Code & data representation

Importing libraries:

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
library(sjstats)
library(ggplot2)
```

Reading the source data:

```
Sys.setlocale(category = "LC_ALL", locale = "English")
```

```
## [1] "LC_COLLATE=English_United States.1252;LC_CTYPE=English_United States.1252;LC_MONETARY=English_U
```

```
data <- read.csv("final.csv", encoding = 'UTF-8')
Sys.setlocale(category = "LC_ALL", locale = "Japanese")
```

```
## [1] "LC_COLLATE=Japanese_Japan.932;LC_CTYPE=Japanese_Japan.932;LC_MONETARY=Japanese_Japan.932;LC_NUM
```

Section 1: Conjugation

Testing for dependence:

```
fisher.test(table(data$conj, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$conj, data$form)
## p-value = 0.05095
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4106897 1.0196977
## sample estimates:
## odds ratio
```

```
##  0.6488755
```

Running the regression:

```
data %>%
  lm(as.numeric(form) ~ conj, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ conj, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7513 -0.6619  0.2487  0.3381  0.3381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.84063    0.07202  25.557   <2e-16 ***
## conj        -0.08936    0.04512  -1.981   0.0483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4549 on 405 degrees of freedom
## Multiple R-squared:  0.009593,   Adjusted R-squared:  0.007147
## F-statistic: 3.923 on 1 and 405 DF,  p-value: 0.04831
```

Section 2: Context postition

Testing for dependence:

```
# without ambiguous contexts
data.unamb_cont <- data[data$context_pos != "unclear", ]
fisher.test(table(data.unamb_cont$context_pos, data.unamb_cont$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data.unamb_cont$context_pos, data.unamb_cont$form)
## p-value = 0.0002458
## alternative hypothesis: two.sided
```

```
# with ambiguous contexts
fisher.test(table(data$context_pos, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$context_pos, data$form)
## p-value = 0.0004691
## alternative hypothesis: two.sided
```

Running the regression:

```
# without ambiguous contexts
data.unamb_cont %>%
```

```
  lm(as.numeric(form) ~ context_pos, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8750 -0.6622  0.1250  0.3378  0.3378
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.87500    0.04933  38.006  < 2e-16 ***
## context_posright -0.21284    0.05754  -3.699 0.000257 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4413 on 300 degrees of freedom
## Multiple R-squared:  0.04362,    Adjusted R-squared:  0.04043
## F-statistic: 13.68 on 1 and 300 DF,  p-value: 0.0002575
```

```
# with ambiguous contexts
data %>%
  lm(as.numeric(form) ~ context_pos, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8750 -0.6622  0.3333  0.3378  0.3378
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.87500    0.05029  37.283  < 2e-16 ***
## context_posright   -0.21284    0.05866  -3.628 0.000322 ***
## context_posunclear -0.20833    0.06676  -3.121 0.001933 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4498 on 404 degrees of freedom
## Multiple R-squared:  0.03396,    Adjusted R-squared:  0.02918
## F-statistic: 7.101 on 2 and 404 DF,  p-value: 0.0009311
```

Section 2.5: Conjugation + context postition

Running the regression:

```
# without ambiguous contexts
data.unamb_cont %>%
  glmer(form ~ context_pos + (1|conj),
```

```
        family = binomial(link="logit"), data = ., control = glmerControl(optimizer = "bobyqa", optCtrl
summary(gl2.5)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: form ~ context_pos + (1 | conj)
##    Data: .
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 10000))
##
##      AIC      BIC   logLik deviance df.resid
##    350.1    361.3   -172.1    344.1      299
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7352 -1.3627  0.3884  0.6907  0.7338
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  conj   (Intercept) 0.01321  0.1149
## Number of obs: 302, groups:  conj, 2
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.9524     0.3485   5.602 2.12e-08 ***
## context_posright   -1.2724     0.3669  -3.468 0.000524 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## cntxt_psrgh -0.895
```

```
icc(gl2.5)
```

```
##
## Intraclass Correlation Coefficient for Generalized linear mixed model
##
## Family : binomial (logit)
## Formula: form ~ context_pos + (1 | conj)
##
##   ICC (conj): 0.0040
```

```
# with ambiguous contexts
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + conj, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos + conj, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.9182 -0.6234  0.1661  0.2923  0.3766
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.00254    0.09118  21.962  < 2e-16 ***
## context_posright -0.21047    0.05739  -3.667  0.00029 ***
## conj             -0.08432    0.05076  -1.661  0.09774 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.44 on 299 degrees of freedom
## Multiple R-squared:  0.05236,    Adjusted R-squared:  0.04602
## F-statistic: 8.261 on 2 and 299 DF,  p-value: 0.0003221
```

Section 3: Sex

Testing for dependence:

```
fisher.test(table(data$sex, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$sex, data$form)
## p-value = 0.003389
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1982449 0.7519020
## sample estimates:
## odds ratio
##  0.3858829
```

Running the regression:

```
data %>%
  lm(as.numeric(form) ~ sex, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ sex, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7306 -0.5106  0.2694  0.2694  0.4894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73056    0.02380   72.70  < 2e-16 ***
## sexm        -0.21992    0.07005   -3.14  0.00182 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4516 on 405 degrees of freedom
## Multiple R-squared:  0.02376,    Adjusted R-squared:  0.02135
```

```
## F-statistic: 9.857 on 1 and 405 DF,  p-value: 0.001816
```

Section 3.5: Sex + context position + conjugation

Running the regression:

```
# without ambiguous contexts
data.unamb_cont %>%
  lm(as.numeric(form) ~ sex + context_pos + conj, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ sex + context_pos + conj, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9288 -0.4757  0.2719  0.3465  0.5243
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.00336    0.09050  22.138  < 2e-16 ***
## sexm              -0.17786    0.07539  -2.359  0.01895 *
## context_posright -0.20067    0.05711  -3.514  0.00051 ***
## conj              -0.07458    0.05055  -1.475  0.14116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4366 on 298 degrees of freedom
## Multiple R-squared:  0.06974,    Adjusted R-squared:  0.06038
## F-statistic: 7.447 on 3 and 298 DF,  p-value: 7.981e-05
```

```
# with ambiguous contexts
data %>%
  lm(as.numeric(form) ~ sex + context_pos + conj, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ sex + context_pos + conj, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9349 -0.5325  0.2641  0.3463  0.5649
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.01697    0.08321  24.238  < 2e-16 ***
## sexm                -0.20332    0.06930  -2.934 0.003537 **
## context_posright   -0.19902    0.05804  -3.429 0.000668 ***
## context_posunclear -0.21440    0.06593  -3.252 0.001242 **
## conj                -0.08210    0.04418  -1.858 0.063869 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.444 on 402 degrees of freedom
## Multiple R-squared:  0.06347,    Adjusted R-squared:  0.05415
## F-statistic: 6.811 on 4 and 402 DF,  p-value: 2.596e-05
```

```
# without ambiguous contexts and with conjugation as a random effect
data.unamb_cont %>%
  glmer(form ~ sex + context_pos + (1|conj),
        family = binomial(link="logit"), data = ., control = glmerControl(optimizer = "bobyqa", optCtrl

summary(gl3.5)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: form ~ sex + context_pos + (1 | conj)
##    Data: .
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 10000))
##
##      AIC      BIC   logLik deviance df.resid
##    346.7    361.6   -169.4    338.7      298
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8179 -0.9682  0.5561  0.6758  1.0329
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  conj   (Intercept) 0.005    0.07071
## Number of obs: 302, groups:  conj, 2
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.0471     0.3483   5.877 4.17e-09 ***
## sexm               -0.8485     0.3607  -2.353 0.018643 *
## context_posright   -1.2382     0.3692  -3.354 0.000796 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) sexm
## sexm        -0.144
## cntxt_psrgh -0.897 -0.020
```

```
icc(gl3.5)
```

```
##
## Intraclass Correlation Coefficient for Generalized linear mixed model
##
## Family : binomial (logit)
## Formula: form ~ sex + context_pos + (1 | conj)
##
##   ICC (conj): 0.0015
```

Section 4: Length

Testing for dependence:

```
fisher.test(table(data$length, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$length, data$form)
## p-value = 0.8113
## alternative hypothesis: two.sided
```

Section 4.5: Length + context position + sex + conjugation

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + length + conj, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + length +
##     conj, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0011 -0.5106  0.1930  0.3090  0.6249
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.84627    0.10649  17.337  < 2e-16 ***
## context_posright -0.19403    0.05656  -3.431 0.000687 ***
## sexm             -0.18037    0.07459  -2.418 0.016204 *
## length            0.13545    0.04972   2.724 0.006823 **
## conj             -0.25155    0.08198  -3.068 0.002350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.432 on 297 degrees of freedom
## Multiple R-squared:  0.09242,    Adjusted R-squared:  0.0802
## F-statistic: 7.561 on 4 and 297 DF,  p-value: 8.199e-06
```

Section 5: Part of speech

Testing for dependence:

```
fisher.test(table(data$part_of_speech, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$part_of_speech, data$form)
## p-value = 3.08e-08
## alternative hypothesis: two.sided
```

Running the regression:

```
data %>%
  lm(as.numeric(form) ~ part_of_speech, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ part_of_speech, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9167 -0.5698  0.2253  0.3277  0.4302
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.77473    0.03250  54.606  < 2e-16 ***
## part_of_speechadv   0.22527    0.09897   2.276   0.0234 *
## part_of_speechpart  0.14194    0.09522   1.491   0.1368
## part_of_speechv    -0.20489    0.04615  -4.439 1.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4385 on 403 degrees of freedom
## Multiple R-squared:  0.08444,    Adjusted R-squared:  0.07762
## F-statistic: 12.39 on 3 and 403 DF,  p-value: 9.121e-08
```

Section 5.5: Part of speech + context position + sex + conjugation

Running the regression:

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + part_of_speech + conj, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + part_of_speech +
##     conj, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9470 -0.4451  0.1056  0.3019  0.6075
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.99955    0.08980  22.267   <2e-16 ***
## context_posright   -0.19626    0.05657  -3.469   0.0006 ***
## sexm               -0.17215    0.07316  -2.353   0.0193 *
## part_of_speechadv   0.28242    0.11371   2.484   0.0136 *
## part_of_speechpart  0.16051    0.09891   1.623   0.1057
## part_of_speechv    -0.13348    0.05355  -2.492   0.0132 *
## conj               -0.05257    0.05009  -1.050   0.2948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4235 on 295 degrees of freedom
## Multiple R-squared:  0.1335, Adjusted R-squared:  0.1159
## F-statistic: 7.578 on 6 and 295 DF,  p-value: 1.42e-07
```

Section 6: Positivity & negativity

Testing for dependence:

```
fisher.test(table(data$pos_neg, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$pos_neg, data$form)
## p-value = 0.0002133
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2739981 0.6934952
## sample estimates:
## odds ratio
##  0.4379414
```

Running the regression:

```
data %>%
  lm(as.numeric(form) ~ pos_neg, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ pos_neg, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7892 -0.6207  0.2108  0.3793  0.3793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.78922    0.03145  56.887  < 2e-16 ***
## pos_negpos  -0.16853    0.04453  -3.784 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4492 on 405 degrees of freedom
## Multiple R-squared:  0.03415,    Adjusted R-squared:  0.03177
## F-statistic: 14.32 on 1 and 405 DF,  p-value: 0.0001775
```

Section 6.5: Positivity & negativity + part of speech + context position + sex + conjugation

Running the regression:

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + part_of_speech + pos_neg + conj, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + part_of_speech +
##     pos_neg + conj, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9449 -0.4549  0.1090  0.3046  0.5991
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.99892    0.08979  22.263  < 2e-16 ***
## context_posright   -0.19561    0.05656  -3.458 0.000624 ***
## sexm               -0.16760    0.07328  -2.287 0.022901 *
## part_of_speechadv   0.28426    0.11371   2.500 0.012967 *
## part_of_speechpart -0.06236    0.23482  -0.266 0.790764
## part_of_speechv    -0.35100    0.21465  -1.635 0.103071
## pos_negpos          0.22411    0.21416   1.046 0.296214
## conj               -0.05397    0.05010  -1.077 0.282275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4235 on 294 degrees of freedom
## Multiple R-squared:  0.1368, Adjusted R-squared:  0.1162
## F-statistic: 6.654 on 7 and 294 DF,  p-value: 2.557e-07
```

Section 7: Past tense

Testing for dependence:

```
fisher.test(table(data$past_tense, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$past_tense, data$form)
## p-value = 0.004539
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2964611 0.8240989
## sample estimates:
## odds ratio
##  0.4932979
```

Running the regression:

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ past_tense, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ past_tense, data = .)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -0.7425 -0.6377  0.2575  0.2575  0.3623
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.74249    0.02942  59.233   <2e-16 ***
## past_tenseyes -0.10481    0.06154  -1.703   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.449 on 300 degrees of freedom
## Multiple R-squared:  0.009575,   Adjusted R-squared:  0.006273
## F-statistic:   2.9 on 1 and 300 DF,  p-value: 0.08961
```

Section 7.5: Past tense + context position + sex + part of speech + conjugation + positivity & negativity

Running the regression:

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + part_of_speech + pos_neg + conj + past_tense, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + part_of_speech +
##     pos_neg + conj + past_tense, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9493 -0.4357  0.1582  0.2971  0.5643
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.99150    0.08973  22.194  < 2e-16 ***
## context_posright   -0.19430    0.05645  -3.442 0.000662 ***
## sexm               -0.16375    0.07317  -2.238 0.025973 *
## part_of_speechadv   0.29861    0.11387   2.622 0.009189 **
## part_of_speechpart -0.07394    0.23445  -0.315 0.752688
## part_of_speechv    -0.37704    0.21490  -1.754 0.080393 .
## pos_negpos          0.26373    0.21534   1.225 0.221672
## conj               -0.04222    0.05061  -0.834 0.404844
## past_tenseyes      -0.09085    0.06065  -1.498 0.135184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4226 on 293 degrees of freedom
## Multiple R-squared:  0.1433, Adjusted R-squared:  0.1199
## F-statistic: 6.127 on 8 and 293 DF,  p-value: 2.616e-07
```

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + part_of_speech + pos_neg + past_tense, data = .) %>%
  summary()
```

```
##
```

```
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + part_of_speech +
##     pos_neg + past_tense, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9371 -0.4504  0.1616  0.2859  0.5496
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.93073    0.05236  36.877  < 2e-16 ***
## context_posright   -0.19412    0.05642  -3.440 0.000665 ***
## sexm               -0.16830    0.07293  -2.308 0.021709 *
## part_of_speechadv   0.29101    0.11345   2.565 0.010810 *
## part_of_speechpart -0.06196    0.23389  -0.265 0.791277
## part_of_speechv    -0.38038    0.21475  -1.771 0.077555 .
## pos_negpos          0.26247    0.21523   1.219 0.223632
## past_tenseyes      -0.09870    0.05988  -1.648 0.100385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4224 on 294 degrees of freedom
## Multiple R-squared:  0.1413, Adjusted R-squared:  0.1208
## F-statistic:  6.91 on 7 and 294 DF,  p-value: 1.275e-07
```

Section 8: Provisional & conditional form

Testing for dependence:

```
fisher.test(table(data$prov_cond, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$prov_cond, data$form)
## p-value = 1.533e-12
## alternative hypothesis: two.sided
```

Running the regression:

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ prov_cond, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ prov_cond, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7658 -0.1500  0.2342  0.2342  0.8500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.76580    0.02588  68.229  < 2e-16 ***
```

```
## prov_condsep -0.61580    0.09838  -6.259 1.34e-09 ***
## prov_condvf  -0.15041    0.12054  -1.248    0.213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4245 on 299 degrees of freedom
## Multiple R-squared:  0.1179, Adjusted R-squared:  0.112
## F-statistic: 19.99 on 2 and 299 DF,  p-value: 7.101e-09
```

Section 8.5: Provisional & conditional form + context position + sex + part of speech + past tense + positivity & negativity

Running the regression:

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + part_of_speech + pos_neg + past_tense + prov_cond, data = .)
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + part_of_speech +
##     pos_neg + past_tense + prov_cond, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9386 -0.1935  0.1467  0.2878  0.8065
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.91650    0.05077  37.747  < 2e-16 ***
## context_posright   -0.15626    0.05525  -2.829   0.0050 **
## sexm               -0.16882    0.07020  -2.405   0.0168 *
## part_of_speechadv   0.26800    0.10938   2.450   0.0149 *
## part_of_speechpart -0.12233    0.22567  -0.542   0.5882
## part_of_speechv    -0.37876    0.20677  -1.832   0.0680 .
## pos_negpos          0.30070    0.20735   1.450   0.1481
## past_tenseyes      -0.08534    0.05920  -1.442   0.1505
## prov_condsep       -0.48869    0.09853  -4.960  1.2e-06 ***
## prov_condvf         0.01826    0.12331   0.148   0.8824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4065 on 292 degrees of freedom
## Multiple R-squared:  0.2099, Adjusted R-squared:  0.1856
## F-statistic:  8.62 on 9 and 292 DF,  p-value: 1.823e-11
```

Section 9: Question

Testing for dependence:

```
fisher.test(table(data$quest, data$form))
```

```
##
##  Fisher's Exact Test for Count Data
##
```

```
## data:  table(data$quest, data$form)
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4450427 3.3132937
## sample estimates:
## odds ratio
##   1.144078
```

Section 9.5: Question + provisional & conditional form + context position + sex + part of speech + past tense + positivity & negativity

Running the regression:

```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + part_of_speech + pos_neg + past_tense + prov_cond + quest, d
  summary()
```

```
##
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + part_of_speech +
##     pos_neg + past_tense + prov_cond + quest, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9308 -0.1905  0.1509  0.2686  0.8095
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.91914    0.05103  37.608  < 2e-16 ***
## context_posright    -0.16182    0.05611  -2.884  0.00422 **
## sexm                -0.16661    0.07038  -2.367  0.01857 *
## part_of_speechadv    0.26534    0.10960   2.421  0.01609 *
## part_of_speechpart -0.11832    0.22602  -0.523  0.60103
## part_of_speechv     -0.37933    0.20700  -1.832  0.06790 .
## pos_negpos           0.29177    0.20814   1.402  0.16205
## past_tenseyes       -0.08171    0.05959  -1.371  0.17133
## prov_condsep        -0.47931    0.09993  -4.796 2.58e-06 ***
## prov_condvf          0.02355    0.12378   0.190  0.84924
## questyes             0.05544    0.09459   0.586  0.55823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.407 on 291 degrees of freedom
## Multiple R-squared:  0.2108, Adjusted R-squared:  0.1837
## F-statistic: 7.774 on 10 and 291 DF,  p-value: 4.784e-11
```

Final model:

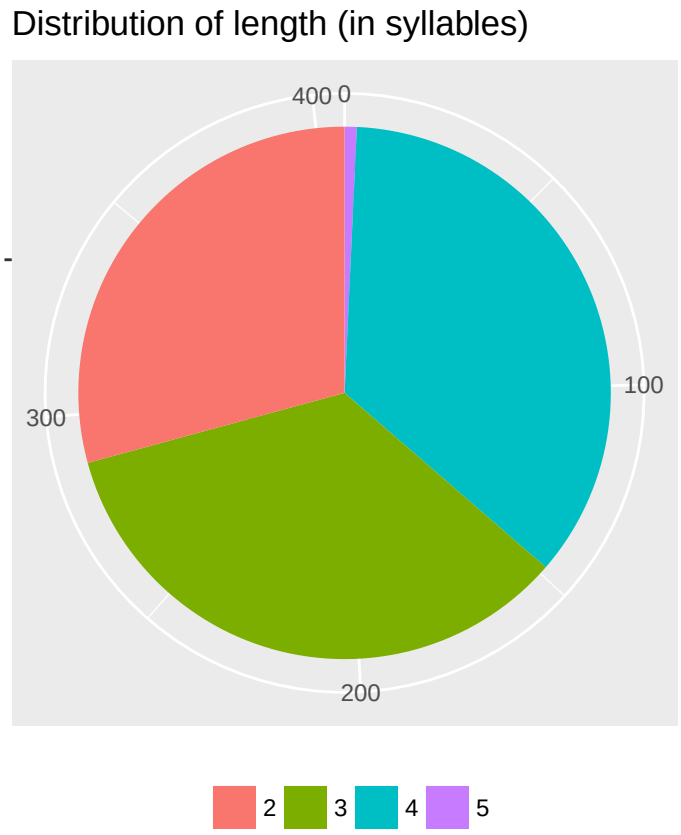
```
data.unamb_cont %>%
  lm(as.numeric(form) ~ context_pos + sex + part_of_speech + pos_neg + past_tense + prov_cond, data = .)
  summary()
```

```
##
```

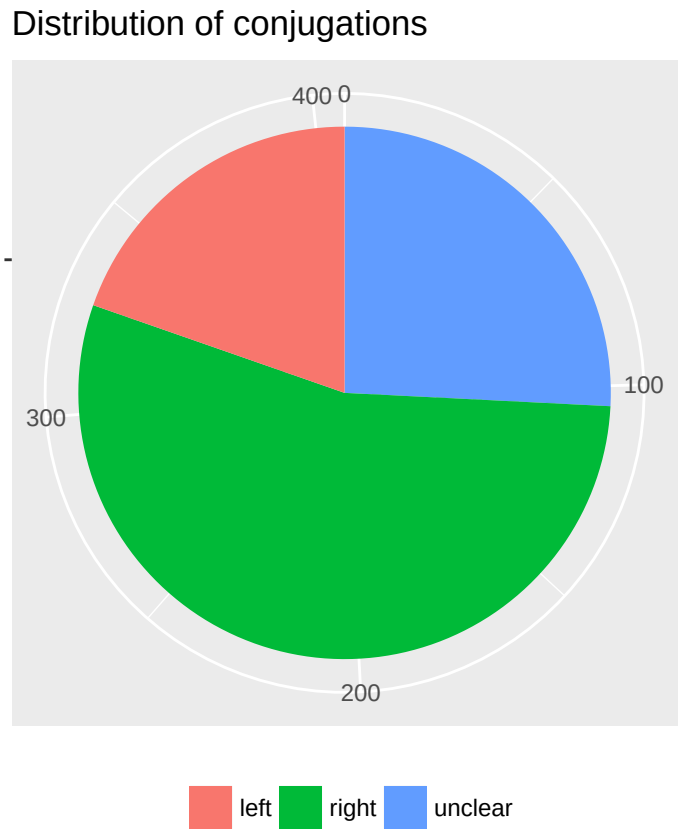
```
## Call:
## lm(formula = as.numeric(form) ~ context_pos + sex + part_of_speech +
##     pos_neg + past_tense + prov_cond, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9386 -0.1935  0.1467  0.2878  0.8065
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.91650    0.05077  37.747  < 2e-16 ***
## context_posright   -0.15626    0.05525  -2.829   0.0050 **
## sexm               -0.16882    0.07020  -2.405   0.0168 *
## part_of_speechadv   0.26800    0.10938   2.450   0.0149 *
## part_of_speechpart -0.12233    0.22567  -0.542   0.5882
## part_of_speechv    -0.37876    0.20677  -1.832   0.0680 .
## pos_negpos          0.30070    0.20735   1.450   0.1481
## past_tenseyes      -0.08534    0.05920  -1.442   0.1505
## prov_condsep       -0.48869    0.09853  -4.960  1.2e-06 ***
## prov_condvf         0.01826    0.12331   0.148   0.8824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4065 on 292 degrees of freedom
## Multiple R-squared:  0.2099, Adjusted R-squared:  0.1856
## F-statistic:  8.62 on 9 and 292 DF,  p-value: 1.823e-11
```

Visualizing data

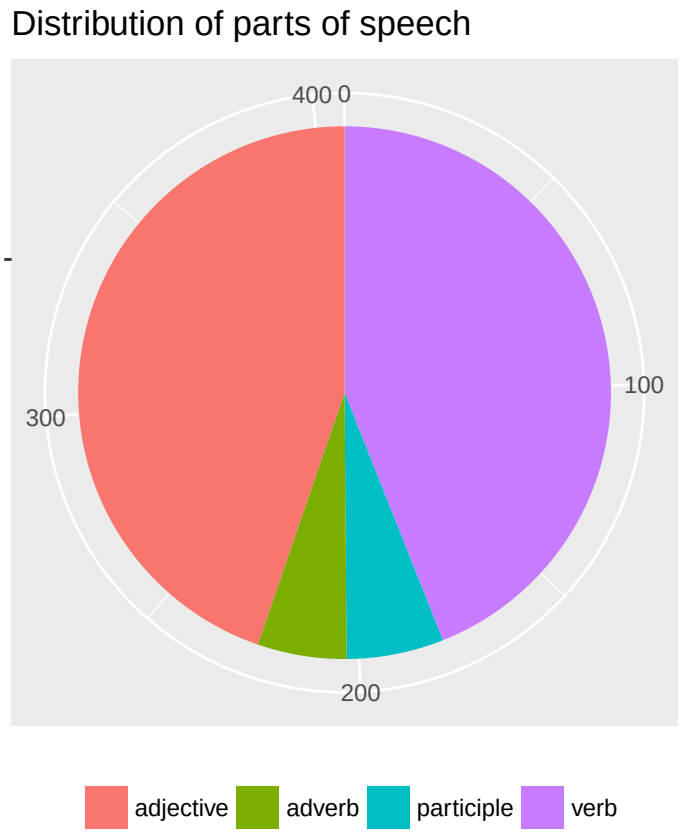
```
data %>%
  count(length) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(length))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  labs(title = "Distribution of length (in syllables)")
```



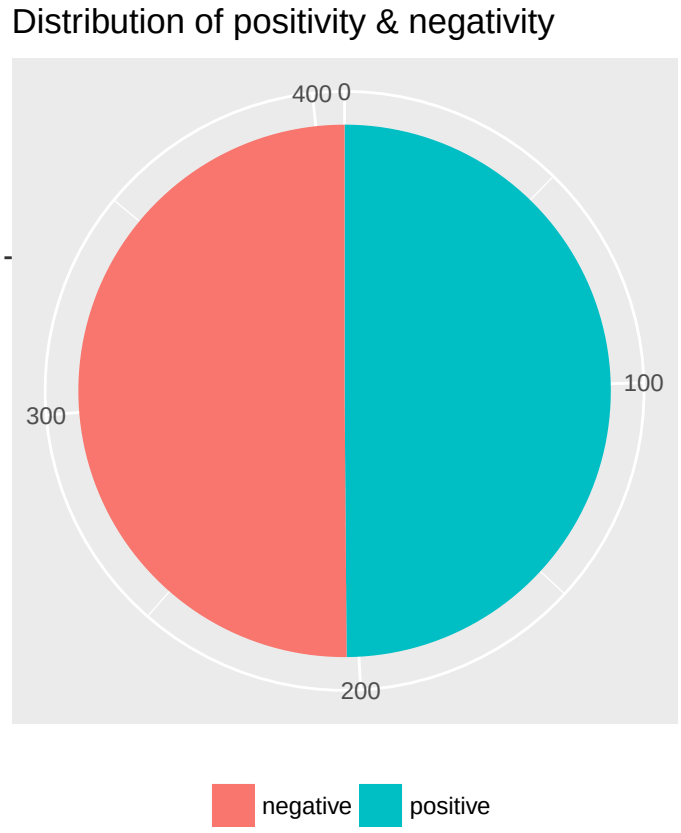

```
data %>%
  count(context_pos) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(context_pos))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  labs(title = "Distribution of conjugations")
```



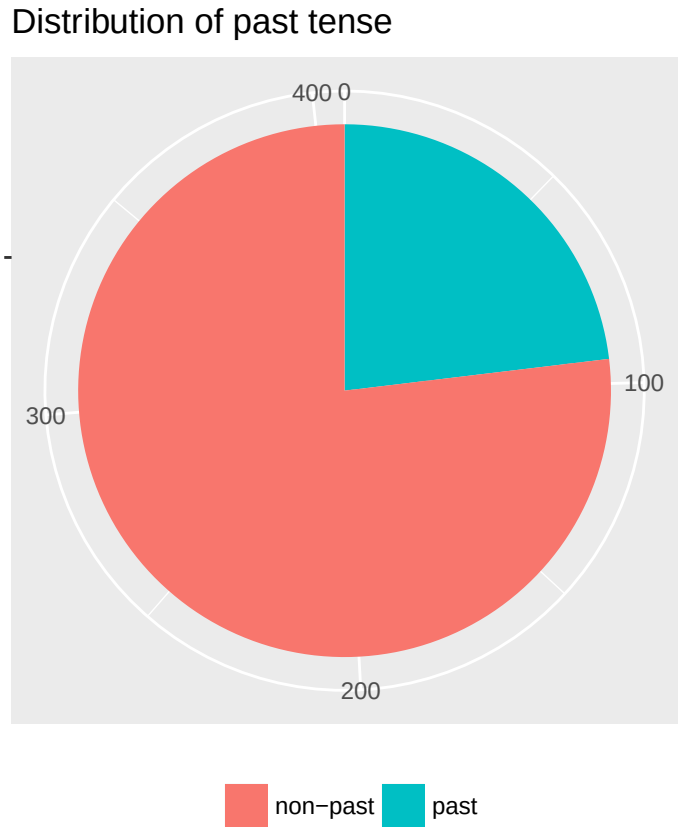

```
data %>%
  count(part_of_speech) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(part_of_speech))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  scale_fill_discrete(labels = c("adjective", "adverb", "participle", "verb")) +
  labs(title = "Distribution of parts of speech")
```



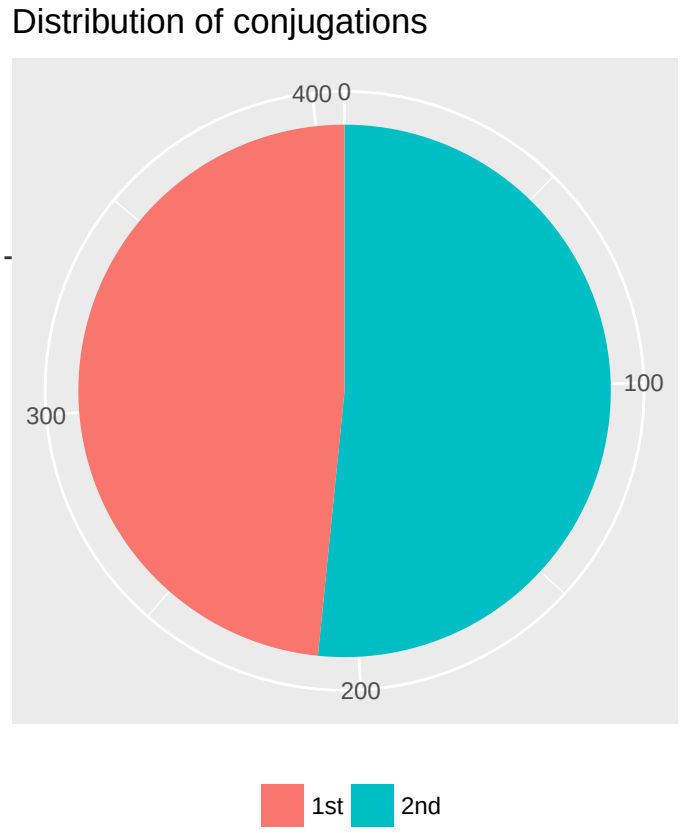

```
data %>%
  count(pos_neg) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(pos_neg))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  scale_fill_discrete(labels = c("negative", "positive")) +
  labs(title = "Distribution of positivity & negativity")
```



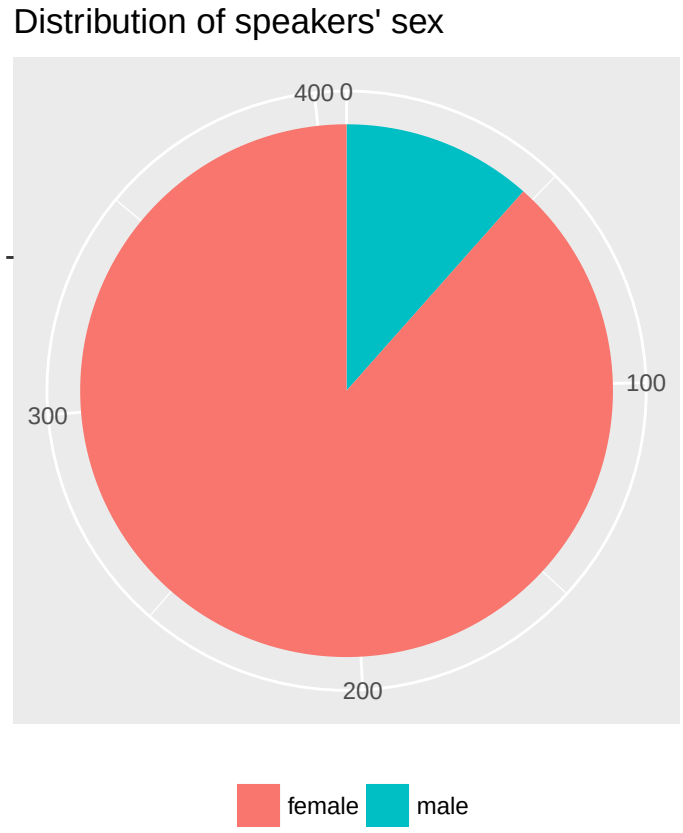

```
data %>%
  count(past_tense) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(past_tense))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  scale_fill_discrete(labels = c("non-past", "past")) +
  labs(title = "Distribution of past tense")
```



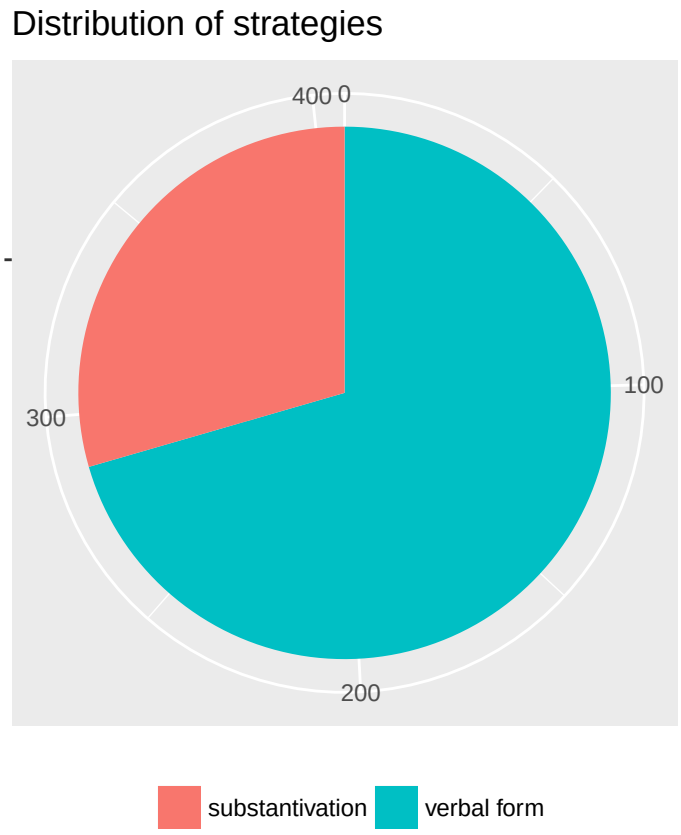

```
data %>%
  count(conj) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(conj))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  scale_fill_discrete(labels = c("1st", "2nd")) +
  labs(title = "Distribution of conjugations")
```



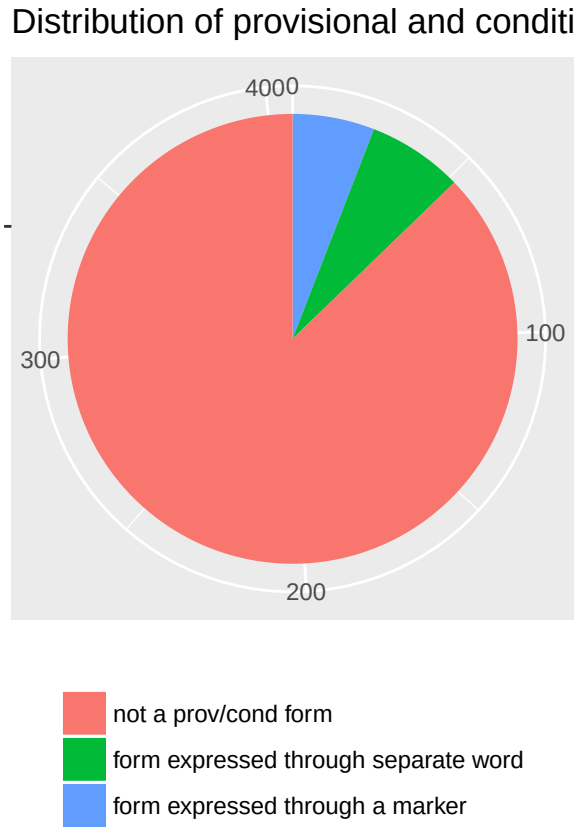

```
data %>%
  count(sex) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(sex))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  scale_fill_discrete(labels = c("female", "male")) +
  labs(title = "Distribution of speakers' sex")
```



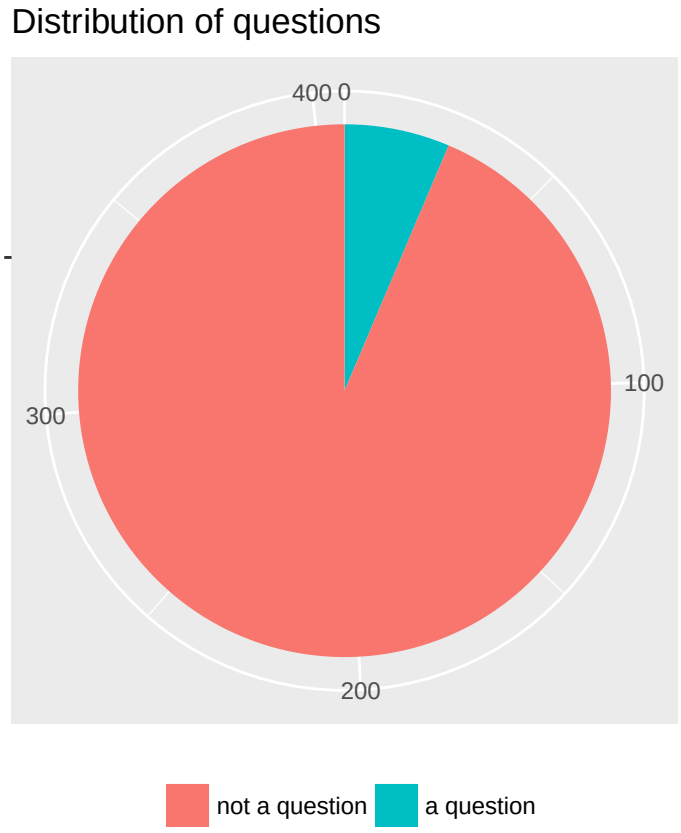

```
data %>%
  count(form) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(form))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  scale_fill_discrete(labels = c("substantivation", "verbal form")) +
  labs(title = "Distribution of strategies")
```




```
data %>%
  count(prov_cond) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(prov_cond))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "vertical", lege
  scale_fill_discrete(labels = c("not a prov/cond form", "form expressed through separate word", "form
  labs(title = "Distribution of provisional and conditional forms")
```




```
data %>%
  count(quest) %>%
  ggplot(., aes(x="", y=n, fill=as.factor(quest))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  theme(axis.title=element_blank(), legend.title = element_blank(), legend.direction = "horizontal", leg
  scale_fill_discrete(labels = c("not a question", "a question")) +
  labs(title = "Distribution of questions")
```


Distribution of questions
400 0
100
200
300
not a question
a question