RuSemShift: a dataset of historical lexical semantic change in Russian

Julia Rodina
National Research University Higher School of Economics
Moscow, Russia
`julia.rodina97@gmail.com`

Andrey Kutuzov
University of Oslo
Oslo, Norway
`andreku@ifi.uio.no`

Abstract

We present RuSemShift, a large-scale manually annotated test set for the task of semantic change modeling in Russian for two long-term time period pairs: from the pre-Soviet through the Soviet times and from the Soviet through the post-Soviet times. Target words were annotated by multiple crowd-source workers. The annotation process was organized following the DURel framework and was based on sentence contexts extracted from the Russian National Corpus. Additionally, we report the performance of several distributional approaches on RuSemShift, achieving promising results, which at the same time leave room for other researchers to improve.

1 Introduction

Language is a constantly changing system by its nature, since it is a method of communication and a social instrument. As such, it should meet the needs of the speakers, and thus it adapts to changes in the society and the ever-changing world. As a part of language, lexical meaning also evolves over time, with words undergoing diachronic (or temporal) semantic shifts (Traugott and Dasher, 2001).

Tracing semantic change can be important either in itself, as a linguistic study, or for practical downstream applications, for example, in socio-linguistic research. Manual analysis of such shifts is time-consuming and laborious even after the emergence of large representative corpora, since one needs to look through a lot of examples and lexicographic resources which often do not record current lexical changes in language due to limited resources. Thus, researchers are trying to model these processes using advanced computational approaches often based on distributional semantics and dense word embeddings (Kutuzov et al., 2018; Tang, 2018).

However, this is still mostly done for English: often simply because manually annotated test data is not available for other languages. Recently, consistently annotated lexical semantic change test sets for multiple languages started to appear; see, for example, Schlechtweg et al. (2020). In this paper, we continue this vein of work by presenting *RuSemShift*. *RuSemShift*[1] is the first historical semantic change dataset for Russian annotated according to the DURel framework (Schlechtweg et al., 2018) using a large crowd-sourcing platform, instead of personal intuitions of individual researchers. It allows to evaluate semantic change detection systems by their ability to estimate the shifts which occurred to Russian words either after 1917 (the fall of the Russian Empire) or after 1990 (the fall of the Soviet Union).

The rest of the paper is organized as follows: in Section 2, we put our research in the context of the related work. In Section 3, we present the employed corpora and the process of the dataset creation. Section 4 describes the annotation itself. In Section 5, we empirically evaluate several existing semantic change detection algorithms (based on static and contextualized embeddings) on *RuSemShift* to check its sanity. Section 6 summarizes our contributions and outlines future research.

---

[1]`https://github.com/juliarodina/RuSemShift`

2 Related work

Works on language change from general linguistics like Traugott and Dasher (2001) or Daniel and Dobrushina (2016) as a rule contain only a small number of hand-picked examples. The *DatSemShift* database (Zalizniak, 2018) features more than 4 000 semantic shifts across 800 languages. But it is focused on cognitive proximities between pairs of linguistic meanings (with a limited set of pre-defined senses): in this paradigm, a semantic shift is just a case of extended polysemy. The *DatSemShift* database is extremely useful for identifying recurring cross-linguistic semantic shifts, but it is difficult to employ it for evaluation of unsupervised semantic change detection systems.

To our knowledge, the first Russian test set to evaluate lexical semantic change detection systems was created by Kutuzov and Kuzmenko (2018). They used prior linguistic work to manually collect Russian words which changed their meaning from the pre-Soviet times through the Soviet times. Kutuzov and Kuzmenko (2018) also employed static word embedding models trained on the corresponding Russian diachronic corpora to detect semantic change for the words from the dataset (distinguishing changed words from stable ones). They concluded that Kendall's τ (Kendall, 1948) and Jaccard similarity (Jaccard, 1901) between nearest neighbor lists worked best in this tasks. More recent work by Fomin et al. (2019) extended this research to more granular time bins (periods of 1 year): they analyzed sequential pairs of word embedding models trained on yearly corpora of Russian news from 2000 up to 2014. To evaluate their system, they created a test set which contained human judgements about how much the meaning of a word has shifted over the given years. Using this ground truth, Fomin et al. (2019) evaluated 5 algorithms for semantic change detection and provided solid foundation for future research in in this direction for Russian.

However, these datasets suffer from two serious issues. First, they are either produced by a single person (Kutuzov and Kuzmenko, 2018) or annotated by the paper authors themselves (Fomin et al., 2019). This makes them inherently subjective. Second, each of this datasets is created and annotated in a different manner, making them inconsistent and not comparable nor to themselves neither to similar efforts made for other languages. In addition, the test set from (Kutuzov and Kuzmenko, 2018) features only binary labels (shifted or stable), making it useless for research in graded semantic change detection.

The *RuSemShift* test set we present in this paper makes partial use of these existing test sets, but all the words are re-annotated from scratch. In *RuSemShift*, we adhere to the language-agnostic *Diachronic Usage Relatedness* (DURel) semantic change annotation methodology proposed in Schlechtweg et al. (2018). The intended use of the DURel framework is the creation of consistently and robustly annotated test sets containing words labeled with the degrees of their diachronic lexical semantic change. Schlechtweg et al. (2018) presented a test set for German annotated this way. In addition, English, German, Latin and Swedish test sets used in the SemEval-2020 shared task on unsupervised lexical semantic change detection (Schlechtweg et al., 2020) were also created following the same approach. Thus, we deem DURel to be the current de-facto standard for the annotation of semantic change datasets.

DURel employs the notion of usage relatedness borrowed from research on word sense disambiguation task (Brown, 2008). The idea is to measure the degree of semantic change as a function of mean relatedness across pairs of word's occurrences in different time periods. The annotators are not required to know anything about diachronic change: they are presented with two sentences (both containing a target word) and asked to estimate how similar are the senses in which the target word is used. They should choose a score from the 4-point scale described below in Section 4.

A sample of target word usage pairs from the first time period t_1 form the so called EARLIER group and the usage pairs from later time period t_2 form the LATER group. To cover the cases that cannot be tracked by comparing mean
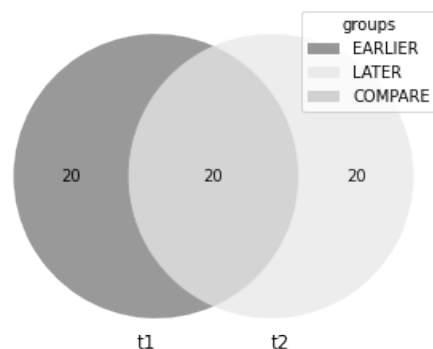

Figure 1: Number of *sentence pairs* for each word from the time period $t1$ (EARLIER), the time period $t2$ (LATER) and from both periods (COMPARE).

relatedness of two time periods, Schlechtweg et al. (2018)
directly compare the 'old' and 'new' meanings by creating the additional COMPARE group which contains pairs
where the first sentences is from the t_1 time period and the second sentence is from the t_2 time period
(see Figure 1). Each group contains 20 sentence pairs randomly sampled from the corresponding corpora.
Target word's *w* degree of semantic change is quantified with two measures:

1. ΔLATER, which is the difference between mean relatedness of the EARLIER group and the LATER group:

   $\Delta LATER(w) = Mean_{LATER}(w) - Mean_{EARLIER}(w);$

2. COMPARE, which is the mean relatedness within the COMPARE group:

   $COMPARE(w) = Mean_{COMPARE}(w).$

Due to space limitations, it is impossible to provide a description of the advantages and disadvantages of these two measures here. We refer the reader to Schlechtweg et al. (2018) for the detailed discussion.

3 Data Sources

In this section, we describe our data sources. It can be looked at a sort of data statement (Bender and Friedman, 2018) for *RuSemShift*. We first present the corpus we employ and its historical sub-corpora, and then move on to the process of preliminary target word selection. The data statement continues in the next Section 4 with the description of our annotators and the annotation workflow in general.

3.1 Corpora

RuSemShift covers 3 time periods of Russian language history, following Kutuzov and Kuzmenko (2018):

1. Texts produced from 1682 to 1916: the period of Russian monarchy before the revolution of 1917; further dubbed **pre-Soviet**.

2. Texts produced from 1918 to 1990: the period of the existence of the Soviet Union; further dubbed **Soviet**.

3. Texts produced from 1991 to 2017: the period after the fall of the Soviet Union: further dubbed **post-Soviet**.

The wide coverage of these periods (several decades or even centuries each) makes it more likely that shifts in word usage will be caused by linguistic factors, and not only by extra-linguistic events. However, it is arguably still strongly influenced by cultural factors. The boundaries between time bins are related to rises and falls of political regimes, and some semantic shifts are inevitably related to that. Note that the first time period (the pre-Soviet times) is substantially longer temporally than the other two. This in theory can lead to multiple meaning shifts occurring *within* this period. But this epoch is already the smallest in terms of corpus size (see below), so it was impossible to further divide it.

Our annotation framework requires diachronic corpora with Russian texts created in the time periods listed above. As a source of such texts, we used the *Russian National Corpus* (RNC).[2] RNC is well balanced and contains Russian texts of diverse genres produced from the middle of the 18th century up to the beginning of the 21st century. Note that the corpus itself is not our contribution: it is a prior work.

The main RNC corpus size is about 320 million word tokens (including punctuation). Since all the texts are annotated with the date of their creation, it is straightforward to separate the corpus into time-specific sub-corpora:

- 94 million tokens in the pre-Soviet sub-corpus;

- 123 million tokens in the Soviet sub-corpus;

- 107 million tokens in the post-Soviet sub-corpus.

---

[2]http://ruscorpora.ru

Value	English judgments from Schlechtweg et al. (2018)	Judgments translated into Russian
0	Cannot decide	Не знаю
1	Unrelated	Значения разные
2	Distantly Related	Значения отдаленно похожи
3	Closely Related	Очень похожие значения
4	Identical	Употреблено в одном значении

Table 1: Original 4-point scale from the DURel and its Russian translation used for the annotation.

3.2 Preliminary word lists

To construct word lists for further annotation, we handpicked words that presumably have undergone semantic changes in the Soviet period compared to the pre-Soviet period or in the post-Soviet compared to the Soviet period. For each of these word lists we also randomly sampled a set of 'filler' or 'distractor' words, trying to reproduce the part of speech and frequency percentile distributions of the original target words as closely as possible. The purpose of fillers (which are assumed to be semantically stable) is to be able to evaluate the performance of semantic change detection systems in a more realistic setup: they are supposed to predict low change scores for fillers (or no change at all).

3.2.1 Pre-Soviet to Soviet dataset ($RuSemShift_1$)

The original target word list for the first pair of time bins was created in Kutuzov and Kuzmenko (2018). It consists of 43 words that have undergone semantic changes through the period from the pre-Soviet to the Soviet times. The words were collected from general linguistic studies on lexical semantic change (Ozhegov, 1953; Daniel and Dobrushina, 2016); there are 38 nouns and 5 adjectives. After the procedure of filler generation, the total number of words for annotation was 71. We emphasize again that only the changed word list was taken from Kutuzov and Kuzmenko (2018), the fillers and annotations were recreated from scratch by us.

3.2.2 Soviet to Post-Soviet dataset ($RuSemShift_2$)

The second test set is entirely our own contribution. It also contains manually chosen target words (35 nouns and 7 adjectives), primarily from the *'New Words and Meanings'* dictionary (Burtseva et al., 2009). The dictionary includes words which acquired new common senses in Russian in the post-Soviet time period (such words have a special label in the dictionary). Note that these words are not neologisms: all of them occurred in the Soviet sub-corpus as well, so semantic change cannot be estimated by frequencies alone (as shown in Section 5). After the filler generation procedure, the test set consists of 69 words.

4 Annotation

The annotation procedure was carried out on the *Yandex.Toloka* crowd-sourcing platform[3], which is more or less equivalent to the global Mechanical Turk or CrowdFlower platforms. Note that in the DURel dataset (Schlechtweg et al., 2018), all five annotators were students of linguistics and two of them had historical background, which cannot be enforced when using crowd-sourcing platforms. However, *Yandex.Toloka* allowed us to improve the quality of annotation by applying various filters to limit who can annotate. We chose 10-30% of the best annotators across the whole platform, keeping only native speakers of Russian, of the age 30 and more (to ensure them being familiar with older word senses) and possessing a university degree. We do not have any knowledge about the annotators' gender distribution.

The annotators who completed the tasks unrealistically fast were automatically filtered out. Furthermore, there were several *control tasks*, consisting of sentence pairs manually annotated by ourselves (with the scores of 1 and 4 only, thus limited to obvious cases). The users who annotated such pairs incorrectly were blocked as well. Also, we accepted the judgments of an annotator only after checking some of them manually to make sure that the annotator had understood the task.

---

[3]https://toloka.yandex.ru/

For each word from our pre-constructed word lists, we extracted all unique sentence contexts from two corpora (*pre-Soviet/Soviet* or *Soviet/post-Soviet*) and randomly sampled 60 sentence pairs, 20 for each group (EARLIER, LATER and COMPARE). For the words that occurred less than 30 times in one of the corpora we decreased the number of its sentence pairs to the closest value, divisible by 3 without a remainder. Final tables for the annotation contained 7846 pairs for $71 + 69 = 140$ words. Each sentence pair was judged by not less than 5 annotators who used the scale presented in Table 1. The annotator interface of *Yandex.Toloka* is shown in Figure 2. For the cases when one sentence context was not enough to understand the meaning of the target word, we implemented an option to show extended contexts from the same corpora. Annotators were provided with detailed guidelines explaining the task and giving examples for each grade.
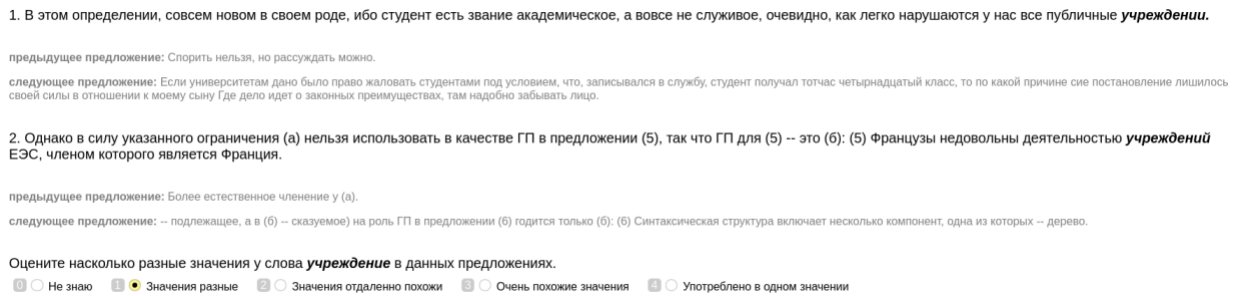

1. В этом определении, совсем новом в своем роде, ибо студент есть звание академическое, а вовсе не служивое, очевидно, как легко нарушаются у нас все публичные **учреждении.**

предыдущее предложение: Спорить нельзя, но рассуждать можно.

следующее предложение: Если университетам дано было право жаловать студентами под условием, что, записывался в службу, студент получал тотчас четырнадцатый класс, то по какой причине сие постановление лишилось своей силы в отношении к моему сыну Где дело идет о законных преимуществах, там надобно забывать лицо.

2. Однако в силу указанного ограничения (а) нельзя использовать в качестве ГП в предложении (5), так что ГП для (5) -- это (6): (5) Французы недовольны деятельностью **учреждений** ЕЭС, членом которого является Франция.

предыдущее предложение: Более естественное членение у (а).

следующее предложение: -- подлежащее, а в (б) -- сказуемое) на роль ГП в предложении (6) годится только (6): (6) Синтаксическая структура включает несколько компонент, одна из которых -- дерево.

Оцените насколько разные значения у слова **учреждение** в данных предложениях.
⓪ ○ Не знаю   ① ● Значения разные   ② ○ Значения отдаленно похожи   ③ ○ Очень похожие значения   ④ ○ Употреблено в одном значении

Figure 2: Interface of the annotation task. In the first sentence, the word 'учреждение' [uchrezhdenie] is used in the sense of 'DECREE, ORDER'. In the second sentence, 'учреждение' denotes 'INSTITUTION'.

After all sentence pairs were annotated, we measured inter-rater agreement as Krippendorff's α (Krippendorff, 2012) with ordinal scale, excluding 0 judgements ('cannot decide'). For $RuSemShift_1$, the agreement coefficient was 0.505, and for $RuSemShift_2$ its value was 0.53. Considering that the task is inherently ambiguous and complex, and also that all five annotators were different for each sentence pair, we believe this score is high enough. However, we also provide the filtered versions of both test sets, where controversial words that were annotated inconsistently (inter-rater agreement less than 0.2) are excluded. 24 words were filtered out from the $RuSemShift_1$ and 19 from the $RuSemShift_2$). We use these filtered test sets for evaluation in Section 5.
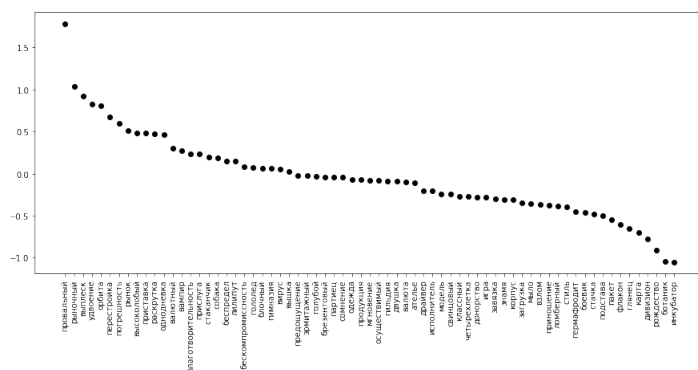
4.1 Analysis

Figure 3: Target words ($RuSemShift_2$) ranked by their ΔLATER values.

As stated in Schlechtweg et al. (2018), both measures have their inherent limitations. ΔLATER can fail to capture semantic change if a word loses an old sense and gains a new one within one time period. COMPARE tends to mix up polysemy with semantic change because of random choice of samples (can arguably be remedied using some kind of normalization). ΔLATER naturally captures the differences between two types of meaning change: innovative shift (negative values) or reducti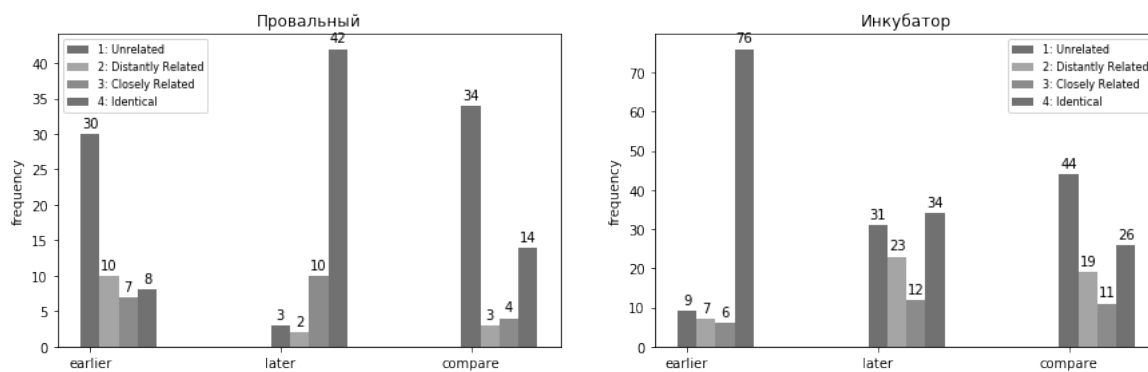ve shift (positive values). But this is true only for high absolute ΔLATER values. Schlechtweg et al. (2018) claim COMPARE to be more suitable for indicating the degree of semantic change, but prospective *RuSemShift* users can choose any of these two measures or implement their own (since the raw annotation data is available).

Figure 4: Distribution of human judgements for the $RuSemShift_2$ target words 'провальный' 'failed' (reductive change) and 'инкубатор' 'incubator' (innovative change).

Figure 3 shows the words from the $RuSemShift_2$ ranked according to their ΔLATER values. For most words, it is close to 0 (no change), but one can also see groups with $\Delta LATER >> 0$ and $\Delta LATER << 0$, indicating changes. Figure 4 compares the distribution of annotators' judgments for two words with the highest and the lowest ΔLATER values (the adjective 'провальный' [proval'nyj] 'failed' and the noun 'инкубатор' [inkubator] 'incubator' correspondingly).

For 'провальный' (the ΔLATER value of 1.78), we can see that judgements are quite diverse in the EARLIER group and the 1-judgement ('unrelated senses') is prevalent, but in the LATER group, the 4-judgement ('identical senses') is considerably more frequent, while the COMPARE group also captures strong change, since the number of 1-judgements is higher than the number of 4-judgements. Almost all context pairs from the EARLIER group (that is, the Soviet time period) shown in Table 2 are related to the literal meaning of 'провал': 'A PLACE WHERE THE SURFACE COLLAPSED INWARD' or the figurative sense of 'LOSS OF CONSCIOUSNESS' used frequently in the set expression 'провальный сон' *('deep dream')*. In all the contexts from the LATER group (the post-Soviet time period), 'провальный' is used in the sense of 'FAILED' which is more common in modern Russian. Thus, we can observe the expansion of this sense. As for the COMPARE group, sentences in pairs from each period support the same observation: there are only two usages from the earlier period that can be interpreted with the 'FAILED' sense. The word 'провальный' did not lose its literal meaning, it just became much less frequent, and the old figurative meaning (as in *'deep dream'*) is almost completely lost.[4] Consequently, we can observe the word losing an old sense and gaining a new one at the same time.

For the word 'инкубатор' (the ΔLATER value of -1.05), the distributions are the opposite. The 4-judgment is prevalent in the EARLIER group and there is diversity in the LATER group. Indeed, in the EARLIER group (the Soviet period), 'инкубатор' is used mostly in its literal meaning of 'INCUBATOR', while in the LATER group (he post-Soviet period) there are many occurrences in the figurative sense of 'BUSINESS INCUBATOR'. Thereby, the word 'инкубатор' is undergoing an innovative meaning change in the post-Soviet times, as its figurative sense is becoming more and more widespread.

5 Evaluation

RuSemShift is mainly intended to be used by other researchers in the field of lexical semantic change detection for Russian. However, below we report the performance of several well-known change detection methods on our datasets, to set the baseline. We solved the task of ranking: that is, given a list of target words and two time-specific corpora, a system should predict semantic change degrees which would position the target words in the order as close to the gold one as possible. First we employed static distributional embeddings trained with the CBOW algorithm (Mikolov et al., 2013). After that, we tried several variations of contextualized embeddings, namely ELMo (Peters et al., 2018).

---

[4]But is kept in the related word 'провал' ('провал в памяти'/*'lapse of memory'*).

Group	Sentence 1	Sentence 2
EARLIER	`<...>` выпили "микстуры"и спят где-нибудь сном провальным, пьяным `<...>` '`<...>` they drank a "potion" and sleeping somewhere in **deep** drunken **sleep** `<...>`' [Viktor Astaf'ev. Pechal'nyj detektiv (1982-1985)]	`<...>` Кщаре доходит до семиде-сяти пяти метров и что вообще здесь много провальных озер `<...>` '`<...>` Kshchare reaches seventy-five meters and in general there are many **deep lakes** `<...>`' [V. A. Solouhin. Vladimirskie prosyolki (1956-1957)]
LATER	`<...>` самые "бюджетные"программы оказываются самыми проваль-ными в исполнении `<...>` '`<...>` the cheapest programs turn out to be the most **disastrous**`<...>`' [Ivan Golikov. Dohodnoe mesto // «Vsluh o...», 2003.05.19]	`<...>` Подготовка к референдуму в Чечне до прошлой недели но-сила провальный характер `<...>` '`<...>`Preparation for referendum in Chech-nya until last week was **disastrous**`<...>`' [Aleksej Makarkin. Krizisnoe upravlenie «chechenizaciej» // «POLITKOM.RU», 2003.03.03]
COMPARE	`<...>` Наденька на минутку за-былась провальным сном и когда открыла глаза `<...>` '`<..>` Nadenka **fainted** for a minute and when she opened her eyes `<..>`' [B. S. ZHitkov. Viktor Vavich. Kniga vtoraya (1941)]	Провальное выступление сбор-ной России на Олимпиаде `<...>` 'The **failed** performance of the Rus-sian team at the Olympics `<...>`' [Sergej Podushkin. Opasnye igry na al'pijskom vozduhe. Gornolyzhniki nachinayut sezon // «Izvestiya», 2002.10.25]

Table 2: Examples of sampled sentence pairs for the word 'провальный' 'failed'.

Dataset	ΔLATER		COMPARE	
	Frequency	word2vec Procrustes	Frequency	word2vec Procrustes
$RuSemShift_1$	-0.275	**0.234**	0.046	**0.403***
$RuSemShift_2$	-0.024	**0.118**	-0.062	**0.269**

Table 3: Spearman ρ correlations of the frequency-based and word2vec-based predictions with *RuSemShift* annotations. * denotes statistical significance at $p < 0.05$.

5.1 Static embeddings

In Table 3, we report the performance of the method based on static embeddings. We also use a very simple frequency-based baseline, where the degree of semantic change is estimated by the difference of target word absolute frequencies between two time periods. The 'word2vec with Procrustes alignment' is the classic method of calculating cosine similarity between target word vectors in two CBOW embedding models trained on different time periods (we used the same splits of the Russian National Corpus as during the creation of *RuSemShift*). The trained models were aligned using Orthogonal Procrustes as described in Hamilton et al. (2016).

We report the correlations between the predictions of these methods and the COMPARE and ΔLATER values from $RuSemShift_1$ and $RuSemShift_2$. Note that we used the absolute values of ΔLATER, leaving the distinguishing of innovative versus reductive semantic change for future work. This means that an ideal system will produce perfectly positive correlations with ΔLATER and perfectly negative correlations with COMPARE (or vice versa, depending on the exact method), since the former increases as the degree of semantic change grows, while the latter increases as word usages become more similar.

As expected, comparing cosine distance of Procrustes-aligned word embeddings far outperforms sim-ply measuring frequency changes (this latter method produces predictions close to random in most cases). It can also be observed that $RuSemShift_2$ is more difficult than $RuSemShift_1$. A possible explana-tion is that Soviet and post-Soviet texts are on average less distant in time from each other than Soviet and pre-Soviet texts: the former pair of time bins lies entirely within 100 years, while the latter covers the time span of about 250 years. Because of that, semantic differences from $RuSemShift_1$ (between Soviet and pre-Soviet lexical meanings) are manifested more clearly in the corpora. Finally, ΔLATER

rankings are more difficult to reproduce than those for COMPARE; no method managed to achieve a statistically significant correlation in this case. This arguably stems from the nature of this measure: even though we used its absolute values, it is still focused rather on the nature of semantic shifts than on their degree, and this is not something which can be easily approximated by cosine similarity between Procrustes-aligned word embeddings. More advanced methods are required to better predict ΔLATER values from data.

5.2 Contextualized embeddings

We trained ELMo (Peters et al., 2018) models[5] on the RNC texts to produce contextualized token representations for each time period. ELMo embeddings are inferred from bidirectional language models trained using two-layer long short-term memory network (LSTM) for next word predictions in both directions. Every token is represented as a linear combination of hidden layers and depends on the context in which the token appears.

All corpora were segmented into sentences, tokenized and lemmatized with a *UDPipe 1.2* (Straka and Straková, 2017) model trained on the SynTagRus treebank (Droganova et al., 2018). It is not yet well known whether contextualized embedding models should be trained on lemmatized or non-lemmatized texts, but Kutuzov and Kuzmenko (2019) showed that at least for Russian, lemmatization improves the performance in word sense disambiguation task. It also excludes word form bias, since we want to trace semantic shifts in lexemes rather than specific word forms.

We trained six ELMo models in three variants:

1. a single model trained on the full RNC corpus with texts from all time periods (differentiation by time periods is made at the inference stage when token embeddings are produced);

2. three models trained separately on each sub-corpus: pre-Soviet, Soviet and post-Soviet models;

3. two models trained incrementally (initialized from the checkpoint of the model trained on texts from the previous time period): Soviet incremental model and post-Soviet incremental model.

We extract ELMo token embeddings for each word's usage in two adjacent time periods and estimate semantic change score for this word using the measures described below. Extracted contextualized embeddings of each target word from two time periods are represented as two time-specific matrices. We explored two semantic change detection measures: *cosine similarity between averaged token embeddings*[6] and *Jensen-Shannon divergence* which requires prior application of the clustering algorithm.

1. *Cosine similarity*. We compute average vectors from usage matrices, which gives us representations which resemble static type embeddings. Then we compute cosine similarity between these average embeddings as a measure of semantic change. The lower is the cosine similarity, the higher is the degree of semantic change. This method is inspired by the PRT technique from Kutuzov and Giulianelli (2020).

2. *Jensen-Shannon divergence* (JSD). In this measure, influenced by Dubossarsky et al. (2015), Martinc et al. (2020) and Giulianelli et al. (2020), word usage matrices from two time periods are first stacked into one matrix. Then, we standardize the vectors and obtain word usage clusters of token embeddings using the Affinity Propagation clustering algorithm (Frey and Dueck, 2007). After obtaining clusters for each word, we calculate usage type (sense) probability distributions for each time period by normalizing counts of word usages in the clusters. Then we compute the JSD score:

$$JSD = \sqrt{\frac{D(p \parallel m) + D(q \parallel m)}{2}} \tag{1}$$

---

[5]Using the implementation from `https://github.com/ltgoslo/simple_elmo` with 3 epochs and the vocabulary of 100 000 most frequent words. The rest hyperparameters were left at their default values.

[6]This was not used with separately trained models, because it does not make sense to directly measure it between vectors from two different spaces.

where D is the Kullback-Leibler divergence, p and q are sense distributions and m is the pointwise mean of p and q. Higher JSD score indicates more intense change in the proportions of clustered word usage types across time periods.

Model type	Dataset	Cosine similarity		JSD	
		ΔLATER	COMPARE	ΔLATER	COMPARE
Single	$RuSemShift_1$	**-0.351***	0.246	-0.158	**0.385***
	$RuSemShift_2$	**-0.300***	**0.541***	-0.147	**0.364***
Incremental	$RuSemShift_1$	0.068	0.055	-0.117	-0.074
	$RuSemShift_2$	0.028	0.186	0.028	0.186
Separate	$RuSemShift_1$	-	-	0.201	-0.283
	$RuSemShift_2$	-	-	-0.064	0.100

Table 4: Spearman ρ correlations of the ELMo-based predictions with *RuSemShift* annotations.

Table 4 shows that incremental and separate ELMo models do not yield significant correlations with human judgments. As for the single model, it consistently outperforms static embeddings on $RuSemShift_2$ (both measures), and $RuSemShift_1$ (ΔLATER): thus, in 3 out of 4 cases. As already observed, COMPARE metrics is easier to approximate than ΔLATER. Also, cosine similarity generally is better than JSD, despite the latter being much heavier computationally. Negative correlations with ΔLATER are normal and caused by the nature of this measure: higher values indicate stronger change.

6 Conclusion

We presented *RuSemShift*, which consists of two publicly available test sets of Russian words manually annotated with the degrees of diachronic semantic change they experienced (the degree is a continuous score). Annotation process was based on the theoretically sound DURel framework (Schlechtweg et al., 2018). One of the datasets is produced from re-annotated list of target words from prior work and another is completely new. The datasets allow to evaluate methods for lexical semantic change detection in Russian: either graded or binary (with any desired binarization threshold). They provide data on 3 large time periods: pre-Soviet (1682-1916), Soviet (1918-1990) and post-Soviet (1991-2017).

This is the first semantic change detection dataset for Russian created in a large-scale crowd-sourcing annotation effort. It is also important that *RuSemShift* is fully compatible with semantic change datasets developed for other languages, for example those presented for the corresponding SemEval-2020 shared task (Schlechtweg et al., 2020). The dataset is available online[7] under a Creative Commons Attribution-ShareAlike 4.0 International License.

As a sanity check (and to establish the baseline performance boundaries), we evaluated several semantic change modeling systems on *RuSemShift*. We managed to achieve significant correlation with human judgments both with static and with contextualized word embeddings, with the latter consistently outperforming the former. At the same time, simple frequency-based baseline failed to achieve any meaningful results, which signals that the dataset lacks simplistic frequency cues.

RuSemShift is limited to nouns and a few adjectives. One of the possible future research directions is to extend the dataset with other parts of speech. Another drawback of the dataset is that it does not distinguish between different types of semantic change (e.g. narrowing, widening, metaphorization etc.). Nevertheless, we hope that in its current state, *RuSemShift* will already be of help to the researchers interested in tracing diachronic semantic shifts in Russian.

Acknowledgments

---

[7]https://github.com/juliarodina/RuSemShift

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Susan Windisch Brown. 2008. Choosing Sense Distinctions for WSD: Psycholinguistic Evidence. In *Proceedings of ACL-08: HLT, Short Papers*, pages 249–252, Columbus, Ohio, June. Association for Computational Linguistics.

Tatiana Burtseva, Evgeniy Levashov, and Yuri Denisenko. 2009. *Novye slova i znacheniya. Slovar'-spravochnik po materialam pressy i literatury 90-h godov XX veka: V 2 tomah [New words and meanings. Reference dictionary by the materials of the press and literature in the 90's of the XX century: in 2 volumes].* Institute for Linguistic Studies of the Russian Academy of Sciences.

Mihael Daniel and Nina Dobrushina. 2016. *Dva veka v dvadtsati slovax [Two centuries in twenty words].* Izdatelskij dom NIU VSHE.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks. In *Proceedings of the 17th international workshop on treebanks and linguistic theories*, pages 52–65, Oslo University, Norway.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*.

Vadim Fomin, Daria Bakshandaeva, Julia Rodina, and Andrey Kutuzov. 2019. Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines. In *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*, pages 203–218, Moscow, Russia.

Brendan Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July. Association for Computational Linguistics.

William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Paul Jaccard. 1901. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines.* Rouge.

Maurice George Kendall. 1948. *Rank correlation methods*. Griffin.

K. Krippendorff. 2012. *Content analysis: An introduction to its methodology*. Sage.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *arXiv preprint arXiv:2005.00050 (to appear in Proceedings of the 14th International Workshop on Semantic Evaluation)*, Barcelona, Spain. Association for Computational Linguistics.

Andrey Kutuzov and Elizaveta Kuzmenko. 2018. Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. *Quantitative Approaches to the Russian Language*, pages 95–112.

Andrey Kutuzov and Elizaveta Kuzmenko. 2019. To Lemmatize or Not to Lemmatize: How Word Normalisation Affects ELMo Performance in Word Sense Disambiguation. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 22–28, Turku, Finland.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397. Association for Computational Linguistics.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the International World Wide Web Conference*, pages 20–24.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Sergey Ozhegov. 1953. K voprosu ob izmenenijax russkogo iazyka v sovetskuju epoxu [About changes in russian language in the soviet era]. *Voprosy yazykoznanija*, 2:71–82.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana, June. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *To appear in Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Elizabeth Traugott and Richard Dasher. 2001. *Regularity in semantic change*. Cambridge University Press.

Anna Zalizniak. 2018. The catalogue of semantic shifts: 20 years later. *Russian Journal of Linguistics*, 22(4):770–787.