

Использование марковских цепей при анализе вариативности в русских диалектных корпусах

Георгий Алексеевич Мороз, Светлана Сергеевна Земичева

Международная лаборатория языковой конвергенции (НИУ ВШЭ)

«Основные приложения математики», НИУ ВШЭ

презентация доступна по ссылке: tinyurl.com/202aha9q

23 марта 2023

План доклада

Лингвистика: мифы и реальность

Обо мне

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

- умеет читать на всех письменностях мира

- умеет читать на всех письменностях мира
- знает все языки на свете

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- пишет без ошибок и знает все правила орфографии

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- пишет без ошибок и знает все правила орфографии
- не знает математики и программирования

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- пишет без ошибок и знает все правила орфографии
- не знает математики и программирования
- все вышеперечисленное, конечно, неправда

Лингвистика

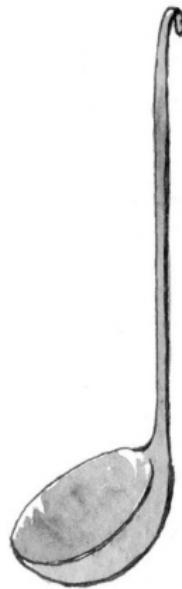
- прескриптивная

Лингвистика

- прескриптивная
- вся остальная (дескриптивная)
 - каталогизация языкового разнообразия, описание языковых контактов
 - исследования грамматики языка
 - исследования распределения грамматических особенностей в языках мира
 - исследования когнитивных способностей человека и других животных, связанных с языком
 - исследования в области синтеза и распознавания речи и языка
 - создание компьютерных инструментов для решения самых разных задач
 - вспомогательные инструменты лингвистического исследования и документации
 - корпусная лингвистика
 - исследования в области NLP, языковых моделей и т. п.
 - симуляционные модели в лингвистике

Прескриптивная vs. дескриптивная лингвистика

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



Прескриптивная vs. дескриптивная лингвистика

Это часть опроса Ивана Левина:



План доклада

Лингвистика: мифы и реальность

Обо мне

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

Обо мне

- полевые исследования (29 поездок)
- фонетист, фонолог, квантитативный лингвист
- езжу на Кавказ
- преподаю статистику и R (язык программирования)
- написал несколько лингвистических пакетов для R
 - `lingtypology`
 - `phonfieldwork`
 - `lingglosses`

План доклада

Лингвистика: мифы и реальность

Обо мне

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

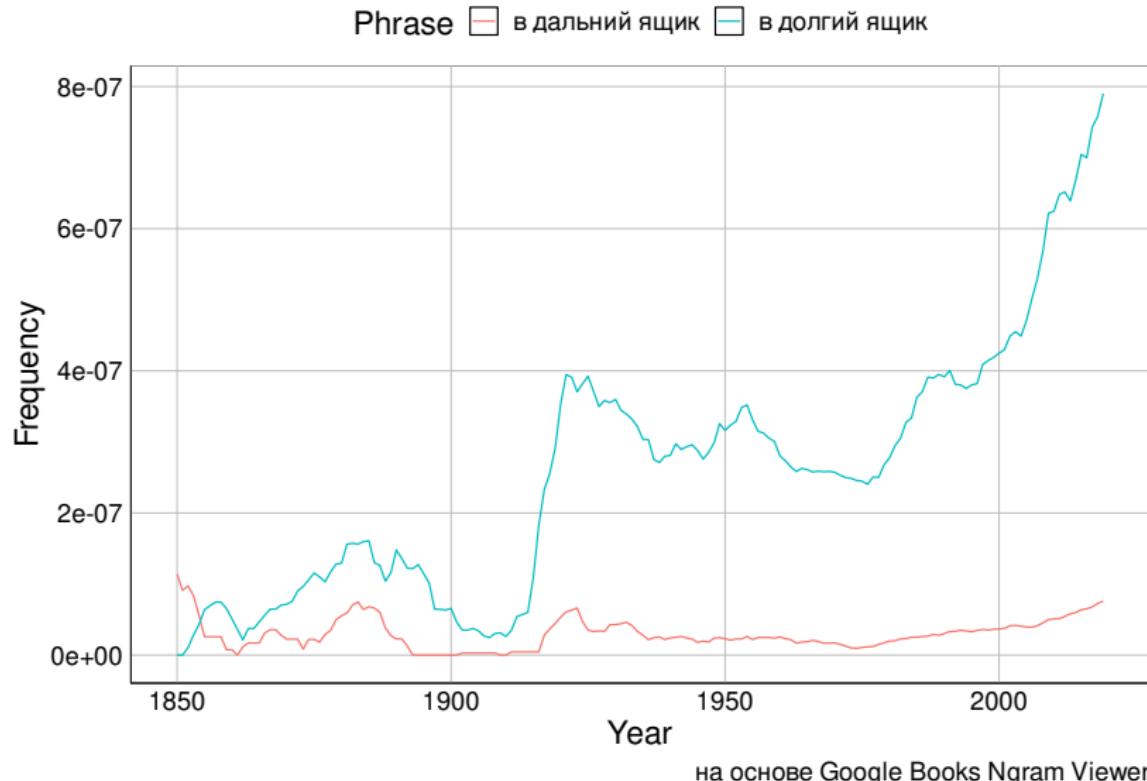
Корпусная лингвистика

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.

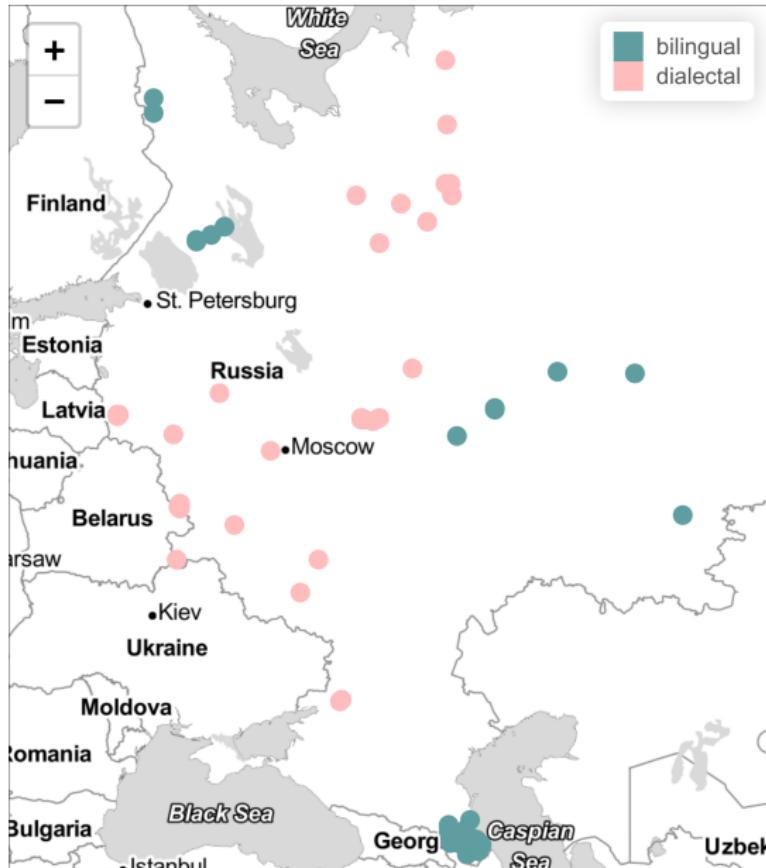
Среди корпусов русского языка можно назвать:

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

Отложить в ... ящик



Диалектные устные корпуса лаборатории языковой конвергенции



Диалектные устные корпуса лаборатории языковой конвергенции

Malinino Corpus

Corpus ▾ Statistics ▾ About ▾ Help | rus en |

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

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

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 - кировские 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

Search in the corpus...

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

Лингвистика: мифы и реальность

Обо мне

Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

Идея исследования

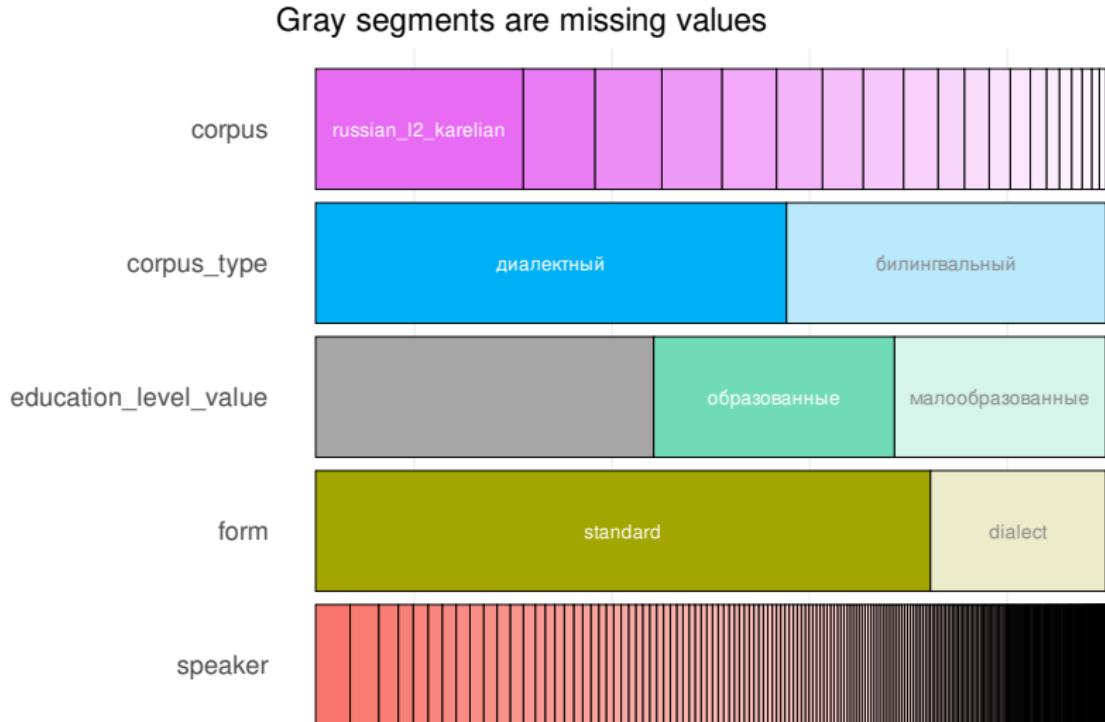
- работа по извлечению примеров и разметке примеров была сделана нашим постдоком Светланой Сергеевной Земичевой
- исследовать использование причастий и деепричастий в устной речи на материале диалектных и билингвальных корпусов
- исследовать соотношение литературных и диалектных форм (например, *евши*)
 - 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



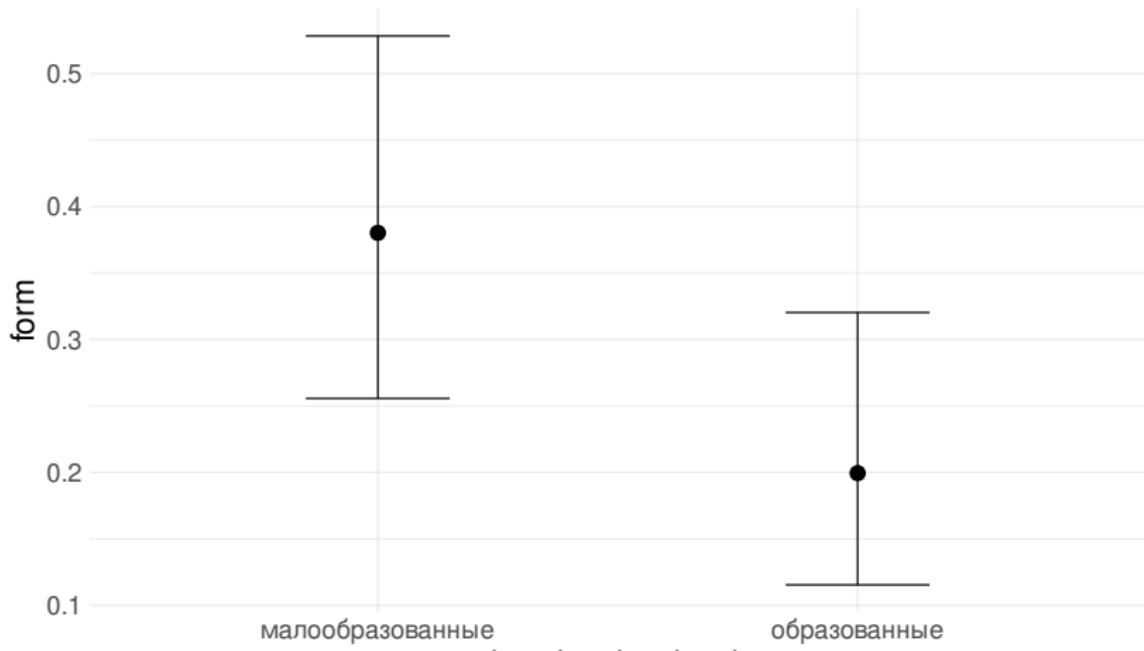
Данные: географическое распределение



Есть ли зависимость между образованием и формой?

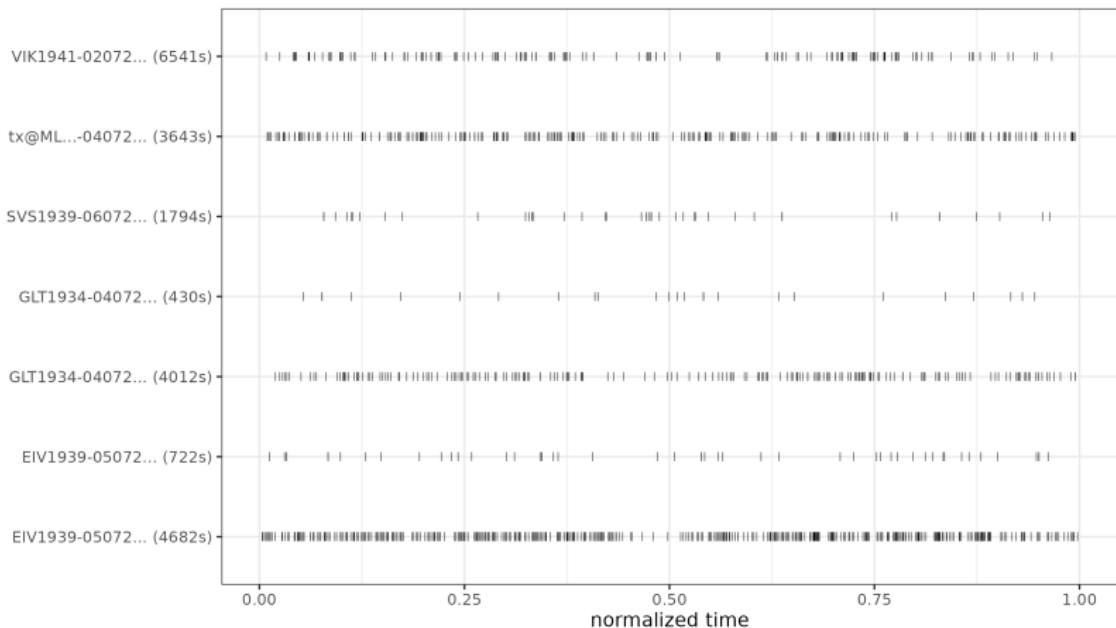
Предсказания байесовской логистической регрессии со смешанным эффектами (80% доверительный интервал):

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



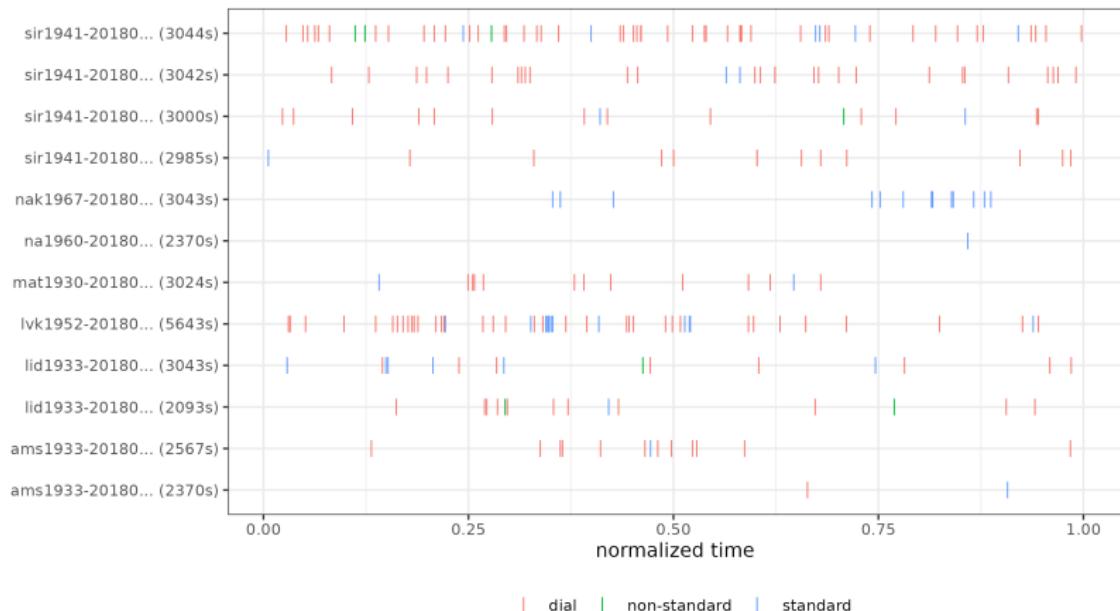
А как эти формы расположены во времени?

Использование *вот* в донском корпусе:



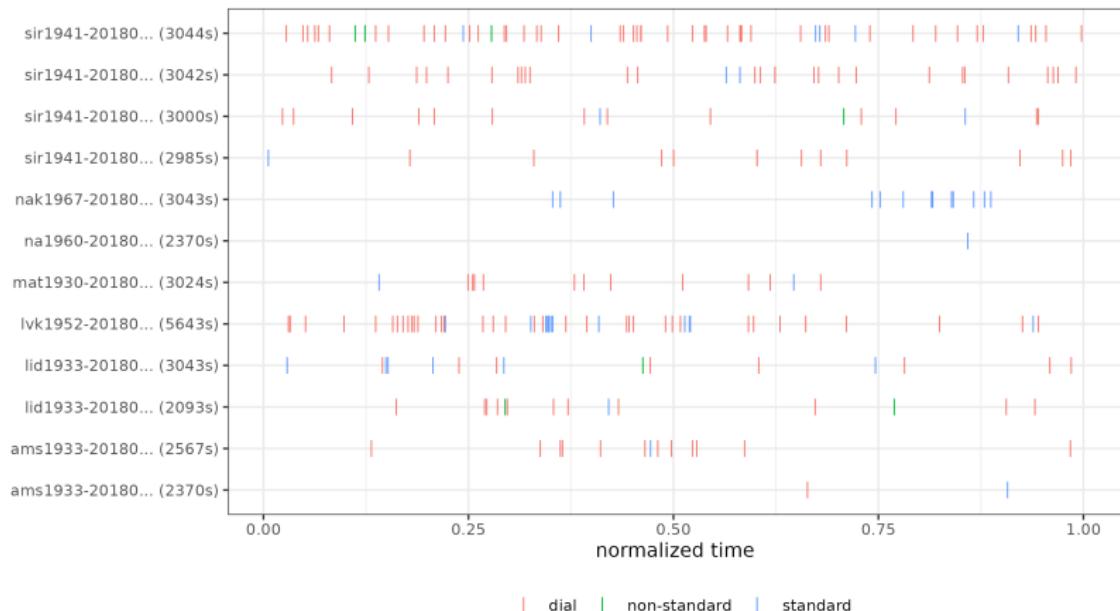
А как эти формы расположены во времени?

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?

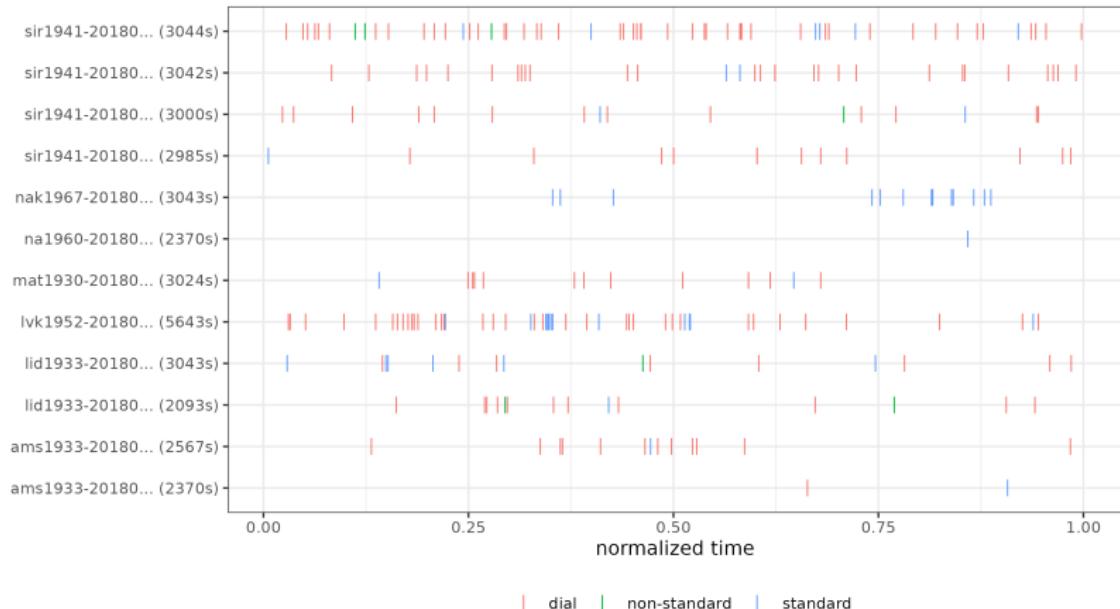


Прайминг

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

Лингвистика: мифы и реальность

Обо мне

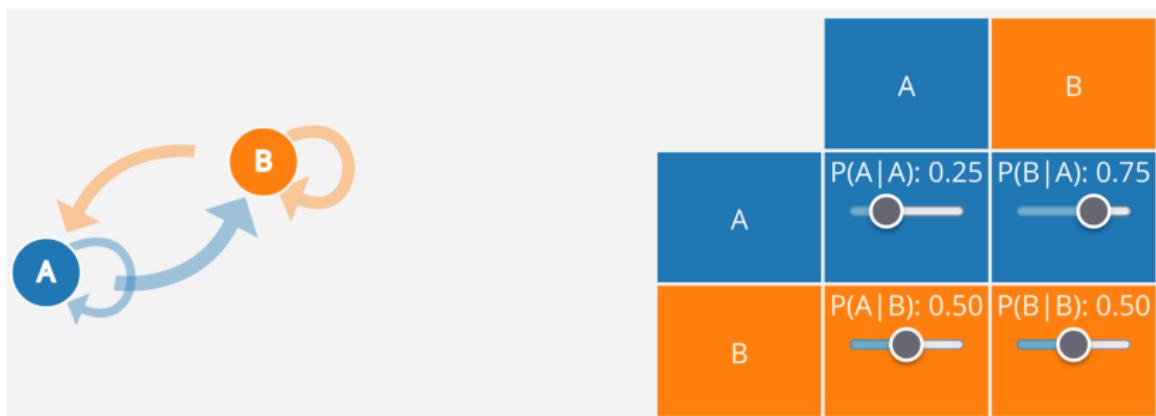
Корпусная лингвистика

Использование причастий и деепричастий в устной речи

Марковский процесс

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:



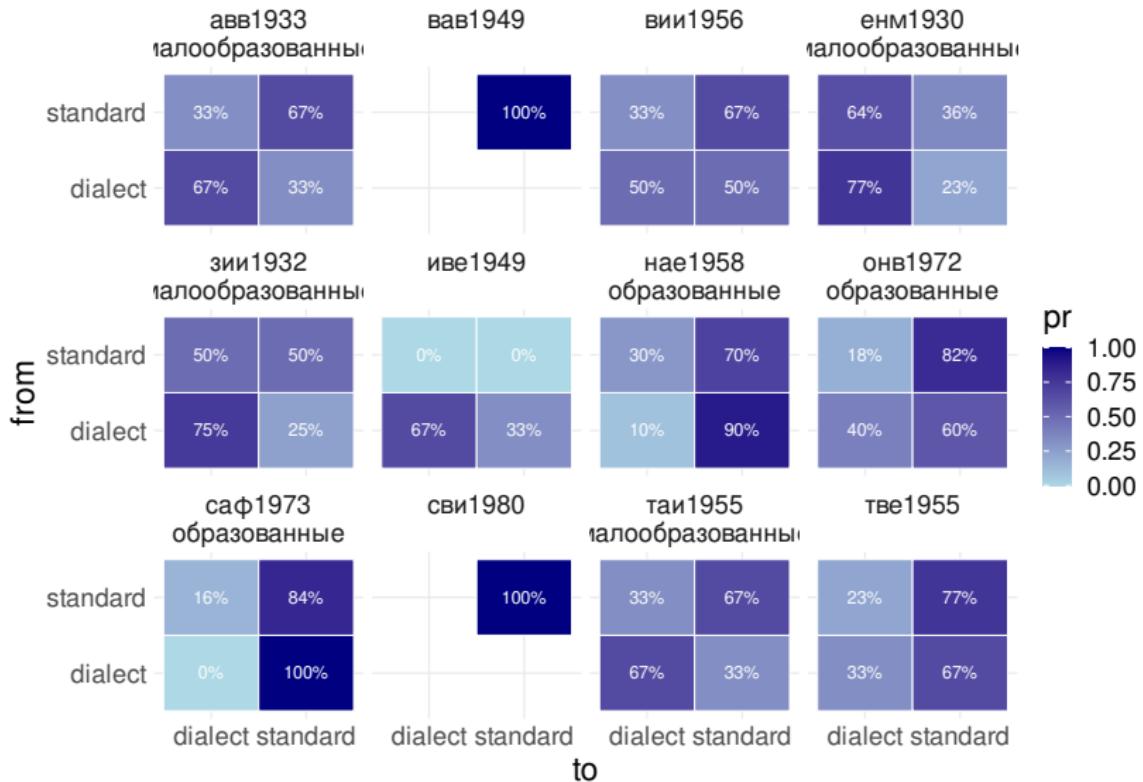
- значения в строчках должны суммироваться до 1

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



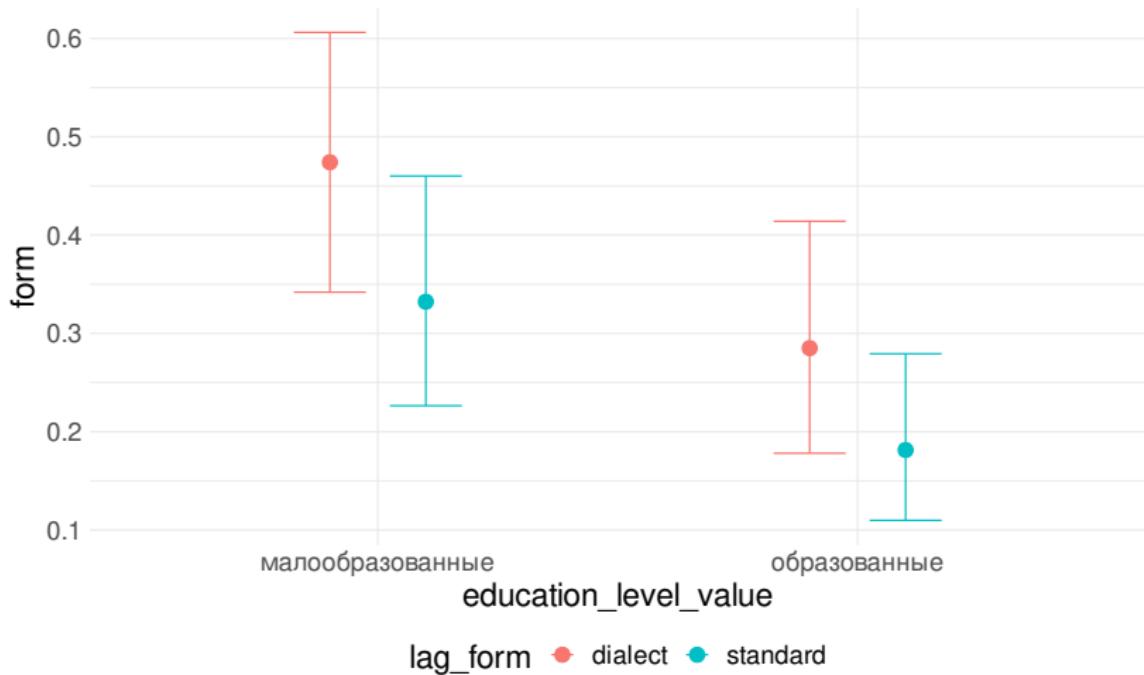
Обобщение по всем корпусам?

- 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):

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



Выводы

- Причастия и деепричастия встречаются в устной речи чаще, чем мы думаем
- Вариативность всюду
- Их можно пытаться моделировать стохастическими моделями разной сложности
- Носители диалектных форм очень разнородны, может быть и не надо пытаться строить какую-то обобщающую модель?

Спасибо за внимание!