Textbook English: A Multi-Dimensional Approach

Online Supplements

Elen Le Foll

2024-03-05

Table of contents

Preface

This Quarto book is **work in progress**. It will eventually contain the online supplements to:

> Le Foll, Elen. to appear. *Textbook English: A Multi-Dimensional Approach* [Studies in Corpus Linguistics]. Amsterdam: John Benjamins.

The book is based on my PhD thesis, which is accessible in Open Access:

> Le Foll, Elen. 2022. *Textbook English: A Corpus-Based Analysis of the Language of EFL textbooks used in Secondary Schools in France, Germany and Spain.* Osnabrück, Germany: Osnabrück University. PhD thesis. https://doi.org/10.48693/278.

1 Introduction

Asked "Where is Brian?", French nationals of a certain generation will immediately reply: "Brian is in the kitchen". Those with a particularly good memory may follow up with: "Where is Jenny, the sister of Brian?" – and, to those in the know, the correct answer is: "Jenny is in the bathroom".[1] There is hardly any need for an in-depth linguistic analysis to conclude that this interaction is highly unlikely to have ever taken place in a real English-speaking family home. To most teachers and learners, it will be evident that it is the result of a none too inspired attempt to model WH-question forms in a textbook dialogue aimed at beginner learners of English as a Foreign Language (EFL). Together with dull gap-fill exercises and photos of out-of-date technology, for many adults, the very mention of the word textbook evokes vivid memories of such artificially sounding, contrived and sometimes even nonsensical dialogues.

This raises the question of the status and nature of textbook language as a specific 'variety' of language, which is at the heart of the present study. It focuses on contemporary EFL textbooks in use in European secondary schools. Situated at the interface between linguistics and foreign language teaching, this study examines the linguistic content of these textbooks and seeks empirical answers to the questions: What kind of English do school EFL textbooks portray? And how far removed is this variety of English from the kind of English that learners can be expected to encounter outside the EFL classroom?

1.1 Research objectives and methodological approach

The above questions are critical because, as many adults' lingering memories of school foreign language lessons testify (see also, e.g., Freudenstein 2002: 55), textbooks play an absolutely central role in classroom-based foreign language learning. In the following, we will see that the dominance of textbooks in EFL school contexts persists to this day. According to Thornbury (2012 in a response to Chong 2012: n.p.), they "(more often [than] not) instantiate the curriculum, provide the texts, and - to a large extent - guide the methodology". In lower secondary EFL instructional contexts, in particular, textbooks constitute a major vector of foreign language input. Yet, numerous studies have shown that "considerable mismatches

[1] Dialogue from *Speak English 6e série verte* (Benhamou & Dominique 1977: 167). It was made popular by stand-up comedian Gad Elmaleh. More information on the context of this textbook dialogue can be found here. An extract of the comedy sketch by Gad Elmaleh that popularised the dialogue can be viewed here with English subtitles: https://youtu.be/11jG7lkwDwU?t=50.

between naturally occurring English and the English that is put forward as a model in pedagogical descriptions" (Römer 2006: 125-26) exist. These mismatches have been observed and sometimes extensively described in textbooks' representations of numerous language features ranging from the use of individual words and phraseological patterns (e.g., Conrad 2004 on the preposition though; Gouverneur 2008 on the high-frequency verbs make and take), to tenses and aspects (e.g., Barbieri & Eckhardt 2007 on reported speech; Römer 2005 on the progressive). More rarely, textbook language studies have also ventured into the study of spoken grammar (e.g., Gilmore 2004) and pragmatics (e.g., Hyland 1994 on hedging in ESP/EAP textbooks).

However, as we will see in Chapter 2, previous EFL textbook studies have tended to focus on one or at most a handful of individual linguistic features. Taken together, they provide valuable insights into "the kind of synthetic English" (Römer 2004b: 185) that pupils are exposed to via their textbooks; yet, what is missing is a more comprehensive, broader understanding of what constitutes 'Textbook English' from a linguistic point of view. Although corpus-based[2] textbook analysis can be traced back to the pioneering work of Dieter Mindt in the 1980s, the language of secondary school EFL textbooks (as opposed to that of general adult EFL or English for Specific Purposes [ESP] coursebooks) remains an understudied area.

The present study therefore sets out to describe the linguistic content of secondary school EFL textbooks and to survey the similarities and most striking differences between 'Textbook English' and 'naturally occurring English' as used outside the EFL classroom, with respect to a wide range of lexico-grammatical features.

To this end, a corpus of nine series of secondary school EFL textbooks (43 textbook volumes) used at lower secondary level in France, Germany, and Spain was compiled (see 4.3.1). In addition, three reference corpora are used as baselines for comparisons between the language input EFL learners are confronted with via their school textbooks and the kind of naturally occurring English that they can be expected to encounter, engage with, and produce themselves on leaving school. Two of these have been built specifically for this project with the aim of representing comparable 'authentic' (for a discussion of this controversial term in ELT, see 2.2) and age-appropriate learner target language.

A bottom-up, corpus-based approach is adopted (e.g., Mindt 1992, 1995a; Biber & Quirk 2012; Biber & Gray 2015; Ronald Carter & McCarthy 2006a). A broad range of linguistic features are considered: ranging from tenses and aspects to negation and discourse markers. We will pay particular attention to the lexico-grammatical aspects of Textbook English that substantially diverge from the target learner language reference corpora and examine these with direct comparisons of textbook excerpts with comparable texts from the reference data.

[2]Here the adjectives 'corpus-based' and 'corpus-driven' are used synonymously (see, e.g., Meunier & Reppen 2015: 499 for further information as to how these terms are sometimes distinguished).

1.2 Outline of the book

The following chapter outlines the background to and motivation behind the present study. Chapter 3 then provides a literature review of state-of-the-art research on the language of school EFL textbooks. It is divided in two parts. Part 1 is a methodological review in which the various methods employed so far to analyse, describe, and evaluate Textbook English are explained and illustrated with selected studies. Part 2 summarises the results of existing studies on various aspects of Textbook English, including lexical, grammatical and pragmatic aspects. Based on the methodological limitations and the gaps identified in the existing literature, Chapter 4 elaborates the specific research questions addressed in the present study. These research questions informed the decision-making processes involved in the compilation of the Textbook English Corpus (TEC) and the selection/compilation of three reference corpora designed to represent learners' target language. These processes and their motivations are explained in the remaining sections of Chapter 4.

Chapter 5 describes the multivariable statistical methods applied to describe the linguistic nature of Textbook English on multiple dimensions of linguistic variation. It begins by explaining the well-established multi-feature/dimensional analysis (MDA) method pioneered by Biber (1988, 1995; see also Berber Sardinha & Veirano Pinto 2014, 2019), before outlining the reasoning for the modified MDA framework applied in the present study. Chapter 6 presents the results of an MDA model of Textbook English which highlights the sources of linguistic variation within EFL textbooks across several dimensions of intra-textbook linguistic variation. Chapter 7 presents the results of a second MDA model that shows how Textbook English is both, in some respects, similar to and, in others, different from the kind of English that EFL learners are likely to encounter outside the classroom.

Chapter 8 explains how the two models contribute to a new understanding of the linguistic characteristics of Textbook English. This, in turn, has implications for teachers, textbook authors, editors, publishers, and policy-makers. These implications are discussed in Chapter 9. It first considers the potential impact of the substantial gaps between Textbook English and the target reference corpora before making suggestions as to how teachers, textbook authors, and editors may want to improve or supplement unnatural-sounding pedagogical texts using corpora and corpus tools. Chapter 10 focuses on the study's methodological strengths and limitations. It explains how the modified MDA framework presented and applied in this study may be of interest to corpus linguists working on a broad range of research questions. Chapter 11 concludes with a synthesis of the most important take-aways from the study. It also points to promising future research avenues.

2 Literature review

This is a tabular overview of the Textbook English studies that I examined as part of my literature review. It presents the results of a non-exhaustive survey of Textbook English studies published over the past four decades, summarising some of the key information on each study, including its main language focus, methodological approach, information on the textbooks investigated, and, if applicable, on any reference corpora used. Empty cells represent fields that are either not applicable to this particular study or for which no information could be found. Intended as a dynamic resource, this interactive, searchable, and filterable table currently lists over 80 studies on the language content of English L2 textbooks, thereby demonstrating the breadth of Textbook English studies published as of early 2022.

3 Corpus data

3.1 Textbook English Corpus (TEC)

A detailed tabular overview of the composition of the Textbook English Corpus (TEC) together with the full bibliographic metadata is available at doi.org/10.5281/zenodo.4922819.

Note that, for copyright reasons, the corpus itself cannot be published. If you are interested in using the corpus for non-commercial research purposes and/or in a potential research collaboration, please get in touch with me via e-mail.

3.2 Reference corpora

3.2.1 Spoken BNC2014

The original corpus files of the Spoken British National Corpus (BNC) 2014 (Love et al. 2017; Love et al. 2019) can be downloaded for free for research purposes from: http://corpora.lancs.ac.uk/bnc2014/signup.php. I used the untagged XML version.

The R script used to pre-process the untagged XML files as explained in Section 4.3.2.2 of the book can be found here: https://github.com/elenlefoll/TextbookEnglish/blob/main/3_Data/BNCspoken_nomark-up_JackJill.R

3.2.2 Informative Texts for Teens Corpus (Info Teens)

For copyright reasons, the corpus itself cannot be made available. Details of its composition can be found in Section 4.3.2.5 of the book. If you are interested in using this corpus for non-commercial research purposes and/or in a potential research collaboration, please get in touch with me via e-mail.

3.2.3 Youth Fiction corpus

For copyright reasons, the corpus itself cannot be made available. The corresponding metadata can be found here: https://github.com/elenlefoll/TextbookEnglish/blob/main/3_Data/3_Youth_Fiction_Index.csv. If you are interested in using this corpus for non-commercial research purposes and/or in a potential research collaboration, please get in touch with me via e-mail.

4 Open Science statement

Another important insight from the methodological part of the literature review (see Section 3.1 in book publication) is that, to the author's best knowledge, no Textbook English study published so far has included (as an appendix or supplementary materials) the data and code necessary to reproduce or replicate the published results. As a result, it is very difficult to evaluate the reliability or robustness of the results reported (see also Le Foll 2024).

Though the terms are sometimes used interchangeably and different (at times incompatible) definitions abound, in computational sciences, 'reproducibility' usually refers to the ability to obtain the same results as an original study using the researchers' data and code, whilst 'replicability' refers to obtaining compatible results with the same method but different data (Association for Computing Machinery 2020; see also Berez-Kroeker et al. 2018).

A major barrier to the reproducibility of (corpus) linguistic research is that it is often not possible for copyright or, when participants are involved, data protection reasons to make linguistic data available to the wider public. However, both research practice and the impact of our research can already be greatly improved if we publish our code or, when using GUI software, methods sections detailed enough to be able to successfully replicate the full procedures. This step can enable others to conduct detailed reviews of our methodologies and conceptual replications of our results on different data.

Aside from data protection and copyright regulations, there are, of course, many reasons why researchers may be reluctant to share their data and code (Berez-Kroeker et al. 2018; McManus 2021). It is not within the scope of this monograph to discuss these; however, it is clear that, in many ways, such transparency makes us vulnerable. At the end of the day: to err is human. Yet, the risks involved in committing to Open Science practices are particularly tangible for researchers working on individual projects, like myself, who have had no formal training in data management or programming and have therefore had to learn "on the job". Nonetheless, I am convinced that the advantages outweigh the risks. Striving for transparency helps both the researchers themselves and others reviewing the work to spot and address problems. As a result, the research community can build on both the mishaps and successes of previous research, thus improving the efficiency of research processes and ultimately contributing to advancing scientific progress.

It is with this in mind that I have decided, whenever possible, to publish all the raw data and code necessary to reproduce the results reported in the present monograph following the FAIR principles (i.e., ensuring that research data are Findable, Accessible, Interoperable and Reusable, see Wilkinson et al. 2016). For copyright reasons, the corpora themselves and

annotated corpus data in the form of concordance lines cannot be made available. However, the outcome of both manual and automatic annotation processes is published in tabular formats in the Online Appendix. These tables allow for the reproduction of all the analyses reported on in the following chapters using the reproducible data analysis scripts also published in the Online Supplements and in the associated Open Science Framework (OSF) repository.

In all chapters of this monograph, full transparency is strived for by reporting on how each sample size was determined and on which grounds data points were excluded, manipulated and/or transformed. Most of these operations were conducted in the open-source programming language and environment R (R Core Team 2022). Most of the data processing and analysis scripts therefore consist of R markdown documents. These were rendered to HTML pages (viewable in the Online Supplements) thus allowing researchers to review the procedures followed without necessarily installing all the required packages and running the code themselves. These scripts also feature additional analyses, tables and plots that were made as part of this study but which, for reasons of space, were not reported on in detail here. Whenever additional software or open-source code from other researchers were used, links to these are also provided in the Online Supplements (in addition to the corresponding references in the bibliography).

5 A Model of Intra-Textbook Linguistic Variation: Data Preparation

This script documents the steps taken to pre-process the Textbook English Corpus (TEC) data that were entered in the multi-dimensional model of intra-textbook linguistic variation (Chapter 6).

5.1 Packages required

The following packages must be installed and loaded to process the data.

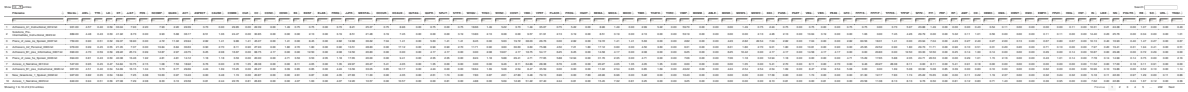
```
#renv::restore() # Restore the project's dependencies from the lockfile to
 ↪  ensure that same package versions are used as in the original study

library(caret) # For its confusion matrix function
library(DT) # To display interactive HTML tables
library(here) # For dynamic file paths
library(knitr) # Loaded to display the tables using the kable() function
library(patchwork) # Needed to put together Fig. 1
library(PerformanceAnalytics) # For the correlation plot
library(psych) # For various useful, stats function
library(tidyverse) # For data wrangling
```

5.2 Data import from MFTE output

The raw data used in this script is a tab-separated file that corresponds to the tabular output of mixed normalised frequencies as generated by the MFTE Perl v. 3.1 (Le Foll 2021a).

```
# Read in Textbook Corpus data
TxBcounts <- read.delim(here("MFTE_data", "Outputs",
 ↪  "TxB900MDA_3.1_normed_complex_counts.tsv"), header = TRUE,
 ↪  stringsAsFactors = TRUE)
TxBcounts <- TxBcounts |>
```

```
  filter(Filename!=".DS_Store") |>
  droplevels()
#str(TxBcounts) # Check sanity of data
#nrow(TxBcounts) # Should be 2014 files
datatable(TxBcounts,
  filter = "top",
) |>
  formatRound(2:ncol(TxBcounts), digits=2)
```



Metadata was added on the basis of the filenames.

```
# Adding a textbook proficiency level
TxBLevels <- read.delim(here("metadata", "TxB900MDA_ProficiencyLevels.csv"),
 ↪  sep = ",")
TxBcounts <- full_join(TxBcounts, TxBLevels, by = "Filename") |>
  mutate(Level = as.factor(Level)) |>
  mutate(Filename = as.factor(Filename))

# Check distribution and that there are no NAs
summary(TxBcounts$Level) |>
  kable(col.names = c("Textbook Level", "# of texts"))
```

Textbook Level	# of texts
A	292
B	407
C	506
D	478
E	331

```
# Check matching on random sample
# TxBcounts |>
#   select(Filename, Level) |>
#   sample_n(20)
```

```
# Adding a register variable from the file names
TxBcounts$Register <- as.factor(stringr::str_extract(TxBcounts$Filename,
 ↪  "Spoken|Narrative|Other|Personal|Informative|Instructional|Poetry")) #
 ↪  Add a variable for Textbook Register
summary(TxBcounts$Register) |>
  kable(col.names = c("Textbook Register", "# of texts"))
```

Textbook Register	# of texts
Informative	364
Instructional	647
Narrative	285
Personal	88
Poetry	37
Spoken	593

```
TxBcounts$Register <- car::recode(TxBcounts$Register, "'Narrative' =
 ↪  'Fiction'; 'Spoken' = 'Conversation'")
#colnames(TxBcounts) # Check all the variables make sense

# Adding a textbook series variable from the file names
TxBcounts$Filename <- stringr::str_replace(TxBcounts$Filename,
 ↪  "English_In_Mind|English_in_Mind", "EIM")
TxBcounts$Filename <- stringr::str_replace(TxBcounts$Filename,
 ↪  "New_GreenLine", "NGL") # Otherwise the regex for GreenLine will override
 ↪  New_GreenLine
TxBcounts$Filename <- stringr::str_replace(TxBcounts$Filename,
 ↪  "Piece_of_cake", "POC") # Shorten label for ease of plotting
TxBcounts$Series <- as.factor(stringr::str_extract(TxBcounts$Filename,
 ↪  "Access|Achievers|EIM|GreenLine|HT|NB|NM|POC|JTT|NGL|Solutions")) #
 ↪  Extract textbook series from (ammended) filenames
summary(TxBcounts$Series)  |>
  kable(col.names = c("Textbook Name", "# of texts"))
```

Textbook Name	# of texts
Access	315
Achievers	240
EIM	180

Textbook Name	# of texts
GreenLine	209
HT	115
JTT	129
NB	44
NGL	298
NM	59
POC	98
Solutions	327

```
# Including the French textbooks for the first year of Lycée to their
 ↪  corresponding publisher series from collège
TxBcounts$Series <-car::recode(TxBcounts$Series, "c('NB', 'JTT') = 'JTT';
 ↪  c('NM', 'HT') = 'HT'") # # Recode final volumes of French series (see
 ↪  Section 4.3.1.1 on textbook selection for details)
summary(TxBcounts$Series) |>
  kable(col.names = c("Textbook Series", "# of texts"))
```

Textbook Series	# of texts
Access	315
Achievers	240
EIM	180
GreenLine	209
HT	174
JTT	173
NGL	298
POC	98
Solutions	327

```
# Adding a textbook country of use variable from the series variable
TxBcounts$Country <- TxBcounts$Series
TxBcounts$Country <- car::recode(TxBcounts$Series, "c('Access', 'GreenLine',
 ↪  'NGL') = 'Germany'; c('Achievers', 'EIM', 'Solutions') = 'Spain'; c('HT',
 ↪  'NB', 'NM', 'POC', 'JTT') = 'France'")
summary(TxBcounts$Country) |>
  kable(col.names = c("Country of Use", "# of texts"))
```

Country of Use	# of texts
France	445
Germany	822
Spain	747

```
# Re-order variables
#colnames(TxBcounts)
TxBcounts <- select(TxBcounts, order(names(TxBcounts))) %>%
  select(Filename, Country, Series, Level, Register, Words, everything())
#colnames(TxBcounts)
```

5.2.1 Corpus size

This table provides some summary statistics about the number of words included in the TEC texts originally tagged for this study.

```
TxBcounts  |>
  group_by(Register) |>
  summarise(totaltexts = n(), totalwords = sum(Words), mean =
   ↪  as.integer(mean(Words)), sd = as.integer(sd(Words)), TTRmean =
   ↪  mean(TTR)) |>
  kable(digits = 2, format.args = list(big.mark = ","))
```

Register	totaltexts	totalwords	mean	sd	TTRmean
Conversation	593	505,147	851	301	0.44
Fiction	285	241,512	847	208	0.47
Informative	364	304,695	837	177	0.51
Instructional	647	585,049	904	94	0.42
Personal	88	69,570	790	177	0.48
Poetry	37	26,445	714	192	0.44

```
#TxBcounts <- saveRDS(TxBcounts, here("processed_data", "TxBcounts.rds"))
```

5.3 Data preparation for PCA

Poetry texts were removed for this analysis as there were too few compared to the other register categories.

```
summary(TxBcounts$Register) |>
  kable(col.names = c("Register", "# texts"))
```

Register	# texts
Conversation	593
Fiction	285
Informative	364
Instructional	647
Personal	88
Poetry	37

This led to the following distribution of texts across the five textbook English registers examined in the model of intra-textbook linguistic variation:

```
TxBcounts <- TxBcounts |>
  filter(Register!="Poetry") |>
  droplevels()

summary(TxBcounts$Register) |>
  kable(col.names = c("Register", "# texts"))
```

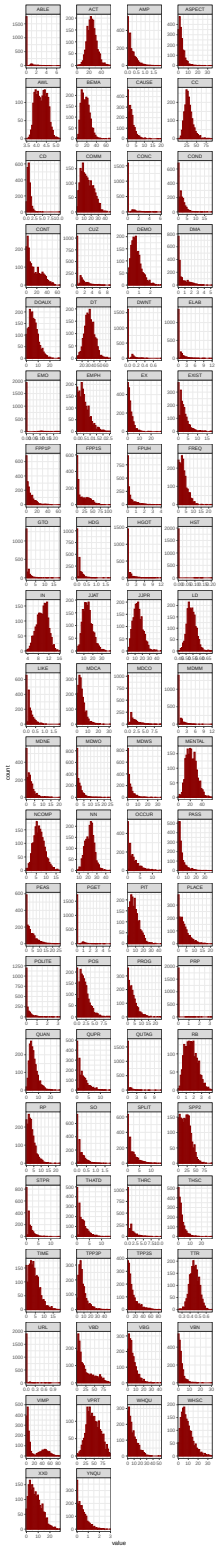
Register	# texts
Conversation	593
Fiction	285
Informative	364
Instructional	647
Personal	88

5.3.1 Feature distributions

The distributions of each linguistic features were examined by means of visualisation. As shown below, before transformation, many of the features displayed highly skewed distributions.

```
TxBcounts |>
  select(-Words) |>
  keep(is.numeric) |>
  tidyr::gather() |>  # This function from tidyr converts a selection of
 ↪  variables into two variables: a key and a value. The key contains the
 ↪  names of the original variable and the value the data. This means we can
 ↪  then use the facet_wrap function from ggplot2
```

```
ggplot(aes(value)) +
  theme_bw() +
  facet_wrap(~ key, scales = "free", ncol = 4) +
  scale_x_continuous(expand=c(0,0)) +
  geom_histogram(bins = 30, colour= "darkred", fill = "darkred", alpha =
   ↪  0.5)
```


ABLE
ACT
AMP
ASPECT
AWL
BEMA
CAUSE
CC
CD
COMM
CONC
COND
CONT
CUZ
DEMO
DMA
DOMJX
DT
DWNT
ELAB
EMO
EMPH
EX
EXIST
FPFUP
FPFUS
FPUH
FREQ
GTO
HDO
HGOT
HST
IN
JJR
JJPR
LO
LIKE
MDCA
MDCO
MDMM
MDNE
MDWO
MDWS
MENTAL
NCOMP
NN
OCCUR
PASS
PEAS
PGET
PIT
PLACE
POLITE
POS
PROG
PRP
QUAN
QUPR
QUTAG
RB
RP
SO
SPLIT
SPP2
STPR
THATD
THRC
THSC
TIME
TPFSP
TPFSS
TTR
URL
VBD
VBG
VBN
VIMP
VPRT
WHQU
WHSC
XX0
YNQU
count
value

```
#ggsave(here("plots", "TEC-HistogramPlotsAllVariablesTEC-only.svg"), width =
 ↪  20, height = 45)
```

5.3.2 Feature removal

A number of features were removed from the dataset as they are not linguistically interpretable. In the case of the TEC, this included the variable CD because numbers spellt out as digits were removed from the textbooks before these were tagged with the MFTE. In addition, the variables LIKE and SO because these are "bin" features included in the output of the MFTE to ensure that the counts for these polysemous words do not inflate other categories due to mistags (Le Foll 2021b).

Whenever linguistically meaningful, very low-frequency features were merged. Finally, features absent from more than third of texts were also excluded. For the analysis intra-textbook register variation, the following linguistic features were excluded from the analysis due to low dispersion:

```
# Removal of meaningless features:
TxBcounts <- TxBcounts |>
  select(-c(CD, LIKE, SO))

# Function to compute percentage of texts with occurrences meeting a
 ↪  condition
compute_percentage <- function(data, condition, threshold) {
  numeric_data <- Filter(is.numeric, data)
  percentage <- round(colSums(condition[, sapply(numeric_data,
 ↪  is.numeric)])/nrow(data) * 100, 2)
  percentage <- as.data.frame(percentage)
  colnames(percentage) <- "Percentage"
  percentage <- percentage |>
    filter(!is.na(Percentage)) |>
    rownames_to_column() |>
    arrange(Percentage)
  if (!missing(threshold)) {
    percentage <- percentage |>
      filter(Percentage > threshold)
  }
  return(percentage)
}

# Calculate percentage of texts with 0 occurrences of each feature
```

```r
zero_features <- compute_percentage(TxBcounts, TxBcounts == 0, 66.6)
#print(zero_features)

# Combine low frequency features into meaningful groups whenever this makes
 ↪  linguistic sense
TxBcounts <- TxBcounts |>
  mutate(JJPR = ABLE + JJPR, ABLE = NULL) |>
  mutate(PASS = PGET + PASS, PGET = NULL)

# Re-calculate percentage of texts with 0 occurrences of each feature
zero_features2 <- compute_percentage(TxBcounts, TxBcounts == 0, 66.6)
print(zero_features2)
```

```
   rowname Percentage
1      GTO      67.07
2     ELAB      69.30
3     MDMM      70.81
4     HGOT      73.75
5     CONC      80.48
6     DWNT      81.44
7    QUTAG      85.99
8      URL      96.51
9      EMO      97.82
10     PRP      98.33
11     HST      99.44
```

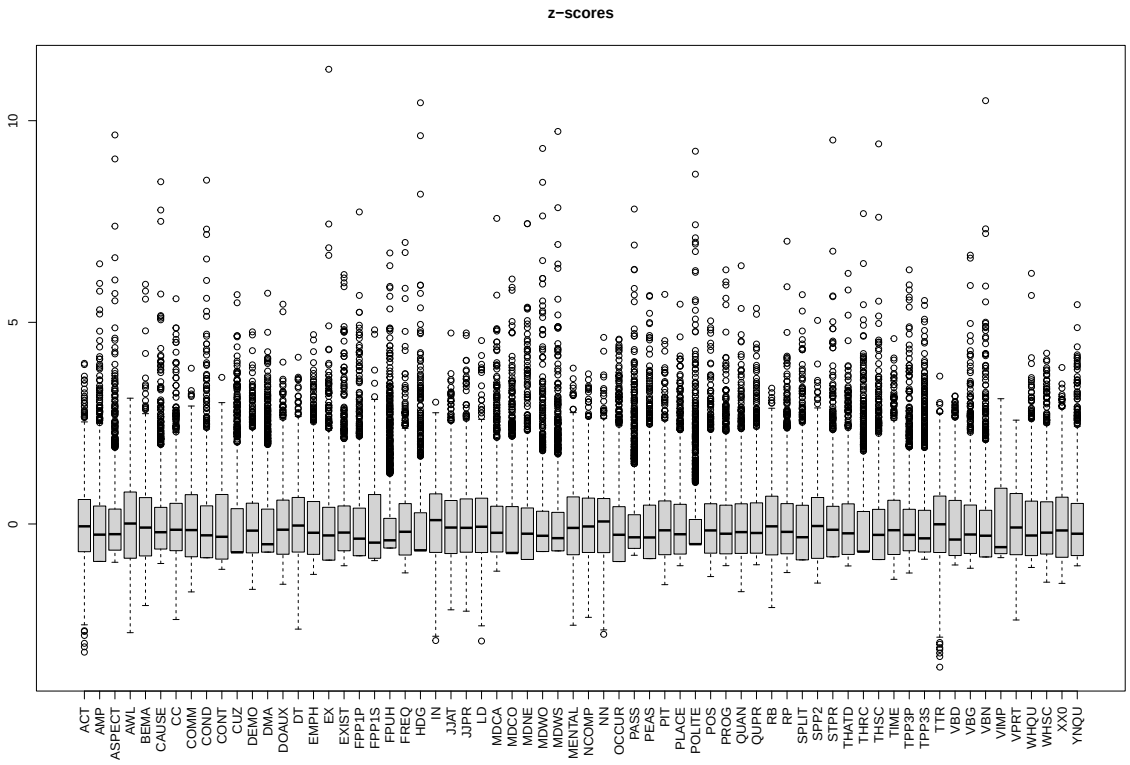
```r
# Drop variables with low document frequency
TxBcounts <- select(TxBcounts, -one_of(zero_features2$rowname))
#ncol(TxBcounts)-8 # Number of linguistic features remaining

# List of features
#colnames(TxBcounts)
```

These feature removal operations resulted in a feature set of 64 linguistic variables.

5.3.3 Identifying potential outlier texts

All normalised frequencies were normalised to identify any potential outlier texts.

```
# First scale the normalised counts (z-standardisation) to be able to compare
 ↪  the various features
TxBcounts |>
  select(-Words) |>
  keep(is.numeric) |>
  scale() ->
  TxBzcounts

boxplot(TxBzcounts, las = 3, main = "z-scores") # Slow to open!
```




```
# If necessary, remove any outliers at this stage.
TxBdata <- cbind(TxBcounts[,1:6], as.data.frame(TxBzcounts))

outliers <- TxBdata |>
  select(-c(Words, LD, TTR)) |>
  filter(if_any(where(is.numeric), ~ .x > 8)) |>
  select(Filename)
```

The following outlier texts were identified and excluded in subsequent analyses.

```
outliers
```

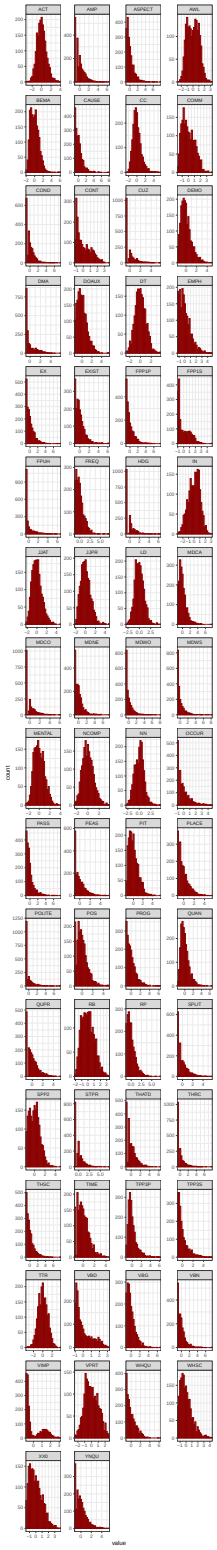
```
                                          Filename
1                            POC_4e_Spoken_0007.txt
2              Solutions_Elementary_Personal_0001.txt
3                        NGL_5_Instructional_0018.txt
4                            Access_1_Spoken_0011.txt
5                               EIM_1_Spoken_0012.txt
6                               NGL_4_Spoken_0011.txt
7       Solutions_Intermediate_Plus_Personal_0001.txt
8            Solutions_Elementary_ELF_Spoken_0021.txt
9                           NB_2_Informative_0009.txt
10        Solutions_Intermediate_Plus_Spoken_0022.txt
11      Solutions_Intermediate_Instructional_0025.txt
12 Solutions_Pre-Intermediate_Instructional_0024.txt
13                            POC_4e_Spoken_0010.txt
14             Solutions_Intermediate_Spoken_0019.txt
15                           Access_1_Spoken_0019.txt
16     Solutions_Pre-Intermediate_ELF_Spoken_0005.txt
```

```
TxBcounts <- TxBcounts |>
  filter(!Filename %in% outliers$Filename)

TxBcounts |>
  select(-Words) |>
  keep(is.numeric) |>
  scale() ->
  TxBzcounts
```

This resulted in 1,961 TEC texts being included in the model of intra-textbook linguistic variation with the following normalised feature distributions.

```
TxBzcounts |>
  as.data.frame() |>
  gather() |>  # This function from tidyr converts a selection of variables
   ↪  into two variables: a key and a value. The key contains the names of
   ↪  the original variable and the value the data. This means we can then
   ↪  use the facet_wrap function from ggplot2
  ggplot(aes(value)) +
    theme_bw() +
```

```
    facet_wrap(~ key, scales = "free", ncol = 4) +
    scale_x_continuous(expand=c(0,0)) +
    geom_histogram(bins = 30, colour= "darkred", fill = "darkred", alpha =
     ↪  0.5)
```


ACT
AMP
ASPECT
AWL
BENIA
CAUSE
CC
COMM
COND
CONT
CUZ
DENO
DMA
DOAUX
DT
EMPH
EX
EXIST
FPFUP
FPP1S
FPUH
FREQ
HDG
IN
INF
JJPR
LD
MDCA
MDCO
MDNE
MDWO
MDWS
MENTAL
NCOMP
NN
OCCUR
PASS
PEAS
PIT
PLACE
POLITE
POS
PROG
QUAN
QUPR
RB
RP
SPLIT
SPIV
STPR
THATD
THRC
THSC
TIME
TPFSP
TPP2S
TTR
VBD
VBG
VBN
VIMP
VPRT
WHQU
WHSC
XX0
YNQU
count
value

```
#ggsave(here("plots", "TEC-zscores-HistogramsAllVariablesTEC-only.svg"),
 ↪  width = 20, height = 45)
```

5.3.4 Signed log transformation

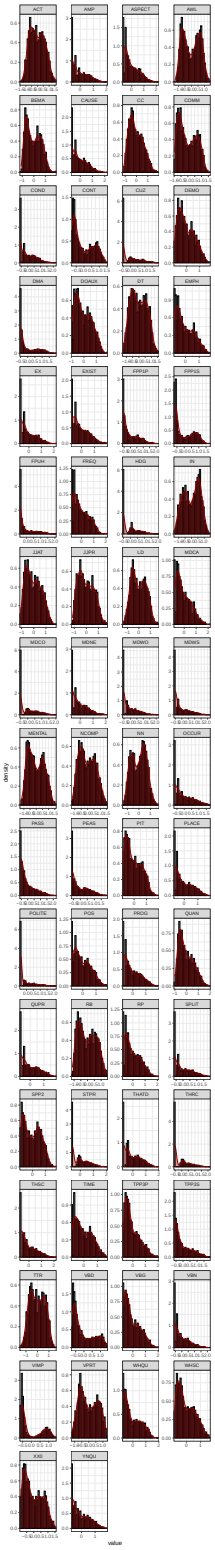
A signed logarithmic transformation was applied to (further) deskew the feature distributions (Diwersy, Evert, and Neumann 2014; Neumann and Evert 2021).

The signed log transformation function was inspired by the SignedLog function proposed in https://cran.r-project.org/web/packages/DataVisualizations/DataVisualizations.pdf

```
# All features are signed log-transformed (note that this is also what
 ↪  Neumann & Evert 2021 propose)
signed.log <- function(x) {
  sign(x) * log(abs(x) + 1)
  }

TxBzlogcounts <- signed.log(TxBzcounts) # Standardise first, then signed log
 ↪  transform

#saveRDS(TxBzlogcounts, here("processed_data", "TxBzlogcounts.rds")) # Last
 ↪  saved 16 Feb 2024
```

The new feature distributions are visualised below.

```
TxBzlogcounts |>
  as.data.frame() |>
  gather() |>  # This function from tidyr converts a selection of variables
   ↪  into two variables: a key and a value. The key contains the names of
   ↪  the original variable and the value the data. This means we can then
   ↪  use the facet_wrap function from ggplot2
  ggplot(aes(value, after_stat(density))) +
  theme_bw() +
  facet_wrap(~ key, scales = "free", ncol = 4) +
  scale_x_continuous(expand=c(0,0)) +
  scale_y_continuous(limits = c(0,NA)) +
  geom_histogram(bins = 30, colour= "black", fill = "grey") +
  geom_density(colour = "darkred", weight = 2, fill="darkred", alpha = .4)
```


ACT
AMP
ASPECT
AWL
BRKA
CAUSE
CC
COMM
COND
CONT
CUZ
DEMO
DMA
DOAUX
DT
EMPH
EX
EXIST
FPFSP
FPFSS
FPUM
FREQ
HDG
IN
JJAT
JJPR
LID
MODA
MOCO
MONE
MONO
MOWS
NEATAL
NCOMP
NU
OCCUR
PAGS
PEAS
PIT
PLACE
POLITE
POS
PROG
QUAN
QUPR
RB
RP
SPLIT
SPP2
STPR
THATD
THEC
THSC
TIME
TPFSP
TPFSS
TTR
VBD
VBG
VBN
VIMP
VINIT
WHQU
WHSC
XXD
YNQU
density
value

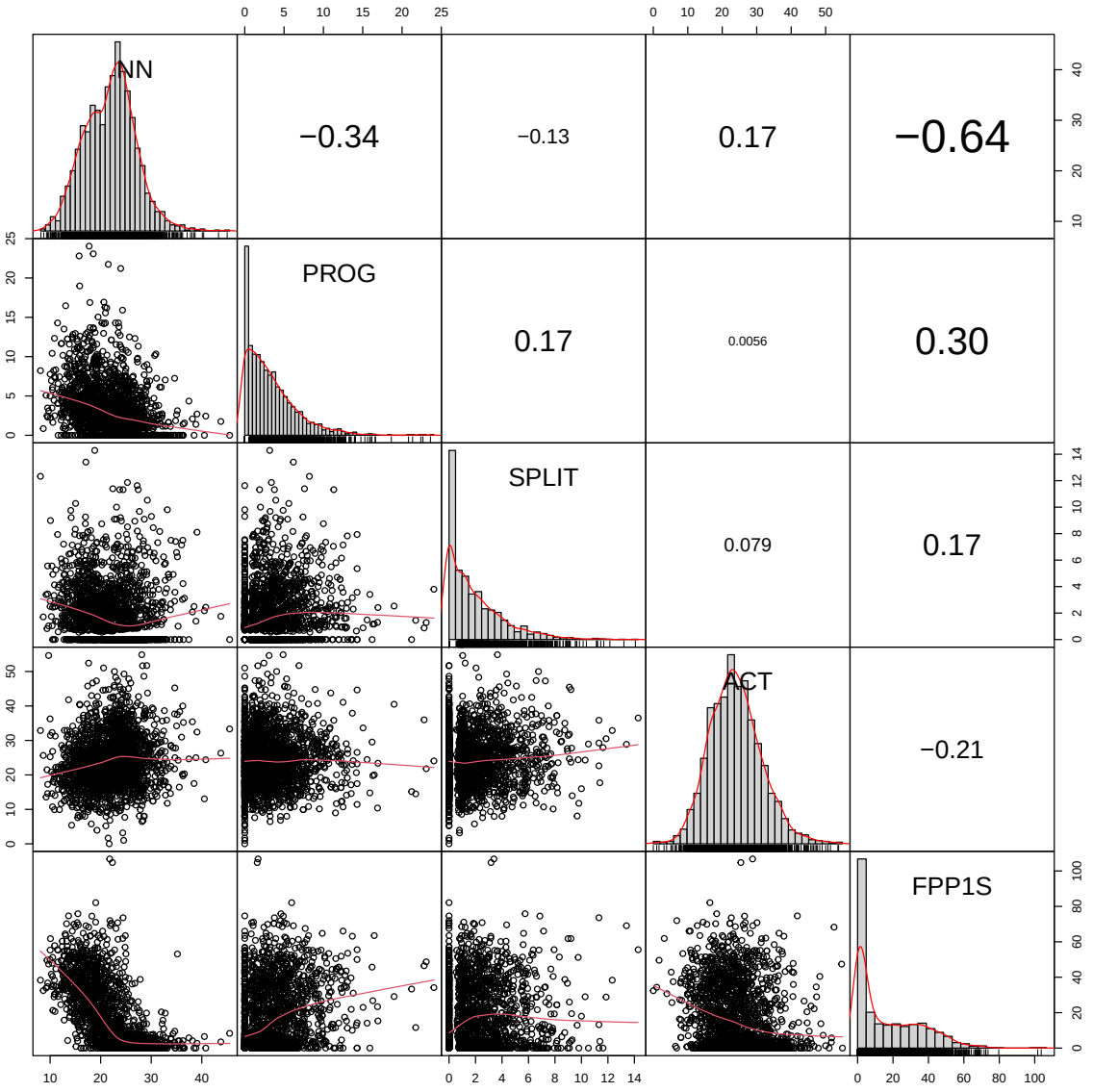
```
#ggsave(here("plots", "DensityPlotsAllVariablesSignedLog-TEC-only.svg"),
 ↪  width = 15, height = 49)
```

The following correlation plots serve to illustrate the effect of the variable transformations performed in the above chunks.

Example feature distributions before transformations:

```
# This is a slightly amended version of the
 ↪  PerformanceAnalytics::chart.Correlation() function. It simply removes the
 ↪  significance stars that are meaningless with this many data points (see
 ↪  commented out lines below)

chart.Correlation.nostars <- function (R, histogram = TRUE, method =
 ↪  c("pearson", "kendall", "spearman"), ...) {
  x = checkData(R, method = "matrix")
  if (missing(method))
    method = method[1]
  panel.cor <- function(x, y, digits = 2, prefix = "", use =
 ↪  "pairwise.complete.obs", method = "pearson", cex.cor, ...) {
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- cor(x, y, use = use, method = method)
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste(prefix, txt, sep = "")
    if (missing(cex.cor))
      cex <- 0.8/strwidth(txt)
    test <- cor.test(as.numeric(x), as.numeric(y), method = method)
    # Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
    #                  cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1), symbols =
     ↪  c("***",
    #
     ↪  "**", "*", ".", " "))
    text(0.5, 0.5, txt, cex = cex * (abs(r) + 0.3)/1.3)
    # text(0.8, 0.8, Signif, cex = cex, col = 2)
  }
  f <- function(t) {
    dnorm(t, mean = mean(x), sd = sd.xts(x))
  }
  dotargs <- list(...)
  dotargs$method <- NULL
```

```
  rm(method)
  hist.panel = function(x, ... = NULL) {
    par(new = TRUE)
    hist(x, col = "light gray", probability = TRUE,
         axes = FALSE, main = "", breaks = "FD")
    lines(density(x, na.rm = TRUE), col = "red", lwd = 1)
    rug(x)
  }
  if (histogram)
    pairs(x, gap = 0, lower.panel = panel.smooth, upper.panel = panel.cor,
          diag.panel = hist.panel)
  else pairs(x, gap = 0, lower.panel = panel.smooth, upper.panel = panel.cor)
}

# Example plot without any variable transformation
example1 <- TxBcounts |>
  select(NN,PROG,SPLIT,ACT,FPP1S)

#png(here("plots", "CorrChart-TEC-examples-normedcounts.png"), width = 20,
 ↪  height = 20, units = "cm", res = 300)
chart.Correlation.nostars(example1, histogram=TRUE, pch=19)
```

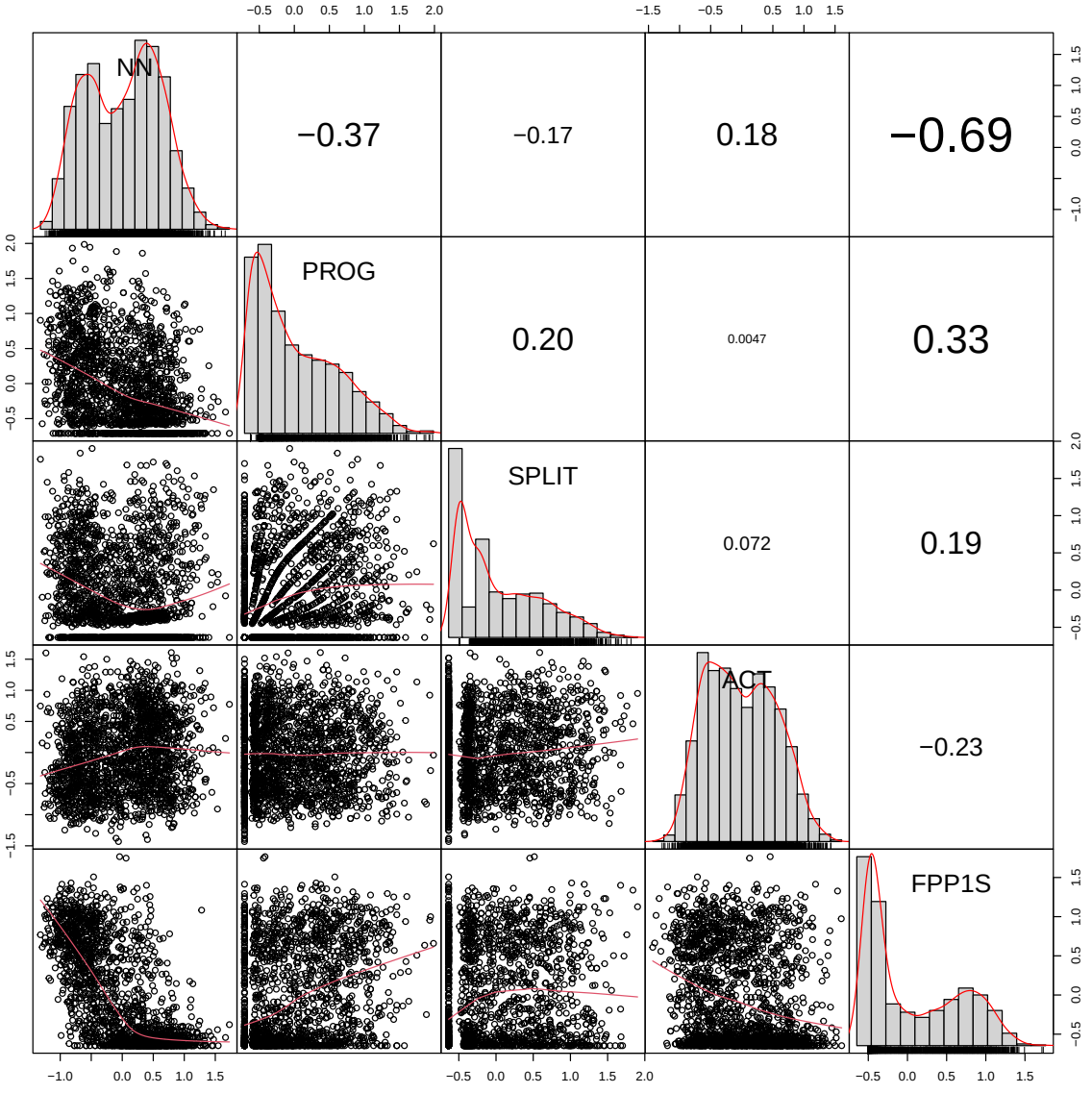



```
#dev.off()
```

Example feature distributions after transformations:

```
# Example plot with transformed variables
example2 <- TxBzlogcounts |>
  as.data.frame() |>
  select(NN,PROG,SPLIT,ACT,FPP1S)
```

```
#png(here("plots", "CorrChart-TEC-examples-zsignedlogcounts.png"), width =
 ↪  20, height = 20, units = "cm", res = 300)
chart.Correlation.nostars(example2, histogram=TRUE, pch=19)
```

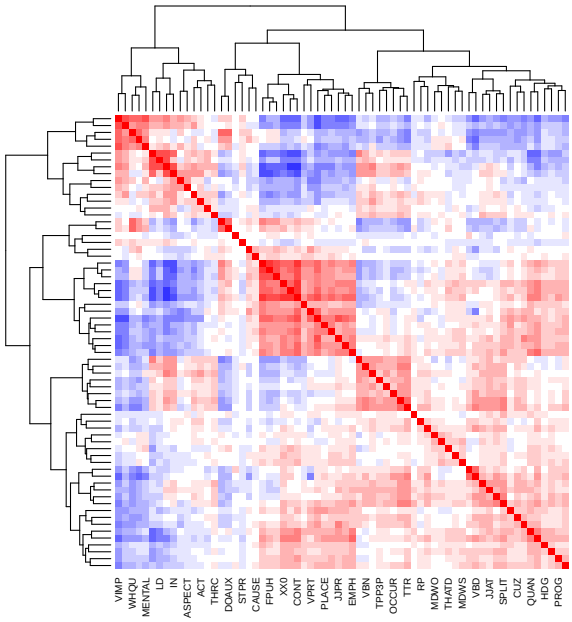



```
#dev.off()
```

5.3.5 Feature correlations

The correlations of the transformed feature frequencies can be visualised in the form of a heatmap. Negative correlations are rendered in blue, whereas positive ones are in red.

```
# Simple heatmap in base R (inspired by Stephanie Evert's SIGIL code)
cor.colours <- c(
  hsv(h=2/3, v=1, s=(10:1)/10), # blue = negative correlation
  rgb(1,1,1), # white = no correlation
  hsv(h=0, v=1, s=(1:10/10))) # red = positive correlation

#png(here("plots", "heatmapzlogcounts-TEC-only.png"), width = 30, height= 30,
 ↪  units = "cm", res = 300)
heatmap(cor(TxBzlogcounts),
        symm=TRUE,
        zlim=c(-1,1),
        col=cor.colours,
        margins=c(0,0))
```


VIMP
WHQU
MENTAL
LD
IN
ASPECT
ACT
THRC
DOAUX
STPR
CAUSE
PPUH
XX0
COMT
VPRT
PLACE
JJR
BMPH
VBN
TPP3P
OCCUR
TTR
RP
MDWO
THATD
MDWS
VBD
JJAT
SPLIT
CUZ
QUAN
HDG
PROG

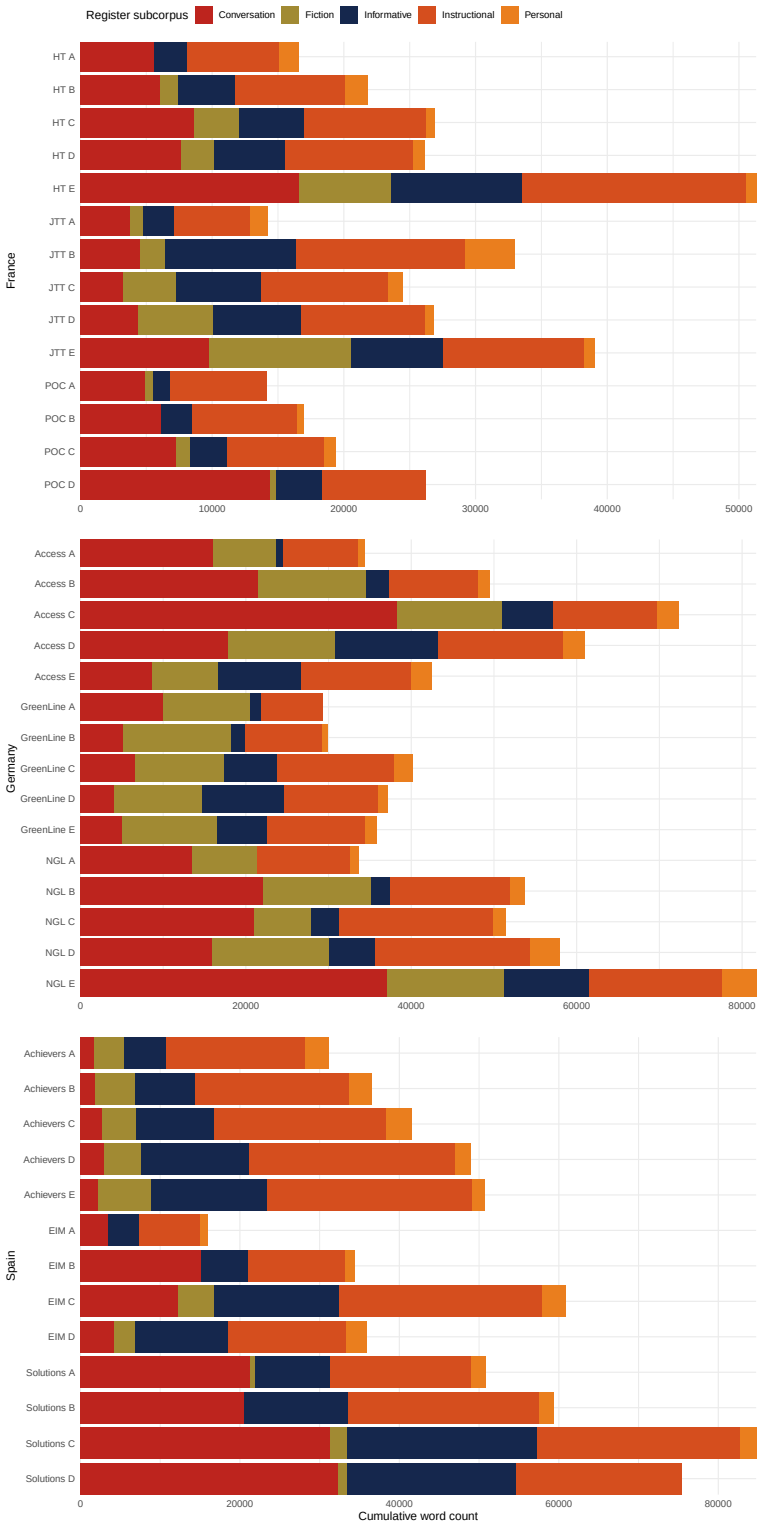
```
#dev.off()

# Calculate the sum of all the words in the tagged texts of the TEC
totalwords <- TxBcounts |>
  select(Words) |>
  sum() |>
  format(big.mark=",")
```

5.4 Composition of TEC texts/files

These figures and tables provide summary statistics on the texts/files of the TEC that were entered in the multi-dimensional model of intra-textbook linguistic variation. In total, the TEC texts entered amounted to 1,693,650 words.

```
metadata <- TxBcounts |>
  select(Filename, Country, Series, Level, Register, Words) |>
  mutate(Volume = paste(Series, Level)) |>
  mutate(Volume = fct_rev(Volume)) |>
  mutate(Volume = fct_reorder(Volume, as.numeric(Level))) |>
  group_by(Volume) |>
  mutate(wordcount = sum(Words)) |>
  ungroup() |>
  distinct(Volume, .keep_all = TRUE)

# Plot for book
metadata2 <- TxBcounts |>
  select(Country, Series, Level, Register, Words) |>
  mutate(Volume = paste(Series, Level)) |>
  mutate(Volume = fct_rev(Volume)) |>
  #mutate(Volume = fct_reorder(Volume, as.numeric(Level))) |>
  group_by(Volume, Register) |>
  mutate(wordcount = sum(Words)) |>
  ungroup() |>
  distinct(Volume, Register, .keep_all = TRUE)

# This is the palette created above on the basis of the suffrager pakcage
 ↪  (but without needed to install the package)
palette <- c("#BD241E", "#A18A33", "#15274D", "#D54E1E", "#EA7E1E",
 ↪  "#4C4C4C", "#722672", "#F9B921", "#267226")
```

```r
PlotSp <- metadata2 |>
  filter(Country=="Spain") |>
  #arrange(Volume) |>
  ggplot(aes(x = Volume, y = wordcount, fill = fct_rev(Register))) +
    geom_bar(stat = "identity", position = "stack") +
    coord_flip(expand = FALSE) + # Removes those annoying ticks before each
     ↪  bar label
    theme_minimal() + theme(legend.position = "none") +
    labs(x = "Spain", y = "Cumulative word count") +
    scale_fill_manual(values = palette[c(5,4,3,2,1)],
                      guide = guide_legend(reverse = TRUE))

PlotGer <- metadata2 |>
  filter(Country=="Germany") |>
  #arrange(Volume) |>
  ggplot(aes(x = Volume, y = wordcount, fill = fct_rev(Register))) +
    geom_bar(stat = "identity", position = "stack") +
    coord_flip(expand = FALSE) +
    labs(x = "Germany", y = "") +
    scale_fill_manual(values = palette[c(5,4,3,2,1)], guide =
     ↪  guide_legend(reverse = TRUE)) +
    theme_minimal() + theme(legend.position = "none")

PlotFr <- metadata2 |>
  filter(Country=="France") |>
  #arrange(Volume) |>
  ggplot(aes(x = Volume, y = wordcount, fill = fct_rev(Register))) +
    geom_bar(stat = "identity", position = "stack") +
    coord_flip(expand = FALSE) +
    labs(x = "France", y  = "", fill = "Register subcorpus") +
    scale_fill_manual(values = palette[c(5,4,3,2,1)], guide =
     ↪  guide_legend(reverse = TRUE, legend.hjust = 0)) +
    theme_minimal() + theme(legend.position = "top", legend.justification =
     ↪  "left")

PlotFr /
PlotGer /
PlotSp
```


Register subcorpus
Conversation
Fiction
Informative
Instructional
Personal
France
HT A
HT B
HT C
HT D
HT E
JTT A
JTT B
JTT C
JTT D
JTT E
POC A
POC B
POC C
POC D
0
10000
20000
30000
40000
50000
Germany
Access A
Access B
Access C
Access D
Access E
GreenLine A
GreenLine B
GreenLine C
GreenLine D
GreenLine E
NGL A
NGL B
NGL C
NGL D
NGL E
0
20000
40000
60000
80000
Spain
Achievers A
Achievers B
Achievers C
Achievers D
Achievers E
EIM A
EIM B
EIM C
EIM D
Solutions A
Solutions B
Solutions C
Solutions D
0
20000
40000
60000
80000
Cumulative word count

```
#ggsave(here("plots", "TEC-T_wordcounts_book.svg"), width = 8, height = 12)
```

The following table provides information about the proportion of instructional language featured in each textbook series.

```
metadataInstr <- TxBcounts |>
  select(Country, Series, Level, Register, Words) |>
  filter(Register=="Instructional") |>
  mutate(Volume = paste(Series, Register)) |>
  mutate(Volume = fct_rev(Volume)) |>
  mutate(Volume = fct_reorder(Volume, as.numeric(Level))) |>
  group_by(Volume, Register) |>
  mutate(InstrWordcount = sum(Words)) |>
  ungroup() |>
  distinct(Volume, .keep_all = TRUE) |>
  select(Series, InstrWordcount)

metaWordcount <- TxBcounts |>
  select(Country, Series, Level, Register, Words) |>
  group_by(Series) |>
  mutate(TECwordcount = sum(Words)) |>
  ungroup() |>
  distinct(Series, .keep_all = TRUE) |>
  select(Series, TECwordcount)

wordcount <- merge(metaWordcount, metadataInstr, by = "Series")

wordcount |>
  mutate(InstrucPercent = InstrWordcount/TECwordcount*100) |>
  arrange(InstrucPercent) |>
  mutate(InstrucPercent = round(InstrucPercent, 2)) |>
  kable(col.names = c("Textbook Series", "Total words", "Instructional
   ↪  words", "% of textbook content"),
        digits = 2,
        format.args = list(big.mark = ","))
```

Textbook Series	Total words	Instructional words	% of textbook content
Access	259,679	60,938	23.47
NGL	278,316	79,312	28.50
GreenLine	172,267	54,263	31.50

Textbook Series	Total words	Instructional words	% of textbook content
Solutions	270,278	87,829	32.50
JTT	137,557	48,375	35.17
HT	142,676	51,550	36.13
POC	76,714	30,548	39.82
EIM	147,185	59,928	40.72
Achievers	208,978	109,886	52.58

6 Summary

In summary, this book has no content whatsoever.

References

Association for Computing Machinery, (ACM). 2020. "Artifact Review and Badging Version 1.1." https://www.acm.org/publications/policies/artifact-review-and-badging-current.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. "Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field." *Linguistics* 56 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Diwersy, Sascha, Stephanie Evert, and Stella Neumann. 2014. "A Weakly Supervised Multivariate Approach to the Study of Language Variation." In, edited by Benedikt Szmrecsanyi and Bernhard Wälchli, 174–204. Berlin: De Gruyter.

Le Foll, Elen. 2021a. *Introducing the Multi-Feature Tagger of English (MFTE)*. Osnabrück University. https://github.com/elenlefoll/MultiFeatureTaggerEnglish.

———. 2021b. *Introducing the Multi-Feature Tagger of English (MFTE)*. Osnabrück University. https://github.com/elenlefoll/MultiFeatureTaggerEnglish.

———. 2024. "Why We Need Open Science and Open Education to Bridge the Corpus Research–practice Gap." In, edited by Peter Crosthwaite, 142–56. London: Routledge.

Love, Robbie, Vaclav Brezina, Tony McEnery, Abi Hawtin, Andrew Hardie, and Claire Dembry. 2019. "Functional Variation in the Spoken BNC2014 and the Potential for Register Analysis." *Register Studies* 1 (2): 296–317. https://doi.org/10.1075/rs.18013.lov.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. "The Spoken BNC2014." *International Journal of Corpus Linguistics* 22 (3): 319–44. https://doi.org/https://doi.org/10.1075/ijcl.22.3.02lov.

McManus, Kevin. 2021. "Are Replication Studies Infrequent Because of Negative Attitudes? Insights from a Survey of Attitudes and Practices in Second Language Research." *Studies in Second Language Acquisition*, December, 1–14. https://doi.org/10.1017/S0272263121000838.

Neumann, Stella, and Stephanie Evert. 2021. "A Register Variation Perspective on Varieties of English." In, edited by Elena Seoane and Douglas Biber, 144178. Studies in Corpus Linguistics 103. Amsterdam: Benjamins.

R Core Team. 2022. "R: A Language and Environment for Statistical Computing." Vienna, Austria. https://www.R-project.org/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.