

Автоматическое определение степени семантических изменений

НИС «Компьютерная семантика», 2022

Даша Попова, Даша Рыжова

Группа задач, связанных с семантическими сдвигами

- Поиск метафор
- Поиск метонимий
- Снятие семантической неоднозначности (Word Sense Disambiguation)
- Поиск семантической неоднозначности (Word Sense Induction)

+ Определение степени семантических изменений (semantic shift evaluation, semantic change detection) – насколько изменился набор значений слова за определенный период времени

Определение степени семантических изменений

Зачем и кому это нужно?

- Теоретической лингвистике
- Истории и культурологии (важные исторические, культурные, политические события отражаются на особенностях употребления некоторых ключевых слов, ср. *Косово* или *новичок*)
- Определение авторского стиля
- Определение особенностей текстов разных жанров
- Оценка степени адекватности языковой модели?..
- И т.п. ...

В теории

- В.В. Виноградов «История слов» (1999)
- Н.Р. Добрушина и др. «Два века в двадцати словах» (2016)
- Лексика меняется очень быстро и по очень разнообразным сценариям. Часто кажется, что пути этих изменений неисповедимы. Можно ли говорить о каких-то общих законах?
- Если такие законы существуют, то можно ли их применять в задачах NLP?

О законах

(Hamilton, Leskovec, Jurafsky, 2016): есть ли связь между частотностью, многозначностью и склонностью к изменению?

- Материал: 6 диахронических корпусов для английского, французского, немецкого и китайского (временной охват – ок. 200 лет)
- Методы: PPMI, SVD, word2vec
- Каждый корпус разбит на декады, для каждой декады – своя модель
- Степень семантического изменения оценивается через косинусное расстояние между векторными представлениями одного и того же слова в разные периоды

О законах

(Hamilton, Leskovec, Jurafsky, 2016): есть ли связь между частотностью, многозначностью и склонностью к изменению?

- Связь между частотностью слова и степенью его изменения к следующему временному периоду:

the *law of conformity* — the rate of semantic change scales with an inverse power-law of word frequency

О законах

(Hamilton, Leskovec, Jurafsky, 2016): есть ли связь между частотностью, многозначностью и склонностью к изменению?

- Связь между многозначностью слова и степенью его изменения к следующему временному периоду:

the *law of innovation* — independent of frequency, words that are more polysemous have higher rates of semantic change

(определение степени многозначности через контекстное разнообразие: кластеризация ближайших соседей слова в зависимости от того, являются ли они близкими соседями друг для друга)

О законах

Dubossarsky et al. 2017

- Google N-grams, N = 5
- вектора сочетаемости, окно ± 2
- разбиение на декады (от 1900 до 2000)
- кластеризация с указанием расстояния от центра кластера (2000 кластеров)

Cluster 1, dist	Cluster 2, dist	...
chamber, 0.04	shutters, 0.04	
room, 0.04	windows, 0.05	
drawing, 0.05	doors, 0.08	
bedroom, 0.06	curtains, 0.1	
kitchen, 0.07	blinds, 0.11	
apartment, 0.1	gates, 0.13	

О законах

Dubossarsky et al. 2017

- Косинус между векторами для одного и того же слова, но в разные временные периоды
- Корреляция между степенью изменения и расстоянием от центра кластера
- Чем больше расстояние от центра кластера, тем выше вероятность того, что в следующий временной период произойдет сдвиг значения
- Глаголы меняются быстрее, чем существительные

Как определить степень сдвига

- Дистрибутивные модели как основной метод
- Контекстуальные лучше (ELMo, BERT)
- Видимо, state of the art на данный момент – supervised методы (предобученный BERT, модели с использованием словарей)
- См. **Kutuzov 2020**

Определение степени
семантических изменений для
русского языка: RuShiftEval-2021

Соревнование на Диалоге-2021 (RuShiftEval)

- Датасеты:

Корпуса: три подкорпуса НКРЯ: досоветский (1700-1916), советский (1918-1990) и постсоветский (1991-2016)

Наборы слов: тренировочный (указана степень изменения), валидационный, тестовый

- Метрика оценки:

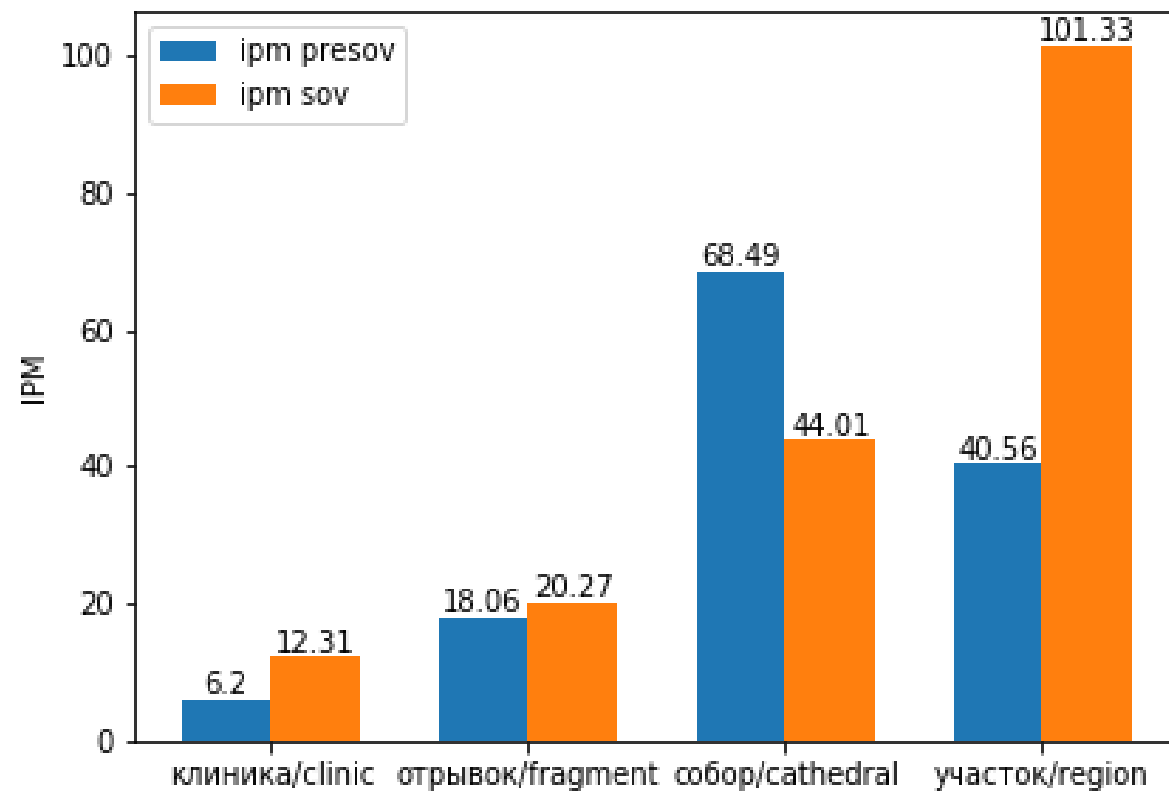
Метрика COMPARE (Schlechtweg et al., 2018)

- Задача: ранжировать слова из тестовой выборки в зависимости от степени семантических изменений, в них произошедших

- Лучшие результаты: корреляция Спирмана ~ 0.7

Корпуса

Период	Размер (в токенах)
досоветский (1700-1916)	73542513
советский (1918-1990)	95043479
постсоветский (1991-2016)	83269542



Датасеты

Тренировочный	
RuSemShift1	44 сущ.
RuSemShift2	43 сущ.

RuSemShift1:

досоветский – советский

RuSemShift2:

советский – постсоветский

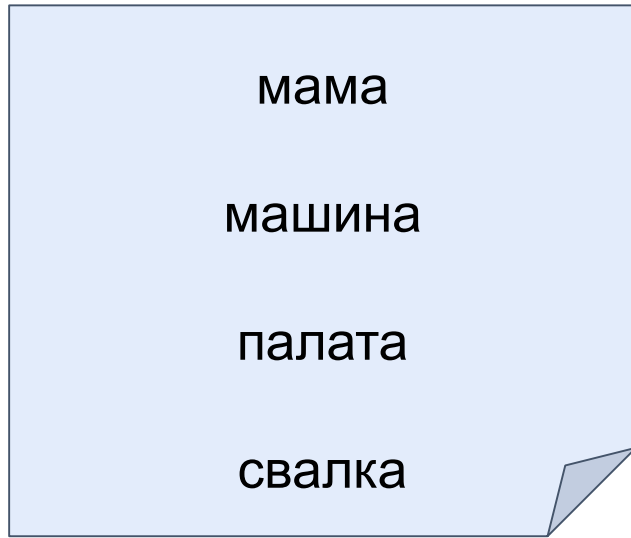
RuSemShift3:

досоветский – постсоветский


Валидационный		
RuSemShift1	RuSemShift2	RuSemShift3
12 существительных		

Тестовый		
RuSemShift1	RuSemShift2	RuSemShift3
99 существительных		

Задача



ранжирование



The word 'ранжирование' (ranking) is centered above a white arrow with a black outline that points to the right.

word	score
мама	3.69
машина	2.12
свалка	1.9
палата	1.46



Оценка качества

prediction
word1
word2
word3
word4

Корреляция Спирмена



Gold standard
word2
word1
word3
word4

Золотой стандарт: метрика COMPARE

- Случайные пары предложений с одним и тем же словом из корпусов разных периодов
- Краудсорсинговые аннотаторы должны оценить, в одном и том же значении слово употреблено в двух предложениях или нет (от 1 до 4)
 - 0: Cannot decide
 - 1: Unrelated
 - 2: Distantly related
 - 3: Closely Related
 - 4: Identical

Метрика COMPARE

1. За два дня до открытия "земского **собора**" (так выражались иные о затее Керенского) это "совещание общественных деятелей" против нескольких голосов приняло резолюцию, предложенную Милюковым.

2. Их Императорские Величества следуют мимо **собора** двенадцати Апостолов из Успенского монастыря в Чудов монастырь. неизвестный.

аннотатор 1	3
аннотатор 2	1
аннотатор 3	1
аннотатор 4	1
аннотатор 5	1

среднее: 1.4

Метрика COMPARE

Пара 1: ***машина***

Удрав с Радио до срока, покатыл к старым доблестным моим морякам -- зенитчикам на Глинка-стрит, дабы просить **машину** под дрова.

На заводе, как на настоящей фабрике-кухне, весь процесс формирования продукции осуществляется в закрытых **машинах**.

Пара 2: ***богадельня***

Долго крутились мы по узким улочкам, искали **богадельню**, наконец нашли.

Моя **богадельня** сделалась вольным домом.

Методы

- Статические эмбединги: word2vec
- Контекстуализованные эмбединги: ELMo, BERT
- Грамматические профили
- Дополнительные ресурсы: словарные данные

Word2Vec

1. lemmas  emb training  Procrustes alignment 
similarity metrics

1. tokens  emb training  Procrustes alignment 



- average of all word forms
- the most frequent word form
- the most changed word form



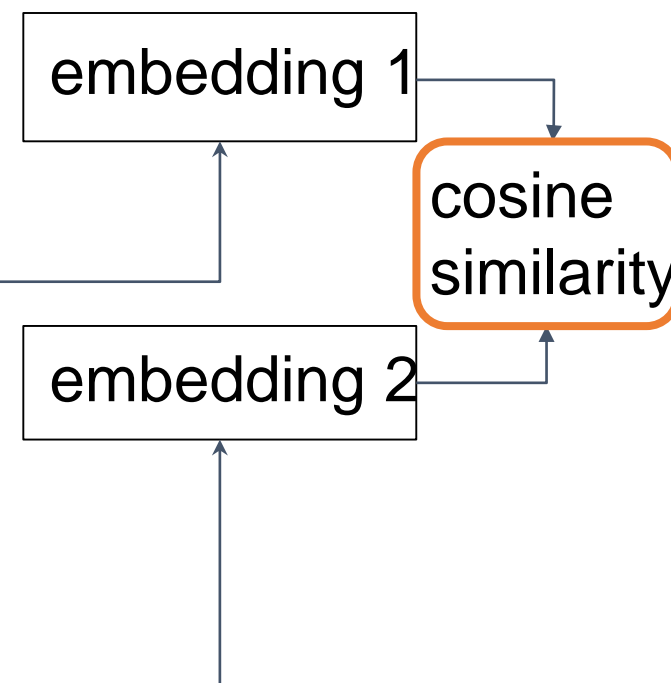
similarity metrics

ELMo, BERT

Пары предложений из разных временных периодов;
по 100 случайных пар

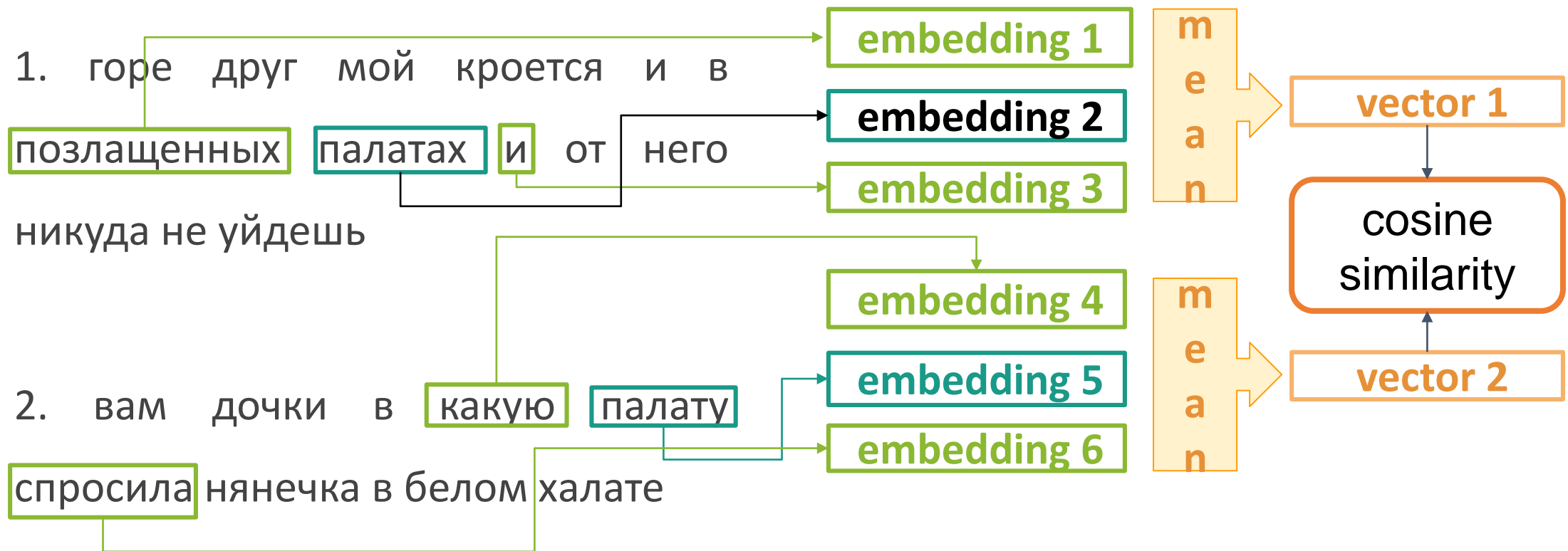
1. по нашему мнению лучше всего употреблять кислое молоко
приготовленное при помощи чистых культур молочнокислых
бактерий а также эти культуры в виде мягкой **мази** которую
можно смешивать с вареньем

2. но лавочник иван семенов еще торговал с черного хода твердыми
как камень мятными пряниками ландрином и колесной **мазью**



ELMo, BERT: с ближайшим контекстом

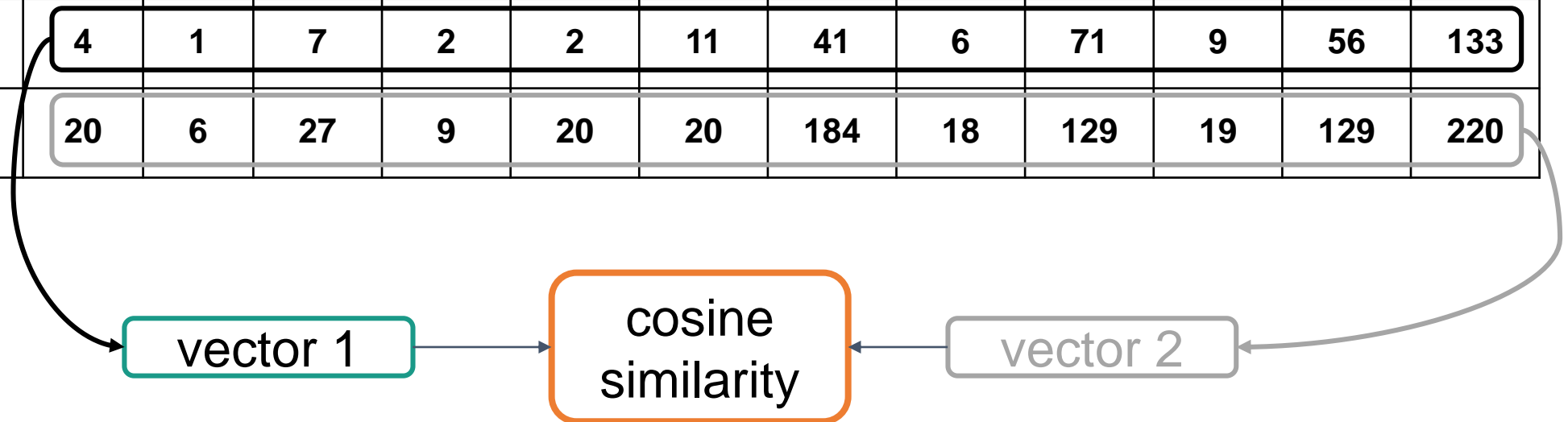
Пары предложений из разных временных периодов; по 100 случайных пар;
ближайший контекст длины 1



Грамматические профили

Грамматические профили для слова *свалка*:

Число	Plural						Singular					
Падеж	Acc	Dat	Gen	Ins	Loc	Nom	Acc	Dat	Gen	Ins	Loc	Nom
досоветский	4	1	7	2	2	11	41	6	71	9	56	133
советский	20	6	27	9	20	20	184	18	129	19	129	220



Результаты (одной из частных реализаций)

Модель	Корреляция для RuSemShift 1	Корреляция для RuSemShift 2	Корреляция для RuSemShift 3
word2vec on lemmas	0.141	0.246*	0.330*
ELMo lemmas	0.469*	0.450*	0.453*
ELMo tokens + context, window = 1	0.430*	0.451*	0.469*
RuBERT	0.380*	0.429*	0.448*
grammatical vectors	0.157	0.199*	0.343*
linear regression	0.480*	0.487*	0.560*

* p-value < 0.05

Почему грамматические профили работают?

Механизмы грамматикализации (Kuteva et al. 2019):

- extension (or context generalization) – use in new contexts,
- desemanticization (or “semantic bleaching”) – loss in meaning content,
- **decategorialization – loss in morphosyntactic properties characteristic of lexical or other less grammaticalized forms**, and
- erosion (or “phonetic reduction”) – loss in phonetic substance.

Почему грамматические профили работают?

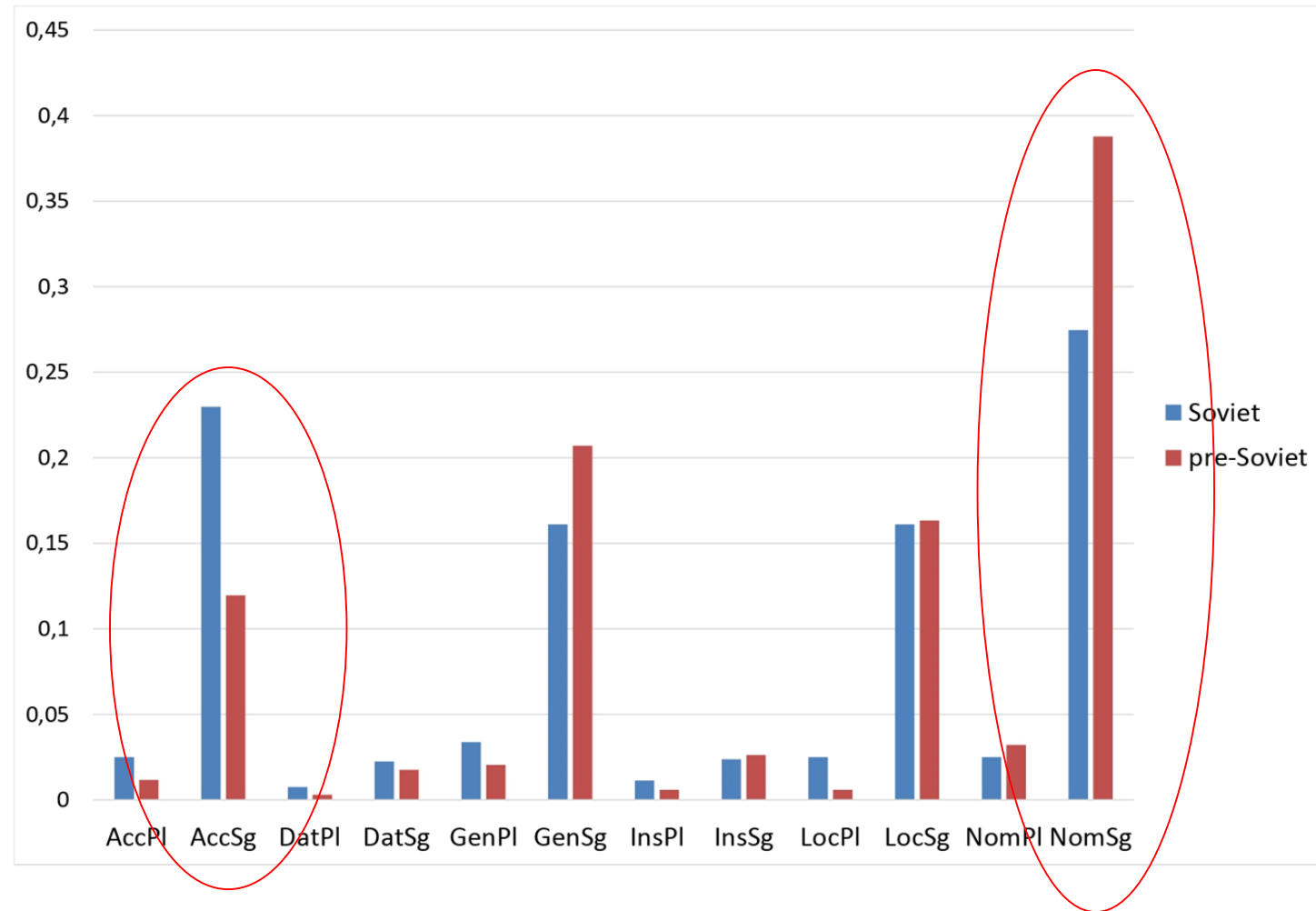
(Рахилина 2020): *почтение*

- до 18 века: действие
многия дары и почтения
- после 18 века: отношение => редукция парадигмы (потеря форм множественного числа)

(Ryzhova et al. 2021):

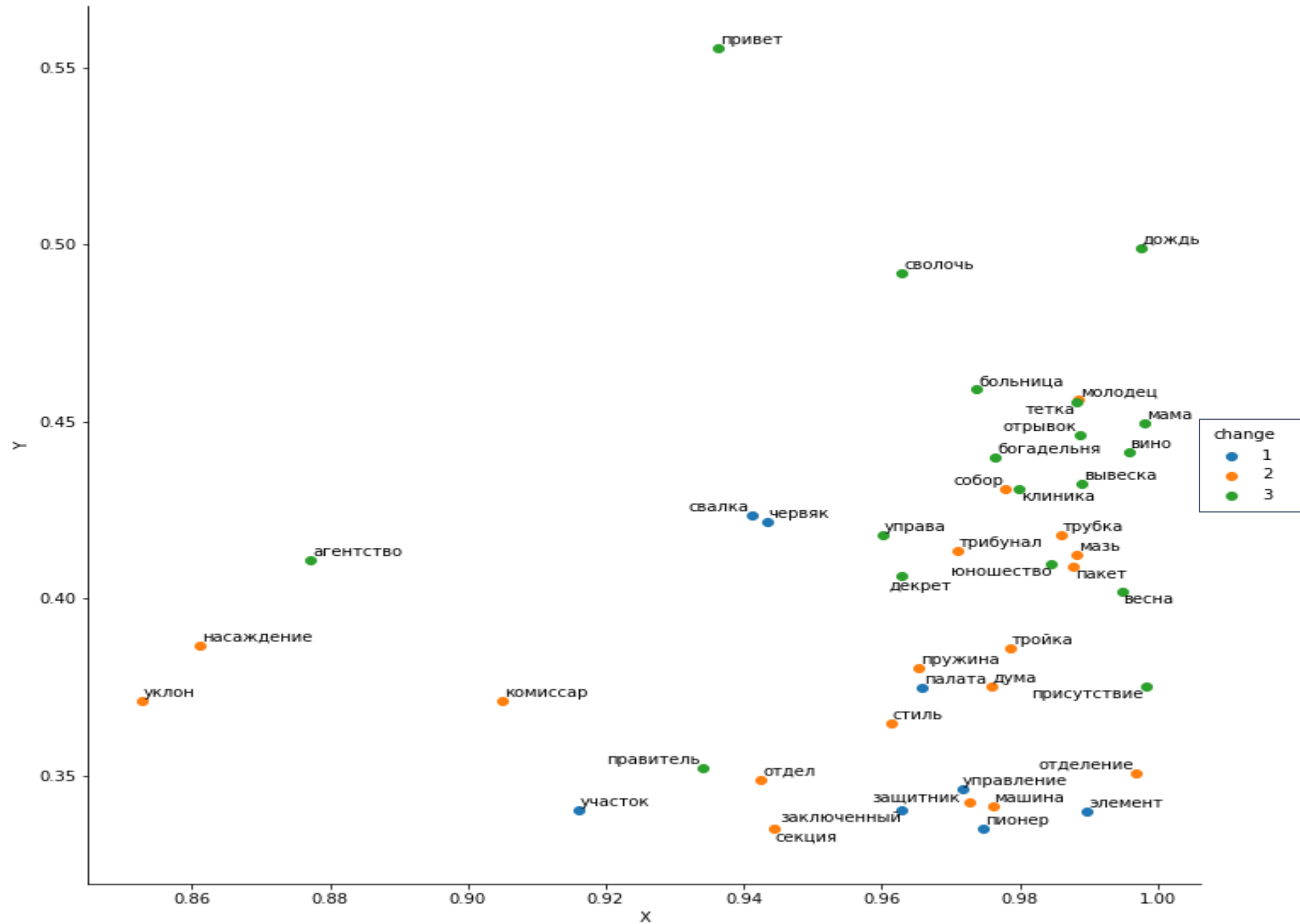
- Изменения в семантике систематически коррелируют с изменениями в грамматическом поведении
- Эти изменения часто не видны невооруженным глазом, но их можно зафиксировать статистически
- Чем длиннее временной отрезок, тем заметнее грамматические изменения

свалка

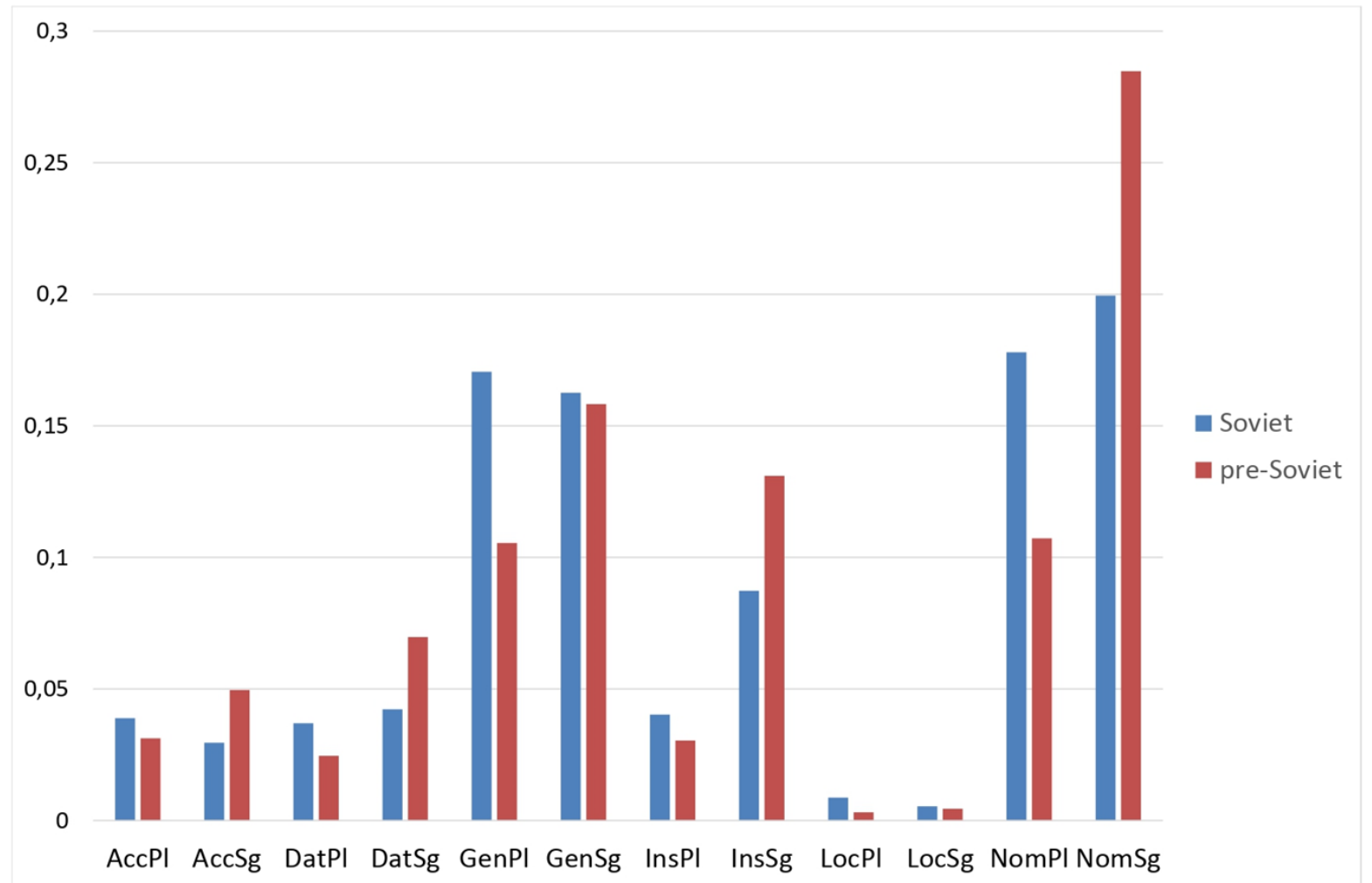


свалка

- (1) *Тотчас же закипела **свалка[NomSg]**, и десятки тел смешались в одну общую кричащую массу.*
- (2) *Говорят шефу: станок сломался. Он верит, волокут станок на **свалку[AccSg]**.*



правитель



Проблемные места

- С лингвистической точки зрения – очень много разных явлений собраны в одну кучу:
 - Разные типы сдвигов
 - Языковые vs. культурные изменения
 - Нужны ли разные методы для изменений разных типов?
- С технической: мало диахронических корпусов
- Неясно, как должна выглядеть идеальная метрика оценки