Textbook English: A Multi-Dimensional Approach

Online Supplements

Elen Le Foll

2024-03-01

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

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

2 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).

2.1 Packages required

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

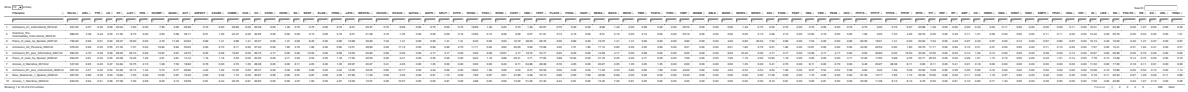
```
#renv::restore() # Restore the project's dependencies from the lockfile to ensure that same p

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
```

2.2 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
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|O
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' = '
#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_M
TxBcounts$Filename <- stringr::str_replace(TxBcounts$Filename, "New_GreenLine", "NGL") # Othe
TxBcounts$Filename <- stringr::str_replace(TxBcounts$Filename, "Piece_of_cake", "POC") # Sho
TxBcounts$Series <- as.factor(stringr::str_extract(TxBcounts$Filename, "Access|Achievers|EIM
summary(TxBcounts$Series)  |>
  kable(col.names = c("Textbook Name", "# of texts"))
```

Textbook Name	# of texts
Access	315
Achievers	240
EIM	180
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 publishe
TxBcounts$Series <-car::recode(TxBcounts$Series, "c('NB', 'JTT') = 'JTT'; c('NM', 'HT') = 'H
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') = 'Germa
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)
```

2.2.1 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 = a
  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"))
```

2.3 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:

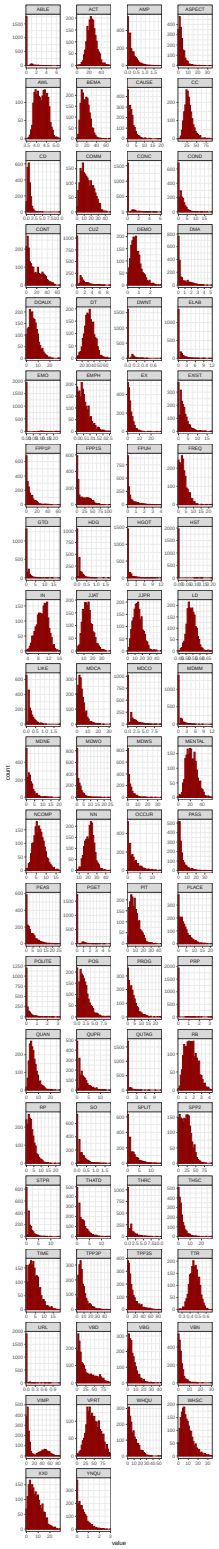
```
TxBcounts <- TxBcounts |>
  filter(Register!="Poetry") |>
  droplevels()

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

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

2.3.1 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
  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
DWK
DOMJK
DT
DWNT
ELAB
EMO
EMPH
EX
EXIST
FPPSP
FPPSS
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
THEC
THSC
TIME
TPPSP
TPP2S
TTR
URL
VBD
VBG
VBN
VIMP
VPRT
WHQU
WHSC
XXE
YNQU
count
value

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

2.3.2 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
  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
zero_features <- compute_percentage(TxBcounts, TxBcounts == 0, 66.6)
#print(zero_features)

# Combine low frequency features into meaningful groups whenever this makes linguistic sense
```

```r
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)
```

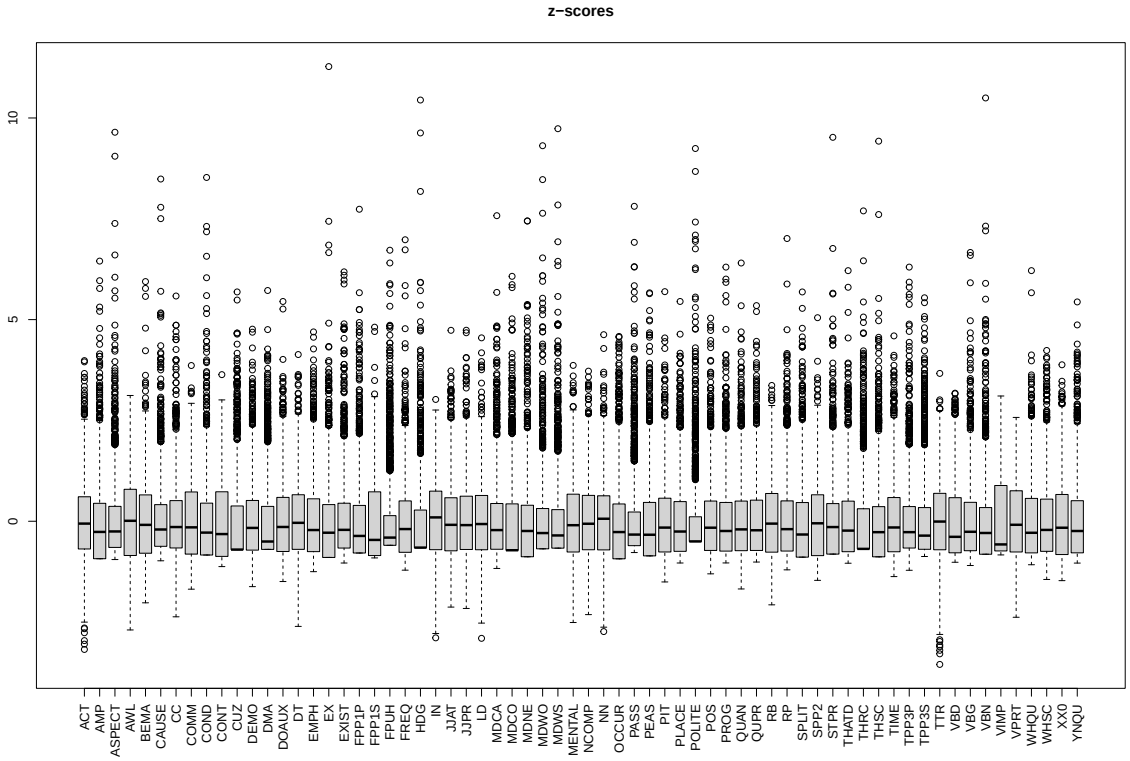
These feature removal operations resulted in a feature set of 64 linguistic variables.

2.3.3 Identifying potential outlier texts

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

```r
# First scale the normalised counts (z-standardisation) to be able to compare the various fea
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
```

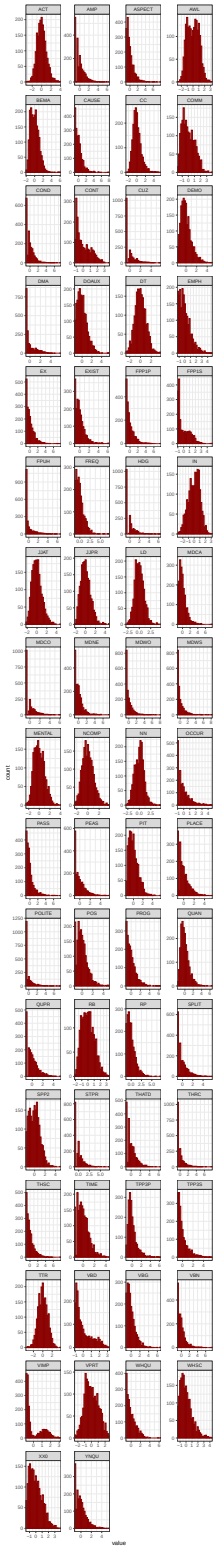
```
                                    Filename
1                      POC_4e_Spoken_0007.txt
2          Solutions_Elementary_Personal_0001.txt
3                NGL_5_Instructional_0018.txt
```

```
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 variable
  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
FPFUS
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
THEC
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 =
```

2.3.4 Signed log transformation

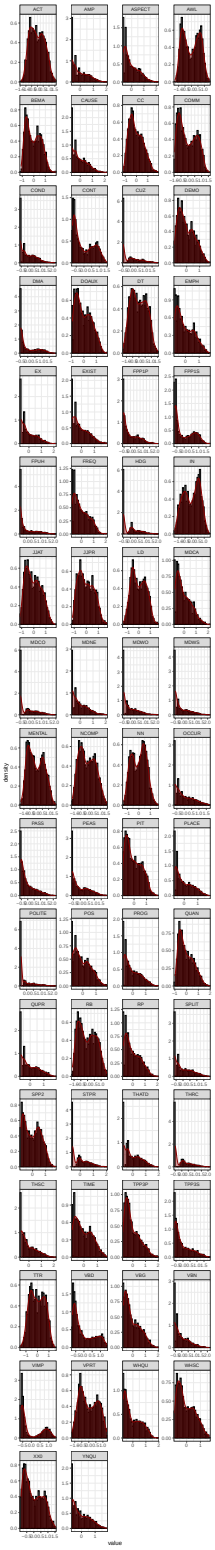
A signed logarithmic transformation was applied to (further) deskew the feature distributions (Diwersy, Evert, and Neumann 2014; Neumann and Evert 2021).

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 p
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 variable
  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
BRAA
CAUSE
CC
COMM
COND
CONT
CUZ
DEMO
DMA
DOAUX
DT
EMPH
EX
EXIST
FPFSP
FPFSS
FPUH
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

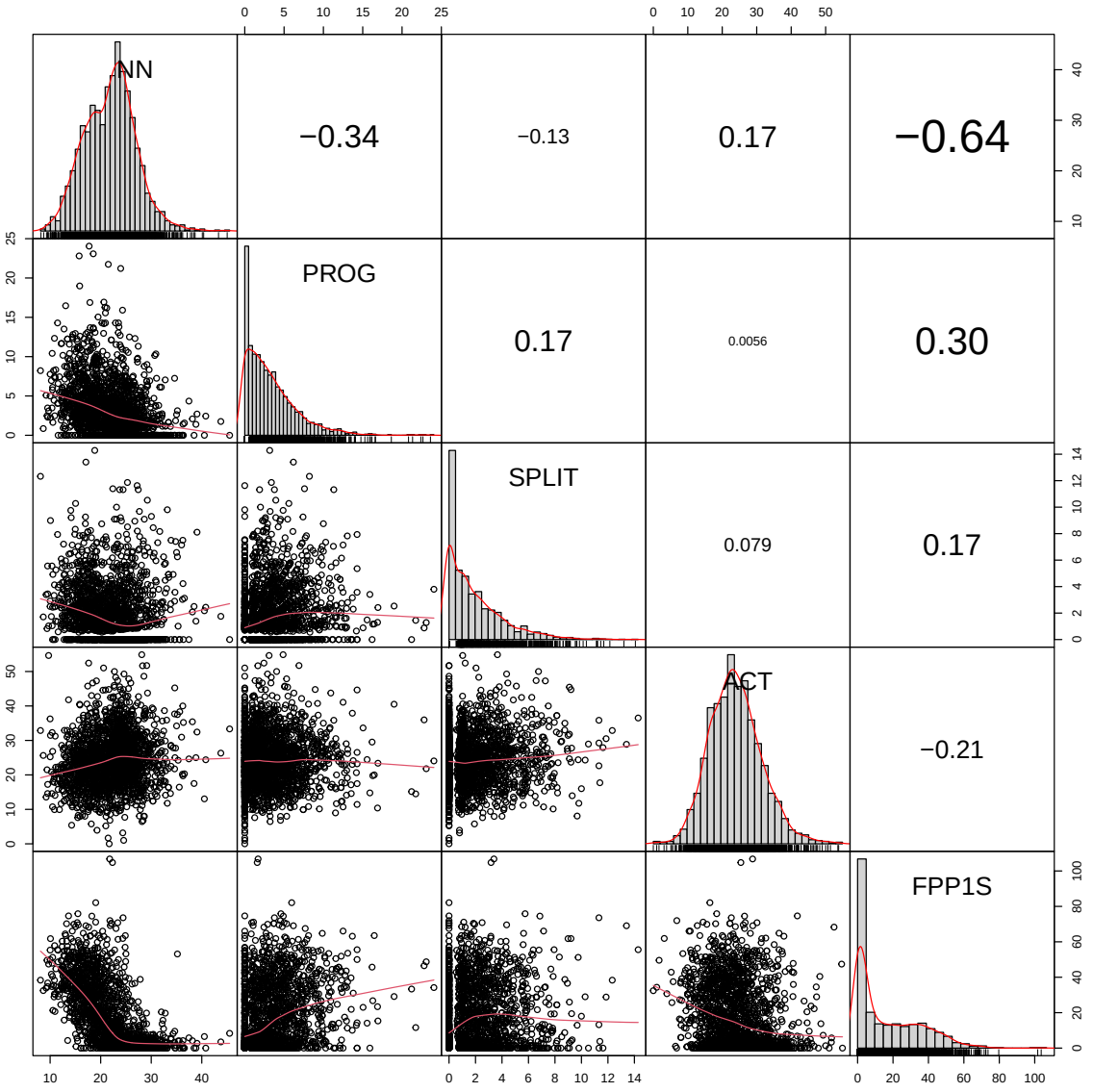
```
#ggsave(here("plots", "DensityPlotsAllVariablesSignedLog-TEC-only.svg"), width = 15, height =
```

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() functio

chart.Correlation.nostars <- function (R, histogram = TRUE, method = c("pearson", "kendall",
  x = checkData(R, method = "matrix")
  if (missing(method))
    method = method[1]
  panel.cor <- function(x, y, digits = 2, prefix = "", use = "pairwise.complete.obs", method
    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
chart.Correlation.nostars(example1, histogram=TRUE, pch=19)
```

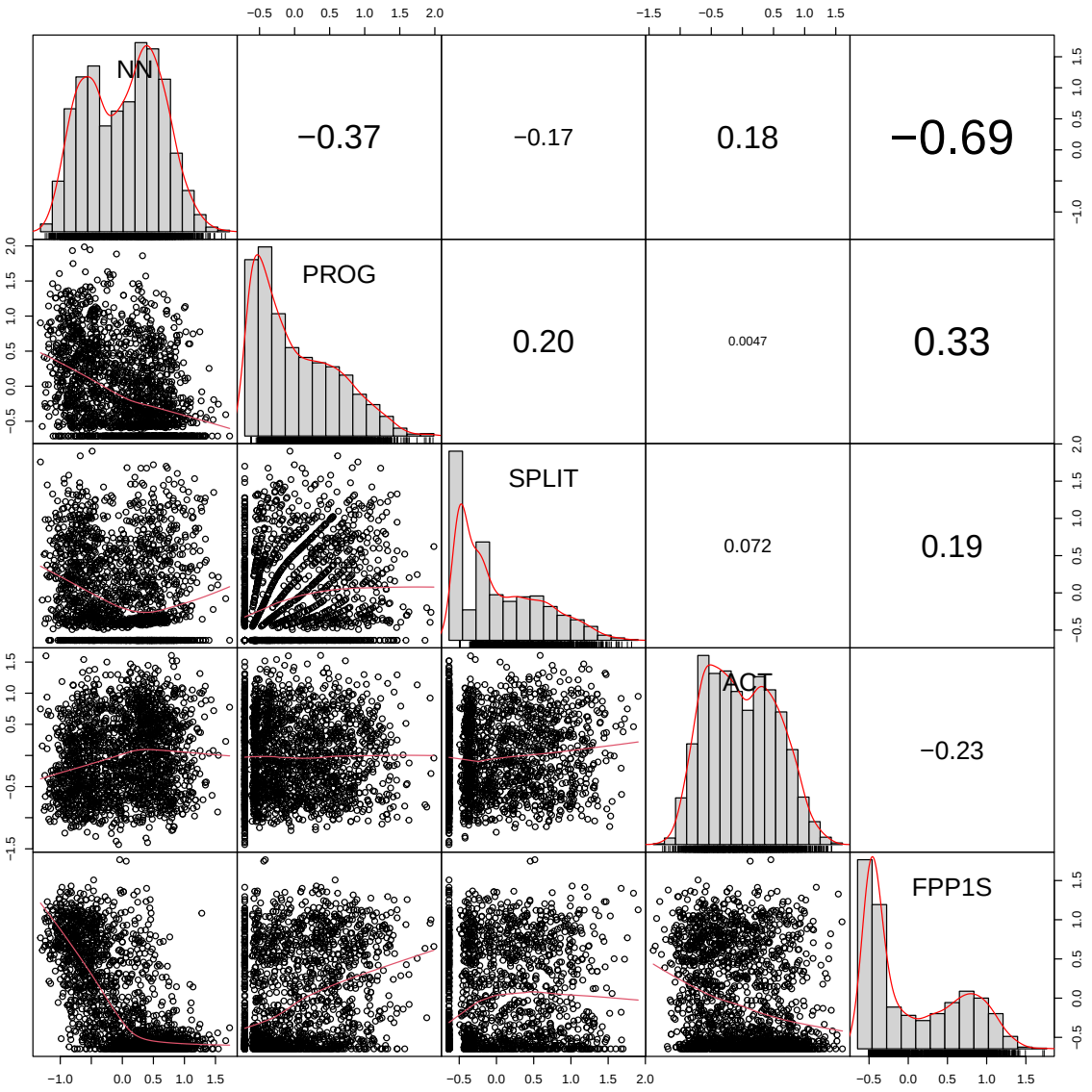



```
#dev.off()
```

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, u
chart.Correlation.nostars(example2, histogram=TRUE, pch=19)
```

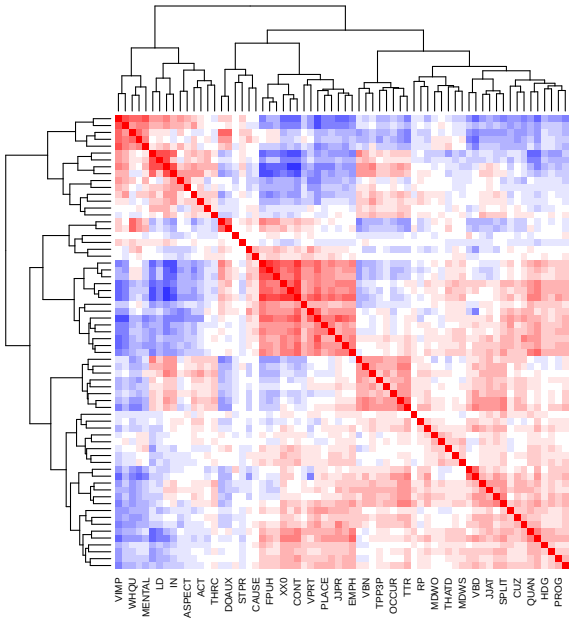



```
#dev.off()
```

2.3.5 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", r
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

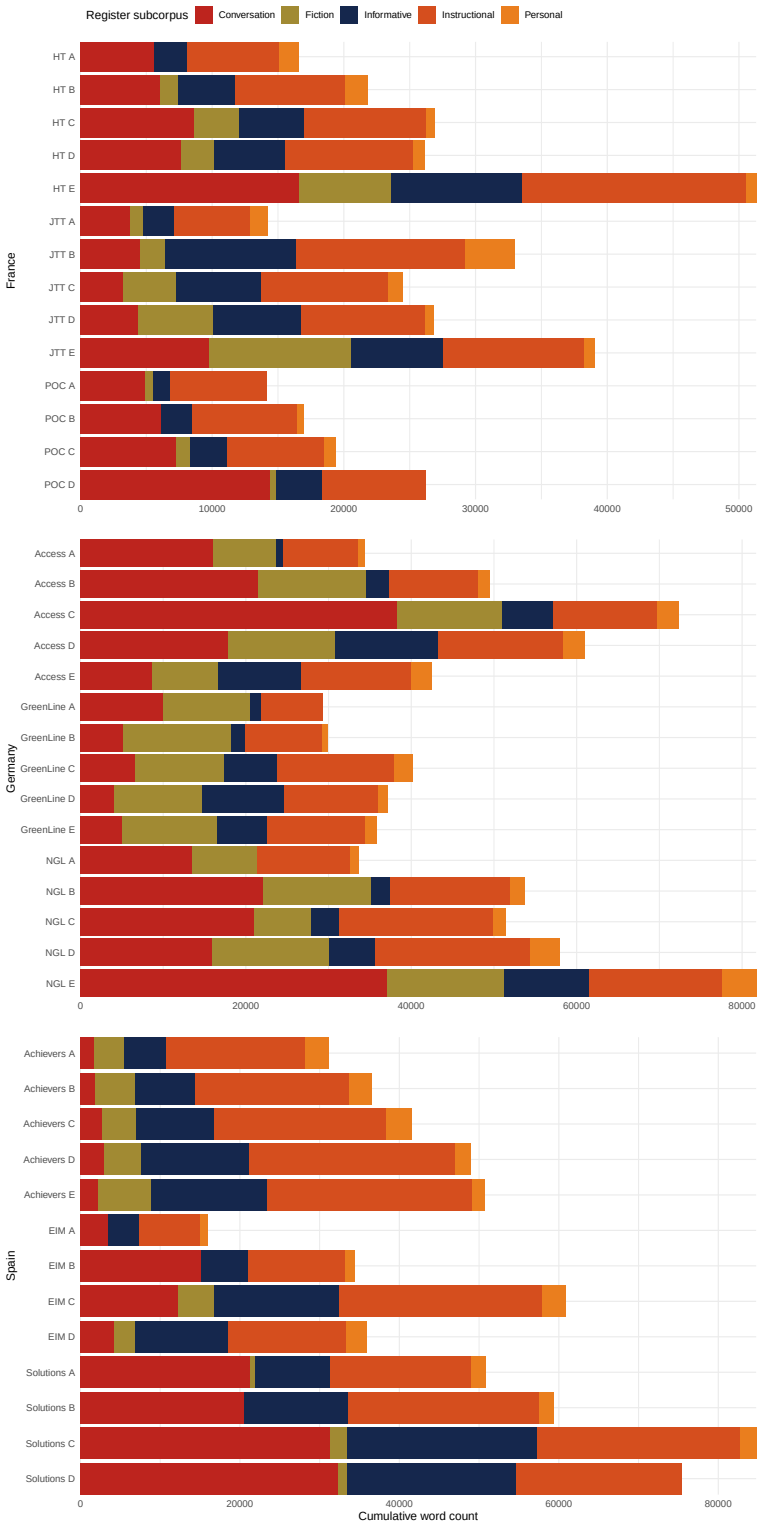
```
#dev.off()

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

2.4 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
palette <- c("#BD241E", "#A18A33", "#15274D", "#D54E1E", "#EA7E1E", "#4C4C4C", "#722672", "#

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, le
    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
        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
Solutions	270,278	87,829	32.50

Textbook Series	Total words	Instructional words	% of textbook content
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

3 Summary

In summary, this book has no content whatsoever.

References

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.

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

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.

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.