, admittedly imperfectly, captured in the textbook series variable).

Table 5. Estimated differences between estimated mean scores for each textbook register

Contrasts	**Estimates [95% CI]**	**SE**	**df**	***p*-value**
Conversation – Fiction	10.67 [7.84–13.49]	0.80	7.30	< **.001**
Conversation – Informative	21.16 [17.24–25.07]	1.20	10.28	< **.001**
Conversation – Instructional	20.39 [16.95–23.85]	1.05	9.84	< **.001**
Conversation – Personal Correspondence	6.39 [2.80–9.98]	1.06	8.56	0.002
Fiction – Informative	10.49 [5.49–15.48]	1.53	10.26	< **.001**
Fiction – Instructional	9.72 [6.17–13.27]	1.11	11.65	< **.001**
Fiction – Personal Correspondence	−4.28 [−8.88–0.33]	1.40	9.86	0.072
Informative – Instructional	−0.77 [−5.62–4.09]	1.48	10.21	0.983
Informative – Personal Correspondence	−14.76 [−18. 51−−11.02]	1.13	9.82	< **.001**
Instructional – Personal Correspondence	−14.00 [−19. 22−−8.78]	1.58	9.82	< **.001**

3.2 The specificities of textbook English registers

Having examined the extent of register variation *within* school EFL textbooks, this section compares three major textbook registers: Conversation, Fiction and Informative texts, with comparable target language reference corpora (see Section 2.2) on Biber's Dimension 1. The distribution of scores, as calculated with the MAT, is illustrated in Figure 3.

Although Textbook Conversation scored highest among the textbook registers, the Spoken BNC2014 displays considerably higher scores than Textbook Conversation ($\bar{x}=15.75$, $SD=7.89$ vs. $\bar{x}=26.02$, $SD=4.04$). Crucially, this difference is, in fact, even greater because the results plotted in Figure 3 correspond to the

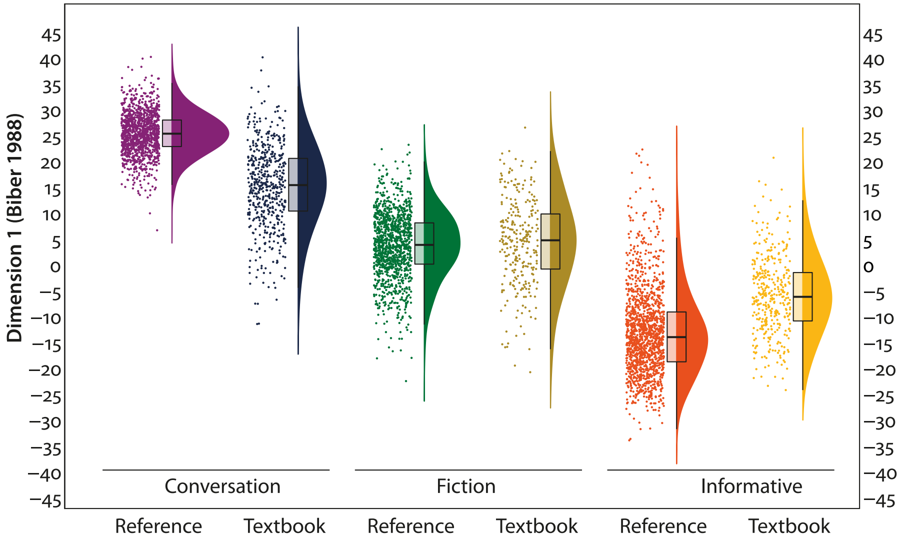

Figure 3. Comparison of the conversation, fiction and informative texts from the TEC with the three corresponding target language reference corpora on Biber's (1988) Dimension 1 (as calculated by the MAT) (Le Foll 2021. Zenodo. Retrieved on 7 May 2021. http://doi.org/10.5281/zenodo.4732334)

unaltered MAT output, in which Dimension 1 scores of the Spoken BNC2014 are artificially deflated: this is caused by the absence of punctuation marks in the Spoken BNC2014. Indeed, the Biber Tagger and, as its faithful "copy", also the MAT, require the presence of punctuation marks and/or prosodic boundary markers to identify five of the 22 features with positive loadings on Biber's Dimension 1: stranded prepositions, discourse particles, non-phrasal clause coordination, sentence relatives and direct WH-questions (Biber 1988: Appendix II). The transcription scheme of the Spoken BNC2014, however, does not include any punctuation signs except question marks (Love, Hawtin, & Hardie 2018: 37–38). Thus, for example, following the operationalisation of the discourse marker variable used in Biber's original MDA, only discourse particles preceded by a punctuation mark are tagged and counted.

Consequently, the five aforementioned features that rely on punctuation and/or prosodic boundary markers had to be excluded from the Dimension 1 scores of the Spoken BNC2014, and for comparability reasons, also from those of Textbook Conversation. This means that, in this particular case, it is not possible to apply Biber's (1988) model one-to-one and, consequently rely solely on Nini's (2014) MAT tool to compare Textbook Conversation with transcriptions of authentic conversation, unless the latter include punctuation marks. In order to bypass this limitation, adjusted scores were calculated in R by adding the *z*-scores (which the MAT helpfully outputs as a tab-separated file) of all the unproblematic features with positive loadings and subtracting those with negative loadings.

The model summarised in Table 6 takes these new adjusted comparable Dimension 1 scores as the outcome variable for the Textbook Conversation and the Spoken BNC2014 corpora. The model's reference levels are Corpus [Textbook] and Register [Conversation]. Hence, Table 6 shows that the estimated Dimension 1 score for naturally-occurring conversation is 16.01 higher than the score estimated for Textbook Conversation (the intercept), i.e., 30.66. The estimated score for the ITTC is –7.5, i.e., 22.15 lower than the intercept.

Table 6. Summary of the model: $\text{Dim1}_{\text{adjusted}}$ ~ 1 + Corpus + Register + Corpus: Register + (Register|Source)

Predictors	**Estimates**	**CI**	***p*-value**
(Intercept)	14.65	12.29 – 17.02	**< 0.001**
Corpus [Reference]	16.01	8.72 – 23.30	**< 0.001**
Register [Fiction]	−10.75	−12.56−−8.93	**< 0.001**
Register [Informative]	−20.72	−22.70−−18.74	**< 0.001**
Corpus [Reference] * Register [Fiction]	−15.48	−22.64−−8.33	**< 0.001**
Corpus [Reference] * Register [Informative]	−22.15	−29.71−−14.59	**< 0.001**
	Random Effects		
σ^2	35.32		
$\tau_{00\ \text{Source}}$	12.34		
$\tau_{11\ \text{Source.RegisterFiction}}$	4.87		
$\tau_{11\ \text{Source.RegisterInformative}}$	7.32		
ρ_{01}	0.40		
	0.12		
ICC	0.35		
N_{Source}	325		
Observations	5,033		
R^2_{marginal} and $R^2_{\text{conditional}}$		0.829 / 0.889	

3.2.1 *Textbook conversation*

As illustrated in Figure 4, the exclusion of the features that rely on punctuation for their operationalisation further widens the gap between naturally-occurring conversation and textbook dialogues (see Table 7).

Table 7. Estimated differences between the estimated mean scores of the model summarised in Table 6

Contrast	Estimate [95% CI]	SE	p-value
Textbook Conversation – Spoken BNC2014	−16.01 [−20.54–−11.48]	3.72	< .0001
Textbook Fiction – Youth Fiction	−0.53 [−3.18–2.12]	1.71	0.759
Textbook Informative – ITTC	6.14 [5.39–6.88]	2.01	0.002

Note. Degrees-of-freedom method: asymptotic.

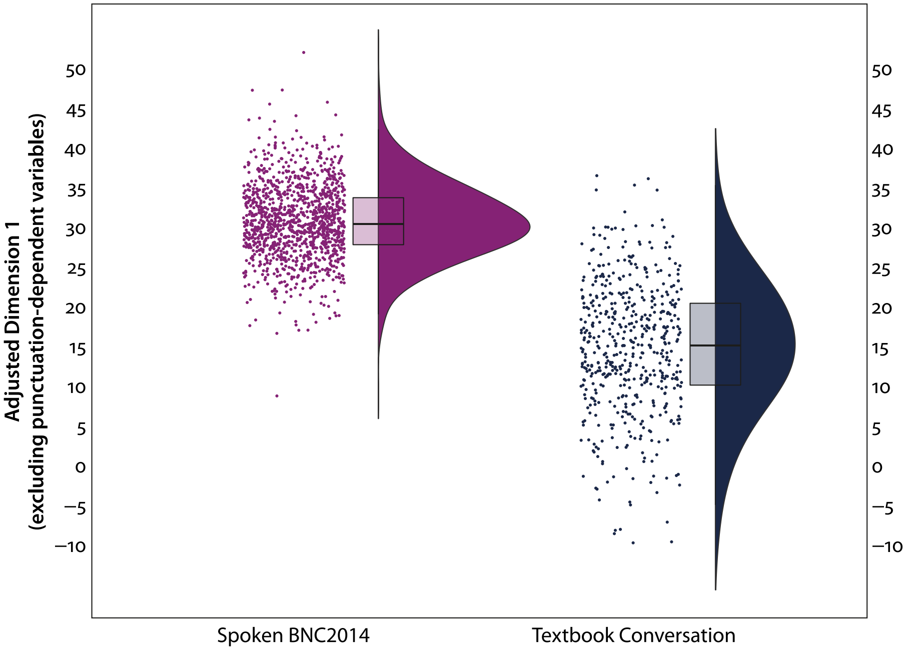

Figure 4. Comparison of modified dimension 1 scores of the spoken BNC2014 and textbook Conversation (Le Foll 2021. Zenodo. Retrieved on 7 May 2021. http://doi.org/10.5281/zenodo.4732343)

Table 8 sheds light on the linguistic features which most contribute to these strikingly low Dimension 1 scores for Textbook Conversation. All the features listed in Table 8 except amplifiers, possibility modals, second person pronouns and indefinite pronouns, contribute to textbook dialogues obtaining lower scores on this dimension.

As compared to the Spoken BNC2014, the greatest underuses in Textbook Conversation are observed in the frequency of hedges (e.g., *sort of*), *that*-deletion

Table 8. Normalised counts for the features loading on Dimension 1

Features on Biber's (1988) Dimension 1	Textbook conversation mean	Textbook conversation SD	Spoken BNC2014 mean	Spoken BNC2014 SD	Comparison mean % difference
Hedges	0.11	0.15	0.26	0.18	−0.81*
That-deletions	0.43	0.35	0.91	0.26	−0.72*
WH-clauses	0.12	0.14	0.22	0.09	−0.59*
Pronoun *it*	1.76	0.80	3.21	0.66	−0.58*
Nouns	24.40	4.34	14.44	1.73	0.51*
Causative subordination	0.14	0.17	0.23	0.16	−0.49*
DO as a main verb	0.34	0.29	0.55	0.21	−0.47*
Emphatics	1.14	0.61	1.71	0.51	−0.40*
Analytic negation	1.55	0.75	2.19	0.48	−0.34*
Contractions	4.24	1.53	5.79	0.86	−0.31*
Amplifiers	0.31	0.30	0.23	0.15	0.30*
Demonstrative pronouns	0.76	0.43	1.02	0.30	−0.29*
Private verbs	2.00	0.86	2.56	0.59	−0.25*
Prepositions	6.58	1.62	5.40	0.74	0.20*
Type/token ratio	0.46	0.05	0.40	0.03	0.14*
Attributive adjectives	4.02	1.33	3.60	0.60	0.11*
Average word length	3.94	0.26	3.65	0.10	0.08*
Present tense	8.66	2.27	9.41	1.22	−0.08*
1st person pronouns	5.98	2.14	5.56	1.19	0.07*
Possibility modals	0.90	0.53	0.85	0.28	0.06
BE as a main verb	3.28	1.07	3.31	0.45	−0.01
2nd person pronouns	3.09	1.41	3.10	0.84	< 0.00
Indefinite pronouns	0.05	0.10	0.05	0.04	< 0.00

Note.
Features with positive loadings in red, with negative loadings in blue. Significance testing was performed with independent two-tailed Wilcoxon tests ($p < .001$ after Holm correction = *)

(marked [THATD] in the example below) and the use of the pronoun *it*. Furthermore, WH-clauses (e.g., *do you know what I mean*), causatives (e.g. *because, cos*), DO as a main verb, emphatics (e.g., *just, really*), analytic negation, contractions, demonstrative pronouns and private verbs (e.g., *THINK, KNOW, BELIEVE, SEE, MEAN*) are also considerably more frequent in naturally-occurring conversation (e.g., Excerpt (1)) than in textbook representations thereof (e.g., Excerpt (2)).

(1) **it**'s the the erm whatever you call **it**
greenfly

yes **it's** er s that **sort of**
greenflies
yes **it's it's** erm something from the greenflies I **think** rather than it'**s not** the tree itself **it's**
the fact that **it's** the aphids erm producing something
do you **think** they drink too drink too much of **this** and **it** makes them ill?
I **think [THATD]** they go they go too too mad on the on the sap and **it just** produces all **this** sticky goo
oh gosh I **didn't know** <BNC2014: SRWD>

Nouns, on the other hand, appear to be considerably overrepresented in pedagogical dialogues (as in Excerpt (2)). These high noun counts correlate positively with high frequencies of prepositional phrases, attributive adjectives, higher type/token ratios and longer words – all of which weigh negatively on this dimension. These features, together with relatively low frequencies of the features with positive loadings discussed above, frequently make textbook dialogues sound like rather unlikely transcripts of real-life conversations, e.g.:

(2) **Man:** Is that your **favourite British dish**?

Woman: Well, I like **roast beef** a lot. But my **real favourite** is waking up **in** the **morning to** the smell **of** a **full English breakfast.** Or **Welsh breakfast**, or the **full Irish breakfast**. Or the **Ulster fry**. Or the **Scottish breakfast. Eggs, bacon** and lots **of** other **tasty things**. It's more or less **the same** wherever you go **in** the **British Isles**. It's just the **name** that changes.

Man: Is that what you have **for breakfast** every day?

Woman: Well, not every day, but sometime **at weekends**. And of course, **at** hotels you can usually have the **full cooked breakfast** if you like. Tastes great **with** a **nice cup of tea**. By the way, did you know that **people in** the **British Isles** drink **around** three **kilos of tea** every **year**.

Man: Three **kilos**?

Woman: Yes, that's over ten **times** as much **tea** as **people** in **Germany** drink. Can you pass the **milk** and **sugar**, please? <TEC: Access G 3>

By contrast, textbook conversations with comparatively high Dimension 1 scores feature more verbal features, such as present tense forms, contractions, negation, first and second person and *it* pronouns, as well as higher normalised counts of discourse markers, amplifiers, hedges, direct WH-questions and stranded prepositions than the majority of textbook dialogues, e.g.:

(3) **Jack:** Lily, there'**s no way I'm** going to **recognise** a model, **it doesn't** matter how famous she **is**. But **I** tell **you what – I bet it isn't** her. What'**s** a famous model going to **be doing** in a shopping mall in **our** town?

Lily: **I think it is** her, **you know**! And she**'s** going into that shop. Come on – **let's** go in too.

Jack: **No way**. Even if **it is** her – **leave** her alone, she **just wants** to do some shopping. And **anyway, what are you** going to **do** – ask her for her autograph or something?

Lily: **I don't know. Maybe I'll just** go up and say hello. **What** do **you reckon**?
<TEC: English in Mind 4>

The model summarised in Table 4 does not lend support to the hypothesis that the dialogues featured in more advanced textbooks have higher, hence more authentic-like, Dimension 1 scores. In fact, some of the Level A textbook dialogues score comparatively high on Dimension 1 owing to their restricted vocabulary, shorter utterances and frequent turns leading to lower type/token and higher verb/noun ratios (e.g., Excerpt (4)). By contrast, many of the texts intended to represent spoken interactions in more advanced textbooks are characterised by a much more nominal style with high informational density, thus featuring high type/token ratios, many prepositions and longer words (e.g., Excerpt (5)).

(4) **Lucy:** **Hey**, watch **out**!

Sam: **Oh**, sorry! **Hey, you're** at Plymstock School.

Lucy: **So**?

Sam: **I'm** at Plymstock school too.

Lucy: **You aren't** from Plymouth!

Sam: **No, I'm not. I'm** new here. **I'm** from London.

Lucy: **OK**.

Sam: **I'm** in Year 7 in class 7EB. What about **you**?

Lucy: **I'm** in 7EB too.

Sam: **Hey**, that's cool. <TEC: Access 1>

(5) **P:** Thanks **for** your **input**, and **good luck**! Now, let's ask someone else. Hello, can I ask you what you think **of** the **American Dream**?

B: Hello! Well, my **ancestors** moved **to** the **United States** long ago, **in** 1846, **during** the **Irish potato famine**. They were **in dire straits** and wanted to escape **poverty**. They had to take care **of** themselves. They worked hard, and slowly they got richer and managed to build a **new life**. They saw the **US as** a **land of freedom** and **opportunity**, where everyone could work hard and be successful. <TEC: Piece of Cake 3^{e}>

3.2.2 *Textbook informative texts*

In contrast to Textbook Conversation, which appears to be considerably less "oral" than the Spoken BNC2014 data, Informative texts in EFL textbooks tend to score higher on Biber's 1988 first dimension than the Informative Texts for Teens Corpus (ITTC) ($\bar{x}$ estimated difference = 6.14, $p = 0.002$). The features which most contribute to this mean difference are first and second person pronouns, *DO* as a main verb, contractions and amplifiers. The prevalence of these features reflects the often informal, "chatty" tone of the Informative texts featured in school EFL textbooks, e.g.:

(6) So how can **you** help yourself to remember things better in the long term? Well, there are several things **you** can **do**. One of them is to make sure **you** pay attention and take in the information properly in the first place. Others are to **do** with the effort **you** make to remember it afterwards. [...] **Don't** wait to revise until exam time – by then **it's** too late!
Although the human brain is **amazingly** powerful, most people only use a tiny amount of its power. The brain is like a muscle. If **you don't** exercise it, it loses its strength and deteriorates. If **you** want to develop and improve **your** mind and make the most of it, **you** need to do regular mental exercises. In spite of all **our** potential brain power, **we** can easily forget 80% of what **we** learn in hours unless **we** make a special attempt to remember it. <TEC: Achievers B2>

The text from which Excerpt (6) was extracted corresponds to the mean Dimension 1 score of the Textbook Informative subcorpus. By way of comparison, Excerpt (7) scores around the mean score of the ITTC. The latter is characterised by more nouns, prepositions, attributive adjectives and longer words.

(7) **Ayanna Pressley** has won her **election**, making her the **first black woman** to represent **Massachusetts in** the **House** of **Representatives, Boston.com** reports. She ran unopposed **in Massachusetts**'s 7th **district**.
Before the **polls** closed **on election day**, she urged **people on Twitter** to vote. "Today, we are powerful. There are only a **few hours** left to get out the **vote**. Go #vote for **progressive candidates** who will fight **for equity** and **justice**," she tweeted. "Vote **for activist leaders** who will work **in** and **with community**. Vote, because this is your **democracy** and your **voice** matters."
<ITTC: teenvogue.com>

In both the textbook and the reference corpus, Informative texts that score lowest on Dimension 1 tend to include bullet point lists and thus feature a high proportion of nominal sentences, as well as many attributive adjectives, a high type/ token ratio and longer words, e.g.:

(8) **Name: Arthur Conan Doyle**
Birth: 2nd **May** 1859 in **Edinburg, Scotland. Death**: 7th **July** 1930 (aged 31) in **England.**
Occupation: Novelist, poet and **doctor**.
Nationality: Scottish
Literary genre: Detective fiction, historical novels. Childhood and **studies**: Very **strict boarding school** from 1868 to 1875. **Medical school**.
Adult life: A **doctor**. Interested in writing **stories**. <TEC: Join the Team 4^{e}>

3.2.3 *Textbook fiction*

In contrast to the two textbook registers discussed above, the difference in mean Dimension 1 scores between Textbook Fiction and the reference Youth Fiction Corpus (YFC) is not significant ($\bar{x}$ estimated difference −0.53, $SE = 1.71$, $p = 0.78$). Fiction usually consists of alternating narration and fictional speech. Thus, novels with a high proportion of dialogues inevitably score high on Biber's first dimension, whilst those with longer descriptive passages score lower. Indeed, additive MDAs of 19th century novels have shown large significant differences on Biber's Dimension 1 between narrative passages, which are more associated with features corresponding to the informational end of the scale, and fictional speech, which is more associated with features characteristic of involvement and interaction (Egbert & Mahlberg 2020: 85; cf. Biber & Finegan 1994). These findings imply that this dimension is not best suited to examine the potentially defining characteristics of Textbook Fiction (cf. Le Foll in preparation, for comparisons on Biber's (1988) other dimensions). That said, the non-significant difference in Dimension 1 scores for Textbook Fiction and the YFC does suggest that they feature similar proportions of narration to fictional speech.

In addition, the model estimates for the Dimension 1 scores of Textbook English registers listed in Table 4 make clear that the small, but significant effect of textbook level on Dimension 1 scores is driven by its interactions with the Fiction register: Textbook Fiction tends towards marginally lower Dimension 1 scores as the proficiency level of the textbooks increases. Though statistically significant, this finding must be approached with caution: not only are the effect sizes very small, Figure 2 also shows some missing data in the Fiction register. Nonetheless, beginner textbooks tend to feature more dialogue-heavy fictional writing, leading to a greater use of first and second personal pronouns, verbal contractions, negation and demonstrative pronouns (see Excerpt (9)), than more advanced teaching materials (Excerpt (10)) or youth fiction novels (Excerpt (11)), which, on average, both feature many more prepositions, nouns and attributive adjectives. Moreover, beginner textbooks that have not yet introduced past tense forms rely on present-tense narration, which also contributes to higher Dimension 1 scores (e.g., Excerpt (9)) in contrast to the narrative texts of more advanced textbooks (e.g., Excerpt (10)) and the majority of novels sampled in the YFC, which largely feature past-tense narration (e.g., Excerpt (11)).

(9) 'Very funny,' Lucy **says.** '**I** think **this** is just a silly trick. **I don't** believe a word.'
'A silly trick?' the Time Lord **laughs.** 'Ha, ha, ha, just look at **this, you** silly girl!'
The lights in the Planetarium **flicker** again, and on the huge screen, Lucy, Sandy and Asim **can** see pictures of Greenwich – and it already **looks** very different. There **aren't** many old people any more, and children **are looking** down at clothes that **are** too big for them.
Then they **hear** the scary voice again.
'So, children. The future of the human race lies in **your** hands. See **this** hourglass here? When the sand is through, **your** time will be up. [...]'
<TEC: New Green Line 1>

(10) The **mountains** stretched **into** infinity: **exquisite shades of** green, grey and brown **against** a **deep azure, cloudless sky. Along** the wall, here and there, were **small groups of tourists basking in** the **wonder of** their **surroundings.** But **the strangest sight of** all was a **table** and **four plastic chairs beneath** a **huge red parasol**, and a **man** selling **bottled water** and **cans of chilled drinks from** an **icebox.** <TEC: Achievers B2>

(11) The **skin around** his **eyes was** darkening **to** a **thin glaze.** The **children looked away.** They **knew** the **signs.** 'You don't wear them, do you? No, you don't. She doesn't like them, so you're not allowed to wear them.' 'I wear mine,' **Natalie offered.** <YFC: Anne Fine 1987 Madame Doubtfire>

4. Conclusion and recommendations

This study has demonstrated that Biber's 1988 model of General English can successfully be used as a baseline to explore register variation within secondary school EFL textbooks. The fact that register explains 63% of the variance observed in Dimension 1 scores across six major registers of the TEC confirms the need to account for register in textbook language studies. Mixed effect models were used to explore additional factors that could potentially explain some of the variation observed, notably the style of the authors, editors and/or publishers of specific textbook series, as well as the proficiency levels of the textbooks. Compared to register, these were shown to only play a marginal role in mediating textbook language variation (RQ1). The only significant interaction between textbook register and proficiency level was observed in the Fiction register, which is easily explained by the fact that the past tense is not featured in beginner level textbooks, meaning that these rely on present tense narration instead – thus leading to higher Dimension 1 scores than narrative texts from more advanced textbooks.

In answer to RQ2, the most striking differences between the textbook and reference registers were observed in the Conversation register: on Biber's (1988) Dimension 1, Textbook Conversation scores considerably lower than the Spoken BNC2014. This is largely due to the much more nominal style of textbook dialogues, which also tend to feature longer speaker turns, longer words and higher type/token ratios. Thus, textbook dialogues appear to primarily function as reinforcers of the vocabulary students are expected to learn, rather than as models of realistic spontaneous spoken interactions. Excluding the features that rely on punctuation for their operationalisations, the most underrepresented Dimension 1 features in Textbook Conversation are hedges, *that*-deletions, WH-clauses and *it* pronouns.

On average, the Informative texts of school EFL textbooks were found to be more interactional and spoken-like than the texts featured on informative websites targeted at English-speaking teenagers; they tend to feature considerably more present tense verbs, contractions, and first and second personal pronouns.

Textbook Fiction scores closest to its corresponding reference corpus of Youth Fiction novels. Tellingly, the fictional, narrative texts featured in secondary school EFL textbooks are the most likely to be extracts or adaptations of works that were not originally penned for pedagogical purposes, i.e., extracts of original novels or short stories of the kind included in the YFC. In addition, some publishers (e.g., Klett, personal communication) contract experienced fiction authors to write such texts. However, the analysis also made clear that further explorations of this register ought to be made on other dimensions of Biber's (1988) model: Dimension 2, 'Narrative vs. Non-narrative Concerns', in particular, may yield more salient results (Le Foll in preparation).

From a methodological point of view, a number of issues in applying Biber's (1988) Dimension 1, 'Involved vs. Informational', to the registers of secondary school EFL textbooks and comparable target language registers have been highlighted. Solutions to overcome issues related to the non-independence of texts from the same textbook series (Section 2.3.4), text length (Section 2.1.1) and the punctuation-dependent operationalisation of some of the features (Section 3.2.1) were discussed and implemented. The latter two issues have made clear that, in spite of the availability and ease of use of the MAT, Biber's (1988) model of spoken and written English cannot be applied to secondary school EFL textbook registers "out of the box". First, we have seen that it requires careful considerations (and coding skills) to extract individual texts from the textbooks, calculate the length of each text and collate shorter texts in order to reach the 400-word threshold needed to calculate the token/type ratio that loads onto Biber's Dimension 1.

Second, the fact that Biber's first dimension includes five linguistic features with operationalisations that rely on punctuation is also clearly a limitation for comparing the dialogues of textbooks to naturally occurring conversation. Either one chooses a reference corpus of spoken English that includes punctuation (with all the transcription reliability issues that this implies), or, as was chosen here, the offending variables must be manually removed and the adjusted dimension scores must be calculated outside of the MAT. Finally, there is a risk that the results of this multi-feature analysis of Textbook Conversation were skewed because Biber's noun variable aggregates common and proper nouns and many textbook dialogues include the name of the person speaking at the start of every turn (see Excerpts (12) and (13)). This will undoubtably have inflated the relative frequencies of nouns in these textbook dialogues. Thus, for a more precise investigation of register variation in secondary school EFL textbooks, future projects include conducting a full MDA with more appropriate linguistic features (e.g., excluding some very rare features and adding more salient ones) and feature operationalisations (e.g., removing the need for punctuation and separating proper nouns from the total noun count).

Nonetheless, the relative simplicity of conducting additive MDAs and the availability of the MAT (Nini 2014), which largely automates the process (see Section 2.3.1), bears the advantage of making the methodology accessible beyond academia. Given its potential for the evaluation of textbook language, it is hoped that the method may be of interest to textbook authors, editors, publishers and representatives of educational authorities. Though it is by no means claimed that it could or should be used as a unique solution, Biber's (1988) framework has been shown to provide a valuable synthesis of the relative frequencies of many relevant linguistic features that can help to distinguish particularly unnatural-sounding texts from more natural-sounding ones. Since it captures functional variation along an involved/oral vs. informational/literate continuum, Dimension 1 lends itself particularly well to the examination of representations of spoken language. Thus, a high score on Biber's Dimension 1, such as that scored by the dialogue quoted in Excerpt (12) (Dim1=38.19 as calculated by the MAT; items that contributed to this high score are in bold), points to a pedagogical text that is likely to paint a more authentic picture of natural conversation than one with a much lower score, e.g., Excerpt (13).

(12) **Amy:** **Hi**, Nick.

Nick: **Hi**, Amy. Amy, **is** this **your** backpack on the floor?

Amy: That's right.

Nick: **Well, could you perhaps** put **it** somewhere else? **It's kind of** in the way.

Amy: **No, it's not. It's** where **I** always **leave it.**

Nick: **Yes, I know you** always **leave it** there. **And it's** always in the way. This **is** a pretty small place, Amy. **So perhaps just** for once **you could** put **your** backpack somewhere where **it isn't** in the way, **hmm**?

Amy: **You don't own** this place, Nick. **So don't try** and **tell me** what to do. **I** came in early to get some things done. **I** put **my** backpack on the floor. **You deal** with **it**! <TEC: English in Mind 4>

Thus, textbook dialogues that score particularly low could be flagged as potentially worth re-examining or revising. For example, Excerpt (13) scored –6.10 on Dimension 1, which is the result of its considerably higher type/token ratio and longer average word length than most natural conversations, as well as the fact that it features many complex nominal phrases, which lead to high relative frequencies of prepositions and attributive adjectives – all of which contribute to negative Dimension 1 scores.

(13) **Journalist:** This is **Sally Gordon** here in **Leicester Square, London**. I'm right **in** the **middle** of **sports fans**. Excuse me, Sir. Who is your **favourite sports hero**?

Dwayne: Definitely, **Chris Hoy**, the **British track cyclist** – won two **gold medals**. He represents **strength** and **courage**, he never gave up.

Journalist: What about you? Who is the **best representative of** your **country**?

Donna: **Kobe Bryant** for sure. I'm American and we are very patriotic when it comes **to sport**. He has shown the **world** we remain the **dominant leaders in basketball**, no doubt. And **Michael Phelps** of course.

Journalist: Why?

Donna: Why? He has just won four **golds** and two **silver medals** and he is a **record holder**. The **dream** came true. Incredible. That's why he is nicknamed "the **Baltimore Bullet**". He **symbolises determination, generosity, hope… great values**. You see, he's a **role model**! He will be remembered forever. <TEC: New Mission 2^{e}>

Crucially, whilst it can be said that textbook dialogues such as Excerpt (12) expose learners to interactional, genuinely conversation-like language that they are likely to encounter outside the classroom, texts such as Excerpt (13) cannot be considered realistic models for EFL learners to acquire spontaneous spoken language comprehension and/or production skills. Such texts, can, of course, be argued to serve other pedagogical purposes, e.g., the high lexical diversity of Excerpt (13) may be specifically aimed at increasing learners' passive vocabulary range. However, where the aim is to present learners with spontaneous, spoken English, low Dimension 1 scores can act as a helpful warning sign that revision ought to be considered. Inversely, when textbook Informative texts score particularly high on

Dimension 1, this is a sign that they are unlikely to be of use as models for students to acquire the skills necessary to write their own informative texts or read for information independently outside the classroom; hence, here too, corpus-informed revisions should be considered.

For example, Excerpt (13) could be improved by consulting a corpus of spoken language, such as the Spoken BNC2014 (Love et al. 2017), and adding some of the frequent lexico-grammatical features of spontaneous, interactional speech. The resulting, revised version is likely to include higher relative frequencies of the features that contribute to high scores on Biber's (1988) Dimension 1 (see Table 8). For example, the proposed revised dialogue printed below as Example (14) features more private verbs (e.g. THINK, FORGET), *that*-deletions, contractions, present tense verbs, first and second person pronouns, analytic negations (*didn't he*), emphatics (*really*), causative subordination (*because),* discourse participles (*well, you know*), hedges (*kind of*), sentence relatives, WH-questions, possibility modals, non-phrasal coordination and final prepositions than the original textbook dialogue in (13). As Excerpt (14) shows, such additions will also naturally lead to revised dialogues with lower type/token ratios, shorter average word lengths and, in particular, lower noun/verb ratios, which all contribute to high Dimension 1 scores, too.

(14) **Journalist: I'm** Sally Gordon, reporting from Leicester Square in London **and** the place **is** full of sports fans. **Let's** see **who we can** talk **to. Excuse me**, Sir. **Can I** ask **you who's your** sports hero**?**

Dwayne: Erm, for **me, it'd definitely** have to be Chris Hoy, **you know**, the British track cyclist **who** won two gold medals. **I think [THATD]** he **really stands** for strength **and and I really admire** his courage **because, well**, he **just** never **gives up**.

Journalist: Sure. And erm **what** about **you**? **Who would you** say **is your** national hero**?**

Donna: Erm, **actually, I'm** American **so** Kobe Bryant, **for sure. We're kind of very** patriotic, **especially** when **it** comes to sports, if **you know what I mean**.

Journalist: And would you say [THATD] basketball is **your** sport **then**?

Donna: Yeah I am into basketball **and that and, you know, I think [THATD]** he**'s really** shown the world **we're** still the best at **it**!

Journalist: Mm.

Donna: Oh **and I** should**n't forget** Michael Phelps, **of course**.

Journalist: Uhu. **What makes you say** that**?**

Donna: You kidding? I mean, he's **just** won **like** four gold medals and two silver.

> **Journalist: Right**, he did, did**n't** he?
>
> **Donna: And** he's a record holder! **I guess** what **I'm saying is** the the dream came true.
>
> **Journalist: Right**.
>
> **Donna: Yeah**, he's **just** incredible. **I mean** that's why **we** call him "the Baltimore Bullet" **because** he's **all** about determination, generosity, hope… he**'s all** about **all** these **really** great values. **You see**, he's he's a role model! **And we'll** never **forget** him, **that's for sure**.

The present results indicate that textbook dialogues with high Dimension 1 scores are more likely to be appropriate models for EFL learners to acquire the skills necessary to navigate natural conversation. In particular, this includes the competent use of a variety of fluency-enhancing strategies to overcome planning phases and manage turn-taking in spontaneous conversation. Previous learner corpus research has shown that EFL learners significantly underuse discourse and vagueness markers as compared to native speakers and tend to rely more on filled and unfilled pauses and/or a very limited set of such markers, instead (e.g., Müller 2005; Götz 2013; Gilquin 2016; Dumont 2018). It has already been suggested that this oft-observed underuse of discourse markers in learner speech "might stem from the fact that an explicit teaching of discourse markers as a fluency-enhancing strategy has not been systematically integrated into EFL textbooks" (Wolk, Götz, & Jäschke 2020: 4; cf. Römer 2005; Gilquin 2016). The results outlined in Section 3.2.1, in which the dialogues of 43 secondary school EFL textbooks were compared to the transcriptions of naturally occurring native-speaker conversations along Biber's (1988) Dimension 1, lends support to this hypothesis.

To conclude, this paper has demonstrated that textbook authors, editors, publishers and educational authorities may want to consider applying additive MDA as part of a wide range of methods for textbook evaluation and revision purposes. However, given the limitations highlighted above, further research is needed to arrive at a comprehensive model of the linguistic specificities of the different registers of secondary school EFL textbooks as compared to situationally similar target language registers. Nonetheless, this preliminary study based on the first dimension of Biber's (1988) model has confirmed that Textbook English cannot be adequately modelled without considering register-based linguistic variation. It has also shown that robust statistical methods must be employed to additionally account for any linguistic variation inherent to the proficiency levels of the textbooks, as well as the idiosyncrasies of individual textbook series (and thereby of their authors, editors and/or publishers).

Acknowledgements

I am grateful to Dirk Siepmann, the two anonymous reviewers, and Bethany Grey, as one of the editors of *Register Studies*, for their highly valuable and detailed feedback on earlier versions of this paper. The usual disclaimers apply.

References

Al-Surmi, M. (2012). Authenticity and TV Shows: A Multidimensional Analysis Perspective. *TESOL Quarterly*, 671–694. https://doi.org/10.1002/tesq.33

Barbieri, F., & Eckhardt, S. E. (2007). Applying Corpus-Based Findings to Form-Focused Instruction: The Case of Reported Speech. *Language Teaching Research*, 11(3), 319–346. https://doi.org/10.1177/1362168807077563

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berber Sardinha, T., & Biber, D. (Eds.). (2014). *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.60

Berber Sardinha, T., & Veirano Pinto, M. (2017). American television and off-screen registers: A corpus-based comparison. *Corpora*, 12(1), 85–114. https://doi.org/10.3366/cor.2017.0110

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019). *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic. https://doi.org/10.5040/9781350023857

Berber Sardinha, T., Veirano Pinto, M., Mayer, C., Zuppardi, M. C., & Kauffmann, C. H. (2019). Adding Registers to a Previous Multi-Dimensional Analysis. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues* (pp. 165–188). New York, NY: Bloomsbury. https://doi.org/10.5040/9781350023857.0017

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

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37. https://doi.org/10.1515/cllt-2012-0002

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly*, 36(1), 9. https://doi.org/10.2307/3588359

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., … Urzua, A. (2004). *Representing Language Use in the University: Analysis of the TOEFFL 2000 Spoken and Written Academic Language Corpus*. Princeton, NJ: Educational Testing Service.

Biber, D., & Finegan, E. (1994). Multi-dimensional analyses of authors' styles: Some case studies from the eighteenth century. In D. Ross & D. Brink (Eds.), *Research in humanities computing* (Vol. 3, pp. 3–17). Oxford: Oxford University Press.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781316410899

Chujo, K. (2004). Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List. *Language and Computers*, 51(1), 231–249.

Clarke, I., & Grieve, D.J. (2017). Dimensions of Abusive Language on Twitter. *Proceedings of the First Workshop on Abusive Language Online*, 1–10. Retrieved from https://www.aclweb.org/anthology/W17-3001.pdf. https://doi.org/10.18653/v1/W17-3001

Conrad, S. (2004). Corpus variety: Corpus linguistics, language variation, and language teaching. In J. McH. Sinclair (Ed.), *Studies in Corpus Linguistics* (Vol. 12, pp. 67–85). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.12.08con

Conrad, S. (2013). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 94–107). (Original work published 2001)

Conrad, S., & Biber, D. (Eds.). (2013). *Variation in English: Multi-Dimensional Studies*. New York: Routledge. (Original work published 2001) https://doi.org/10.18820/9781920689094

Conrad, S.M. (1996). Academic discourse in two disciplines: Professional writing and student development in biology and history (PhD dissertation). Northern Arizona University.

Crossley, S.A., Kyle, K., & Römer, U. (2019). Examining Lexical and Cohesion Differences in Discipline-Specific Writing Using Multi-Dimensional Analysis. In T.B. Sardinha & M.V. Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues* (pp. 189–216). Bloomsbury Academic. https://doi.org/10.5040/9781350023857.0019

Crossley, S., Allen, L.K., & McNamara, D. (2014). A Multi-Dimensional analysis of essay writing: What linguistic features tell us about situational parameters and the effects of language functions on judgments of quality. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Studies in Corpus Linguistics* (Vol. 60, pp. 197–238). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.60.07cro

Dumont, A. (2018). Fluency and disfluency: A corpus study of non-native and native speaker (dis)fluency profiles (PhD dissertation). Université catholique de Louvain, Louvain. Retrieved from http://hdl.handle.net/2078.1/198393

Egbert, J., & Mahlberg, M. (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

Egbert, J., & Staples, S. (2019). Doing Multi-Dimensional Analysis in SPSS, SAS, and R. In T.B. Sardinha & M.V. Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues* (pp. 125–144). Bloomsbury Academic. https://doi.org/10.5040/9781350023857.0015

Ellis, N., & Collins, L. (2009). Input and Second Language Acquisition: The Roles of Frequency, Form, and Function Introduction to the Special Issue. *The Modern Language Journal*, 93(3), 329–335. https://doi.org/10.1111/j.1540-4781.2009.00893.x

European Council (Ed.). (2004). *Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)* (6. pr). Stuttgart: Klett.

Forchini, P. (2012). *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Peter Lang. https://doi.org/10.3726/978-3-0351-0325-0

Friginal, E., & Hardy, J.A. (2014). Conducting Multi-Dimensional Analysis Using SPSS. In T.B. Sardinha & D. Biber (Eds.), *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber* (pp. 297–316). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.60.10fri

Gabrielatos, C. (2013). If-conditionals in ICLE and the BNC: A success story for teaching or learning? In S. Granger, F. Meunier, & G. Gilquin (Eds.), *Twenty Years of Learner Corpus Research: Looking back, moving ahead* (Presses Universitaires de Louvain, pp. 155–166).

Gilmore, A. (2004). A Comparison of Textbook and Authentic Interactions. *ELT Journal*, 58(4), 363–374. https://doi.org/10.1093/elt/58.4.363

Gilquin, G. (2016). Discourse markers in L2 English: From classroom to naturalistic input. In O. Timofeeva, A.-C. Gardner, A. Honkapohja, & S. Chevalier (Eds.), *Studies in Language Companion Series* (Vol. 177, pp. 213–249). Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.177.09gil

Götz, S. (2013). *Fluency in Native And Nonnative English Speech*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.53

Gouverneur, C. (2008). The Phraseological Patterns of High-frequency Verbs in Advanced English for General Purposes: A Corpus-driven Approach to EFL Textbook Analysis. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 223–243). Amsterdam: John Benjamins. https://doi.org/10.1075/z.138.17gou

Gray, B. (2015). *Linguistic Variation in Research Articles: When discipline tells only part of the story*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.71

Gray, B., & Egbert, J. (2019). Editorial: Register and register variation. *Register Studies*, 1(1), 1–9. https://doi.org/10.1075/rs.00001.edi

Gries, S. Th. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125. https://doi.org/10.3366/cor.2015.0068

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, 13(3), 239–256. https://doi.org/10.1016/0889-4906(94)90004-3

Johansson, S., Leech, G.N., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computer*. Department of English, University of Oslo.

Le Foll, E. (2020, October). *Issues in Compiling and Exploiting Textbook Corpora*. Presented at the Japanese Association for English Corpus Studies 2020, Tokyo. https://doi.org/10.13140/RG.2.2.32006.60487

Le Foll, Elen. (2021). Bibliographic metadata of the Textbook English Corpus (TEC) (Version v. 1.1) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.4922819

Le Foll, E. (in preparation). *Textbook English: A Corpus-Based Analysis of the Language of EFL textbooks used in Secondary Schools in France, Germany and Spain*.

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means* [Manual]. Retrieved from https://CRAN.R-project.org/package=emmeans

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*, 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Love, R., Hawtin, A., & Hardie, A. (2018, September). *The British National Corpus 2014: User manual and reference guide*. Retrieved from http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf

Lüdecke, D. (2020). *sjPlot: Data visualization for statistics in social science* [Manual]. Retrieved from https://CRAN.R-project.org/package=sjPlot

Meunier, F., & Gouverneur, C. (2009). New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (Ed.), *Studies in Corpus Linguistics* (Vol. 33, pp. 179–201). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.33.16meu

Miller, D. (2011). ESL Reading Textbooks vs. University Textbooks: Are We Giving Our Students the Input They May Need? *Journal of English for Academic Purposes*, 10(1), 32–46. https://doi.org/10.1016/j.jeap.2010.12.002

Mindt, D. (1987). *Sprache, Grammatik, Unterrichtsgrammatik: Futurischer Zeitbezug im Englischen I* (first). Frankfurt am Main: Diesterweg.

Mindt, D. (1995). Schulgrammatik vs. Grammatik der englischen Sprache. In *Perspektiven des Grammatikunterrichts* (Vol. 404).

Muhammad, S. (2020). A corpus based comparison of variation in online registers of Pakistani English using MD analysis (PhD dissertation). University of Münster.

Müller, S. (2005). *Discourse Markers in Native and Non-native English Discourse*. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.138

Mundry, R., & Sommer, C. (2007). Discriminant function analysis with nonindependent data: Consequences and an alternative. *Animal Behaviour*, 74(4), 965–976. https://doi.org/10.1016/j.anbehav.2006.12.028

Nini, A. (2014). Multidimensional Analysis Tagger (MAT) (Version 1.3). Retrieved from http://sites.google.com/site/multidimensionaltagger

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues* (pp. 67–96). New York: Bloomsbury. https://doi.org/10.5040/9781350023857.0012

Quaglio, P. (2009). *Television Dialogue: The sitcom Friends vs. natural conversation*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.36

Rautionaho, P., & Deshors, S. C. (2018). Progressive or not progressive?: Modeling the constructional choices of EFL and ESL writers. *International Journal of Learner Corpus Research*, 4(2), 225–252. https://doi.org/10.1075/ijlcr.16019.rau

Römer, U. (2004). A Corpus-Driven Approach to Modal Auxiliaries and Their Didactics. In J. McH. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 185–199). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.12.14rom

Römer, U. (2005). *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to English Progressive Forms, Functions, Contexts, and Didactics*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.18

Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 121–134. https://doi.org/10.1515/zaa-2006-0204

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation* (Vol. 56). Studentlitteratur.

Usó-Juan, E., & Martínez-Flor, A. (2010). The teaching of speech acts in second and foreign language instructional contexts. In *Pragmatics across Languages and Cultures* (pp. 423–442). Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110214444.3.423

Veirano Pinto, M. (2019). Using Discriminate Function Analysis in Multi-Dimensional Analysis. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues* (pp. 217–230). Bloomsbury Academic. https://doi.org/10.5040/9781350023857.0020

Vellenga, H. (2004). Learning Pragmatics from ESL & EFL Textbooks: How Likely? *TESL-EJ Teaching English as a Second or Foreign Language*, 8(2), n. p.

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. New York: Routledge. https://doi.org/10.4324/9781315165547

Wolk, C., Götz, S., & Jäschke, K. (2020). Possibilities and Drawbacks of Using an Online Application for Semi-automatic Corpus Analysis to Investigate Discourse Markers and Alternative Fluency Variables. *Corpus Pragmatics*. https://doi.org/10.1007/s41701-019-00072-x

Zuppardo, M.C. (2013). A linguagem da aviação: Um estudo de manuais aeronáuticos baseado na Análise Multidimensional. *[Aviation Language: A Study of Aeronautical Handbooks Based on Multi-Dimensional Analysis]*, 11, 6–25.

Appendix. Examples of the six textbook registers examined

Conversation

Nice of you to let us come to your barbie, Mike.

No worries. Great you're here. – Hey Cam! Come and meet a couple of new mates. They're staying at the hostel. Hey, how're you doing?

Hi. I'm Tanya. Nice to meet you.

You're a Kiwi, right? From your accent? And you're, let me guess … American? No, I'm from Israel. Moshe.

OK, cool. You know Mike can trace his ancestors right back to the first British convicts in Australia?

Come on, Cam. Not that joke again! The story is: My ancestor Bill was walking down the road when a man bumped into him. The man was being chased by a police officer because he'd stolen a gold necklace from a jewellery shop. When the man ran off, the police officer stopped Bill and found the gold necklace in his pocket. Then Bill was arrested and given a sentence of 20 years in Australia.

OK, he was just a victim. But many of the convicts were real criminals. […]

<TEC: New Green Line 5>

Informative text

English is an official language in over seventy-five countries in the world.

More than two billion people speak English. Fifty-four English-speaking countries are members of the Commonwealth of Nations, an association of independent countries. Queen Elizabeth II is head of the Commonwealth.

31% (percent) of the world's population live in the Commonwealth.

Six people out of ten in the United Kingdom have a relative in a Commonwealth country.

<TEC: Hi There 5^{e}>

Instructional text

Plymouth, my hometown

a. In the film, the girl shows us her hometown. Watch the film. What did she show us? Choose A or B.
b. Watch the film again.

What other things and places can you see in the film?
Make a list. <TEC: Access 1>

Fiction

With my backpack in my hands, I stepped off the train onto the crowded platform. It was 7:30 in the evening. People were hurrying home. A mother and her two young children were sitting on a bench. The mother was talking to the boy, but he wasn't looking at her. The girl was singing quietly and playing with a toy. Around them, travellers were shouting greetings, waving goodbye, carrying heavy bags or running to catch trains. A very tall man was standing completely still near the exit. Why was he wearing summer clothes in this weather? And why was he looking straight at me? <TEC: Solutions Pre-intermediate>

Personal correspondence

Ally McKoene > WestHigh Bros
December 1 near University Heights, IA via mobile
Your best feature is definitely your kindness and I'm sure everyone else agrees! You have tons of kindness in your heart and your compliments can light up anyone's face. You guys are some of the kindest people I've met and I'm so glad that you guys do what you do. Your compliments can make anyone's day :) keep it up!
Like – Comment
nh [OCR error: Facebook-style thumbs up symbol] West High Bros likes this.
<TEC: New Mission 2e>

Poetry

School friends
Welcome to my school!
Welcome to my school!
Come in, and be cool!
Good morning, you can all sit down!
I'm Mister Parker
Yes, I'm your teacher
Good morning, Good morning Sir!
Can you repeat? I think I don't know I don't understand… Can I open the window?
Can you come to the board? Can you write the date? Yes sir! Yes sir!
It's a piece of cake!
Welcome to my school!
Let's all be cheerful!
Take your pens and write this down I'm Mister Parker!
Listen to your teacher!
Yes Sir! No Sir! Thank you Sir! <TEC: Join the Team 6e>

Address for correspondence

Elen Le Foll
Institute of English and American Studies
Osnabrück University
Neuer Graben 40
49069 Osnabrück
Germany

elen@fridu.net

elefoll@uos.de
https://orcid.org/0000-0002-5839-8010

Publication history

Date received: 18 September 2020
Date accepted: 6 July 2021
Published online: 23 November 2021