Лингвистические исследования вариативности

НИС Непараметрика и другие сюжеты статистики

Г. А. Мороз

11.03.2024

Мифы о лингвистике

#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

Лингвистика

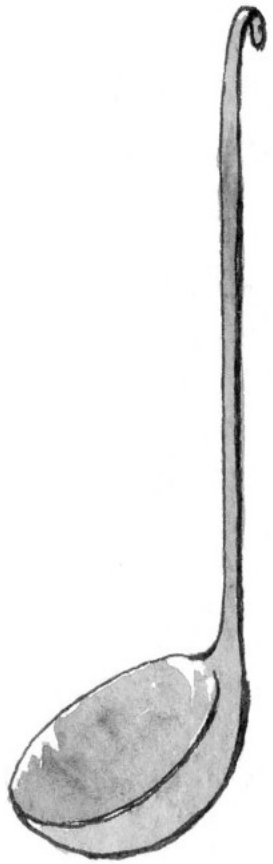
- прескриптивная

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 распознования речи и языка
  - исследования в области NLP, пробинг языковых моделей и т. п.

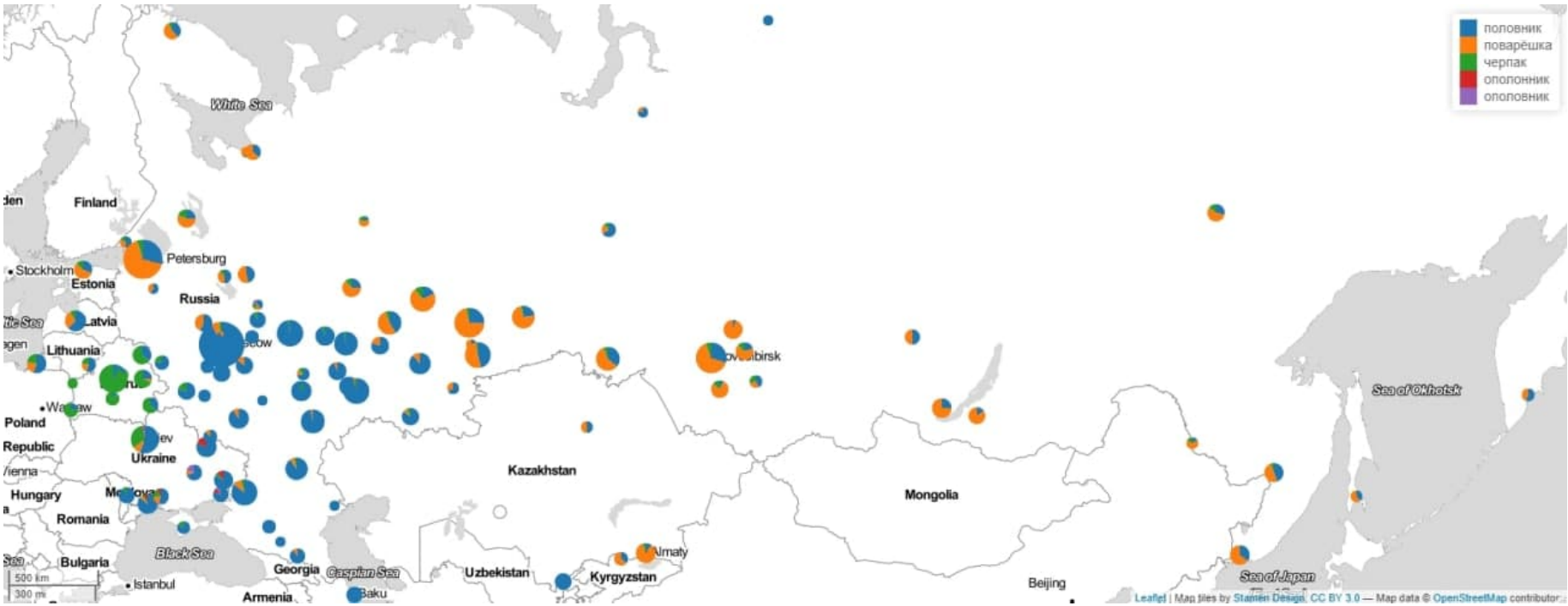
Прескриптивная vs. дескриптивная лингвистика

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



Прескриптивная vs. дескриптивная лингвистика

Это часть опроса И. Левина:



Прескриптивная vs. дескриптивная лингвистика

Запишите где-нибудь, как бы вы заполнили пробелы в предложении:

Я позвал одну мо_ подругу на мо_ день рождения.

Обо мне

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`
- руковожу Международной лабораторией языковой конвергенции

Вариативность в андийском языке

Вдохновение

- “Two equally interesting questions are at the heart of this book: how an extraordinary degree of idiosyncratic linguistic variation can coexist with an extraordinarily homogeneous speaker population, and how linguists might overlook the possibility of their coexistence.” [Dorian, 2010, 3]
- Самира Ферхеес

Данные

Данные были собраны у:

- 44 носителей андийского языка (нахско-дагестанская семья) во время полевого исследования (Ботлихский район, Дагестан) в 2019 году

Данные

Данные были собраны у:

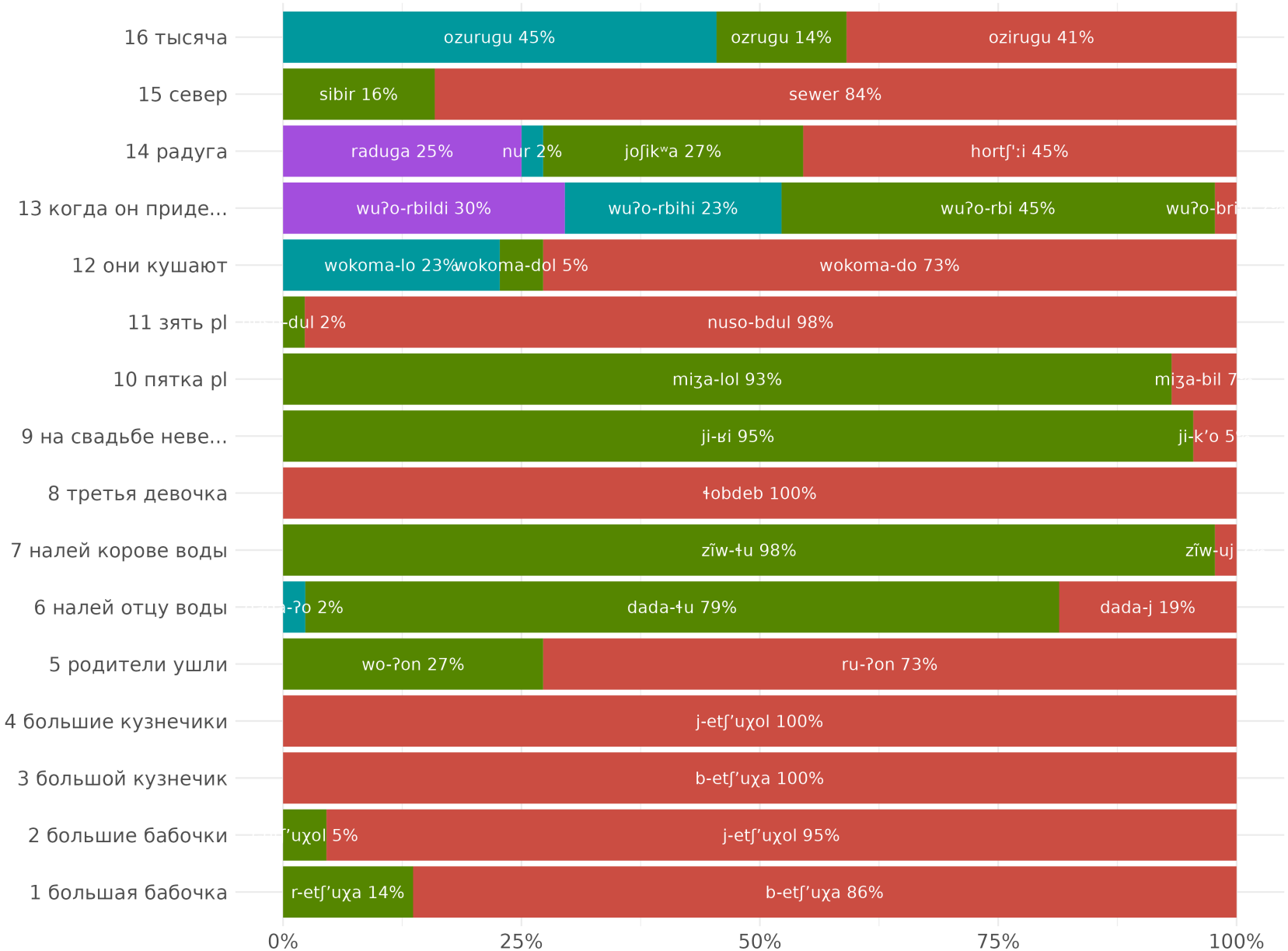
- 44 носителей андийского языка (нахско-дагестанская семья) во время полевого исследования (Ботлихский район, Дагестан) в 2019 году

- и 23 исследователей нахско-дагестанских языков при помощи

44 носителей перевели следующие предложения:

(1) ‘большая бабочка’
(2) ‘большие бабочки’
(3) ‘большой кузнечик’
(4) ‘большие кузнечики’
(5) ‘родители ушли’
(6) ‘налей отцу воды’
(7) ‘налей своей корове воды’
(8) ‘третья девочка’
(9) ‘на свадьбе невеста была красивая’
(10) ‘пятки’
(11) ‘зятья’
(12) ‘они едят’
(13) ‘когда он придет, мы будем есть’
(14) ‘радуга’
(15) ‘север’
(16) ‘тысяча’

Зиловский опрос (44 носителей)



Информационная энтропия

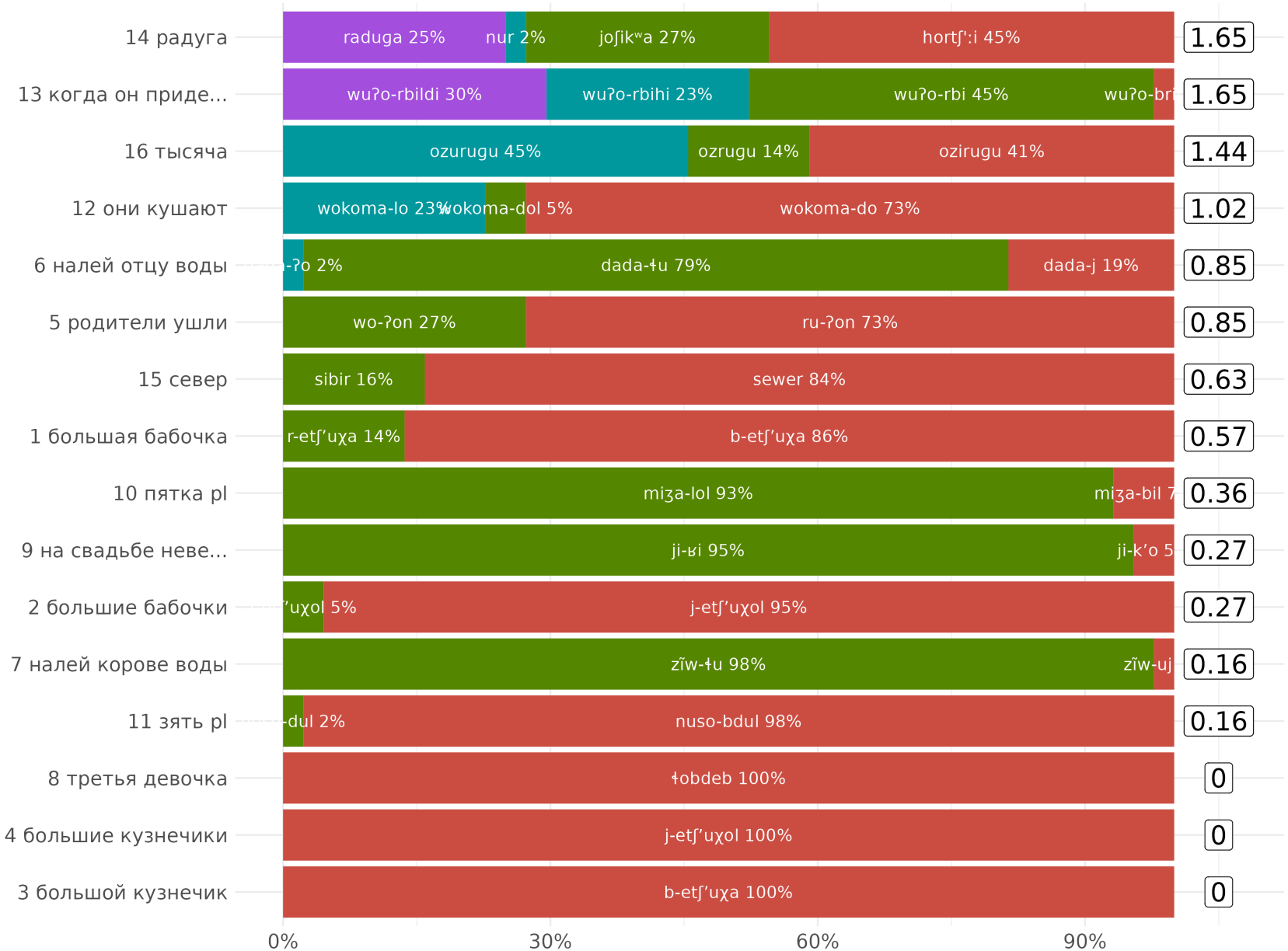
Чтобы измерить вариативность каждого вопроса, мы решили использовать информационную энтропию [Shannon, 1948]:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \times \log_2 P(x_i)$$

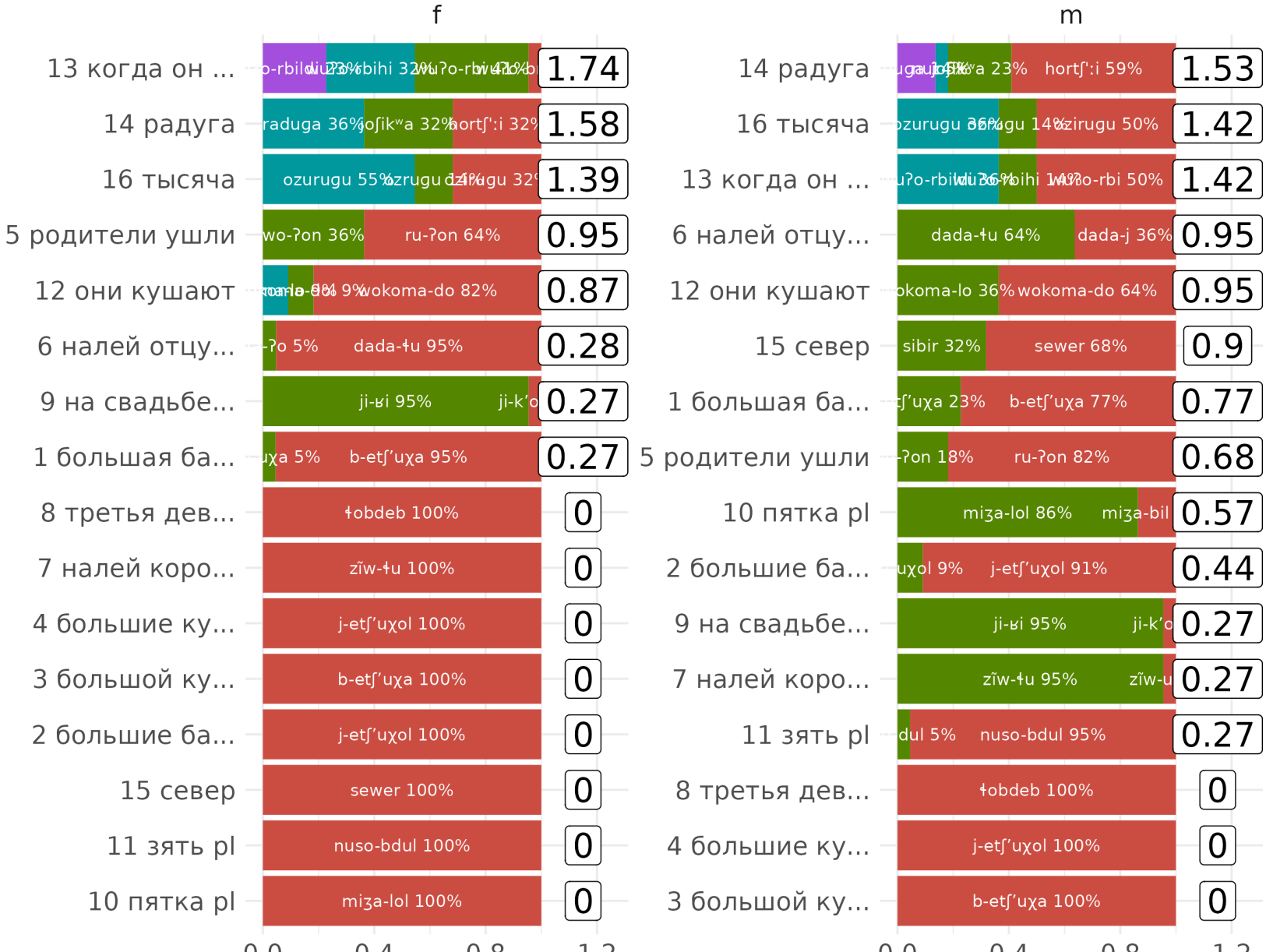
Область значения энтропии $H(X) \in [0, +\infty]$:

данные	энтропия
A-A-A-A-A	0.00
A-A-A-A-B	0.72
A-A-A-B-B	0.97
A-A-B-B-B	0.97
A-A-B-B-C	1.52
A-B-C-A-B	1.52

Зиловский опрос (44 носителей): энтропия справа



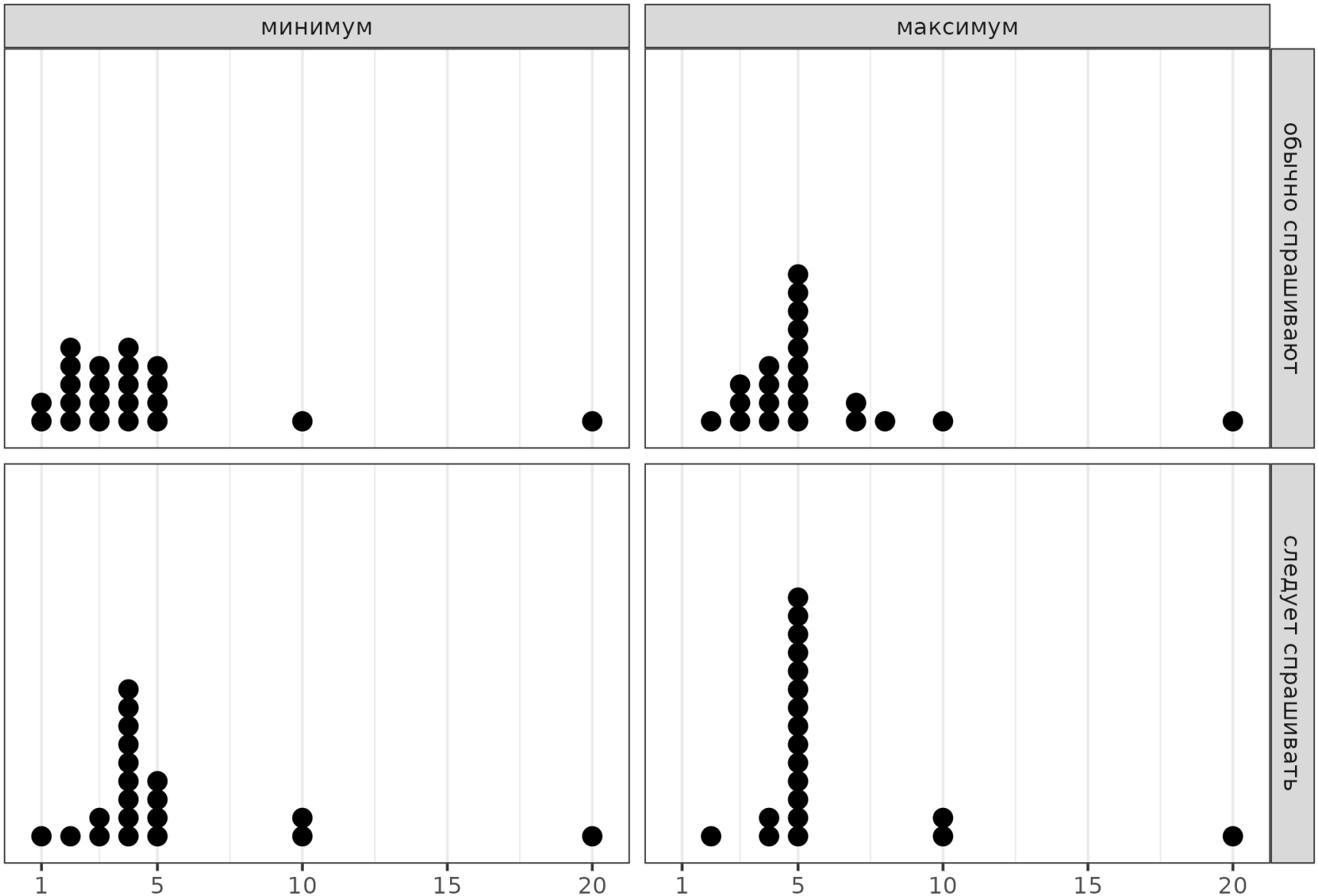
Зиловский опрос (44 носителей): гендер



23 нахско-дагестанских исследователей заполнили следующую анкету:

- образование
- лингвистические интересы
- изучалась ли лингвистика в университете
- участие в полевой работе в качестве студента
- год получения степени
- место учебы/работы
- предпочтительное количество людей в полевой работе
- цели полевой работы
- **количество носителей, которые, согласно мнению исследователя, *следует* опрашивать**
- **количество носителей, которые исследователь *обычно* опрашивает**
- ...

Количество носителей

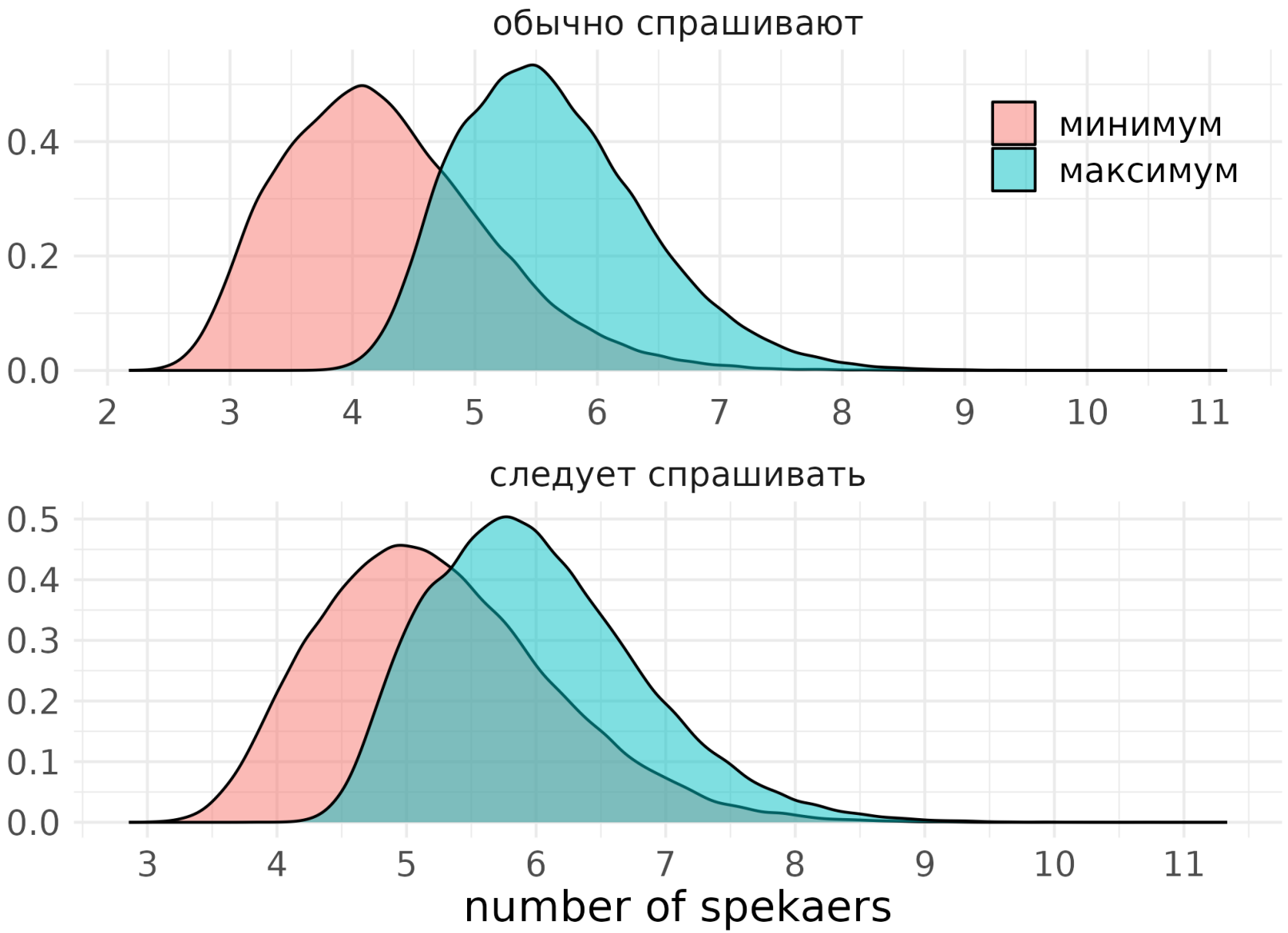


Бутстрэп

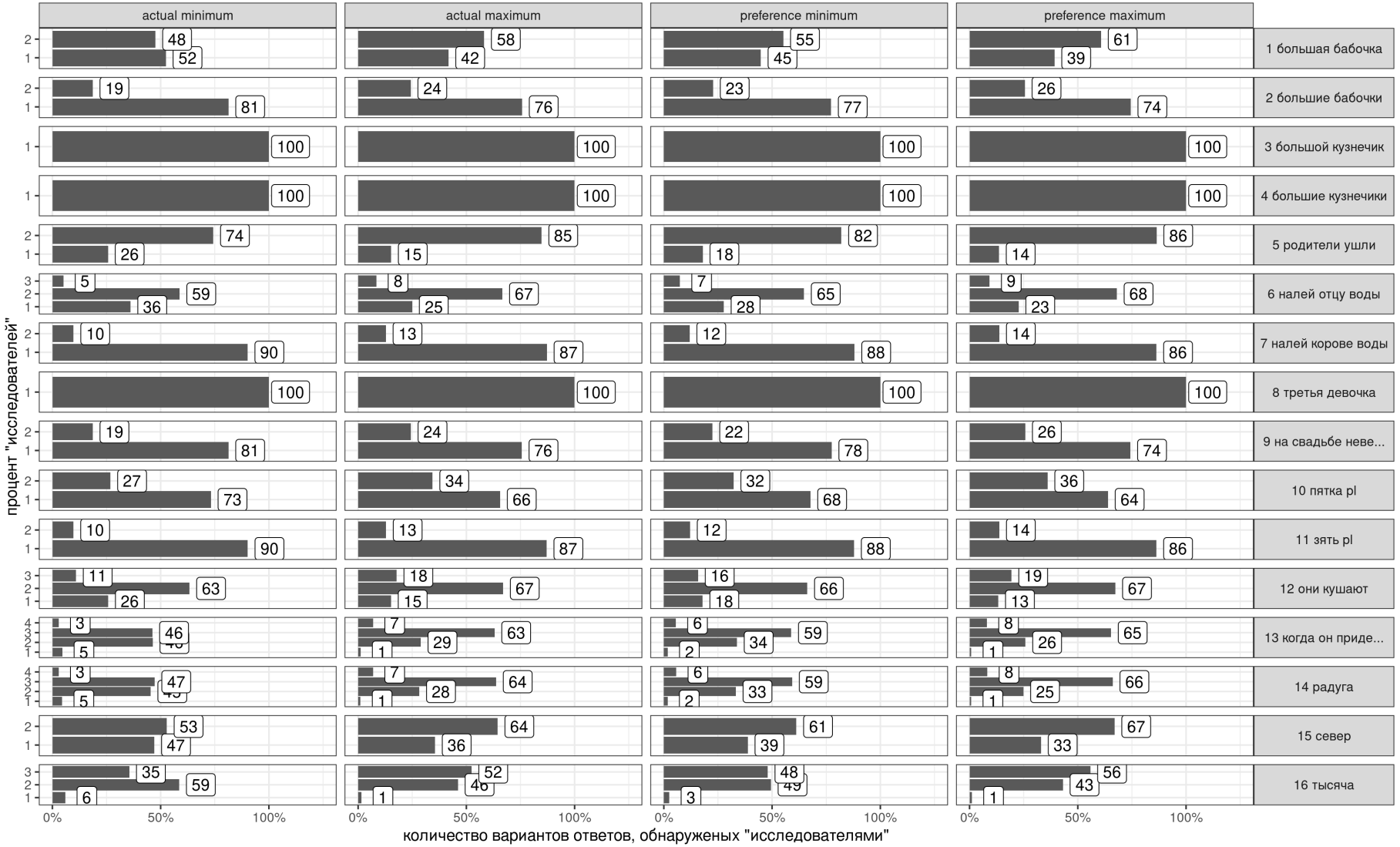
“To pull oneself over a fence by one’s bootstraps”.

Бутстрэп — это такой статистический подход, в рамках которого некоторый статистический параметр оценивается на основе большого количества выборок из имеющихся данных с повторением (т. е. каждое наблюдение может встретиться в выборке 0 раз, 1 раз, 2 раза и т. д.). В результате, вместо одной оценки параметра получается столько оценок, сколько у нас выборок, а все эти оценки формируют распределение.

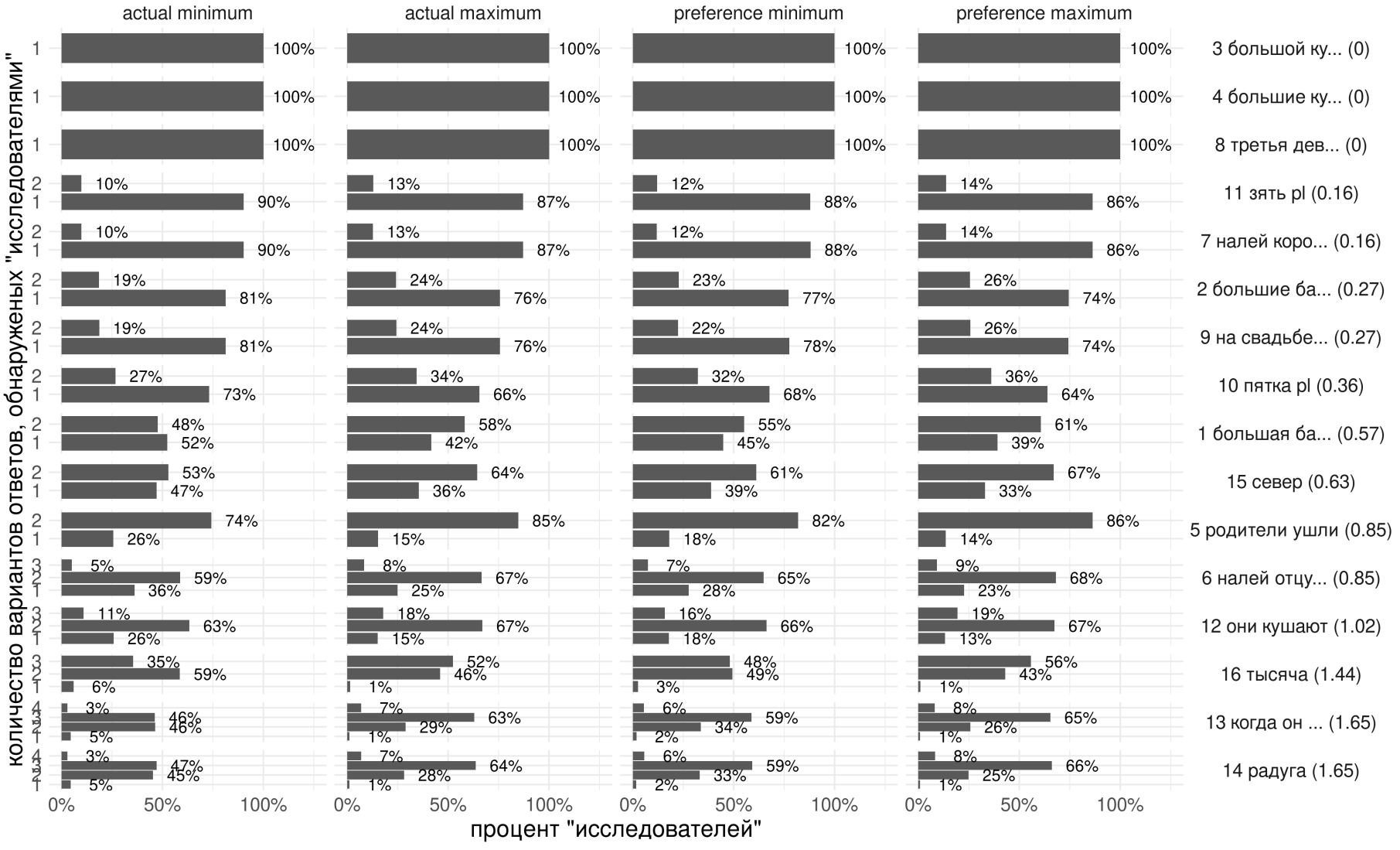
Бустрэп количества опрашиваемых носителей (10^5 iter.)



Что если 10^5 исследователей наведаются в Зило?



Что если 10^5 исследователей наведаются в Зило?



Заключение

- вариативность можно описывать при помощи энтропии
- “среднестатистического” исследователя — осмысленная единица метаанализа, которую следует дальше исследовать
- естественно: количество обнаруженной любыми исследователями зависит от энтропии вопроса

Исследование билингвов

Группа DiaL2

- работа сделана вместе с К. Наккарато
- другие члены группы: М. Ермолова, С. Земичева, Н. Кошелюк, А. Яковлева

Корпусная лингвистика

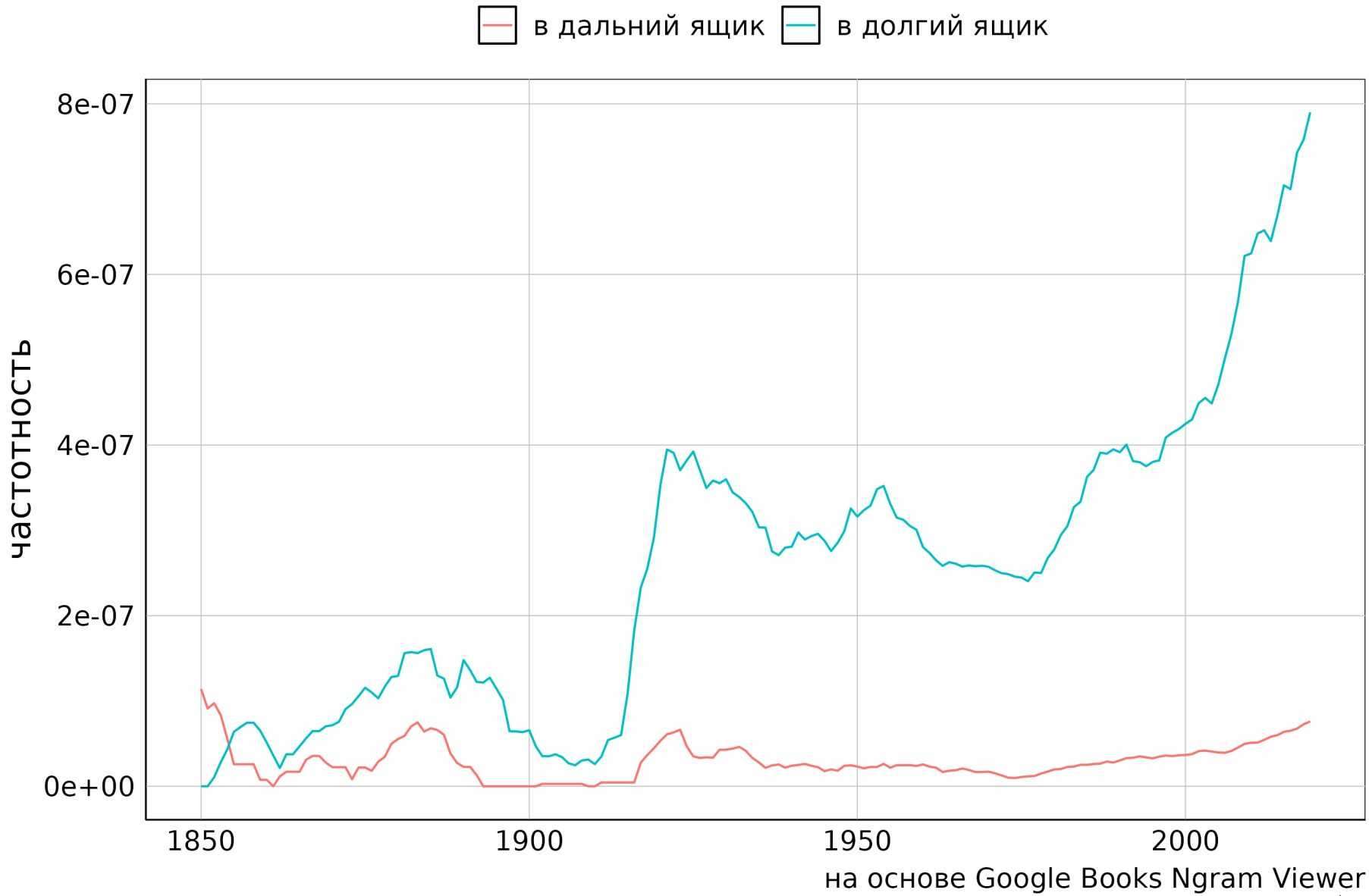
Корпусная лингвистика — это область лингвистики, которая занимается исследованием языковых явлений на материале некоторых собраний языкового материала. В большинтсве случаев это письменные тексты, однако это может быть аудио и даже видео корпуса.

Среди корпусов русского языка можно назвать:

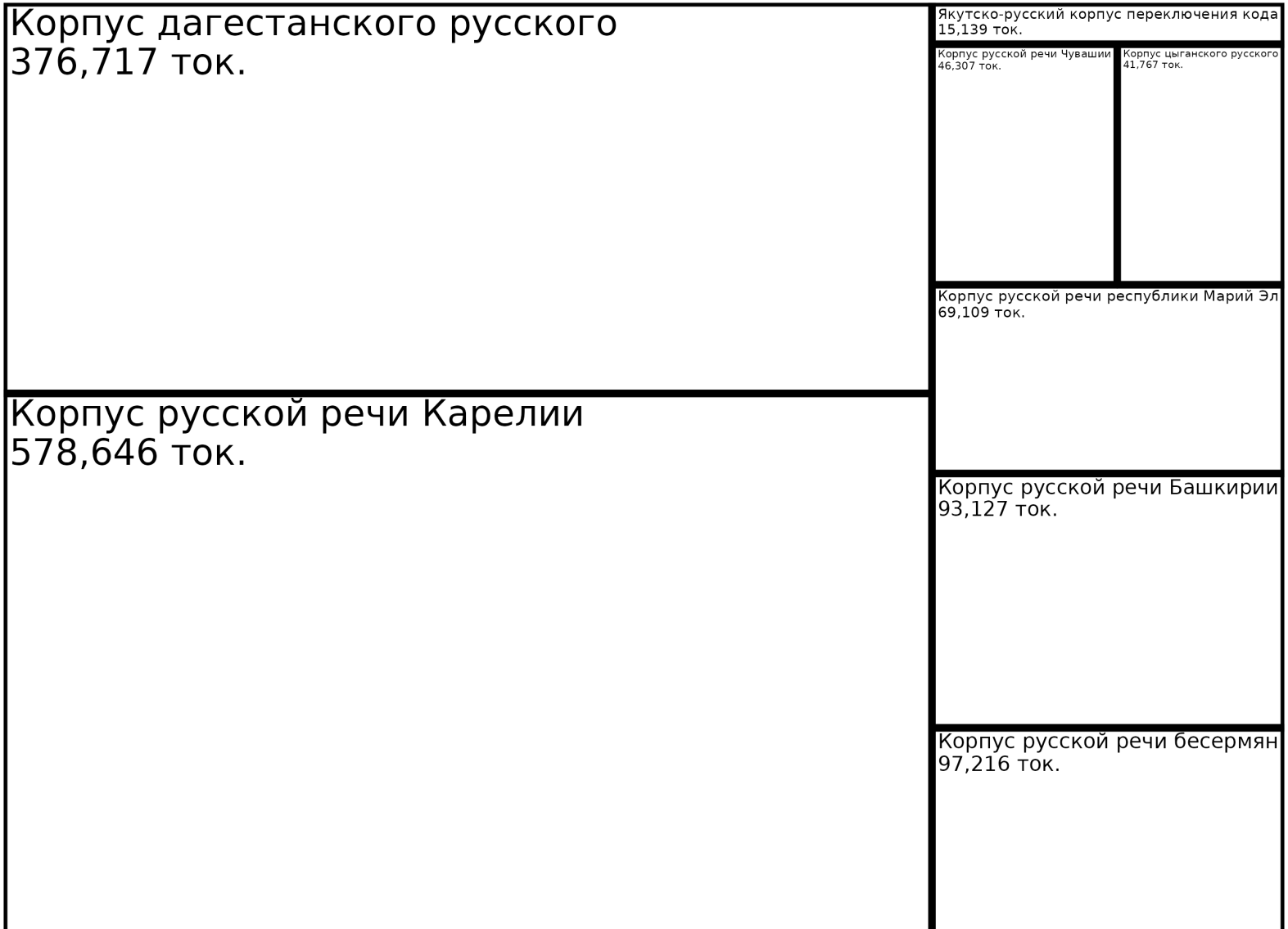
- Национальный корпус русского языка
  - более 1.5 млрд слов
  - много подкорпусов (газетный, устный, параллельный, диалектный, поэтический, исторические)
- Google Books Ngram Viewer
- ...

Отложить в ... ящик

Отложить в ... ящик



Билингвальные корпуса Международной лаборатории языковой конвергенции



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 [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

(17) *Пешком ходил Верхний Дженгутай **пять километра**.* (дагест.)

(18) *Этот меньше, после **двое аборт** делала одну.* (марийский)

- Примеры размечены по некотрым 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

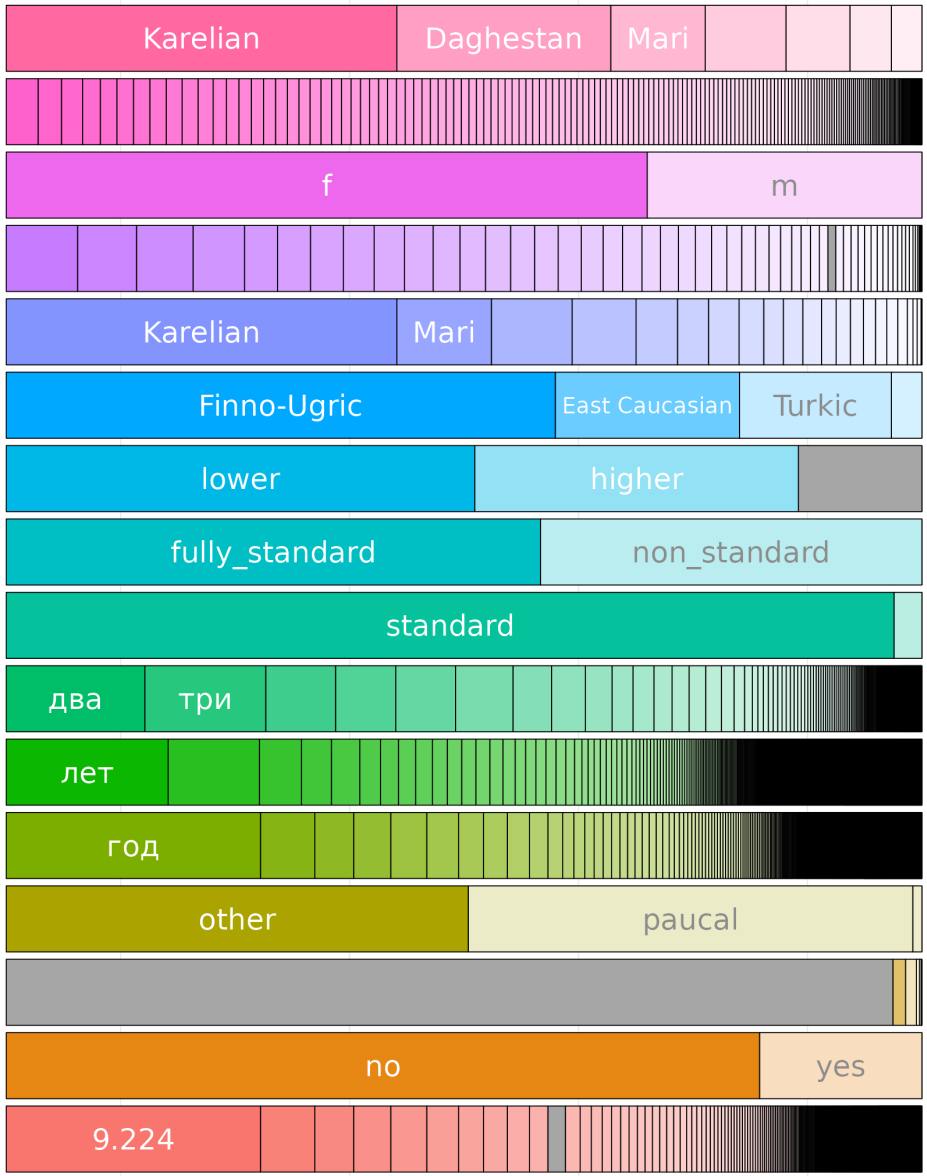
(2.4) noun lemma

(2.5) numeral type

(2.6) noun type

(2.7) ambiguous

(2.8) dice coefficient



Список литературы I

Nancy C Dorian. *Investigating variation: The effects of social organization and social setting*. Oxford University Press, 2010.

C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

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