Supplementary materials to the paper ‘Adaptation of Russian borrowings in Andic languages’

21.04.2025

Table of contents

1 Data

Read the dataset. The whole document is based on manipulations of different parts of the same dataset.

```r
library(tidyverse)

read_csv("data.csv", show_col_types = FALSE) |>
  mutate(language_ref = str_c(language, ": ", reference)) ->
  df
```

The dataset consists of 21749 observations with the following columns:

- `language`: language
- `reference`: source of data
- `dictionary_translation`: unified dictionary translation
- `lemma_frequency_ipm`: frequency of the dictionary translation in RNC;

- `russian_ipa`: modified IPA transcription of the Russian word or part of the word;
- `target_ipa`: IPA transcription of target language word;
- `change`: binary coding for the change;
- `type_of_change`: coding of the type of change (e.g., apocope, epenthesis, metathesis, and others);
- `total`: total number of units in the analysis;
- `changes`: number of observed changes;
- `time_of_borrowing`: approximate time of borrowing based on data from the RNC.

Here is a most frequent values in the subsample of variables (variables `dictionary_translation`, `lemma frequency_ipm`, `russian_ipa` and `target_ipa` are ommited due to the huge amount of values).

```r
library(inspectdf)

df |>
  mutate_all(as.factor) |>
  select(language, reference, change, total_units_per_word, changes_per_word, time_of_borrowing, type_of_change) |>
  rename(`(1) language` = language,
         `(2) reference` = reference,
         `(3) change` = change,
         `(4) total units per word` = total_units_per_word,
         `(5) changes per word` = changes_per_word,
         `(6) type of change` = type_of_change,
         `(7) time of borrowing` = time_of_borrowing) |>
  inspect_cat()   |>
  show_plot()+
  labs(title = NULL, subtitle = NULL, text = element_text(size = 30))+
  theme(axis.text.y = element_text(hjust = 0))
```

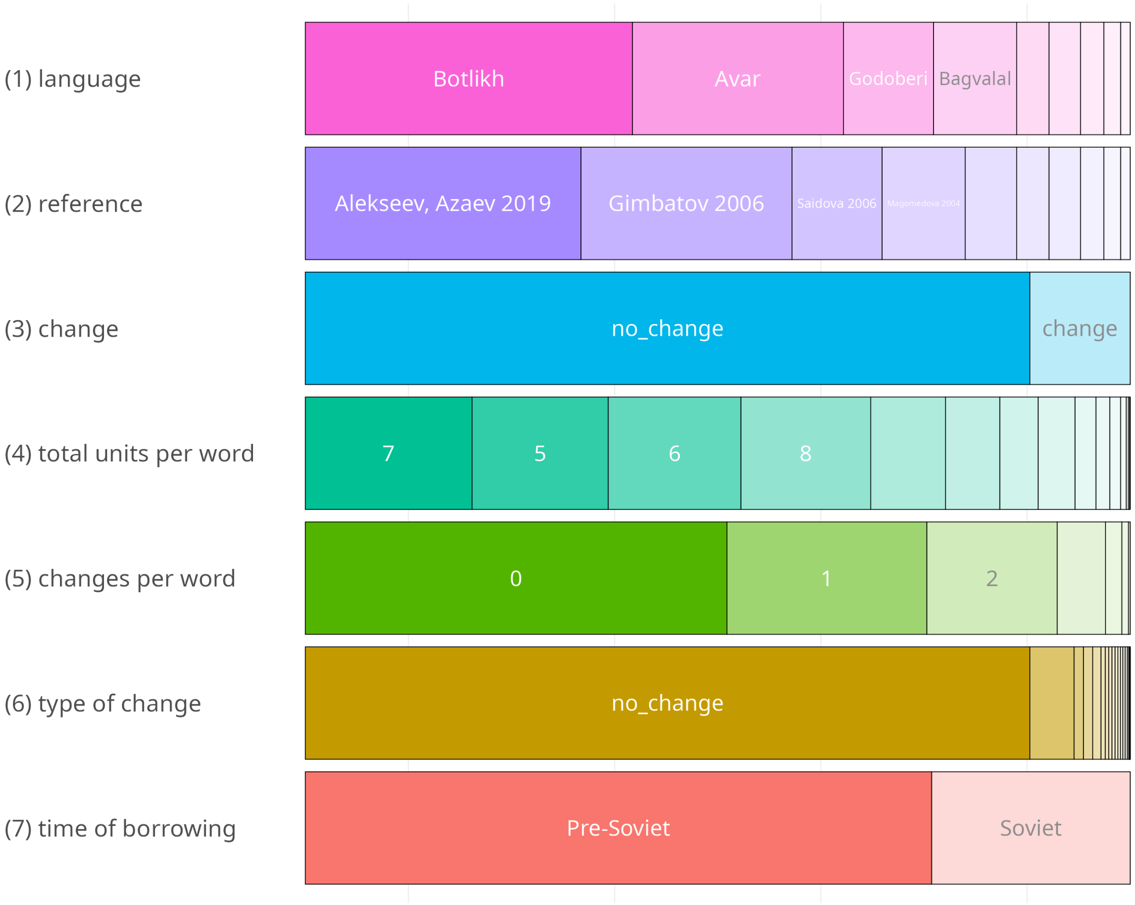

Figure 1: Frequency of values for each variable in the dataset. Gray segments are missing values.

2 Distribution of changes across languages

```
df ▷
  select(language_ref, type_of_change, time_of_borrowing) ▷
  filter(type_of_change ≠ "no_change",
         type_of_change ≠ "other") ▷
  mutate(type_of_change = str_split(type_of_change, ", ")) ▷
  unnest(type_of_change) ▷
  add_count(language_ref, type_of_change) ▷
  count(language_ref, type_of_change, time_of_borrowing, n) ▷
  mutate(type_of_change = tidytext::reorder_within(type_of_change,  n, language_ref),
         time_of_borrowing = fct_relevel(time_of_borrowing, "Soviet")) ▷
```

```r
  ggplot(aes(nn, type_of_change, fill = time_of_borrowing))+
  geom_col()+
  facet_wrap(~language_ref, scales = "free", ncol = 2)+
  tidytext::scale_y_reordered()+
  scale_x_continuous(breaks = scales::breaks_pretty())+
  labs(x = NULL,
       y = NULL,
       fill = NULL)+
  theme_minimal()+
  theme(legend.position = "bottom")
```

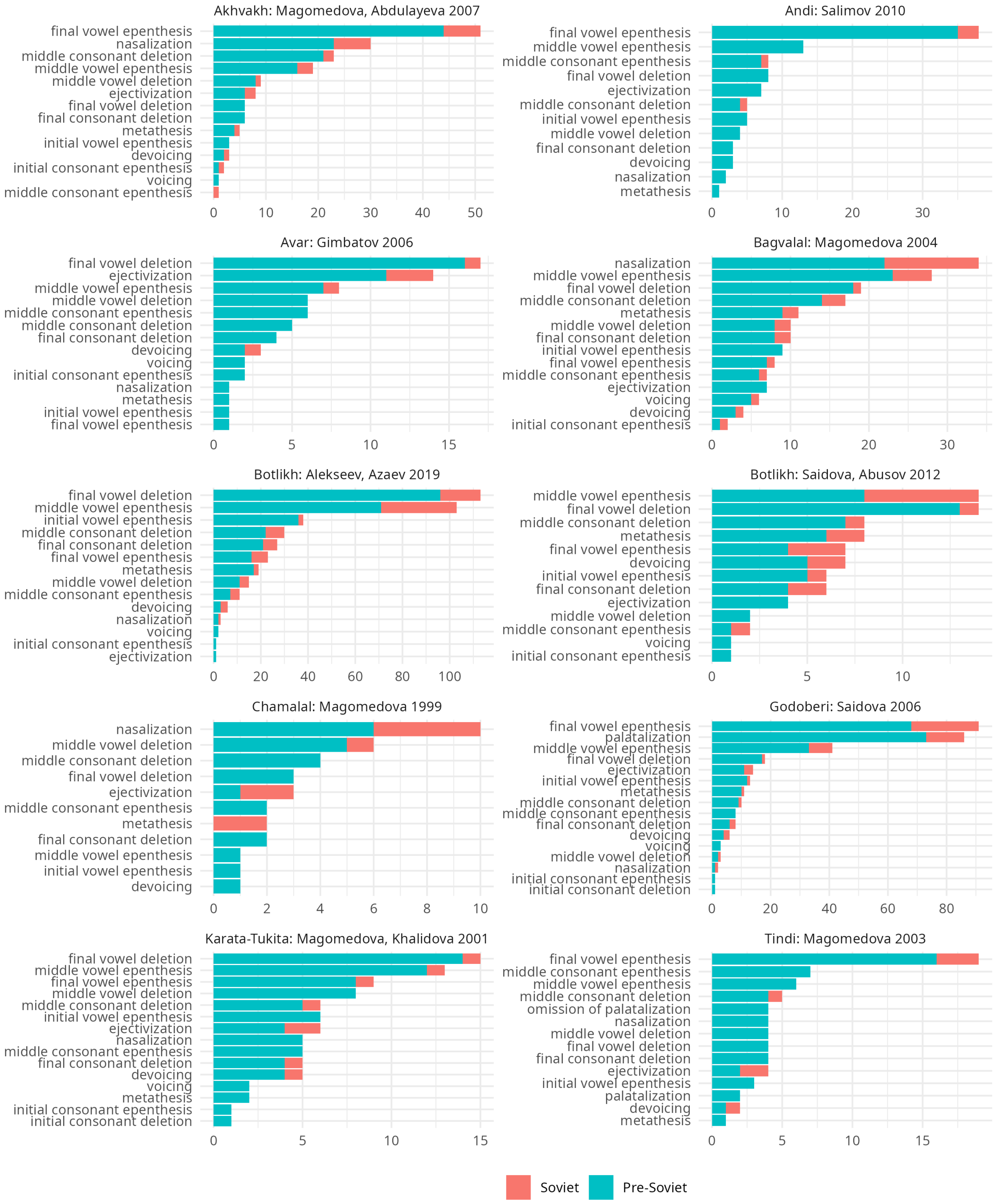

Figure 2: Frequency of each change by language

3 Modeling

We decided to create a model that predicts the average number of changes based on the dictionary and approximated time of borrowing. To do so, we applied a mixed effect logistic regression model. The models were generated with the R (R Core Team 2024) package `lme4` (Bates et al. 2015) with the following formula:

```
change ~ dictionary*approximated time of borrowing + dictionary translation frequency + (1|dictionary lemma translation)
```

The random effect of the model is unified dictionary lemma translation. We included in the model interaction of two variables: language resource and approximated time of borrowing. Since this model will compare values with some baseline Soviet borrowings from Avar dictionary (Gimbatov 2006) were used as a baseline. Differences between all dictionaries turn out to be statistically significant. Approximated time of borrowing (p-value = 0.07570) and dictionary translation frequency (p-value = 0.05672) turned out not to be statistically significant. Just a few interactions of language resources variable with the time of borrowing variable turn out to be statistically significant: for Andi and for Tindi.

```r
library(lme4)
library(lmerTest)
```

```

Attaching package: 'lmerTest'

The following object is masked from 'package:lme4':

    lmer

The following object is masked from 'package:stats':

    step
```

```r
df |>
  mutate(change = if_else(change == "no_change", 0, 1),
         lemma_frequency_ipm = if_else(is.na(lemma_frequency_ipm), log(0.0001), log(lemma_frequency_ipm)),
         time_of_borrowing = fct_relevel(time_of_borrowing, "Soviet"),
         language_ref = fct_relevel(language_ref, "Avar: Gimbatov 2006")) |>
  lmer(change ~ time_of_borrowing*language_ref+lemma_frequency_ipm + (1|russian_source_lexeme), data = _) ->
  fit

summary(fit)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: change ~ time_of_borrowing * language_ref + lemma_frequency_ipm +
    (1 | russian_source_lexeme)
   Data:
mutate(df, change = if_else(change == "no_change", 0, 1), lemma_frequency_ipm = if_else(is.na(lemma_frequency_ipm),
    log(1e-04), log(lemma_frequency_ipm)), time_of_borrowing = fct_relevel(time_of_borrowing,
    "Soviet"), language_ref = fct_relevel(language_ref, "Avar: Gimbatov 2006"))

REML criterion at convergence: 11149.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8462 -0.4593 -0.2548 -0.0446  3.3500

Random effects:
 Groups                Name        Variance Std.Dev.
 russian_source_lexeme (Intercept) 0.005434 0.07372
 Residual                          0.093594 0.30593
Number of obs: 21749, groups:  russian_source_lexeme, 1343

Fixed effects:
                                                                                    Estimate
(Intercept)                                                                        1.474e-02
time_of_borrowingPre-Soviet                                                        1.985e-02
language_refAkhvakh: Magomedova, Abdulayeva 2007                                   2.387e-01
language_refAndi: Salimov 2010                                                     1.072e-01
language_refBagvalal: Magomedova 2004                                              1.109e-01
language_refBotlikh: Alekseev, Azaev 2019                                          5.987e-02
language_refBotlikh: Saidova, Abusov 2012                                          8.655e-02
language_refChamalal: Magomedova 1999                                              2.724e-01
language_refGodoberi: Saidova 2006                                                 1.512e-01
language_refKarata-Tukita: Magomedova, Khalidova 2001                              1.671e-01
language_refTindi: Magomedova 2003                                                 3.659e-01
lemma_frequency_ipm                                                               -2.107e-03
time_of_borrowingPre-Soviet:language_refAkhvakh: Magomedova, Abdulayeva 2007       1.314e-02
time_of_borrowingPre-Soviet:language_refAndi: Salimov 2010                         7.443e-02
time_of_borrowingPre-Soviet:language_refBagvalal: Magomedova 2004                  1.493e-03
time_of_borrowingPre-Soviet:language_refBotlikh: Alekseev, Azaev 2019              2.062e-02
time_of_borrowingPre-Soviet:language_refBotlikh: Saidova, Abusov 2012              7.669e-03
time_of_borrowingPre-Soviet:language_refChamalal: Magomedova 1999                 -8.648e-02
time_of_borrowingPre-Soviet:language_refGodoberi: Saidova 2006                     1.964e-02
time_of_borrowingPre-Soviet:language_refKarata-Tukita: Magomedova, Khalidova 2001  3.908e-02
```

```
time_of_borrowingPre-Soviet:language_refTindi: Magomedova 2003                    -1.264e-01
                                                                                  Std. Error
(Intercept)                                                                        9.661e-03
time_of_borrowingPre-Soviet                                                        1.117e-02
language_refAkhvakh: Magomedova, Abdulayeva 2007                                   3.094e-02
language_refAndi: Salimov 2010                                                     3.280e-02
language_refBagvalal: Magomedova 2004                                              1.813e-02
language_refBotlikh: Alekseev, Azaev 2019                                          1.115e-02
language_refBotlikh: Saidova, Abusov 2012                                          1.985e-02
language_refChamalal: Magomedova 1999                                              4.739e-02
language_refGodoberi: Saidova 2006                                                 1.695e-02
language_refKarata-Tukita: Magomedova, Khalidova 2001                              4.184e-02
language_refTindi: Magomedova 2003                                                 5.681e-02
lemma_frequency_ipm                                                                1.105e-03
time_of_borrowingPre-Soviet:language_refAkhvakh: Magomedova, Abdulayeva 2007       3.364e-02
time_of_borrowingPre-Soviet:language_refAndi: Salimov 2010                         3.539e-02
time_of_borrowingPre-Soviet:language_refBagvalal: Magomedova 2004                  2.034e-02
time_of_borrowingPre-Soviet:language_refBotlikh: Alekseev, Azaev 2019              1.299e-02
time_of_borrowingPre-Soviet:language_refBotlikh: Saidova, Abusov 2012              2.294e-02
time_of_borrowingPre-Soviet:language_refChamalal: Magomedova 1999                  5.282e-02
time_of_borrowingPre-Soviet:language_refGodoberi: Saidova 2006                     1.921e-02
time_of_borrowingPre-Soviet:language_refKarata-Tukita: Magomedova, Khalidova 2001  4.439e-02
time_of_borrowingPre-Soviet:language_refTindi: Magomedova 2003                     5.929e-02
                                                                                          df
(Intercept)                                                                        5.524e+03
time_of_borrowingPre-Soviet                                                        5.279e+03
language_refAkhvakh: Magomedova, Abdulayeva 2007                                   1.839e+04
language_refAndi: Salimov 2010                                                     1.695e+04
language_refBagvalal: Magomedova 2004                                              1.871e+04
language_refBotlikh: Alekseev, Azaev 2019                                          1.891e+04
language_refBotlikh: Saidova, Abusov 2012                                          1.767e+04
language_refChamalal: Magomedova 1999                                              2.037e+04
language_refGodoberi: Saidova 2006                                                 1.784e+04
language_refKarata-Tukita: Magomedova, Khalidova 2001                              2.059e+04
language_refTindi: Magomedova 2003                                                 2.006e+04
lemma_frequency_ipm                                                                1.782e+03
time_of_borrowingPre-Soviet:language_refAkhvakh: Magomedova, Abdulayeva 2007       1.848e+04
time_of_borrowingPre-Soviet:language_refAndi: Salimov 2010                         1.726e+04
time_of_borrowingPre-Soviet:language_refBagvalal: Magomedova 2004                  1.898e+04
time_of_borrowingPre-Soviet:language_refBotlikh: Alekseev, Azaev 2019              1.923e+04
time_of_borrowingPre-Soviet:language_refBotlikh: Saidova, Abusov 2012              1.814e+04
time_of_borrowingPre-Soviet:language_refChamalal: Magomedova 1999                  2.033e+04
time_of_borrowingPre-Soviet:language_refGodoberi: Saidova 2006                     1.835e+04
```

```
time_of_borrowingPre-Soviet:language_refKarata-Tukita: Magomedova, Khalidova 2001  2.053e+04
time_of_borrowingPre-Soviet:language_refTindi: Magomedova 2003                     1.998e+04
                                                                                  t value
(Intercept)                                                                         1.526
time_of_borrowingPre-Soviet                                                         1.777
language_refAkhvakh: Magomedova, Abdulayeva 2007                                    7.715
language_refAndi: Salimov 2010                                                      3.270
language_refBagvalal: Magomedova 2004                                               6.117
language_refBotlikh: Alekseev, Azaev 2019                                           5.368
language_refBotlikh: Saidova, Abusov 2012                                           4.361
language_refChamalal: Magomedova 1999                                               5.747
language_refGodoberi: Saidova 2006                                                  8.925
language_refKarata-Tukita: Magomedova, Khalidova 2001                               3.994
language_refTindi: Magomedova 2003                                                  6.441
lemma_frequency_ipm                                                                -1.907
time_of_borrowingPre-Soviet:language_refAkhvakh: Magomedova, Abdulayeva 2007        0.391
time_of_borrowingPre-Soviet:language_refAndi: Salimov 2010                          2.103
time_of_borrowingPre-Soviet:language_refBagvalal: Magomedova 2004                   0.073
time_of_borrowingPre-Soviet:language_refBotlikh: Alekseev, Azaev 2019               1.587
time_of_borrowingPre-Soviet:language_refBotlikh: Saidova, Abusov 2012               0.334
time_of_borrowingPre-Soviet:language_refChamalal: Magomedova 1999                  -1.637
time_of_borrowingPre-Soviet:language_refGodoberi: Saidova 2006                      1.022
time_of_borrowingPre-Soviet:language_refKarata-Tukita: Magomedova, Khalidova 2001   0.880
time_of_borrowingPre-Soviet:language_refTindi: Magomedova 2003                     -2.132
                                                                                  Pr(>|t|)
(Intercept)                                                                        0.12704
time_of_borrowingPre-Soviet                                                        0.07570
language_refAkhvakh: Magomedova, Abdulayeva 2007                                  1.28e-14
language_refAndi: Salimov 2010                                                     0.00108
language_refBagvalal: Magomedova 2004                                             9.73e-10
language_refBotlikh: Alekseev, Azaev 2019                                         8.06e-08
language_refBotlikh: Saidova, Abusov 2012                                         1.30e-05
language_refChamalal: Magomedova 1999                                             9.23e-09
language_refGodoberi: Saidova 2006                                                 < 2e-16
language_refKarata-Tukita: Magomedova, Khalidova 2001                             6.51e-05
language_refTindi: Magomedova 2003                                                1.22e-10
lemma_frequency_ipm                                                                0.05672
time_of_borrowingPre-Soviet:language_refAkhvakh: Magomedova, Abdulayeva 2007       0.69604
time_of_borrowingPre-Soviet:language_refAndi: Salimov 2010                         0.03547
time_of_borrowingPre-Soviet:language_refBagvalal: Magomedova 2004                  0.94146
time_of_borrowingPre-Soviet:language_refBotlikh: Alekseev, Azaev 2019              0.11249
time_of_borrowingPre-Soviet:language_refBotlikh: Saidova, Abusov 2012              0.73812
time_of_borrowingPre-Soviet:language_refChamalal: Magomedova 1999                  0.10160
```

```
time_of_borrowingPre-Soviet:language_refGodoberi: Saidova 2006                     0.30660
time_of_borrowingPre-Soviet:language_refKarata-Tukita: Magomedova, Khalidova 2001  0.37862
time_of_borrowingPre-Soviet:language_refTindi: Magomedova 2003                     0.03299

(Intercept)
time_of_borrowingPre-Soviet                                                       .
language_refAkhvakh: Magomedova, Abdulayeva 2007                                  ***
language_refAndi: Salimov 2010                                                    **
language_refBagvalal: Magomedova 2004                                             ***
language_refBotlikh: Alekseev, Azaev 2019                                         ***
language_refBotlikh: Saidova, Abusov 2012                                         ***
language_refChamalal: Magomedova 1999                                             ***
language_refGodoberi: Saidova 2006                                                ***
language_refKarata-Tukita: Magomedova, Khalidova 2001                             ***
language_refTindi: Magomedova 2003                                                ***
lemma_frequency_ipm                                                               .
time_of_borrowingPre-Soviet:language_refAkhvakh: Magomedova, Abdulayeva 2007
time_of_borrowingPre-Soviet:language_refAndi: Salimov 2010                        *
time_of_borrowingPre-Soviet:language_refBagvalal: Magomedova 2004
time_of_borrowingPre-Soviet:language_refBotlikh: Alekseev, Azaev 2019
time_of_borrowingPre-Soviet:language_refBotlikh: Saidova, Abusov 2012
time_of_borrowingPre-Soviet:language_refChamalal: Magomedova 1999
time_of_borrowingPre-Soviet:language_refGodoberi: Saidova 2006
time_of_borrowingPre-Soviet:language_refKarata-Tukita: Magomedova, Khalidova 2001
time_of_borrowingPre-Soviet:language_refTindi: Magomedova 2003                    *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Correlation matrix not shown by default, as p = 21 > 12.
Use print(x, correlation=TRUE)  or
    vcov(x)        if you need it
```

The model predictions are visualized with effect plots.

```r
library(ggeffects)

df |>
    distinct(language_ref, russian_source_lexeme) |>
    count(language_ref) |>
    rename(x = language_ref,
           word_list_size = n) ->
    word_list_size

fit |>
  ggpredict(terms = c("language_ref", "time_of_borrowing")) |>
  as_tibble() |>
  left_join(word_list_size) |>
  mutate(x = str_c(x, " (", word_list_size, " lemmata)"),
         x = fct_reorder(x, predicted)) |>
  ggplot(aes(predicted, x, color = group))+
  geom_linerange(aes(xmin = conf.low, xmax = conf.high), position = position_dodge(width = 0.5)) +
  geom_point(show.legend = FALSE, position = position_dodge(width = 0.5))+
  theme_minimal()+
  labs(x = "model prediction of the probability of change",
       y = NULL,
       color = NULL)+
  theme(text = element_text(size = 19),
        legend.position = "bottom")
```

4 Packages

In the following table, we list all R packages and R version used in the project:

Table 1: The list of versions of R packages used in the project

package	version	citation
ggeffects	2.2.1	Lüdecke (2018)
inspectdf	0.0.12.1	Rushworth (2024)
lme4	1.1.37	Bates (2015)
quarto	1.4.4	Allaire (2024)
scales	1.3.0	Wickham (2023)
tidytext	0.4.2	Silge (2016)
tidyverse	2.0.0	Wickham (2019)
R	4.3.3	R Core Team (2024)

Allaire, JJ, and Christophe Dervieux. 2024. *Quarto: R Interface to 'Quarto' Markdown Publishing System*. https://CRAN.R-project.org/package=quarto.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

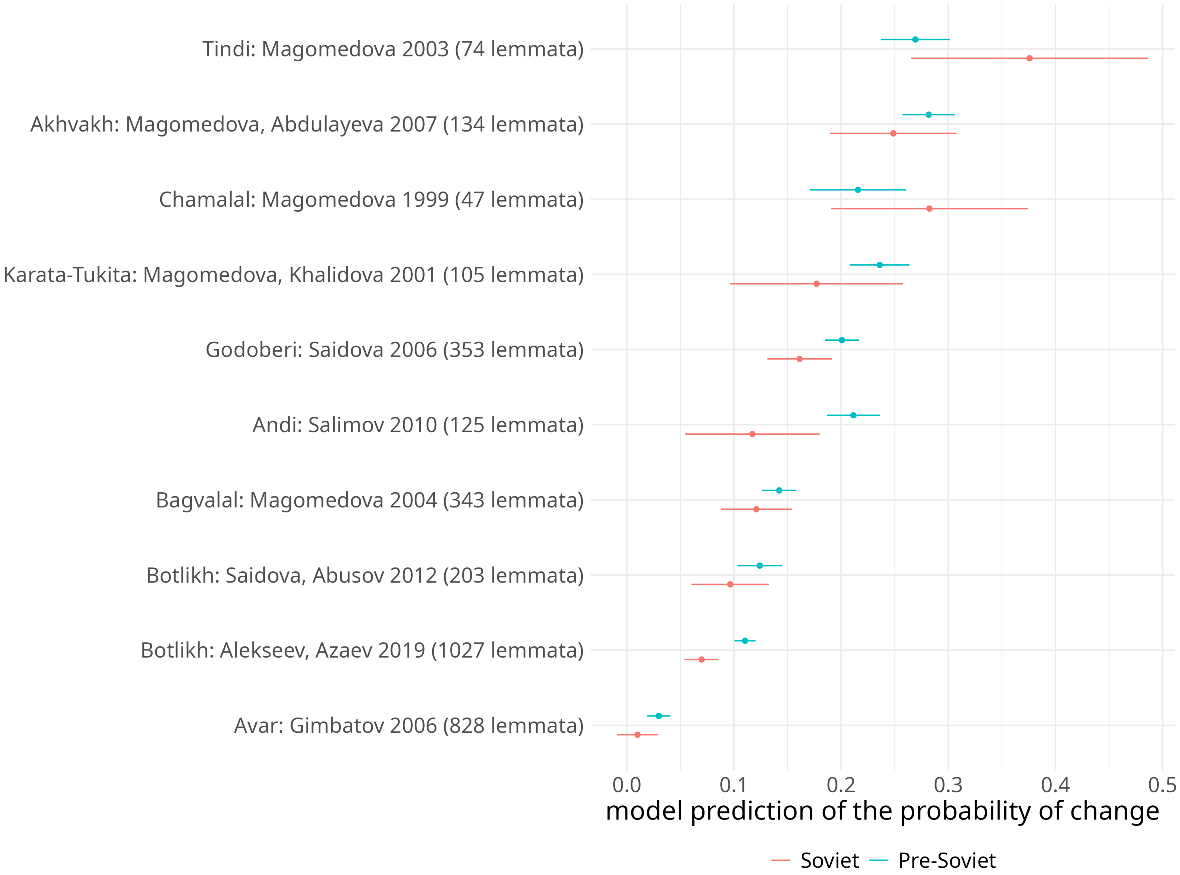

Figure 3: Probabilities of change by language, source and approximate time of borrowing with 95% confidence intervals.

Lüdecke, Daniel. 2018. "Ggeffects: Tidy Data Frames of Marginal Effects from Regression Models." *Journal of Open Source Software* 3 (26): 772. https://doi.org/10.21105/joss.00772.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rushworth, Alastair. 2024. *Inspectdf: Inspection, Comparison and Visualisation of Data Frames*. https://CRAN.R-project.org/package=inspectdf.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in r." *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.