Appendix: 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" by Chiara Naccarato, Anastasia Panova, Natalia Stoynova.

R version:

```
getRversion()
```

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

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

Data preparation

Import the data in R

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

```
gen <- read.csv("dag_rus.csv", stringsAsFactors=TRUE)
```

Set the correct reference levels

```
gen$head_lexical_class <- relevel(gen$head_lexical_class, "non_kinship")
gen$gen_lexical_class <- relevel(gen$gen_lexical_class, "non_human")
gen$gender <- relevel(gen$gender,"m")
gen$gen_referentiality <- relevel(gen$gen_referentiality,"non_definite")
gen$gen_length <- relevel(gen$gen_length,"one-word")
gen$head_length <- relevel(gen$head_length,"multi-word")
gen$givenness <- relevel(gen$givenness,"other")
gen$year_of_birth <- relevel(gen$year_of_birth, "<1950")
```

Logistic regression

Full model

```
library("lme4")
```

The full model includes all sociolinguistic, lexico-semantic and formal parameters used for the analysis of the data. Information about the speakers is included as a random effect to take into account possible idiosyncrasies. The response variable is the position of the genitive in noun phrases, i.e. left or right.

```
model1 <- glmer (position ~ gen_lexical_class + head_lexical_class + education +
                   gender + gen_referentiality + language_family + year_of_birth +
                   gen_individuation + gen_length + head_length + givenness +
                   (1|speaker), data = gen, family ="binomial",
                 control = glmerControl(optimizer ="bobyqa"))
summary(model1)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: position ~ gen_lexical_class + head_lexical_class + education +
    gender + gen_referentiality + language_family + year_of_birth +
    gen_individuation + gen_length + head_length + givenness +
    (1 | speaker)
   Data: gen
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   333.7    396.3   -151.8    303.7      467

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.4222  0.1350  0.2147  0.3230  3.6105

Random effects:
 Groups  Name        Variance Std.Dev.
 speaker (Intercept) 0.0606   0.2462
Number of obs: 482, groups:  speaker, 40

Fixed effects:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    4.18911    0.69196   6.054 1.41e-09 ***
gen_lexical_classhuman_other  -0.53454    0.55150  -0.969  0.33243
gen_lexical_classkinship      -1.53134    0.53677  -2.853  0.00433 **
gen_lexical_classproper_human -1.02893    0.65519  -1.570  0.11631
head_lexical_classkinship     -2.30205    0.50952  -4.518 6.24e-06 ***
educationlower                -0.45161    0.37045  -1.219  0.22281
genderf                       -0.67656    0.34882  -1.940  0.05243 .
gen_referentialitydefinite    -0.91080    0.46021  -1.979  0.04780 *
language_familyTurkic         -0.25240    0.40048  -0.630  0.52854
year_of_birth>1950             0.05908    0.34311   0.172  0.86328
gen_individuationsg            0.36664    0.41896   0.875  0.38151
gen_lengthmulti-word          -0.62974    0.32873  -1.916  0.05540 .
head_lengthone-word           -0.61041    0.45841  -1.332  0.18300
givennesshnew_ggiven           0.14707    0.39492   0.372  0.70960
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Correlation matrix not shown by default, as p = 14 > 12.
Use print(x, correlation=TRUE)  or
    vcov(x)        if you need it
```

Step-wise selection procedure

Given that p-values alone do not constitute a sufficient reason to remove a variable from the model, we resort to AIC (Akaike Information Criterion) comparison, i.e. we compare the AIC (Akaike Information Criterion) of the model including all predictors and of other models lacking one of the predictors, and remove (one at a time) the predictors whose presence in the model increases the AIC value.

```
drop1(model1)
```

```
boundary (singular) fit: see ?isSingular
```

```
Single term deletions

Model:
position ~ gen_lexical_class + head_lexical_class + education +
    gender + gen_referentiality + language_family + year_of_birth +
    gen_individuation + gen_length + head_length + givenness +
    (1 | speaker)
                  npar    AIC
<none>                 333.67
gen_lexical_class     3 335.45
head_lexical_class    1 353.34
education             1 333.10
gender                1 335.30
gen_referentiality    1 335.81
language_family       1 332.06
year_of_birth         1 331.70
gen_individuation     1 332.43
gen_length            1 335.38
head_length           1 333.55
givenness             1 331.81
```

The lowest AIC is associated with the model not including the predictor year_of_birth, so we remove it first.

```
model2 <- glmer (position ~ gen_lexical_class + head_lexical_class + education +
                   gender + gen_referentiality + language_family +
                   gen_individuation + gen_length + head_length + givenness +
                   (1|speaker), data = gen, family ="binomial",
                 control = glmerControl(optimizer ="bobyqa"))
drop1(model2)
```

```
boundary (singular) fit: see ?isSingular
```

```
Single term deletions

Model:
position ~ gen_lexical_class + head_lexical_class + education +
    gender + gen_referentiality + language_family + gen_individuation +
    gen_length + head_length + givenness + (1 | speaker)
                  npar    AIC
<none>                 331.70
```

```
gen_lexical_class     3 333.51
head_lexical_class    1 351.49
education             1 331.29
gender                1 333.30
gen_referentiality    1 333.81
language_family       1 330.08
gen_individuation     1 330.46
gen_length            1 333.40
head_length           1 331.56
givenness             1 329.83
```

The lowest AIC is associated with the model not including the predictor givenness, so we remove it.

```
model3 <- glmer (position ~ gen_lexical_class + head_lexical_class + education +
                   gender + gen_referentiality + language_family +
                   gen_individuation + gen_length + head_length +
                   (1|speaker), data = gen, family ="binomial",
                 control = glmerControl(optimizer ="bobyqa"))
drop1(model3)
```

```
Single term deletions

Model:
position ~ gen_lexical_class + head_lexical_class + education +
    gender + gen_referentiality + language_family + gen_individuation +
    gen_length + head_length + (1 | speaker)
                   npar    AIC
<none>                  329.83
gen_lexical_class     3 331.56
head_lexical_class    1 351.16
education             1 329.41
gender                1 331.73
gen_referentiality    1 331.88
language_family       1 328.21
gen_individuation     1 328.51
gen_length            1 331.58
head_length           1 329.81
```

The lowest AIC is associated with the model not including the predictor language_family, so we remove it.

```
model4 <- glmer (position ~ gen_lexical_class + head_lexical_class + education +
                   gender + gen_referentiality +
                   gen_individuation + gen_length + head_length +
                   (1|speaker), data = gen, family ="binomial",
                 control = glmerControl(optimizer ="bobyqa"))
drop1(model4)
```

```
Single term deletions

Model:
position ~ gen_lexical_class + head_lexical_class + education +
    gender + gen_referentiality + gen_individuation + gen_length +
```

```
    head_length + (1 | speaker)
                   npar    AIC
<none>                  328.21
gen_lexical_class     3 330.96
head_lexical_class    1 349.23
education             1 327.57
gender                1 329.75
gen_referentiality    1 330.22
gen_individuation     1 327.04
gen_length            1 330.07
head_length           1 328.18
```

The lowest AIC is associated with the model not including the predictor gen_individuation, so we remove it.

```
model5 <- glmer (position ~ gen_lexical_class + head_lexical_class + education +
                   gender + gen_referentiality +
                   gen_length + head_length +
                   (1|speaker), data = gen, family ="binomial",
                 control = glmerControl(optimizer ="bobyqa"))
drop1(model5)
```

```
Single term deletions

Model:
position ~ gen_lexical_class + head_lexical_class + education +
    gender + gen_referentiality + gen_length + head_length +
    (1 | speaker)
                   npar    AIC
<none>                  327.04
gen_lexical_class     3 329.08
head_lexical_class    1 347.36
education             1 326.38
gender                1 328.74
gen_referentiality    1 328.35
gen_length            1 329.43
head_length           1 327.00
```

The lowest AIC is associated with the model not including the predictor education, so we remove it.

```
model6 <- glmer (position ~ gen_lexical_class + head_lexical_class +
                   gender + gen_referentiality +
                   gen_length + head_length +
                   (1|speaker), data = gen, family ="binomial",
                 control = glmerControl(optimizer ="bobyqa"))
drop1(model6)
```

```
Single term deletions

Model:
position ~ gen_lexical_class + head_lexical_class + gender +
    gen_referentiality + gen_length + head_length + (1 | speaker)
```

```
                   npar    AIC
<none>                  326.38
gen_lexical_class     3 329.81
head_lexical_class    1 346.61
gender                1 327.56
gen_referentiality    1 327.37
gen_length            1 328.51
head_length           1 326.20
```

The lowest AIC is associated with the model not including the predictor head_length, so we remove it.

```
model7 <- glmer (position ~ gen_lexical_class + head_lexical_class +
                   gender + gen_referentiality +
                   gen_length +
                   (1|speaker), data = gen, family ="binomial",
                 control = glmerControl(optimizer ="bobyqa"))
```

```
drop1(model7)
```

```
Single term deletions

Model:
position ~ gen_lexical_class + head_lexical_class + gender +
    gen_referentiality + gen_length + (1 | speaker)
                   npar    AIC
<none>                  326.20
gen_lexical_class     3 330.01
head_lexical_class    1 346.76
gender                1 327.31
gen_referentiality    1 326.73
gen_length            1 328.39
```

The lowest AIC is associated with the model including all remaining predictors, which we will further consider as the minimal adequate model.

Minimal adequate model

```
summary(model7)
```

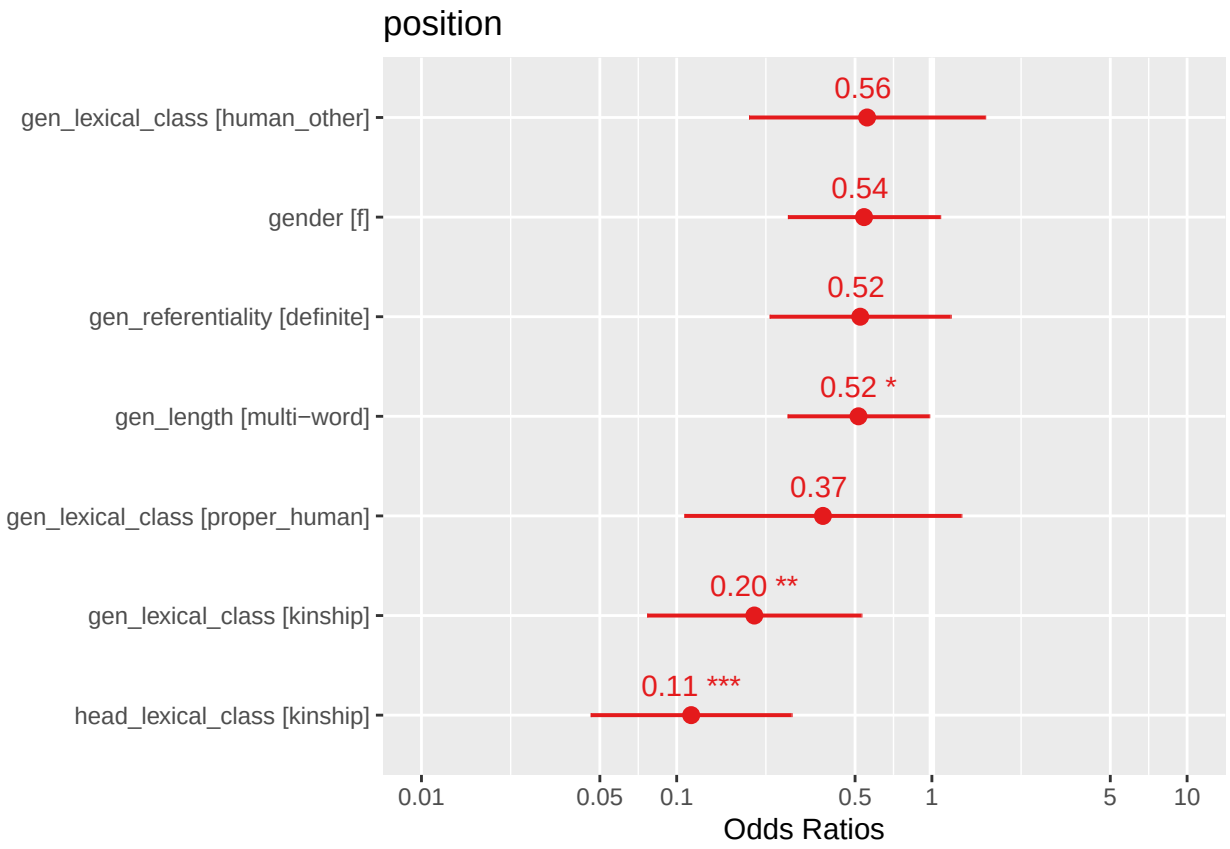
```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: position ~ gen_lexical_class + head_lexical_class + gender +
    gen_referentiality + gen_length + (1 | speaker)
   Data: gen
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   326.2    363.8   -154.1    308.2      473
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.4974  0.1599  0.2298  0.3150  3.1782

Random effects:
 Groups  Name        Variance Std.Dev.
 speaker (Intercept) 0.1122   0.335
Number of obs: 482, groups:  speaker, 40

Fixed effects:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.5554     0.4276   8.314  < 2e-16 ***
gen_lexical_classhuman_other  -0.5845     0.5436  -1.075  0.28230
gen_lexical_classkinship      -1.6019     0.4930  -3.249  0.00116 **
gen_lexical_classproper_human -0.9832     0.6374  -1.542  0.12298
head_lexical_classkinship     -2.1717     0.4623  -4.698 2.63e-06 ***
genderf                       -0.6120     0.3505  -1.746  0.08084 .
gen_referentialitydefinite    -0.6466     0.4167  -1.552  0.12076
gen_lengthmulti-word          -0.6624     0.3260  -2.032  0.04214 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) gn_lxcl_clssh_ gn_lx_ gn_lxcl_clssp_ hd_lx_ gendrf gn_rfr
gn_lxcl_clssh_ -0.370
gn_lxcl_cls    -0.084  0.360
gn_lxcl_clssp_ -0.036  0.293         0.547
hd_lxcl_cls    -0.084 -0.314        -0.631 -0.515
genderf        -0.359  0.025        -0.169 -0.149          0.027
gn_rfrntlty    -0.657  0.193        -0.161 -0.187         -0.016  0.106
gn_lngthml-    -0.237  0.048         0.072  0.126          0.134 -0.100 -0.245
```

Visualization of the estimates in the minimal adequate model

```
library(sjPlot)
library(ggplot2)
```

```
plot_model(model7, type = "est", show.values = TRUE, sort.est = TRUE, value.offset = .3)
```



Obtaining the C value for the minimal adequate model

```
library(Hmisc)
```

```
somers2(binomial()$linkinv(fitted(model7)), as.numeric(gen$position) -1)
```

```
         C         Dxy           n     Missing
 0.9002869   0.8005738 482.0000000   0.0000000
```

Calculating the proportion of correctly predicted values

```
library("gmodels")
```

```
fitted <- fitted(model7)
predicted <- ifelse(fitted >= .5, 1,0)
a <- data.frame(gen, predicted)
CrossTable(gen$position, a$predicted)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
```

```
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  482


             | a$predicted
gen$position |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
        left |        67 |        36 |       103 |
             |   128.642 |    27.937 |           |
             |     0.650 |     0.350 |     0.214 |
             |     0.779 |     0.091 |           |
             |     0.139 |     0.075 |           |
-------------|-----------|-----------|-----------|
       right |        19 |       360 |       379 |
             |    34.961 |     7.593 |           |
             |     0.050 |     0.950 |     0.786 |
             |     0.221 |     0.909 |           |
             |     0.039 |     0.747 |           |
-------------|-----------|-----------|-----------|
Column Total |        86 |       396 |       482 |
             |     0.178 |     0.822 |           |
-------------|-----------|-----------|-----------|
```

Investigating whether multicollinearity is a problem for the predictors in the model

```
library(languageR)
```

```
collin.fnc(getME(model7, "X")[, -1])$cnumber
```

```
[1] 4.973687
```

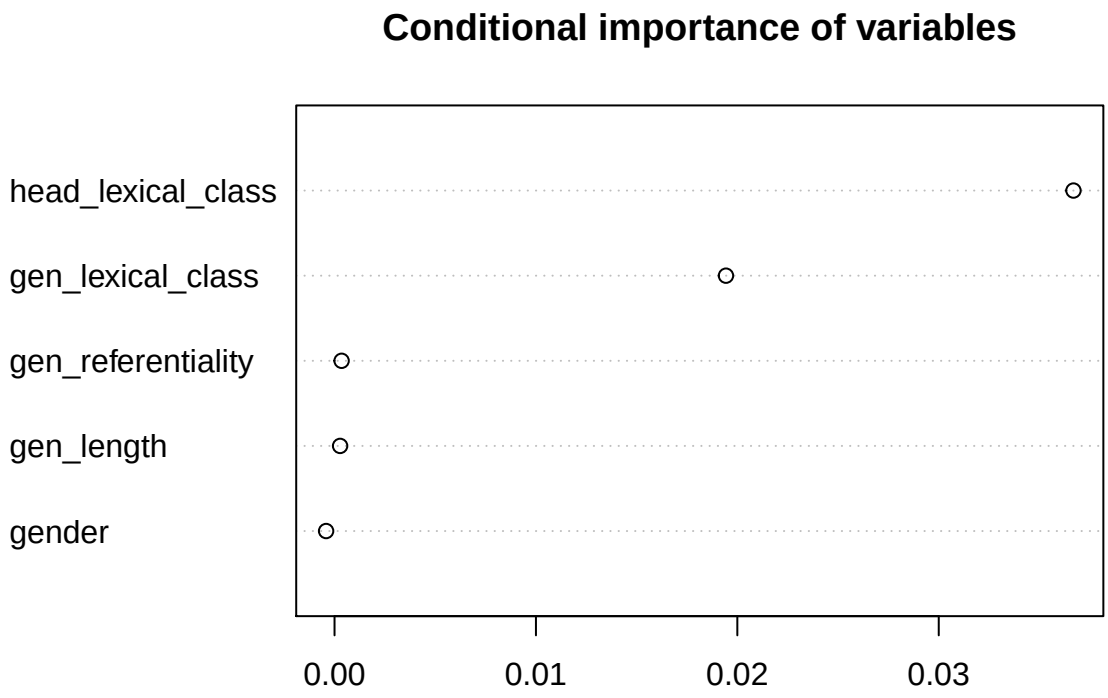
Random forest

For the set of predictors included in the minimal adequate model, we fit a random forest analysis, which can be useful when the number of observations is relatively small, but the number of predictors is large, as in our case.

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

Fit a random forest and visualize the conditional importance of variables

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



Versions of the packages used in the analysis:

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

```
     party strucchange    sandwich         zoo  modeltools     mvtnorm
   "1.3-7"     "1.5-2"     "3.0-0"     "1.8-9"    "0.2-23"     "1.1-1"
 languageR     gmodels       Hmisc     Formula    survival     lattice
   "1.5.0"    "2.18.1"     "4.5-0"     "1.2-4"    "3.2-10"   "0.20-41"
    sjPlot        lme4      Matrix     forcats     stringr       dplyr
   "2.8.7"    "1.1-26"     "1.3-2"     "0.5.1"     "1.4.0"     "1.0.5"
     purrr       readr       tidyr      tibble     ggplot2   tidyverse
   "0.3.4"     "1.4.0"     "1.1.3"     "3.1.0"     "3.3.3"     "1.3.0"
```