Использование текстовых корпусов для изучения русско-горномарийского переключения кодов

Вадим Дьячков, ИЯз РАН
hyppocentaurus@mail.ru
Ирина Хомченкова, ИРЯ РАН / МГУ
irina.khomchenkova@yandex.ru

Введение

- Исследование переключения кодов в горномарийском языке с использованием разных типов корпусов
- **Луговой марийский**: вставка русской лексики вместо исконной марийской при обозначении дней недели, цветов, чисел и терминов родства, использовании русских дискурсивных маркеров, дублировании и окказиональных заимствованиях [Гаврилова 2012; Гаврилова 2013]
- **Горномарийский**: до настоящего момента оставался за рамками исследований о переключении кодов.

Материал исследования

- **Корпус** горномарийского языка, созданный участниками проекта под руководством Е. В. Кашкина в 2016–2018 гг. (63522 токенов, диалогическая и монологическая речь) в с. Кузнецово.
- **Подкорпус** (ELAN): диалоги с большим количеством включений на русском языке.
- Данные **элицитации**.
- Мы покажем, как разные виды материала позволяют решать различные задачи, стоящие перед исследователями переключения кодов в горномарийском языке.

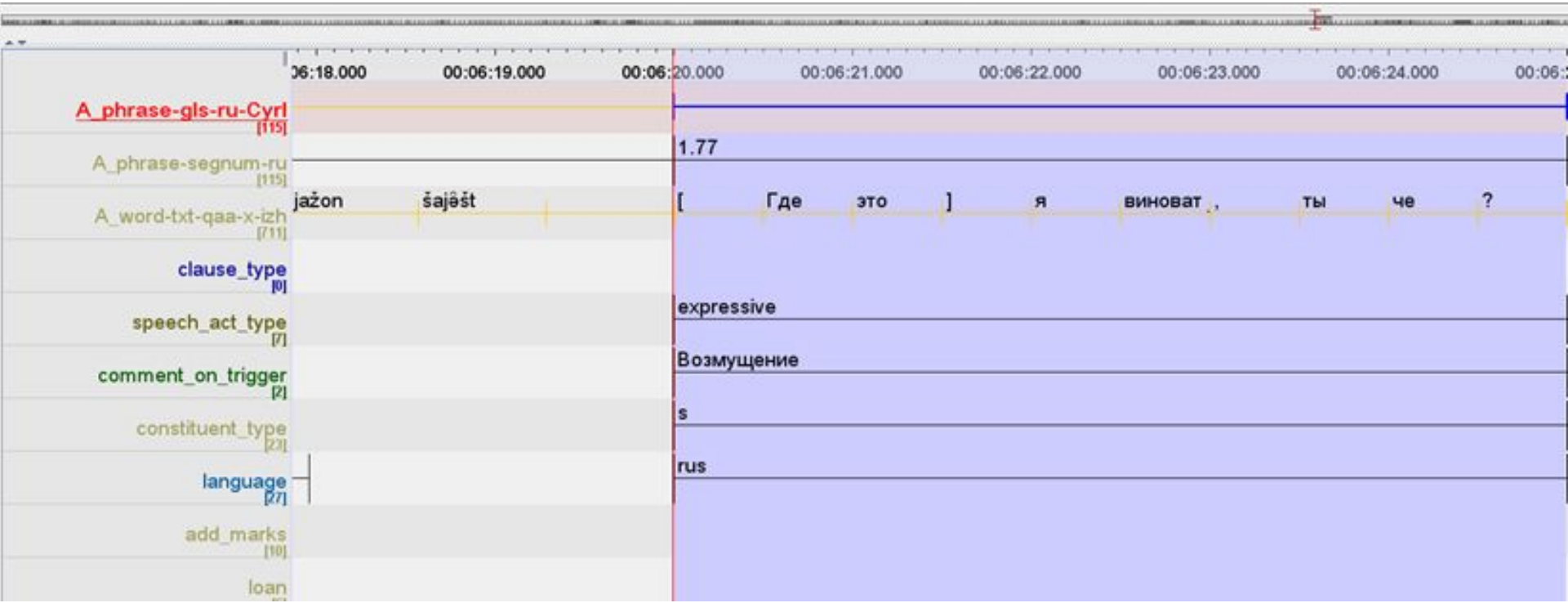
Подкорпус с переключением кодов

- Большой корпус не всегда позволяет выявить значимые синтаксические характеристики переключения кодов
- С этой целью мы отобрали 6 текстов с наибольшим количеством переключений

Подкорпус: разметка

- Около 47 минут звучащей речи
- Разметка выполнена в среде ELAN
- Имеется перевод и слой с разбиением на токены

Разметка


00:06:18.000
00:06:19.000
00:06:20.000
00:06:21.000
00:06:22.000
00:06:23.000
00:06:24.000
A_phrase-gls-ru-Cyrl [115]
A_phrase-segnum-ru [115]
A_word-txt-qaa-x-izh [711]
clause_type [0]
speech_act_type [7]
comment_on_trigger [2]
constituent_type [23]
language [27]
add_marks [10]
loan
1.77
jažon šajə̂št
[Где это] я виноват , ты че ?
expressive
Возмущение
s
rus

Подкорпус: разметка

Слои разметки:

- *clause_type* — тип придаточного предложения
- *speech_act_type* — тип речевого акта, отражающий некоторую иллокутивную цель говорящего
- *comment_on_trigger* — комментарий в свободной форме
- *constituent_type* — синтаксический тип переключаемой составляющей (= XP);
- *language*
- *loan*

Подкорпус: разметка

Набор тэгов:

- *discm* — дискурсивные маркеры
  - (*в общем, может, видишь…*)
- *pred* — предикативные единицы, отличные от глаголов
  - *А-а-а* , ***все понятно****, jara , jara,* ***все понятно***
- *adv* — адвербиалы
- *pp* — предложная группа
- *s* — клауза (главная или зависимая)

Типы составляющих

	s	pp	adv	pred	discm	conjp	vp
Text 1_a	6	0	2	0	3	0/1	0
Text 2	7	4	4	4	5	1/0	0
Text 3	6	1	3	5	3	0	0
Text 4	4	1	6	3	4	0	0
Text 5_a	4	4	7	1	6	0	7
Text 5_b	9	3	5	1	1	0	3
итого	**36**	**13**	**27**	**14**	**22**	**2**	**10**

Подкорпус: результаты

- Русские речевые клише (*в гробу сто раз перевернулся*)
- Наибольшее количество переключений — адвербиалы и дискурсивные маркеры
- Практически нет переключений внутри именной и глагольной групп

Подкорпус: плюсы и минусы

+

1. Можно зафиксировать некоторые синтаксические **тенденции**
2. Можно отметить некоторые **типы речевых актов**, при которых производится переключение

—

1. Имеющиеся в распоряжении тексты — **не спонтанные**
2. Ничего нельзя сказать о **причинах употребления** некоторых конструкций: например, неясно, действительно ли PP часто переключаеются?

Роль русскоязычных включений

- Мнение носителя:
  нормально сказать *диссертациям досрочно защищаяш*
- Другие наблюдения:
  {Разговор на горномарийском в магазине между продавцом и покупателем.}
  Покупатель:
  - *Третьим будешь?*

Элицитация переключения кодов

- Полученная в ходе анализа текстовой коллекции информация послужила основой для экспериментов, в ходе которых мы протестировали некоторые ограничения на русскоязычные включения

Элицитация переключения кодов

- Не проверяем синтаксические ограничения внутри изолированного предложения
- Но строим анкеты с **включением целевых фрагментов** нужного типа
- Анкеты строятся как диалогические тексты на хорошо понятную информантам тематику
- Методология описана в [François 2019]

Эксперимент

- Тексты на две темы:
    1. «Рассматривание фотографии»
    2. «Покупка товаров в магазине»
- 14 готовых текстов от 10 разных носителей
- Инструкции носителям: переводить не задумываясь, можно использовать любые слова и переводить не дословно

Эксперимент: плюсы и минусы

+

1. Информанты легко представляют себе моделируемую ситуацию
2. Нет ограничений на способ произнесения выражений
3. Даже на ограниченном материале удалось выявить несколько новых структурных типов переключения кодов

—

1. Элицитация — все еще не естественная, а моделируемая речь
2. Велика вероятность прайминга

Эксперимент: “Покупка товаров в магазине”

— Здравствуйте! А можно мне вот эти два куска сыра?
— Тебе каких, побольше, поменьше?
— Давай тот, который поменьше. **А** еще колбасы немного. **«Кремлевский сервелат»** есть или закончился?
— Нет, вчера **студенты из Москвы** пришли, все раскупили. **Вот** есть немного **сырокопченой**.
— **Ну давай, ладно**. **А** еще есть сгущенка и тушенка?
— Надо посмотреть. **А** сколько банок нужно?
— Штук 10.
— А зачем тебе столько?
— Рабочие приехали дом разбирать, их кормить надо. А еще сын приехал на выходные. Я стряпать сейчас буду. Еще **приправу для курицы** надо.
— Нету, наверно. Сейчас посмотрю. Нет, закончилась.

Эксперимент: “Покупка товаров в магазине”

— Еще **чай черный** надо **в пачках**.
— Тебе какой?
— А какой там остался?
— Есть **обычный чай**, есть **ароматизированный,** вот **с абрикосом** одна упаковка осталась.
— А **в пакетиках** какой чай есть?
— **В пакетиках — «Лесные ягоды», «Малина», «Черника»**.
— Давай тогда **в пакетиках, «Лесные ягоды»**, три упаковки. Еще сочников возьму. Они свежие?
— Свежие, **конечно**! **Вот только что** привезли. Бери еще пирожки, **сосиску в тесте**. Вот вчерашние пирожки, они подешевле. А вот здесь сегодняшние.
— Нет, пирожки не буду сегодня брать. Давай мне еще печенье **«Топленое молоко»**, полкило.
— Всего **576 рублей 70 копеек**. Ой, у тебя только тысяча? Надо разменять. Ладно, сейчас посмотрю в той кассе. Вот, есть сдача.

Эксперимент

- Фиксировались
  - прямые русские включения
  - паузы при порождении речи (“!”)

Эксперимент: результаты

	rus	!	всего	
Кремлевский сервелат	9	0	9	100%
сырокопченый колбаса	9	0	9	100%
топленое молоко	9	0	9	100%
сосиска в тесте	8	0	9	88,9%
а	8	0	9	88,9%
наверно	8	0	9	88,9%
конечно	6	0	9	66,6%
сейчас	6	0	9	66,6%
только что	5	0	9	55,6%
приправа для курицы	3	4	9	33,3%
“Лесные ягоды”, “Малина”, “Черника”	3	3	9	33,3%
(в) пакетиках	2	0	9	22,2%

Эксперимент: результаты

- Адвербиалы и дискурсивные частицы воспроизводятся в неизменном виде
- Вариативность фрагментов с адъективными составляющими
- Несколько “незапланированных” структур с включением русскоязычных фрагментов

Адъективные составляющие

— А в пакетиках какой чай есть?

— В пакетиках — “Лесные ягоды”, “Малина”, “Черника”

- “Остров включенного языка” (в модели Майерс-Скоттон)

  a *paket-an* *čaj-vlä* ***лесные ягоды,*** *малина, черника* *ulə̑*
  пакет-PROP чай-PL EX

- ML+EL составляющая (встраивание русских лексем в горномарийскую рамку)

  paket-ə̑štə̑ ***лесной ягода****, малина, черника*
  пакет-IN

“Незапланированные” структуры

- *Ну давай, ладно.* {сырокопченую колбасу}

1.	Nu	jara	…	2
	ну	хорошо		
2.	Ну	давай(те)	…	5
3.	Ну	ладно	…	1
4.	Ну	давайте	тогда…	1

“Незапланированные” структуры

- Эффекты правостороннего ветвления?

— *Давай тогда в пакетиках, «Лесные ягоды», три упаковки.*

1. Давай(те) тогда… 5
2. Давай(те)… 2

- NB: davaj(te) - показатель гортатива!

“Незапланированные” структуры

The Double Morphology Principle (Myers-Scotton 1993: 133, Hok-Shing Chan 2009)

- [при советской власти] -štə̂
  -IN
  Предложная фраза (rus) - послеложная фраза (hm)

- — *А можно мне вот эти два куска сыра?*

1.	…	kok	[laštə̂k	сыр]	-ə̂m
2.		kok	[кусок	сыр]	-ə̂m
		два			-ACC

“Незапланированные” структуры

- *Бери еще пирожки, сосиску в тесте.*

1.	…	[сосиска	в	тесте]	-m -ACC	3
2.	…	[сосиски	в	тесте]		2
3.	…	[сосиску	в	тесте]		1

Основной корпус

- Для некоторых целей нужен не подкорпус, а весь объем материала
- Например, проверить, что может влиять на переключение кодов в какой-либо конкретной конструкции

Основной корпус

- Конструкции с количественными и порядковыми числительными
  - Относительно легкий поиск по всему корпусу как русских, так и горномарийских числительных
  - Русские и горномарийские конструкции отличаются: неконгруэнтная структура → переключение кодов

mən'ə	*mar-lan*	*ke-n-äm*	***təžem***	***əndekš***	***šüdə***	
я	мужчина-DAT	идти-PRET-1SG	тысяча	девять	сто	
kändäkš	***lu-šə***	***i-n***	***s'emnadcatə̑j***	***maj-ə̑n***		
восемь	десять-ORD	год-GEN		-GEN		

‘Я вышла замуж в 1980 году, 17 мая’.

Числительные в русском и горномарийском

Количественные

- Русский: числительные приписывают генитив существительному
- Горномарийский: не влияют на падеж существительного
  - *kə̑m ə̈rvezäš* / **ə̈rvezäš-ə̈n* <три мальчик мальчик-GEN> ‘(Ко мне подошли) три мальчика’
- Разные правила числового маркирования

st'ip'end'ij-žə̑ ***двенадцать рублей*** *ə̑l'ə̑* ***в месяц***
стипендия-POSS.3SG быть-AOR.3SG
‘Стипендия была двенадцать рублей в месяц’.

Числительные в русском и горномарийском

Порядковые

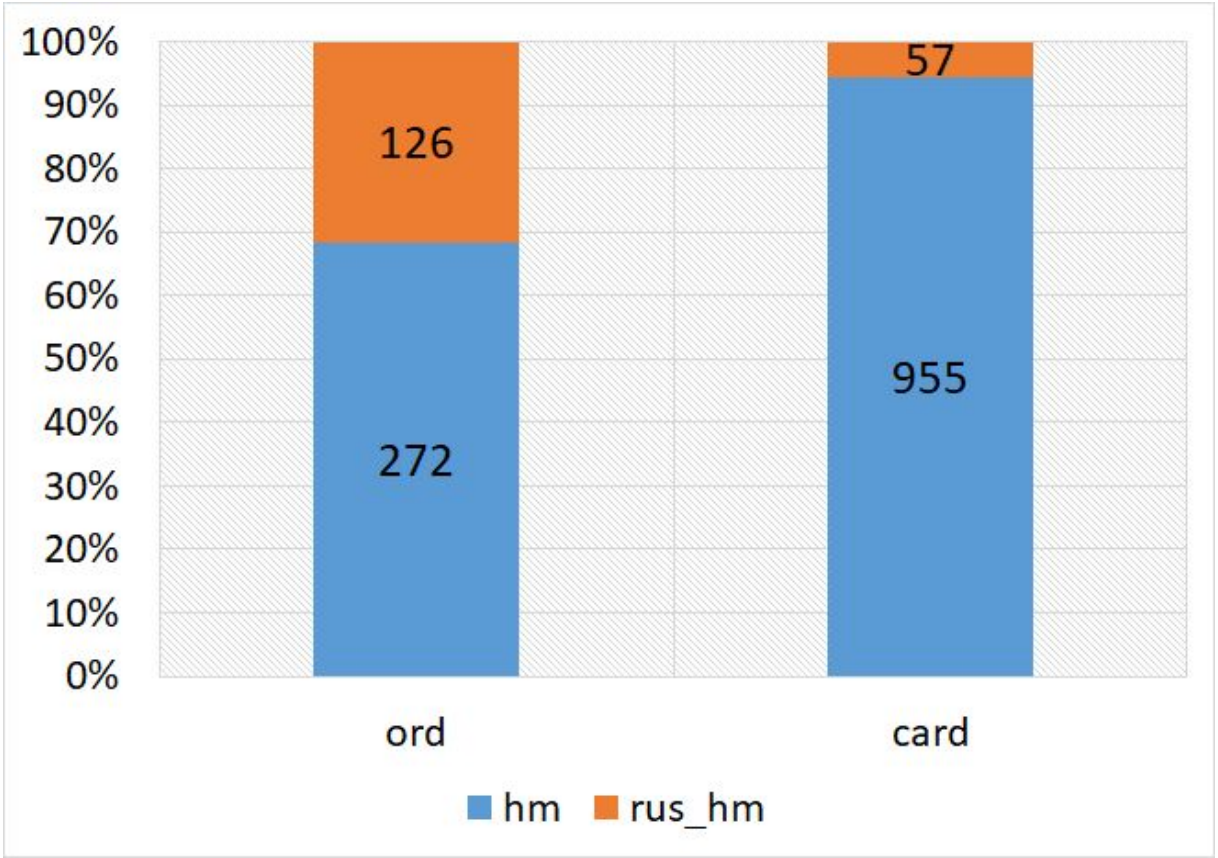
- Русский: числительные имеют формы всех трех родов (*третий*, *третья*, *третье)*, согласуются с существительным по числу и падежу (*о третьем человеке)*.
- Горномарийский: категории рода нет, числительные не согласуются с существительными
  - *kə̑m-šə̑ / *kə̑m-šə̑-m jamdar-ə̑m* <три-ORD три-ORD-ACC бутылка-ACC>
    ‘(Дед) третью бутылку (в руках держал)’.

p’ervə̑j ***sm’enə̈***-*m=ät* *nänge-ä*
первый смена-ACC=ADD вести-NPST.3SG
‘Он и первую смену ведет’.

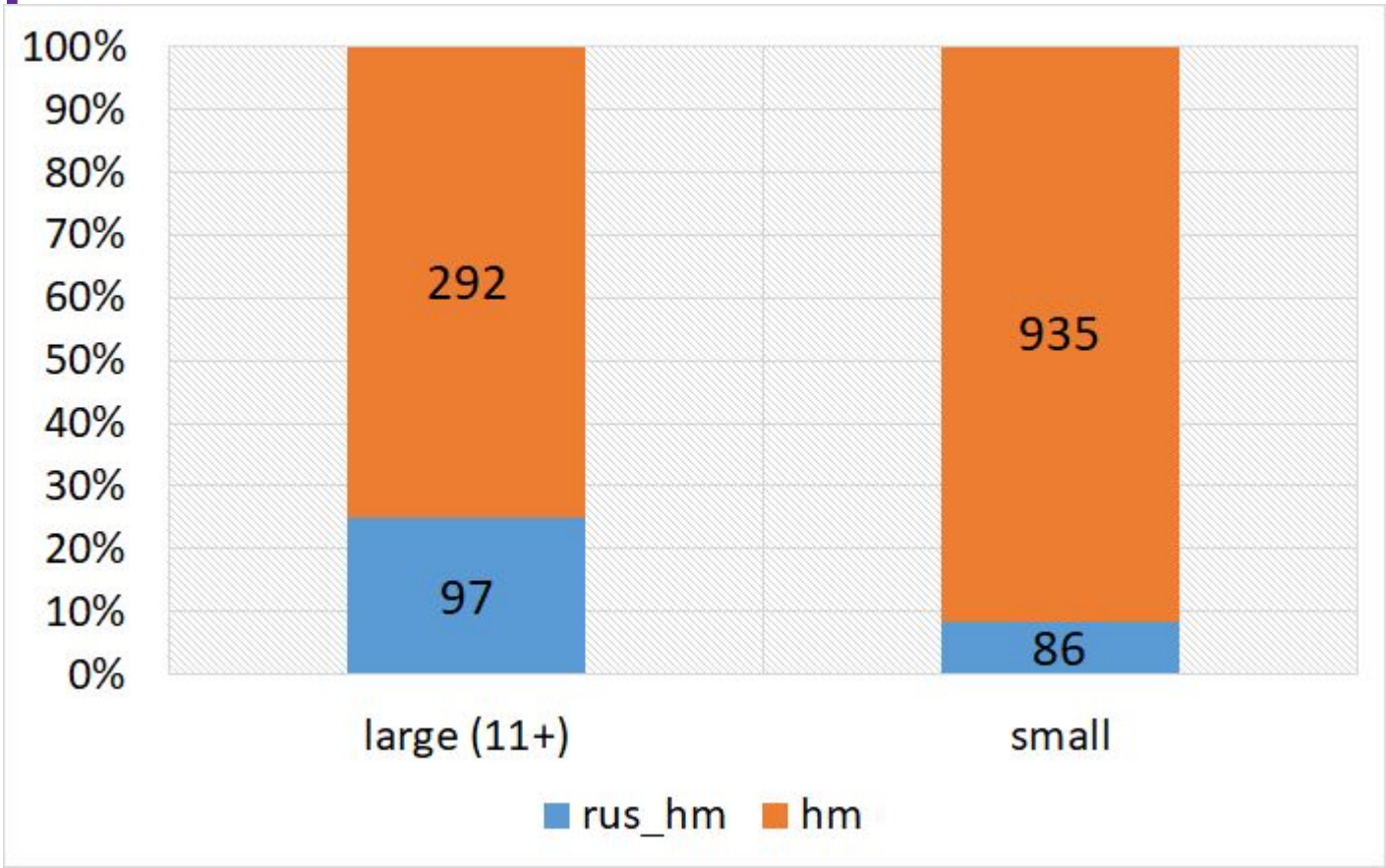
Выбор языка: факторы

- Тип числительного
  - количественное vs порядковое
- Арифметическое значение
  - русский используется для больших числительных, например, 11+ в коми-пермяцком [Максимов 2017]
- Семантический тип контекста
  - Удмуртский: русские числительные в контекстах времени, даты, денег… [Максимов 2017]
- **NB** Эти признаки во многом пересекаются: *year = ordinal & large*

Синтаксический тип числительного

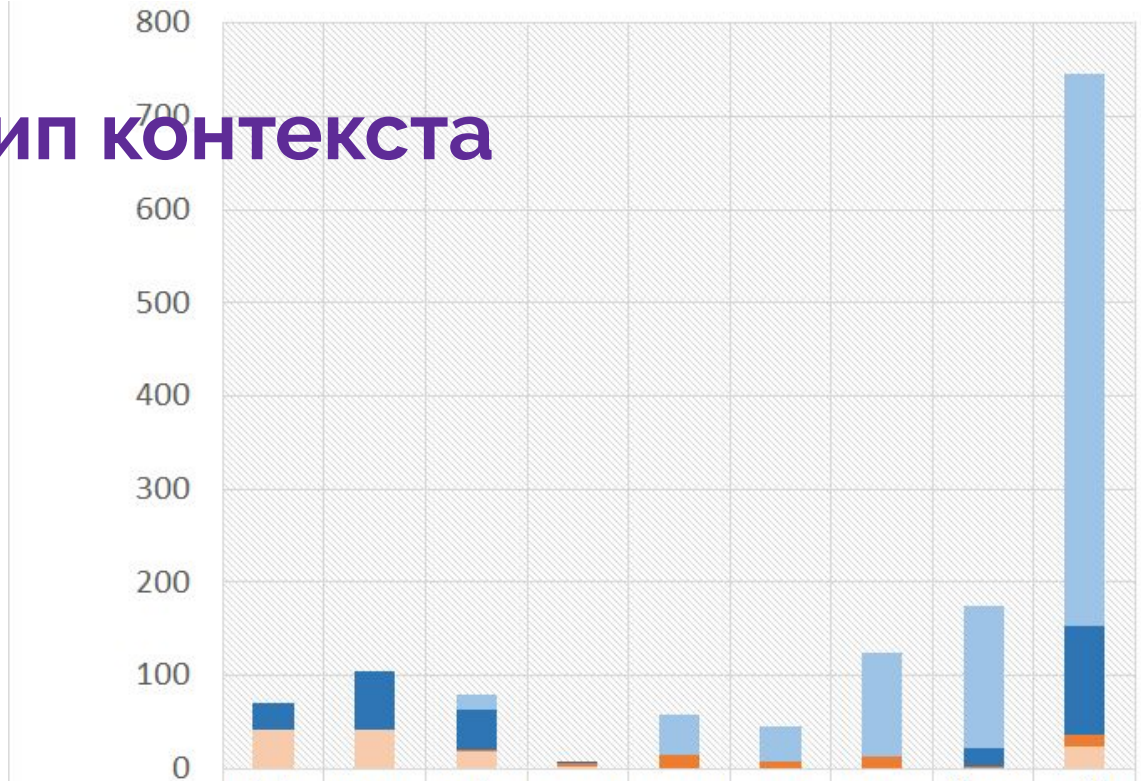


Арифметическое значение числительного



Семантический тип контекста

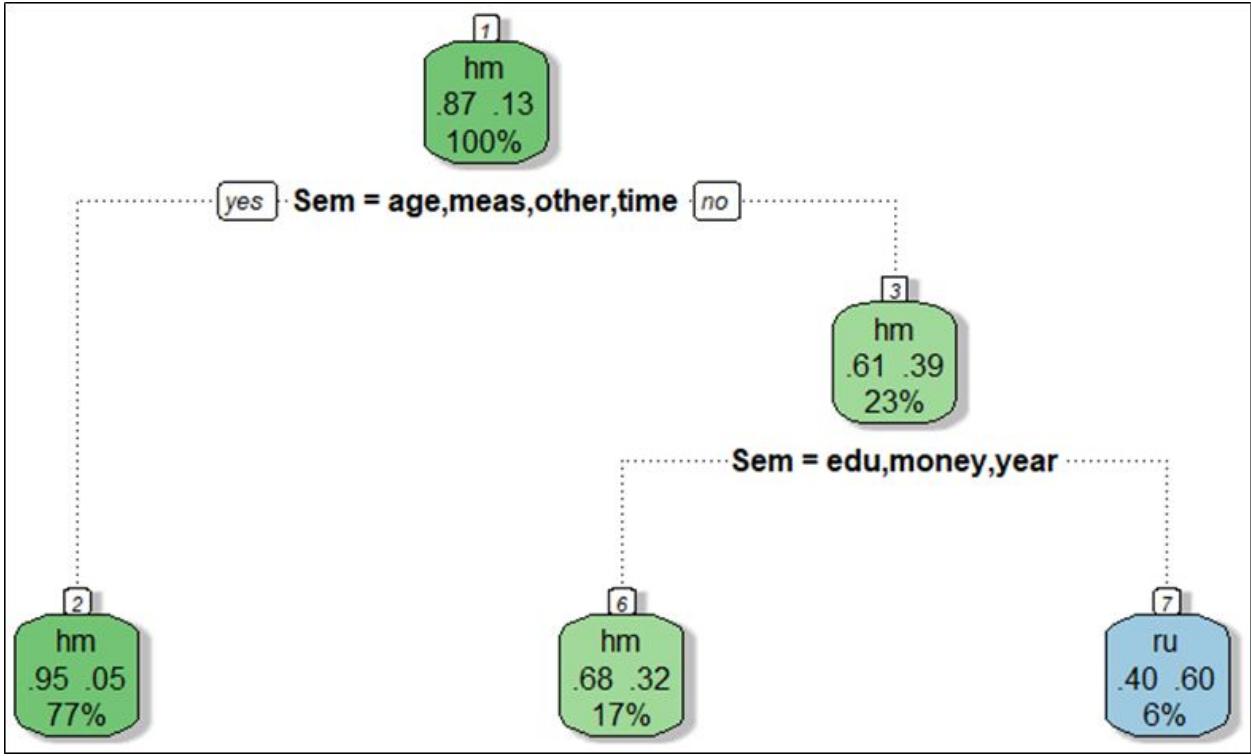
- **Контексты**: year (*1973, 19th cent.*), date (*July, 21st*), education (*5th grade, lessons*), num (*phone*), money, measure (*g, km*), time (*60 years, first day, for two days*), age (*6 y.o.*), other
- **Результат**: number (75%), date (59%), year (40%), education (25%), money (26%), age (16%).



	date	year	edu	num	mon	age	meas	time	oth
hm-card	0	0	17	0	43	38	112	153	592
hm-ord	29	63	43	2	0	0	0	19	116
r-card	0	0	2	4	15	7	12	3	14
r-ord	41	42	18	2	0	0	0	0	23

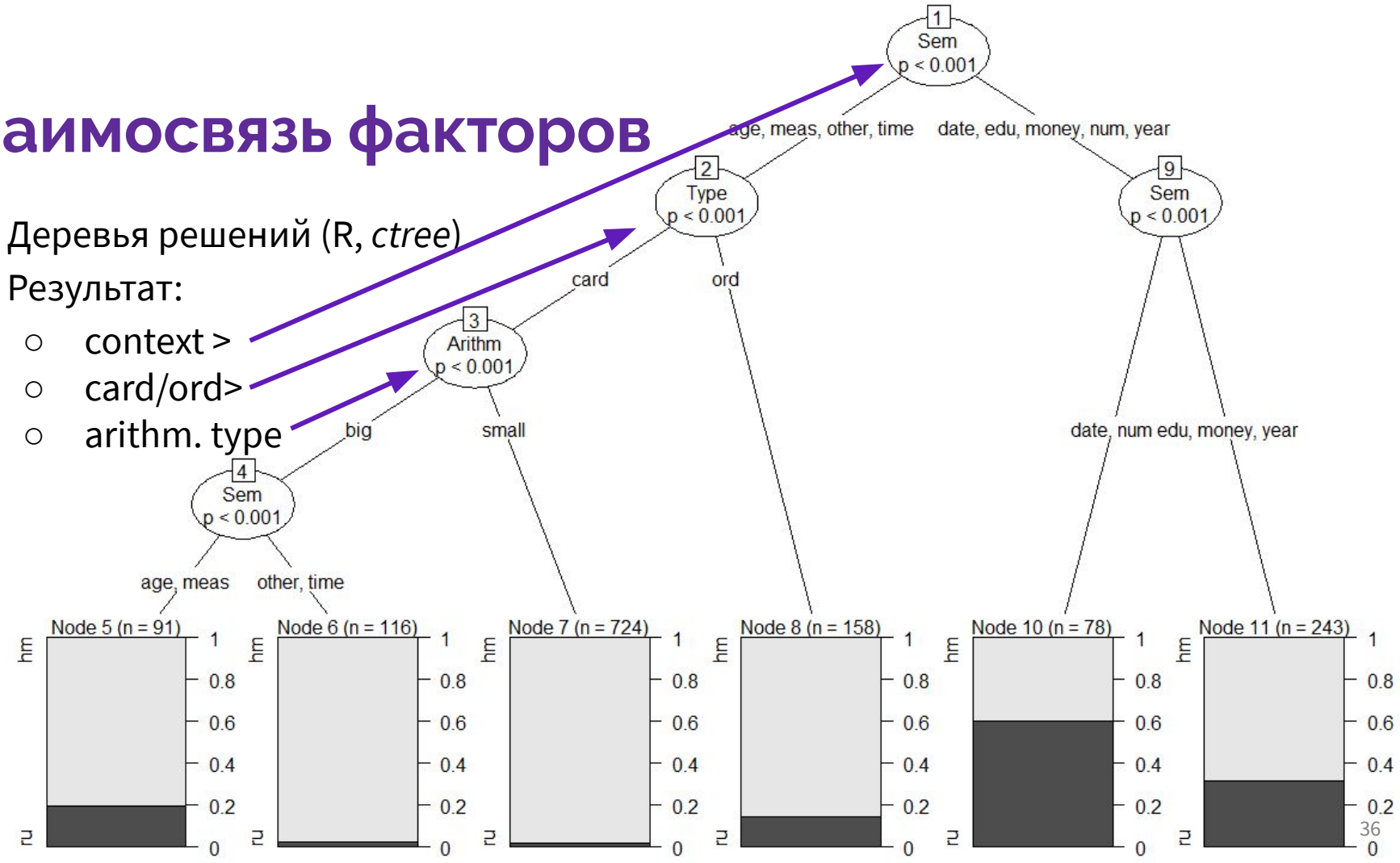
Взаимосвязь факторов

- Деревья решений в R (*rpart*)
- Точность: 0.879
- Результат: влияние контекста (date → ru)
- Проблемы: результат зависит от метода + корреляция между факторами (*year: ord & large*)



Взаимосвязь факторов

- Деревья решений (R, *ctree*)
- Результат:
  - context >
  - card/ord>
  - arithm. type



Выводы

- Каждый из видов исследования позволяет выявить разные аспекты переключения кодов
  - Подкорпус с большим количеством переключения кодов — выявить базовые ограничения ПК
  - Метод элицитации связных текстов — получать контексты, в которых носители никогда не признались бы при обычной элицитации
  - Большой корпус горномарийского языка — строить статистические модели