

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, Artyom Orekhov, Kirill Sidorov, Angelina Stepanova

2 May 2023

Outline of the talk

Introduction

Our team

Data collection

Data analysis

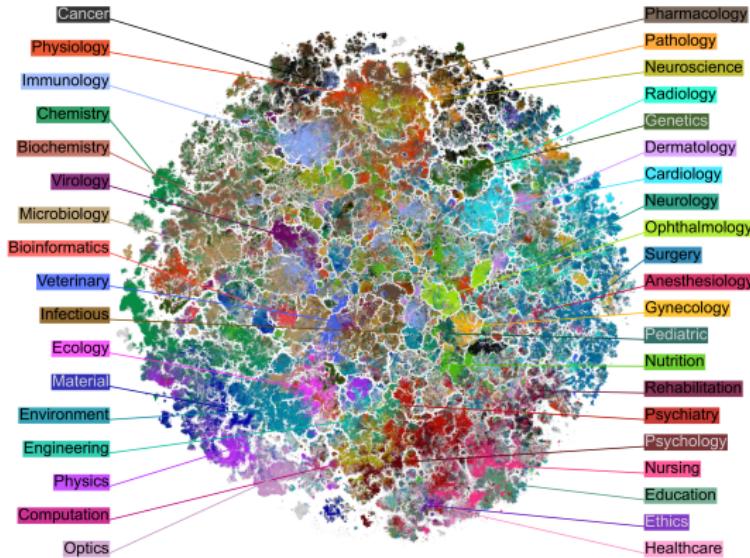
Conclusions

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

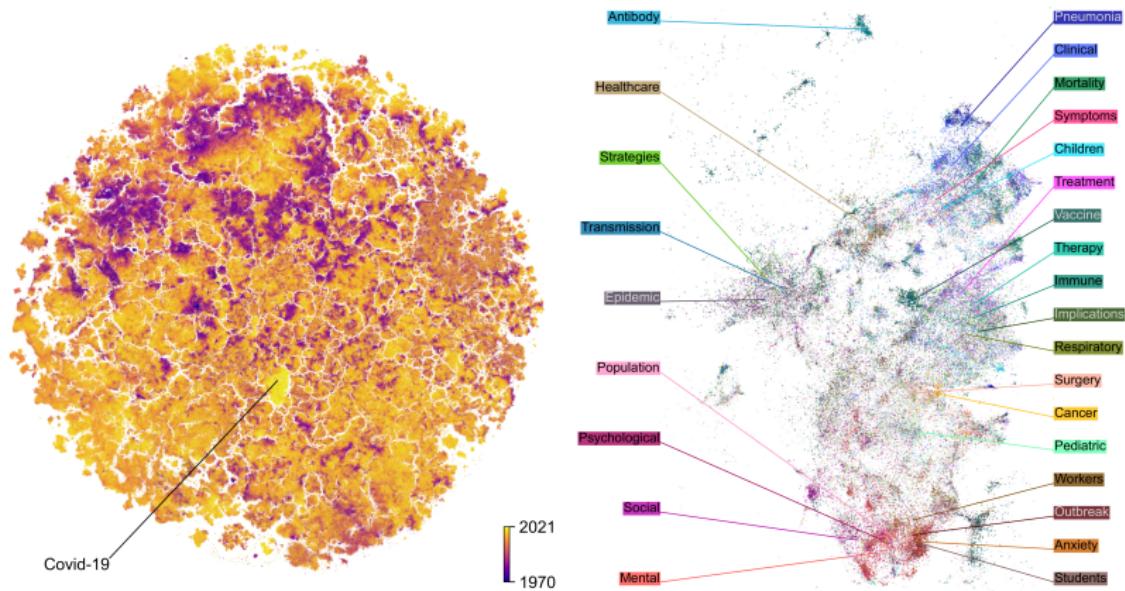
The number of publications in biomedicine and life sciences has rapidly grown over the last decades, with over 1.5 million papers now published every year. This makes it difficult to keep track of new scientific works and to have an overview of the evolution of the field as a whole. Here we present a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

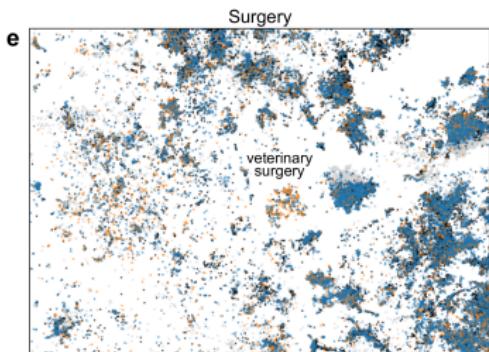
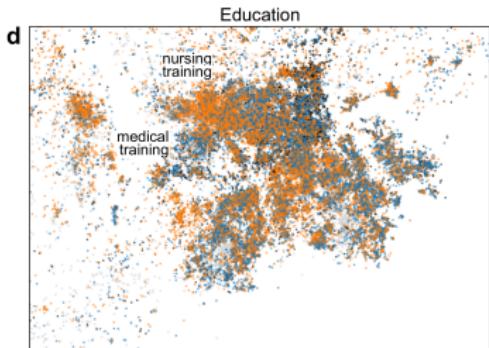
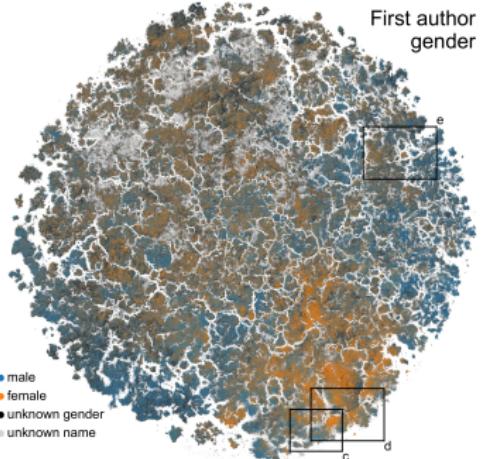
This is a preprint and has not been certified by peer review!



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.

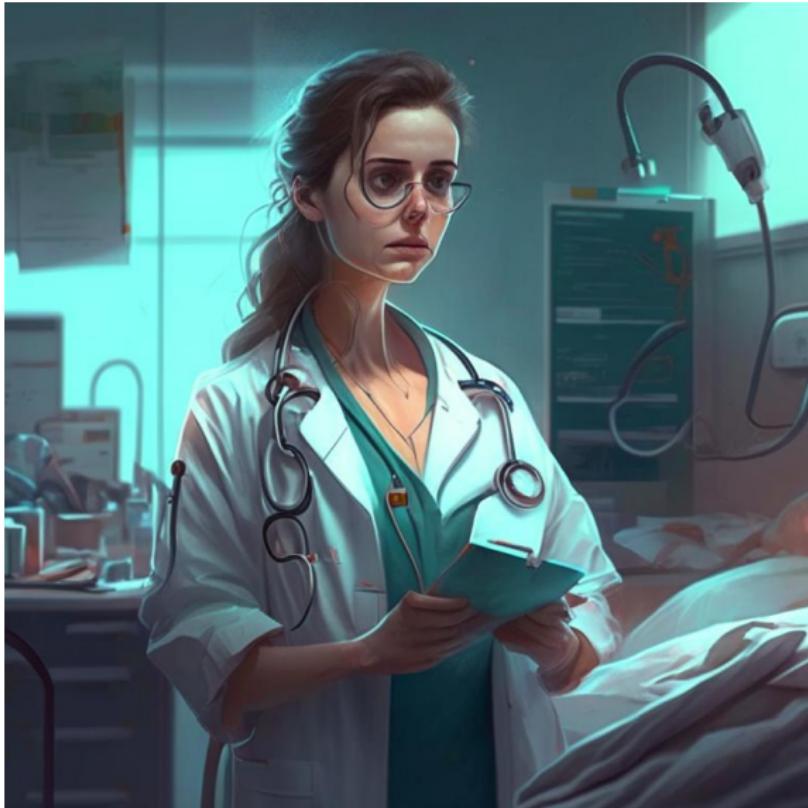


Covid-19 region of the map. Colours are assigned using labels based on paper titles. Unlabeled Covid papers are shown in the background in gray. This region in the embedding also contained some non-Covid papers (~15%) about other respiratory epidemics, they are not shown.



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

Our team

Data collection

Data analysis

Conclusions

Our team

- conceptualization: George Moroz, Boris Orekhov
- team leadership: George Moroz, Asya Alekseeva, Kirill Sidorov
- data curation, data analysis: George Moroz
- data gathering and annotation:
 - DH masters: Kirill Sidorov and Artyom Orekhov
 - bachelor students: Asya Alekseeva, Timofei Dedov and Angelina Stepanova
 - group of bachelor students who choose this project as a summer practice (2-week internship)

Outline of the talk

Introduction

Our team

Data collection

Data analysis

Conclusions

Journal lists

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

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

- 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 personal beliefs on what is linguistics, and what is not (tags: linguistics (232), interdisciplinary (203), language_learning (26) and some others).

There is also a side-project on annotation of journals for literature science (tags: literary_studies (168), interdisciplinary (253), philology (12) and some others).

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

- We planed to scrap all papers' metadata from the journals' web-pages

Abstracts extraction

- We planed to scrap all papers' metadata from the journals' web-pages
- Then we discovered the `Crossref` database and the `rcrossref` package ([Chamberlain et al. 2022](#)) for R

Abstracts extraction

- We planed to scrap all papers' metadata from the journals' web-pages
- Then we discovered the `Crossref` database and the `rcrossref` package ([Chamberlain et al. 2022](#)) for R
- Then we discovered the `OpenA1ex` database and the `openalexR` package ([Aria and Le 2023](#)) for R

Abstracts extraction: scrapping journal web-pages

Since journals can be grouped by the publishers it is possible to write a web-scraper for the group of journals from one publisher. However there were some problems:

- blocking the IP address
- a lot of old publication are digitized in the form of a pdf of the first page of the paper
- change of the name and ISSN² of the journal
- change of the publisher of the journal
- there are still some mistakes on the journal websites

The screenshot shows a SpringerLink journal article page. At the top, there's a logo for SpringerLink, the URL <https://doi.org/10.1007/BF02743731>, a search bar, and a login link. Below the header, the page navigation includes Home > Russian Linguistics > Article. It shows the journal title 'Russian Linguistics' and the volume and issue information 'Articles | Published: June 1990' and 'БЪЛГАРСКА / БЪЛГАРСКИ ДУХОВ НОИ КУЛЬТУРЫ'. The article title is 'БЪЛГАРСКА / БЪЛГАРСКИ ДУХОВ НОИ КУЛЬТУРЫ'. Below this, there are links for 'Buy article PDF' (39.95 €), which includes VAT for the Russian Federation, and 'Instant access to the full article PDF'. There's also a link to 'Read this article via DeepDyve.'

²ISSN (International Standard Serial Number) — eight-digit serial number used to uniquely identify a serial publication.

Abstracts extraction: Crossref

Crossref is a digital object identifier (DOI) Registration Agency that provide an API to their database. Crossref does not provide a full text or abstracts of papers, however at the begining we used it in order to gather papers' metadata.

The last change of the pipeline for gathering data (30.04.2023) made this database non-relevant to the project.

Abstracts extraction: OpenAlex

[OpenAlex](#) is a free and open catalog of the world's scholarly papers, researchers, journals, and institutions. [OpenAlex](#) provide free well-documented API that make it possible to download all journal papers' metadata and abstracts using a list of journals' ISSNs (40 lines of R code).

In the further work we will discuss the result based on 222 linguistic journals downloaded from [OpenAlex](#).

The main advantage of the [OpenAlex](#) database is that they provide a concept information for each paper.

Abstracts extraction: OpenAlex concepts

Magnusson, E., Nauclér, K. (1993). The development of linguistic awareness in language-disordered children. *First language*, 13(37), 93–111.

This paper concerns the development of linguistic awareness in language-disordered children, and the role of linguistic and cognitive factors in this development. A longitudinal study was conducted with 39 language-disordered and a matched group of 39 normally-speaking children. Phonological awareness and other abilities were tested from preschool to Grade 4. Various comparisons were made of the abilities of the two groups. The findings indicate the importance of language comprehension and of syntactic production, but not of phonology, for the development of phonological awareness.

concepts	level	score
Linguistics	1	0.7674719
Psychology	0	0.5576790
Language development	2	0.5576040
Philosophy	0	0.1505756

Abstracts extraction: OpenAlex concepts

Arkadiev, P. M. (2023). Finiteness in morphology and syntax: Evidence from Abaza. *Voprosy Jazykoznanija*, (1), 103–131.

The article surveys morphological and syntactic ramifications of finiteness and nonfiniteness of verbal forms in the polysynthetic Abaza language (Northwest Caucasian, Karachay-Cherkessia). In Abaza, a whole range of morphological phenomena exhibit more or less robust correlations with dependent resp. subordinate status of the clause. However, I show that none of these phenomena can be treated as an unequivocal indicator of (non)finiteness, paying particular attention to several types of verbal forms that combine morphological features of both finiteness and nonfiniteness. Hypotheses motivating these “paradoxical” constellations of properties are put forward.

concepts	level	score
Syntax	2	0.7972705
Linguistics	1	0.7173848
Morphology (biology)	2	0.6633695
Range (aeronautics)	2	0.4956700
Psychology	0	0.4723184
Morpheme	2	0.4599627
Mathematics	0	0.3435510
Philosophy	0	0.2382300
Biology	0	0.0801864
Genetics	1	0.0000000
Materials science	0	0.0000000
Composite material	1	0.0000000

Abstracts extraction: OpenAlex concepts

OpenAlex has about 65k concepts that are listed in [this spreadsheet](#). Concepts are hierarchical and are stored in six layers. It is modified version of the similar tree made in ([Shen et al. 2018](#)) (based on Microsoft Academic Graph ([Sinha et al. 2015](#))).

Each work is tagged by a classifier with multiple concepts, based on the title, abstract, and the journal name. The classifier was trained on corpus of Microsoft Academic Graph.

Outline of the talk

Introduction

Our team

Data collection

Data analysis

Conclusions

Data structure

Outline of the talk

Introduction

Our team

Data collection

Data analysis

Conclusions

Thank you for your attention!

References

- 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.
- 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>.

References

- 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.
- Zhihong Shen, Hao Ma, and Kuansan Wang. A web-scale system for scientific knowledge exploration. *arXiv preprint arXiv:1805.12216*, 2018.
- Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015.

References

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