Appendix 2: Data Analysis Documentation

This appendix describes the R-code that was used for the analysis in the paper "Word-order variation in a contact setting: A corpus-based investigation of Russian spoken in Daghestan".

R version:

```
getRversion()
```

```
## [1] '4.0.5'
```

Versions of the packages used in the analysis are specified at the very end of the document.

Preparation of the data

Import the data in R

```
library("tidyverse")
```

```
gen_std <- read.csv("std_rus.csv", stringsAsFactors=TRUE)
```

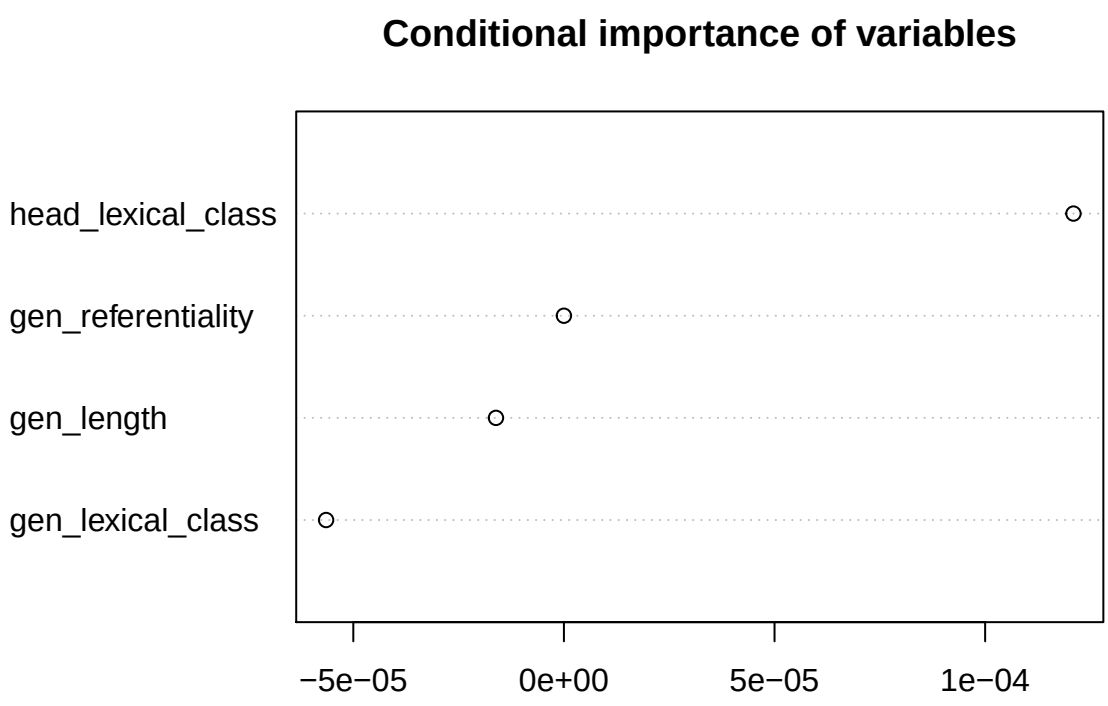
Set correct reference levels

```
gen_std$head_lexical_class <- relevel(gen_std$head_lexical_class, "non_kinship")
gen_std$gen_lexical_class <- relevel(gen_std$gen_lexical_class, "non_human")
gen_std$gen_referentiality <- relevel(gen_std$gen_referentiality,"non_definite")
gen_std$gen_length <- relevel(gen_std$gen_length,"one-word")
```

Random forest

```
library("party")
```

```
gen_std_rf <- cforest(position ~ gen_lexical_class + head_lexical_class +
                        gen_referentiality + gen_length,
                      data = gen_std, controls = cforest_unbiased(ntree = 1000, mtry = 2))
gen_std_varimp <- varimp(gen_std_rf, conditional = TRUE)
dotchart(sort(gen_std_varimp), main = "Conditional importance of variables")
```




```
library(Hmisc)
```

```
gen_std_rf.pred <- unlist(treeresponse(gen_std_rf))[c(FALSE,TRUE)]
```

```
somers2(gen_std_rf.pred, as.numeric(gen_std$position) -1)
```

```
##           C         Dxy           n     Missing
##   0.8771261   0.7542522 675.0000000   0.0000000
```

```
table(predict(gen_std_rf), gen_std$position)
```

```
##
##         left right
##   left     0     0
##   right   19   656
```

```
656/675
```

```
## [1] 0.9718519
```

Versions of the packages used in the analysis:

```
installed.packages()[names(sessionInfo()$otherPkgs), "Version"]
```

```
##       Hmisc     Formula    survival     lattice       party strucchange
##     "4.5-0"     "1.2-4"    "3.2-10"   "0.20-41"     "1.3-7"     "1.5-2"
##    sandwich         zoo  modeltools     mvtnorm     forcats     stringr
##     "3.0-0"     "1.8-9"    "0.2-23"     "1.1-1"     "0.5.1"     "1.4.0"
##       dplyr       purrr       readr       tidyr      tibble     ggplot2
##     "1.0.5"     "0.3.4"     "1.4.0"     "1.1.3"     "3.1.0"     "3.3.3"
##   tidyverse
##     "1.3.0"
```