

# Встреча о сотрудничестве Факультета Математики и Школы Лингвистики

Георгий Мороз

21.02.2024

# Моделирование вариативности на материале билингвальных корпусов

## 8 устных билингвальных корпусов

Корпус дагестанского русского  
376,717 ток.

Корпус русской речи Карелии  
578,646 ток.

Якутско-русский корпус переключения кода  
15,139 ток.

Корпус русской речи Чувашии  
46,307 ток.

Корпус шаганского русского  
41,767 ток.

Корпус русской речи республики Марий Эл  
69,109 ток.

Корпус русской речи Башкирии  
93,127 ток.

Корпус русской речи бесермян  
97,216 ток.

# 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)

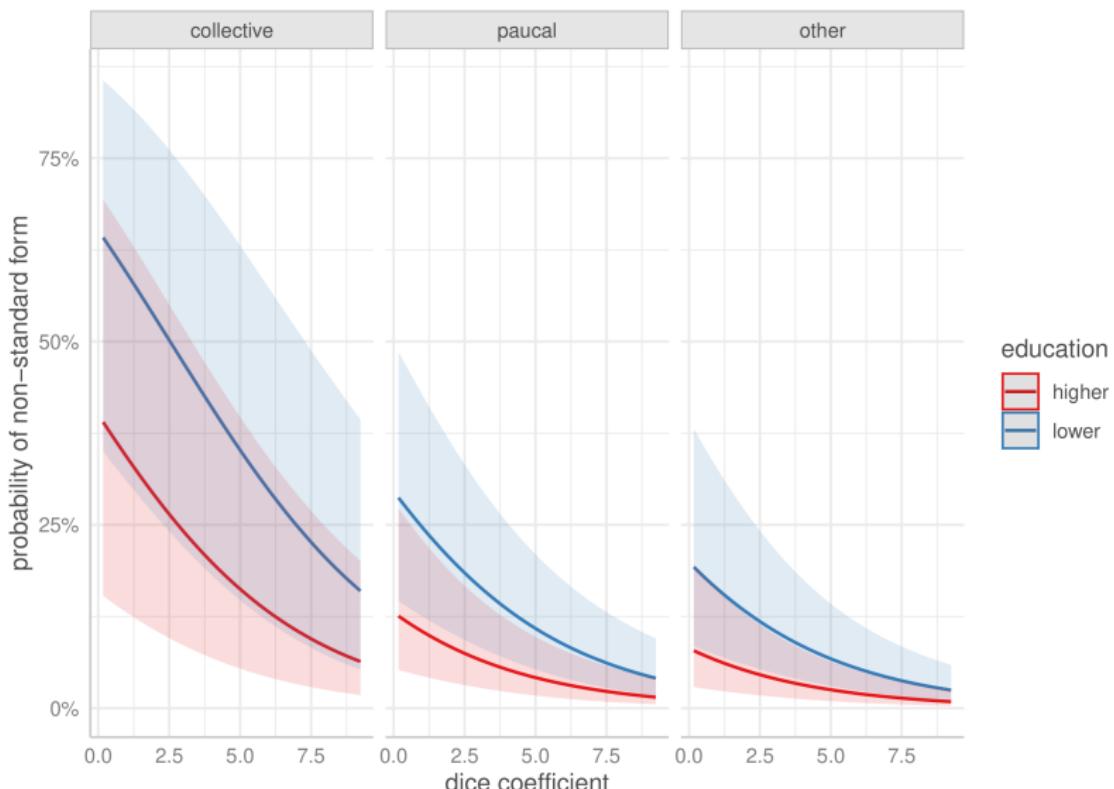
- Примеры размечены по некоторым параметрам
  - лингвистическим
    - коллокационность комбинации числительного + существительного
    - тип числительного (собирательные *двое, троє, паукальные два, три*, другие)
  - социолингвистическим
    - год рождения
    - пол
    - образование
    - первый язык

## 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

## Результаты

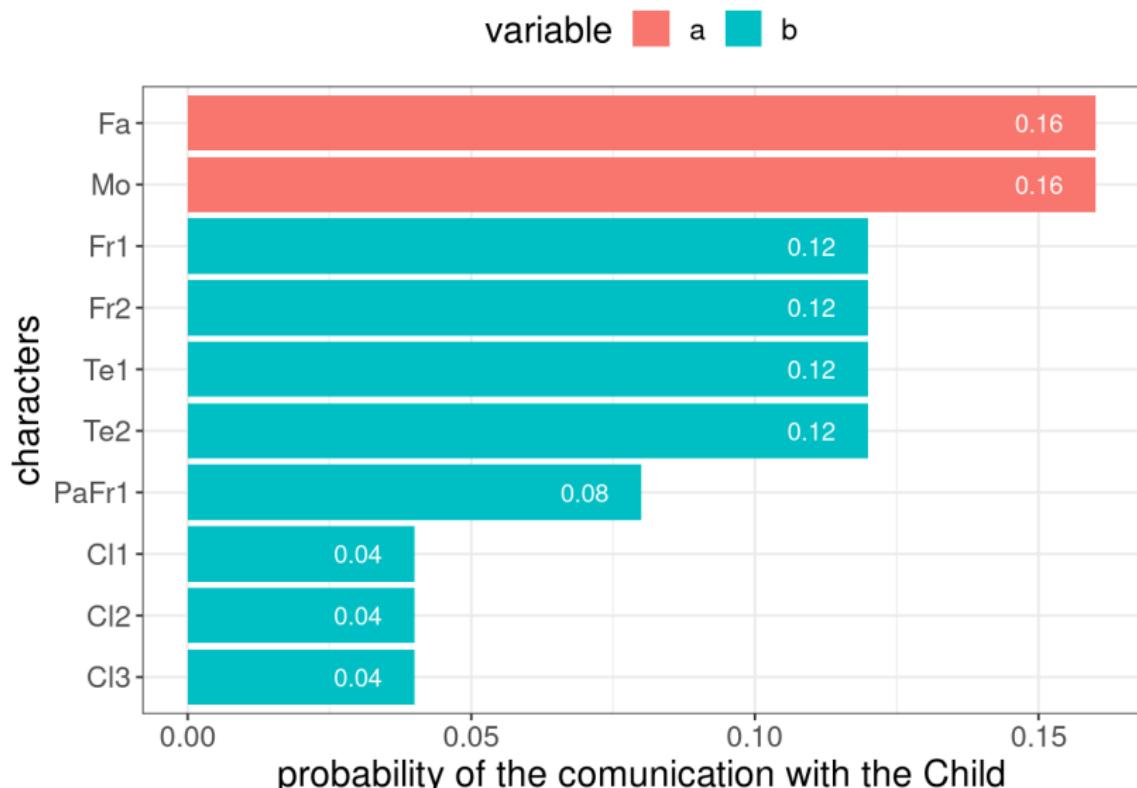


# А зачем нам математики?

## Моделирование изменений вариативности

- ребенок (Ch), переменная  $a$ ;
- двое родителей (Fa, Mo), переменная  $a$ ;
- два друга из школы (Fr1, Fr2), переменная  $b$ ;
- три одноклассника (C11, C12, C13), переменная  $b$ ;
- два преподавателя (Te1, Te2), переменная  $b$ ;
- два знакомых родителей (PaFr1, PaFr2), переменная  $b$ .

## Моделирование изменений вариативности

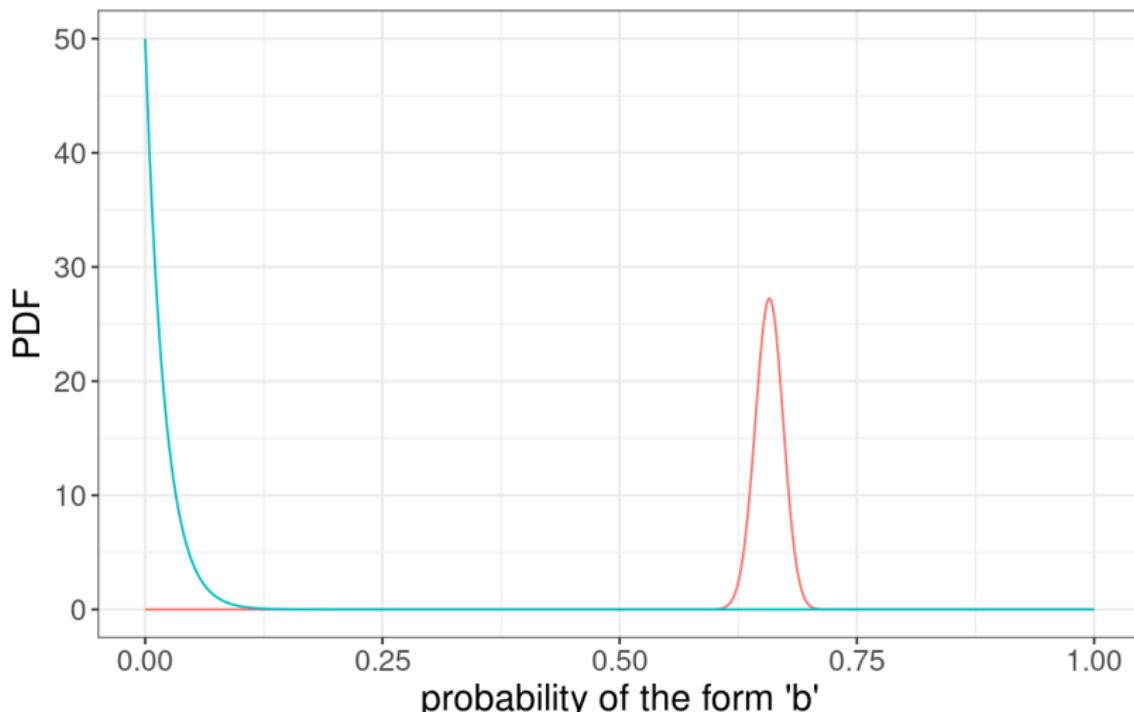


## Моделирование изменений вариативности

- Симулируем 1000 разговоров ребенка
- В каждом разговоре ребенок наблюдает либо форму  $a$ , либо форму  $b$
- Описывать вероятность использование формы  $a$  или  $b$  можно при помощи бета-распределения
- Априорное распределение можно взять, скажем Beta(1, 50)
- В ходе симуляции ребенок наблюдает 310 форм  $a$  и 690 форм  $b$
- Результат изменений можно представить в виде байесовского апдейта бета распределения

# Моделирование изменений вариативности

— after 500 conversations Beta(691, 360) — initial state Beta(1, 50)

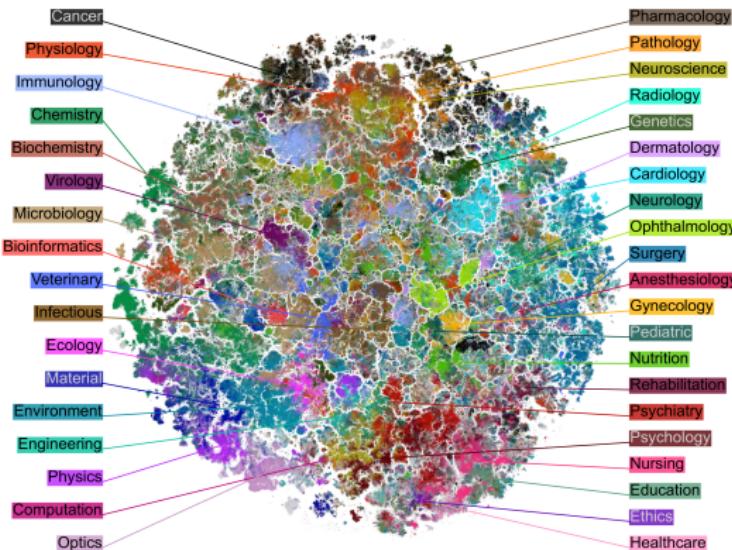


## Моделирование изменений вариативности

- разве человек может упомянуть все формы, что он слышал в жизни? наверное, нет
- можно считать, что у него в голове есть лишь параметры бета распределений, которые он обновляет с каждым новым встреченным употреблением
- а можно ввести функция забывания, чтобы он помнил лишь последние, скажем, 300 наблюдений
- ...

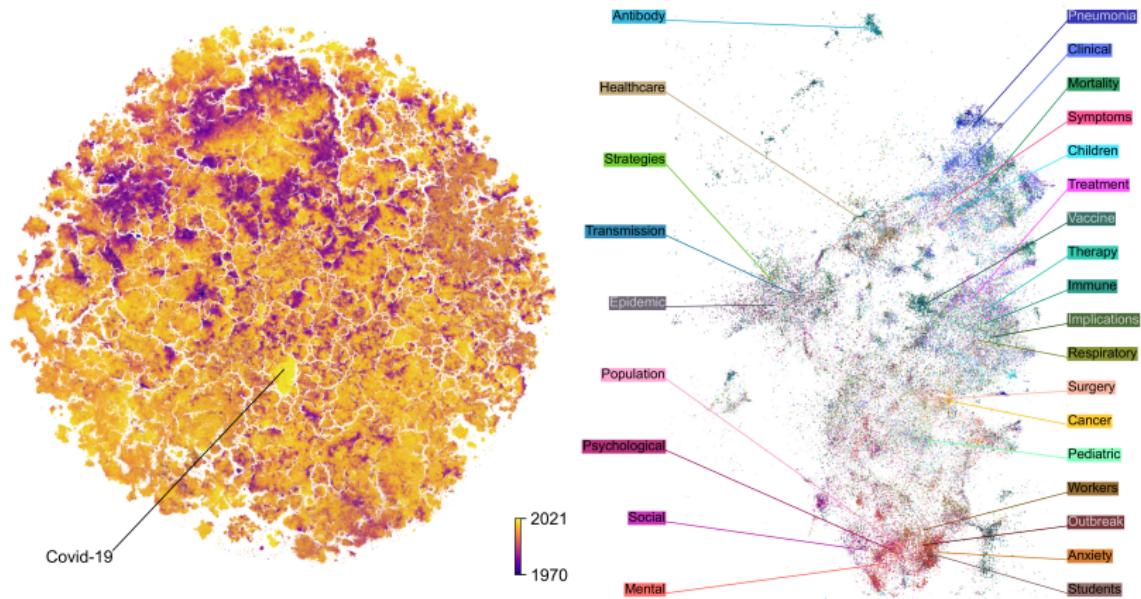
# Ландшафт науки (математика?)

[Gonzalez-Marquez et al., 2023]



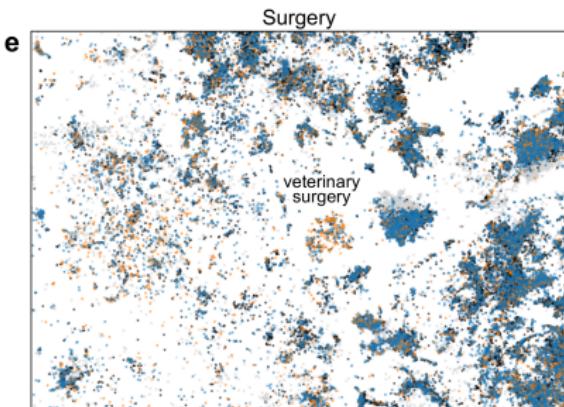
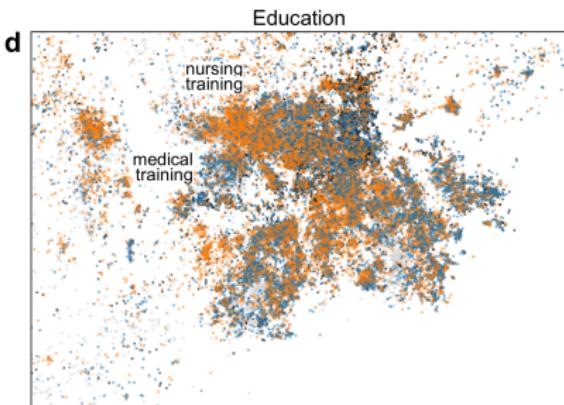
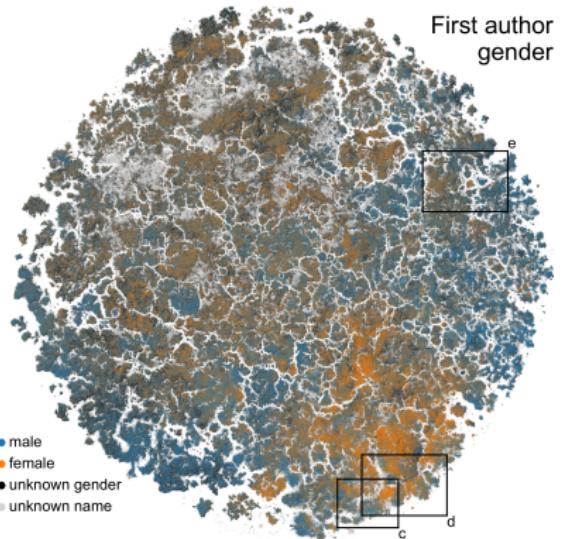
2D эмбеддинги на основе 21 миллиона аннотаций, которые были трансформированы в 768-мерное векторное пространство при помощи PubMedBERT [Gu et al., 2021], а дальше сплюснутая в 2D при помощи t-SNE [Van der Maaten and Hinton, 2008]. Цвета основаны на названиях журналов. [Вот тут интерактивная версия](#).

## [Gonzalez-Marquez et al., 2023]

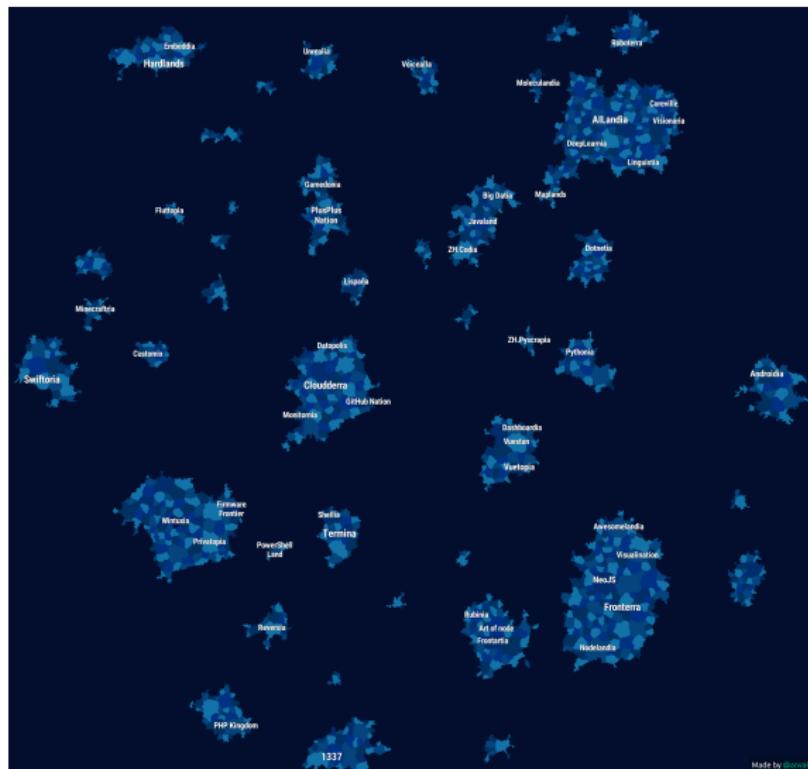


Регион карты, посвященный Covid-19. Цвета приписаны на основе названий работ. Кроме того здесь есть около 15% работ не посвященных короновирусу.

## [Gonzalez-Marquez et al., 2023]



## Карта репозиториев гитхаба (Андрей Кашча)



<https://anvaka.github.io/map-of-github/>

## Ландшафт лингвистики

- выбрать список журналов для анализа
- извлечь аннотации для всех работ из выбранных журналов
- использовать векторизатор и метод уменьшения размерностей для преобразования пространства аннотаций в 2D

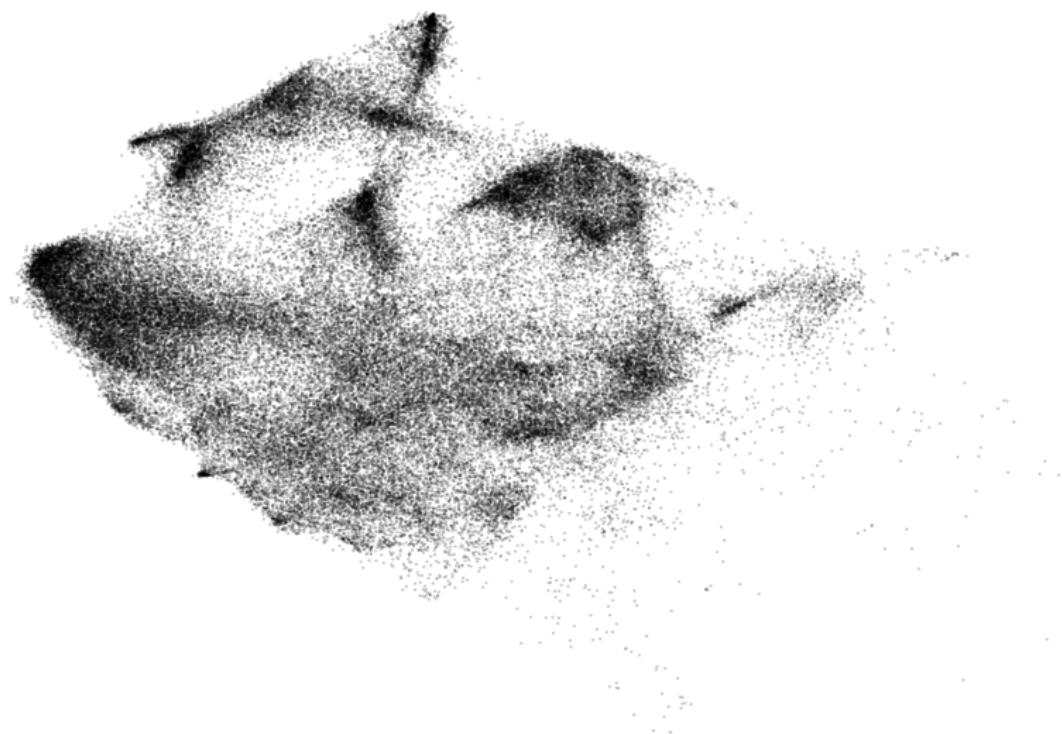
## Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках
- сообщения об отсутствии аннотации
- acknowledgments вместо аннотации
- библиографическое описание книги (в случаях рецензии)
- начало статьи вместо аннотации (характерно для старых статей)
- ошибки распознавания
- слишком короткие/длинные аннотации

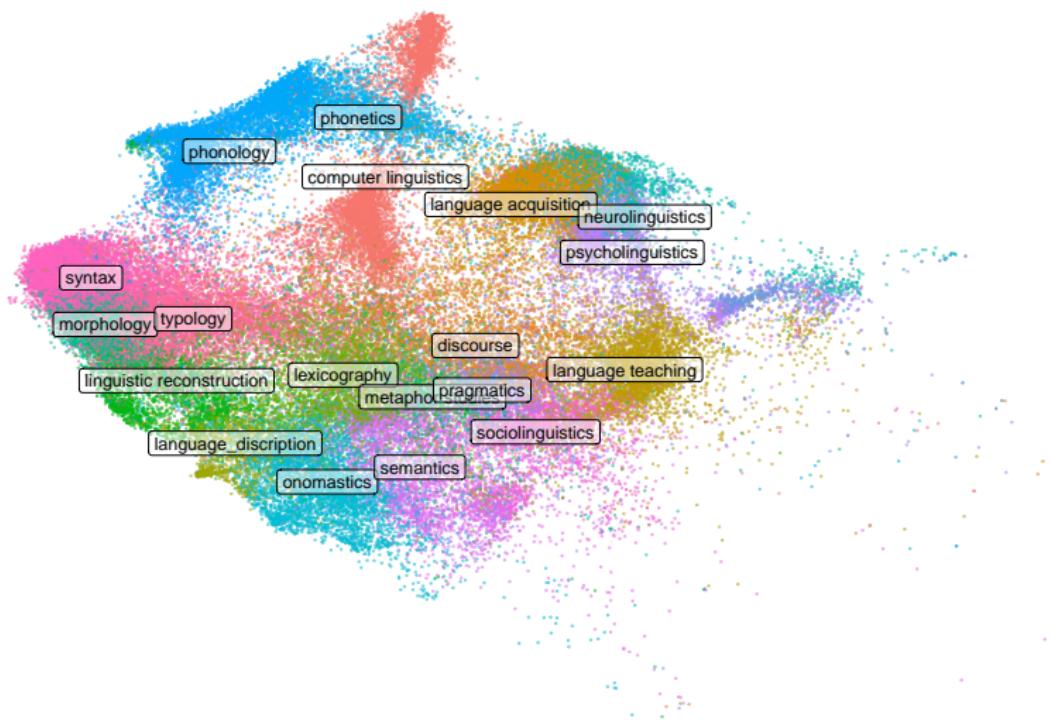
## Структура данных: 78962 строчек, 31 колонок

- **id:** <https://openalex.org/W3040611730>
- **doi:** <https://doi.org/10.1075/fol.18056.dob>
- **author:** Nina Dobrushina
- **title:** Negation in complement clauses of fear-verbs
- **publication\_year:** 2021
- **journal:** Functions of Language
- **issn\_l:** 0929-998X
- **first\_page:** 121
- **last\_page:** 152
- **volume:** 28
- **issue:** 2
- **is\_retracted:** FALSE
- **cited\_by\_count:** 1
- **abstract:** Complement clauses of verbs of fear often contain expletive negation, which is negative marking without negative meaning. <...>
- **concepts:** Negation; Complement (music); Linguistics; Verb; Meaning (existential); Psychology; Mathematics;

# Ландшафт лингвистических исследований



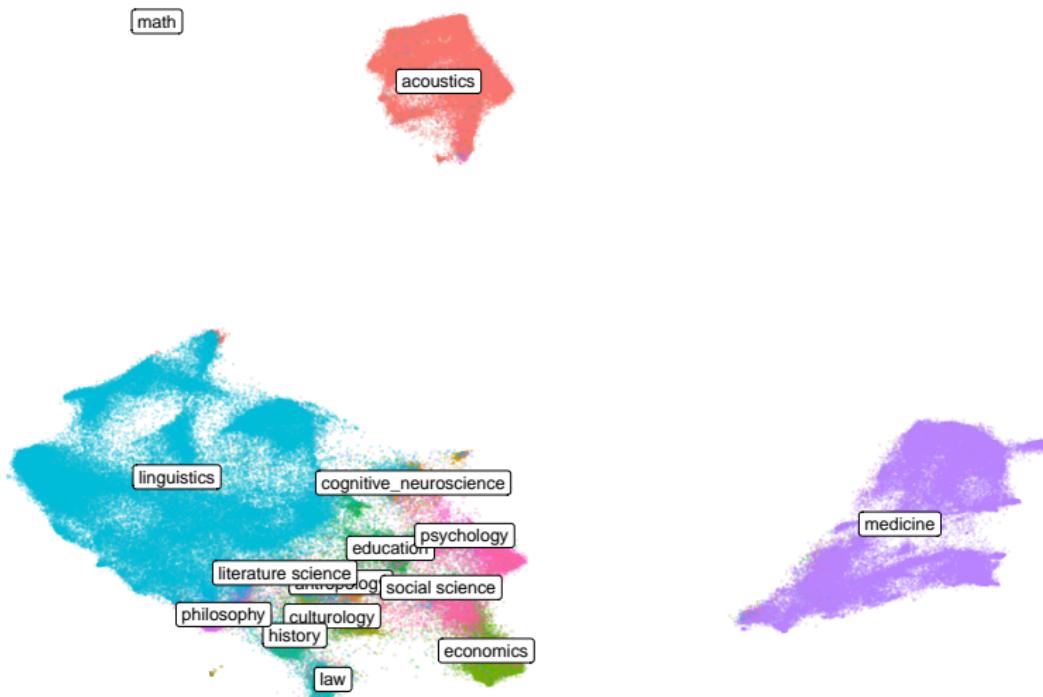
# Наша полуавтоматическая разметка аннотаций



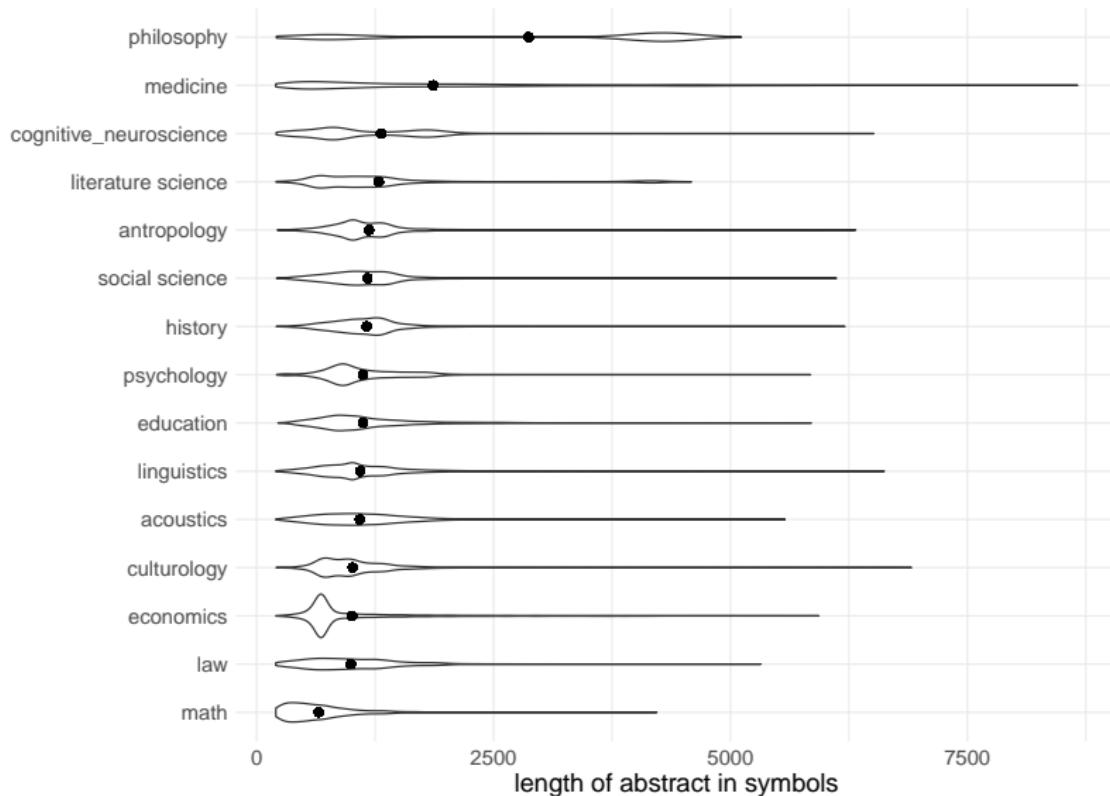
# А что если добавить других дисциплин?

field	journal	n
social science	Administrative Science Quarterly	1539
social science	American Sociological Review	3932
social science	Journal of Organizational Behavior	2012
psychology	Annual Review of Psychology	871
psychology	Journal of Applied Psychology	3523
psychology	Psychological Bulletin	1894
philosophy	American Philosophical Quarterly	350
philosophy	Journal of the History of Philosophy	2062
medicine	The Lancet	50736
math	Annals of Mathematics	1118
math	Journal of the American Mathematical Society	632
literature science	American Journal of Philology	683
literature science	Poetics	1406
law	American Journal of International Law	2758
law	Berkeley Journal of International Law	106
law	European Journal of International Law	1339
history	Annales. Histoire, Sciences Sociales	330
history	History	578
history	The Historical Journal	2297
education	Educational Research Review	1313
education	Educational Researcher	1579
education	Review of Educational Research	1953

## А что если добавить других дисциплин?



# Математика



## Список литературы I

Rita Gonzalez-Marquez, Luca Schmidt, Benjamin M Schmidt, Philipp Berens, and Dmitry Kobak. The landscape of biomedical research. *bioRxiv*, 2023. doi: <https://doi.org/10.1101/2023.04.10.536208>.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

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.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

## Список литературы II

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