Корпусная лингвистика

для курса ‘Качественные исследования религиозных сообществ’

Г. А. Мороз

21.03.2024

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

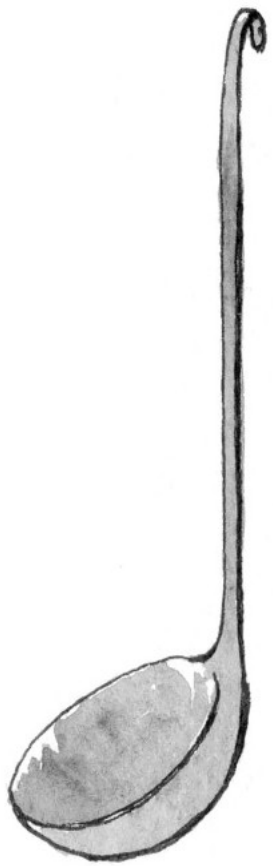
- прескриптивная

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.

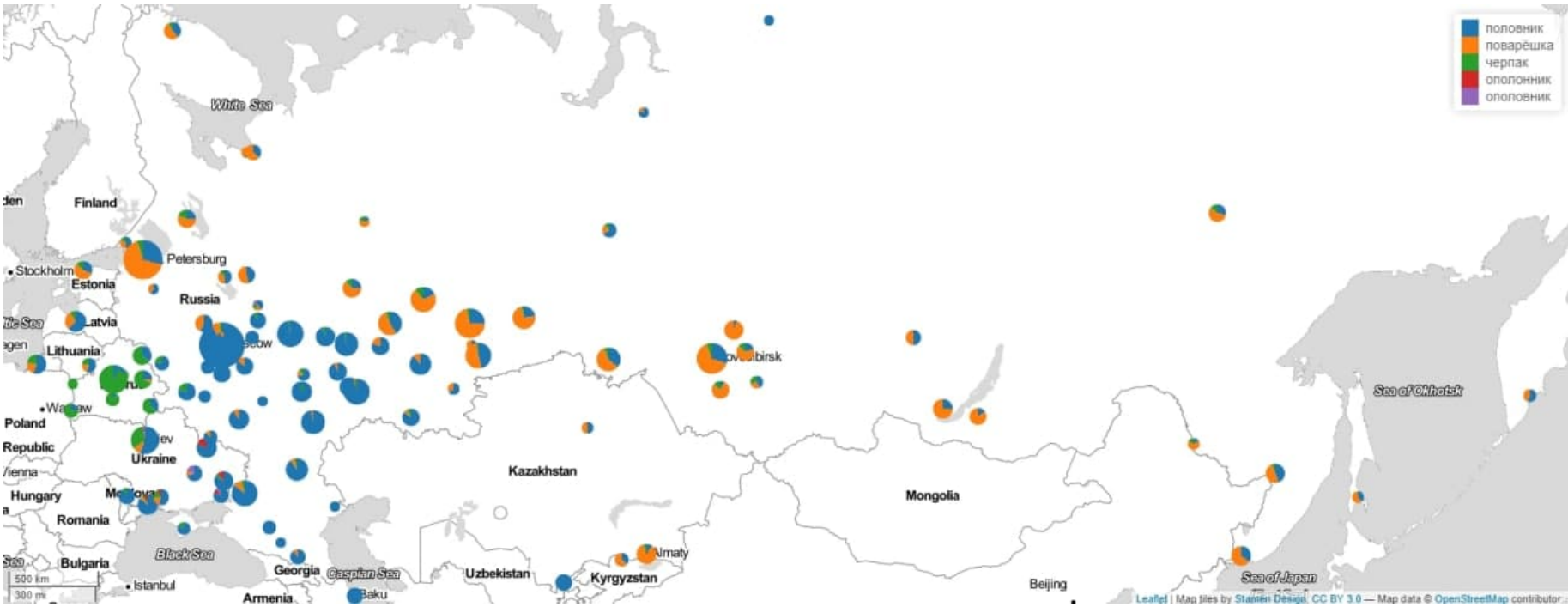
Прескриптивная vs. дескриптивная лингвистика

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



Прескриптивная vs. дескриптивная лингвистика

Это часть опроса И. Левина:



Корпусная лингвистика

Корпусная лингвистика

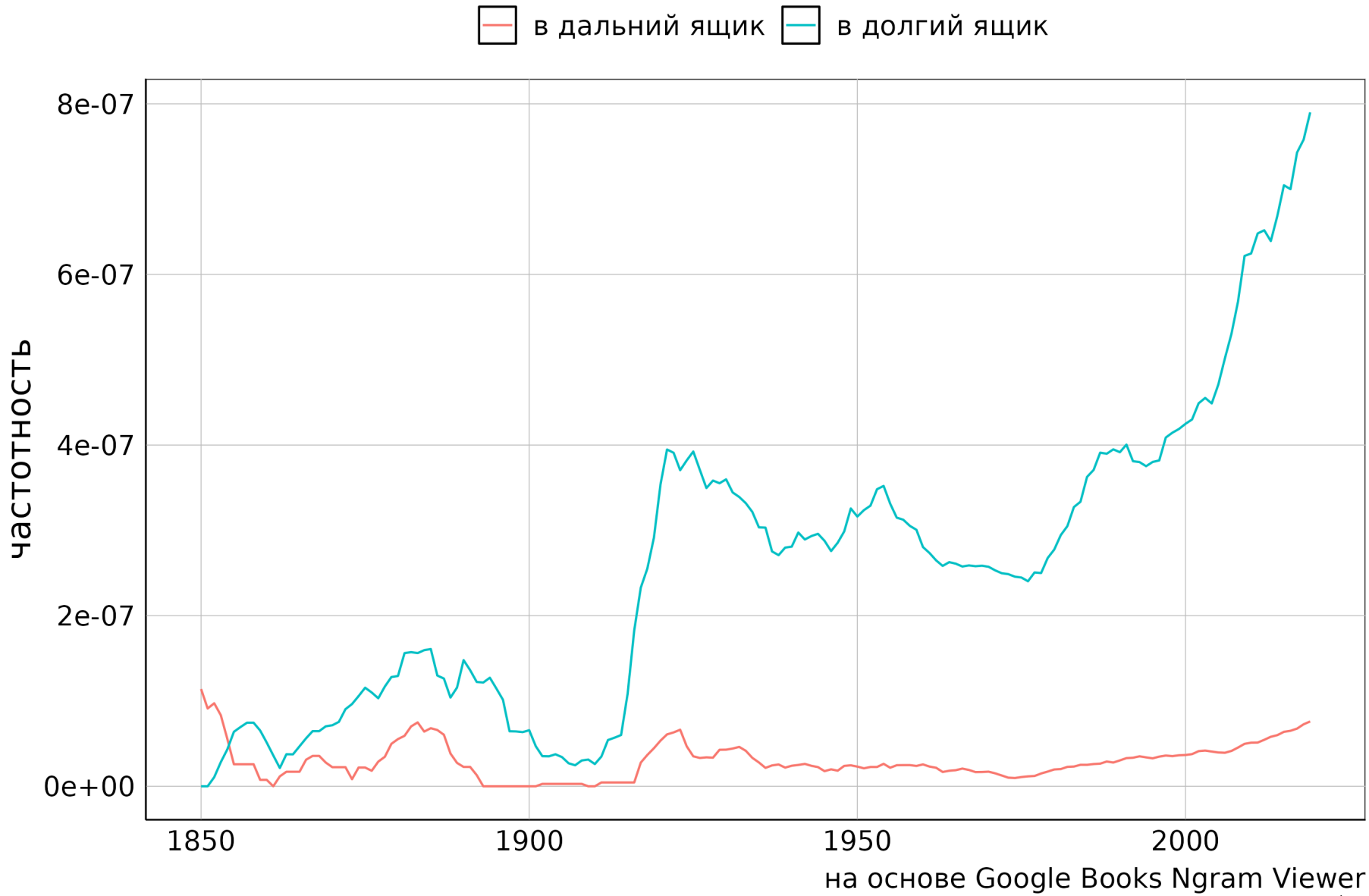
Корпусная лингвистика — это область лингвистики, которая занимается исследованием языковых явлений на материале некоторых собраний языкового материала. В большинстве случаев это письменные тексты, однако это может быть аудио и даже видео корпуса.

Среди корпусов русского языка можно назвать:

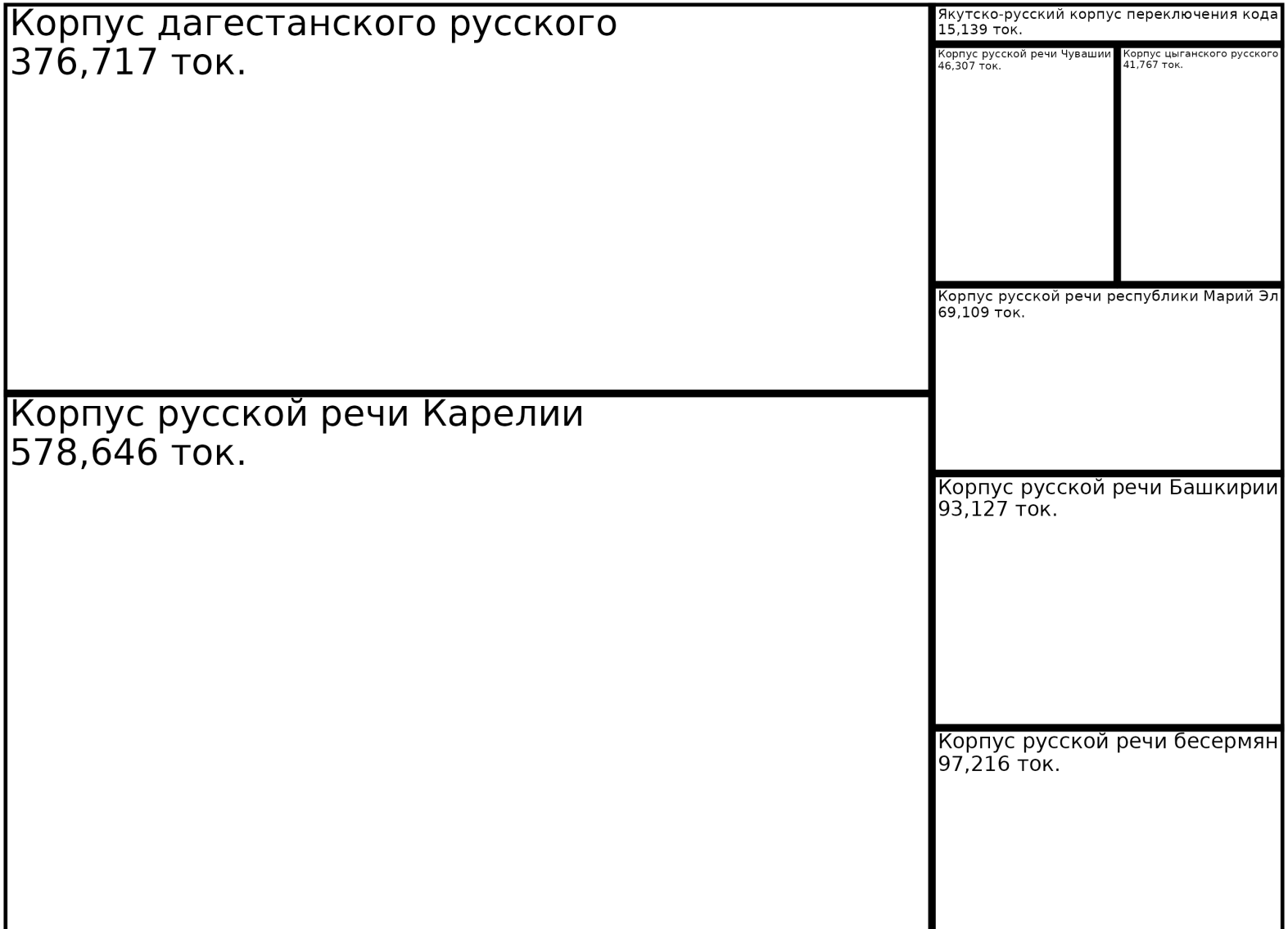
- Национальный корпус русского языка
  - более 1.5 млрд слов
  - много подкорпусов (газетный, устный, параллельный, диалектный, поэтический, исторические)
- Google Books Ngram Viewer
- ...

Отложить в ... ящик

Отложить в ... ящик



Билингвальные корпуса Международной лаборатории языковой конвергенции



Исследования

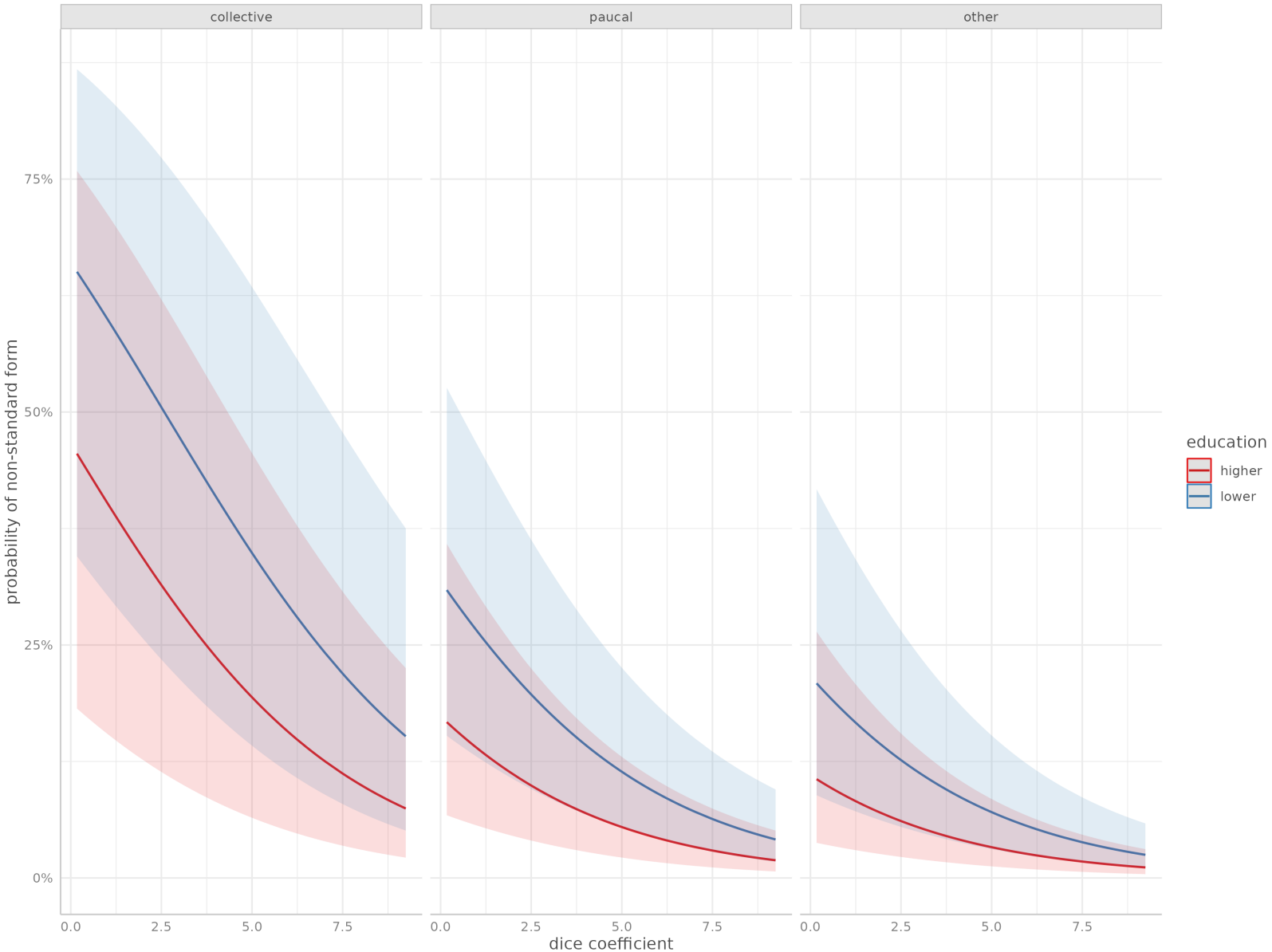
- исследование числительных [Naccarato, Moroz forthcoming]
  - *Пешком ходил Верхний Дженгутай **пять километра.*** (дагестанский русский)
  - *Этот меньше, после **двое аборт** делала одну.* (марийский русский)
- выпадение предлогов [Yakovleva, Koshelyuk, Moroz in preparation]
  - *Со второго курса что ли практика началась, _ больнице.* (марийский русский)
  - *Вот, отремонтировал _ трудом пополам, китайские часы -то.* (бесермянский русский)
  - *я пошёл, _ начальнику дал предложение.* (бесермянский русский)

Моделирование

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

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

Предсказания модели



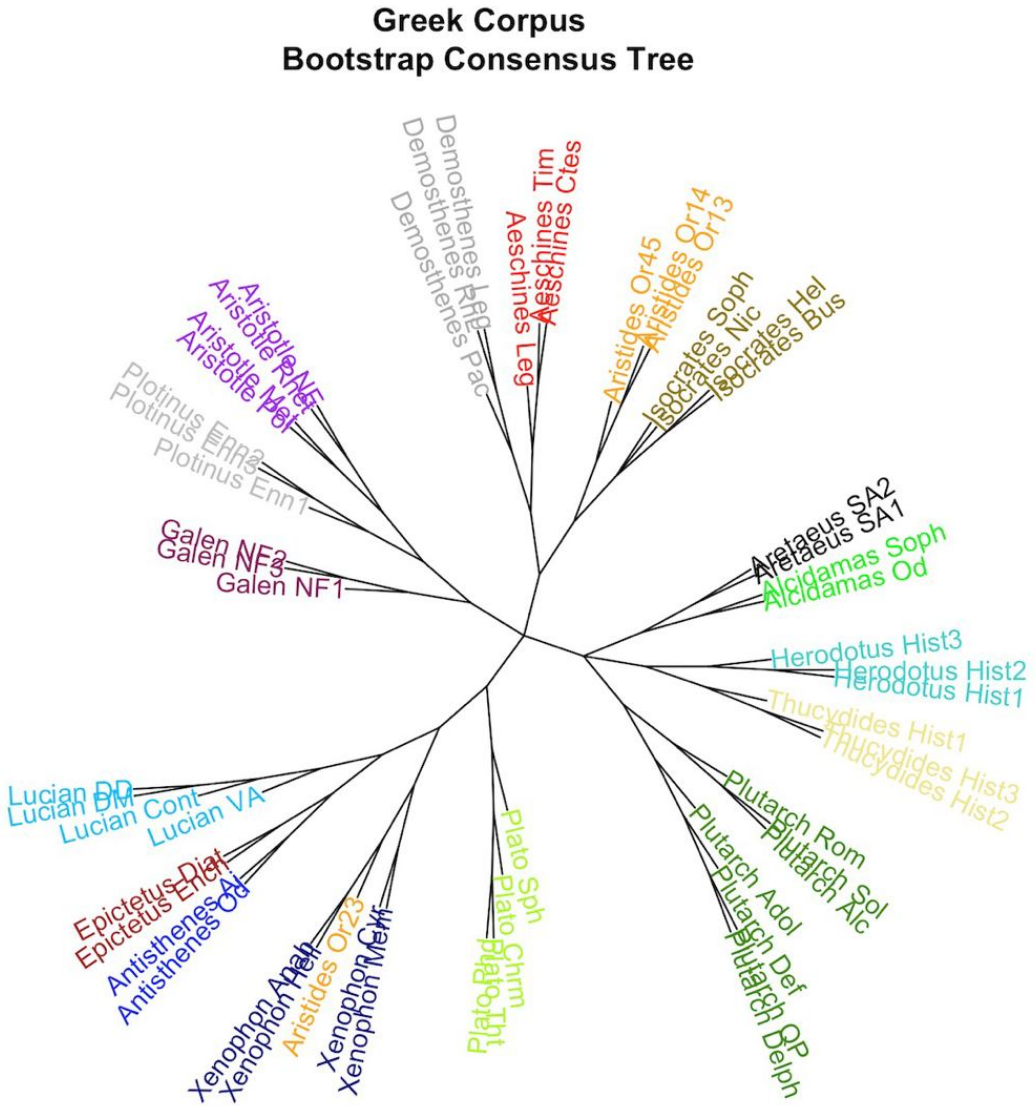
Digital Humanities

Digital Humanities

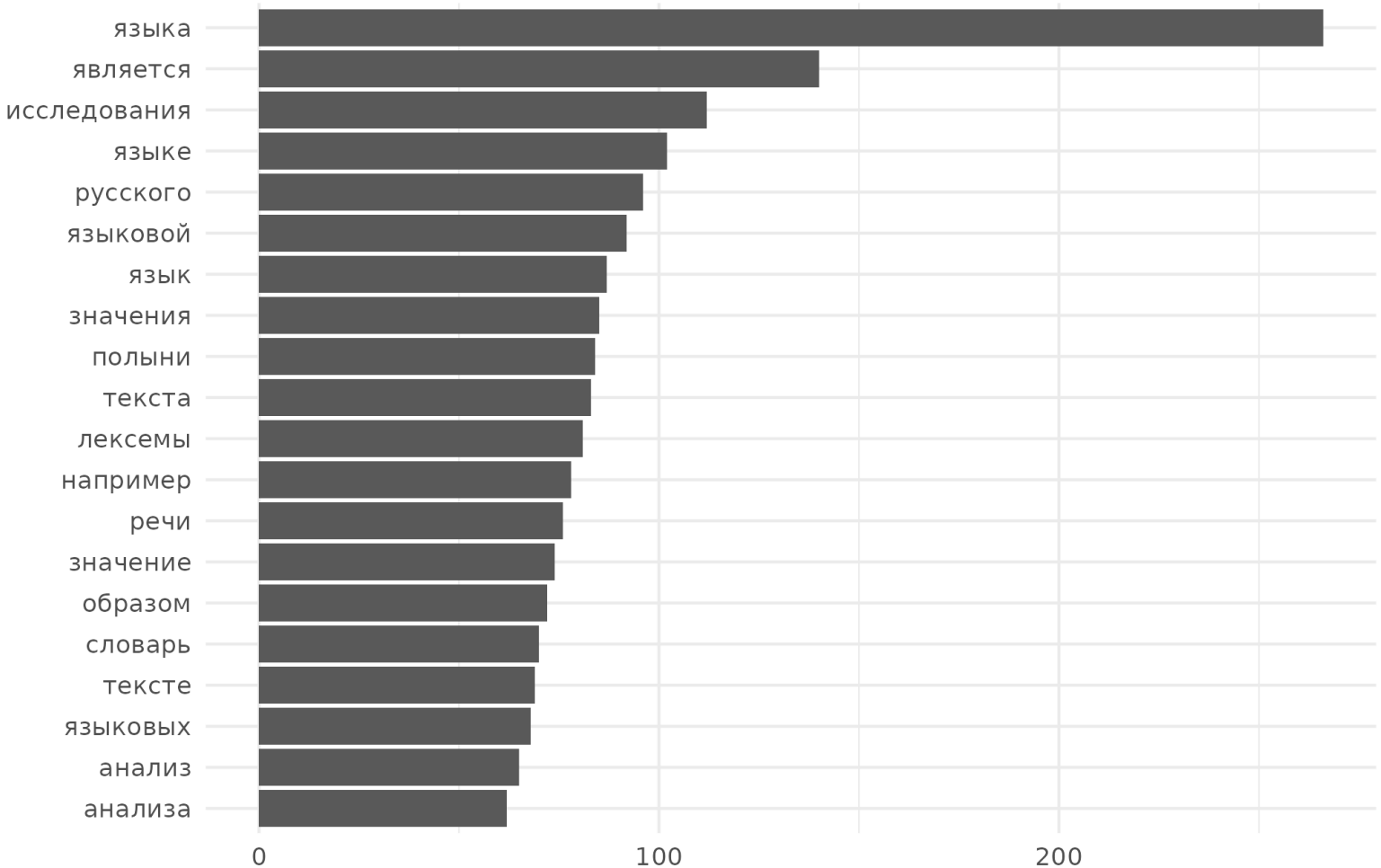
Это направление цифровых гуманитарных исследований, которое сложно хорошо очертить: они скорее объединены методом. Мы поговорим про два аспекта, которые относят к DH:

- определение авторства
- дальнее чтение (distant reading)

Греческий корпус из тг канала О. В. Алиевой

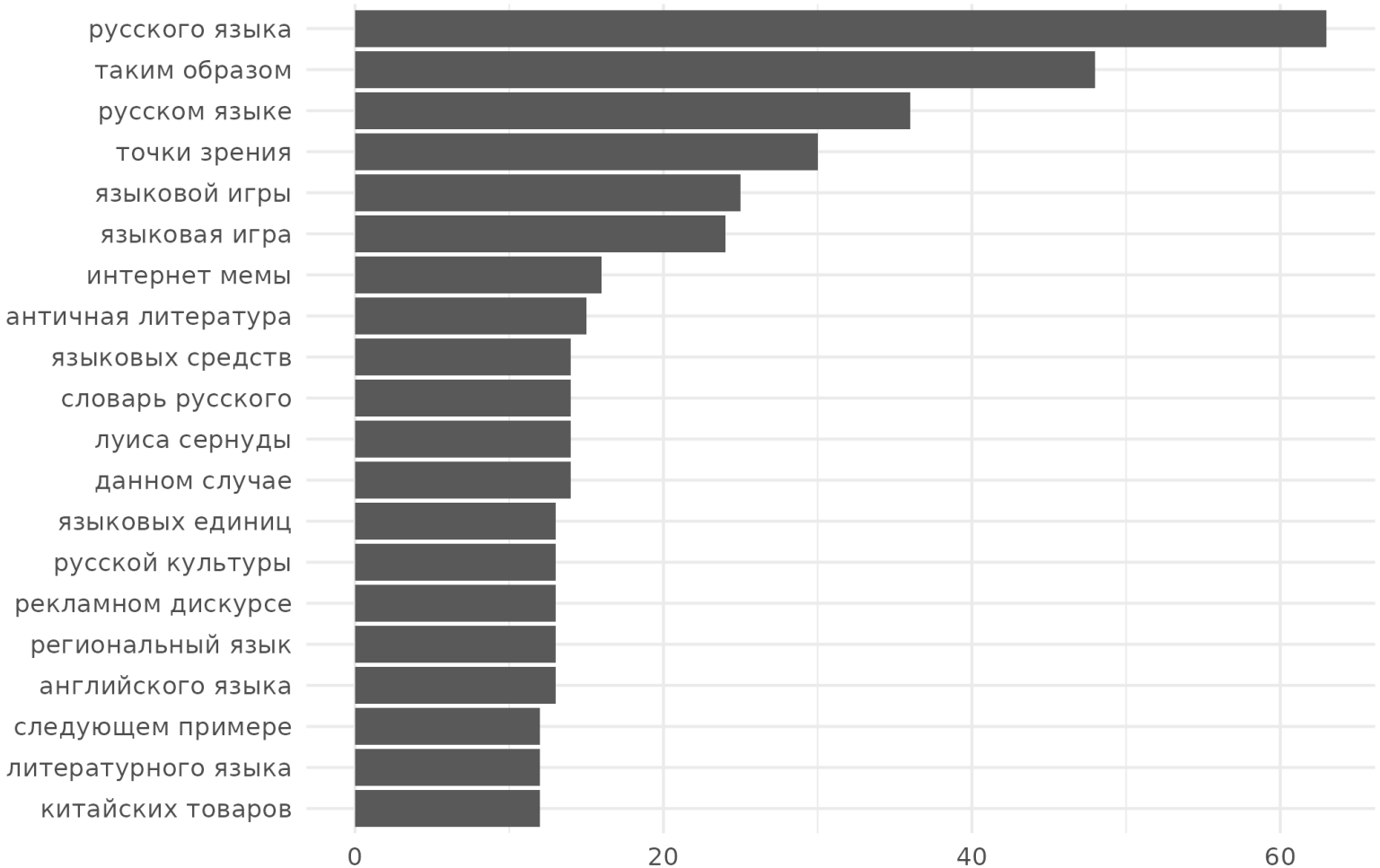


30 статей из Известий ЮФУ. Филологические науки (2022-2023)



пришлось добавить в стоп-слова: дата, обращения, канд, филол, наук, дис, канд, южного, федерального, университета, гос, ун, список, источников, научная, статья, известия, юфу, др, филологические, науки, ключевые, слова, электронный, ресурс

30 статей из Известий ЮФУ. Филологические науки (2022-2023)



пришлось добавить в стоп-слова: дата, обращения, канд, филол, наук, дис, канд, южного, федерального, университета, гос, ун, список, источников, научная, статья, известия, юфу, др, филологические, науки, ключевые, слова, электронный, ресурс