

What do we know about linguistic journals?

Applying NLP methods to a dataset of abstracts from linguistic journals
(work in progress)

George Moroz, Asya Alekseeva, Timofei Dedov, Artem Orekhov, Kirill Sidorov, Angelina Stepanova

2 May 2023

Outline of the talk

Introduction

Data collection

Data analysis

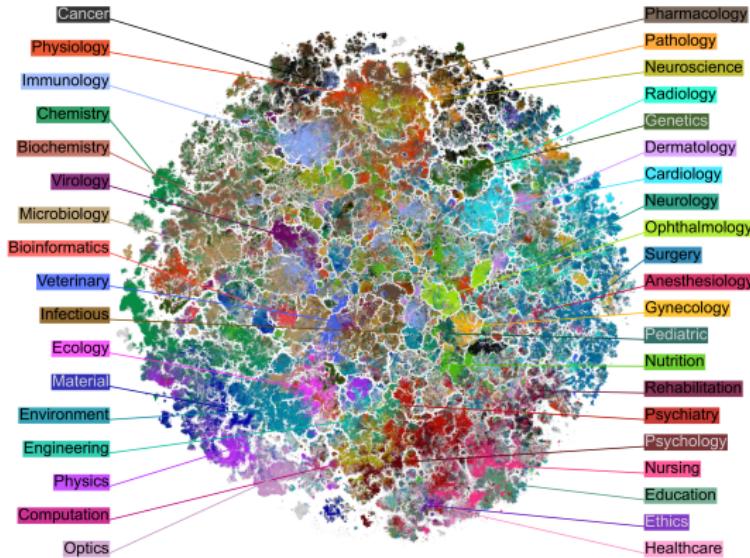
Conclusions

(Gonzalez-Marquez et al. 2023) The landscape of biomedical research

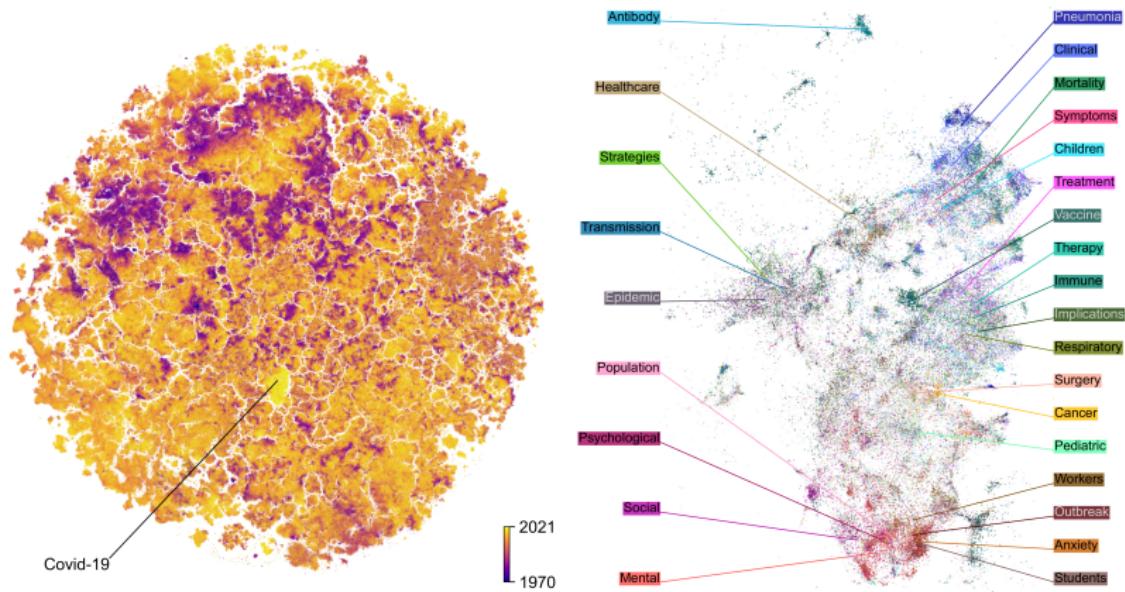
The number of publications in biomedicine and life sciences has rapidly grown over the last decades, with over 1.5 million papers now published every year. This makes it difficult to keep track of new scientific works and to have an overview of the evolution of the field as a whole. Here we present a 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> (interactive version)

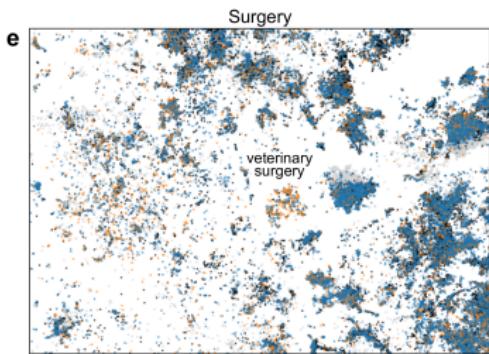
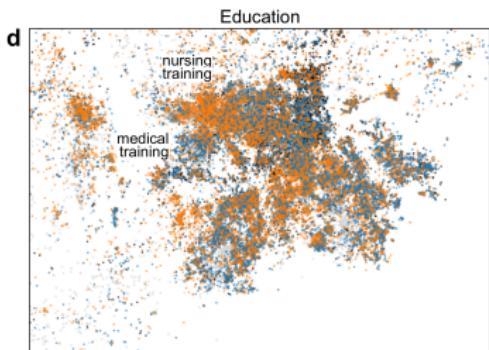
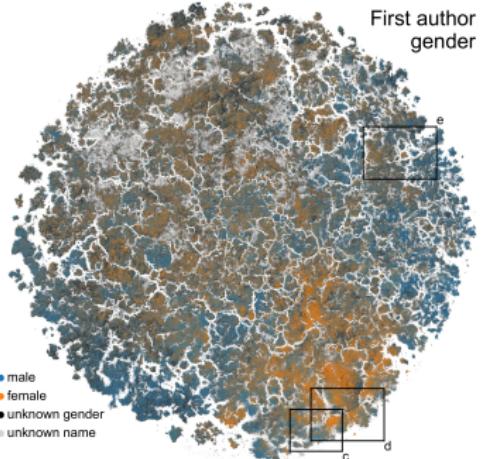
This is a preprint and has not been certified by peer review!



2D embedding of the PubMed dataset. Paper abstracts ($n = 21\text{ M}$) were transformed into 768-dimensional vectors with PubMedBERT (Gu et al. 2021) and then embedded in 2D with t-SNE (Van der Maaten and Hinton 2008). Coloured using labels based on journal titles. Unlabeled papers are shown in gray and are displayed in the background.

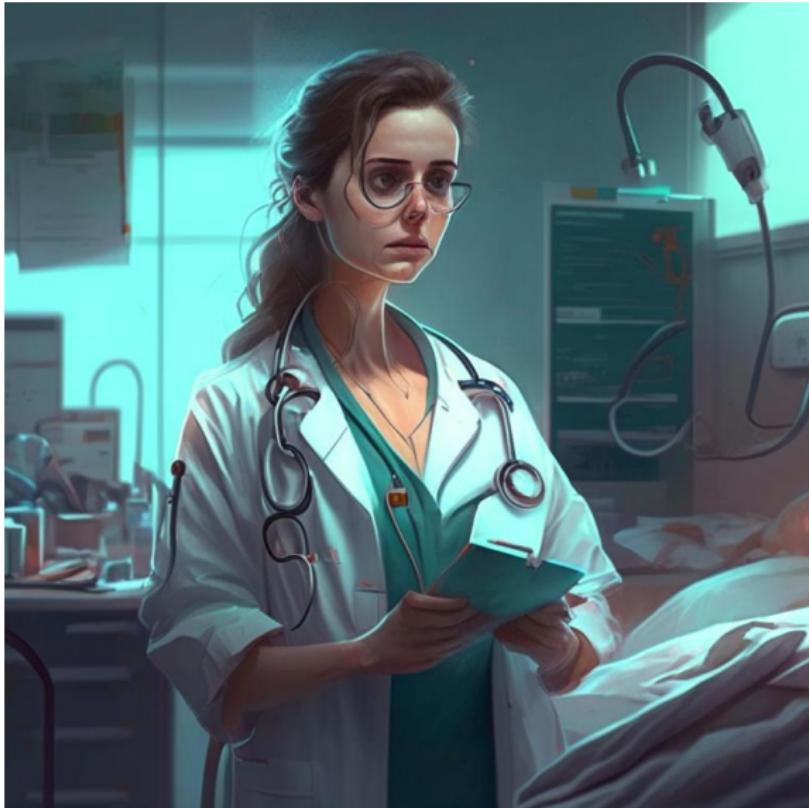


Covid-19 region of the map. Colours are assigned using labels based on paper titles. Unlabeled Covid papers are shown in the background in gray. This region in the embedding also contained some non-Covid papers (~15%) about other respiratory epidemics; they are not shown.



Papers coloured by the inferred gender of their first authors.(d–e)
Regions of the map showing within-label heterogeneity in the distribution of first authors' gender.

Reaction of professional doctors and pharmacologists



AI generated picture

How wonderful would it be to have something like this for linguistics?

- zoning of different subfields of linguistics
- interaction of different subfields of linguistics
 - can we see a boundary between morphology and syntax?
 - can we see cancer-like behavior of the Computer linguistics among all subfields of modern linguistics?
 - can we visualize the emergence of Computer linguistics?
 - can we show how strong/vague are boundaries between philology, Digital Humanity and linguistics?
 - can we visualize the history of linguistics?
 - can we see the scientific track/preferences of different researchers? E. g. if the person is a phonetician will this person also be into the computational linguistics?
 - ...

Plan

- extract abstracts from publications in linguistic journals
- use them in order to create some space of linguistic publications (one dot – one publication)
- see how linguistic journals are distributed within this space

Outline of the talk

Introduction

Data collection

Data analysis

Conclusions

We have different journal lists

- Tag филология, лингвистика, медиакоммуникации from list of journals HSE uses for Academic Merit Bonus.¹

Списки журналов НИУ ВШЭ / HSE Journal Lists			
Список А / List A	Список В / List B	Список С / List C	Список D / List D
627	43	1941	36

Выберите фильтр / Choose filter

Поиск по Спискам / List Search:

А В С Д Е

Поиск по Категории / Category Search:

ФИЛОЛОГИЯ, Л... Е

Название	ISSN	Св...	Категория
AAC: AUGMENTATIVE AND ALTERNATIVE COMMUNIC...	0743-9618; 1477-...	А	БИОЛОГИЯ, МЕДИЦИНА И ЗДРАВООХРАНЕНИЕ; ФИЛО...
ACROSS LANGUAGES AND CULTURES	1565-1923; 1588-...	А	ФИЛОЛОГИЯ, ЛИНГВИСТИКА И МЕДИАКОММУНИКАЦИ...
ACTA BOREALIA	0806-3831; 1503-...	А	МОНОСТРОЮЩИЕ ГУМАНИТАРНЫЕ НАУКИ; ИСТОРИЯ, АРХ...

- Tag 6162 Languages from journal rankings from [Finish Publication Forum](#)

Results 1 - 20 / 1245	First	Previous	Next	Last
LevelTitle				
1	AALITRA REVIEW			
2	ACROSS LANGUAGES AND CULTURES			
1	ACROSS THE DISCIPLINES			
1	ACTA ACUSTICA			

¹As far as I know, we are not allowed to disclosure those lists outside the HSE. However, right now they are available without any checks for HSE affiliation.

Journal lists

After we gathered journal lists we annotated them according to our beliefs on what is linguistics, and what is not.

There is also a side-project on annotation of journals for literature science.

After manual annotation 232 (16%) journals out of 1421 have tag linguistics. 1334 journals have not yet been annotated.

HSE level	Helsenki level				
	a	b	c	d	NA
a	50	51	23	1	1
b	1	3	12	NA	3
c	NA	7	42	1	37

Abstracts extraction

Outline of the talk

Introduction

Data collection

Data analysis

Conclusions

Outline of the talk

Introduction

Data collection

Data analysis

Conclusions

Thank you for your attention!

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.

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