Detection of Semantic Changes in Russian Nouns with Distributional Models and Grammatical Features

Anastasiia Ryzhova

Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Lomonosov Moscow State University

Daria Ryzhova

HSE University

Ilya Sochenkov

HSE University

Introduction

- experiments within RuShiftEval competition
- team *aryzhova*
- 6th place on the leaderboard
- main focus: Russian rich morphology

Task description

мама ‘mother’

машина ‘machine’

палата ‘chamber’

свалка ‘dump’

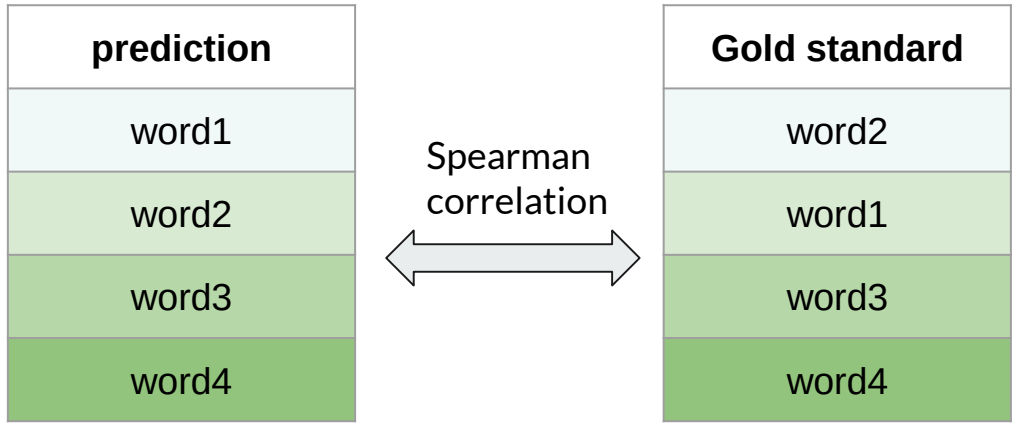
ranking

word	score
мама	3.69
машина	2.12
свалка	1.9
палата	1.46

RuShiftEval competition

Three pairs of time periods:

- pre-Soviet -Soviet (RuSemShift1)
- Soviet - post-Soviet (RuSemShift2)
- pre-Soviet-post-Soviet (RuSemShift3)



COMPARE metric

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.

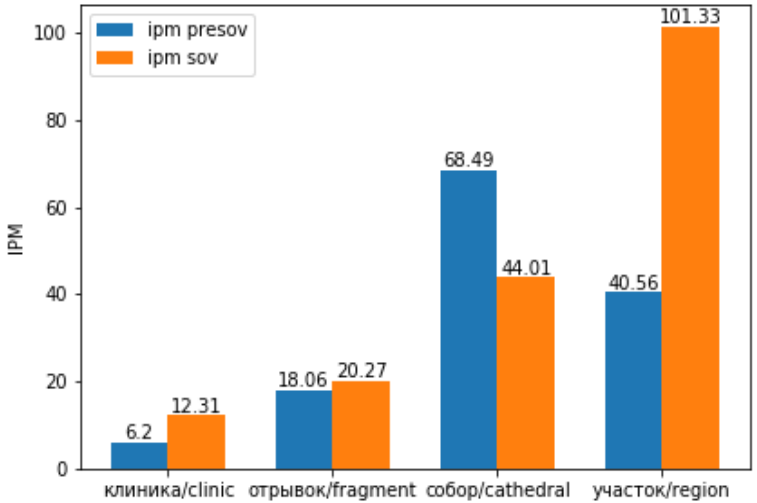
1. Их Императорские Величества следуют мимо ***собора*** двенадцати Апостолов из Успенского монастыря в Чудов монастырь. неизвестный.

annotator 1	3
annotator 2	1
annotator 3	1
annotator 4	1
annotator 5	1

mean: 1.4

Dataset: Russian Diachronic Corpora

Period	Number of tokens
Pre-Soviet (1700-1916)	73542513
Soviet (1918-1990)	95043479
Post-Soviet (1991-2016)	83269542



Evaluation phases

Train	
RuSemShift1	44 nouns
RuSemShift2	43 nouns

Development		
RuSemShift1	RuSemShift2	RuSemShift3
12 nouns		

Test		
RuSemShift1	RuSemShift2	RuSemShift3
99 nouns		

Theory and Algorithms

1. Static word embeddings: **word2vec**
2. Contextualized word embeddings: **ELMO, RuBert**
3. Grammatical vectors
4. Linear regression (two features: cosine similarities of ELMo vectors and cosine similarities of grammatical vectors)

Theory and Algorithms: word2vec

1. lemmas ⟹ emb training ⟹ Procrustes alignment ⟹ cosine similarity

1. tokens ⟹ emb training ⟹ Procrustes alignment ⟹

⟹

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

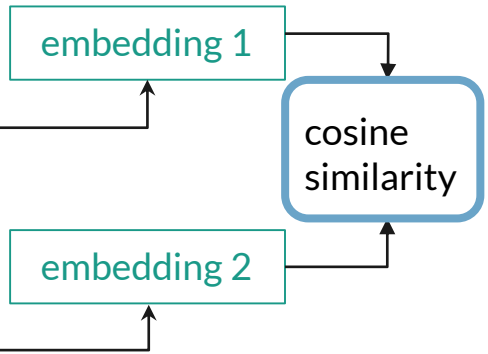
⟹ cosine similarity

Theory and Algorithms: ELMO and RuBERT

Two sentences from different time periods; 100 random pairs

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

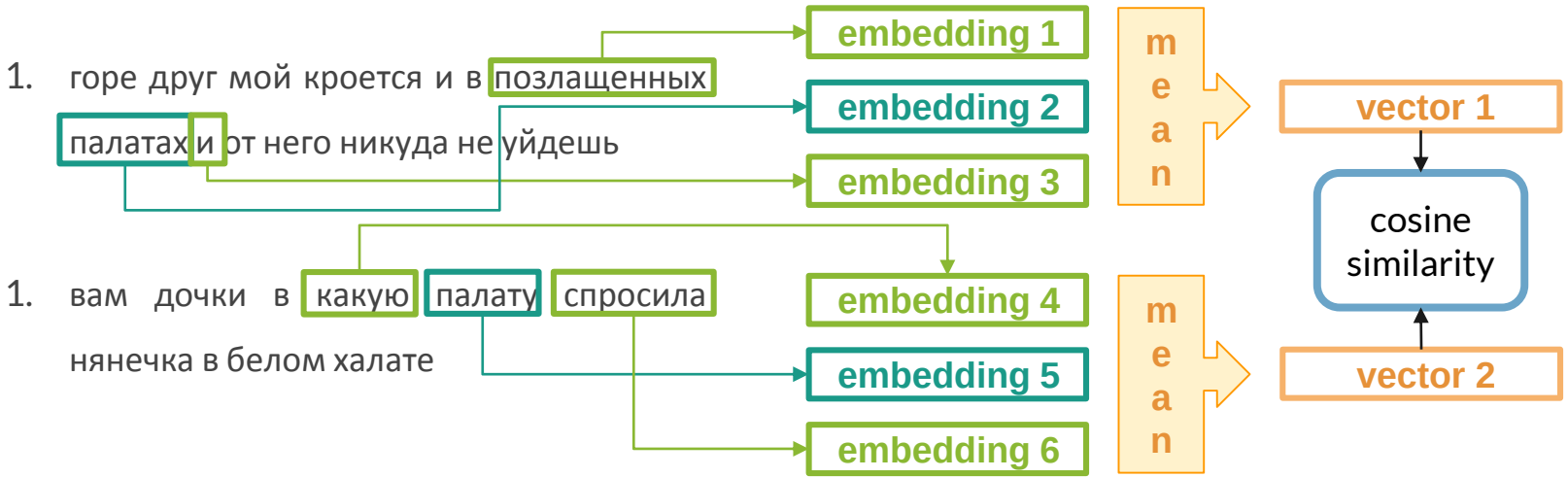
1. но лавочник иван семенов еще торговал с черного хода твердыми как камень мятными пряниками ландрином и колесной **мазью**



ruwikiruscorpora_tokens_elmo_1024_2019; ruwikiruscorpora_lemmas_elmo_1024_2019

Theory and Algorithms: ELMO with context

Two sentences from different time periods; 100 random pairs. Context window = 1

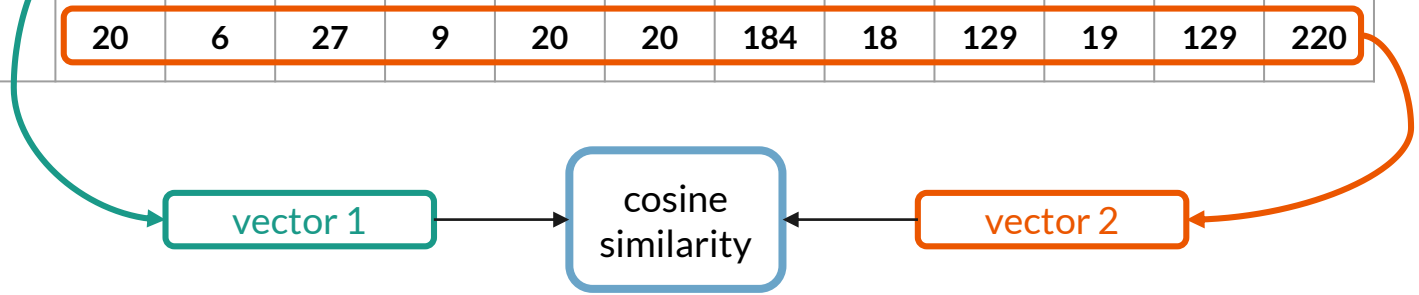


ruwikiruscorpora_tokens_elmo_1024_2019; ruwikiruscorpora_lemmas_elmo_1024_2019

Theory and Algorithms: grammatical features

word *svalka* ‘dump’

Number	Plural						Singular					
Case	Acc	Dat	Gen	Ins	Loc	Nom	Acc	Dat	Gen	Ins	Loc	Nom
pre-Soviet	4	1	7	2	2	11	41	6	71	9	56	133
Soviet	20	6	27	9	20	20	184	18	129	19	129	220



Results of the experiments: test dataset

Model	Spearman correlation 1	Spearman correlation 2	Spearman correlation 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***

* indicates the results where the p-value is lower than 0.05

Why grammatical profiles help?

Mechanisms of grammaticalization (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.

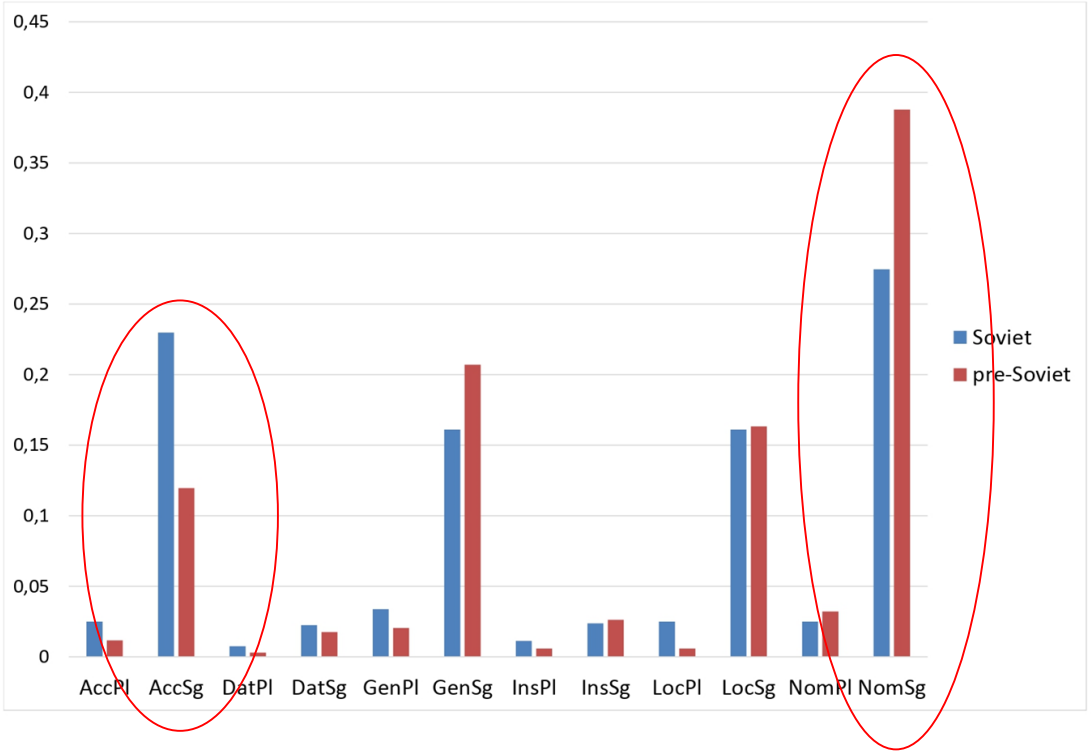
Why grammatical profiles help?

(Rakhilina 2020): *почтение* ‘reverence’

- before the 18th century: an action
  многия дары и почтениа ‘various presents and reverences’
- after the 18th century: an attitude => reduction of a paradigm (loss of all plural forms)

Our findings:
This happens systematically, and these changes are not always obvious -- but can be detected automatically

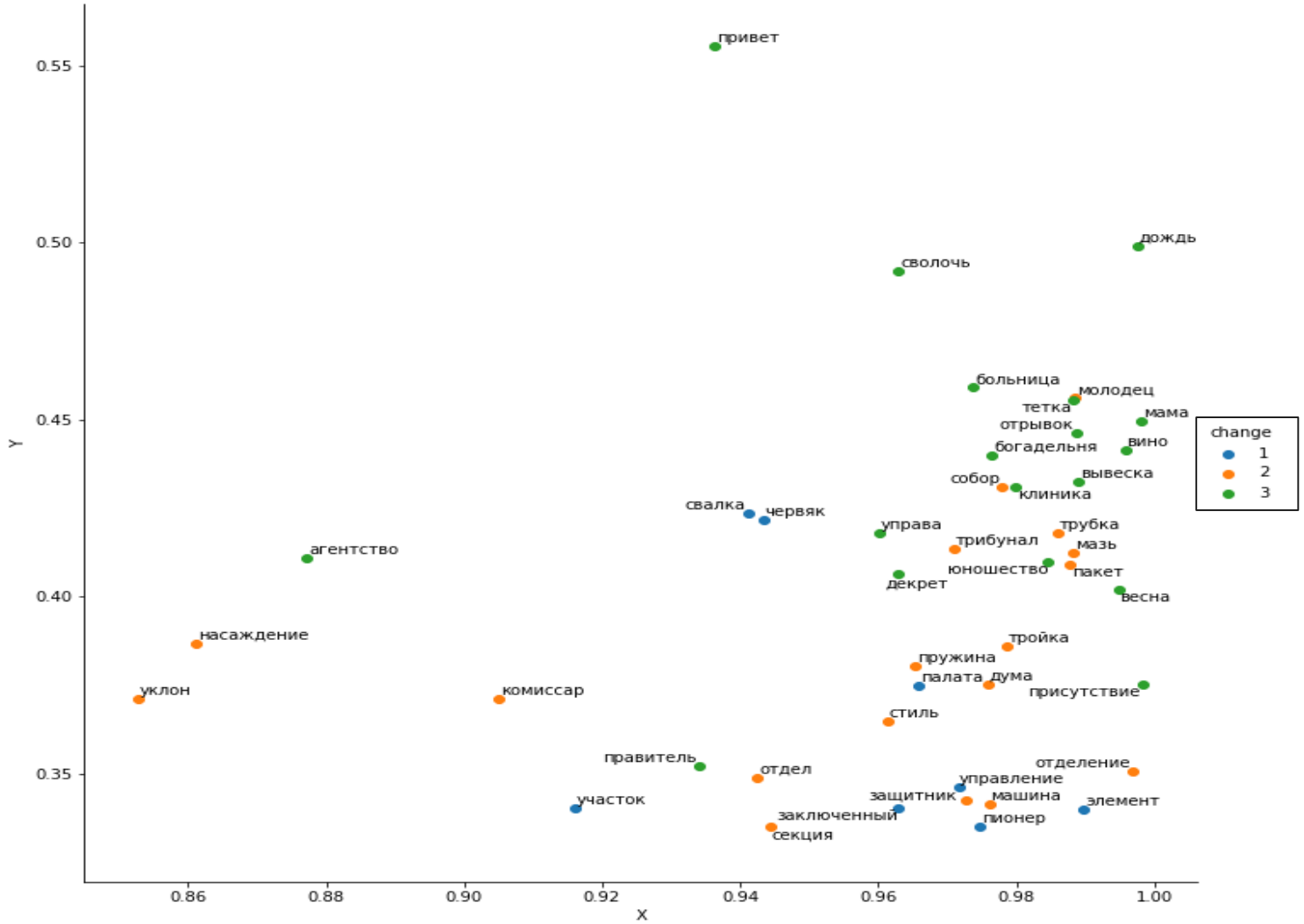
svalka ‘dump’


0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
AccPl
AccSg
DatPl
DatSg
GenPl
GenSg
InsPl
InsSg
LocPl
LocSg
NomPl
NomSg
Soviet
pre-Soviet

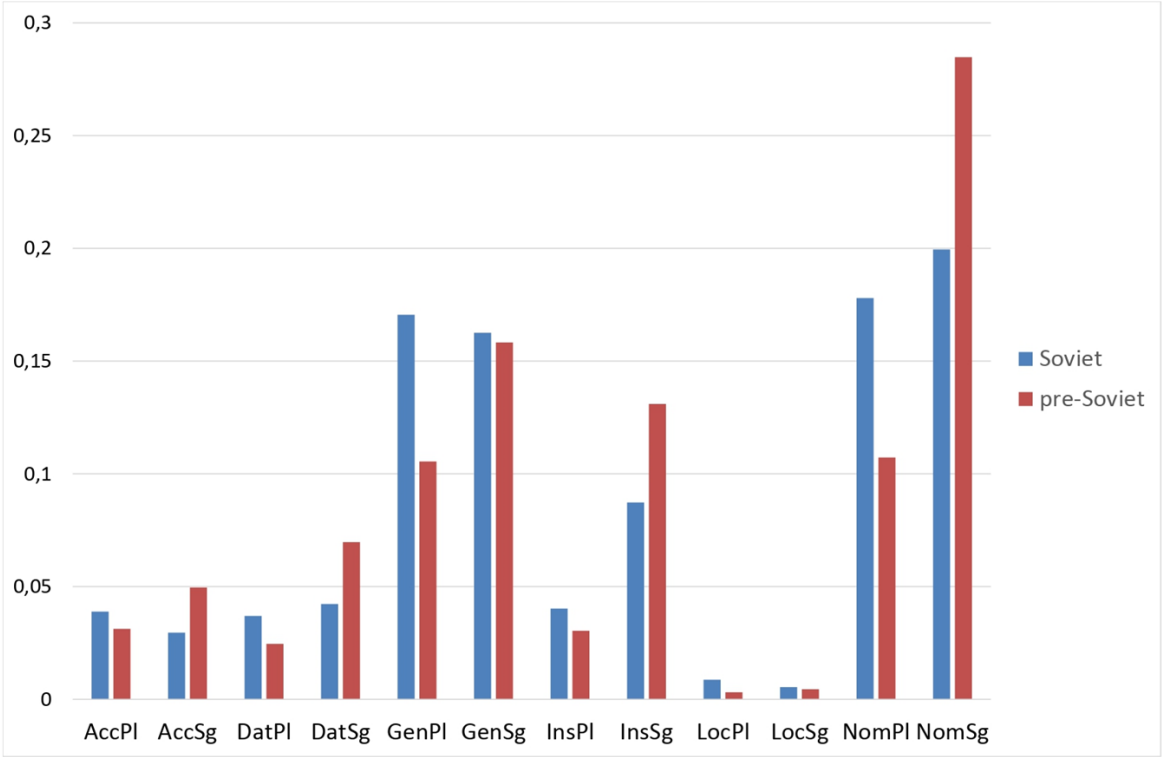
svalka ‘dump’

(1) *Тотчас же закипела* ***свалка*****[NomSg]**, *и десятки тел смешались в одну общую кричащую массу*.
’A **scuffle** ensued, with dozens of women in a bawling, struggling mass on the ground.’ [Aleksandr Kuprin. Olesya (Stepan Apresyan, 1982)]

(2) *Говорят шефу: станок сломался. Он верит, волокут станок на* ***свалку*****[AccSg]**.
’They would tell the boss that a lathe was broken. He would believe them and they would drag the lathe out on to the **rubbish dump**.’ [Anatoly Kuznetsov. Babi Yar (David Floyd, 1970)]


привет
дождь
сволочь
больница
молодец
тетка
мама
отрывок
богадельня
вино
собор
вывеска
клиника
свалка
червяк
управа
трубка
трибунал
мазь
агентство
юношество
пакет
декрет
весна
насаждение
тройка
пружина
палата
дума
присутствие
уклон
комиссар
стиль
правитель
отдел
отделение
управление
защитник
машина
элемент
участок
заключенный
пионер
секция
change
1
2
3
0.55
0.50
0.45
0.40
0.35
y
0.86
0.88
0.90
0.92
0.94
0.96
0.98
1.00
x

pravitel’
‘ruler’



Conclusions

1. Linear regression (two features: cosine similarities of ELMo vectors and cosine similarities of grammatical vectors) showed the best performance
2. Grammatical re-profiling correlates with semantic change
3. Grammatical change is more visible on a longer time span

Thank you for your attention!