Strategies of choice between substantivation and regular verb usage in the context of ability in Japanese

Ekaterina Birjukova, Alexandra Efimova

Hypothesis

There are two different ways of expressing ability to do something in Japanese: one through verb substantivation, another through a verbal form, and there is no difference in meaning between them. We believe that form choice depends on conjugation (we will consider only the first and the second one, leaving out the third one as its ability-expressing strategy is somewhat different), length of the verb that holds the main semantics (the number of syllables that it comprises) and a number of other factors.

Research design

Our research is based on materials from parallel Japanese-English corpora at https://context.reverso.net/ and http://www.manythings.org/corpus/. We also used data from https://www5.atwiki.jp/hmiku/ lyrics database for more precise statistical testing.

Null hypothesis is as follows:

- Choice strategy for expressing ability in Japanese verbs is completely random and does not depend on any of the mentioned factors.

Alternative hypothesis:

- There is a more or less clear pattern that allows actually calling this choice a strategy.

Our data

All data has been collected with web scraping script written in Python. There are currently several hundred examples from the parallel corpora and approximately thirty thousand texts from the database. The data can be accessed on GitHub.

From all the data we have collected, we have gathered information about 52 randomly picked verbs. Each line of our data table contains following properties:

1. **inf** *[character string]* – verb in its infinitive form;
2. **length** *[non-negative integer number]* – length of the verb's infinitive form in syllables;
3. **tr** *[character string]* – translation of the verb;
4. **context_pos** *[factor]* – position of the verb in the context:

- *right*;
- *left*;
- *unclear*;

5. **part_of_speech** *[factor]* – the verb's grammatical type of speech:

- *adj* – adjective;
- *adv* – verbal adverb or state adverb;
- *v* – verb;
- *part* – participle;

6. **pos_neg** *[factor]* – whether the verb or the verb to which the target verb form is attached has or has not negation:

- *pos* – not negated;
- *neg* – negated;

7. **past_tense** *[factor]* – whether the verb or the verb to which the target verb form is attached is in past tense or not:

- *yes*;
- *no*;

8. **conj** *[factor]* – conjugation of the verb:

- *1*;
- *2*;

9. **sex** *[factor]* – sex of the speaker:

- *f* – female;
- *m* – male;

10. **form** *[factor]* – type of the strategy:

- *vf* – verbal form;
- *subst* – substantivation;

11. **prov_cond** *[factor]* – whether the verb is in the provisional or conditional form:

- *no* – not in the provisional or conditional form;
- *vf* – the provisional or conditional form is expressed through a marker;
- *sep* – the provisional or conditional form is expressed through a separate word;

12. **quest** *[factor]* – whether the verb or the verb to which the target verb form is attached has a question marker:

- *yes*;
- *no*.

Context position

Examples of right context position:

(1) 僕　は　歌う
boku wa utau
I THEME to.sing
'I sing'

(2) 僕　は　歌　を　歌う
boku wa uta wo utau
I THEME song ACC to.sing
'I sing a song'

In these examples the verb "to sing" is located to the right from both the subject and the object and denotes an active action, as opposed to its left position:

(3) 僕 が 歌う 歌
boku ga utau uta
I NOM to.sing song
'a song that I sing'

(4) 歌 を 歌う 僕
uta wo utau boku
song ACC to.sing I
'I who sing a song'

In examples (3) and (4) the verb expresses a property, while the actual verbal action of the sentence will be expressed through some other verb. It is often impossible to establish the context type, because Japanese allows for omission, which could be reconstructed in both ways:

(5) 歌わ-ない 踊ら-ない
utawanai odoranai
sing-NEG to.dance-NEG

(5) could be translated both as

(6) '[僕/君/... は/が] 歌わない踊らない'

'[I/you/...] don' t sing, [I/you/...] don' t dance'

and

(7) '歌わない踊らない [僕/君/...]'

'[I/you/...] who don' t sing, [I/you/...] who don' t dance'

In such ambiguous cases we marked the context type as "unclear".

Grammatical parts of speech

Verbal adverbs and participles are clearly related to the verbs and could be assumed to be verbal forms. In Japanese, simple verbal negation in casual speech form behaves like an adjective: it has a past tense (8 and 9), a special form that non-final adjectives receive when there is more than one adjective related to the same word (10):

(8) 見-え-ない 闇
mi-e-nai yami
sing-NEG to.dance-NEG
'darkness in which nothing can be seen'

(9) 深い 闇
fukai yami
deep darkness
'deep darkness'

(10) 見-え-な-くて 深-くて 怖い 闇
mi-e-na-kute fuka-kute kowai yami
to.see-POT-NEG-ENUM deep-ENUM scary darkness
'deep scary darkness in which nothing can be seen'

For these reasons, the grammatical category for such verbal forms has been marked as "adjective".

Entries marked as "verbs" are regular verb forms, like the verb "to sing" in examples (1) and (2).

Positivity and negativity

As we have already shown, adjective-like verbal forms have their own negativity marker. So do state and verbal adverbs:

(11) 疲れ-て　　立-て-な-く-なる
tsukare-te　　ta-te-naku-naru
to.become.tired-PART　to.stand-POT-NEG-ADV-become
'to become unable to stand [due to] having become tired'

(12) 答え　分-れ-ず
kotae　waka-re-zu
answer　to.understand-POT-NEG.ADV
'not being able to understand the answer'

Participles' negation is the one of the verb they are attached to:

(13) 言-え-て　　い-ない　言葉
i-e-te　　i-nai　kotoba
to.say-POT-PART　to.be-NEG　word
'words that cannot be said'

Past tense

There is no future tense in Japanese, so only past and non-past tense has been marked.

Conjugations

In this paper, we refer to the 2 conjugations based on verbal form of strategy expression:

(14) 言-う → 言-える
i-u　　i-eru
'to say → to be able to say'

(15) 食べ-る → 食べ-られる
tabe-ru　　tabe-rareru
'to eat → to be able to eat'

where the first is an example of a first conjugation verb, the second – of a second conjugation verb. Very conveniently, this strategy depends on the number of syllables that are not incorporated in the verbal stem: if there are any, then it is the second conjugation.

Provisional and conditional forms

There are two ways to put a verb into a provisional or conditional form: to add a special marker or a separate word "if". Provisional form of the verb means that something can or will be able to happen, provided that some requirement has been met:

(16) 漢字 を　覚え-られ-たら　　書く
kanji wo　oboe-rare-tara　　kaku
kanji ACC　to.learn-POT-PROV　to.write-POT
'[I] write kanji if (as soon as) I am able to learn them'

(17) 漢字 を　覚え-られる なら 書く
kanji wo　oboe-rare-ru　nara　kaku
kanji ACC　to.learn-POT　if　to.write-POT
'[I] write kanji if (as soon as) I am able to learn them'

Conditional form of the verb means that something can be done if some requirement is met:

(18) 漢字 を 覚え-られ-ば 書く
kanji wo oboe-rare-tara kaku
kanji ACC to.learn-POT-COND to.write-POT
'if [I] can learn kanji, I write them'

The boundary between the two is somewhat vague, so we decided to join them.

Methods of analysis

We are going to use Fisher's test to check our preliminary assumptions regarding the importance of the factors for the choice of the strategy.

We are also going to use logistic regression to see which factors are more important and ultimately to build a model that describes our strategies in the best way possible. First, we will see how much of the variance a factor can explain on its own. Then, we will incorporate it into a multi-variable model and see whether it has much impact on the variance of that model. If it does, we will include it into our model.

We are using Nagelkerke's R^2 to assess our models through the amount of variance that they are able to explain.

Code and results

Importing libraries:

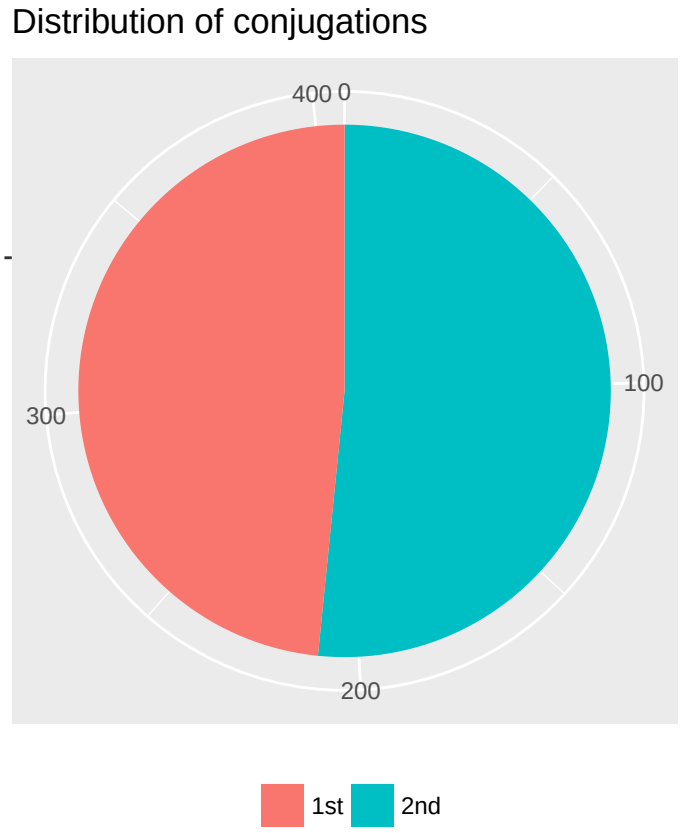
```
library(tidyverse)
library(rms)
library(lme4)
library(sjstats)
library(ggplot2)
```

Reading the source data:

```
# for some reason table layout breaks if the locale is not English
Sys.setlocale(category = "LC_ALL", locale = "English")

data <- read.csv("final.csv", encoding = 'UTF-8')
```

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:

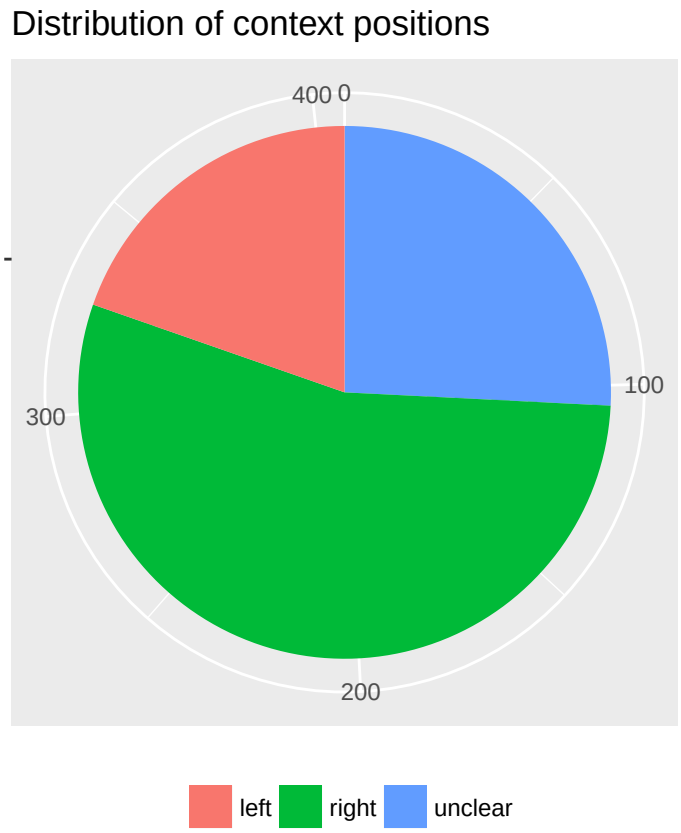
```
data_d <- datadist(data)
options(datadist="data_d")

data %>%
  lrm(form ~ conj, data = .) %>%
  print()
```

```
## Logistic Regression Model
##
```

```
##  lrm(formula = form ~ conj, data = .)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes          Indexes
##  Obs           407    LR chi2      3.92    R2       0.014    C       0.554
##   subst        120    d.f.            1    g        0.217    Dxy     0.107
##   vf           287    Pr(> chi2) 0.0476    gr       1.242    gamma   0.213
##  max |deriv| 1e-08                         gp       0.045    tau-a   0.045
##                                            Brier    0.206
##
##            Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept  1.5390 0.3605  4.27  <0.0001
##  conj      -0.4336 0.2201 -1.97  0.0488
##
```

2. Context position



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$context_pos <- droplevels(data.unamb_cont$context_pos)
data_d.unamb <- datadist(data.unamb_cont)
options(datadist="data_d.unamb")
data.unamb_cont %>%
  lrm(form ~ context_pos, data = .) %>%
  print()
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos, data = .)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes           Indexes
##  Obs           302    LR chi2     14.71    R2       0.068    C       0.602
##   subst         85    d.f.            1    g        0.497    Dxy     0.205
##   vf           217    Pr(> chi2) 0.0001    gr       1.644    gamma   0.562
##  max |deriv| 5e-08                         gp       0.083    tau-a   0.083
##                                            Brier    0.193
##
##                     Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept           1.9459 0.3381  5.76  <0.0001
##  context_pos=right -1.2730 0.3666 -3.47  0.0005
##
```

```
options(datadist="data_d")
# with ambiguous contexts
data %>%
  lrm(form ~ context_pos, data = .) %>%
  print()
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos, data = .)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes           Indexes
##  Obs           407    LR chi2     15.70    R2       0.054    C       0.582
##   subst        120    d.f.            2    g        0.407    Dxy     0.164
##   vf           287    Pr(> chi2) 0.0004    gr       1.502    gamma   0.278
##  max |deriv| 1e-07                         gp       0.068    tau-a   0.068
```

```
##                                           Brier    0.201
##
##                      Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept            1.9459 0.3381  5.76  <0.0001
##  context_pos=right   -1.2730 0.3666 -3.47  0.0005
##  context_pos=unclear -1.2528 0.3964 -3.16  0.0016
##
```

Unsurprisingly, inclusion of ambiguous context produces worse results. From now on we are excluding it from all computations.

Trying a mixed model and a simple multi-variable model:

```
# mixed model
data.unamb_cont %>%
  glmer(form ~ context_pos + (1|conj),
        family = binomial(link="logit"), data = .,
        control = glmerControl(optimizer = "bobyqa",
                               optCtrl = list(maxfun=10000))) -> gl2.5
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
```

We have not been able to assess the mixed model's R^2, but the ICC for the random effect shows that it is quite useless. From now on we will only use multi-variable models with fixed effects.

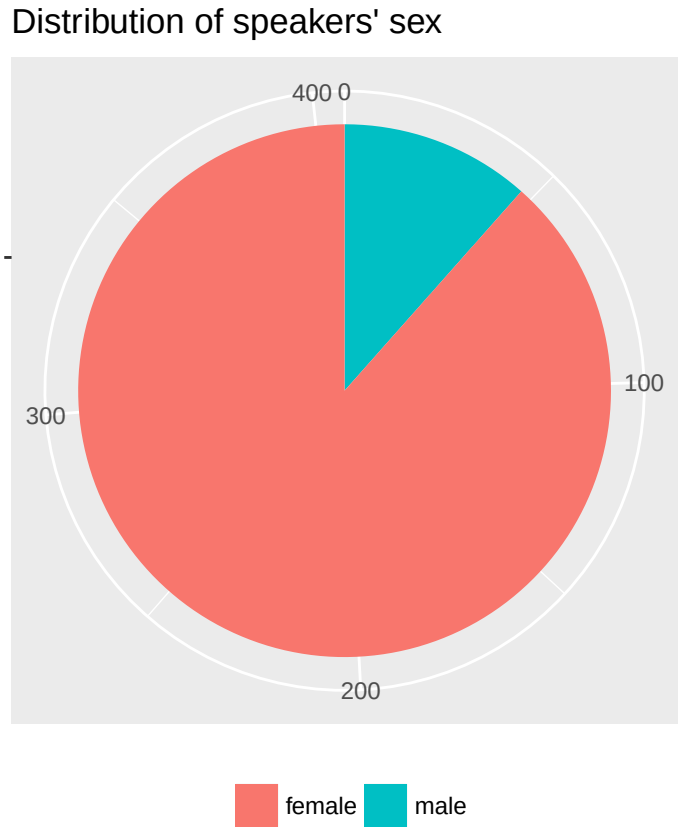
```
# model with all fixed variables
data.unamb_cont %>%
  lrm(form ~ context_pos + conj, data = .) %>%
  print()
```

```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + conj, data = .)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes           Indexes
##  Obs           302    LR chi2     17.49    R2       0.081    C       0.641
##   subst         85    d.f.            2    g        0.636    Dxy     0.282
##   vf           217    Pr(> chi2) 0.0002    gr       1.889    gamma   0.413
##  max |deriv| 1e-07                         gp       0.111    tau-a   0.114
##                                            Brier    0.191
##
##                    Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept          2.6317 0.5434  4.84  <0.0001
##  context_pos=right -1.2708 0.3678 -3.46  0.0005
##  conj              -0.4414 0.2664 -1.66  0.0975
##
```

We are keeping the model with fixed effects that we have obtained here and will be refer to it in the future as "best model".

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.unamb_cont %>%
  lrm(form ~ sex, data = .) %>%
  print()
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ sex, data = .)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
```

```
##                          Ratio Test           Indexes           Indexes
##  Obs           302    LR chi2      6.63    R2       0.031    C       0.557
##   subst         85    d.f.            1    g        0.207    Dxy     0.115
##   vf           217    Pr(> chi2) 0.0100    gr       1.231    gamma   0.430
##  max |deriv| 4e-13                         gp       0.047    tau-a   0.047
##                                            Brier    0.197
##
##            Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept  1.0734 0.1415  7.59  <0.0001
##  sex=m     -0.9193 0.3510 -2.62  0.0088
##
```

Incorporating into the best model:

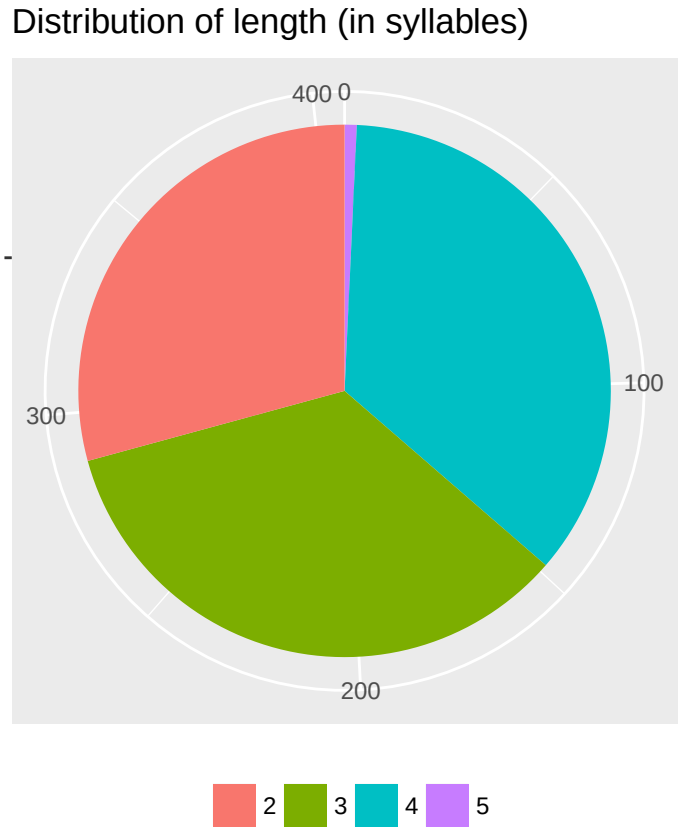
```
data.unamb_cont %>%
  lrm(form ~ sex + context_pos + conj, data = .) %>%
  print()
```

```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ sex + context_pos + conj, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
##  Obs           302    LR chi2      22.52    R2       0.103    C       0.662
##   subst         85    d.f.             3    g        0.727    Dxy     0.323
##   vf           217    Pr(> chi2) <0.0001    gr       2.069    gamma   0.420
##  max |deriv| 3e-07                         gp       0.130    tau-a   0.131
##                                            Brier    0.188
##
##                     Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept           2.6794 0.5518  4.86  <0.0001
##  sex=m              -0.8179 0.3602 -2.27  0.0232
##  context_pos=right -1.2432 0.3704 -3.36  0.0008
##  conj               -0.4077 0.2693 -1.51  0.1301
##
```

Good gain in R^2, updating the best model.

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
```

No need to even run a regression – it will not show anything.

Incorporating into the best model:

```
data.unamb_cont %>%
  lrm(form ~ context_pos + sex + length + conj, data = .) %>%
  print()
```

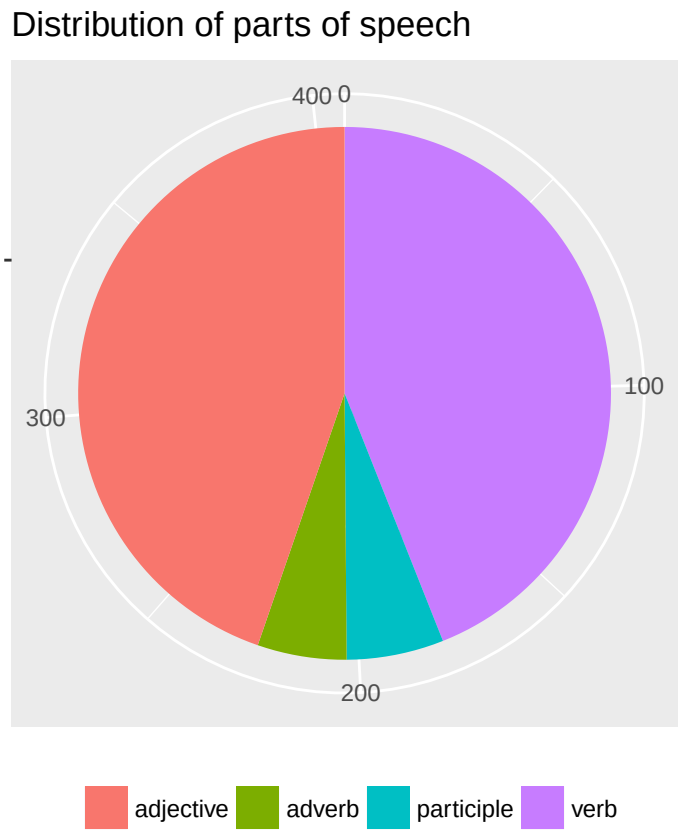
```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + sex + length + conj, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
```

```
##                          Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      30.10    R2       0.136    C       0.686
##   subst         85    d.f.             4    g        0.865    Dxy     0.373
##   vf           217    Pr(> chi2) <0.0001    gr       2.376    gamma   0.428
##  max |deriv| 2e-06                          gp       0.152    tau-a   0.151
##                                             Brier    0.183
##
##                    Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept          1.9065 0.6137  3.11  0.0019
##  context_pos=right -1.2426 0.3752 -3.31  0.0009
##  sex=m             -0.8641 0.3664 -2.36  0.0183
##  length             0.7435 0.2762  2.69  0.0071
##  conj              -1.3970 0.4659 -3.00  0.0027
##
```

Good gain in R^2, updating the best model.

5. Parts 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.unamb_cont %>%
  lrm(form ~ part_of_speech, data = .) %>%
  print()
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ part_of_speech, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      27.01    R2       0.123    C       0.657
##   subst         85    d.f.             3    g        1.381    Dxy     0.315
##   vf           217    Pr(> chi2) <0.0001    gr       3.980    gamma   0.522
##  max |deriv| 0.001                          gp       0.128    tau-a   0.128
##                                             Brier    0.187
##
##                      Coef    S.E.    Wald Z Pr(>|Z|)
##  Intercept            1.2629  0.2140  5.90  <0.0001
##  part_of_speech=adv   8.2140 28.5680  0.29  0.7737
##  part_of_speech=part  1.0397  0.7719  1.35  0.1780
##  part_of_speech=v    -0.8635  0.2760 -3.13  0.0018
##
```

Incorporating into the best model:

```
data.unamb_cont %>%
  lrm(form ~ context_pos + sex + part_of_speech + conj + length, data = .) %>%
  print()
```

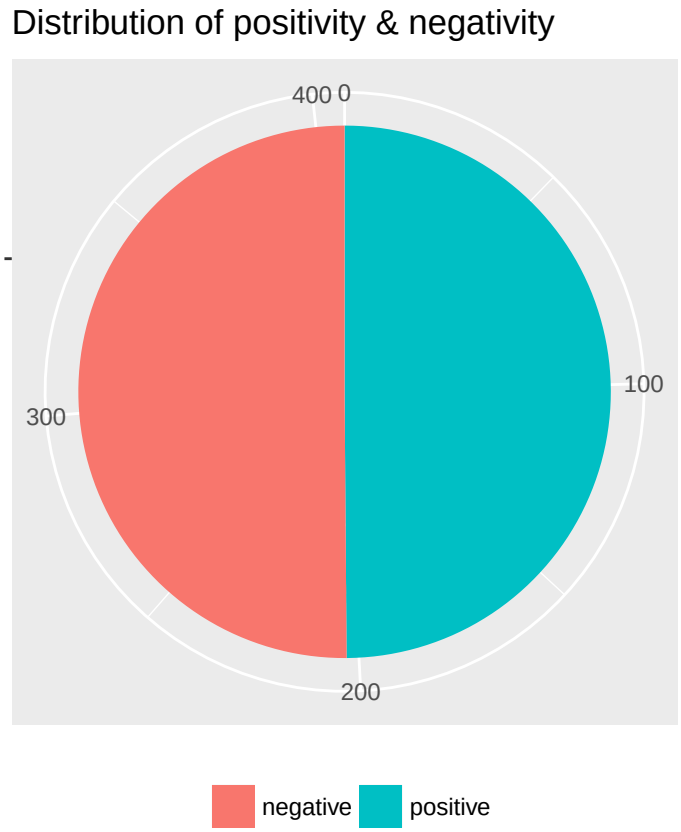
```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + sex + part_of_speech + conj +
##      length, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      56.78    R2       0.246    C       0.762
##   subst         85    d.f.             7    g        1.932    Dxy     0.524
##   vf           217    Pr(> chi2) <0.0001    gr       6.901    gamma   0.553
##  max |deriv| 0.004                          gp       0.209    tau-a   0.213
##                                             Brier    0.169
##
##                      Coef    S.E.    Wald Z Pr(>|Z|)
##  Intercept            1.9677  0.6495  3.03  0.0025
##  context_pos=right   -1.3005  0.3877 -3.35  0.0008
##  sex=m               -0.9863  0.3869 -2.55  0.0108
##  part_of_speech=adv   8.6332 27.2404  0.32  0.7513
```

```
##  part_of_speech=part  1.2871  0.8139  1.58  0.1138
##  part_of_speech=v    -0.7191  0.2941 -2.44  0.0145
##  conj                -1.4227  0.4821 -2.95  0.0032
##  length               0.8313  0.2871  2.90  0.0038
##
```

Very good gain in R^2, updating the best model.

6. Positivity and 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.unamb_cont %>%
  lrm(form ~ pos_neg, data = .) %>%
  print()
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ pos_neg, data = .)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      7.77    R2       0.037    C       0.589
##   subst         85    d.f.            1    g        0.364    Dxy     0.177
##   vf           217    Pr(> chi2) 0.0053    gr       1.439    gamma   0.348
##  max |deriv| 8e-14                         gp       0.072    tau-a   0.072
##                                            Brier    0.197
##
##              Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept    1.3437 0.2050  6.55  <0.0001
##  pos_neg=pos -0.7261 0.2646 -2.74  0.0061
##
```

Incorporating into the best model:

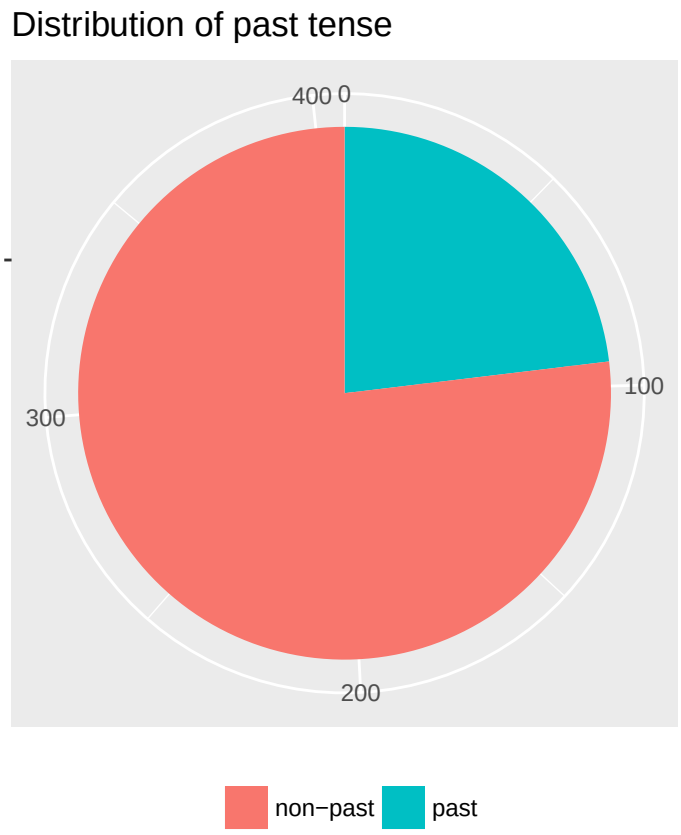
```
data.unamb_cont %>%
  lrm(form ~ context_pos + sex + part_of_speech + conj + length + pos_neg, data = .) %>%
  print()
```

```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + sex + part_of_speech + conj +
##      length + pos_neg, data = .)
##
##                         Model Likelihood     Discrimination    Rank Discrim.
##                            Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      57.79    R2       0.250    C       0.761
##   subst         85    d.f.             8    g        1.944    Dxy     0.521
##   vf           217    Pr(> chi2) <0.0001    gr       6.984    gamma   0.549
##  max |deriv| 0.004                          gp       0.210    tau-a   0.212
##                                             Brier    0.168
##
##                      Coef    S.E.    Wald Z Pr(>|Z|)
##  Intercept            2.0117  0.6556  3.07  0.0022
##  context_pos=right   -1.3141  0.3898 -3.37  0.0007
##  sex=m               -0.9657  0.3889 -2.48  0.0130
##  part_of_speech=adv   8.6391 27.2616  0.32  0.7513
##  part_of_speech=part  0.0157  1.5561  0.01  0.9919
##  part_of_speech=v    -1.9540  1.3263 -1.47  0.1407
##  conj                -1.4268  0.4827 -2.96  0.0031
##  length               0.8202  0.2876  2.85  0.0043
##  pos_neg=pos          1.2683  1.3247  0.96  0.3384
##
```

Gain in R^2 is very insufficient, not updating the best model.

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 %>%
  lrm(form ~ past_tense, data = .) %>%
  print()
```

```
## Logistic Regression Model
```

```
##
##  lrm(formula = form ~ past_tense, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      2.79    R2       0.013    C       0.546
##   subst         85    d.f.            1    g        0.175    Dxy     0.091
##   vf           217    Pr(> chi2) 0.0946    gr       1.191    gamma   0.242
##  max |deriv| 2e-08                         gp       0.037    tau-a   0.037
##                                            Brier    0.200
##
##                 Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept       1.0589 0.1498  7.07  <0.0001
##  past_tense=yes -0.4936 0.2918 -1.69  0.0908
##
```

Incorporating into the best model:

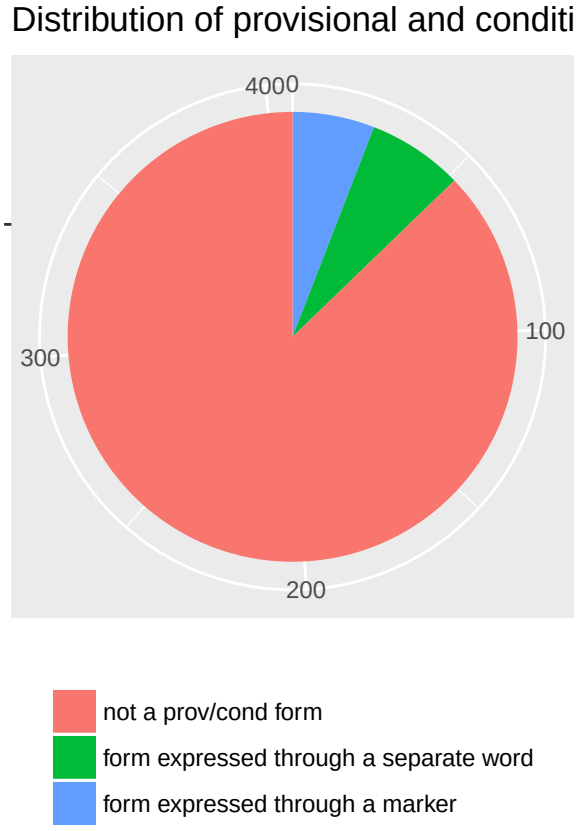
```
data.unamb_cont %>%
  lrm(form ~ context_pos + sex + part_of_speech +
        length + conj + past_tense, data = .) %>%
  print()
```

```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + sex + part_of_speech + length +
##      conj + past_tense, data = .)
##
##                         Model Likelihood     Discrimination    Rank Discrim.
##                            Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      58.39    R2       0.253    C       0.765
##   subst         85    d.f.             8    g        1.958    Dxy     0.529
##   vf           217    Pr(> chi2) <0.0001    gr       7.088    gamma   0.547
##  max |deriv| 0.004                          gp       0.214    tau-a   0.215
##                                             Brier    0.167
##
##                      Coef    S.E.    Wald Z Pr(>|Z|)
##  Intercept            1.9585  0.6519  3.00  0.0027
##  context_pos=right   -1.2970  0.3868 -3.35  0.0008
##  sex=m               -0.9724  0.3876 -2.51  0.0121
##  part_of_speech=adv   8.7568 26.8495  0.33  0.7443
##  part_of_speech=part  1.4008  0.8161  1.72  0.0861
##  part_of_speech=v    -0.6645  0.2985 -2.23  0.0260
##  length               0.8201  0.2893  2.83  0.0046
##  conj                -1.3554  0.4897 -2.77  0.0056
##  past_tense=yes      -0.4254  0.3343 -1.27  0.2032
##
```

Gain in R^2 is very insufficient, not updating the best model.

8. Provisional and conditional forms



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 %>%
  lrm(form ~ prov_cond, data = .) %>%
  print()
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ prov_cond, data = .)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      31.91    R2       0.144    C        0.607
##   subst         85    d.f.             2    g        0.413    Dxy      0.215
##   vf           217    Pr(> chi2) <0.0001    gr       1.512    gamma    0.737
##  max |deriv| 3e-10                          gp       0.087    tau-a    0.087
```

```
##                                           Brier    0.178
##
##                Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept      1.1847 0.1440  8.23  <0.0001
##  prov_cond=sep -2.9193 0.6426 -4.54  <0.0001
##  prov_cond=vf  -0.7147 0.5880 -1.22  0.2241
##
```

Incorporating into the best model:

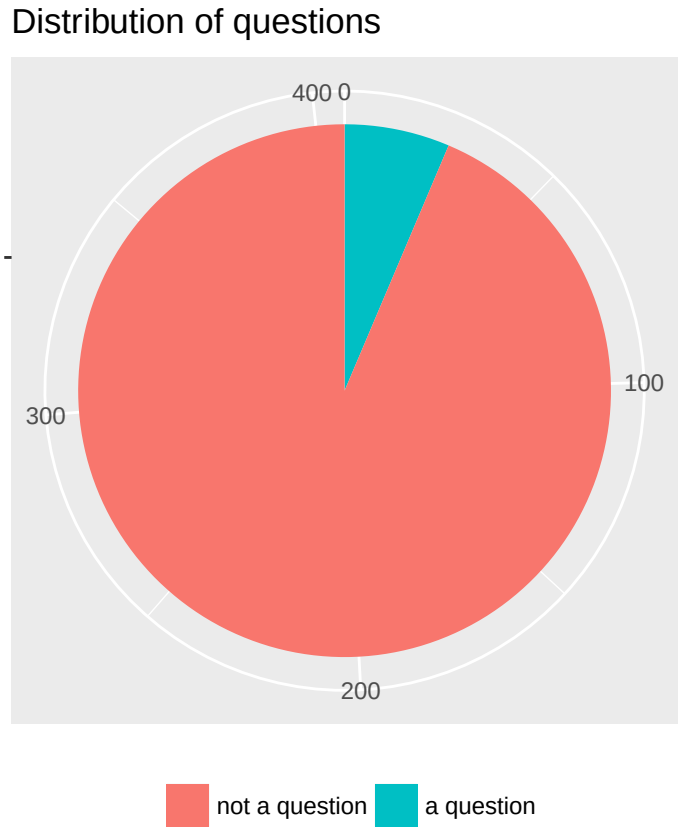
```
data.unamb_cont %>%
  lrm(form ~ context_pos +conj + sex + part_of_speech + length + prov_cond, data = .) %>%
  print()
```

```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + conj + sex + part_of_speech +
##      length + prov_cond, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      76.46    R2       0.322    C       0.789
##   subst         85    d.f.             9    g        2.105    Dxy     0.578
##   vf           217    Pr(> chi2) <0.0001    gr       8.208    gamma   0.598
##  max |deriv| 0.004                          gp       0.236    tau-a   0.235
##                                             Brier    0.155
##
##                      Coef    S.E.    Wald Z Pr(>|Z|)
##  Intercept            1.7415  0.6709  2.60  0.0094
##  context_pos=right   -1.0917  0.3965 -2.75  0.0059
##  conj                -1.5668  0.5060 -3.10  0.0020
##  sex=m               -1.0583  0.4002 -2.64  0.0082
##  part_of_speech=adv   8.5476 27.2519  0.31  0.7538
##  part_of_speech=part  1.2055  0.8209  1.47  0.1419
##  part_of_speech=v    -0.5444  0.3135 -1.74  0.0825
##  length               0.9550  0.3028  3.15  0.0016
##  prov_cond=sep       -2.5257  0.6800 -3.71  0.0002
##  prov_cond=vf         0.0881  0.6323  0.14  0.8892
##
```

Good gain, updating the model.

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
```

No reason to run regression.

Incorporating into the best model:

```
data.unamb_cont %>%
  lrm(form ~ context_pos + conj + sex + length +
        part_of_speech + prov_cond + quest, data = .) %>%
  print()
```

```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
```

```
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + conj + sex + length + part_of_speech +
##      prov_cond + quest, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      76.66    R2       0.322    C       0.789
##   subst         85    d.f.            10    g        2.109    Dxy     0.578
##   vf           217    Pr(> chi2) <0.0001    gr       8.239    gamma   0.594
##  max |deriv| 0.004                          gp       0.236    tau-a   0.235
##                                             Brier    0.155
##
##                      Coef    S.E.    Wald Z Pr(>|Z|)
##  Intercept            1.7551  0.6728  2.61  0.0091
##  context_pos=right   -1.1180  0.4005 -2.79  0.0053
##  conj                -1.5428  0.5093 -3.03  0.0025
##  sex=m               -1.0460  0.4010 -2.61  0.0091
##  length               0.9440  0.3042  3.10  0.0019
##  part_of_speech=adv   8.5364 27.2287  0.31  0.7539
##  part_of_speech=part  1.1941  0.8217  1.45  0.1462
##  part_of_speech=v    -0.5861  0.3262 -1.80  0.0724
##  prov_cond=sep       -2.4824  0.6867 -3.62  0.0003
##  prov_cond=vf         0.1151  0.6355  0.18  0.8562
##  quest=yes            0.2520  0.5621  0.45  0.6539
##
```

No changes to R^2, not updating the model.

Final model:

```
data.unamb_cont %>%
  lrm(form ~ context_pos +conj + sex + part_of_speech + length + prov_cond, data = .) %>%
  print()
```

```
## Warning in Design(eval.parent(m)): Variable context_pos has levels left
## right which do not match levels given to datadist ( left right unclear ).
## datadist values ignored.
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ context_pos + conj + sex + part_of_speech +
##      length + prov_cond, data = .)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes          Indexes
##  Obs           302    LR chi2      76.46    R2       0.322    C       0.789
##   subst         85    d.f.             9    g        2.105    Dxy     0.578
##   vf           217    Pr(> chi2) <0.0001    gr       8.208    gamma   0.598
##  max |deriv| 0.004                          gp       0.236    tau-a   0.235
##                                             Brier    0.155
##
##                      Coef    S.E.    Wald Z Pr(>|Z|)
```

```
##  Intercept             1.7415  0.6709  2.60  0.0094
##  context_pos=right    -1.0917  0.3965 -2.75  0.0059
##  conj                 -1.5668  0.5060 -3.10  0.0020
##  sex=m                -1.0583  0.4002 -2.64  0.0082
##  part_of_speech=adv    8.5476 27.2519  0.31  0.7538
##  part_of_speech=part   1.2055  0.8209  1.47  0.1419
##  part_of_speech=v     -0.5444  0.3135 -1.74  0.0825
##  length                0.9550  0.3028  3.15  0.0016
##  prov_cond=sep        -2.5257  0.6800 -3.71  0.0002
##  prov_cond=vf          0.0881  0.6323  0.14  0.8892
##
```

Our model explains 32% of the dependent variable's variance. It is greater than probability of choosing the substantivation strategy, so it might be a good result.