

# Анализ вариативности на материале диалектных корпусов и корпусов билингвального русского

Георгий Алексеевич Мороз, Светлана Сергеевна Земичева

Международная лаборатория языковой конвергенции (НИУ ВШЭ,  
Москва)

«День науки», Южный федеральный университет

презентация доступна по ссылке: [tinyurl.com/25toys5m](https://tinyurl.com/25toys5m)

06 апреля 2023

# План доклада

Обо мне

Прескриптивная vs. дескриптивная лингвистика

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

Чем донской корпус отличается от остальных корпусов?

Заключение

# Обо мне

- полевой исследователь (29 поездок)
- фонетист, фонолог, квантитативный лингвист
- езжу на Кавказ
- преподаю статистику и R (язык программирования)
- написал несколько лингвистических пакетов для R
  - `lingtypology`
  - `phonfieldwork`
  - `lingglosses`
- заведующий Международной лаборатории языковой конвергенции (НИУ ВШЭ, Москва)

# План доклада

Обо мне

**Прескриптивная vs. дескриптивная лингвистика**

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

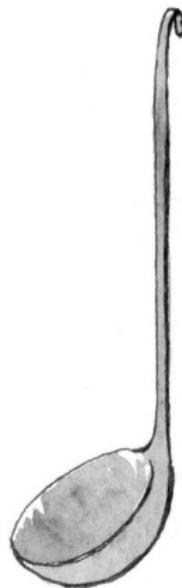
Марковский процесс

Чем донской корпус отличается от остальных корпусов?

Заключение

## Прескриптивная vs. дескриптивная лингвистика

Назовите, пожалуйста, что изображено на картинке.  
(рисунок Тани Пановой)



# Прескриптивная vs. дескриптивная лингвистика

Это часть опроса И. С. Левина:



# План доклада

Обо мне

Прескриптивная vs. дескриптивная лингвистика

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

Чем донской корпус отличается от остальных корпусов?

Заключение

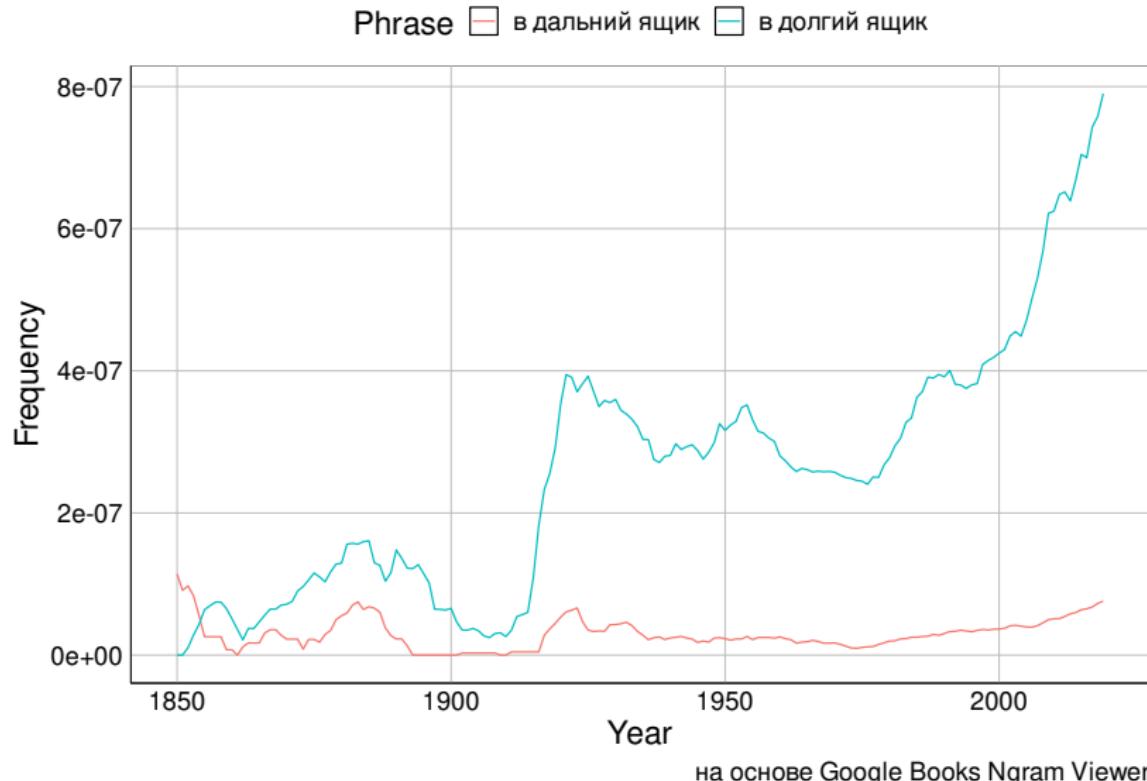
## Корпусная лингвистика

Корпусная лингвистика — это область лингвистики, которая занимается исследованием языковых явлений на материале некоторых собраний языкового материала. В большинстве случаев это письменные тексты, однако это может быть аудио и даже видео корпуса.

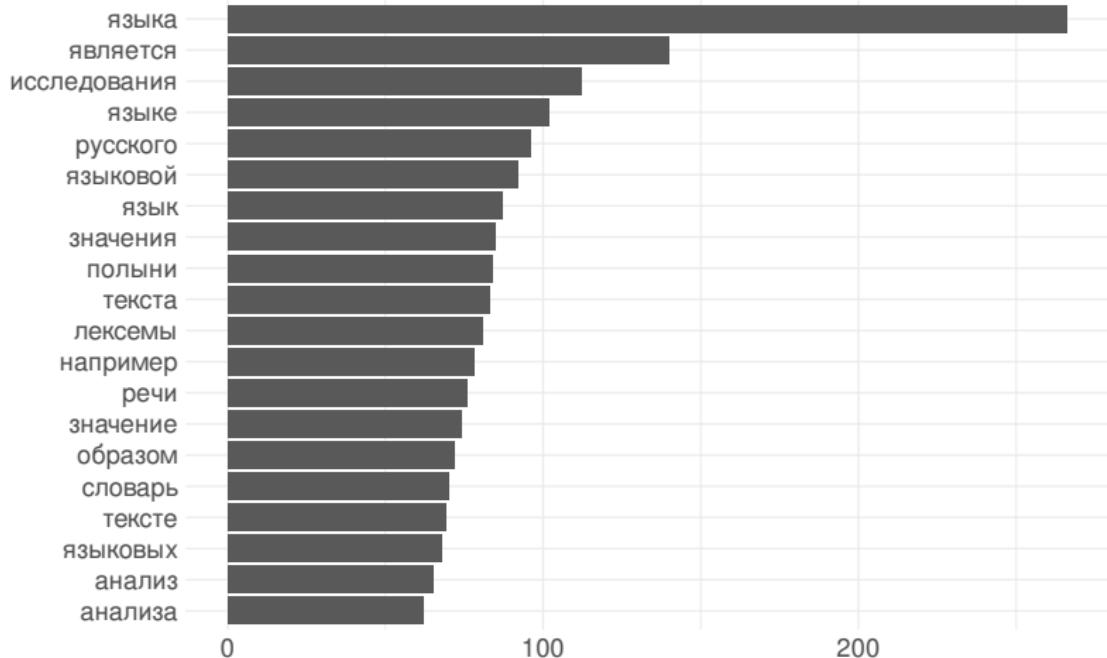
Среди корпусов русского языка можно назвать:

- Национальный корпус русского языка
  - более 1.5 млрд слов
  - много подкорпусов (газетный, устный, параллельный, диалектный, поэтический, исторические)
- Google Books Ngram Viewer
- ...

# Отложить в ... ящик

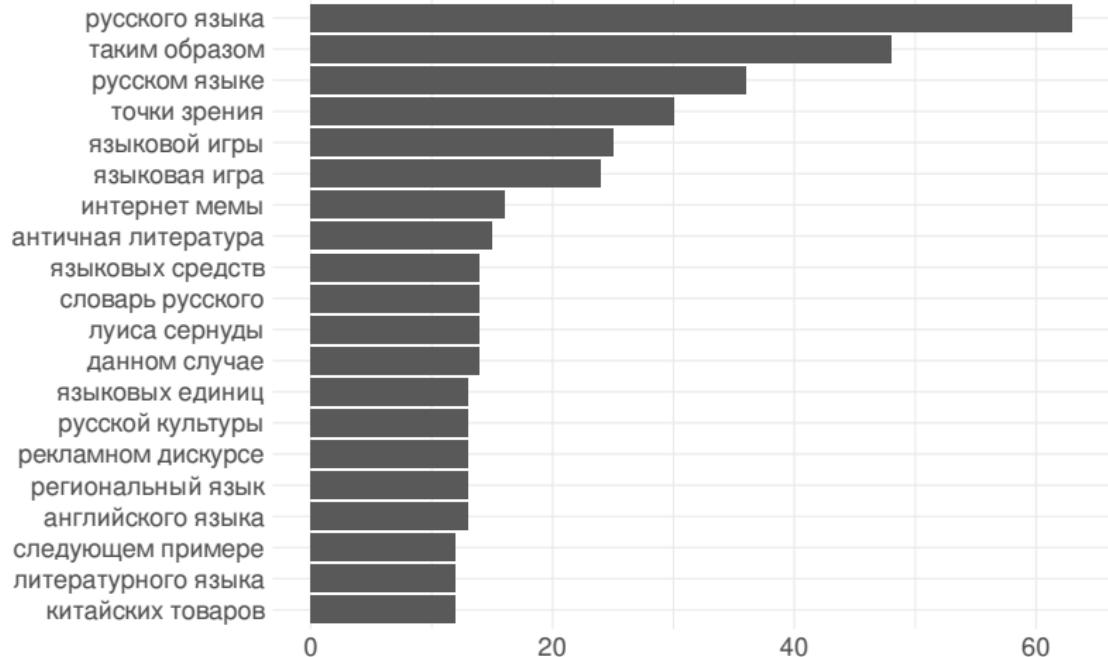


# 30 статей из Известий ЮФУ. Филологические науки (2022-2023)



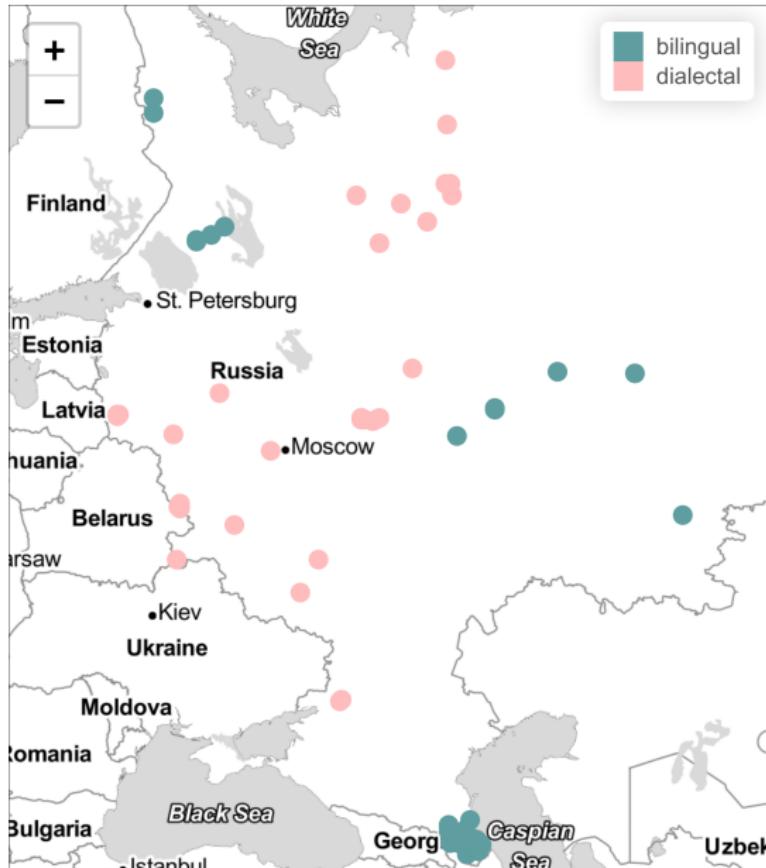
пришлось добавить в стоп-слова: дата, обращения, канд, филол, наук, дис, канд, южного, федерального, университета, гос, ун, список, источников, научная, статья, известия, юфу, др, филологические, науки, ключевые, слова, электронный, ресурс

# 30 статей из Известий ЮФУ. Филологические науки (2022-2023)



пришлось добавить в стоп-слова: дата, обращения, канд, филол, наук, дис, канд, южного, федерального, университета, гос, ун, список, источников, научная, статья, известия, юфу, др, филологические, науки, ключевые, слова, электронный, ресурс

# Диалектные устные корпуса лаборатории языковой конвергенции



# Диалектные устные корпуса лаборатории языковой конвергенции

Malinino Corpus

Corpus ▾ Statistics ▾ About ▾ Help | rus en |

Query / lemma=«рука»%c  
Number of results: 35

Recording	Transcription
	Так , дай <b>руку</b> покажу́й .
	А теперь вёд , и говорю , до того износилось , вот как это работает <b>руки</b> ... < смеется >
	Тоже так , показалася и <b>руки</b> отворялись .
	Без <b>руки</b> осталася , вёд рано пил .
	Под <b>руку</b> тебе чёго , чего любишь ... И еши ...
	Вот на эту , на скамью , им говорим , право чтоб сплыть , и право сразу наберёз , и пот право <b>рукой</b> поставили , и вылезли , блево .
	Вот , потом там выражения тупебеко - членов пот , лежит , прям вот так <b>руки</b> .

201008e\_mgd

[Слушать в одном файле \(mp3\)](#)

First transcription: 0:0:0 Last transcription: 0:18:10

**Interviewer:** <>! ↴  
**mgd:** <Рисовала>... Русалка.! ↴  
**Interviewer:** Ага.! ↴ А масленицу делали у вас?! ↴  
**mgd:** А?! ↴  
**Interviewer:** А на масленицу не было? ↴  
**mgd:** О, а как же! И масленица бывает.! ↴ Тогда , бывало , катались с горы,! ↴  
**Interviewer:** С горы? ↴  
**mgd:** Бывалочки - кворовали лошадиные сани , это сейчас лошадей-то уже нету,! ↴ тогда были лошадиные сани , унесут где-нибудь , и с горы катаются до самой реки . Вон сидят вот оттуда - пошёл!! ↴  
**Interviewer:** Ага! ↴  
**mgd:** До самой реки едут! ↴  
**Interviewer:** На санях на этих , да?! ↴  
**mgd:** Да , на этих санях катались.! ↴  
**Interviewer:** Ага.! ↴ А саму масленицу как-то тоже изображали куклой такой , нет?! ↴  
**mgd:** Да и - как изображали?! ↴

Malinino Corpus

Corpus ▾ Statistics ▾ About ▾ Help | rus en |



Corpus of the Russian dialect spoken in the village Malinino

Collection of Spoken Texts and Recordings

## About the project

The Malinino dialect corpus contains texts recorded in the village Malinino (Khlevenskiy rayon, Lipetsk Oblast) in July-August of 2010. The dialect is part of the interzonal group B, of the South Russian dialect group (following the classification of K.F. Zakharova and V.G. Orlova). It combines

# План доклада

Обо мне

Прескриптивная vs. дескриптивная лингвистика

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

Чем донской корпус отличается от остальных корпусов?

Заключение

## Идея исследования

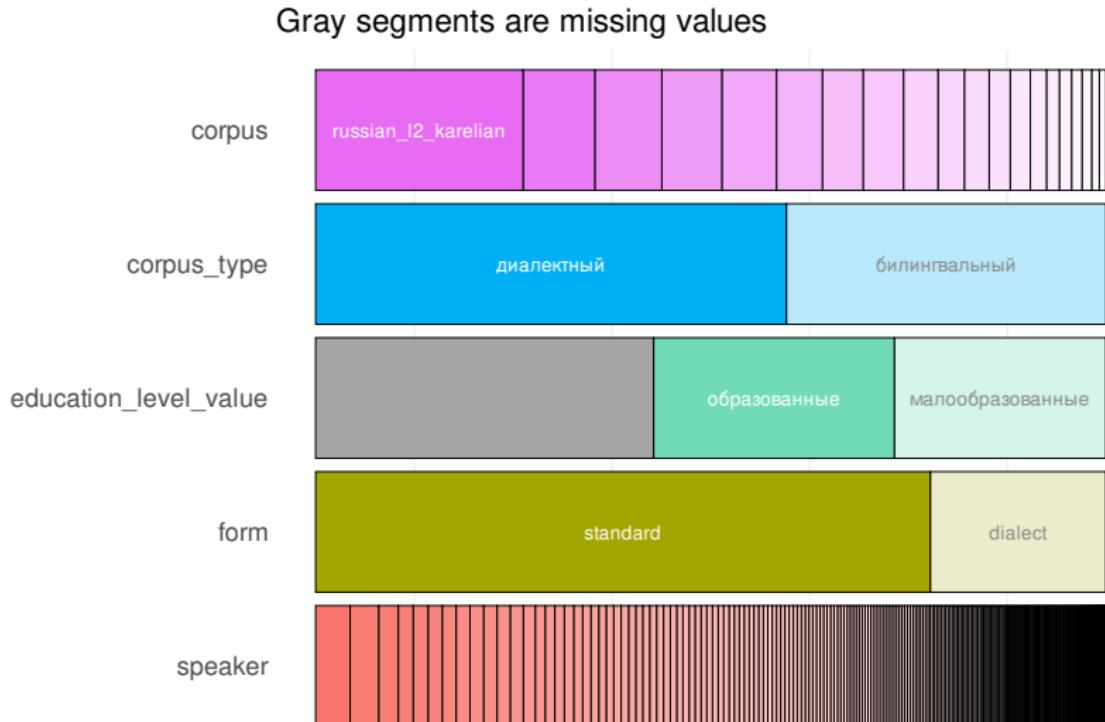
- работа по извлечению примеров и разметке примеров была сделана нашим постдоком Светланой Сергеевной Земичевой
- исследовать использование причастий и деепричастий в устной речи на материале диалектных и билингвальных корпусов
- исследовать соотношение литературных и диалектных форм (например, *евши*)
  - диалектные формы на *-ши* могут иметь разные функции, иногда они ведут себя как финитные формы:

*Пока делаешь, придёшь — он уже вставши.* (Макеево, 1953, f)

- С. С. Земичева изначально выделила три типа форм:
  - стандартные: *вышедший, сделанный, сделав*
  - диалектные: *вышедши, забывши*
  - промежуточные:  *выпивши, доённый*

# Данные: 3185 наблюдений

- 20 корпусов, 273 носителей



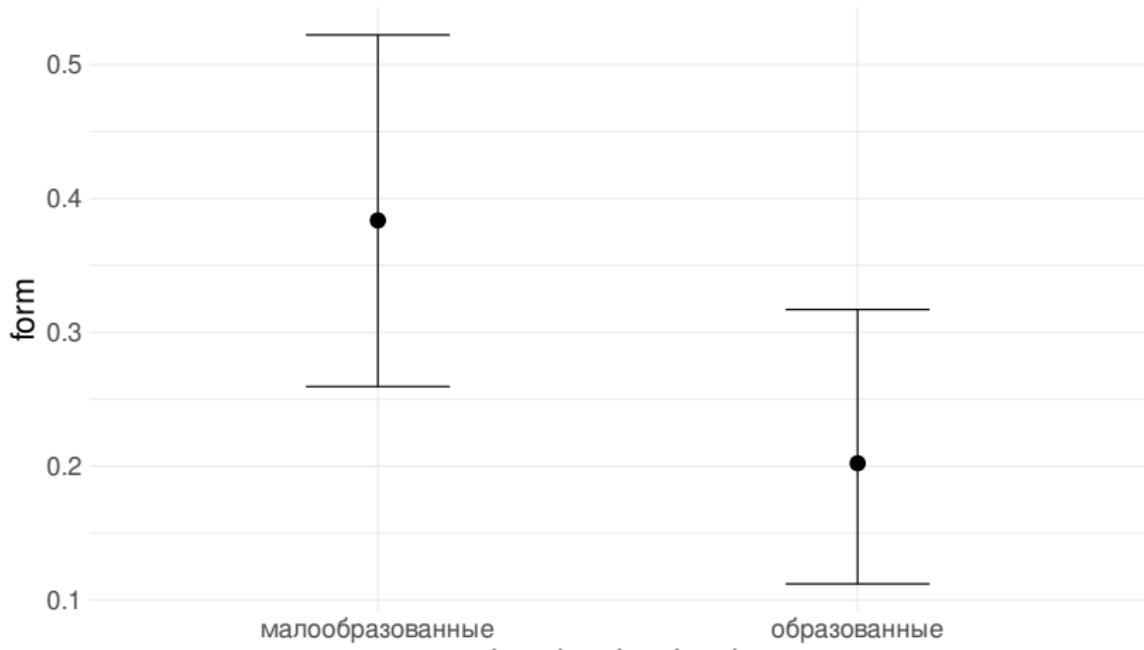
# Данные: географическое распределение



# Есть ли зависимость между образованием и формой?

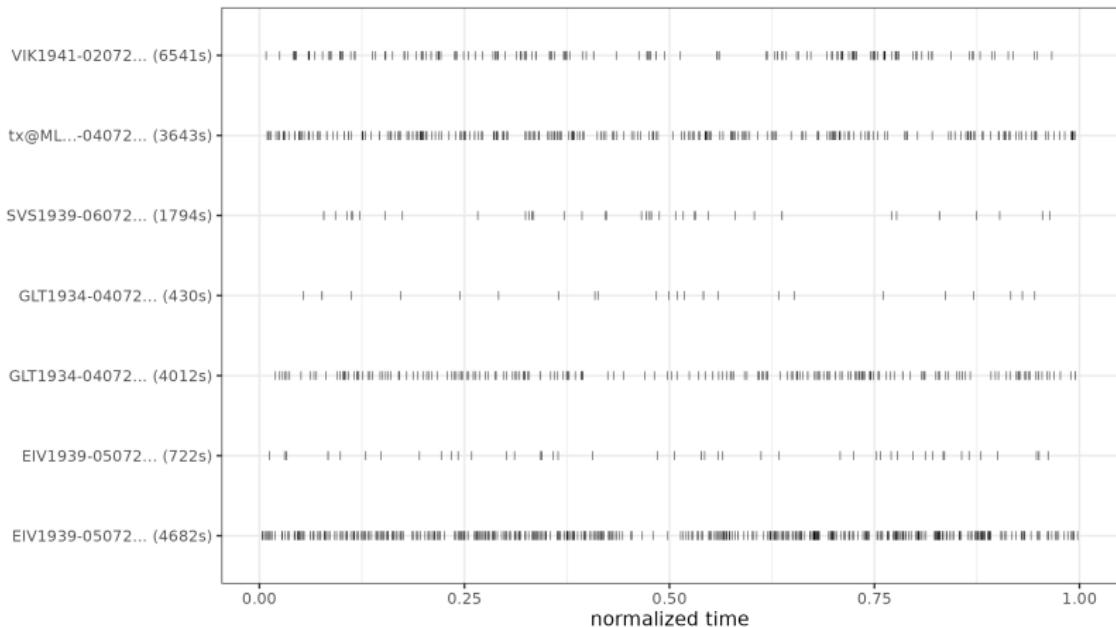
Предсказания байесовской логистической регрессии со смешанным эффектами (80% доверительный интервал):

$\text{form} \sim \text{education} + (\text{1} | \text{corpus/speaker})$



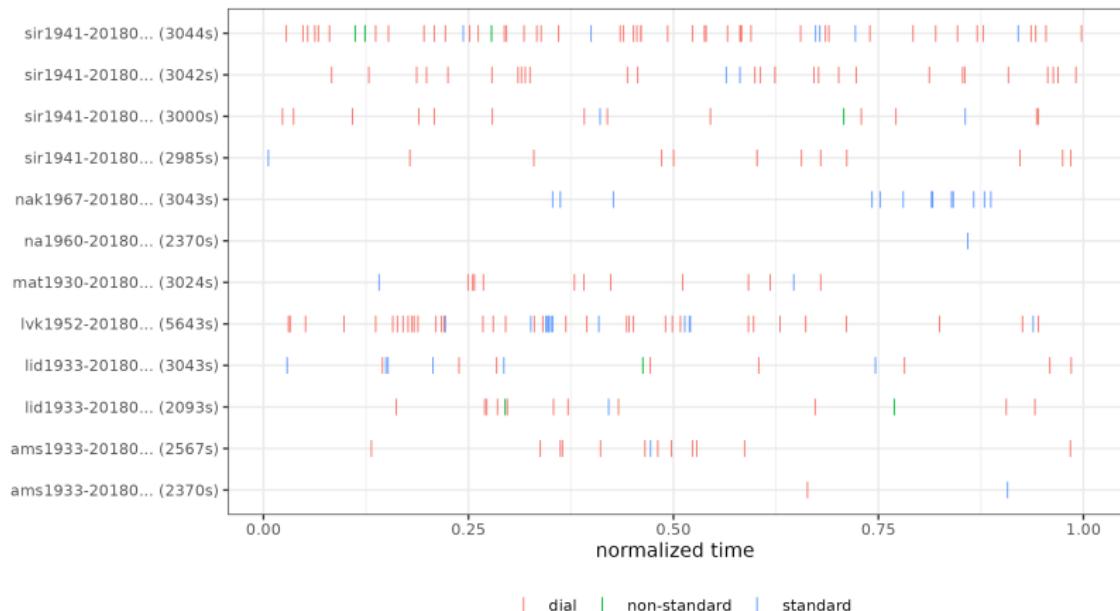
# А как эти формы расположены во времени?

Использование *вот* в донском корпусе:



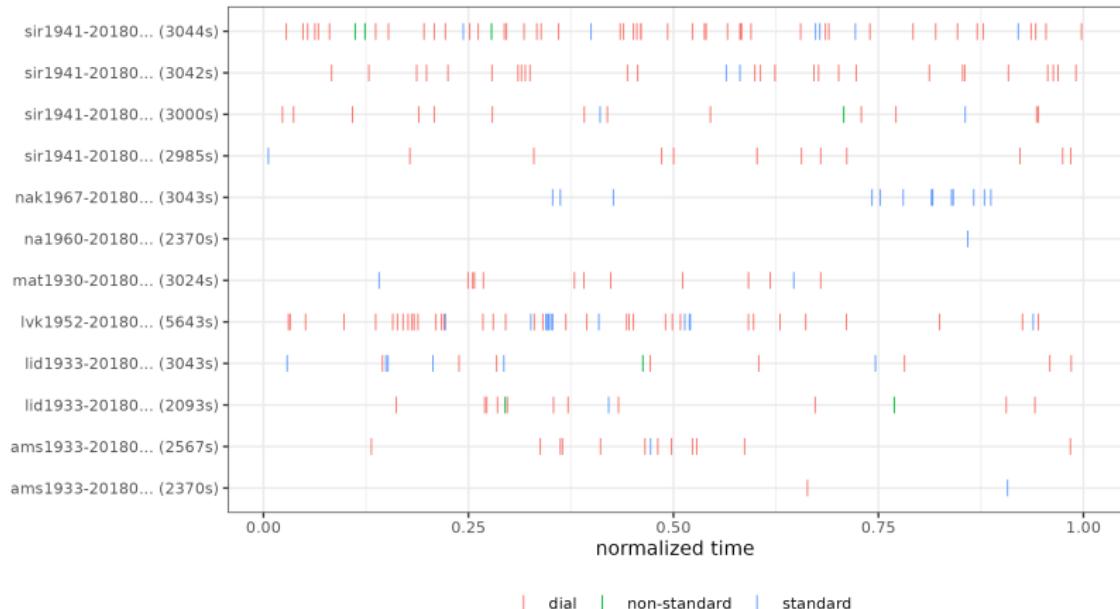
# А как эти формы расположены во времени?

Использование причастий/деепричастий в корпусе Лужникова (разметка С. С. Земичевой):



# А как эти формы расположены во времени?

Может быть можно попробовать смоделировать вероятность перехода от диалектной формы в недиалектную и наоборот?

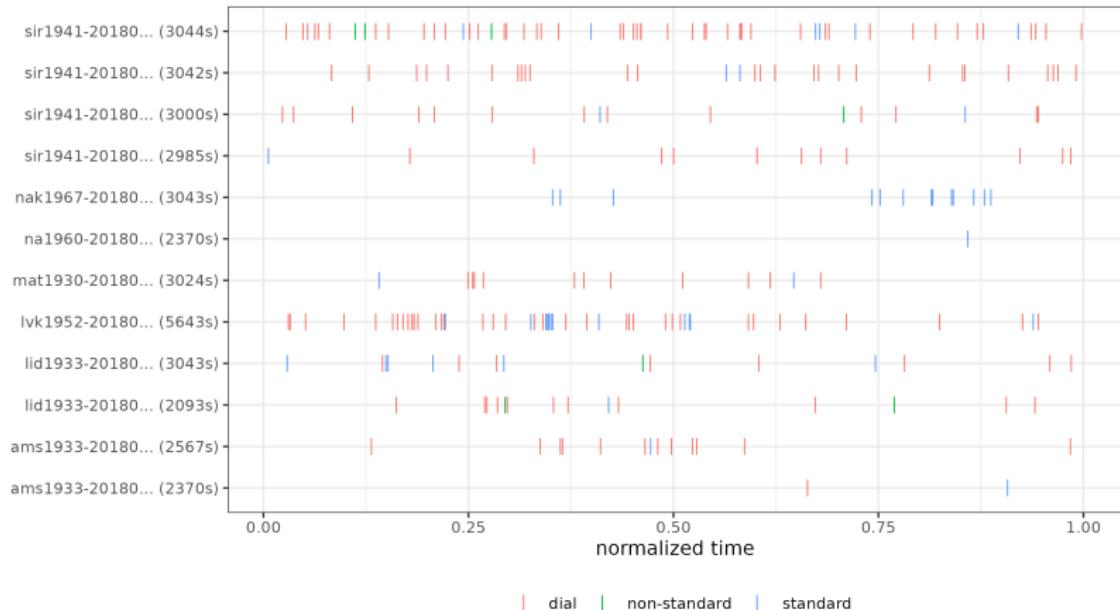


## Прайминг

Праймингом в лингвистике называют эффект, когда говорящие повторяют форму/вариант, которую была использована перед анализируемым речевым актом. Очень походит на эффект якоря или эффект привязки.

# Прайминг?

Посмотрите на lvk1952:



1vk1952: Это полоса от пожарных, это у нас тут тоже пропахано, это от пожаров.

Interviewer: А почему такой полосы не пропахано в Ситниково или вот в Лужникове?

1vk1952: Не пропахано? В Лужниках есть, вот от нас идешь - у складу удобрения пропахано. Да, ну а там тогда не знаю. Не, там с Ситникова идешь, вот как с Ситникова значит, на правой стороне-то, эво тут з мосту перейдешь, поднимешься, там тоже пропахано! Это от пожаров, это каждый год пропахивают у нас. Тут всё пропахано, там за деревней пропахано, там пропахано, Хорёво там, это всё опахано Да, это... чё, трава-то, загорится, ее же, вон у нас вот трава когда загорелась - один, два, три, четыре дома. Сразу сгорели.

# План доклада

Обо мне

Прескриптивная vs. дескриптивная лингвистика

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

Чем донской корпус отличается от остальных корпусов?

Заключение

# Марковский процесс

Марковская цепь — это одна из популярных семей стохастических процессов, которая описывает переход из разных состояний. Их часто представляют в виде графа и матрицы переходов:



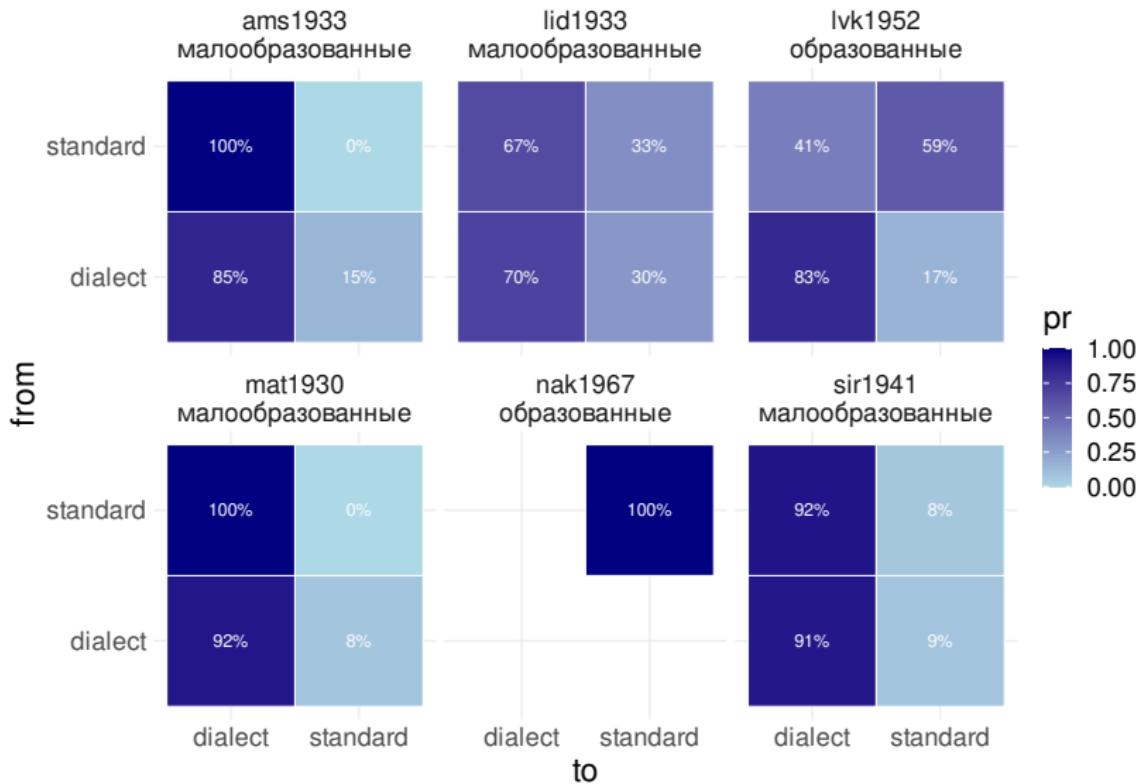
- значения в строчках должны суммироваться до 1

## Марковский процесс на примере одного носителя

Возьмем носителя sir1941 и проанализируем ее 139 форм:

```
## MLE Fit
## A 2 - dimensional discrete Markov Chain defined by the following states:
## dialect, standard
## The transition matrix (by rows) is defined as follows:
##          dialect   standard
## dialect  0.9120000 0.08800000
## standard 0.9230769 0.07692308
```

# Для каждого носителя из корпуса Лужниково



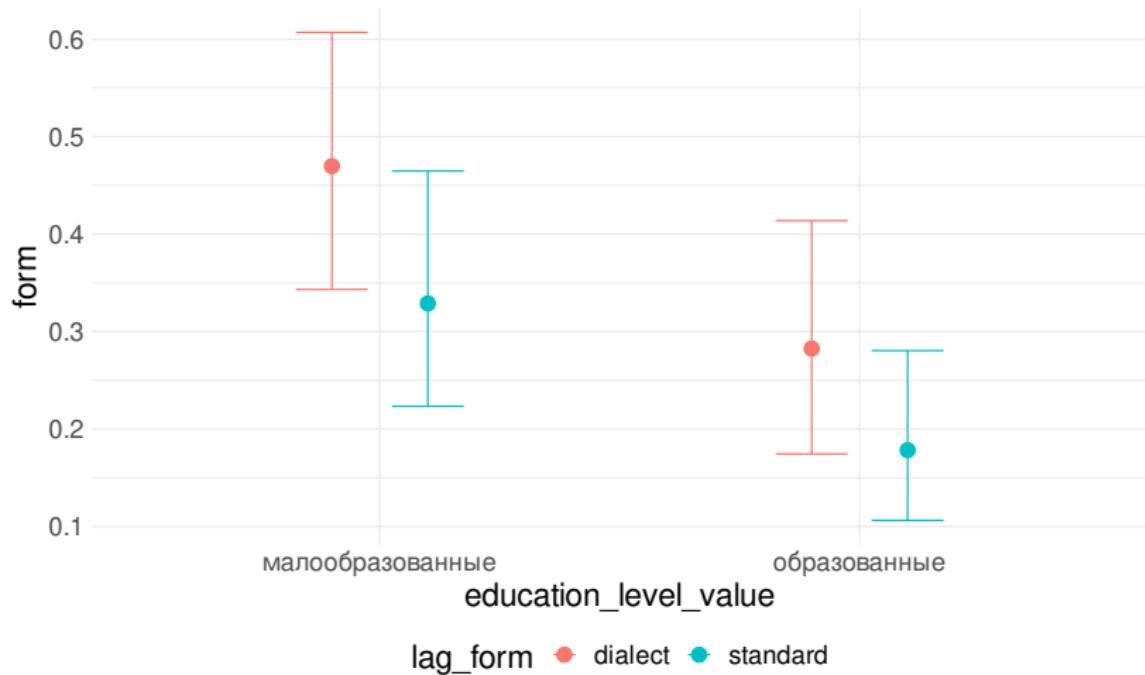
## Обобщение по всем корпусам?

- Хотелось бы чтобы можно было делать иерархические марковские цепи, аналогичные нашей регрессии (носитель, вложен в корпус).
- Кроме того, хотелось бы делать поправку на количество единиц для анализа.
- Вроде это должно покрываться иерархическими марковскими моделями, но я не нашел их реализацию, которую бы мне подходила, так что я заменил все регрессией

# Обобщение по всем корпусам?

Предсказания байесовской логистической регрессии со смешанным эффектами (80% доверительный интервал):

$\text{form} \sim \text{education} + \text{previous form} + (1 | \text{corpus}/\text{speaker})$



# План доклада

Обо мне

Прескриптивная vs. дескриптивная лингвистика

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

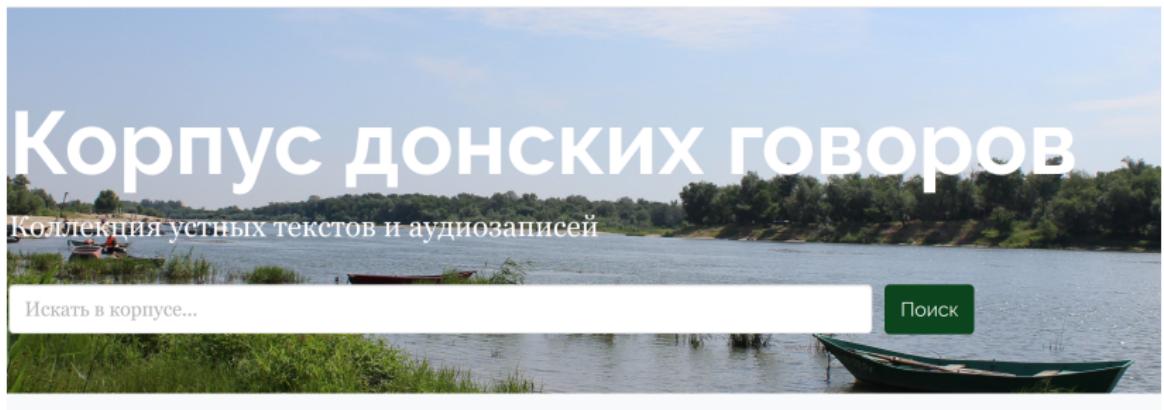
Марковский процесс

Чем донской корпус отличается от остальных корпусов?

Заключение

Корпус донских говоров

Корпус ▾ Статистика ▾ О проекте ▾ Помощь [ rus en ]



Коллекция устных текстов и аудиозаписей

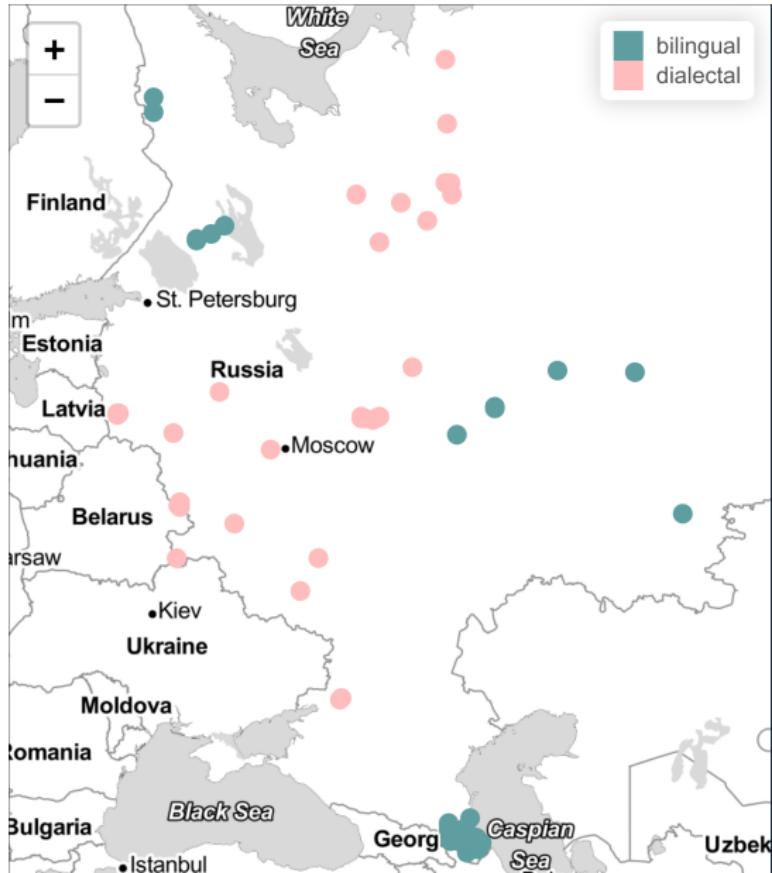
Искать в корпусе...

Поиск

[http://lingconlab.ru/don\\_rnd](http://lingconlab.ru/don_rnd)

Флягина М.В., Калиничева Н.В., Северина Е.М. Корпус донских диалектов. 2022–2023. Москва: Международная лаборатория языковой конвергенции, НИУ ВШЭ. Электронный ресурс:  
[http://lingconlab.ru/don\\_rnd](http://lingconlab.ru/don_rnd), дата обращения об.04.2023.

# Давайте сравним с остальными 17-ю диалектными корпусами



## Мера TF-IDF

TF-IDF — это мера, которую используют для оценки важности слова в контексте документа.

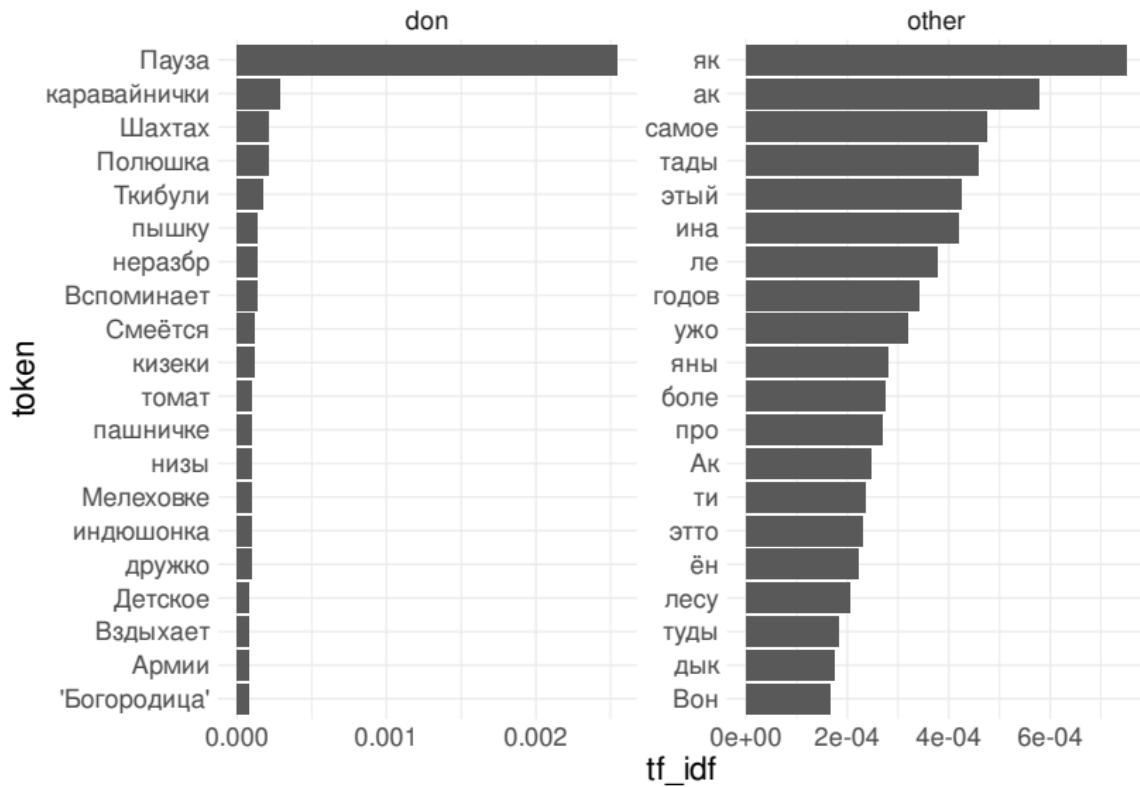
TF (term frequency) — частотность языковой единицы.

Например, слово встречается в тексте длиной  $m$  слов  $n$  раз, тогда его частотность это  $\frac{n}{m}$ .

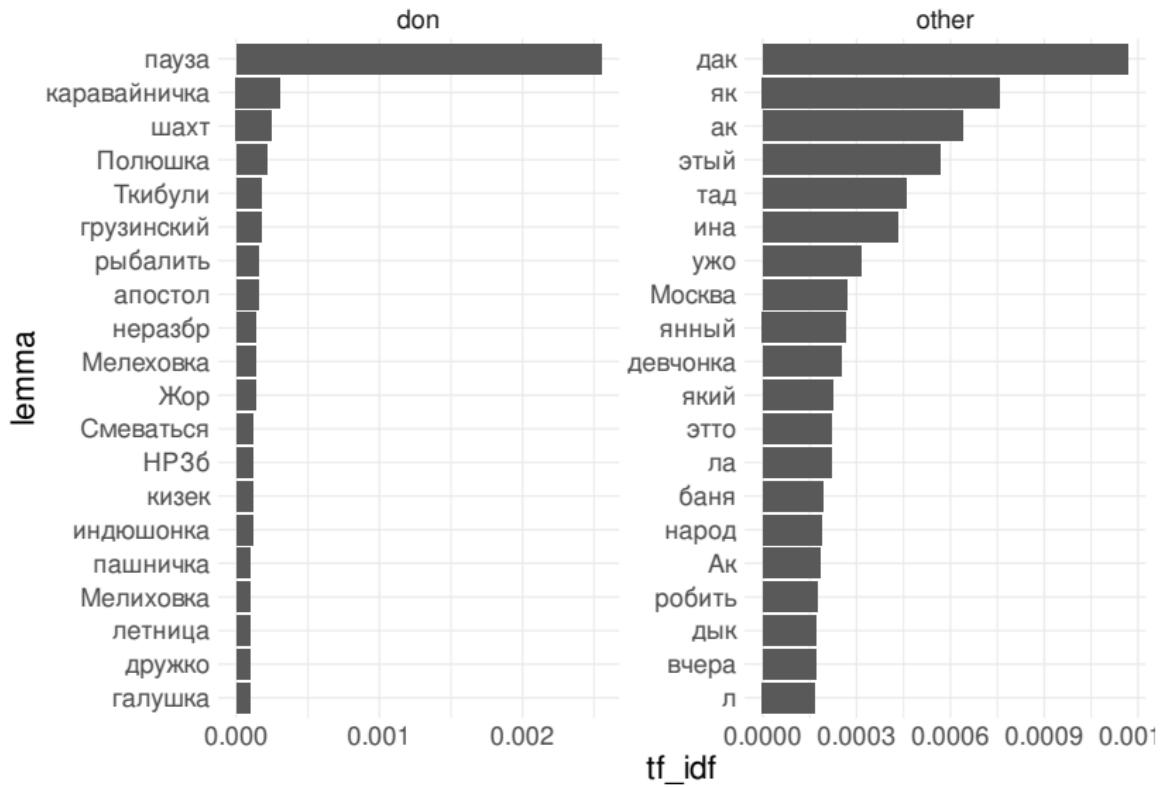
IDF (inverse document frequency) — инверсия частоты, с которой некоторое слово встречается в документах коллекции.

Например, если языковая единица встречается в  $j$  документах из коллекции в  $k$  документов, тогда его обратная частота будет равна  $\log_{10} \frac{j}{k}$ .

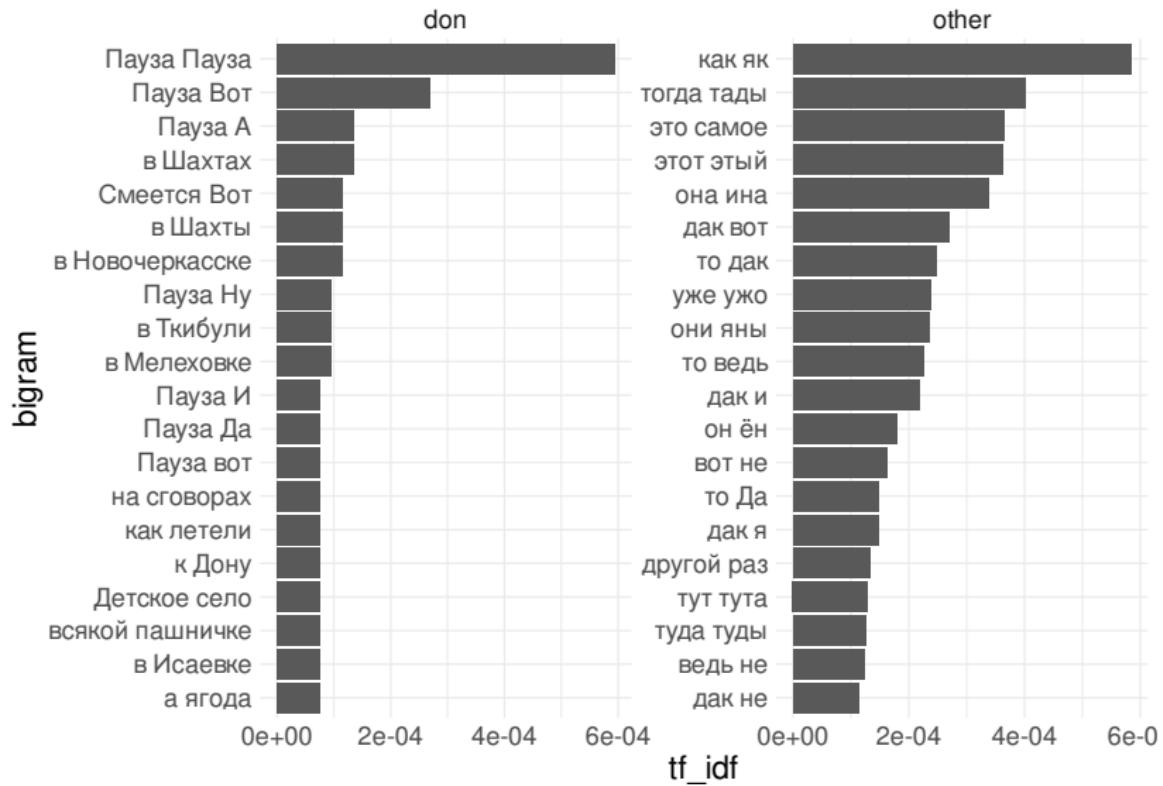
## Давайте сравним с остальными 17-ю диалектными корпусами



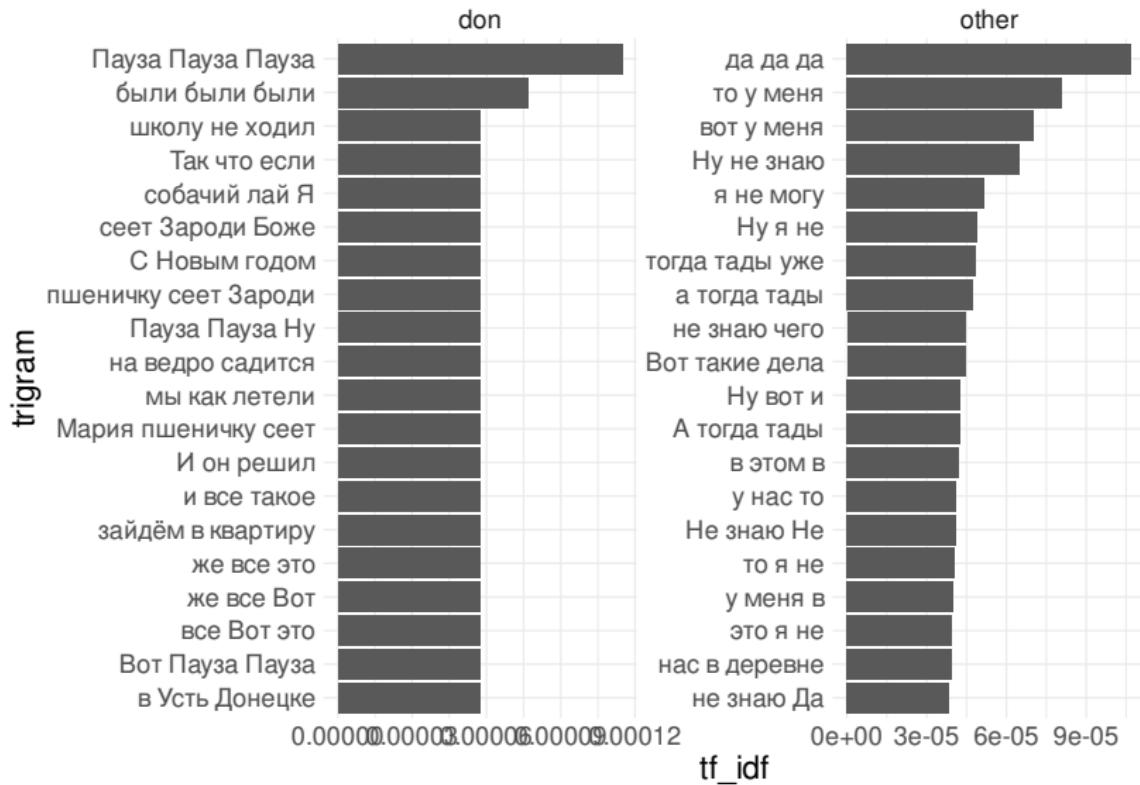
# Давайте сравним с остальными 17-ю диалектными корпусами



# Давайте сравним с остальными 17-ю диалектными корпусами



# Давайте сравним с остальными 17-ю диалектными корпусами



# План доклада

Обо мне

Прескриптивная vs. дескриптивная лингвистика

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

Чем донской корпус отличается от остальных корпусов?

**Заключение**

## Мы с вами посмотрели

- на диахроническое исследование на больших корпусах  
*(отложить в ... ящик)*

## Мы с вами посмотрели

- на диахроническое исследование на больших корпусах  
(*отложить в ... язык*)
- на слепое чтение на основе ограниченного корпуса  
(Известия ЮФУ)

## Мы с вами посмотрели

- на диахроническое исследование на больших корпусах  
(*отложить в ... язык*)
- на слепое чтение на основе ограниченного корпуса  
(Известия ЮФУ)
- на межкорпусное лингвистическое исследование  
(использование причастий и деепричастий)

## Мы с вами посмотрели

- на диахроническое исследование на больших корпусах  
*(отложить в ... язык)*
- на слепое чтение на основе ограниченного корпуса  
*(Известия ЮФУ)*
- на межкорпусное лингвистическое исследование  
*(использование причастий и деепричастий)*
- контрастивное исследование (на примере корпуса  
донских диалектов)

**Спасибо за внимание!**