

# Исследования вариативности в русском языке билингвов

Г. А. Мороз (Международная лаборатория языковой  
конвергенции)

26.03.2024



## Прескриптивное vs. дескриптивное



# Прескриптивная vs. дескриптивная лингвистика

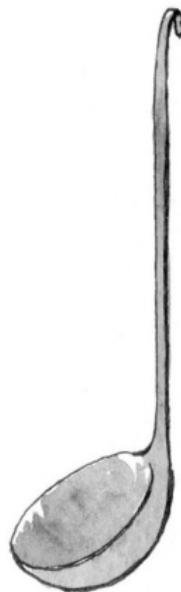
- прескриптивная

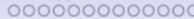
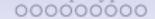
# Прескриптивная vs. дескриптивная лингвистика

- прескриптивная
- вся остальная (дескриптивная)
  - каталогизация языкового разнообразия, описание языковых контактов
  - исследования и документация грамматики, фонетики и лексики конкретных языков
  - исследования распределения грамматических/фонетических/лексических особенностей в языках мира
  - исследования и документация исторических изменений грамматических/фонетических/лексических особенностей языков
  - исследования когнитивных способностей человека и других животных, связанных с языком (усвоение, потеря языка и др.)
  - языковые аспекты исследования мозга
  - исследования в области синтеза и распознавания речи и языка
  - исследования в области NLP, пробинг языковых моделей и т. п.
  - ...

# Прескриптивная vs. дескриптивная лингвистика

Запишите где-нибудь, что изображено на картинке (рис. Т. Пановой).





# Прескриптивная vs. дескриптивная лингвистика

Это часть опроса И. Левина 2021 года:



- Синий (Blue): половник
- Оранжевый (Orange): поварёшка
- Зелёный (Green): черпак
- Красный (Red): ополовник
- Фиолетовый (Purple): ополонник

# Корпусная лингвистика

## Корпусная лингвистика

Корпусная лингвистика — это область лингвистики, которая занимается исследованием языковых явлений на материале некоторых собраний языкового материала. В большинстве случаев это письменные тексты, однако это может быть аудио и даже видео корпуса.

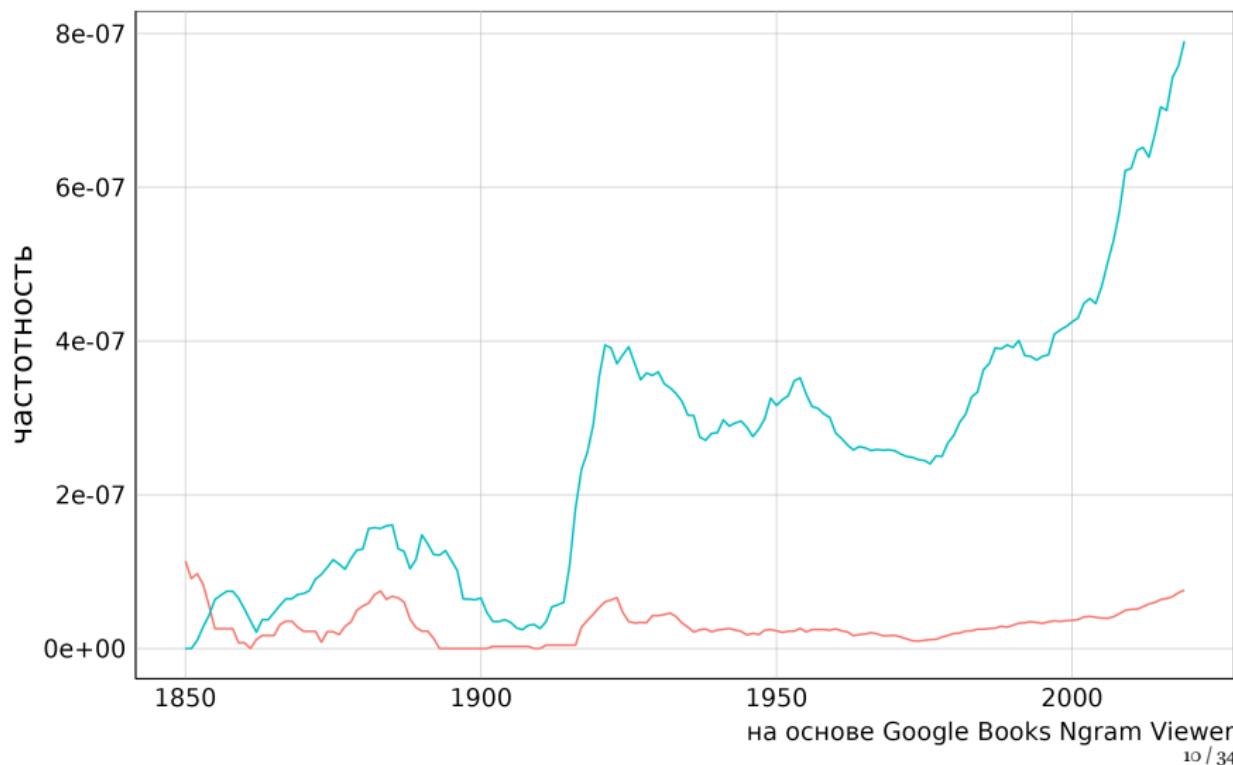
Среди корпусов русского языка можно назвать:

- Национальный корпус русского языка
  - более 1.5 млрд слов
  - много подкорпусов (газетный, устный, параллельный, диалектный, поэтический, исторические)
- Google Books Ngram Viewer
- ...

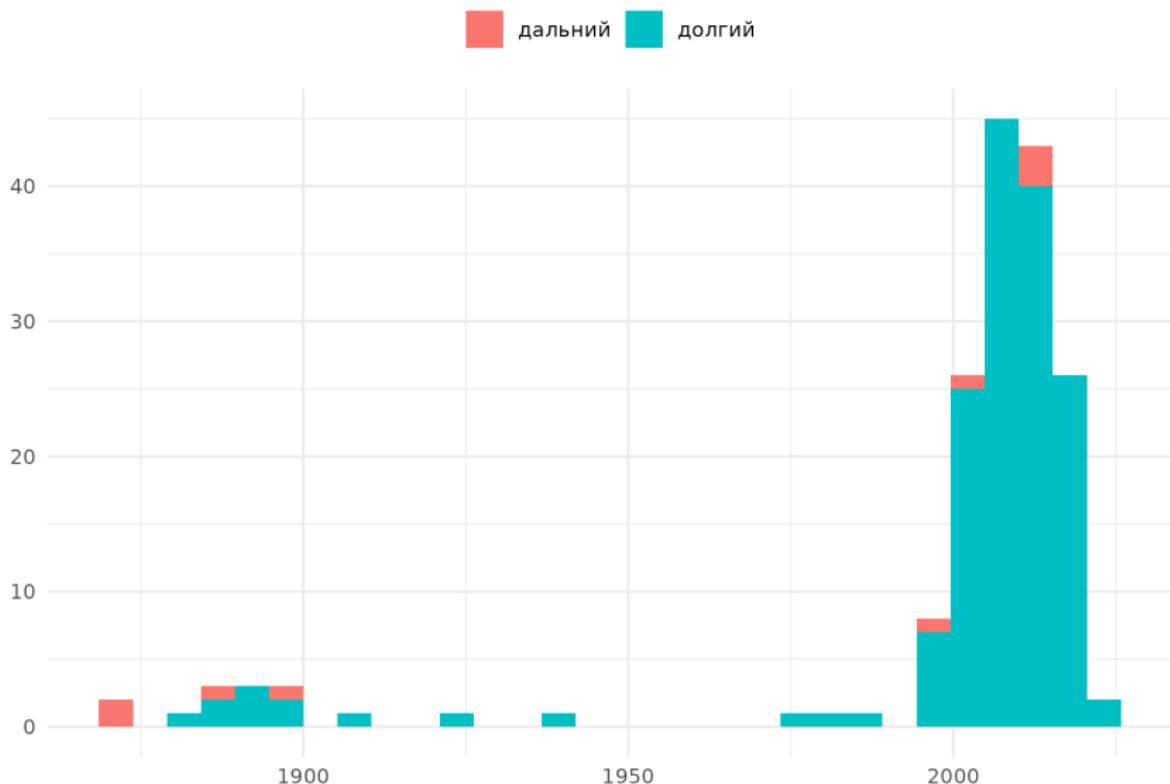
## *Отложить в ... ящик*

## Отложить в ... ящик

■ в дальний ящик ■ в долгий ящик



## *Отложить в ... ящик*



## Наши ресурсы

# Ресурсы Международной лаборатории языковой конвергенции

- [lingconlab.ru](http://lingconlab.ru)
- 22 устных диалектных корпуса
- 8 устных билингвальных корпусов
- 10 корпусов малых языков
- другие
  - словари (мегебский, рутульский, тукитинский, хваршинский, даргинский)
  - Типологический атлас языков Дагестана
  - Атлас многоязычия в Дагестане
  - Атлас рутульских диалектов
  - Корпус Просодии Русских Диалектов (ПРУД)
  - ...

## 22 устных диалектных корпуса

Корпус говора Хиславичского района  
260,793 ток.

Корпус устьянских говоров  
959,782 ток.

Корпус говора села Спиридонова Буда  
70,565 ток.

Корпус говора верхней Пинеги и Вии  
70,803 ток.

Корпус донских говоров  
71,600 ток.

Корпус говора деревни Веегора  
91,514 ток.

Корпус говора Мантуровского района Костромской области  
113,837 ток.

Корпус говоров низовья рек Лух и Теза  
146,350 ток.

Корпус говора села Кеба  
54,535 ток.

Корпус говоров среднего печеня Северной Двины  
68,010 ток.

Корпус опочецких говоров  
68,741 ток.

Корпус говора села Роговатка  
100,047 ток.

Корпус говора села Церковное  
39,469 ток.

Корпус говоров окрестностей Михайловска  
47,576 ток.

Корпус говоров Средней Печени  
63,270 ток.

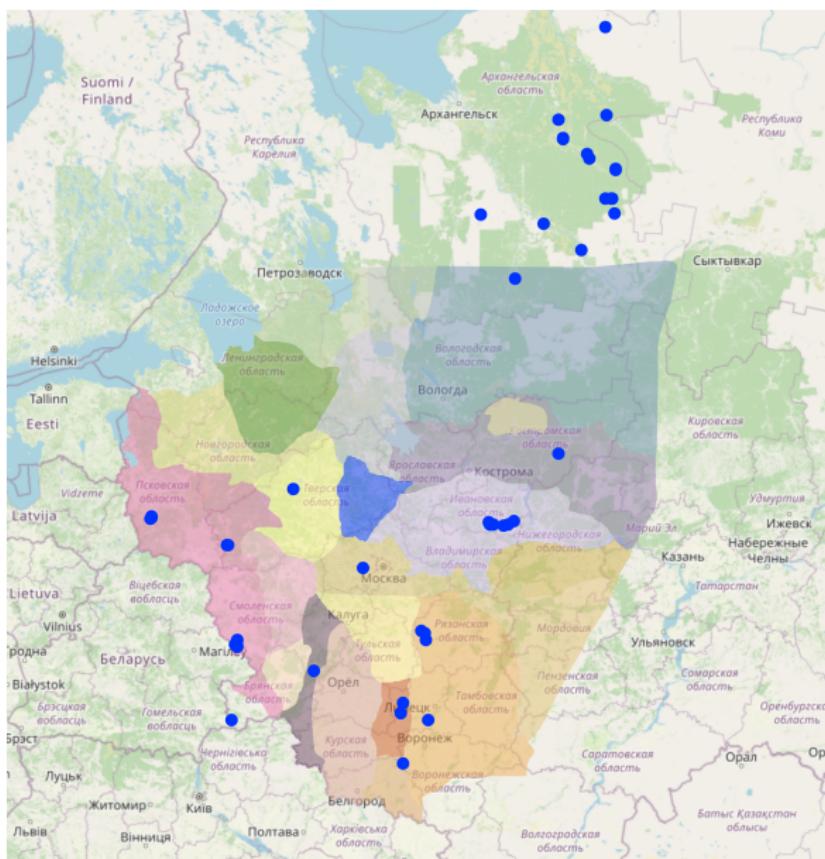
Корпус говора города Звенигород  
68,324 ток.

Корпус говора Средней Пёзы  
79,566 ток.

Корпус говора деревень Шетнево и Маниново  
95,335 ток.

Корпус говора села Малинино  
138,943 ток.

## 22 устных диалектных корпуса



## 8 устных билингвальных корпусов

Корпус дагестанского русского  
376,717 ток.

Якутско-русский корпус переключения кода  
15,139 ток.

Корпус русской речи Чувашии  
46,307 ток.

Корпус чыгаринского русского  
41,767 ток.

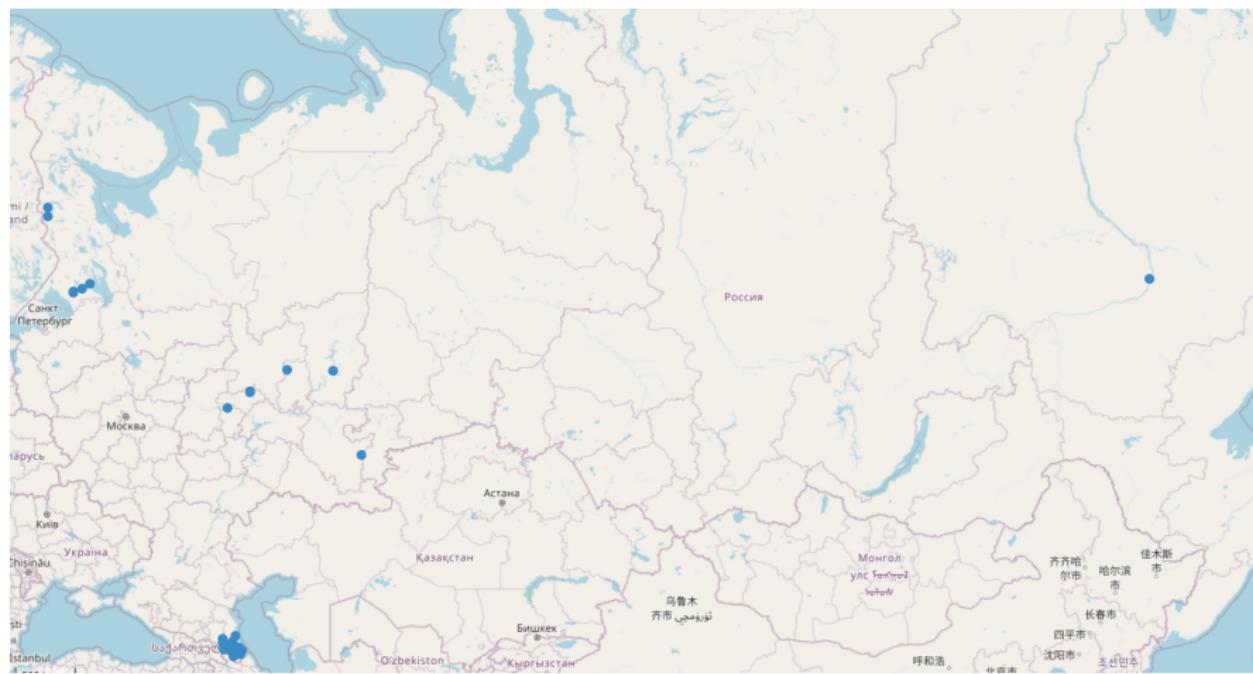
Корпус русской речи Карелии  
578,646 ток.

Корпус русской речи республики Марий Эл  
69,109 ток.

Корпус русской речи Башкирии  
93,127 ток.

Корпус русской речи бесермян  
97,216 ток.

## 8 устных билингвальных корпусов



# 10 корпусов малых языков

Устный корпус башкирского языка  
25,000 ток.

Устный корпус диалектов хакасского языка  
58,000 ток.

Устный корпус цыцильского говора польского языка  
5,113 ток.

Корпус тантынского даргинского  
2,683 ток.

Устный корпус абазинского языка  
3,636 ток.

Устный корпус муиринского даргинского  
6,935 ток.

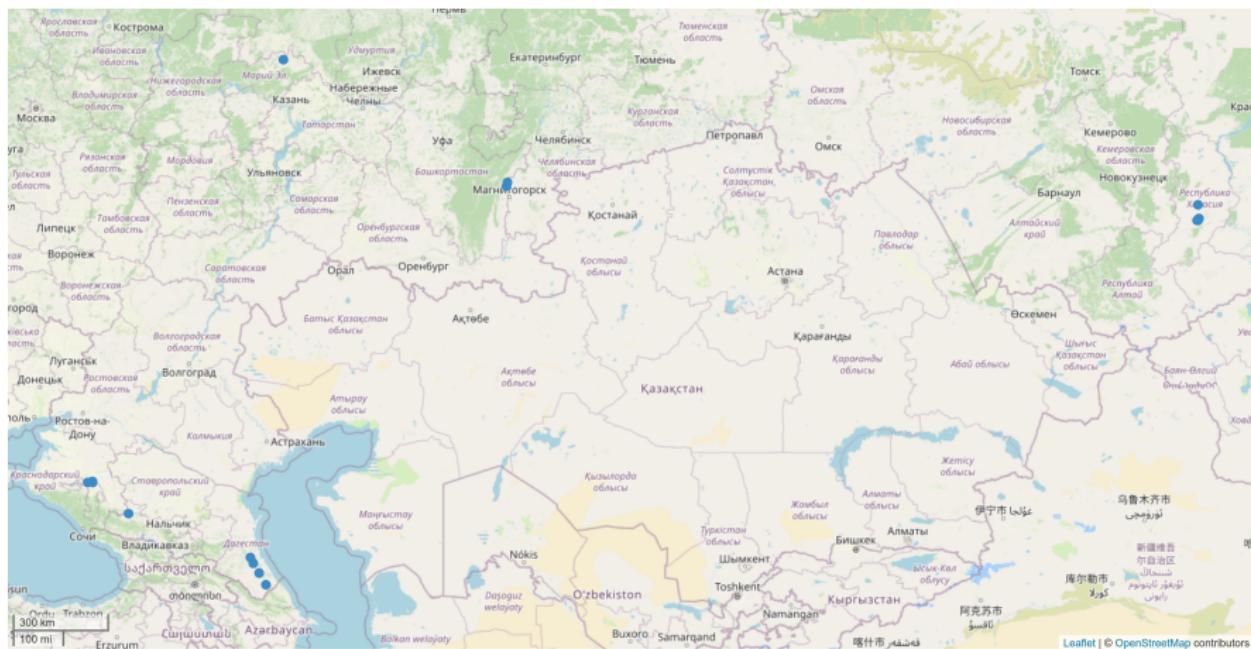
Устный корпус кадарского даргинского  
6,366 ток.

Устный корпус адыгейского языка  
9,128 ток.

Устный корпус баскенского диалекта испано-чересского языка  
7,020 ток.

Устный корпус лугового марийского языка  
11,647 ток.

# 10 корпсов малых языков



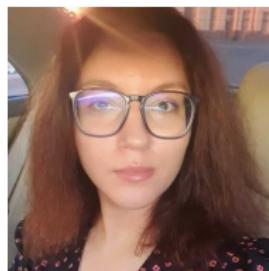
## Группа DiaL2



(a) М. В. Ермолова



(b) С. С. Земичева



(c) Н. А. Кошельюк



(d) Г. А. Мороз



(e) К. Наккарато



(f) А. В. Яковлева

РНФ (24-28-01097) "Исследование вариативности билингвального и  
диалектного русского языка на материале устных корпусов"

# Исследование билингвального русского

# Нестандартные количественные конструкции в речи билингвов

- система русских числительных сложная
- системы числительных в L1 доступных нам корпусах значительно проще
- количественные конструкции в речи билингвов исследовалась в работах [[Stoyanova, 2019, Стойнова, 2021](#)]
- В работе [[Стойнова, 2021](#)] употребление нестандартных конструкций объясняется контактом
- Увидим ли мы такой же эффект на основе данных наших корпусов?

## Данные

- Сначала мы автоматически отобрали **7,376** контекстов
- Для анализа мы отобрали **1,748** примеров

- (1) *Пешком ходил Верхний Дженгутай пять километра.* (дагест.)
- (2) *Этот меньше, после двое **аборт** делала одну.* (марийский)

- Примеры размечены по некоторым параметрам
  - лингвистическим
    - **коллокационность** комбинации числительного + существительного
    - тип числительного (собирательные *двое, трое*, паукальные *два, три*, другие)
  - социолингвистическим
    - год рождения
    - пол
    - образование
    - первый язык

# Данные

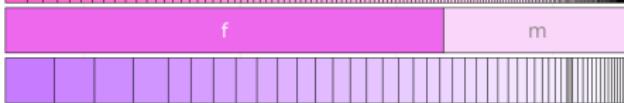
(1.1) corpora: 7



(1.2) speakers: 188



(1.3) gender



(1.4) year of birth



(1.5) L1: 21



(1.6) L1 family



(1.7) education



(1.8) standardness of the speaker



(2.1) marking



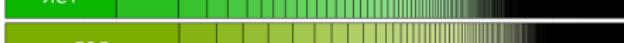
(2.2) numeral



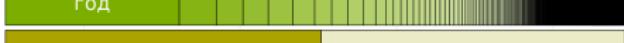
(2.3) noun token



(2.4) noun lemma



(2.5) numeral type



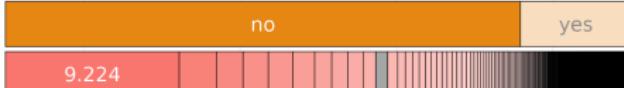
(2.6) noun type



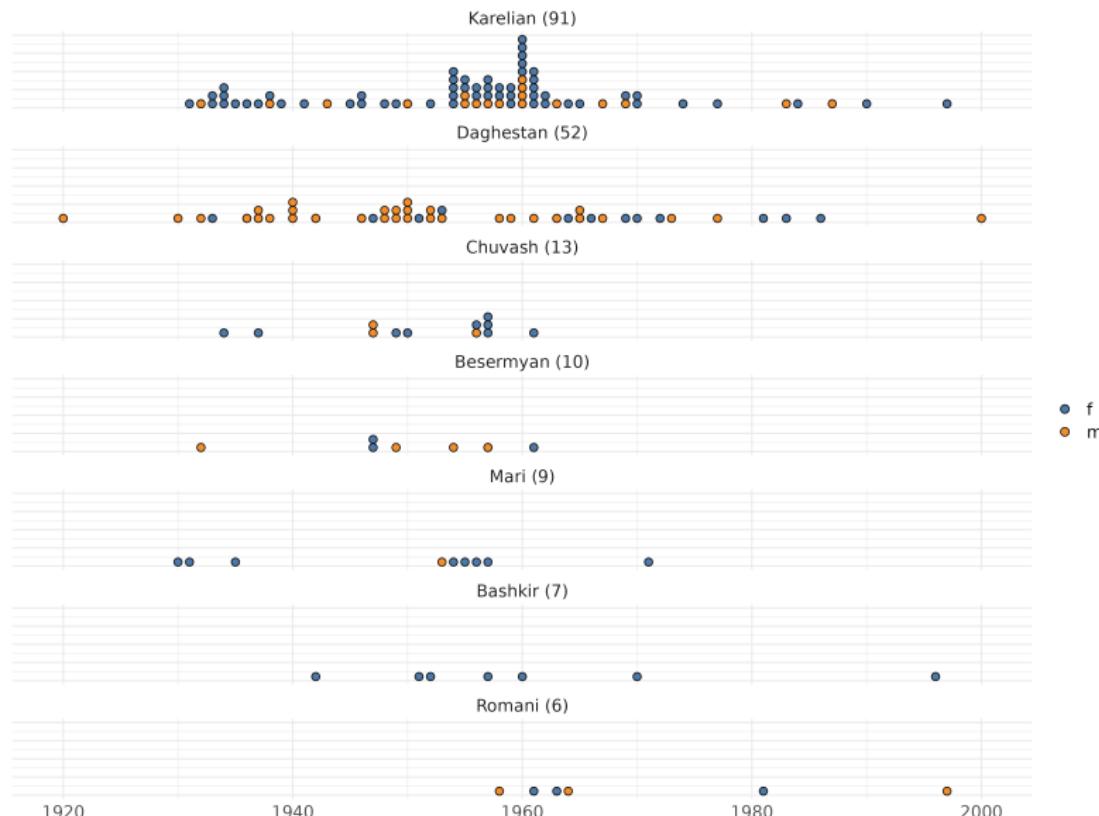
(2.7) ambiguous



(2.8) dice coefficient



# К сожалению, данные очень разнородные

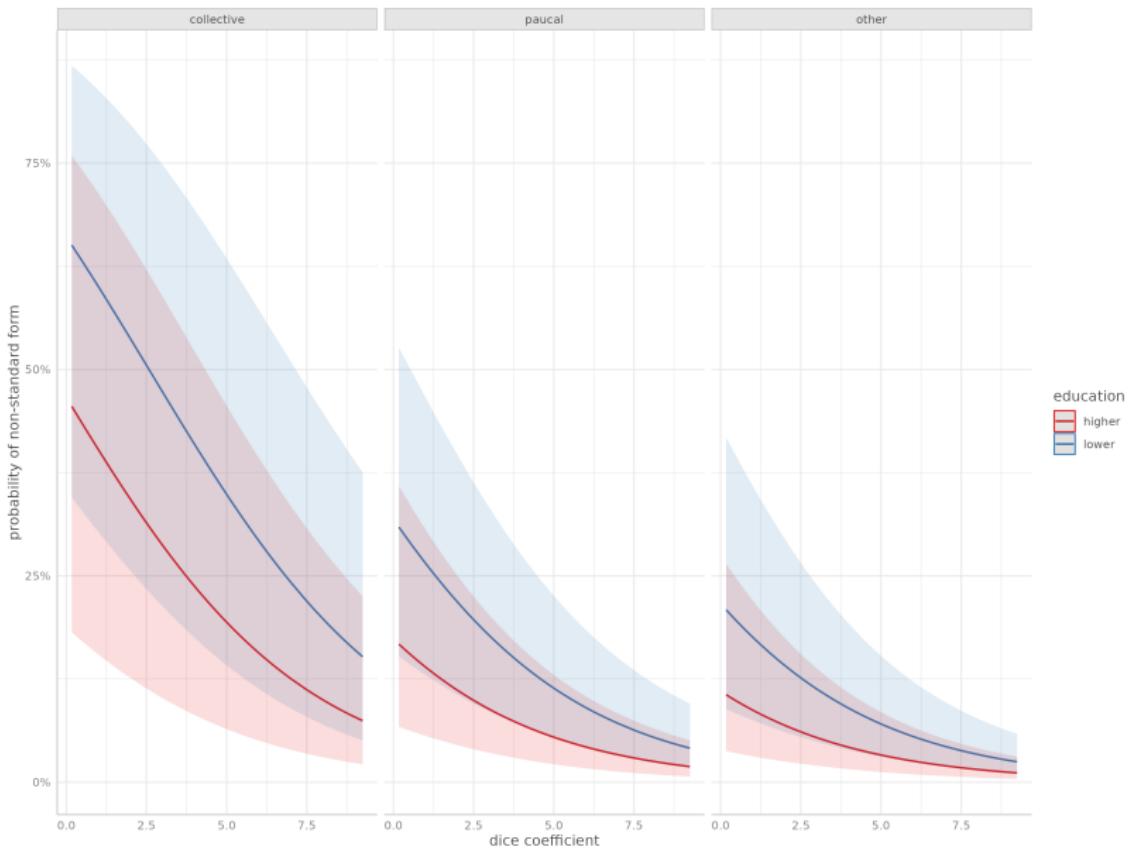


## Моделирование

Мы запустили иерархическую логистическую регрессию со смешанными эффектами, предсказывая вероятность нестандартной формы

- основные эффекты
  - коллокационность \*\*\*
  - тип числительного \*\*\*
  - образование \*
  - год рождения \*
- случайные эффекты
  - носитель вложен в первый язык

# Предсказания модели



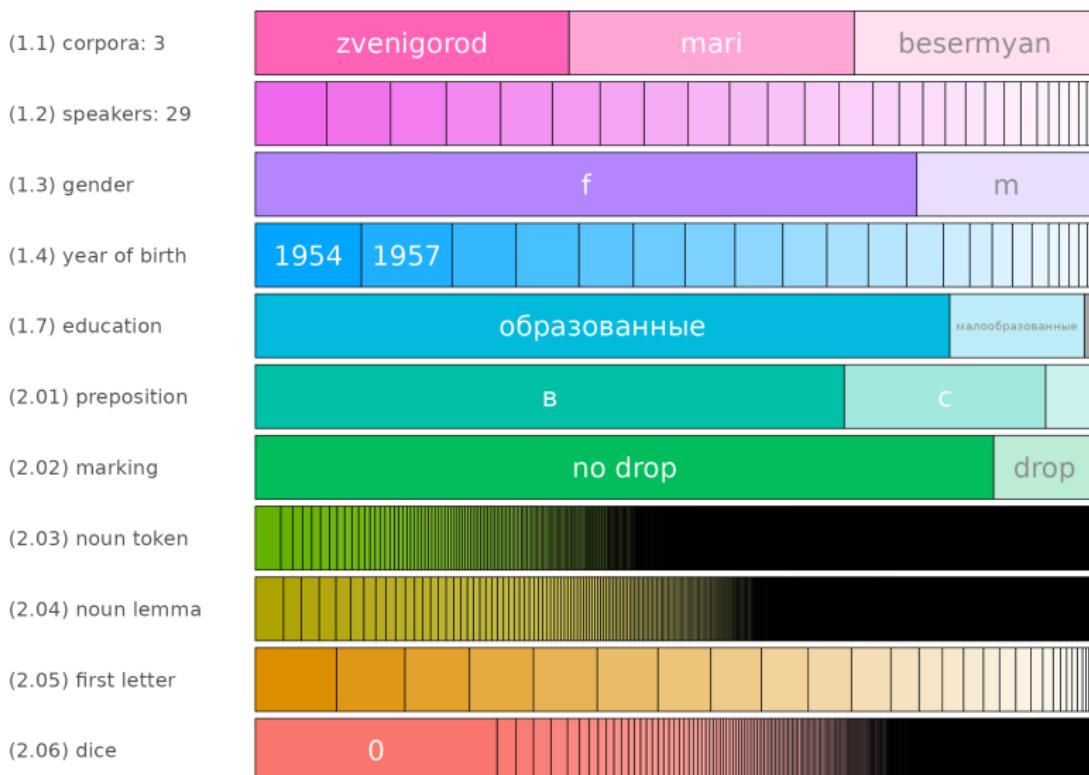
## Выпадение предлогов в речи билингвов

- Для анализа мы отобрали 4990 контекстов из трех корпусов: бесермянского (1435), марийского (1693), звенигородского (1863):
- *Со второго курса что ли практика началась, \_ больнице.*  
(марийский русский)
- *Вот, отремонтировал\_ трудом пополам, китайские часы -то.*  
(бесермянский русский)
- *я пошёл, \_ начальнику дал предложение.* (бесермянский  
русский)

# Выпадение предлогов в речи билингвов

- Примеры размечены по некоторым параметрам
  - есть ли опущение предлога
  - тип предлога: *в, с, к*
  - лингвистическим
    - коллокационность комбинации предлога + существительного
    - первый согласный/гласный существительного
  - социолингвистическим
    - год рождения
    - гендер
    - образование
    - первый язык

# Данные

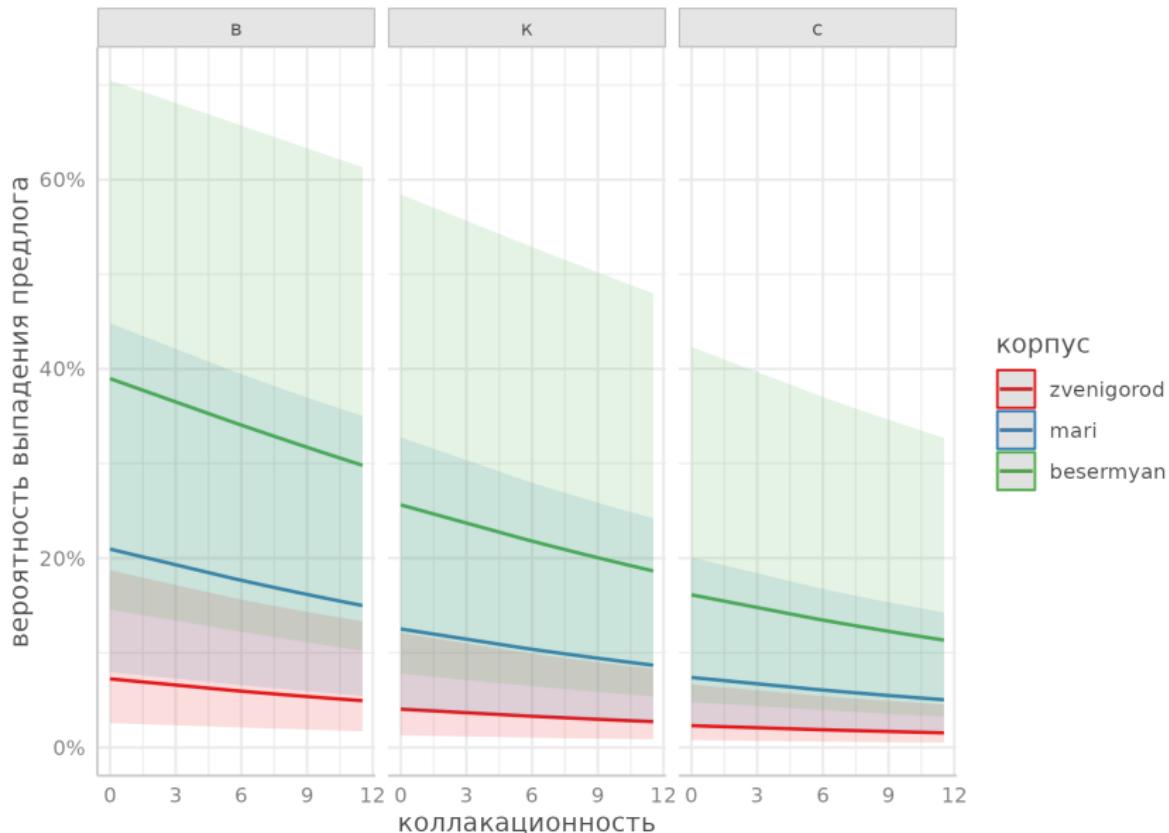


## Моделирование

Мы запустили иерархическую логистическую регрессию со смешанными эффектами, предсказывая вероятность выпадения предлога:

- основные эффекты
  - коллокационность \*\*
  - предлог \*\*
  - год рождения \*\*
  - образование
  - гендер
  - корпус \*
- случайные эффекты
  - носитель

## Предсказания модели



## Заключение

## Список литературы I

- N. Stoynova. Russian in contact with southern tungusic languages:  
Evidence from the contact russian corpus of northern siberia and the  
russian far east. *Slavica Helsingiensia*, 52, 2019.
- Н. Стойнова. Нестандартные количественные конструкции в  
русской речи носителей нанайского и ульчского языков. *Russian  
Linguistics*, 45, 2021.