

# Построение ландшафта лингвистики: первые результаты и поиск стыков с другими науками

Г. А. Мороз

Международная лаборатория языковой конвергенции (НИУ ВШЭ,  
Москва)

«Дизайн междисциплинарных исследований в контексте сближения  
моделей естественно-научного и гуманитарно-социального знания»,  
МФТИ

4 октября 2023

# План доклада

Введение

Наш проект

Сбор данных

Результаты

Заключение

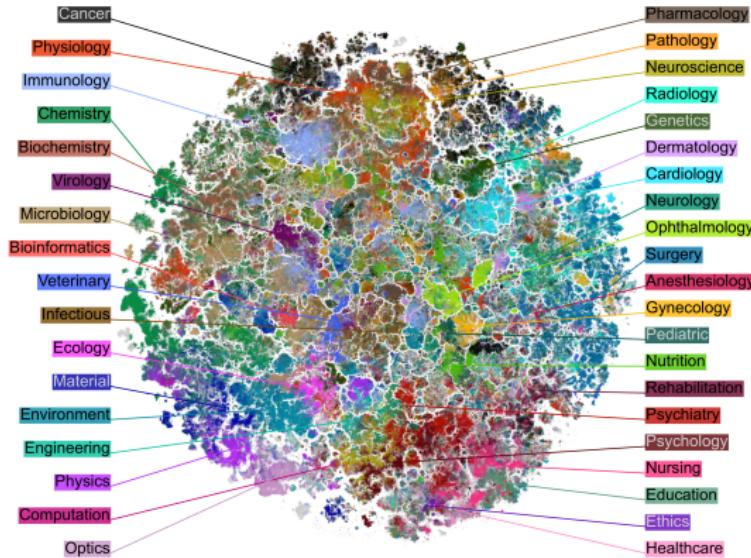
## (Gonzalez-Marquez et al. 2023) ландшафт биомедицинских исследований

*The number of publications in biomedicine and life sciences has rapidly grown over the last decades, with over 1.5 million papers now published every year. This makes it difficult to keep track of new scientific works and to have an overview of the evolution of the field as a whole. Here we present a 2D atlas of the entire corpus of biomedical literature, and argue that it provides a unique and useful overview of the life sciences research. <...>*

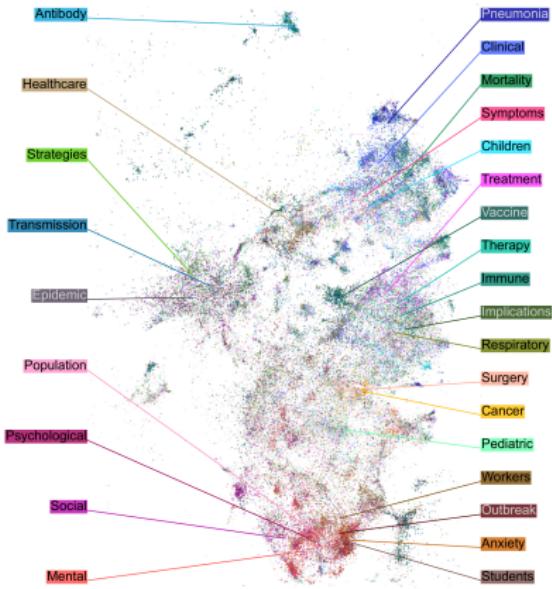
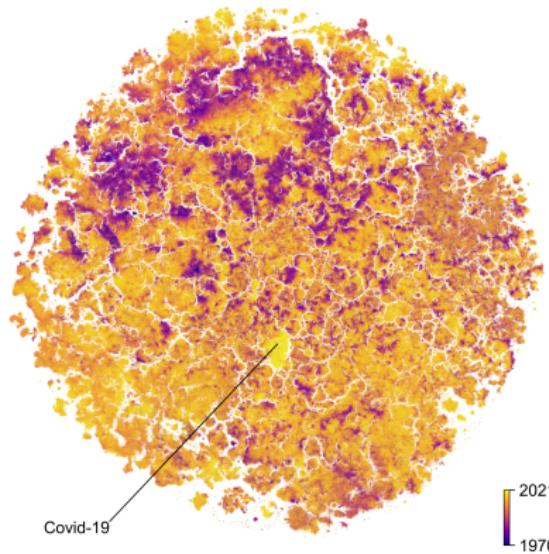
<https://static.nomic.ai/pubmed.html> (интерактивная версия)

Это препринт!

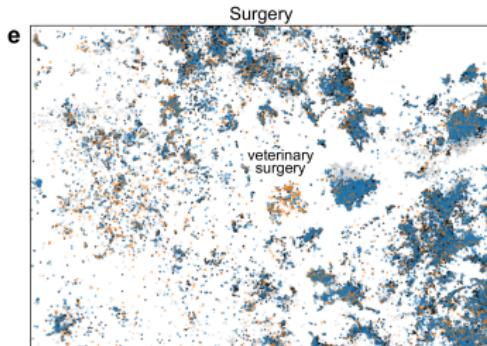
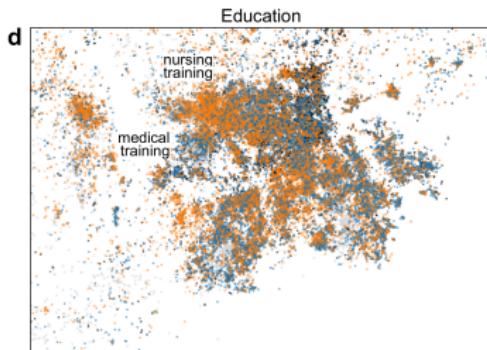
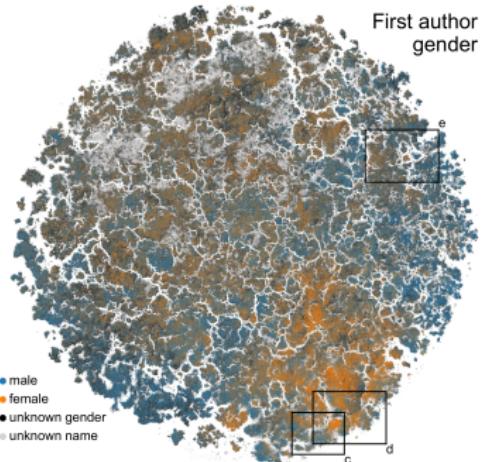
# (Gonzalez-Marquez et al. 2023)



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.

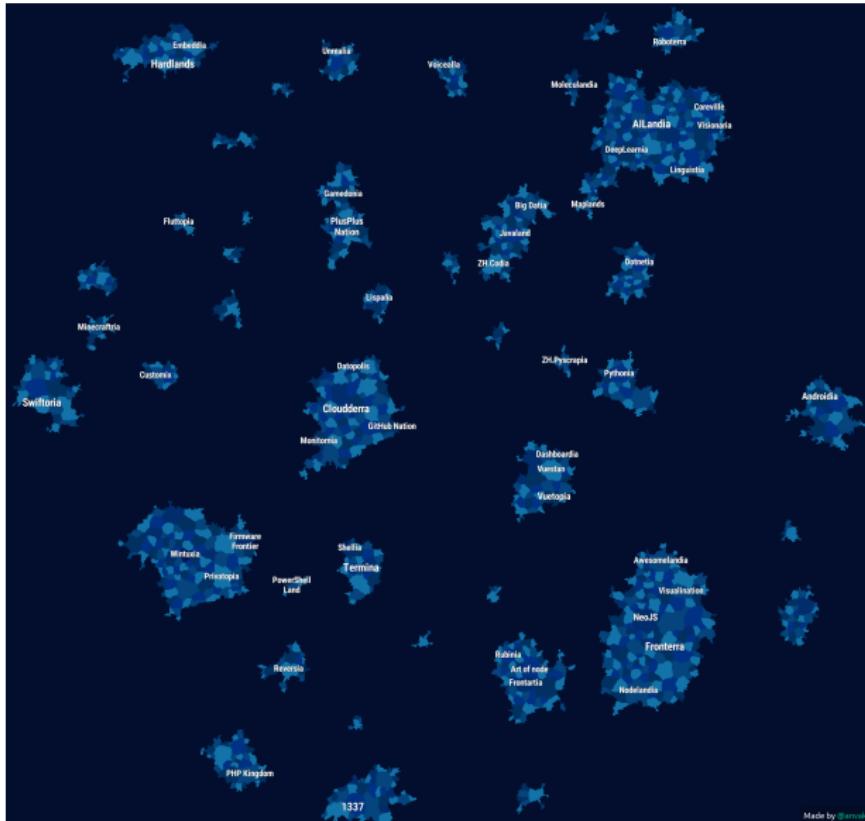


Регион карты, посвященный Covid-19. Цвета приписаны на основе названий работ. Кроме того здесь есть около 15% работ не посвященных короновирусу.



Статьи раскрашены по полу первого автора.

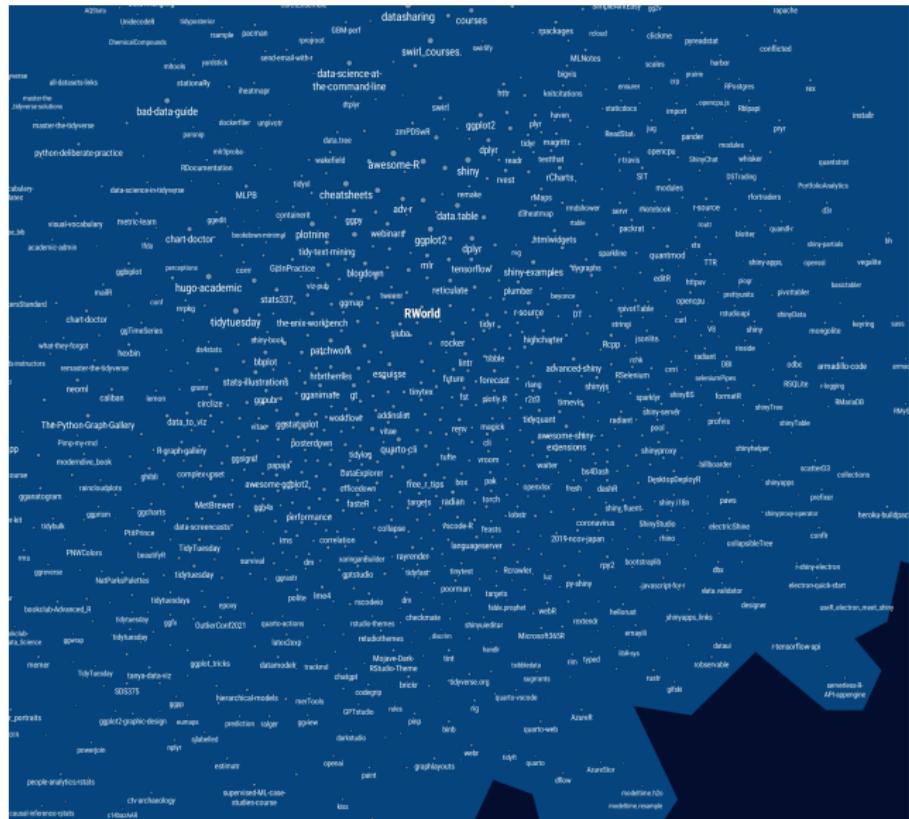
# Карта репозиториев гитхаба (Андрей Кашча)



Made by @anvaka

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

# Карта репозиториев гитхаба (Андрей Кашча)



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

# Библиометрические исследования

Библиометрия – дисциплина, возникшая в конце XIX века, в рамках которой можно встретить разные применения математических методов к исследованию научных работ. Наиболее известные применения:

- графы соавторства
- библиографические ссылки
- ключевые слова
- измерение качества журналов
- и др.

# План доклада

Введение

Наш проект

Сбор данных

Результаты

Заключение

# Команда

- руководители
  - Г. Мороз
- студенты
  - А. Агроскина (б)
  - А. Алексеева (б)
  - Т. Дедов (б)
  - А. Орехов (м)
  - К. Сидоров (м)
  - А. Степанова (б)

## План исследования

- 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
- исследовать, насколько релевантно для лингвистики получившееся пространство
- выявить и исследовать возможные междисциплинарные стыки

# План доклада

Введение

Наш проект

Сбор данных

Результаты

Заключение

# Списки журналов

Мы использовали несколько источников журналов

- Тэг филология, лингвистика, медиакоммуникации в Вышкинском списке журналов

Списки НИУ ВШЭ / HSE Journal Lists

Название	ISSN	Сп... ▾	Категория
АЛГИМЕНТАТИВНАЯ И АЛЬТЕРНАТИВНАЯ КОММУНИКАЦИЯ	0741-4618; 1477-...	А	БИОЛОГИЯ, МЕДИЦИНА И ЗДРАВООХРАНЕНИЕ, ФИЗИО...
ПРОСТЫЕ ЯЗЫКИ И КУЛЬТУРЫ	1565-1923; 1568-...	А	ФИЛОЛОГИЯ, ЛИНГВИСТИКА И МЕДИАКОММУНИКАЦИИ...
АКТЫ БОРЕАЛИИ	0866-3831; 1563-...	А	МОДНОСТЬ И ГУМАНИТАРНЫЕ НАУКИ: ИСТОРИЯ, АРХИВОВА...

- Тэг 6162 Languages в списке журналов из ресурса [Finish Publication Forum](#)

Results 1 - 20 / 1245      First Previous Next Last

Level	Title
A	AALITRA REVIEW
2	ACROSS LANGUAGES AND CULTURES
1	ACROSS THE DISCIPLINES
1	ACTA ACUSTICA

# Списки журналов

После соединения списков журналов мы по своему усмотрению разметили их по некоторым категориям (теги: linguistics (358), interdisciplinary (433), language\_learning (69) и другие).

HSE level	Helsenki level		
	a	b	c
a	42	37	15
b	1	3	10
c	0	24	124
d	0	1	4

## Списки журналов

После соединения списков журналов мы по своему усмотрению разметили их по некоторым категориям (теги: linguistics (358), interdisciplinary (433), language\_learning (69) и другие).

HSE level	Helsenki level		
	a	b	c
a	42	37	15
b	1	3	10
c	0	24	124
d	0	1	4

Разметка “лингвистиченоти” журналов – огромная и слабо автоматизируемая работа, которая требует экспертизы в самых разных областях лингвистики.

# Сбор аннотаций лингвистических исследований

- Мы планировали написать краулер, который бы собирал статьи из желаемых журналов...

# Сбор аннотаций лингвистических исследований

- Мы планировали написать краулер, который бы собирал статьи из желаемых журналов...
- но мы обнаружили базу данных Crossref и соответствующий пакет для R `rcrossref` ([Chamberlain et al. 2022](#))...

# Сбор аннотаций лингвистических исследований

- Мы планировали написать краулер, который бы собирал статьи из желаемых журналов...
- но мы обнаружили базу данных Crossref и соответствующий пакет для R `rcrossref` ([Chamberlain et al. 2022](#))...
- а потом мы обнаружили базу данных OpenAlex и соответствующий пакет для R `openalexR` ([Aria and Le 2023](#))

## Чистка аннотаций

- заметки редактора

## Чистка аннотаций

- заметки редактора
- некрологи и поздравления

## Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций

## Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг

## Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала

## Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации, отмененных (retracted) статей

## 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

## 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

# 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

## 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
- библиографическое описание книги (в случаях реview)

# 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
- библиографическое описание книги (в случаях рецензии)
- начало статьи вместо аннотации (характерно для старых статей)

# 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
- библиографическое описание книги (в случаях рецензии)
- начало статьи вместо аннотации (характерно для старых статей)
- ошибки распознавания

## 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
- библиографическое описание книги (в случаях рецензии)
- начало статьи вместо аннотации (характерно для старых статей)
- ошибки распознавания
- слишком короткие/длинные аннотации

# Примеры проблемных аннотаций



<https://doi.org/10.1007/BF02743731>

Search [Login](#)

[Home](#) > [Russian Linguistics](#) > Article

Articles | [Published: June 1990](#)

## ъжык БытА / ъжыкиИ ДУХОВ НОИ КУЛЬТУРЫ

[Russian Linguistics 14, 129–146 \(1990\)](#) | [Cite this article](#)

18 Accesses | 3 Citations | [Metrics](#)

This is a preview of subscription content, [access via your institution](#).

### ЛИТЕРАТУРА

БАРт, Р.: 1978, 'ЛИНГВИСТИКА ТЕКСТА'.НОВОЕ В ЖАРУБЕ ЕЖНОЙ ЛИНГВИСТИКЕ. Вып. VIII. ЛИНГВИСТИКА ТЕКСТА, МОСКВА, 442–449.

Access via your institution

Access options

Buy article PDF

39,95 €

Price includes VAT (Russian Federation)

Instant access to the full article PDF.

[Rent this article via DeepDyve](#)

# Примеры проблемных аннотаций

## LANGUAGE OF THE TOBACCO MARKET

ROBERT J. FITZPATRICK

*Louisville, Kentucky*

LISTEN to the chant of the tobacco auctioneer.' 'Fo'teen-a-lee-di-leen-a-  
lee-di-leen — — — qwa-qwa-qwa-qwa-aw-aw — — — ha-ha-ha-ha-ha  
— — — three-di-lee-di-lee — — fifteen — American.' How familiar is this  
chant to the listeners of a well-known radio program. Yet how many of  
them could tell if they heard the same jargon at a real tobacco auction  
that the bid on a pile of tobacco had been opened at fourteen dollars a  
hundred pounds, that the buyers had raised the bid to \$14.25, to \$14.50,  
to \$14.75, and that the tobacco had finally been sold at \$15.00?

<https://doi.org/10.2307/486818>

# План доклада

Введение

Наш проект

Сбор данных

Результаты

Заключение

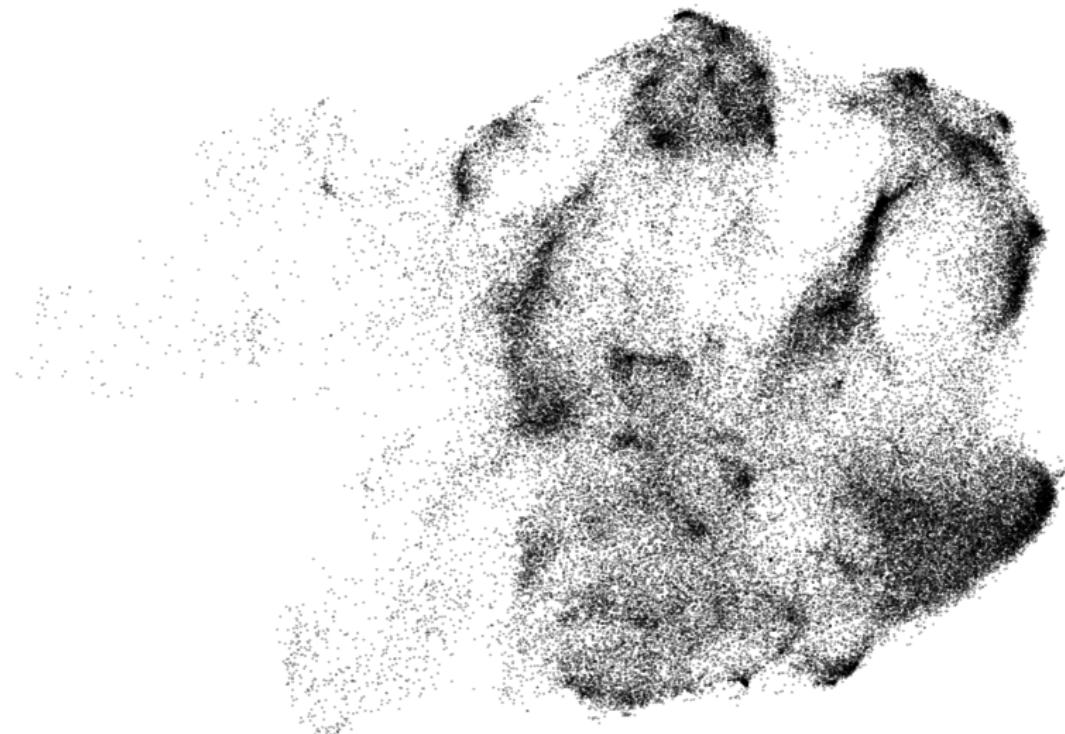
# Структура данных: 82087 строчек, 26 колонок

- **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; Computer science; Philosophy; Biochemistry; Chemistry; Complementation; Psychotherapist; Gene; Phenotype
- **retrieved:** 30-04-2023

# Векторное представление слов

- мы использовали векторизатор doc2vec (Le and Mikolov 2014; Wijffels 2021) (смотрели GloVe (Pennington et al. 2014), думаем в сторону BERT (Devlin et al. 2018) и RoBERTa (Liu et al. 2019))
- полученное 50-мерное пространство мы сократили до 2D при помощи t-SNE (Van der Maaten and Hinton 2008)

# Ландшафт лингвистических исследований



# Аннотации журналов, в названиях которых содержится *phon*



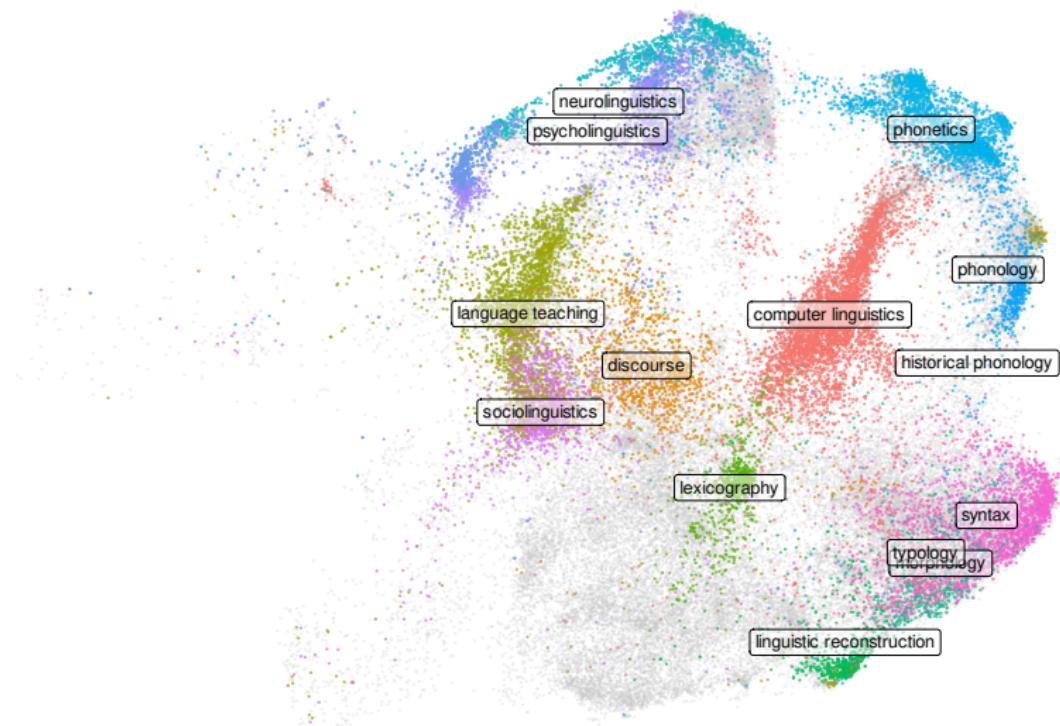
## Аннотации журналов, в названиях которых содержится *psych*



## Аннотации, в которых упоминается *Noam Chomsky*



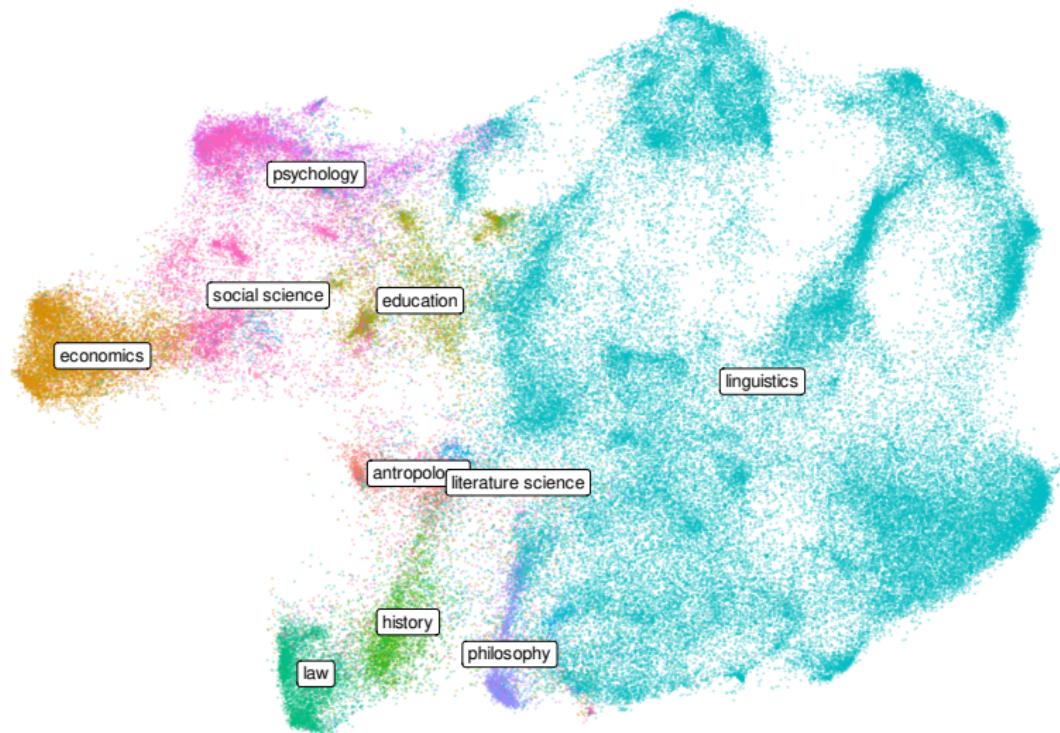
# Наша ручная аннотация статей



# А что если добавить других дисциплин?

field	journal	n
antropology	Annual Review of Anthropology	915
antropology	Current Anthropology	1719
economics	Journal of Political Economy	2654
economics	The American Economic Review	6417
education	Educational Research Review	1313
education	Educational Researcher	1579
education	Review of Educational Research	1953
history	Annales. Histoire, Sciences Sociales	330
history	History	578
history	The Historical Journal	2297
law	American Journal of International Law	2758
law	Berkeley Journal of International Law	106
law	European Journal of International Law	1339
literature science	American Journal of Philology	684
literature science	Poetics	1407
philosophy	American Philosophical Quarterly	350
philosophy	Journal of the History of Philosophy	2066
psychology	Annual Review of Psychology	871
psychology	Journal of Applied Psychology	3526
psychology	Psychological Bulletin	1894
social science	Administrative Science Quarterly	1568
social science	American Sociological Review	3932
social science	Journal of Organizational Behavior	2012

# А что если добавить других дисциплин?



# План доклада

Введение

Наш проект

Сбор данных

Результаты

Заключение

## Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета

## Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать местастыка и области междисциплинарных исследований

## Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать местастыка и области междисциплинарных исследований
- Необходимо продолжить исследование полученного пространства

## Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать местастыка и области междисциплинарных исследований
- Необходимо продолжить исследование полученного пространства
- Разметка журналов нуждается в дополнительной проверке

## Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать местастыка и области междисциплинарных исследований
- Необходимо продолжить исследование полученного пространства
- Разметка журналов нуждается в дополнительной проверке
- Необходимо доразметить аннотации

# Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать местастыка и области междисциплинарных исследований
- Необходимо продолжить исследование полученного пространства
- Разметка журналов нуждается в дополнительной проверке
- Необходимо доразметить аннотации
- Интересно посмотреть другие векторизаторы (BERT, RoBERTa)

# Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать местастыка и области междисциплинарных исследований
- Необходимо продолжить исследование полученного пространства
- Разметка журналов нуждается в дополнительной проверке
- Необходимо доразметить аннотации
- Интересно посмотреть другие векторизаторы (BERT, RoBERTa)
- Интересно сравнить разные методы уменьшения размерности (t-SNE vs UMAP)

# Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать местастыка и области междисциплинарных исследований
- Необходимо продолжить исследование полученного пространства
- Разметка журналов нуждается в дополнительной проверке
- Необходимо доразметить аннотации
- Интересно посмотреть другие векторизаторы (BERT, RoBERTa)
- Интересно сравнить разные методы уменьшения размерности (t-SNE vs UMAP)
- Не стоит обобщать знания, полученные на основе журналов на всю лингвистику

# Заключение

- Наши методы позволяют смотреть на предметную область с высоты птичьего полета
- Наши методы позволяют искать места стыка и области междисциплинарных исследований
- Необходимо продолжить исследование полученного пространства
- Разметка журналов нуждается в дополнительной проверке
- Необходимо доразметить аннотации
- Интересно посмотреть другие векторизаторы (BERT, RoBERTa)
- Интересно сравнить разные методы уменьшения размерности (t-SNE vs UMAP)
- Не стоит обобщать знания, полученные на основе журналов на всю лингвистику
- Не все области гуманитарного и социального знания одинаково представлены в журнальных публикациях

Спасибо за внимание!

## Литература

Massimo Aria and Trang Le. *openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API*, 2023. URL <https://CRAN.R-project.org/package=openalexR>. R package version 1.0.2.9.

Scott Chamberlain, Hao Zhu, Najko Jahn, Carl Boettiger, and Karthik Ram. *rcrossref: Client for Various 'CrossRef' APIs*, 2022. URL <https://CRAN.R-project.org/package=rcrossref>. R package version 1.2.0.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

## Литература

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.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

## Литература

- Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.
- Jan Wijffels. *doc2vec: Distributed Representations of Sentences, Documents and Topics*, 2021. URL  
<https://CRAN.R-project.org/package=doc2vec>. R package version 0.2.0.