Исследование вариативности в русской речи билингвов на корпусном материале

Н. М. Стойнова, stoynova@yandex.ru

Международный съезд учителей и преподавателей русской словесности
МГУ, 11-13.11.2021

Исследование поддержано грантом РНФ № 17-18-01649

Введение

Введение

- Русская речь носителей миноритарных языков России:
  - исследования на материале корпусов (больших аннотированных коллекций текстов)
- Корпус русской речи носителей языков Севера Сибири и Дальнего Востока (http://web-corpora.net/ruscontact/)
- В этом докладе: речь носителей южно-тунгусских языков: нанайского и ульчского
- Языки под угрозой исчезновения: используются только старшим поколением. Все носители свободно владеют русским языком и активно его используют.
- При этом: русская речь самых пожилых носителей заметно отличается от речи монолингвов

Введение

(1) Езжайте Булаву/ Спросите людям/ где Черный озер\ (vsg)

- В корпусе размечены грамматические особенности, предположительно связанные с копированием грамматических структур родного языка или с нестандартным усвоением русской системы (напр., опущение предлога, необычная модель управления и нестандартный выбор типа склонения и согласования по роду в (1))
- В речи разных рассказчиков разные типы таких явлений представлены неоднородно.
- Разметку корпуса можно использовать для исследования такого рода вариативности между рассказчиками.
- Какие грамматические особенности в этом смысле более стабильны (одинаково представлены в речи всех рассказчиков), а какие менее; какие из них оказываются похожи, т.е. частотны/низкочастотны в речи одних и тех же рассказчиков?
- И наоборот: на какие группы разбиваются говорящие по набору характерных для них грамматических контактных явлений, насколько эти группы соответствуют интуиции и коррелируют с социолингвистическими характеристиками?

Введение

→ В этом докладе:

- попытка оценить такого рода вариативность
- статистическими методами
- на основании существующей корпусной разметки

→ За пределами доклада:

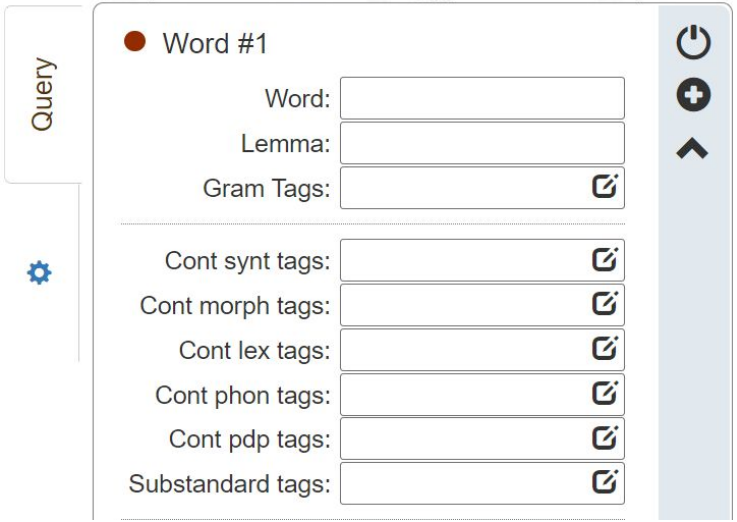
- возможности практического использования самого корпуса / такого рода исследований при преподавании русского языка детям-билингвам?
- для нанайского и ульчского сообщества: скорее неактуально (дети - носители русского, монолингвы), в других ситуациях: возможно

В рамках проекта: “Динамика языковых контактов в циркумполярном регионе” (ИЯз РАН, рук. О. В. Ханина, В. Ю. Гусев)
Подготовка корпуса: П. С. Плешак, Н. М. Стойнова, И. А. Хомченкова

Корпус

THE CORPUS OF CONTACT-INFLUENCED RUSSIAN

of Northern Siberia and The Russian Far East



- http://web-corpora.net/ruscontact/corpus.html
- небольшая коллекция звучащей русской речи
- от двуязычных носителей языков Севера Сибири и Дальнего Востока (в основном самодийские и тунгусо-маньчжурские)
- автоматическая морфологическая разметка
- ручная разметка **“контактных явлений”**
- Объем:
  - всего: **262 159** слов
  - доступен поиск по контактным явлениям: **180 105** слов

Разметка контактных явлений

- **морфология** (словообразование, словоизменение, использование грамматических категорий): 10 помет
- **синтаксис** (простое предложение): 24 пометы: наиболее детальная разметка, чаще всего используется
- **pdp** (остальное): 4 пометы (полипредикация) + 5 помет (дискурс) + 3 пометы (интонация)
- **фонетика**: 18 помет (NB не последовательная, а иллюстративная разметка)
- **лексика:** 3 пометы

⇒ 67 помет

- **“substandard”** (не контактные явления: диалектное, региональное, связанное со стилем речи): 6 помет

Контактные явления: в этом исследовании

- **Только морфосинтаксические**:
  - наиболее последовательно размечены (ср. фонетические с “иллюстративной разметкой”)
  - менее чувствительны к размеру корпуса (ср. лексические)
  - взяты только частотные (≥100 вхождений)
  - некоторые объединены / модифицированы
  - **всего 10** контактных явлений

Контактные явления: в этом исследовании (10)

FEATURE	COMMENT
agr_gender_all	объединено: agr_adj_gender&agr_verb_gender&agr_anaph_gender
agr_num_all	объединено: agr_adj_num&agr_verb_num&agr_anaph_num
asp	
bare_form	объединено и частично переразмечено: gov_caseless&agr_adj_case&pp (частично)
gov	
gov_dom	
infl_n	
num	
number	
prep_drop	
[refl]	исключено: слабо коррелирует с остальными (возможно, некорректно размечено)

Контактные явления: примеры

- agr_gender_all

(1a) А гадюка ядовитый (ekx) ← agr_adj_gender
(1b) Папа рыбачила\ (fna) ← agr_verb_gender
(1c) Какая-то шапка\ у него было, у папина маме (fna) ← agr_anaph_gender

- agr_num_all

(2a) Так парни\ молодой\ парни (vsg) ← agr_adj_num
(2b) Некоторые остался\ тут, ну/ (fna) ← agr_verb_num
(2c) Мама давай\/ его <японцев> кормить (fna) ← agr_anaph_num

Контактные явления: примеры

- asp

(1a) Пораньше звонила/ бы - я бы бода\ варила бы (fna)
(1b) Будем мы его разделить\ (vsg)

- bare_forms

(2a) И там... Харпичане умер от рак\ (vsg) ← pp
(2b) Школа\ стала ходить (vsg) ← gov_caseless
(2c) Тоже хорошая рыбу\ ели\ (eia) ← agr_adj_case

Контактные явления: примеры

- gov

(1a) На оморочке сел/ и поехала\ (vsg)
(1b) Мы туда рыбачили\/ (fna)
(1c) С амурскими отличается\ (lfs)

- gov_dom

(2) Как нарта\ таскать тяжело\ (oek)

- infl_n

(3a) Комсомольский билета девки спрятали/ (fna)

(3b) Она только уборщицем работала/ (fna)

Контактные явления: примеры

- num

(1) А пальцы ... две пальцы\ (pld)

- number

(2a) Мы месте\ китайцем жили тут (fna) = ‘вместе с китайцами’
(2b) Ну раньше\ как ... в наших дееетствах\/ (ekx)

- prep_drop

(3a) А сейчас люди __ моторе\ ходят (eia)
(3b) __ Сорок первом году начАла война/ (fna)

Тексты и рассказчики

Тексты:

- тунгусские записи: от носителей нанайского и ульчского языков (Хабаровский край: вдоль р. Амур)
- только тексты с разметкой контактных явлений
- 54 318 словоупотреблений (ок. 10 часов)

Рассказчики:

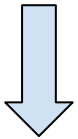
- старшего поколения: 1917-1957 г.р.
- 27 рассказчиков

Анализ

Анализ

→ **Анализ основных компонентов** (Principal Component Analysis, PCA)

- показывает как группируются контактные явления = какие из них одинаково представлены в речи одних и тех же рассказчиков
- декомпозиция переменных (контактные явления) в многомерное пространство “основных компонентов” (осей), таких что они оптимальным образом описывают вариативность между субъектами (рассказчиками)



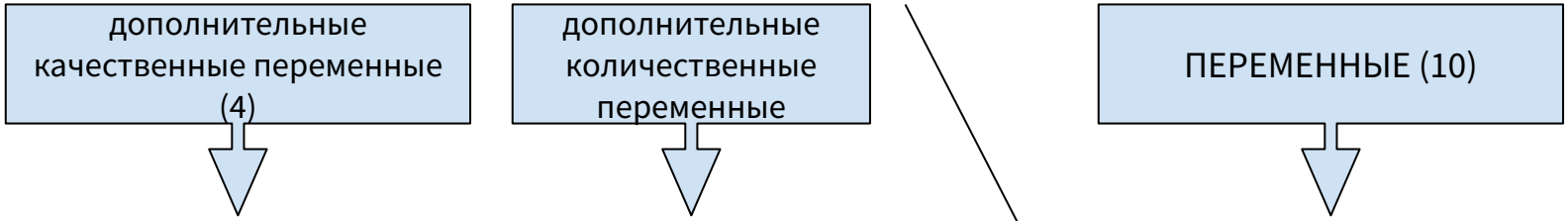
→ **Иерархическая кластеризация на основе принципиальных компонентов** (HCPC)

- кластеризация рассказчиков: для каких рассказчиков характерны одни и те же контактные явления

Анализ

Переменные:

- 10 частотных морфосинтаксических контактных явлений
- agr_gender_all, agr_num_all, asp, bare_form, gov, gov_dom, infl_n, num, number, prep_drop
- нормализованные частоты: N вхождений на 1 000 сл.-употр. (для каждого носителя)



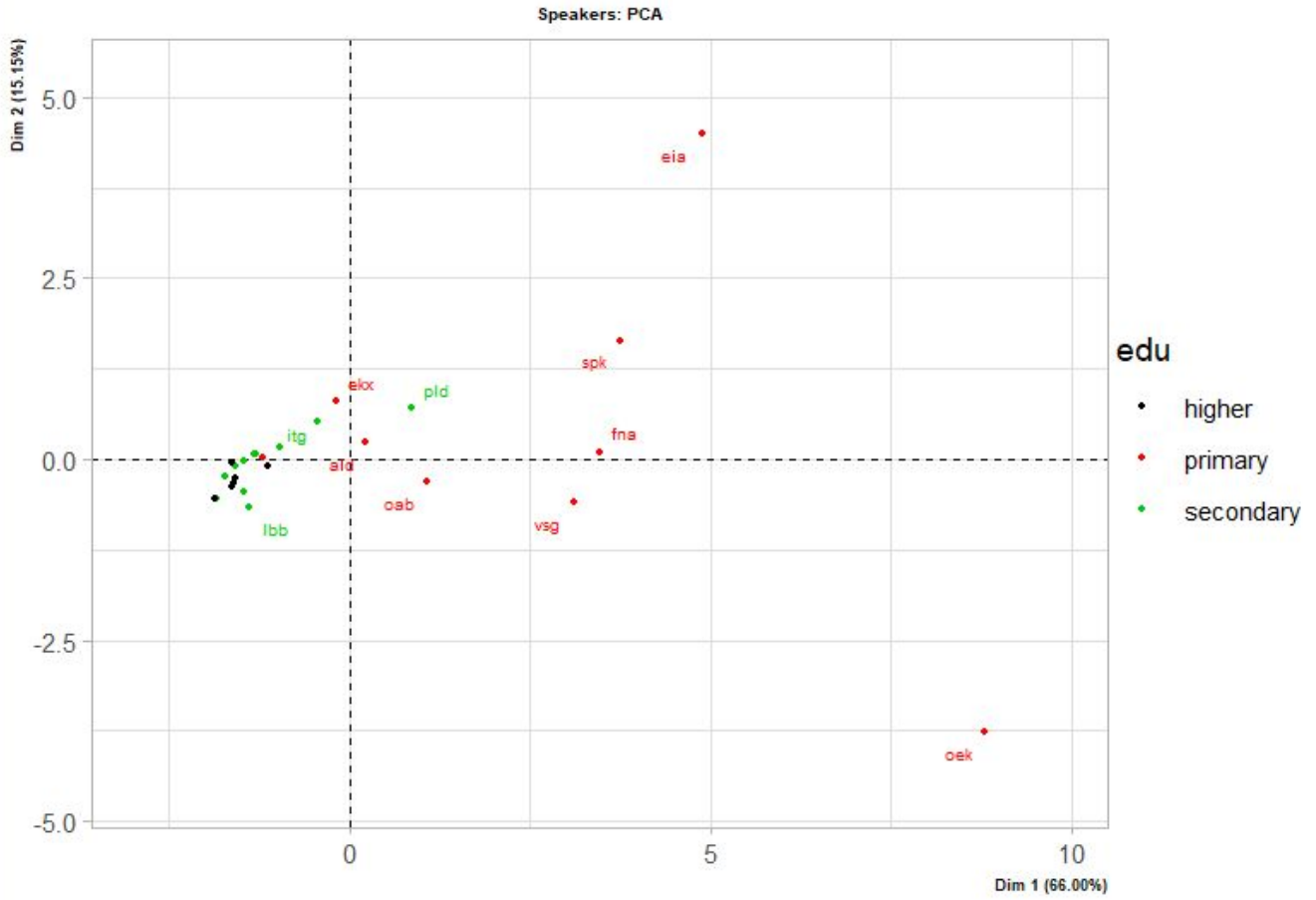
speaker	pob	sex	edu	lang	yob	tokens	agr_gender	agr_num_a	bare_form	gov
gpb	Dzhonka	F	primary	Nanai	1935	297	0	0	3.367003367	3.367003367
oab	Dudi	F	primary	Ulcha	1935	4608	16.71006944	3.038194445	2.821180556	5.425347222
lpd	Aori	F	higher	Ulcha	1939	148	0	0	0	0
rchk	Achan	F	secondary	Nanai	1942	1740	2.873563219	0.5747126437	0	4.597701149
itg	Sira camp	F	secondary	Nanai	1945	262	0	0	0	11.45038168

Результаты: PCA

Рассказчики

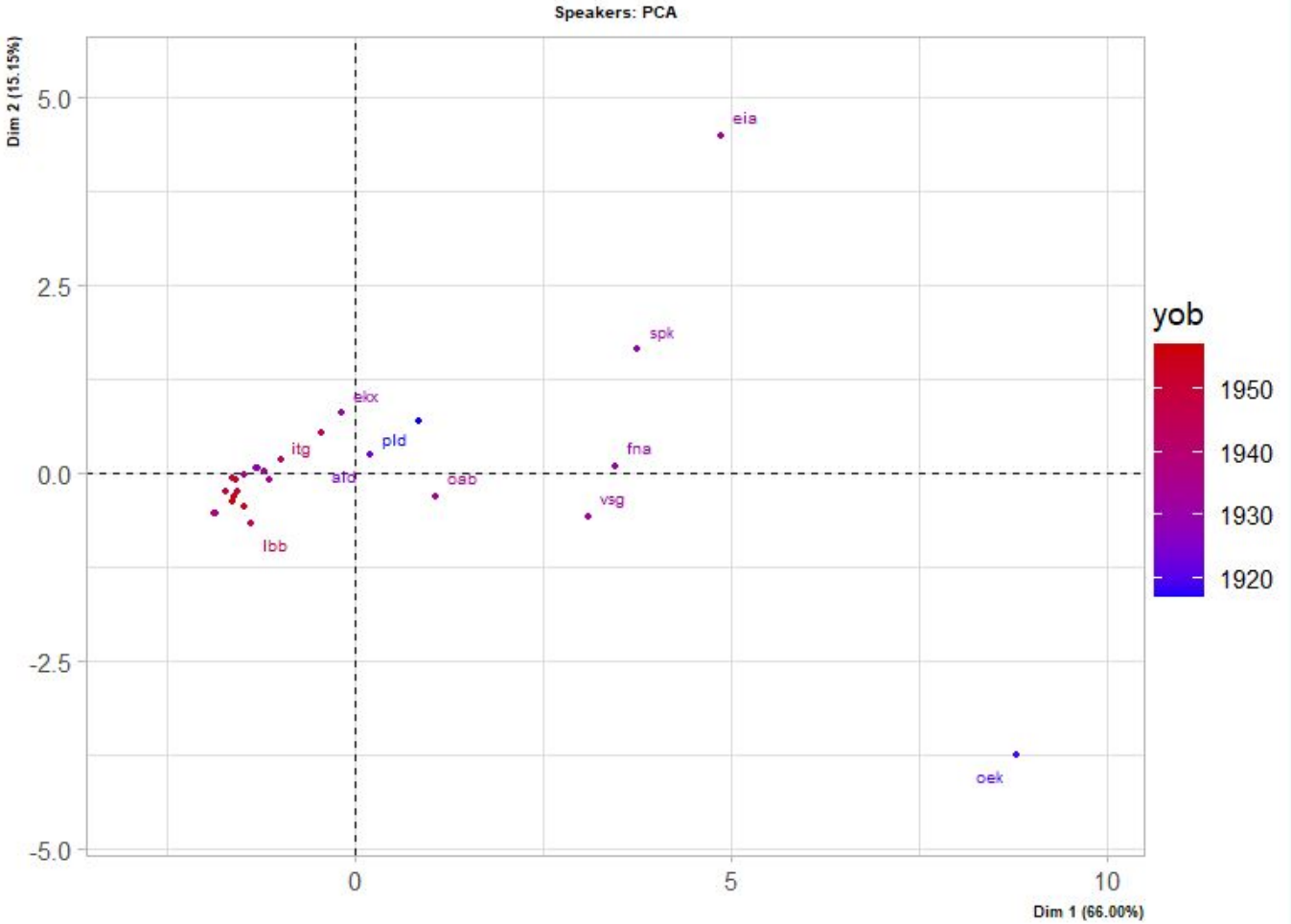
→ несколько рассказчиков, которые заметно отличаются от остальных и друг от друга vs. большая однородная группа рассказчиков

→ видимая корреляция с уровнем образования



Рассказчики

→ видимая корреляция с годом рождения

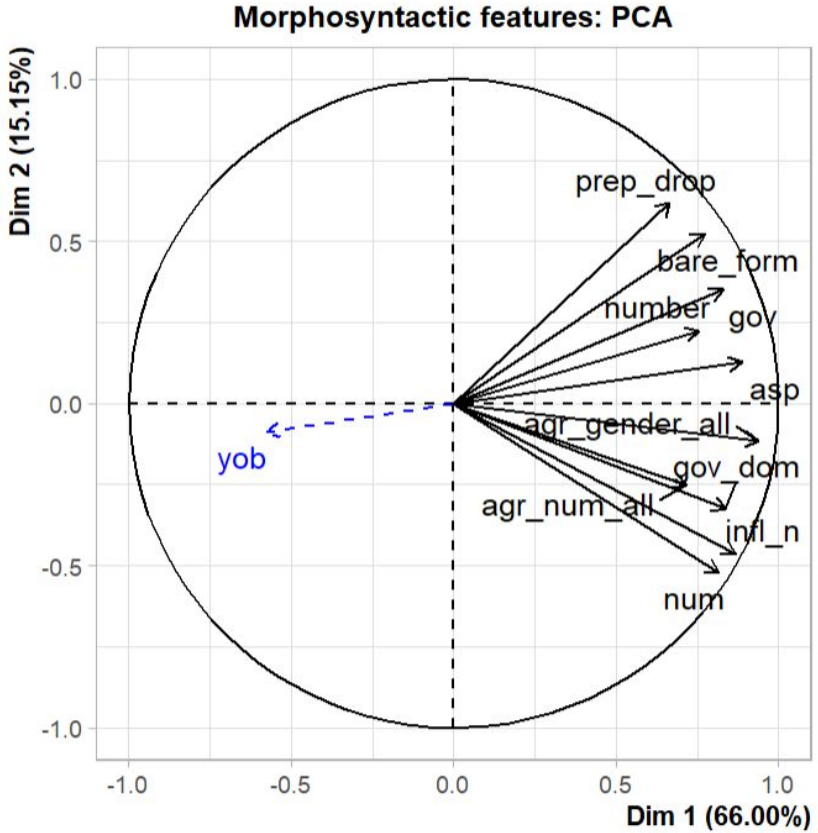


Контактные явления: морфосинтаксис

→ разные контактные явления ведут себя неожиданно сходным образом

- (т.е. у всех похожие значения по Оси 1, которая описывает больше 50% вариативности между носителями)
- здесь тоже видна корреляция с годом рождения (синий вектор yob)

ОСЬ 1 (неполное усвоение русской системы): **agr_gender_all > asp > infl_n > gov_dom > gov > num > bare_form > number > agr_num_all > prep_drop** > 0 + edu, yob



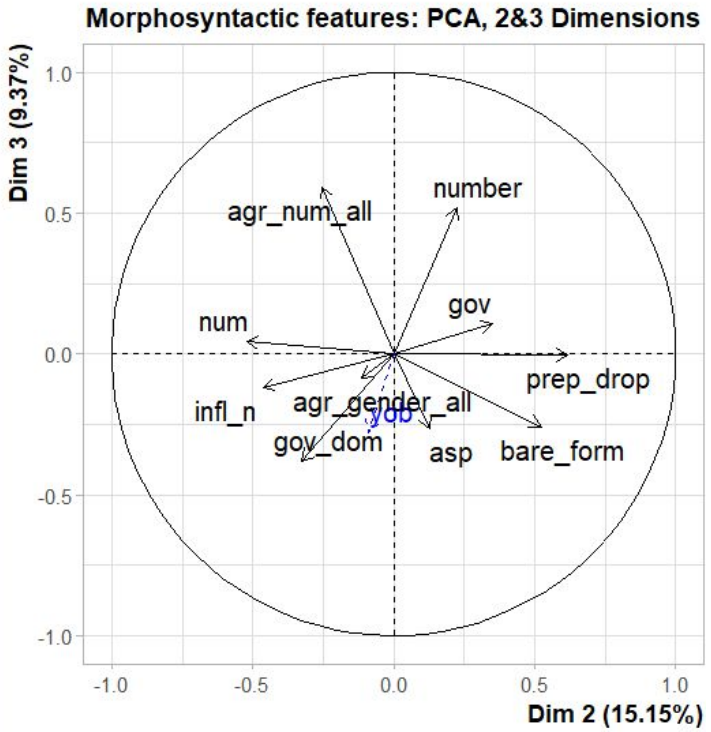
Контактные явления: морфосинтаксис

ОСЬ 2 (упрощение): **prep_drop** > **bare_form** > gov > number > asp > 0 > agr_gender_all > agr_num_all > gov_dom > **infl_n** > **num**

ОСЬ 3 (число): **agr_num_all** > **number** > gov > num > 0 > prep_drop > agr_gender_all > infl_n > asp > bare_form > **gov_dom**

→ Отчетливо не группируются явления, связанные с копированием морфосинтаксической модели (gov, num, number, ?gov_dom, ?prep_drop)

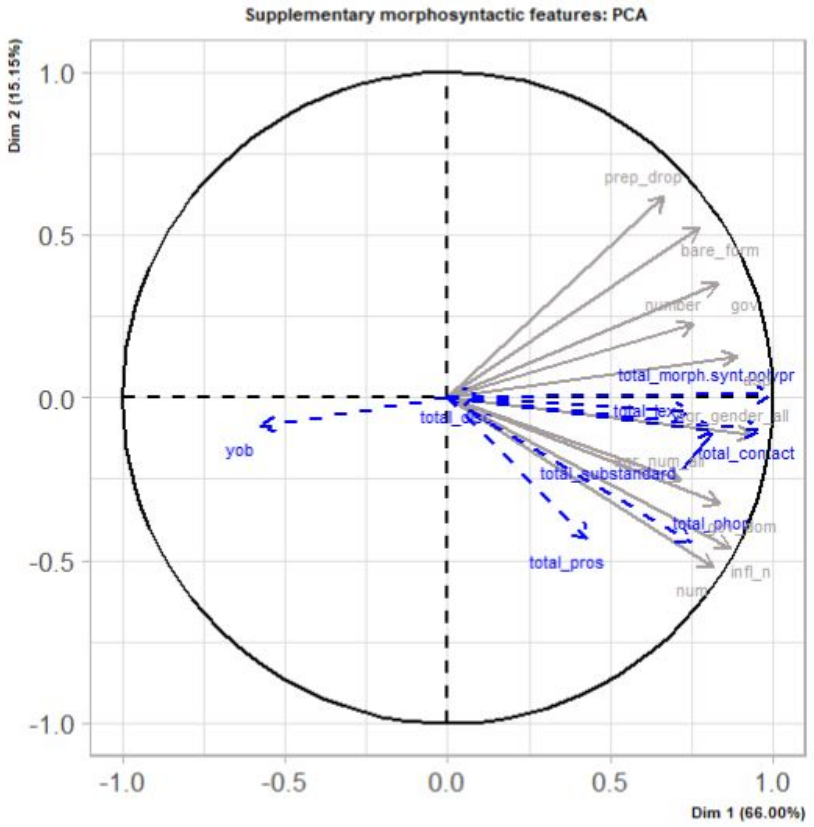
→ Почему число?



Контактные явления: прочие

NB Не вошли в анализ

- все положительно скоррелированы с Осью 1
- все отрицательно скоррелированы с Осью 2
- substandard (неконтактное): ведет себя как контактные явления



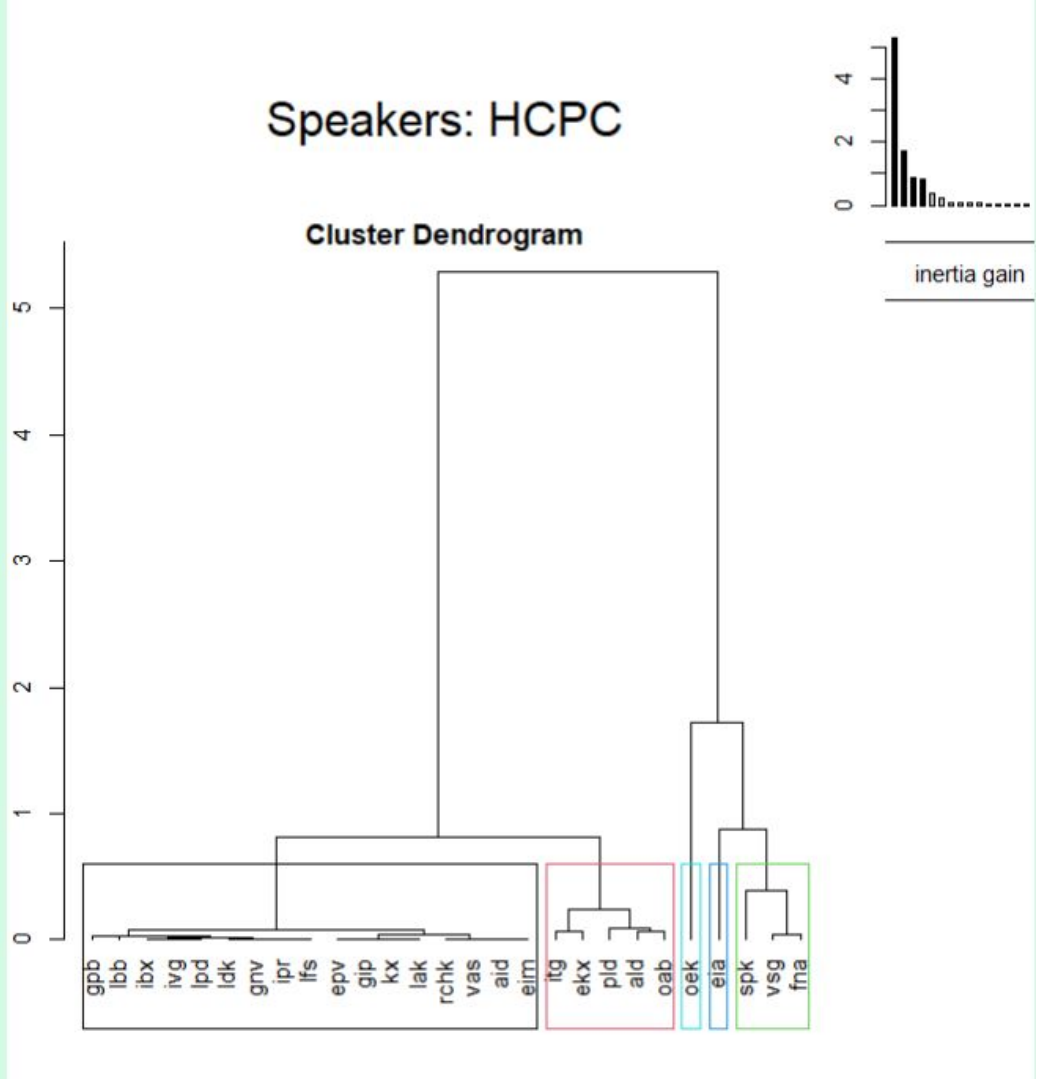
Результаты: HCPC

Рассказчики: кластеризация

⇒ Кластеризация по тому, насколько похожи у рассказчиков свойственные им наборы контактных явлений

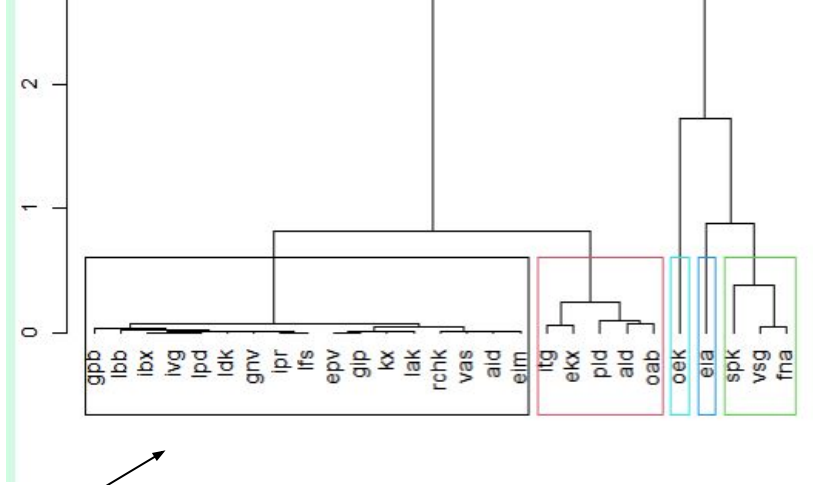
→ 5 кластеров:

1 большой кластер + 4 очень маленьких кластера



Кластеры: описание

feature	cluster 1 (v-test)
gov	-4,5
number	-4,33
asp	-3,42
agr_gender_all	-3,38
agr_num_all	-3,22
bare_form	-2,89
prep_drop	-2,81
num	-2,58
infl_n	-2,51
gov_dom	-2,05



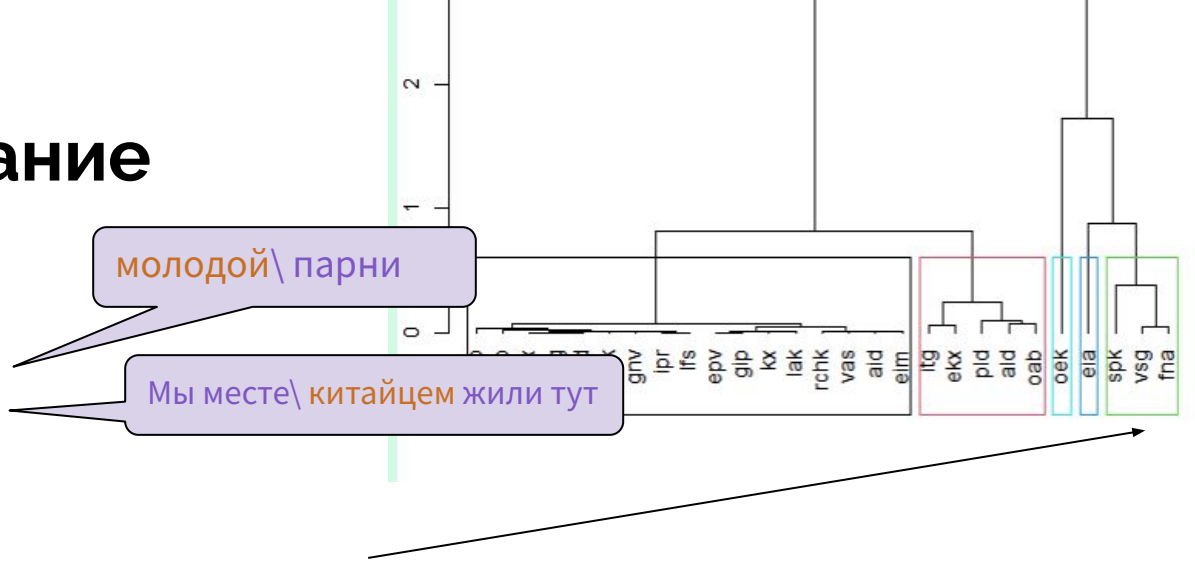
КЛАСТЕР 1: “стандартные рассказчики”

→ самый большой кластер

- значимо низкий % всех контактных явлений
- корреляция с годом рождения (молодые) и уровнем образования (+ высшее, - начальное)

Кластеры: описание

feature	cluster 2 (v-test)
agr_num_all	3,61
number	2,84
num	2,07
agr_gender_all	2,04
gov	2,01
asp	2
prep_drop	1,93
bare_form	1,91
infl_n	1,31
gov_dom	0,289



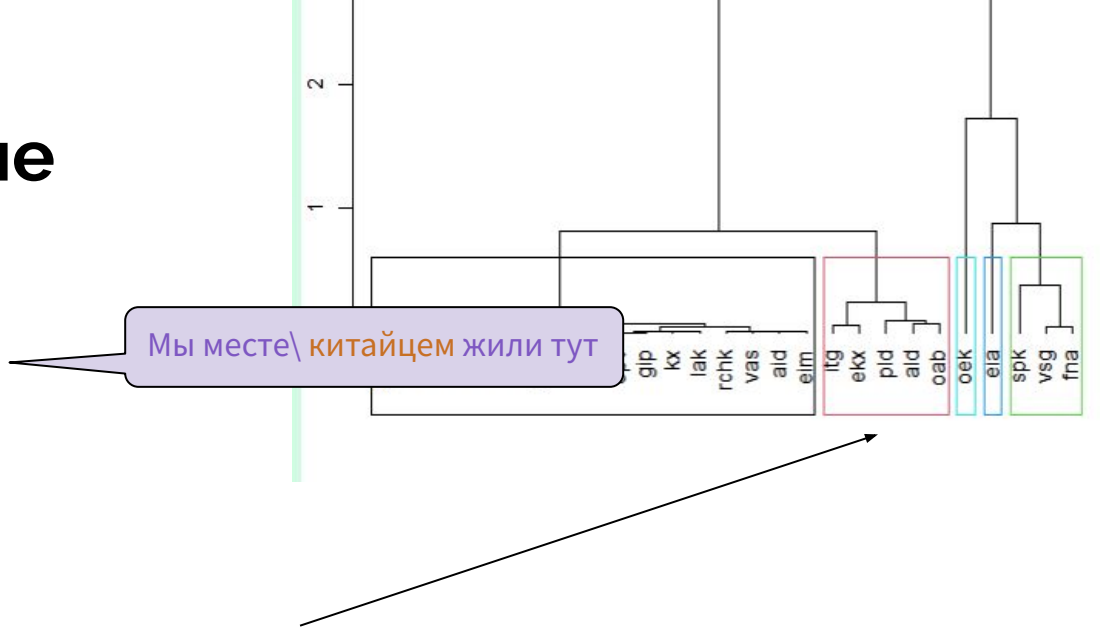
КЛАСТЕР 2: “нестандартные рассказчики”

→ spk, vsg, fna

- значимо высокий % многих контактных явлений (NB особенно arg_num и number)
- корреляция с годом рождения (старшие) и уровнем образования (начальное)

Кластеры: описание

feature	cluster 3 (v-test)
number	2,04
gov	1,95
agr_num_all	0,375
asp	0,148
agr_gender_all	0,0202
num	-0,271
prep_drop	-0,28
infl_n	-0,343
bare_form	-0,49
gov_dom	-0,719



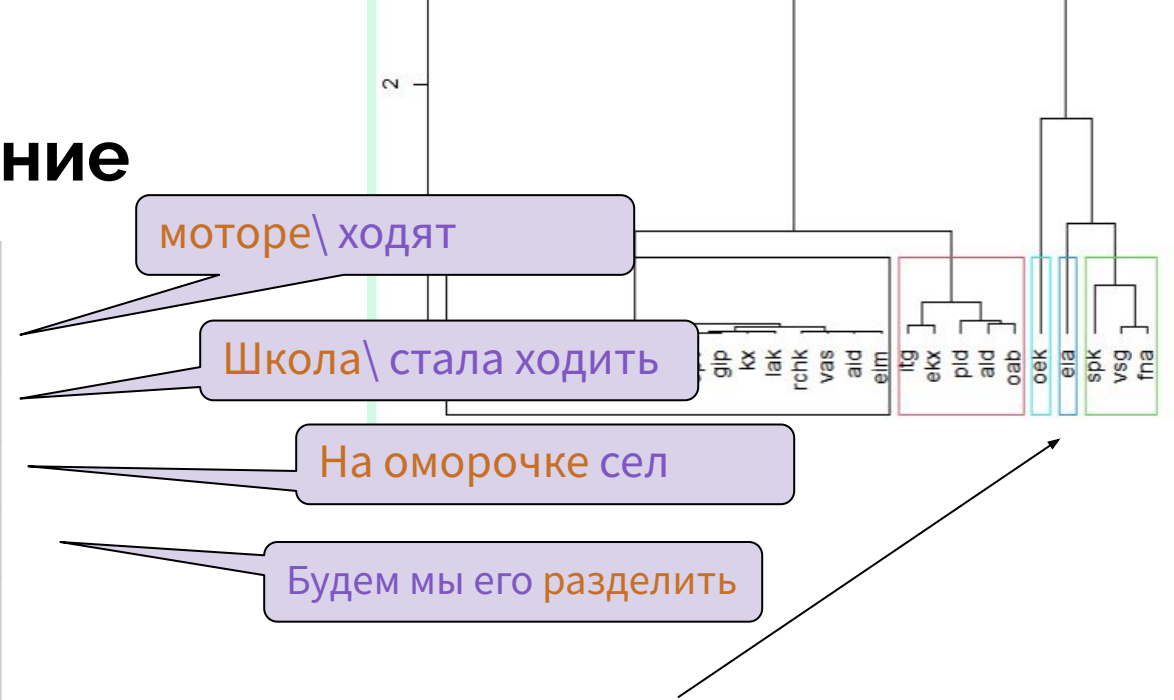
КЛАСТЕР 3: “нестандартный выбор числа”

→ itg, ekx, pld, ald, oab

- значимо высокий процент **number**

Кластеры: описание

feature	cluster 4 (v-test)
prep_drop	4,19
bare_form	3,88
gov	2,41
asp	2
agr_gender_all	1,59
gov_dom	1,47
number	1,02
infl_n	0,219
num	-0,427
agr_num_all	-0,518



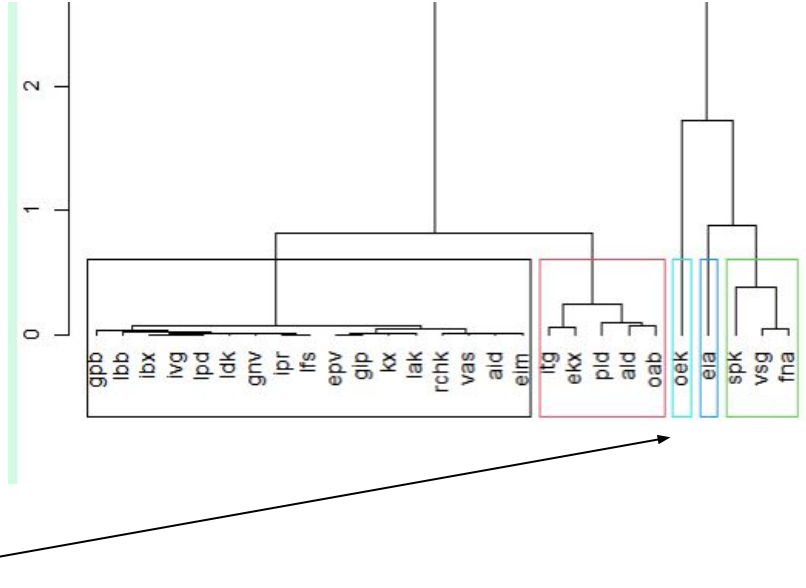
КЛАСТЕР 4: “упрощение”

→ eia

значимо высокий % prep_drop, bare_form, gov, asp

Кластеры: описание

feature	cluster 5 (v-test)
gov_dom	4,76
infl_n	4,71
num	4,15
agr_gender_all	3,63
asp	3,1
agr_num_all	1,97
gov	1,73
bare_form	1,34
number	1,13
prep_drop	0,36



КЛАСТЕР 5: “особый носитель”

- значимо высокий процент целого набора контактных явлений
- набор не такой, как для кластера 2
- не до конца понятно, что объединяет эти явления

Обсуждение

Некоторые обобщения

→ Какие грамматические особенности в этом смысле более стабильны (одинаково представлены в речи всех рассказчиков), а какие менее?

- ВСЕ РАССМОТРЕННЫЕ КОНТАКТНЫЕ ЯВЛЕНИЯ ВНОСЯТ ВКЛАД В ВАРИАТИВНОСТЬ МЕЖДУ РАССКАЗЧИКАМИ

→ Как группируются контактные явления? Какие из них оказываются похожи, т.е. частотны/низкочастотны в речи одних и тех же рассказчиков?

- РАЗНЫЕ КОНТАКТНЫЕ ЯВЛЕНИЯ ВЕДУТ СЕБЯ ДОСТАТОЧНО ПОХОЖЕ
- НАМЕЧЕННЫЕ ГРУППЫ ЧЕРТ: СВЯЗАННЫЕ С НЕПОЛНЫМ УСВОЕНИЕМ / С УПРОЩЕНИЕМ / ЧИСЛО И СОГЛАСОВАНИЕ ПО ЧИСЛУ (ПОЧЕМУ?)
- но не очень отчетливые тенденции
- неожиданно не группируются ЯВЛЕНИЯ, СВЯЗАННЫЕ С КОПИРОВАНИЕМ МОДЕЛИ

Некоторые обобщения

→ Как кластеризуются рассказчики (носители) по характерным для их речи контактным явлениям?

- БОЛЬШАЯ ГРУППА “СТАНДАРТНЫХ” НОСИТЕЛЕЙ VS. НЕСКОЛЬКО МАЛЕНЬКИХ ГРУПП С БОЛЬШОЙ ВАРИАТИВНОСТЬЮ МЕЖДУ НОСИТЕЛЯМИ

→ Соответствуют ли получившиеся кластеры исследовательской интуиции?

- ДА, САМИ КЛАСТЕРЫ ВЫГЛЯДЯТ ОЧЕНЬ ЕСТЕСТВЕННО
- НО, ГРАММАТИЧЕСКИЕ ЯВЛЕНИЯ, РЕЛЕВАНТНЫЕ ДЛЯ КЛАСТЕРИЗАЦИИ, ВЫГЛЯДЯТ БОЛЕЕ НЕОЖИДАННО

→ Есть ли корреляции с социолингвистическими параметрами?

- ДА: ГОД РОЖДЕНИЯ, УРОВЕНЬ ОБРАЗОВАНИЯ

Дополнительные слайды

Problems: annotation of contact features

→ Does the existing corpus annotation capture the degree of deviation from monolingual benchmark and inter-speaker variation well?

- Basically, the annotation reflects the “locus” of the feature and not its “nature”

some tags cover more or less homogenous features (e.g. NEG), some other do not

for instance, GOV = something non-standard in argument encoding

(1) Клей делали... этот... **кета\ шкурой** (vsg) - INS instead of iz + GEN: copying? (INS in Tungusic)
(2) **Хабаровске**\ ездит, ну\/ (fna) - LOC instead of *v* + ACC: simplification?
(3) **С амурскими** отличается\ (lfs) - *s* + INS instead of *ot* + GEN: incomplete acquisition? (cf. с русскими одинаковый)

Problems: annotation of contact features

→ Does the existing corpus annotation capture the degree of deviation from monolingual benchmark and inter-speaker variation well? -- TO RE-ANNOTATE data for this specific study?

- The “nature” of some features reflected in the annotation is unclear

for instance, NUM = copying or simplification?

(1) Вот там **пять дом**/ (vsg) NUM + NOM.SG instead of NUM + GEN.PL

NUM.SG is expected in Nanai ⇒ copying?

NUM.SG is simpler

- The current annotation contains some errors… TO FIX!

Problems: lack of data

A very small number of speakers

Very few data from each speaker

Does the PCA-analysis give correct results?

Maybe, some other methods will work better?