Analysis of Available Data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
load_KUK_subcorpus_metadata <- function(crp) {
  read_tsv(paste(c(
    "../corpora/KUK_1.0/metadata/", crp, "_DocumentFileFormat.tsv"
  ), collapse = "")) %>%
    filter(FileFormat == "TXT") %>%
    full_join(
      read_tsv(paste(c(
        "../corpora/KUK_1.0/metadata/",
        crp,
        "_DocumentIdentificationGenreProperties.tsv"
      ), collapse = "")),
      by = "KUK_ID"
    ) %>%
    mutate(across(where(is.numeric), as.character)) %>%
    mutate(subcorpus = crp) %>%
    select(KUK_ID, FileName, FileFormat, FolderPath, subcorpus, everything())
}

kuky_orig <- fromJSON("../corpora/KUKY/argumentative.json")$documents %>%
  as.data.frame() %>%
  bind_rows(
    fromJSON("../corpora/KUKY/normative.json")$documents %>% as.data.frame()
  ) %>%
  rename(KUK_ID = doc_id) %>%
  select(!c(plainText, doc_name)) %>%
  select(KUK_ID, everything())
```

```
kuky_kuk <- load_KUK_subcorpus_metadata("KUKY") %>%
  filter(FolderPath != "data/KUKY/MD_TXT")
```

```
## Rows: 448 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): KUK_ID, FileName, FileFormat, FolderPath
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 224 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (8): KUK_ID, SourceDB, Anonymized, RecipientType, RecipientIndividuation...
## lgl (4): SourceID, DocumentTitle, ClarityPursuit, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
kuky <- kuky_kuk %>% full_join(kuky_orig, by = "KUK_ID")
czcdc <- load_KUK_subcorpus_metadata("CzCDC")
```

```
## Rows: 237723 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): KUK_ID, FileName, FileFormat, FolderPath
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 237723 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (10): KUK_ID, SourceDB, SourceID, DocumentTitle, Anonymized, RecipientTy...
## lgl  (2): ClarityPursuit, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
eso <- load_KUK_subcorpus_metadata("ESO")
```

```
## Rows: 11230 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): KUK_ID, FileFormat, FolderPath
## dbl (1): FileName
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 5615 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (10): KUK_ID, SourceDB, SourceID, DocumentTitle, Anonymized, RecipientTy...
## lgl  (2): ClarityPursuit, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
frbo <- load_KUK_subcorpus_metadata("FrBo")
```

```
## Rows: 638 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): KUK_ID, FileName, FileFormat, FolderPath
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 319 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (10): KUK_ID, SourceDB, SourceID, DocumentTitle, Anonymized, RecipientTy...
## lgl  (2): ClarityPursuit, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
lifrlaw <- load_KUK_subcorpus_metadata("LiFRLaw")
```

```
## Rows: 36 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): KUK_ID, FileName, FileFormat, FolderPath
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 18 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (9): KUK_ID, SourceDB, SourceID, DocumentTitle, Anonymized, Recipient Ty...
## lgl (2): ClarityPursuit, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ombuflyers <- load_KUK_subcorpus_metadata("OmbuFlyers")
```

```
## Rows: 234 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): KUK_ID, FileName, FileFormat, FolderPath
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 117 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (8): KUK_ID, DocumentTitle, Anonymized, RecipientType, RecipientIndividu...
## lgl (4): SourceDB, SourceID, ClarityPursuit, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
df <- bind_rows(kuky, czcdc) %>%
  bind_rows(eso) %>%
  bind_rows(frbo) %>%
  bind_rows(lifrlaw) %>%
  bind_rows(ombuflyers)

str(df)
```

```
## tibble [244,016 x 35] (S3: tbl_df/tbl/data.frame)
##  $ KUK_ID                   : chr [1:244016] "671918e2c6537d54ff0626db" "671918e2c6537d54ff0626dc" "67
##  $ FileName                 : chr [1:244016] "orig_Certifikáty autorizovaných inspektorů" "red_Co je t
##  $ FileFormat               : chr [1:244016] "TXT" "TXT" "TXT" "TXT" ...
##  $ FolderPath               : chr [1:244016] "data/KUKY/TXT" "data/KUKY/TXT" "data/KUKY/TXT" "data/KU
##  $ subcorpus                : chr [1:244016] "KUKY" "KUKY" "KUKY" "KUKY" ...
##  $ SourceDB                 : chr [1:244016] "SourceDB" "SourceDB" "SourceDB" "SourceDB" ...
##  $ SourceID                 : chr [1:244016] NA NA NA NA ...
##  $ DocumentTitle            : chr [1:244016] NA NA NA NA ...
##  $ ClarityPursuit           : logi [1:244016] NA NA NA NA NA NA ...
##  $ Anonymized.x             : chr [1:244016] "No" "No" "No" "No" ...
##  $ RecipientType.x          : chr [1:244016] "natural person" "natural person" "natural person" "natu
##  $ RecipientIndividuation.x: chr [1:244016] "public" "public" "public" "public" ...
##  $ AuthorType.x             : chr [1:244016] "individual" "individual" "individual" "authority" ...
##  $ Objectivity.x            : chr [1:244016] "quasiobjective" "quasiobjective" "quasiobjective" "quasi
##  $ LegalActType.x           : chr [1:244016] "normative" "normative" "normative" "normative" ...
##  $ Bindingness.x            : logi [1:244016] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Readability              : chr [1:244016] "low" "high" "low" "low" ...
##  $ SyllogismBased           : chr [1:244016] "false" "false" "false" "false" ...
##  $ DocumentVersion          : chr [1:244016] "Original" "Redesign" "Original" "Original" ...
##  $ ParentDocumentID         : chr [1:244016] NA NA NA NA ...
##  $ LegalActType.y           : chr [1:244016] "normative" "normative" "normative" "normative" ...
##  $ Objectivity.y            : chr [1:244016] "quasiobjective" "quasiobjective" "quasiobjective" "quasi
##  $ Bindingness.y            : logi [1:244016] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ AuthorType.y             : chr [1:244016] "individual" "individual" "individual" "authority" ...
##  $ RecipientType.y          : chr [1:244016] "natural person" "natural person" "natural person" "natu
##  $ RecipientIndividuation.y: chr [1:244016] "public" "public" "public" "public" ...
##  $ Anonymized.y             : chr [1:244016] "No" "No" "No" "No" ...
##  $ Anonymized               : chr [1:244016] NA NA NA NA ...
##  $ RecipientType            : chr [1:244016] NA NA NA NA ...
##  $ RecipientIndividuation   : chr [1:244016] NA NA NA NA ...
##  $ AuthorType               : chr [1:244016] NA NA NA NA ...
##  $ Objectivity              : chr [1:244016] NA NA NA NA ...
##  $ LegalActType             : chr [1:244016] NA NA NA NA ...
##  $ Bindingness              : logi [1:244016] NA NA NA NA NA NA ...
##  $ Recipient Type           : chr [1:244016] NA NA NA NA ...
```

```r
readable <- df %>% filter(Readability %in% c("high", "medium"))
unreadable <- df %>% filter(!(Readability %in% c("high", "medium")))

str(readable)
```

```
## tibble [186 x 35] (S3: tbl_df/tbl/data.frame)
##  $ KUK_ID                   : chr [1:186] "671918e2c6537d54ff0626dc" "673b7a37c6537d54ff062b8d" "673b
##  $ FileName                 : chr [1:186] "red_Co je to územní plánování_final_přidat odkaz na manuál
##  $ FileFormat               : chr [1:186] "TXT" "TXT" "TXT" "TXT" ...
##  $ FolderPath               : chr [1:186] "data/KUKY/TXT" "data/KUKY/TXT" "data/KUKY/TXT" "data/KUKY/
```

```
##  $ subcorpus               : chr [1:186] "KUKY" "KUKY" "KUKY" "KUKY" ...
##  $ SourceDB                : chr [1:186] "SourceDB" "SourceDB" "SourceDB" "SourceDB" ...
##  $ SourceID                : chr [1:186] NA NA NA NA ...
##  $ DocumentTitle           : chr [1:186] NA NA NA NA ...
##  $ ClarityPursuit          : logi [1:186] NA NA NA NA NA NA ...
##  $ Anonymized.x            : chr [1:186] "No" "No" "No" "No" ...
##  $ RecipientType.x         : chr [1:186] "natural person" "natural person" "natural person" "natural
##  $ RecipientIndividuation.x: chr [1:186] "public" "public" "public" "public" ...
##  $ AuthorType.x            : chr [1:186] "individual" "individual" "authority" "authority" ...
##  $ Objectivity.x           : chr [1:186] "quasiobjective" "quasiobjective" "quasiobjective" "quasiob
##  $ LegalActType.x          : chr [1:186] "normative" "normative" "normative" "normative" ...
##  $ Bindingness.x           : logi [1:186] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Readability             : chr [1:186] "high" "high" "high" "medium" ...
##  $ SyllogismBased          : chr [1:186] "false" "false" "false" "false" ...
##  $ DocumentVersion         : chr [1:186] "Redesign" "Redesign" "Redesign" "Original" ...
##  $ ParentDocumentID        : chr [1:186] NA NA NA NA ...
##  $ LegalActType.y          : chr [1:186] "normative" "normative" "normative" "normative" ...
##  $ Objectivity.y           : chr [1:186] "quasiobjective" "quasiobjective" "quasiobjective" "quasiob
##  $ Bindingness.y           : logi [1:186] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ AuthorType.y            : chr [1:186] "individual" "individual" "authority" "authority" ...
##  $ RecipientType.y         : chr [1:186] "natural person" "natural person" "natural person" "natural
##  $ RecipientIndividuation.y: chr [1:186] "public" "public" "public" "public" ...
##  $ Anonymized.y            : chr [1:186] "No" "No" "No" "No" ...
##  $ Anonymized              : chr [1:186] NA NA NA NA ...
##  $ RecipientType           : chr [1:186] NA NA NA NA ...
##  $ RecipientIndividuation  : chr [1:186] NA NA NA NA ...
##  $ AuthorType              : chr [1:186] NA NA NA NA ...
##  $ Objectivity             : chr [1:186] NA NA NA NA ...
##  $ LegalActType            : chr [1:186] NA NA NA NA ...
##  $ Bindingness             : logi [1:186] NA NA NA NA NA NA ...
##  $ Recipient Type          : chr [1:186] NA NA NA NA ...
```

```
str(unreadable)
```

```
## tibble [243,830 x 35] (S3: tbl_df/tbl/data.frame)
##  $ KUK_ID                  : chr [1:243830] "671918e2c6537d54ff0626db" "671918e2c6537d54ff0626dd" "67
##  $ FileName                : chr [1:243830] "orig_Certifikáty autorizovaných inspektorů" "orig_finan
##  $ FileFormat              : chr [1:243830] "TXT" "TXT" "TXT" "TXT" ...
##  $ FolderPath              : chr [1:243830] "data/KUKY/TXT" "data/KUKY/TXT" "data/KUKY/TXT" "data/KU
##  $ subcorpus               : chr [1:243830] "KUKY" "KUKY" "KUKY" "KUKY" ...
##  $ SourceDB                : chr [1:243830] "SourceDB" "SourceDB" "SourceDB" "SourceDB" ...
##  $ SourceID                : chr [1:243830] NA NA NA NA ...
##  $ DocumentTitle           : chr [1:243830] NA NA NA NA ...
##  $ ClarityPursuit          : logi [1:243830] NA NA NA NA NA NA ...
##  $ Anonymized.x            : chr [1:243830] "No" "No" "No" "No" ...
##  $ RecipientType.x         : chr [1:243830] "natural person" "natural person" "natural person" "natu
##  $ RecipientIndividuation.x: chr [1:243830] "public" "public" "public" "public" ...
##  $ AuthorType.x            : chr [1:243830] "individual" "individual" "authority" "authority" ...
##  $ Objectivity.x           : chr [1:243830] "quasiobjective" "quasiobjective" "quasiobjective" "quasi
##  $ LegalActType.x          : chr [1:243830] "normative" "normative" "normative" "normative" ...
##  $ Bindingness.x           : logi [1:243830] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Readability             : chr [1:243830] "low" "low" "low" "low" ...
##  $ SyllogismBased          : chr [1:243830] "false" "false" "false" "false" ...
##  $ DocumentVersion         : chr [1:243830] "Original" "Original" "Original" "Original" ...
##  $ ParentDocumentID        : chr [1:243830] NA NA NA NA ...
```

```
##  $ LegalActType.y          : chr [1:243830] "normative" "normative" "normative" "normative" ...
##  $ Objectivity.y           : chr [1:243830] "quasiobjective" "quasiobjective" "quasiobjective" "quasi
##  $ Bindingness.y           : logi [1:243830] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ AuthorType.y            : chr [1:243830] "individual" "individual" "authority" "authority" ...
##  $ RecipientType.y         : chr [1:243830] "natural person" "natural person" "natural person" "natu
##  $ RecipientIndividuation.y: chr [1:243830] "public" "public" "public" "public" ...
##  $ Anonymized.y            : chr [1:243830] "No" "No" "No" "No" ...
##  $ Anonymized              : chr [1:243830] NA NA NA NA ...
##  $ RecipientType           : chr [1:243830] NA NA NA NA ...
##  $ RecipientIndividuation  : chr [1:243830] NA NA NA NA ...
##  $ AuthorType              : chr [1:243830] NA NA NA NA ...
##  $ Objectivity             : chr [1:243830] NA NA NA NA ...
##  $ LegalActType            : chr [1:243830] NA NA NA NA ...
##  $ Bindingness             : logi [1:243830] NA NA NA NA NA NA ...
##  $ Recipient Type          : chr [1:243830] NA NA NA NA ...
```