Вариативность в русском языке билингвов

для курса «Основные приложения математики», НИУ ВШЭ

Г. А. Мороз
Международная лаборатория языковой конвергенции

23.04.2024

Прескриптивное vs. дескриптивное

#тыжлингвист

- знает как правильно: писать, употреблять слова/выражения, ...

#тыжлингвист

- знает как правильно: писать, употреблять слова/выражения, ...
- умеет читать на всех письменностях мира

#тыжлингвист

- знает как правильно: писать, употреблять слова/выражения, ...
- умеет читать на всех письменностях мира
- знает все языки на свете

#тыжлингвист

- знает как правильно: писать, употреблять слова/выражения, ...
- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух

#тыжлингвист

- знает как правильно: писать, употреблять слова/выражения, ...
- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова

#тыжлингвист

- знает как правильно: писать, употреблять слова/выражения, ...
- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- не может разбираться в статистике и программировании

#тыжлингвист

- знает как правильно: писать, употреблять слова/выражения, ...
- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- не может разбираться в статистике и программировании
- все перечисленное выше — чушь

Обо мне

- полевой исследователь (30 поездок, почти все на Кавказ)
- фонетист, фонолог, квантитативный лингвист, занимаюсь линвистической географией
- преподаю статистику и R (язык программирования)
- написал несколько лингвистических пакетов для R
  - `lingtypology`
  - `phonfieldwork`
  - `lingglosses`

Прескриптивная vs. дескриптивная лингвистика

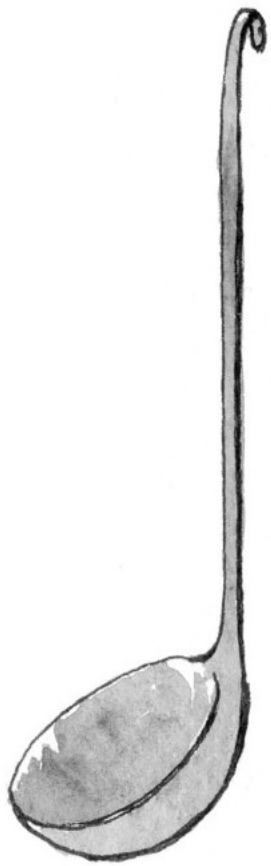
- прескриптивная

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.
  - ...

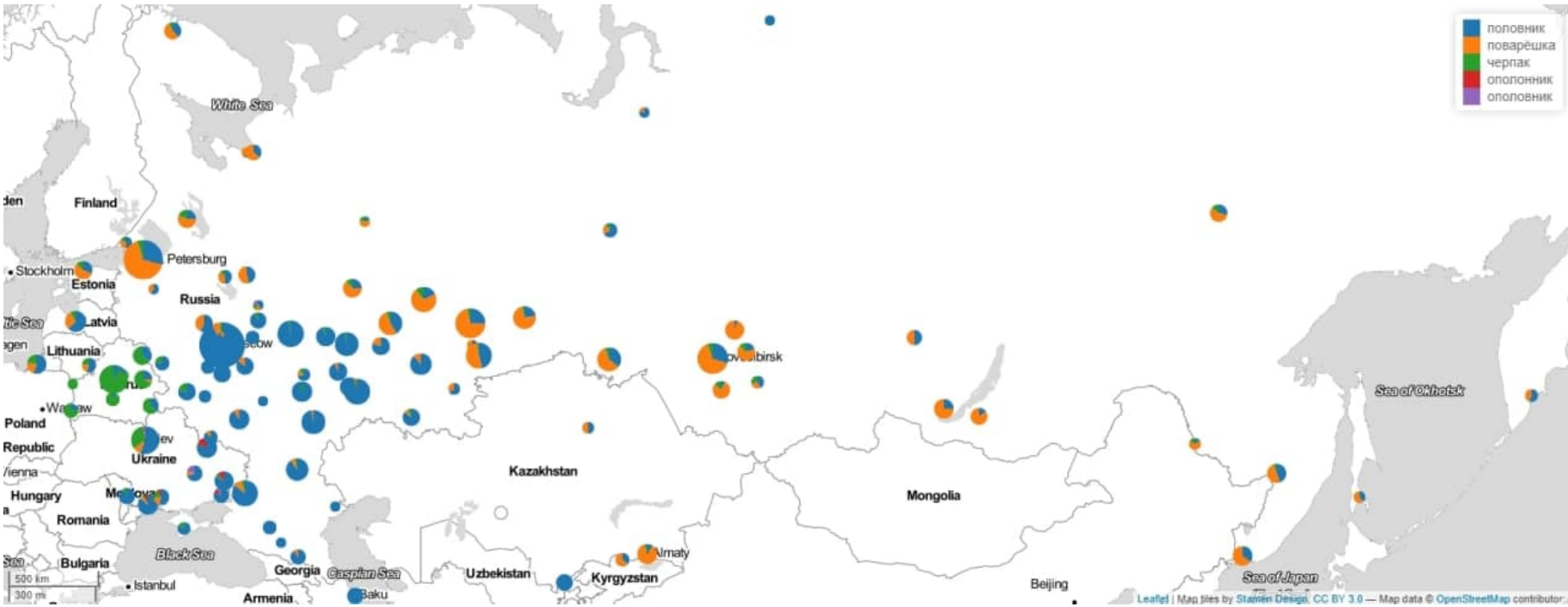
Прескриптивная vs. дескриптивная лингвистика

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



Прескриптивная vs. дескриптивная лингвистика

Это часть опроса И. Левина 2021 года:



половник
поварёшка
черпак
ополовник
ополонник

Корпусная лингвистика

Корпусная лингвистика

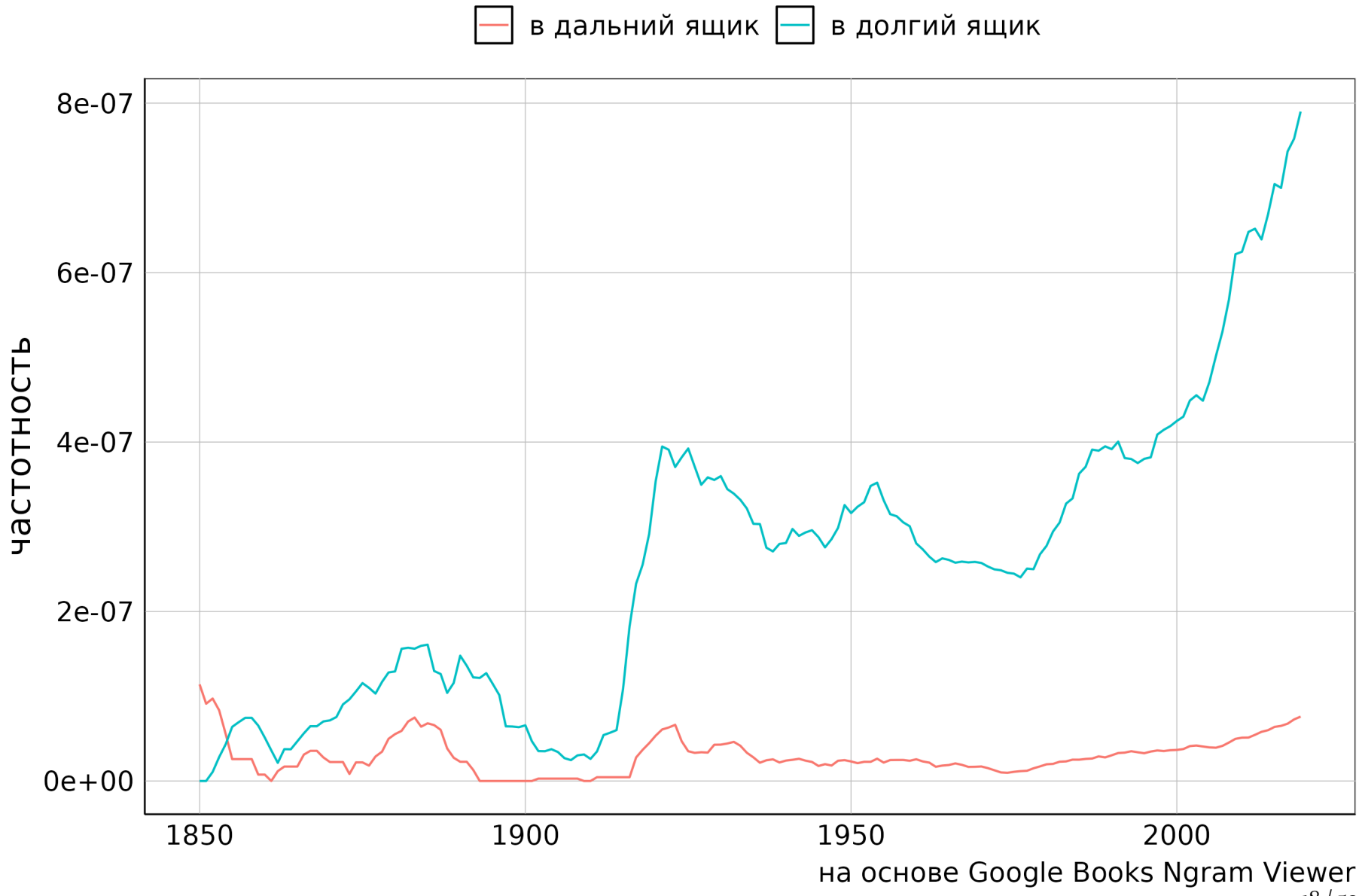
Корпусная лингвистика — это область лингвистики, которая занимается исследованием языковых явлений на материале некоторых собраний языкового материала. В большинстве случаев это письменные тексты, однако это может быть аудио и даже видео корпуса.

Среди корпусов русского языка можно назвать:

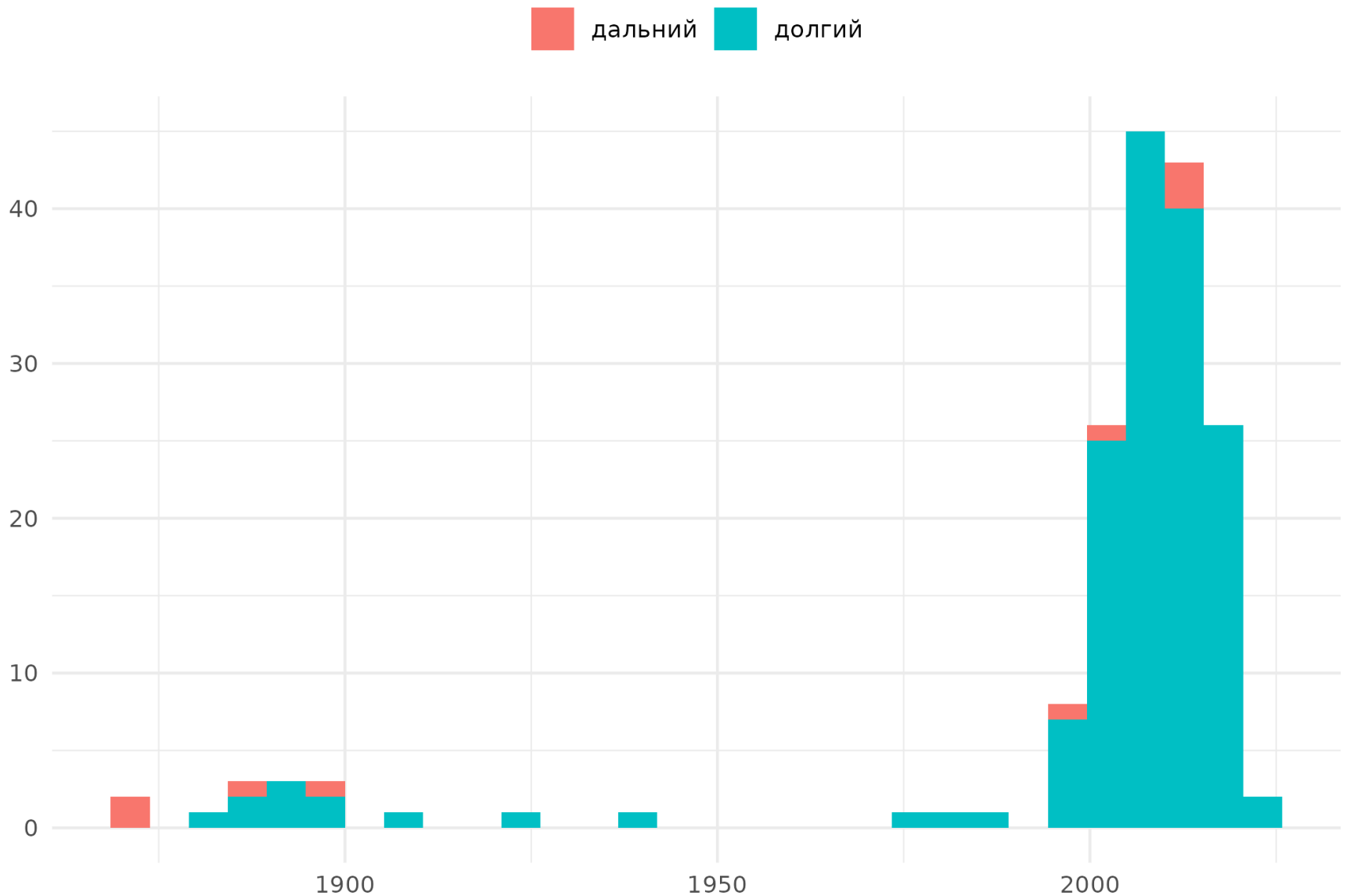
- Национальный корпус русского языка
  - более 1.5 млрд слов
  - много подкорпусов (газетный, устный, параллельный, диалектный, поэтический, исторические)
- Google Books Ngram Viewer
- ...

Отложить в ... ящик

Отложить в ... ящик



Отложить в ... ящик



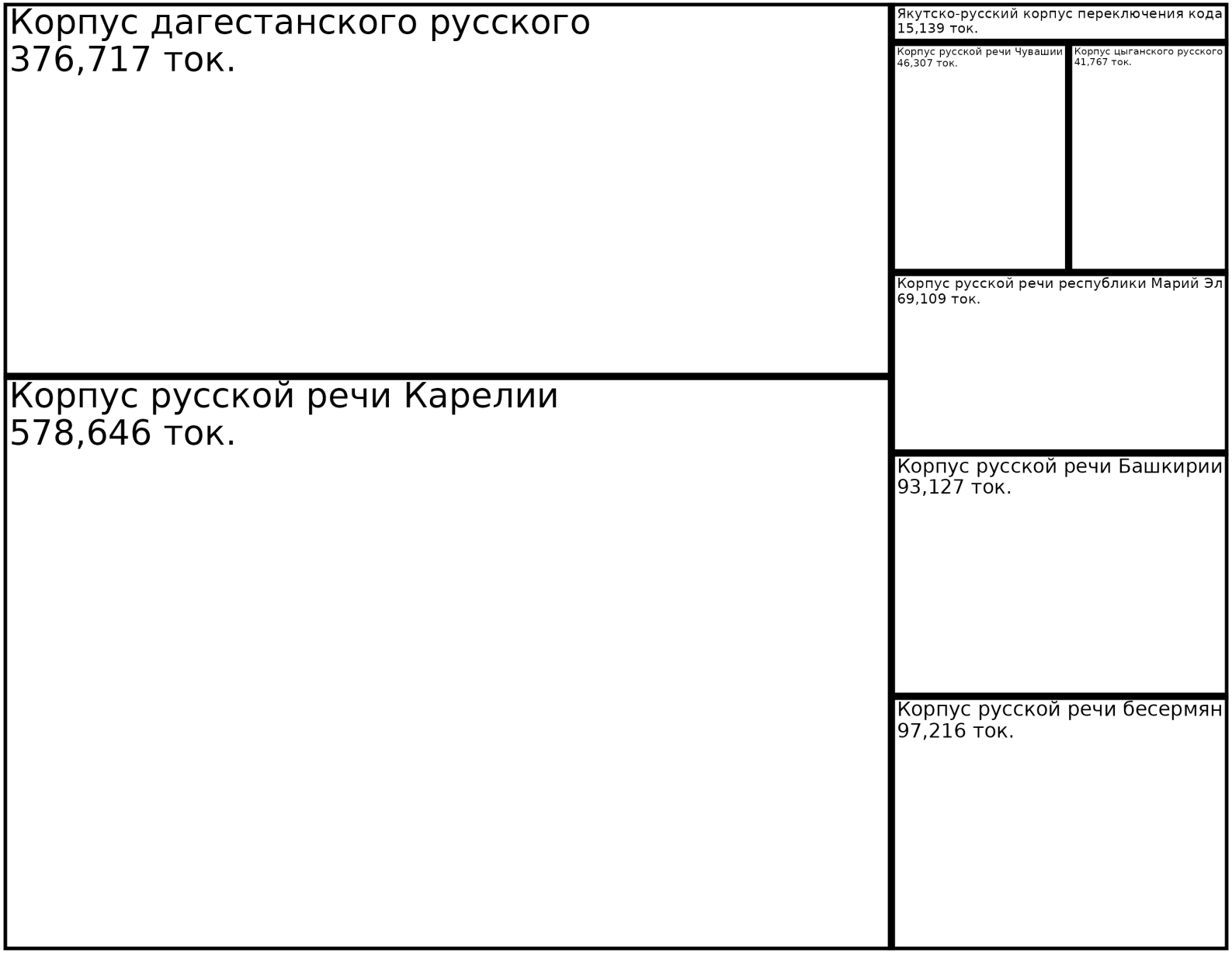
на основе Национального корпуса русского языка

Наши ресурсы

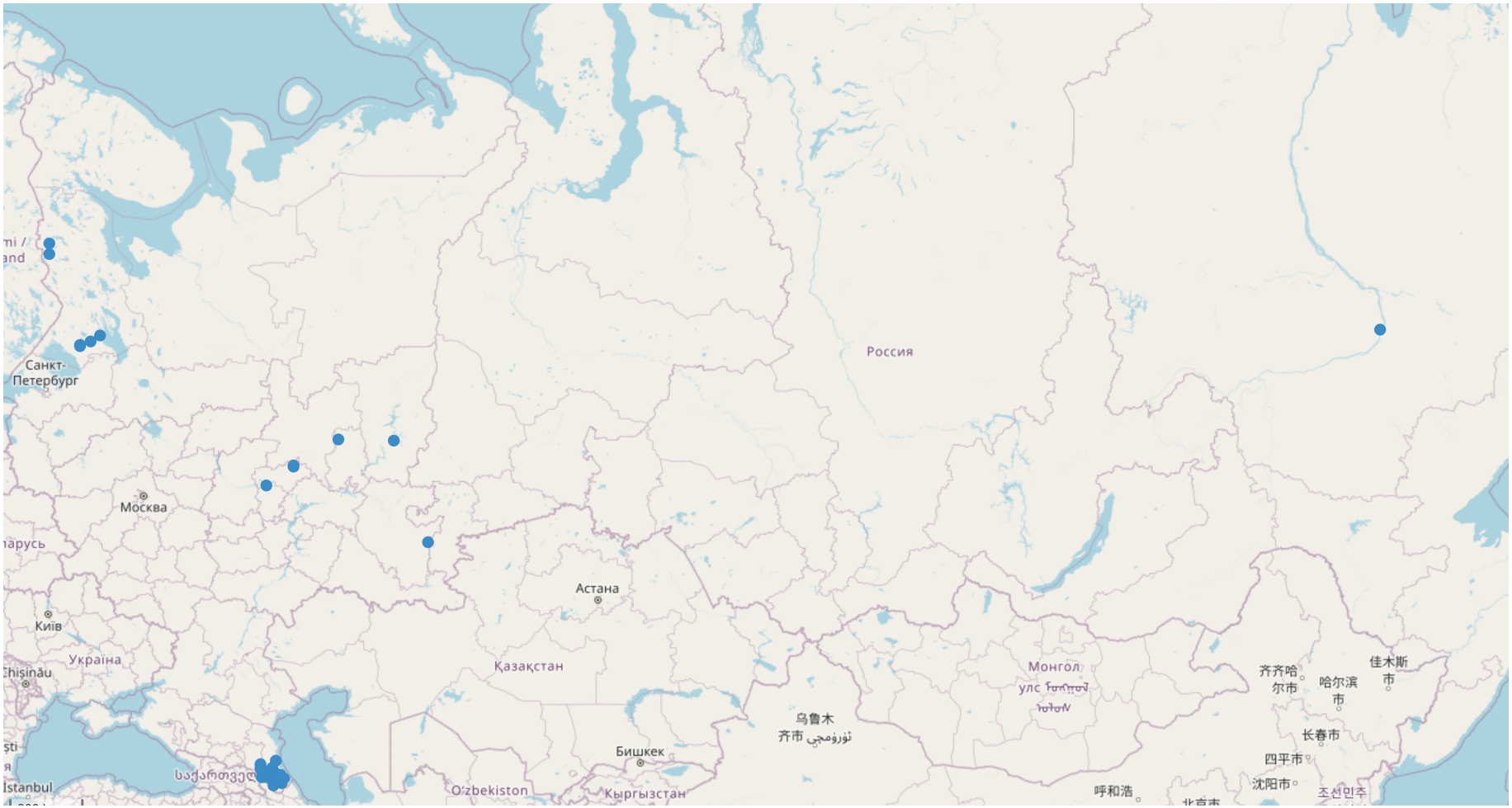
Ресурсы Международной лаборатории языковой конвергенции

- 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)
  - ...

8 устных билингавльных корпусов



8 устных билингавльных корпусов



Группа DiaL2

(a) М. В. Ермолова

(b) С. С. Земичева

(c) Н. А. Кошелюк

(d) Г. А. Мороз

(e) К. Наккарато

(f) А. В. Яковлева

Исследование билингвального русского

Нестандартные количественные конструкции в речи билингвов

- система русских числительных сложная
- системы числительных в L1 доступных нам корпусов значительно проще
- количественные конструкции в речи билингвов исследовались в работах [Stoynova, 2019, Стойнова, 2021]
- В работе [Стойнова, 2021] употребление нестанадартных 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

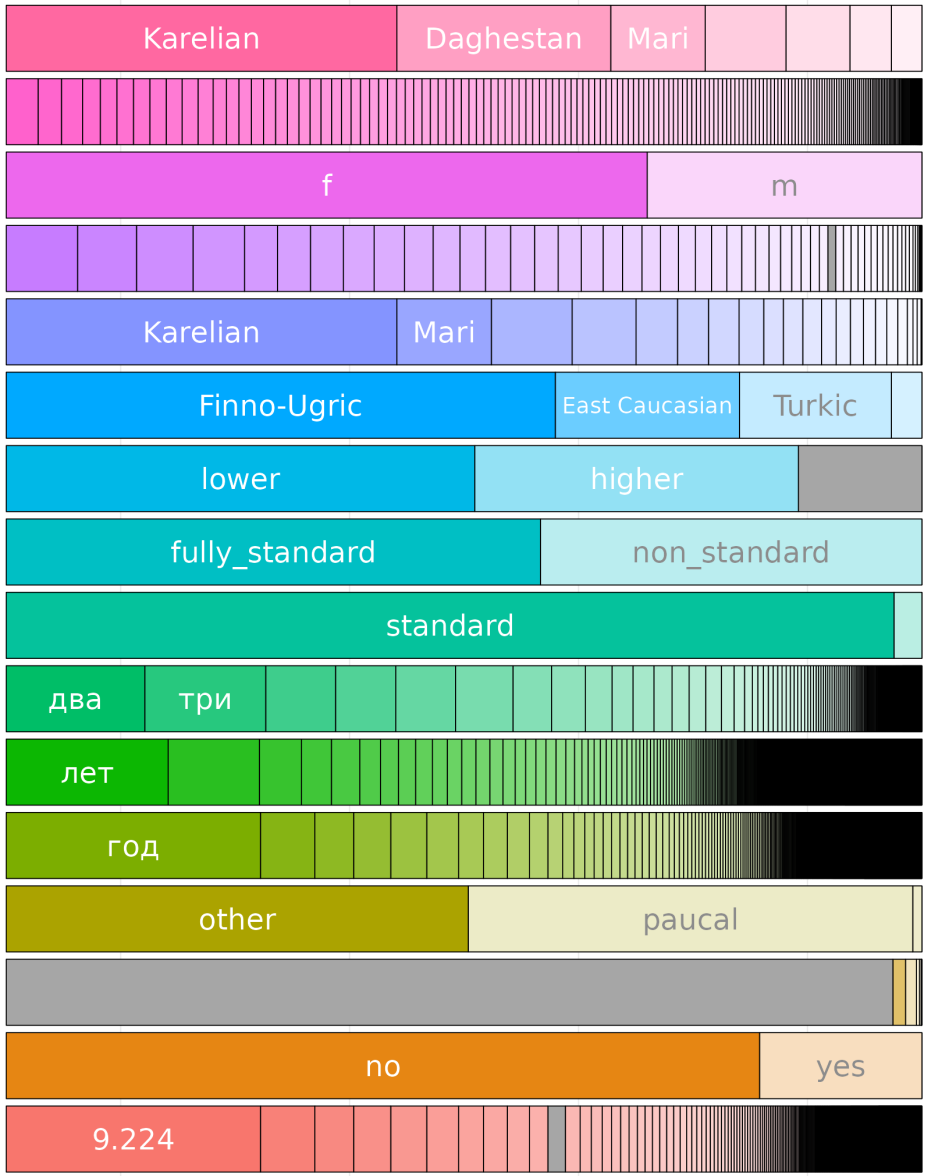
(2.4) noun lemma

(2.5) numeral type

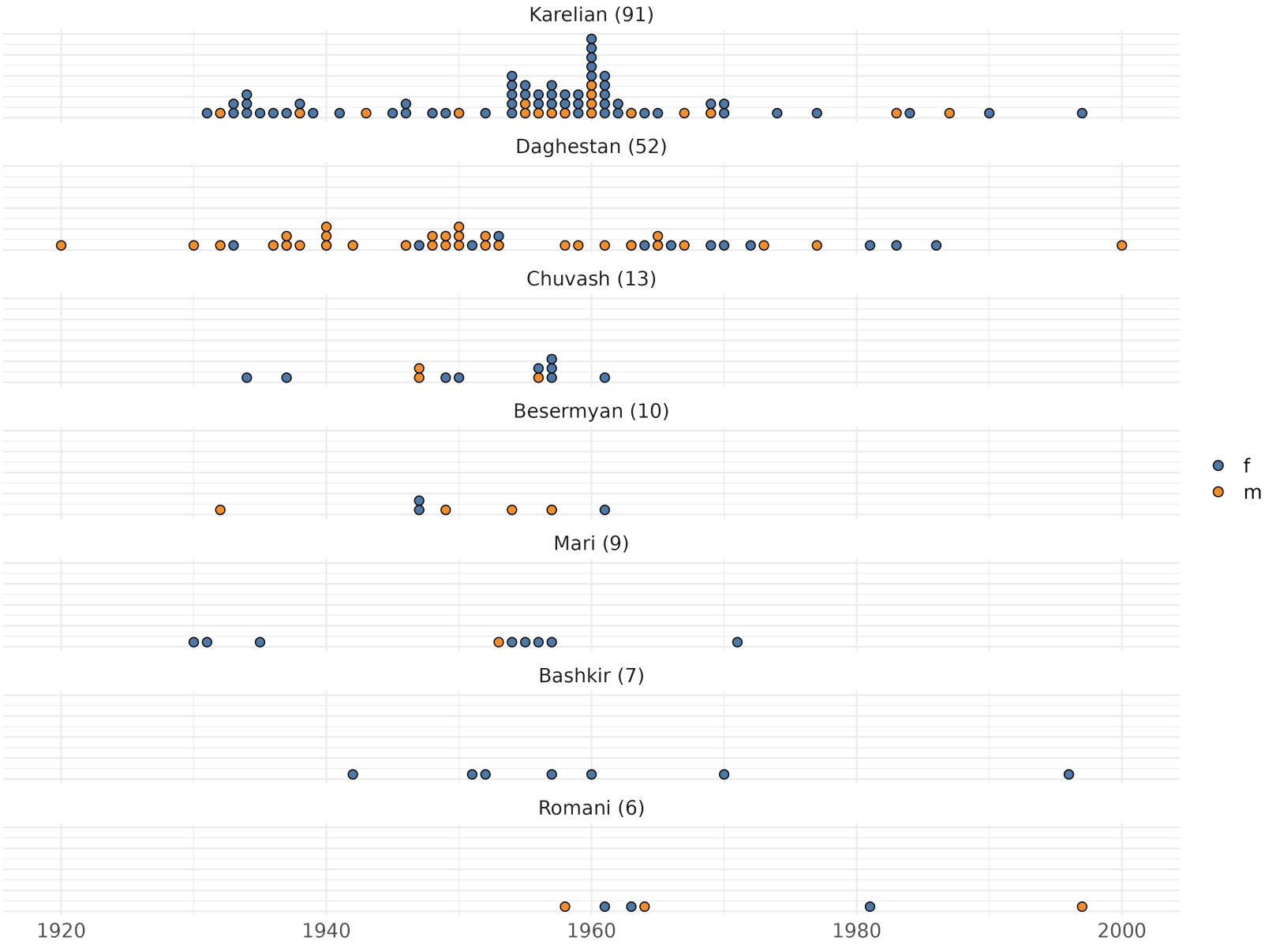
(2.6) noun type

(2.7) ambiguous

(2.8) dice coefficient



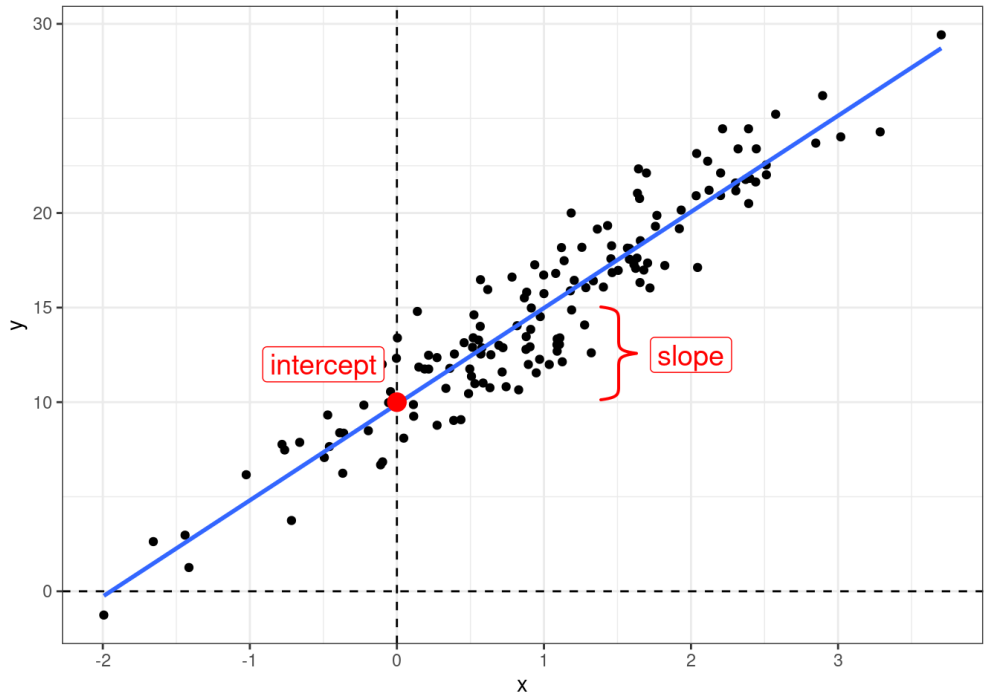
К сожалению, данные очень разнородные



Моделирование

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

Моделирование: регрессия



- свободный член (intercept) – значение y при $x = 0$;
- угловой коэффициент (slope) – изменение y при изменении x на одну единицу.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_i + \epsilon_i$$

Моделирование: множественная регрессия

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_{1i} + \cdots + \hat{\beta}_n \times x_{ni} + \epsilon_i,$$

- x_{ki} — i-ый элемент векторов значений $X_1, \ldots, X_n$;
- y_i — i-ый элемент вектора значений Y;
- $\hat{\beta}_0$ — оценка случайного члена (intercept);
- $\hat{\beta}_k$ — коэфциент при переменной X_k;
- ϵ_i — i-ый остаток, разница между оценкой модели $(\hat{\beta}_0 + \hat{\beta}_1 \times x_i)$ и реальным значением y_i; весь вектор остатков иногда называют случайным шумом.

Моделирование: регрессия со смешенными эффектами

$$y_i = X_i \times \hat{\beta} + Z_i \times \hat{b}_n + \epsilon_i,$$

- y_i — i-ый элемент вектора значений Y;
- X_i — i-ый элемент из множества переменных $X_1, \dots, X_n$;
- Z_i — i-ый элемент из множества группирующих переменных случайных эффектов $Z_1, \dots, Z_n$;
- $\hat{\beta}$ — оценка коэффициентов при переменных X;
- $\hat{b}_i$ — случайный нормальнораспределенный шум Z_i;
- ϵ_i — i-ый остаток, разница между оценкой модели и реальным значением y_i; весь вектор остатков иногда называют случайным шумом.

Моделирование: логистическая регрессия

В регрессии мы хотим чего-то такого:

$$\underbrace{y}_{[-\infty,+\infty]} = \underbrace{\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \varepsilon_i}_{[-\infty,+\infty]}$$

Однако если y — бинарные ответ, то, получается чушь. Поэтому используют логарифм шансов:

Шансы — отношение количества успехов к количеству неудач:

$$odds = \frac{p}{1-p} = \frac{p(\text{успеха})}{p(\text{неудачи})}, odds \in [0, +\infty]$$

Натуральный логарифм шансов:

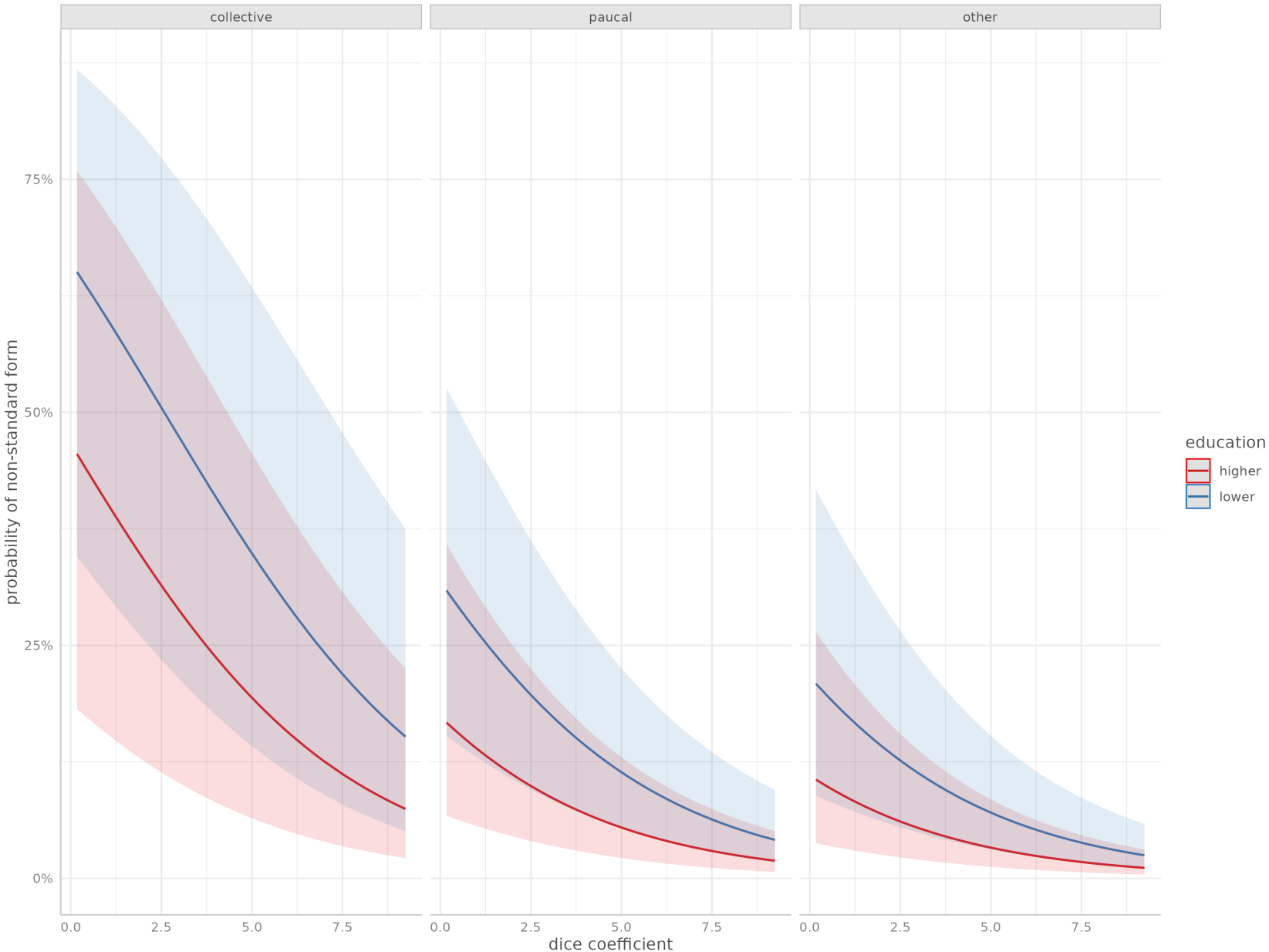
$$\log(odds) \in [-\infty, +\infty]$$

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 5005 контекстов из трех корпусов: бесермянского (1438), марийского (1707), звенигородского (1860):
- *Со второго курса что ли практика началась, _ больнице.* (марийский русский)
- *Вот, отремонтировал _ трудом пополам, китайские часы -то.* (бесермянский русский)
- *Жених приходит _ бабе.* (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 звук следующего слова
  - социолингвистическим
    - год рождения
    - пол
    - количество лет обучения
    - первый язык

Данные

(1.1) corpora: 3

(1.2) speakers: 29

(1.3) gender

(1.4) year of birth

(1.5) years of education

(2.1) preposition lemma

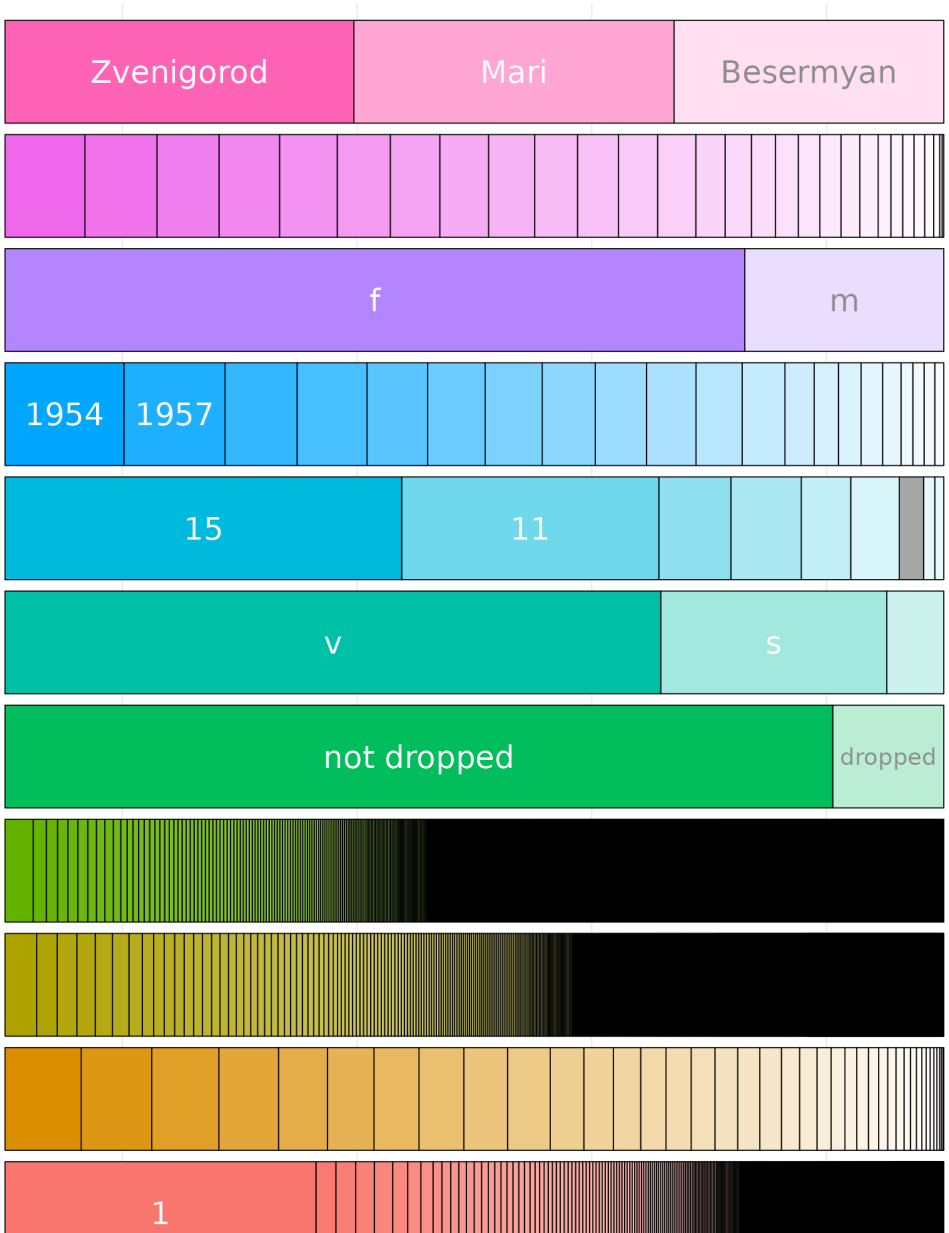
(2.2) preposition drop

(2.3) folowing wordform's token

(2.4) folowing wordform's lemma

(2.5) folowing wordform's first sound

(2.6) dice coefficient

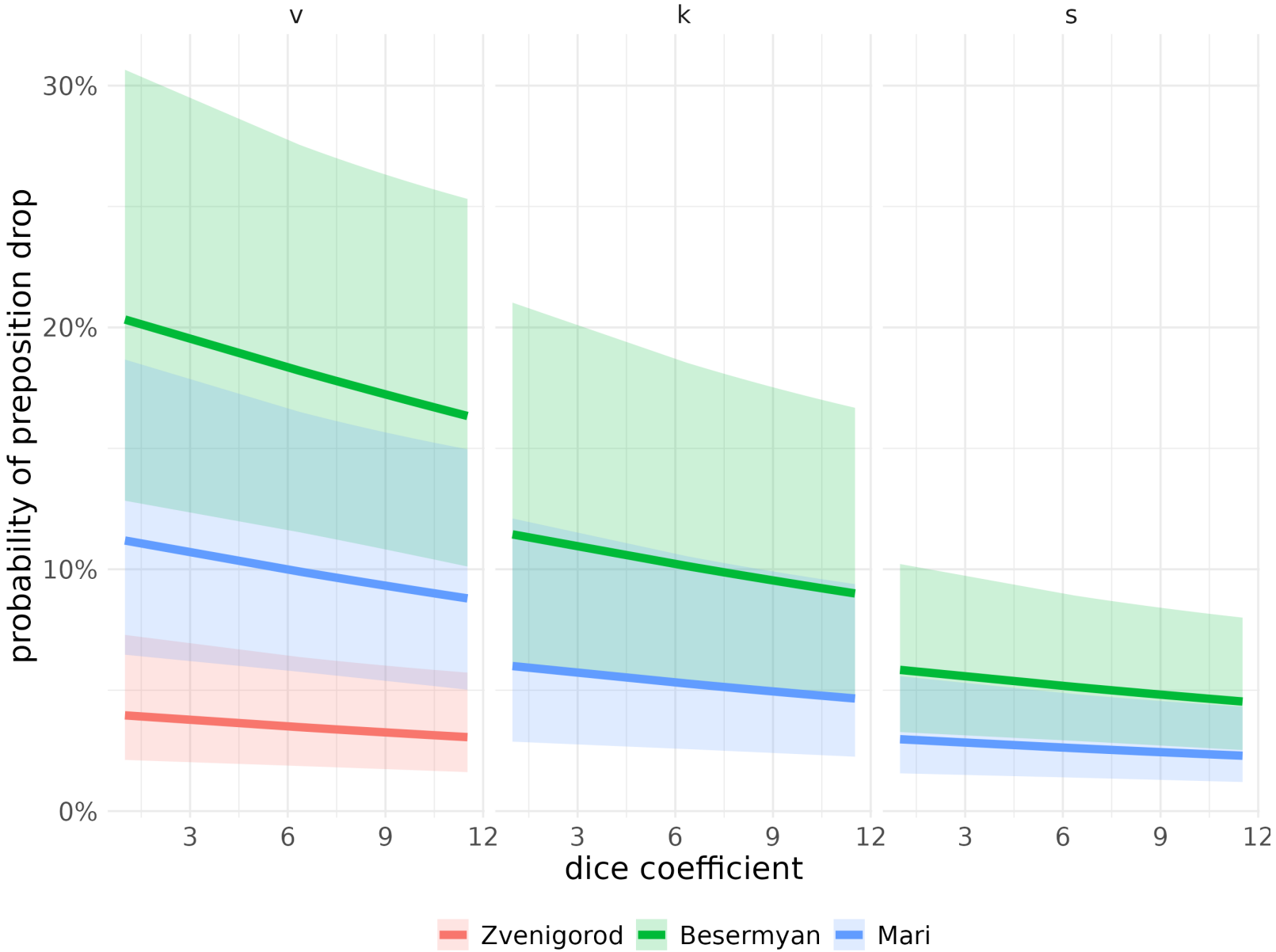


Моделирование

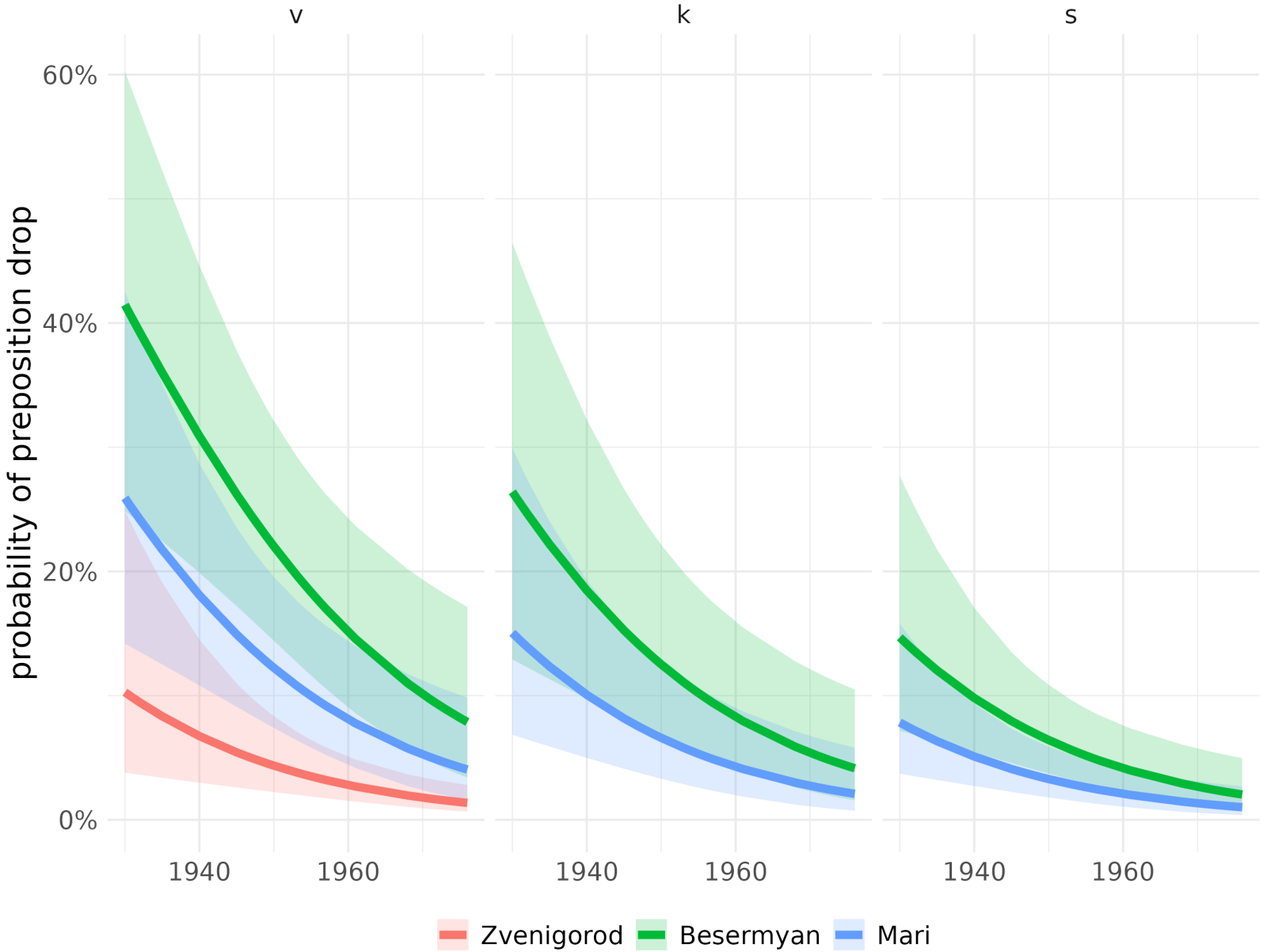
Мы запустили иерархическую логистическую регрессию со смешанными эффектами, предсказывая **вероятность выпадения предлога:**

- основные эффекты
  - коллокационность **
  - предлог **
  - год рождения **
  - корпус *
- случайные эффекты
  - носитель

Предсказания модели



Предсказания модели



Согласие разметчиков

Согласие разметчиков

Когда мы анализировали выпадение предлогов, мы слушали аудиозаписи.

- сравнение того, насколько решение аннотатоторов было единообразное делается при помощи мер согласия (каппа Коэна, каппа Фляйса, Intra-class correlation coefficient)

Согласие разметчиков

Когда мы анализировали выпадение предлогов, мы слушали аудиозаписи.

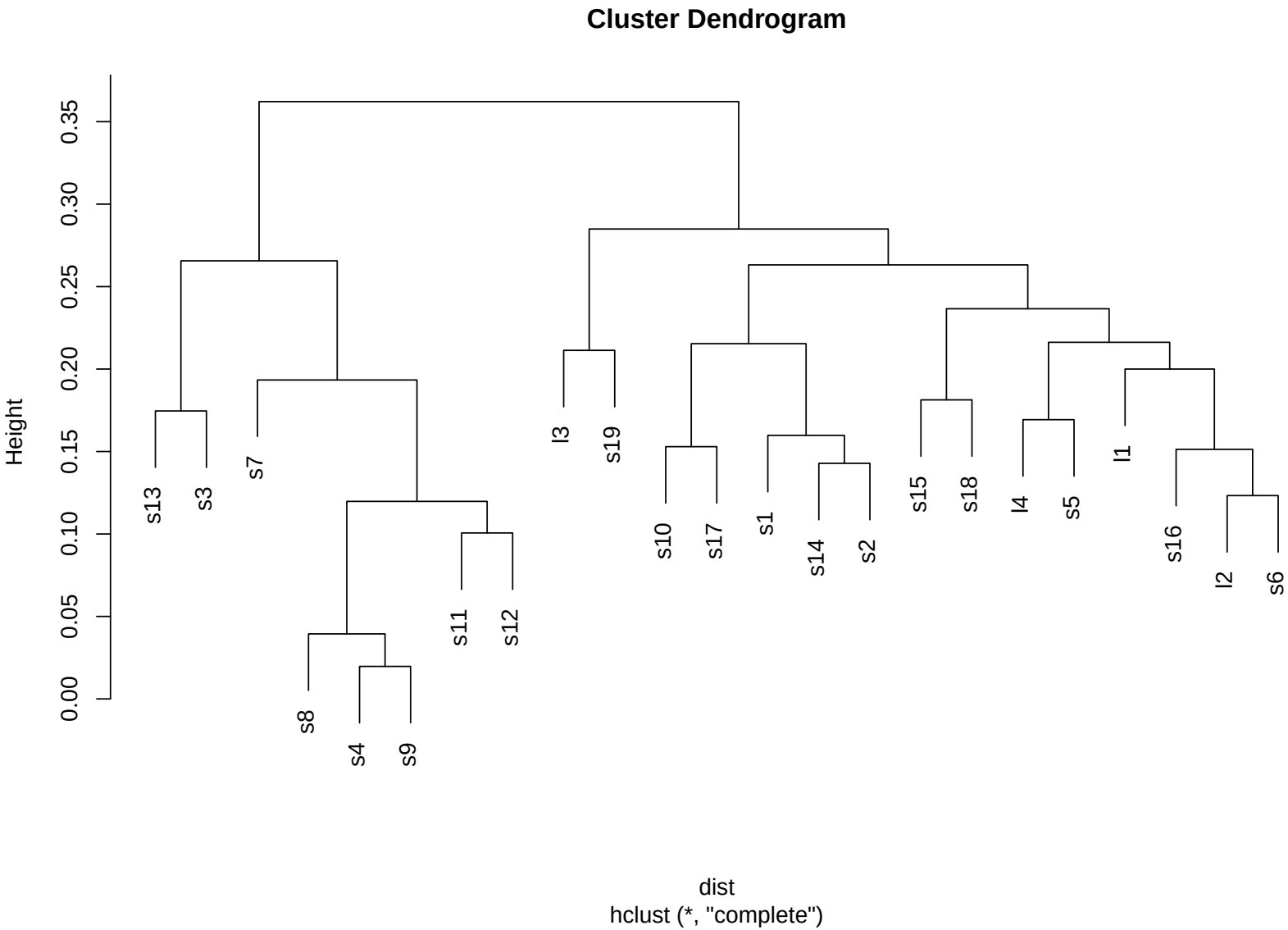
- сравнение того, насколько решение аннотатоторов было единообразное делается при помощи мер согласия (каппа Коэна, каппа Фляйса, Intra-class correlation coefficient)
- мы попросили студентов прошлого года курса “Основные приложения математики” поучаствовать с разметкой

Согласие разметчиков

Когда мы анализировали выпадение предлогов, мы слушали аудиозаписи.

- сравнение того, насколько решение аннотатоторов было единообразное делается при помощи мер согласия (каппа Коэна, каппа Фляйса, Intra-class correlation coefficient)
- мы попросили студентов прошлого года курса “Основные приложения математики” поучаствовать с разметкой
- а потом мы запустили кластеризацию

Согласие разметчиков



Заключение

- Методы корпусной лингвистики позволяют получать и анализировать данные для документации языковой вариативности

Заключение

- Методы корпусной лингвистики позволяют получать и анализировать данные для документации языковой вариативности
- В обоих независимых исследованиях мера коллокационности оказывалась существенной при моделировании вариативности

Заключение

- Методы корпусной лингвистики позволяют получать и анализировать данные для документации языковой вариативности
- В обоих независимых исследованиях мера коллокационности оказывалась существенной при моделировании вариативности
- Это соотносится с идеей, что частотность единиц играет важную роль в становлении языка у детей и в изменении языка [Bybee, 2006, Bybee and Hopper, 2001, Tomasello, 2005, Wolter and Gyllstad, 2013].

Список литературы I

J. L. Bybee and P. J. Hopper. Introduction. In J. L. Bybee and P. J. Hopper, editors, *Frequency and the emergence of linguistic structure*, pages 1–26. John Benjamins Publishing Company, 2001.

Joan Bybee. *Frequency of use and the organization of language*. Oxford University Press, 2006.

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.

Michael Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press, 2005.

Список литературы II

Brent Wolter and Henrik Gyllstad. Frequency of input and l2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition*, 35(3):451–482, 2013.

Н. Стойнова. Нестанадртные количественные конструкции в русской речи носителей нанайского и ульчского языков. *Russian Linguistics*, 45, 2021.