

Анализ вариативности в билингвальных корпусах русского языка: числительные и выпадение предлогов

Г. А. Мороз

Международная лаборатория языковой конвергенции (НИУ ВШЭ,
Москва)

«Студенческая конференция ИЛ РГГУ»

28–29 октября 2023

План доклада

Ресурсы международной лаборатории языковой
конвергенции

Нестандартные количественные конструкции в речи
билингвов

Выпадение предлогов в речи билингвов

Ресурсы международной лаборатории языковой конвергенции

- lingconlab.ru
- 22 устных диалектных корпуса
- 8 устных билингвальных корпусов
- 10 корпусов малых языков
- другие
 - словари (мегебский, рутульский, тукитинский, хваршинский, даргинский)
 - Типологический атлас языков Дагестана
 - Атлас многоязычия в Дагестане
 - Атлас рутульских диалектов
 - Корпус Просодии Русских Диалектов (ПРУД)
 - ...

22 устных диалектных корпуса

Корпус говора Хиславичского района
260,793 ток.

Корпус говора села Спиридонова Буда
70,565 ток.

Корпус говора села Кеба
54,535 ток.

Корпус говора села Церковное
39,469 ток.

Корпус устьянских говоров
959,782 ток.

Корпус говора верхней Пиннели и Валдайской возвышенности
70,803 ток.

Корпус говоров среднего течения Оки и Клязьмы
68,010 ток.

Корпус говоров Средней Пинеги
43,070 ток.

Корпус донских говоров
71,600 ток.

Корпус опочечских говоров
68,741 ток.

Лужниковский корпус
68,666 ток.

Корпус говора города Заинск
68,324 ток.

Корпус говора деревни Веегора
91,514 ток.

Корпус говора деревни Нехочина
88,965 ток.

Корпус говора Средней Пинеги
79,566 ток.

Корпус говора Мантуровского района Костромской области
113,837 ток.

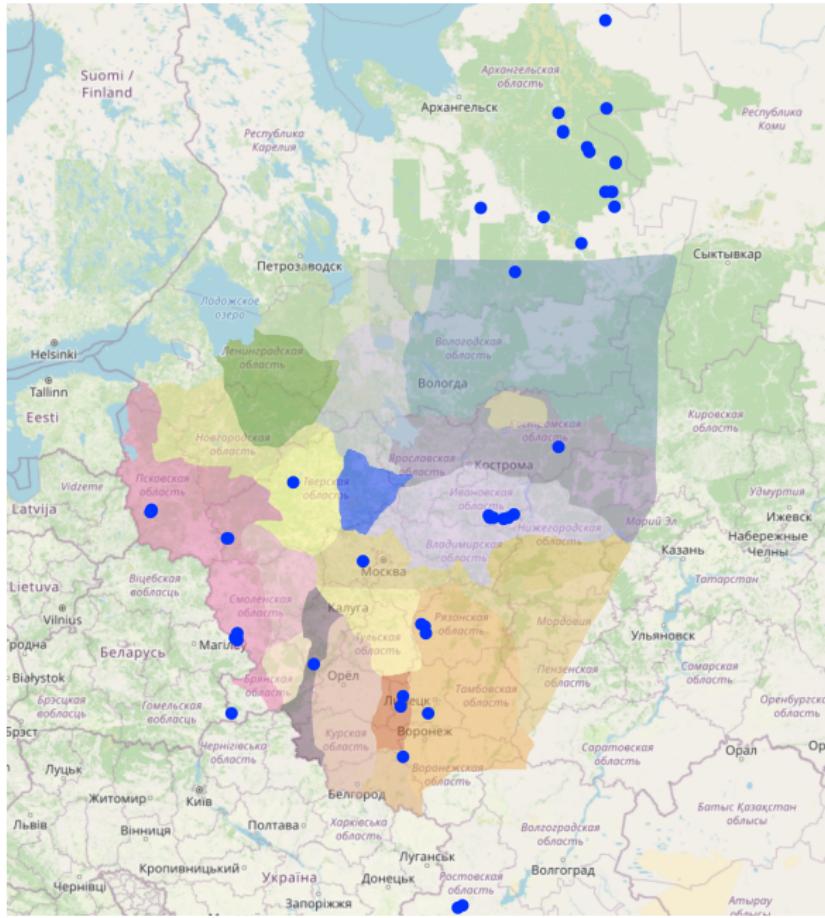
Корпус говора села Роговатка
100,047 ток.

Корпус говора деревни Шаптаково и Манево
65,336 ток.

Корпус говоров низовья рек Лух и Теза
146,350 ток.

Корпус говора села Малинино
138,943 ток.

22 устных диалектных корпуса



8 устных билингвальных корпусов

Корпус дагестанского русского
376,717 ток.

Якутско-русский корпус переключения кода
15,139 ток.

Корпус русской речи Чувашии
46,307 ток.

Корпус чеченского русского
41,767 ток.

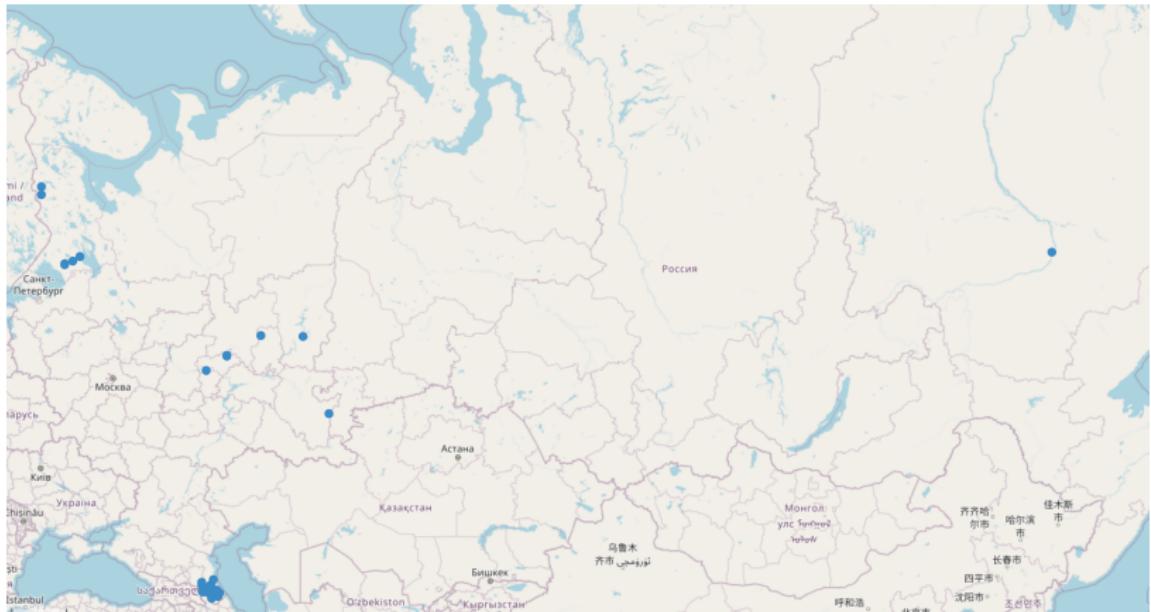
Корпус русской речи Карелии
578,646 ток.

Корпус русской речи Республики Марий Эл
69,109 ток.

Корпус русской речи Башкирии
93,127 ток.

Корпус русской речи бесермян
97,216 ток.

8 устных билингвальных корпусов

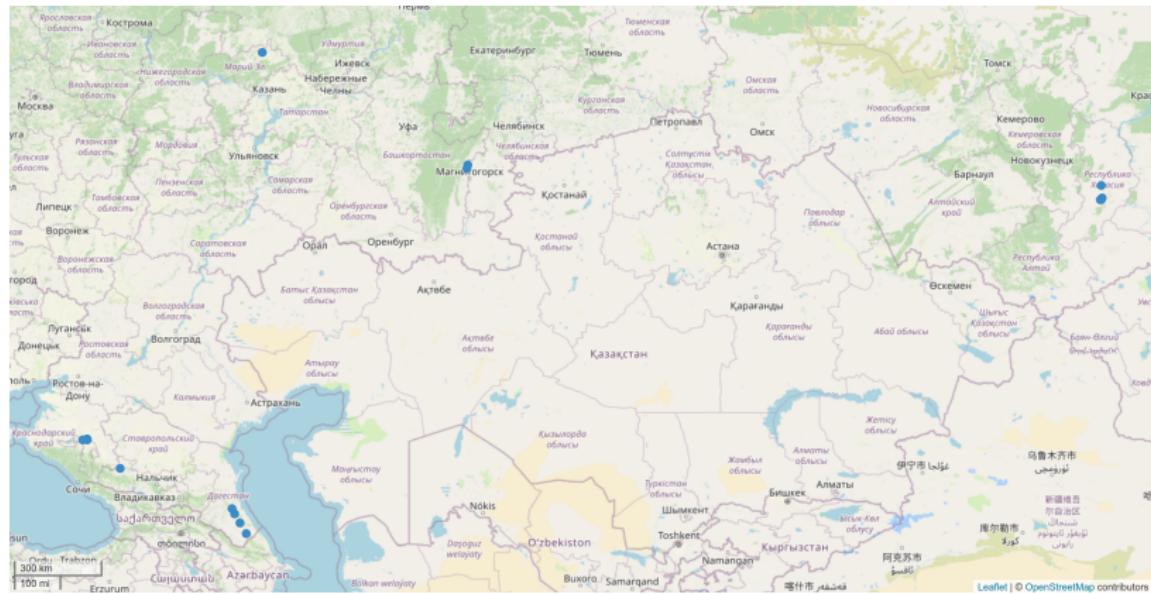


10 корпусов малых языков

Корпус литературного даргинского
703,988 ток.

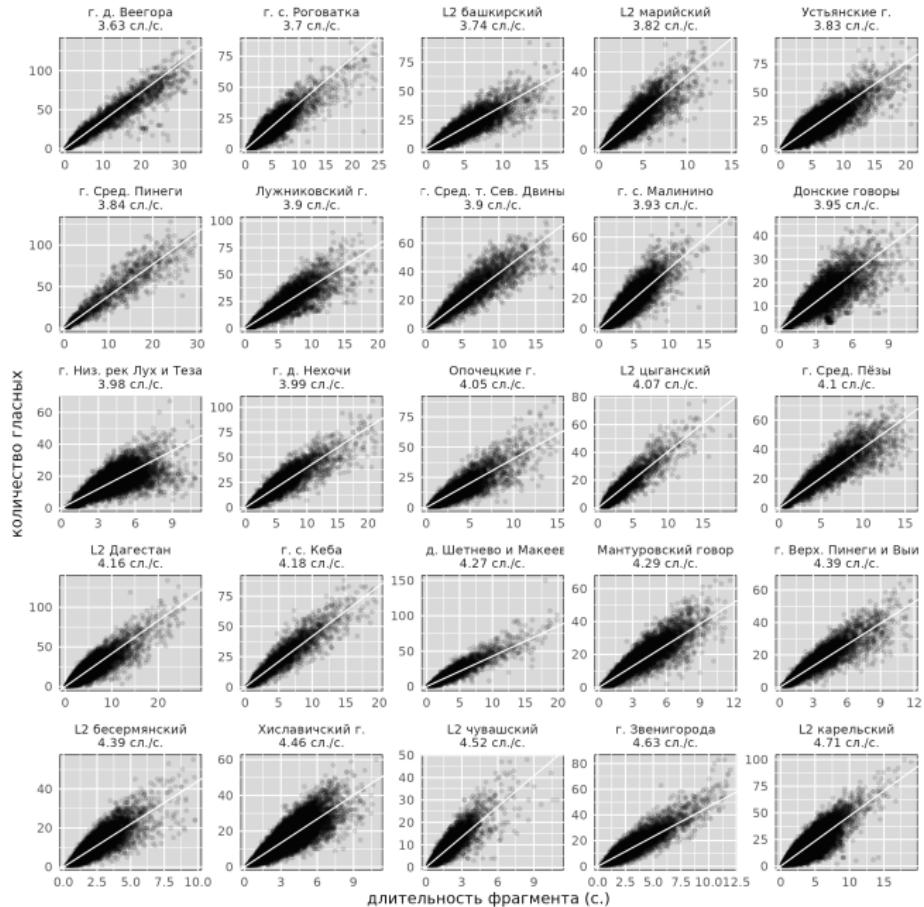
Истный корпус лугового марийского языка 11,647 ток.	
Устный корпус кадарского даргинского 12,654 ток.	
Устный корпус башкирского языка 25,000 ток.	
Истный корпус диалектов хакасского языка 59,000 ток.	

10 корпусов малых языков



Что делать со всеми этими корпусами?

Скорость речи (Мороз 2023: 382)



Корпус просодии русских диалектов (Князев et al. 2024)

PRuD главная о проекте информация о говорах ритмическая структура карта данные



intonation_type	transcription	transcription2	image	speaker	idiom	village	district	region	со
All	со м	All	All	All	All	All	All	All	All
Незавершённость	Дети были со мною маленькие...	H*!H*!H*H%	?	ЕАЖ1935	Говор_2	Роговатое	Старооскольский	Белгородская	
Утверждение	Пошли на солому.	%L L+H* L* L%	?	?	Говор_3	Клиновское	Коношский	Архангельская	
Незавершённость; завершённость	Беру вот такую чесовочку льну - несу домой.	L+H* H% %L L+H* L* L%	?	?	Говор_3	Клиновское	Коношский	Архангельская	
Незавершённость	Само по себе дерево посохло...	L*+H H- H%	?	МАН1910	Говор_51	Мосеево	Мезенский	Архангельская	
Утверждение	Сортируем лес.	L* L%	?	РИИ1915	Говор_55	Веегора	Пинежский	Архангельская	
Утверждение, взкий фокус	Время семь часов!	H* L%	?	РУК	Говор_55	Веегора	Пинежский	Архангельская	

Проект Dial2



М. В. Ермолова, С. С. Земичева, Н. А. Кошелюк



Г. А. Мороз, К. Наккарато, А. В. Яковлева

Проект Dial2

- фокус на данных русского языка в диалектных и билингвальных корпусах
- документация и моделирование вариативности в нестандартных вариантах русского языка

План доклада

Ресурсы международной лаборатории языковой
конвергенции

**Нестандартные количественные конструкции в речи
билингвов**

Выпадение предлогов в речи билингвов

Почему мы ожидаем вариативность в числительных?

- система русских числительных сложная
- системы числительных в L1 доступных нам корпусах значительно проще
- количественные конструкции в речи билингвов исследовалась в работах ([Stoyanova 2019; Стойнова 2021](#))
- В работе ([Стойнова 2021](#)) употребление нестандартных конструкций объясняется контактом

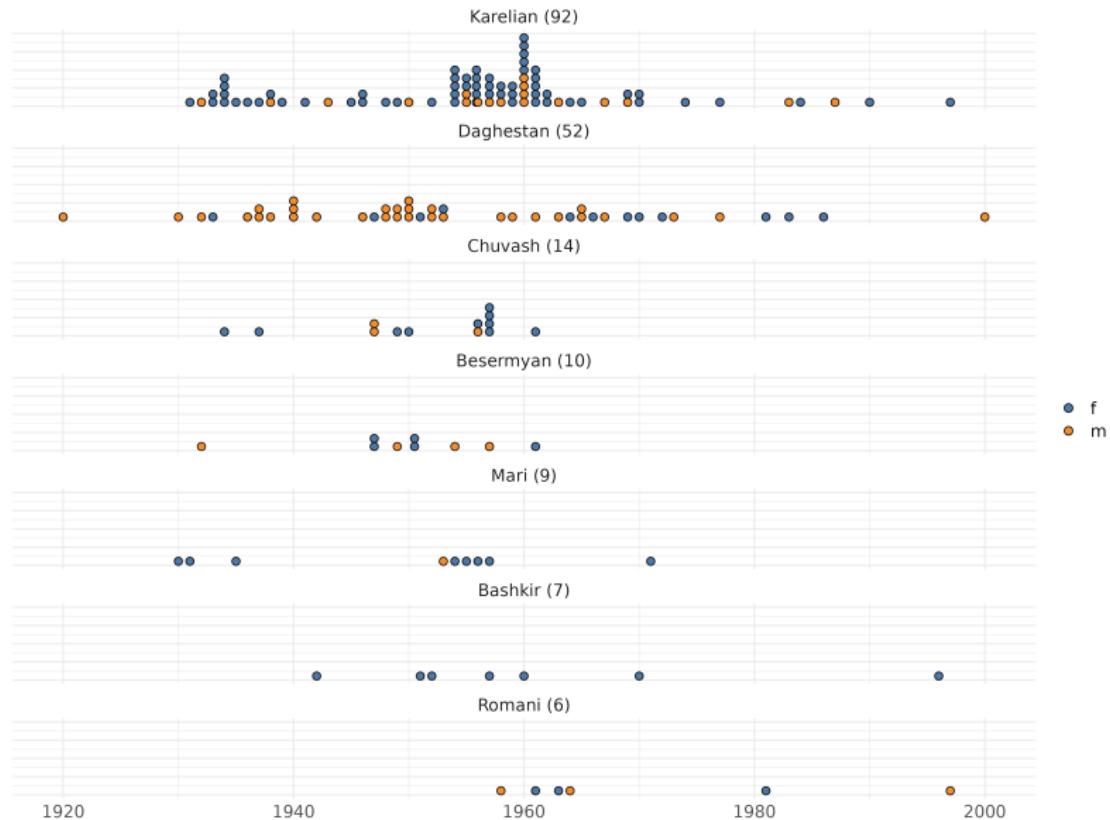
Почему мы ожидаем вариативность в числительных?

- система русских числительных сложная
- системы числительных в L1 доступных нам корпусах значительно проще
- количественные конструкции в речи билингвов исследовалась в работах ([Stoyanova 2019; Стойнова 2021](#))
- В работе ([Стойнова 2021](#)) употребление нестандартных конструкций объясняется контактом
- Увидим ли мы такой же эффект на основе данных наших корпусов?

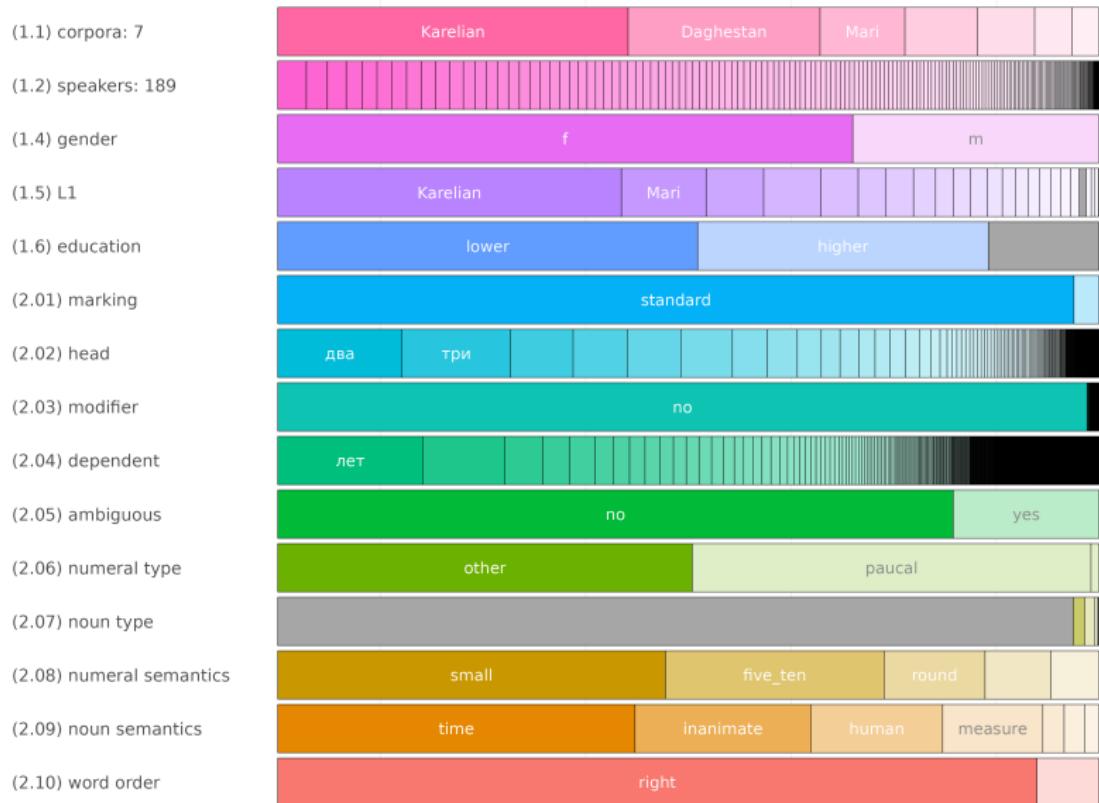
Данные

- Сначала мы отобрали 7,376 контекстов
- Мы исключили:
 - конструкции с порядковыми числительными (*в шестьдесят первом году*)
 - исключили неядерные падежи (*с двумя детьми*)
 - конструкции в которых есть неизменяемые существительные (*два медресе*)
 - конструкции без выраженного имени (*восемь было*)
 - конструкции с числительным *один* и существительными, которые имеют свойства имени (*тысяча, миллион*, и др.)
 - примеры, где нет социолингвистической информации о носителе
- Осталось: 4,196 примеров:
 1. *Пешком ходил Верхний Дженгутай пять километра.*
(Корпус дагестанского русского)
 2. *Этот меньше, после двое abort делала одну.* (Корпус марийских билингвов)

Носители

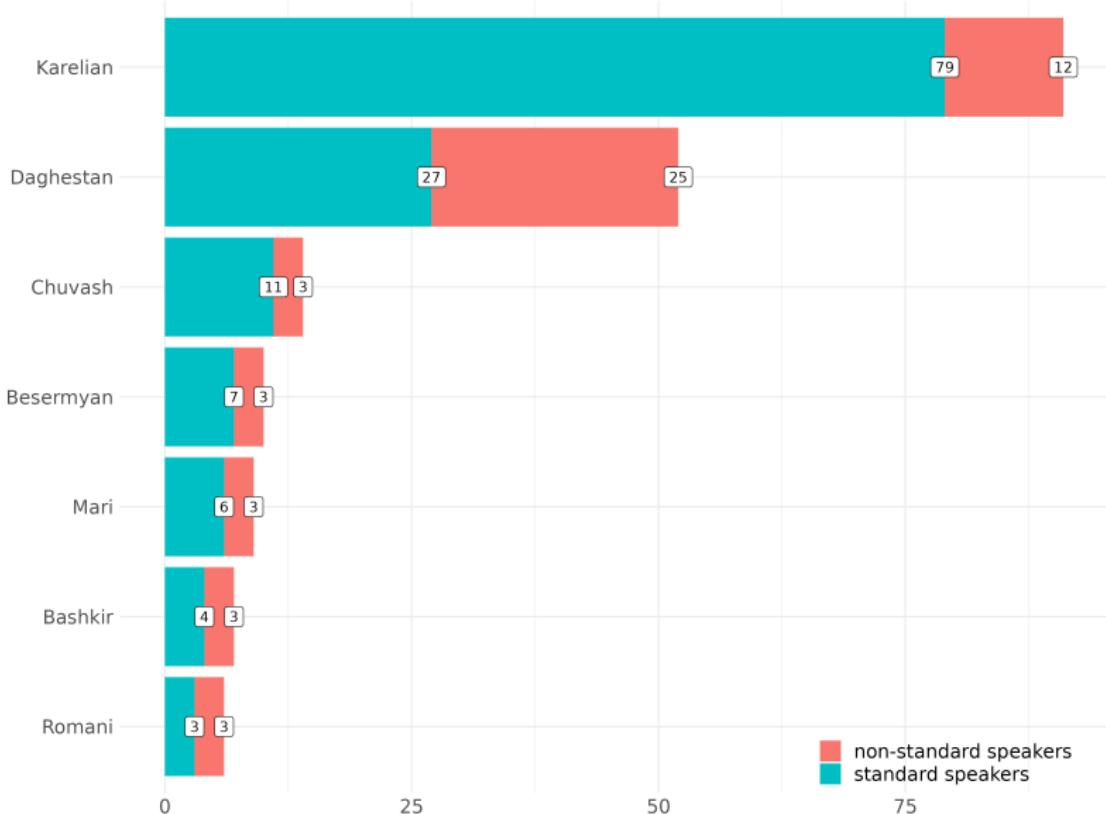


Структура данных

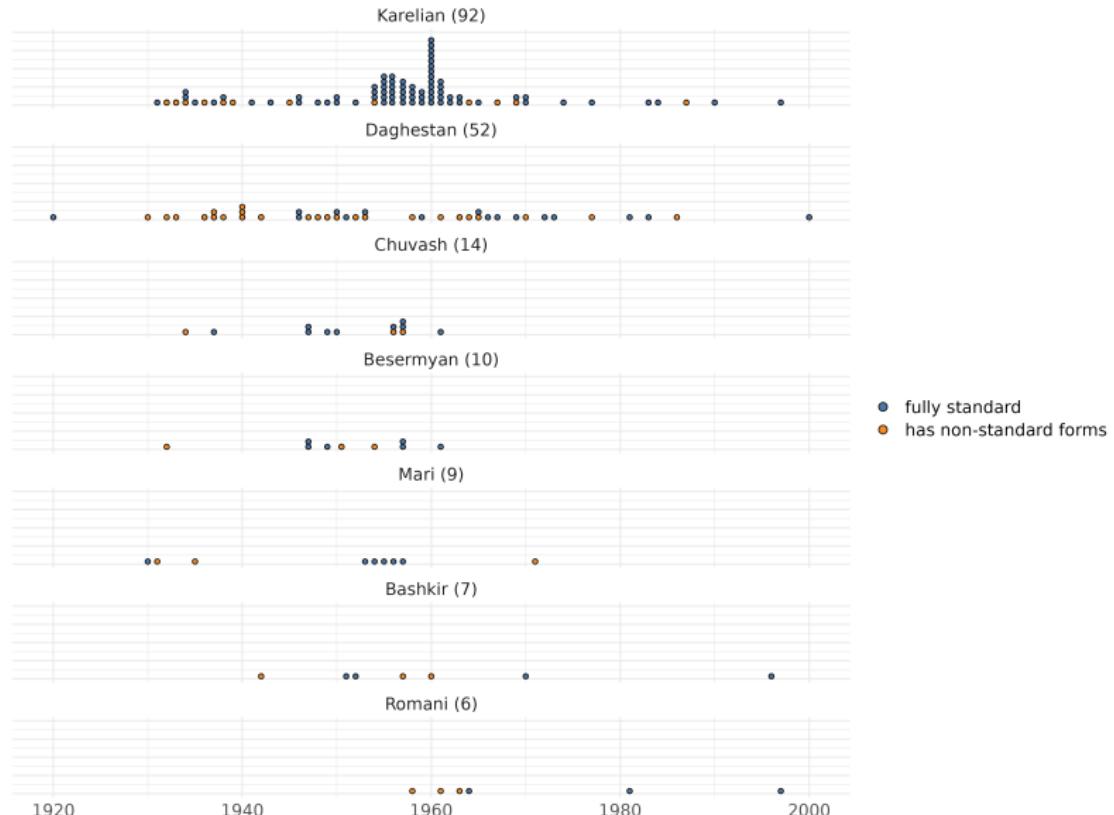


Gray segments are missing values

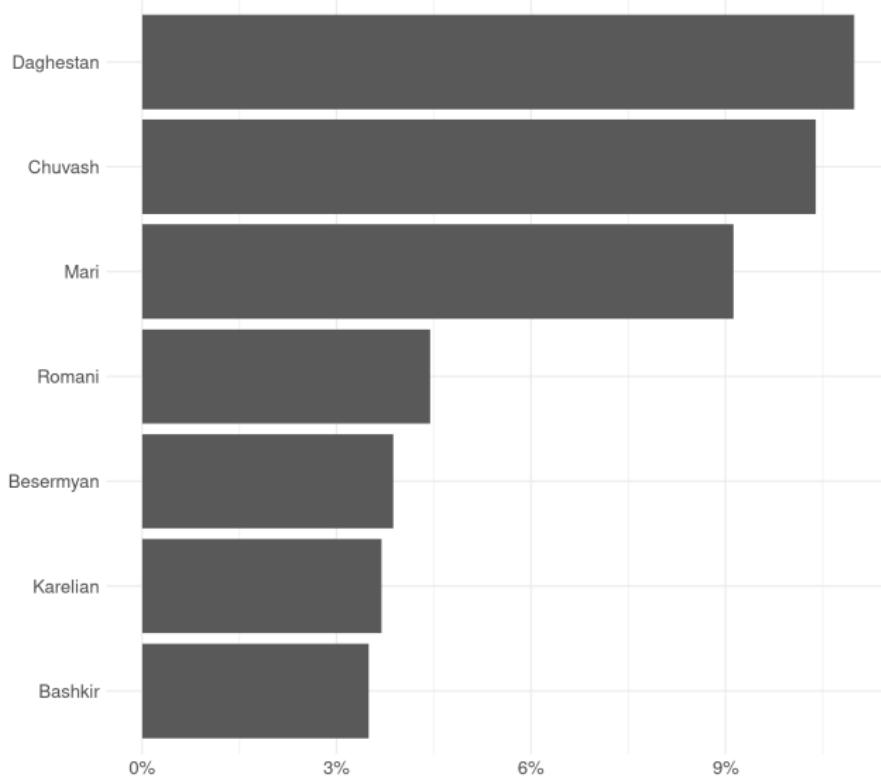
Только 52 из 189 носителей используют нестандартные конструкции (28%)



Только 52 из 189 носителей используют нестандартные конструкции (28%)



Процент конструкций, если оставить только носителей, которые используют нестандартные формы



Можно ли объяснить наблюдаемую вариативность контактом?

	paucals	other numerals
Russian	genitive singular	genitive plural
East Caucasian		nominative singular
Turkic		nominative singular
Mari and Udmurt		nominative singular
Karelian		partitive singular
Romani		nominative plural

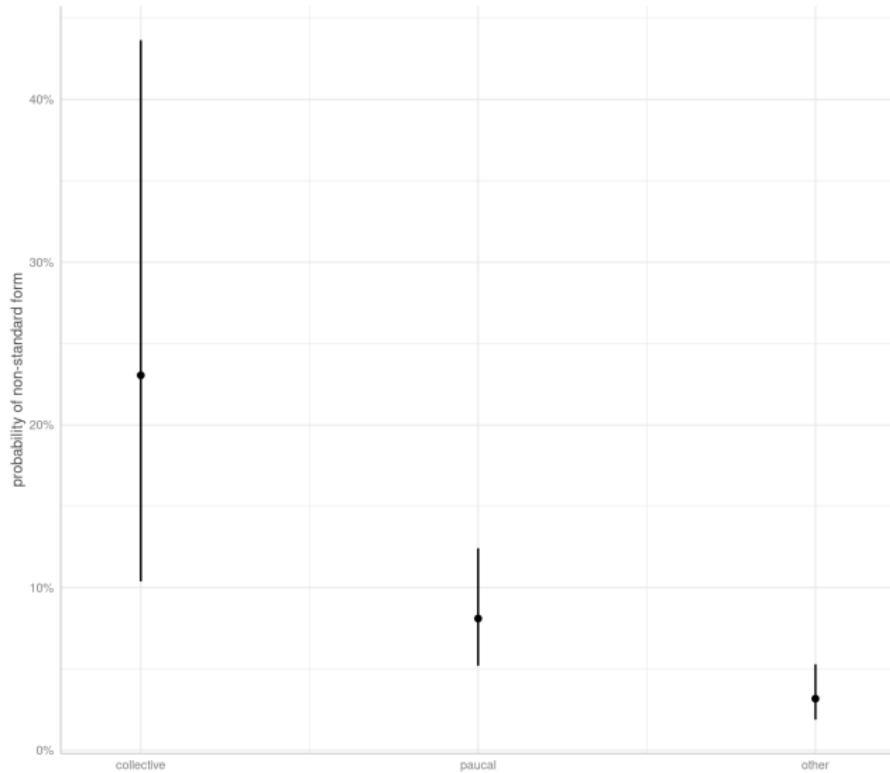
Можно ли объяснить наблюдаемую вариативность контактом?

	paucals	other numerals
Russian	genitive singular	genitive plural
East Caucasian		nominative singular
Turkic		nominative singular
Mari and Udmurt		nominative singular
Karelian		partitive singular
Romani		nominative plural

- Нет
 - Если бы наблюдавшее было просто калькирование, то мы бы ожидали больше случаев совпадения ожидаемой конструкции из L1, а такое происходит лишь в половине случаев.

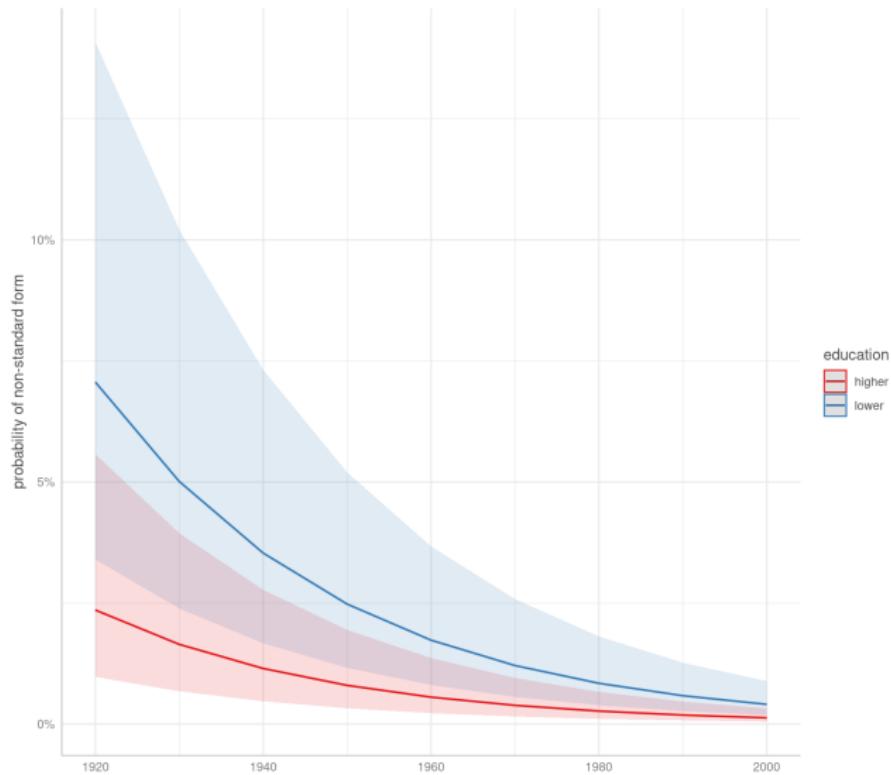
Логистическая регрессия со случайными эффектами

$P(\text{non-st. form}) \sim \text{type} + (1|\text{corpus/speaker})$



Логистическая регрессия со случайными эффектами

$$P(\text{non-st. form}) \sim \text{age} + \text{education} + (1|\text{corpus/speaker})$$



Выводы

- Конструкции с числительными иллюстрируют значительную вариативность в билингвальных корпусах, однако уровень вариативности разный в разных корпусах (самый большой в Дагестане 11%), но все же мы отмечаем меньший процент чем у носителей нанайского и ульчского (25.3%, ([Стойнова 2021: 316](#)))
- Наблюдаемые конструкции лишь в половине случаев объяснимы калькированием из L1

План доклада

Ресурсы международной лаборатории языковой
конвергенции

Нестандартные количественные конструкции в речи
билингвов

Выпадение предлогов в речи билингвов

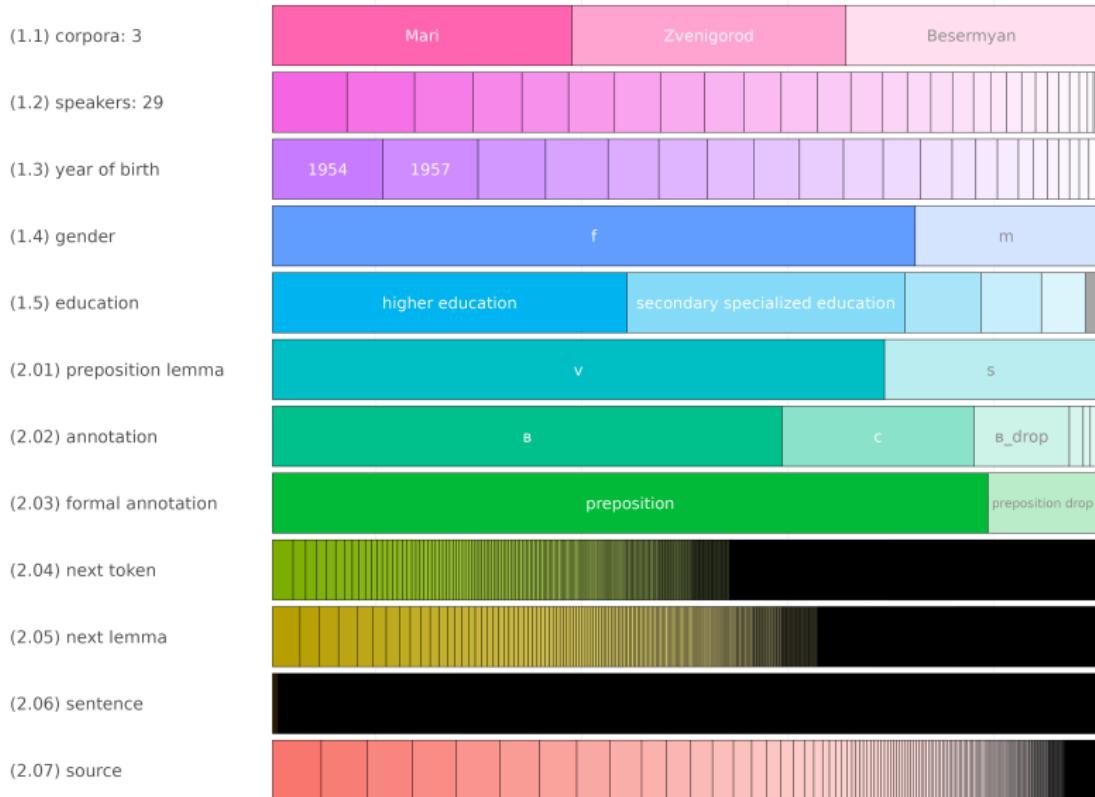
Теоретический контекст

Выпадение предлогов наблюдается и в других языках, например, в новогреческом, аглийском и в разных нестандартных вариантах русского

Больше чем корпусное исследование

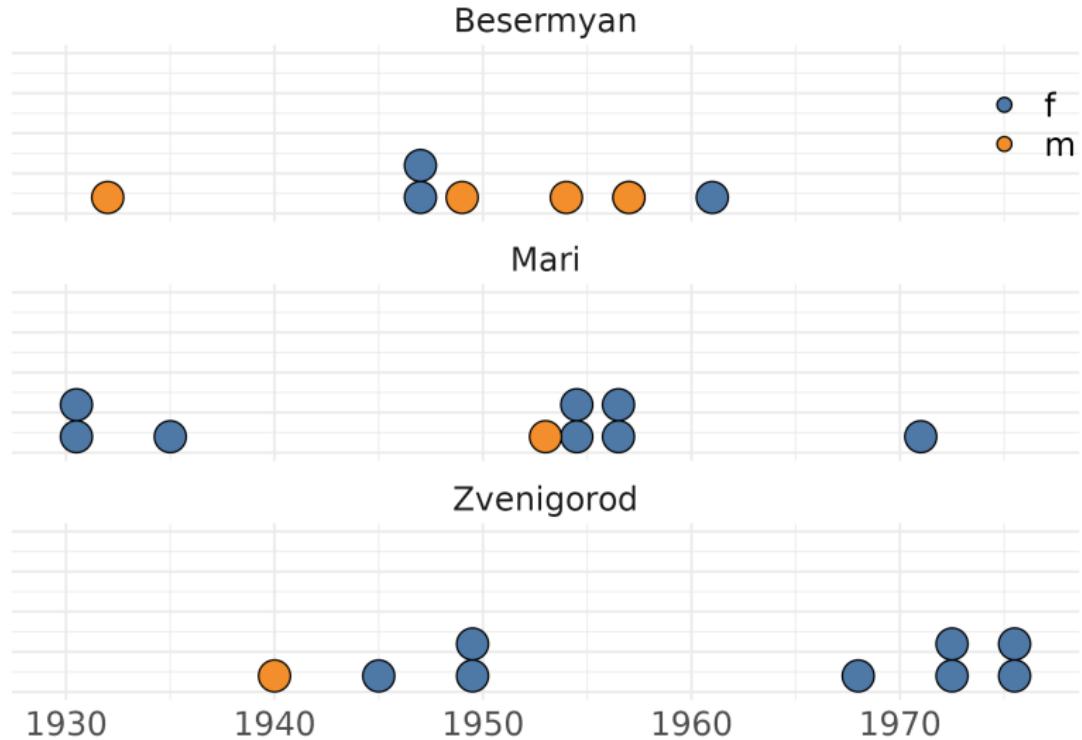
- При исследовании выпадения предлогов, мы не можем полагаться на разметчиков, так как корпуса размечались по нормам литературного языка.
- В связи с этим, мы вынуждены были переслушивать все контексты, которые нас интересовали.

Структура данных (4393 наблюдения)

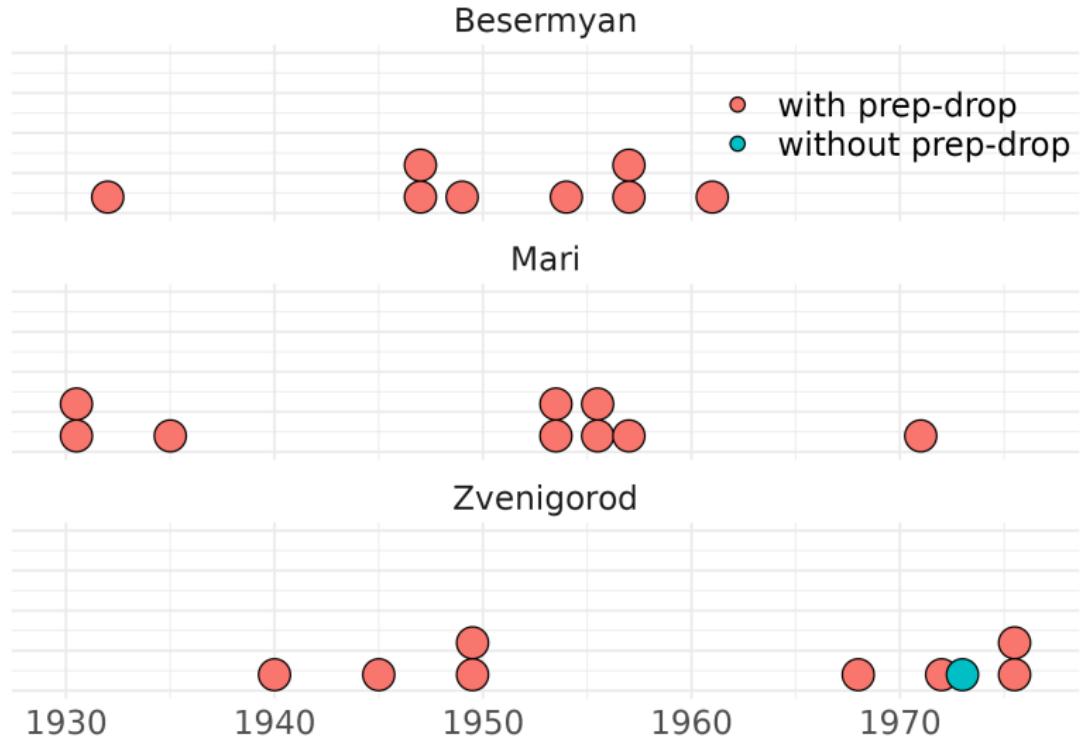


Gray segments are missing values

Носители

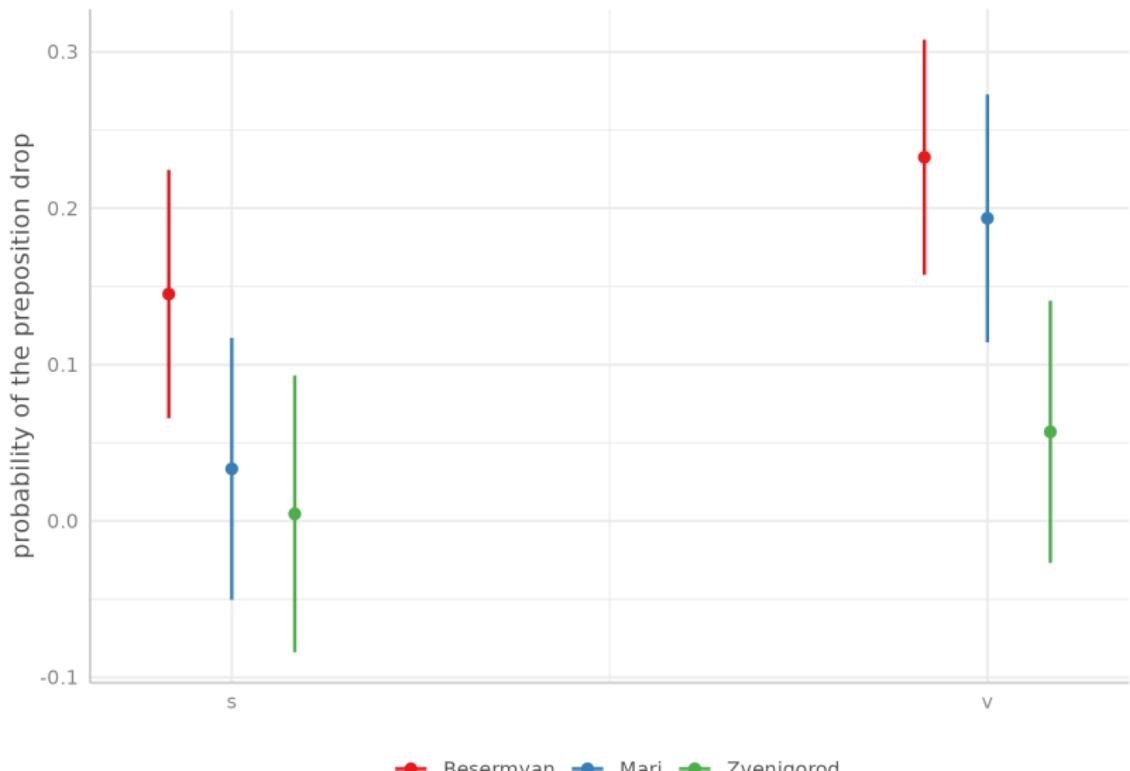


Носители, у которых наблюдается выпадение

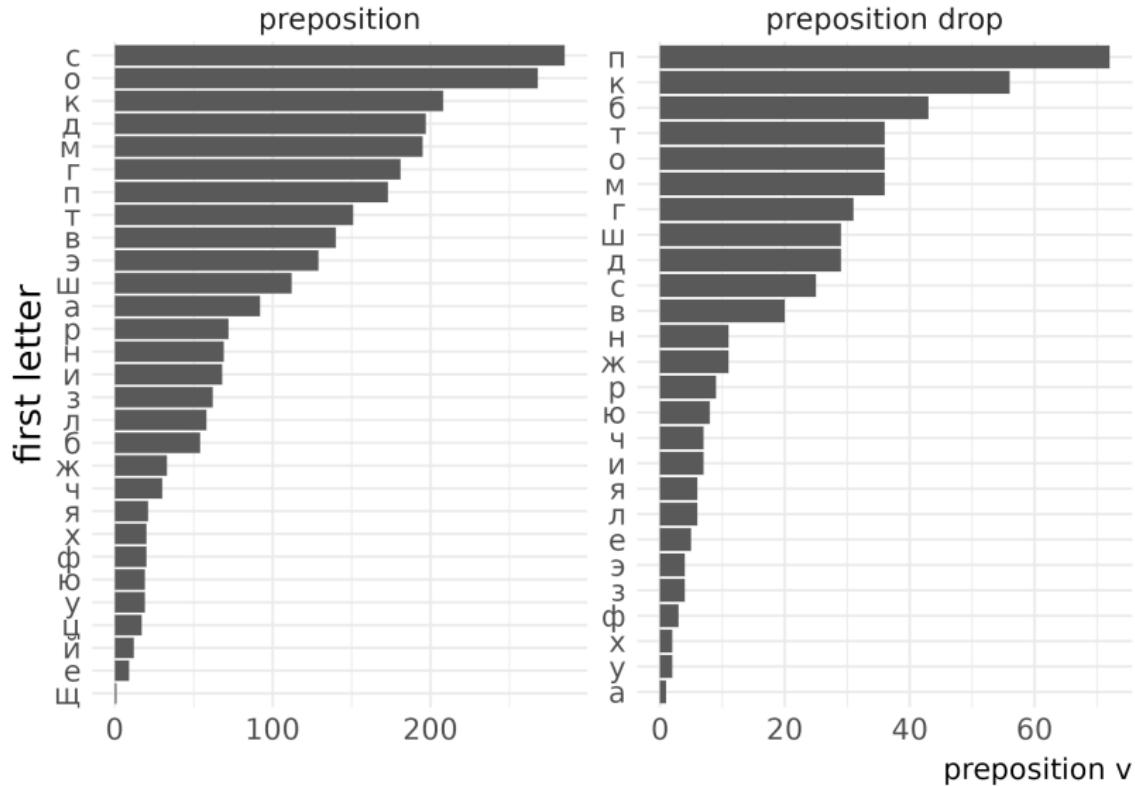


Логистическая регрессия со случайными эффектами

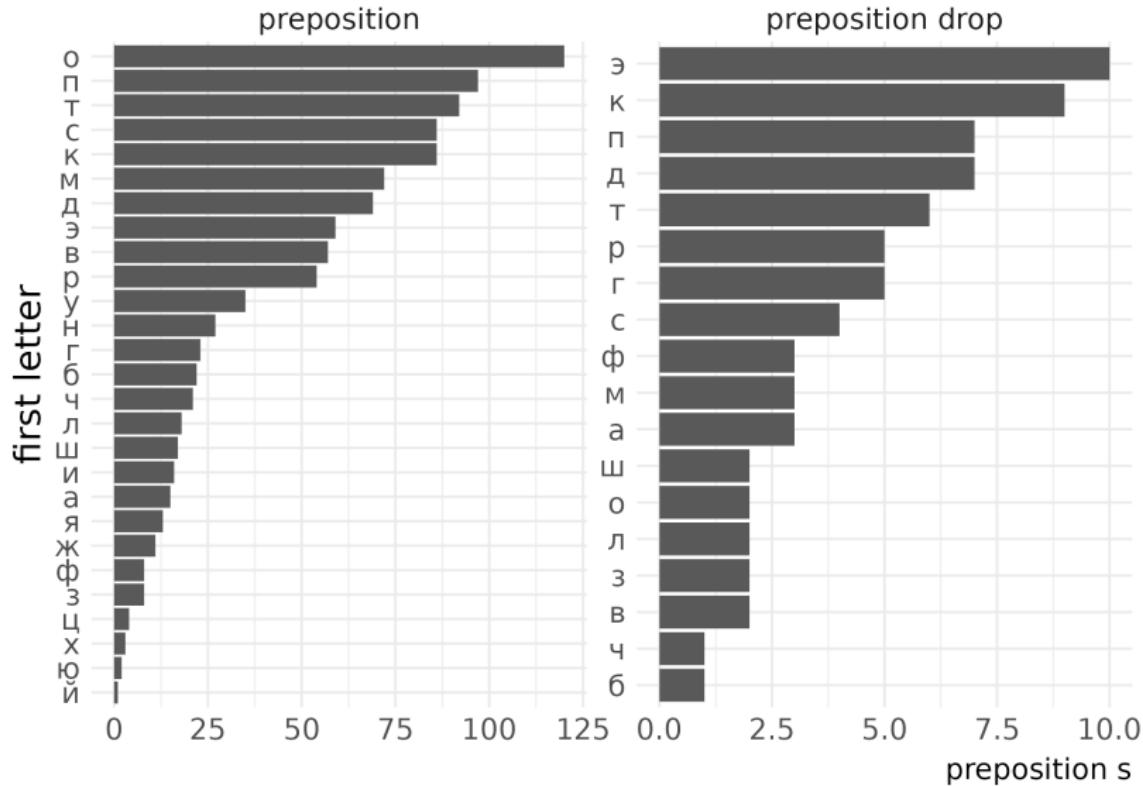
preposition drop ~ corpus * preposition + (1|speaker)



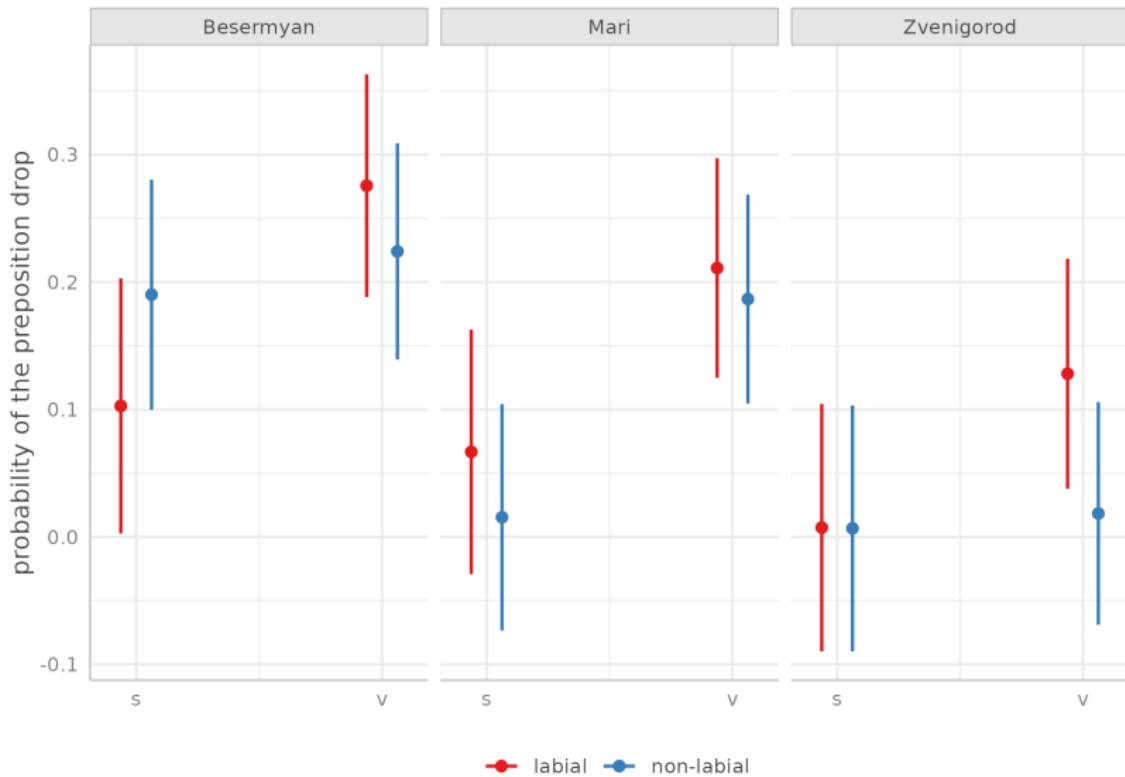
Может быть фонология играет какую-то роль?



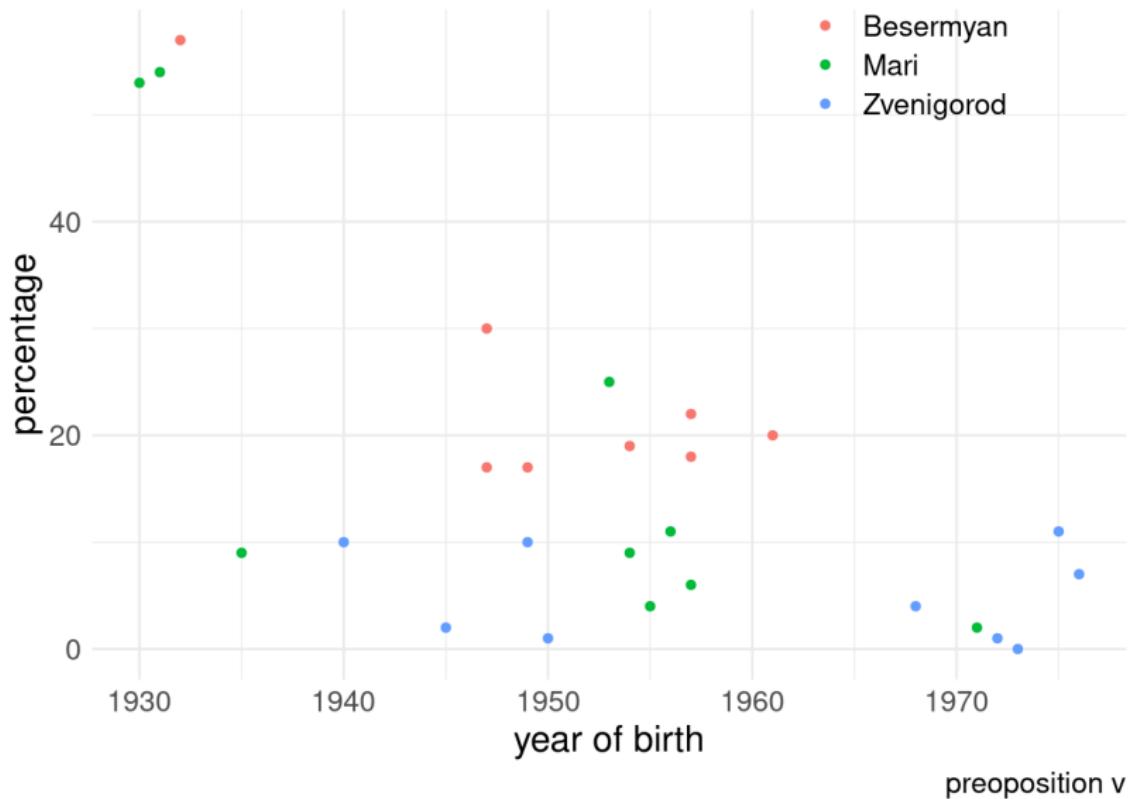
Может быть фонология играет какую-то роль?



Может быть фонология играет какую-то роль?

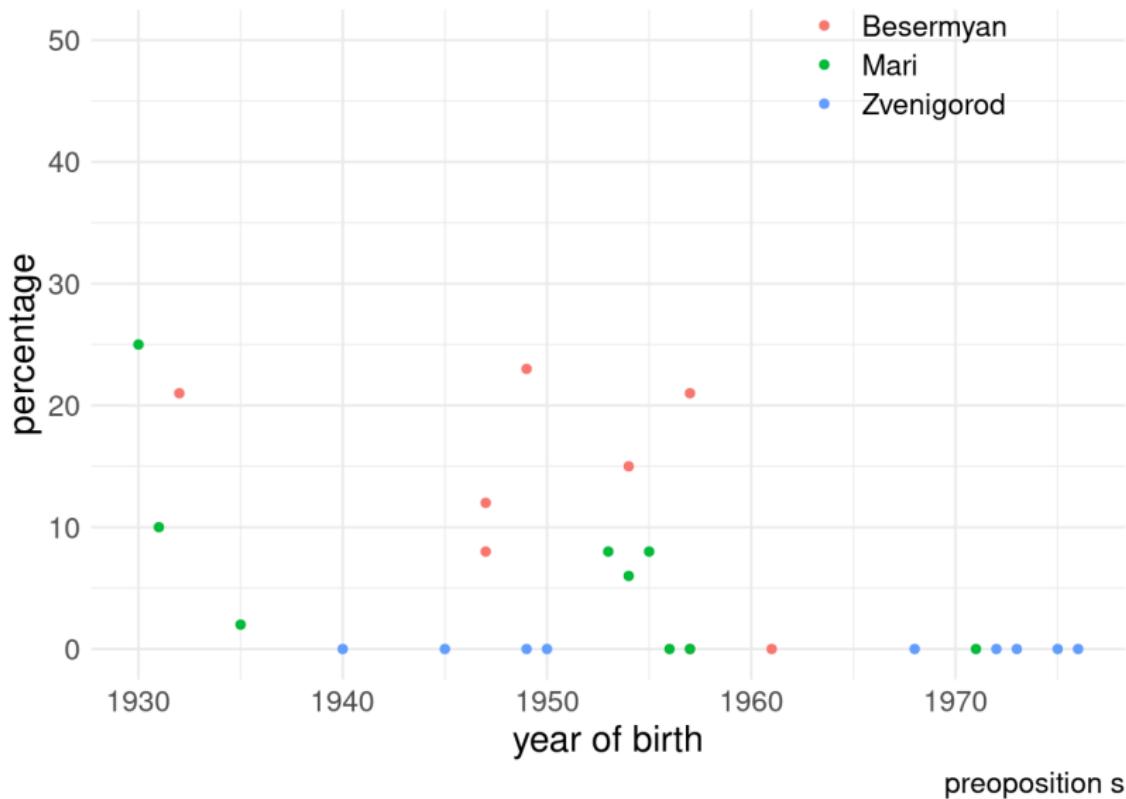


В каком корпусе больше всего выпадений?



preposition v

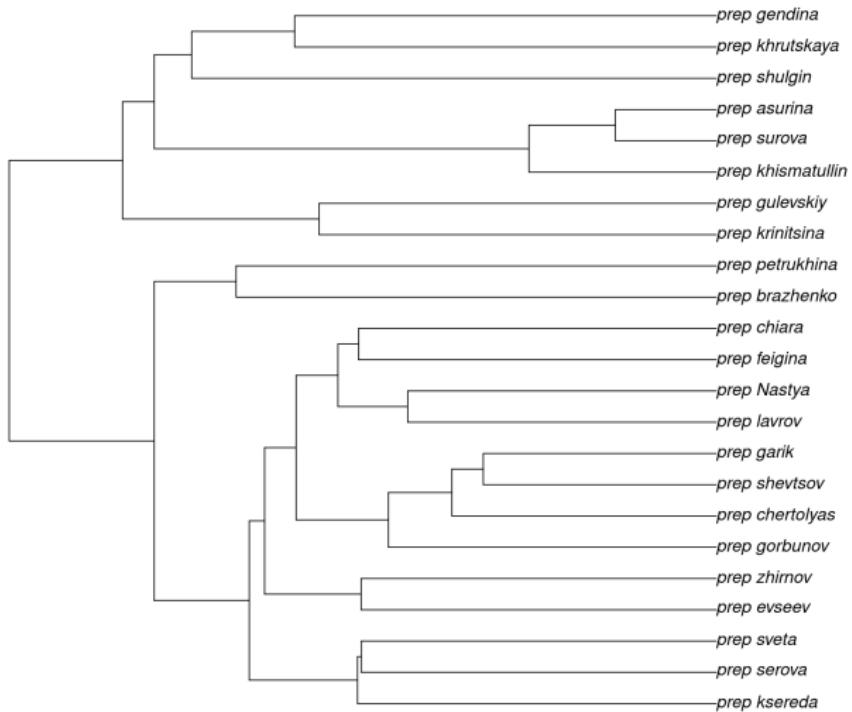
В каком корпусе больше всего выпадений?



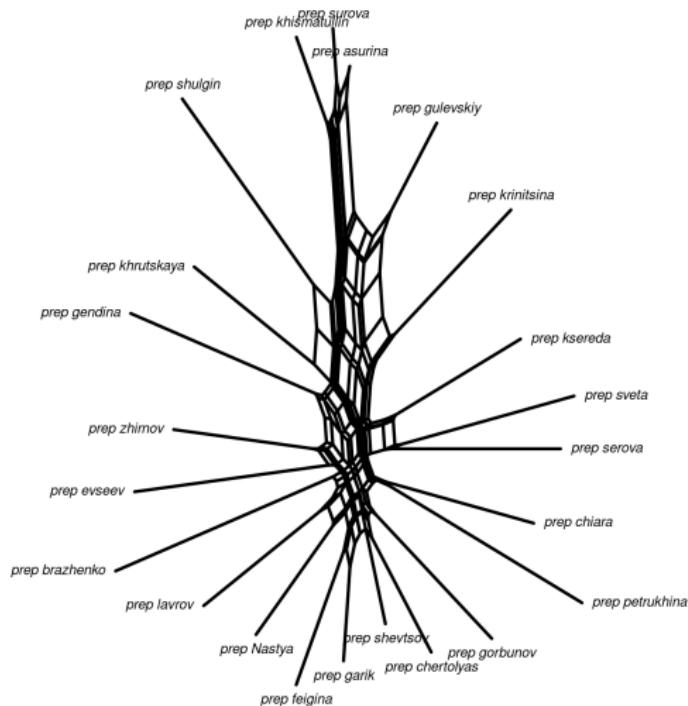
Анализ согласия

- каждый раз мы делаем вывод на основе прослушивания
- в какой-то момент мы решили проанализировать, а насколько одинаково мы принимаем решения
- потом мы сделали выборку в 200 примеров
- раздали 23 аннотаторам
- унифицировали разметку: бинарное (v-drop, v, NA)
- посчитали расстояния Жаккара и построили иерархическую кластеризацию и нейборнет

Анализ согласия



Анализ согласия



Выводы

- Носители из билингвальных корпусов демонстрируют больше выпадения предлогов;
- К сожалению, данных для моделирования возраста и образования недостаточно;
- Фонетика следующего слова может влиять на выпадение предлогов
- Иногда выпадения предлогов можно объяснить лексикализацией (*в общем*)

Спасибо за внимание!

Литература

- N. Stoynova. Russian in contact with southern tungusic languages: Evidence from the contact russian corpus of northern siberia and the russian far east. *Slavica Helsingiensia*, 52, 2019.
- C. В. Князев, Г. А. Мороз, and Дьяченко С. В. Корпус Просодии Русских Диалектов (ПРУД), 2024. URL
<https://LingConLab.github.io/PRuD/>.
- Г. А. Мороз. Скорость русского речи на основе билингвальных и диалектных устных корпусов. In Н. А. Коротаев and Н. Р. Сумбатова, editors, *Состав науки: Сборник статей к юбилею Веры Исааковны Подлесской*. Буки Веди, М., 2023.
- Н. Стойнова. Нестандартные количественные конструкции в русской речи носителей нанайского и ульчского языков. *Russian Linguistics*, 45, 2021.