

Текст как Big Data: моделирование конвергентных процессов в языке и речи цифровыми методами

Г. А. Мороз

16.01.2024

План

О проекте

Задачи и результаты первого этапа проекта

Апробация результатов

Планы продолжения проекта

Участники

- (Москва) Международная лаборатория языковой конвергенции



(a) Г. А. Мороз



(b) Б. В. Орехов

- (Санкт-Петербург) Лаборатория языковой конвергенции



(a) Т. Ю. Шерстинова



(b) А. В. Колмогорова

О проекте

Актуальность проекта состоит в применении к большим текстовым массивам современных средств обработки естественного языка. Применения методов компьютерной лингвистики становится опытным полигоном для выявления границ и возможностей группы компьютерных методов в исследовании языка и текста.

Подпроекты (Москва):

- Построение ландшафта лингвистики на основе аннотаций
- Культуронимическое исследование
- Тематическое моделирование на корпусе текстов «Прожито»
- Создание языковых моделей для решения задач computational literary studies

Подпроекты (Санкт-Петербург):

- Моделирование картин мира писателя и поколения посредством компьютерного анализа больших коллекций художественных текстов (корпуса русского рассказа, корпуса текстов личной библиотеки С.Довлатова и его собственных сочинений, корпуса фанфиков)
- Создание устного корпуса речи молодежи и дополнение корпуса русского рассказа
- Методология описания “социального настроения эпохи” посредством компьютерного анализа коллекций текстов, находящихся на периферии идеологичности: корпуса открыток, корпуса текстов учебников по истории России, корпуса советских песен
- Разработка методов эмоциональной разметки текстов разной жанровой природы
- Исследование категории естественности устной и письменной речи в контексте задачи автоматической генерации текста

План

О проекте

Задачи и результаты первого этапа проекта

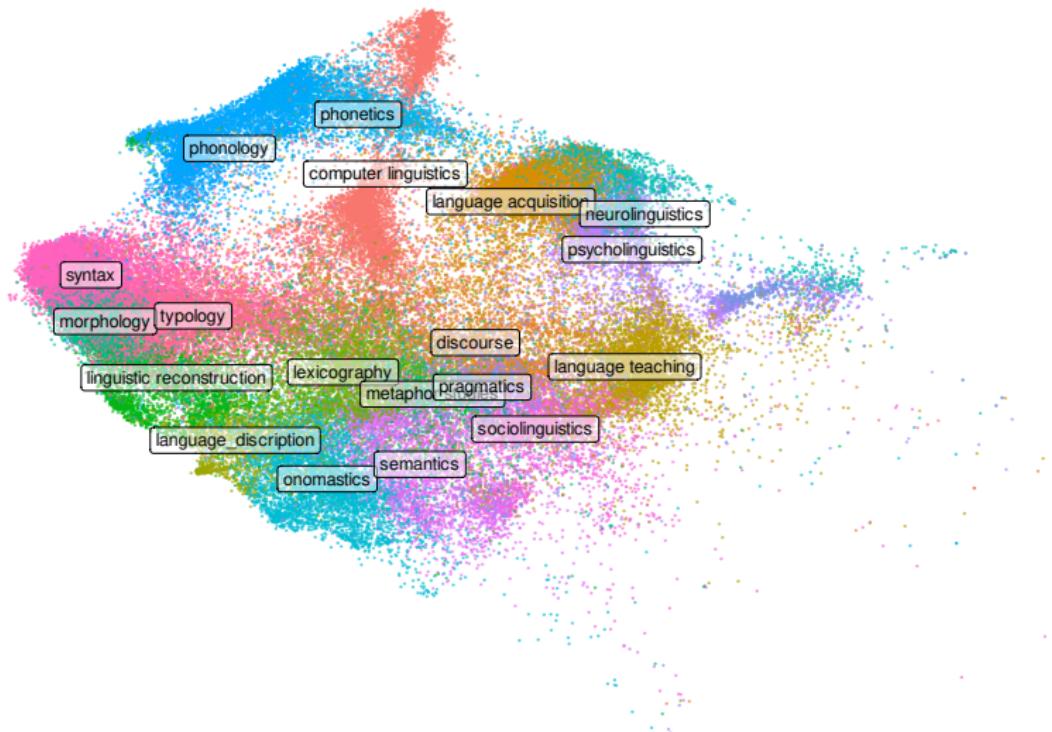
Апробация результатов

Планы продолжения проекта

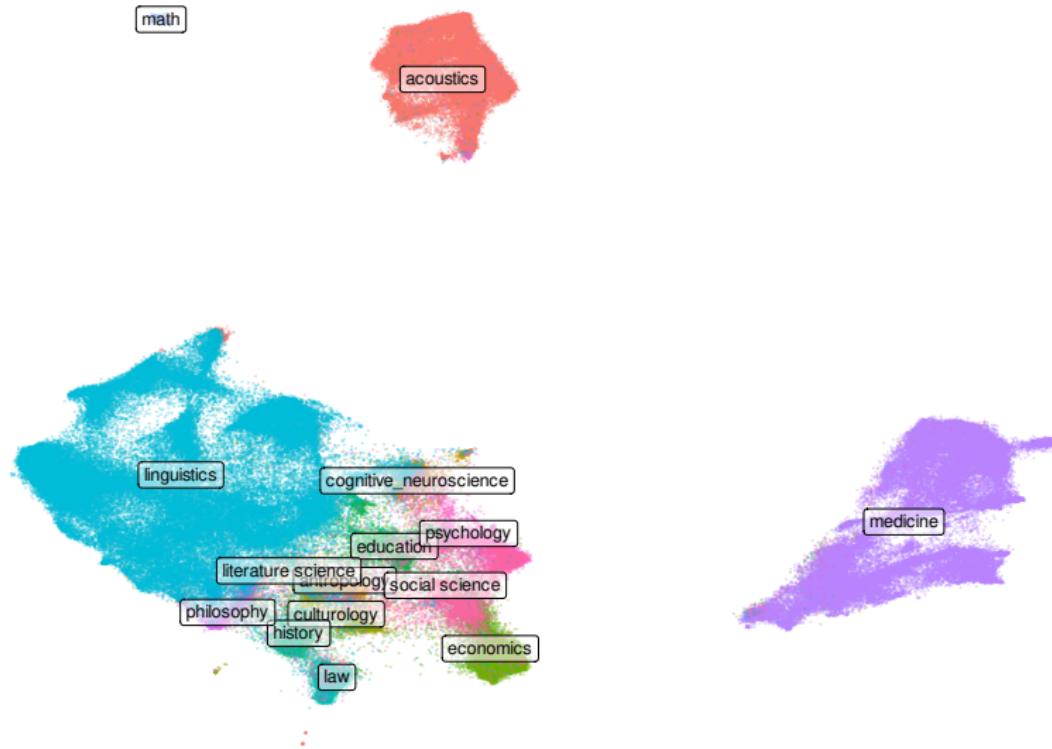
Построение ландшафта лингвистики

- сбор, чистка и разметка англоязычных аннотаций из **лингвистических журналов**
 - собрано более 80 000 аннотаций из 330 журналов
- сбор, чистка и разметка англоязычных аннотаций из **нелингвистических журналов**
 - собрано более 150 000 аннотаций из 32 журналов
- настройка методов векторизации и уменьшения размерности

Построение ландшафта лингвистики: результаты



Построение ландшафта лингвистики: результаты



Культуронимическое исследование

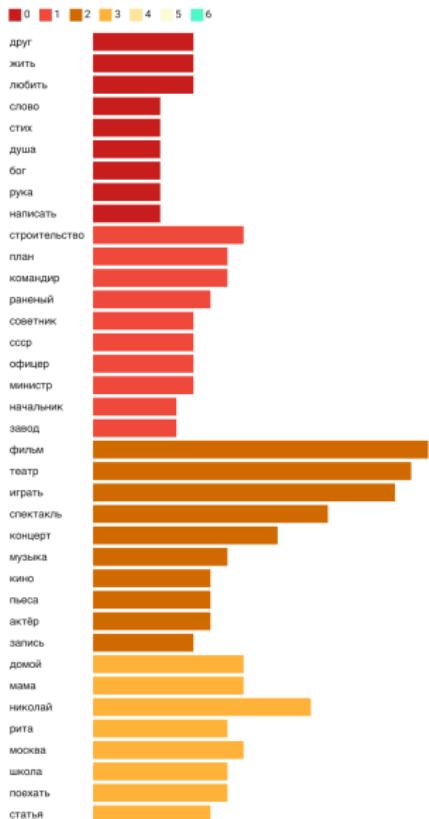
- создание статистического алгоритма для выявления культурономически значимых выбросов частотности
 - написан код на Python, данные в процессе обсчета

Тематическое моделирование на корпусе текстов «Прожито»

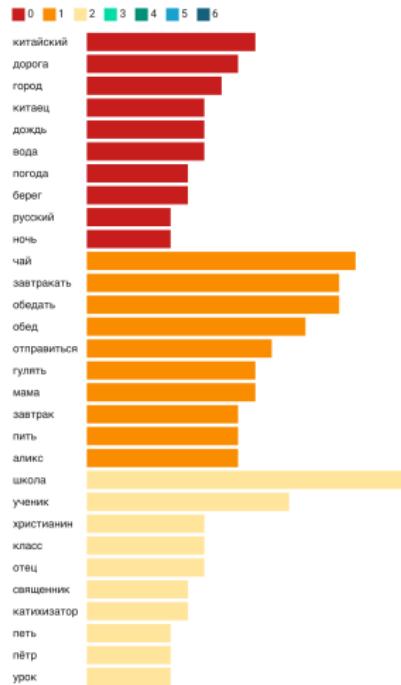
- построение тематических моделей на релевантных выборках
 - построено 6 моделей

Тематическое моделирование на корпусе текстов «Прожито»

Тематическое моделирование на записях, сделанных в 1971 – 1980 гг.



Тематическое моделирование на записях, сделанных в 1891 – 1900 гг.



Создание языковых моделей

- сбор корпуса, выбор задач, метрик оценки и архитектур
 - собран корпус художественной прозы на русском языке объемом 10 млрд словоупотреблений

План

О проекте

Задачи и результаты первого этапа проекта

Апробация результатов

Планы продолжения проекта

- **4 октября:** Международная конференция «Дизайн междисциплинарных исследований в контексте сближения моделей естественно-научного и гуманитарно-социального знания», МФТИ. Г. А. Мороз «Построение ландшафта лингвистики: первые результаты и поиск стыков с другими науками»
- **19 октября:** Международная конференция «Русская и зарубежная филология в диалоге культур», Южный федеральный университет. Г. А. Мороз «Построение ландшафта лингвистики: первые результаты»

Формы взаимодействия с кампусом СПб

- **16 июня:** открытый очно-заочный семинар, посвященный открытию Лаборатории языковой конвергенции в Санкт-Петербурге и презентации междисциплинарного проекта
- **24 октября:** круглый стол «Естественное мышление vs искусственный интеллект через призму исследований больших языковых данных»
- **19–21 октября:** Всероссийская научно-практическая конференция «Русская и зарубежная филология в диалоге культур»
- Ведение закрытого и открытого Телеграм-каналов для обсуждения текущих вопросов, онлайн и очные встречи (март, июнь, сентябрь, ноябрь, декабрь).

План

О проекте

Задачи и результаты первого этапа проекта

Апробация результатов

Планы продолжения проекта

Планы продолжения проекта

- активно взаимодействовать с коллегами из кампуса СПб
- продолжать исследования в рамках описанных направлений
- подготовить публикации по заявленным направлениям:

название	2023	2024	2025
Публикации в научных журналах, входящих в Список А	0	0	1
Публикации в научных журналах, входящих в Список В, С, D	2	2	2
Прочие публикации в научных журналах, входящих в ядро РИНЦ	2	2	2
Главы в рецензируемых монографиях РИД	0	1	0
	1	0	1