Three-part diachronic semantic change dataset for Russian

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

Lidia Pivovarova
University of Helsinki
Finland
lidia.pivovarova@helsinki.fi

Abstract

We present a manually annotated lexical semantic change dataset for Russian: RuShiftEval. Its novelty is ensured by a single set of target words annotated for their diachronic semantic shifts across three time periods, while the previous work either used only two time periods, or different sets of target words. The paper describes the composition and annotation procedure for the dataset. In addition, it is shown how the ternary nature of RuShiftEval allows to trace specific diachronic trajectories: 'changed at a particular time period and stable afterwards' or 'was changing throughout all time periods'. Based on the analysis of the submissions to the recent shared task on semantic change detection for Russian, we argue that correctly identifying such trajectories can be an interesting sub-task itself.

1 Introduction

This paper describes *RuShiftEval*: a new dataset of diachronic semantic changes for Russian words. Its novelty in comparison with prior work is its multi-period nature. Until now, semantic change detection datasets focused on shifts occurring between **two** time periods. On the other hand, *RuShiftEval* provides human-annotated degrees of semantic change for a set of Russian nouns over **three** time periods: pre-Soviet (1700-1916), Soviet (1918-1990) and post-Soviet (1992-2016). Notably, it also contains 'skipping' comparisons of pre-Soviet meanings versus post-Soviet meanings. Together, this forms three subsets: *RuShiftEval*-1 (pre-Soviet VS Soviet), *RuShiftEval*-2 (Soviet VS post-Soviet) and *RuShiftEval*-3 (pre-Soviet VS post-Soviet).

The three periods naturally stem from the Russian history: they were radically different in terms of life realities and writing and practices, which is reflected in the language. As an example, the word дядька lost its 'tutor of a kid in a rich family' sense in the Soviet times, with only the generic 'adult man' sense remaining. Certainly, language development never stops and Russian also gradually evolved within those periods as well, not only on their boundaries. However, in order to create a usable semantic change dataset, one has to draw the boundaries somewhere, and it is difficult to come up with more fitting 'changing points' for Russian.

RuShiftEval can be used for testing the ability of semantic change detection systems to trace long-term multi-point dynamics of diachronic semantic shifts, rather than singular change values measured by comparing two time periods. As such, *RuShiftEval* was successfully employed in a recent shared task on semantic change detection for Russian (Kutuzov and Pivovarova, 2021).

2 Related work

Automatic detection of word meaning change is a fast growing research area (Kutuzov et al., 2018; Tahmasebi et al., 2018). Evaluation of this task is especially challenging; *inter alia*, it requires gold standard annotation covering multiple word usages.

The common practice is to annotate pairs of sentences as using a target word in either the same or different senses. It was introduced for the word sense disambiguation task in (Erk et al., 2013), while (Schlechtweg et al., 2018) proposed methods to aggregate pairwise annotations for semantic change modeling; one of them, the COMPARE metrics, is used in *RuShiftEval*.

A similar approach was used for the SemEval'20 shared task on semantic change detection (Schlechtweg et al., 2020): annotators labeled pairs of sentences, where some pairs belonged to the same periods and some to different ones. This annotation resulted in a diachronic word usage graph, which was then clustered to obtain sepa-

Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, pages 7–13

rate word senses and their distributions between time periods (Schlechtweg et al., 2021).

The pairwise sentence annotation has been used in creating another semantic change dataset for Russian, *RuSemShift* (Rodina and Kutuzov, 2020). We use the same annotation procedure and rely on the same corpus, i.e. Russian National Corpus (RNC) split into pre-Soviet, Soviet and post-Soviet subcorpora. However, *RuSemShift* features two sets of words: one for the changes between the pre-Soviet and Soviet periods, and another for the Soviet and post-Soviet periods. The new *RuShiftEval* dataset, which we present in this paper, uses a *joint word set* allowing for tracing each word across three time periods. In addition, we directly annotate semantic change between the pre-Soviet and post-Soviet periods, skipping the Soviet one.

3 Dataset Construction

3.1 Word List Creation

In building the dataset, we relied on the graded view on word meaning change (Schlechtweg et al., 2021): for each word in the dataset, we measure a *degree of change* between pairs of periods, rather than making a binary decision on whether its sense inventory changed over time. The measure relies on pairwise sentence annotations, where each pair of sentences is processed by at least three annotators.

Compiling the target-word set, we needed to ensure two main conditions: (i) the dataset contains many 'interesting' words, i.e. words that changed their meaning between either pair of periods; (ii) not all words in the dataset actually changed their meaning. We followed the same procedure as in (Kutuzov and Kuzmenko, 2018; Rodina and Kutuzov, 2020; Schlechtweg et al., 2020): first, select changing words, and then augment them with *fillers*, i.e. random words following similar frequency distribution across three time periods.

Technically, it was possible to populate the target word set automatically, using any pre-trained language model (LM) for Russian and some measure of distance between word representations in different corpora. However, we wanted our target words choice to be motivated linguistically rather than influenced by any LM architecture. Therefore, to find changing words, we first consulted several dictionaries of outdated or, on the contrary, the most recent Russian words, such as (Novikov, 2016; Basko and Andreeva, 2011; Skljarevsky, 1998). Unfortunately, dictionaries provided less examples than we needed: they often contain archaisms, neologisms, multi-word expressions, and words which are infrequent in the corpus or not used in the meanings specified in the dictionaries.

However, we discovered that some changing words could be found in papers on specific linguistics problems. For example, the word облако ('cloud') was found in a paper on the Internet language (Baldanova and Stepanova, 2016); стол ('table/diet')—in an article discussing the language of one story by Pushkin (M., 2016). Finally, to find some of the target words, we used our intuition as educated native speakers. Out of 50 words, 13 were found in dictionaries, 10 invented by ourselves and the rest 27 found in articles on more specific topics. Regardless the initial word origin, we manually checked that all words occur at least 50 times in each of the three sub-corpora and that the distinctive sense is used several times.

Fillers (selected for each target word) are sampled so that they belong to the same part of speech—nouns in our case—and their frequency percentile is the same as the target word frequency percentile in all three periods. The aim here is to ensure that frequency cannot be used to distinguish the target words from fillers.[1] For *RuShiftEval*, we sampled two filler words for each target word.

The final dataset consists of 111 Russian nouns, where 12 words form a development set and 99 words serve as a test set. Since the annotation procedure is the same as for *RuSemShift* (Rodina and Kutuzov, 2020), one can use one of these resources as a training set and then evaluate on another.

3.2 Annotation

Annotators' guidelines were identical to those in *RuSemShift* (Rodina and Kutuzov, 2020). To generate annotation tasks, we sampled 30 sentences from each sub-corpus and created sentence pairs. We ran this sampling independently for all three period pairs. The sentences were accompanied by one preceding and one following sentence, to ease the annotators' work in case of doubt. The task was formulated as labeling on a 1-4 scale, where 1 means the senses of the target word in two sentences are unrelated, 2 stands for 'distantly related', 3 stands for 'closely related', and 4 stands for 'senses are identical' (Hätty et al., 2019). Annotators were also allowed to use the 0 ('cannot decide') judgments.

[1] Indeed, there is no significant correlation between frequency differences and the aggregated relatedness scores from our gold annotation.

Time bins	α	ρ	JUD	0-JUD
Test set (99 words)				
RuShiftEval-1	0.506	0.521	8 863	42
RuShiftEval-2	0.549	0.559	8 879	25
RuShiftEval-3	0.544	0.556	8 876	31
Development set (12 words)				
RuShiftEval-1	0.592	0.613	1 013	7
RuShiftEval-2	0.609	0.627	1 014	3
RuShiftEval-3	0.597	0.632	1 015	2

Table 1: *RuShiftEval* statistics. α and ρ are inter-rater agreement scores as calculated by Krippendorff's α (ordinal scale) and mean pairwise Spearman ρ. JUD is total number of judgments and 0-JUD is the number of 0-judgments ('cannot decide').

They were excluded from the final datasets, but their number was negligible anyway: about 100 out of total 30 000.

The annotation was carried out on the Yandex.Toloka crowd-sourcing platform.[2] We employed native speakers of Russian, older than 30, with a university degree. To ensure the annotation quality, the authors themselves annotated about 20 control examples for each pair of periods. We chose the most obvious cases of 1 and 4 for this; annotators who answered incorrectly (not with the exactly matching grade), were banned from the task for 24 hours. The inter-rater agreement statistics and the number of judgments in each *RuShiftEval* subset are shown in Table 1. The agreement is on par with other semantic change annotation efforts: (Schlechtweg et al., 2020) report Spearman correlations ranging from 0.58 to 0.69, (Rodina and Kutuzov, 2020) report Krippendorff's α ranging from 0.51 to 0.53.[3] Each subset was annotated by about 100 human raters, more or less uniformly 'spread' across annotation instances, with the only constraint being that each instance must be annotated by three different persons.

Finally, the degrees of semantic change for each word between a pair of periods were calculated using the COMPARE metrics (Schlechtweg et al., 2018), which is the average of pairwise relatedness scores. Interestingly, some words initially sampled as fillers—e.g. ядро ('cannonball or core/nucleus')—ended up among most changed according to the annotation. Also some words from the initial set were annotated as relatively stable. This happened because the distinctive sense was rare or because annotators' opinion diverged from linguistic knowledge in the dictionaries. For example, for the word бригада ('brigade/gang/team') dictionaries list two distinct senses—a military and a civil one. However, in most cases the annotators considered these senses identical or closely related.

The dataset is publicly available, including the raw scores assigned by annotators.[4]

4 Diachronic trajectory types

RuShiftEval allows tracing multi-hop dynamics of semantic change. A similar analysis of diachronic word embedding series or 'trajectories' was conducted in (Kulkarni et al., 2015) and (Hamilton et al., 2016b), but the former focused on change point detection, and the latter on finding general laws of semantic change. With manually annotated *RuShiftEval* dataset we were able to move further and identify at least three different types of changing trajectories: 1) changes in every period pair; 2) change in the Soviet period as compared to the pre-Soviet period; 3) change in the post-Soviet period as compared to the Soviet period.

Since approximately a half of the words in the dataset did not change their meaning they exhibit a fourth, trivial type of trajectory, where all three distances are small. In principle there could be a fifth type of trajectory, where difference between pre-Soviet and post-Soviet periods is substantially smaller than between other period pairs, which would mean that a word was used in a new sense during the Soviet time but then came back to its original meaning. However, we did not find any words following this trajectory type and not sure whether this behavior is theoretically plausible.

Table 2 shows examples of nouns belonging to three non-stable trajectory types. Below we explain the semantic change processes for them.

1. The word закладка belongs to the type 1. Its dominant sense in the pre-Soviet period was 'foundation' (as in '*The foundation of the new church building took place yesterday*'). In the Soviet times, the 'bookmark' sense emerged (it was already present before, but very rare). Then, the post-Soviet time period saw the emergence of two

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

[3] Note it does not make much sense to report correlations for individual annotators ('data columns'), since in our crowdworking setup, the columns are not associated with particular persons.

[4] https://github.com/akutuzov/rushifteval_public

Type	Examples	Baseline	Top
1	закладка (‘foundation/bookmark/hidden artifact’), линейка (‘carriage/ruler/series of goods’), центр (‘center’)	0.5	1.0
2	дядька (‘tutor/adult man’), живот (‘life/belly/stomach’), лох (‘salmon/silver-berry/easy victim, stupid person’), роспись (‘list/painting’), ядро (‘cannonball/core/nucleus’)	1.0	1.0
3	полоса (‘stripe/ribbon/lane/runway’), связка (‘ligament/vocal cords/mutual connection’), спутник (‘fellow traveler/satellite/sputnik’), ссылка (‘exile/link’), тачка (‘wheelbarrow/car’), формат (‘format’)	0.4	0.8-1.0

Table 2: Semantic change trajectory types in *RuShiftEval* and the percentage of words with correctly captured type for the baseline and the 4 best shared task submissions (see 4.1).

new senses, both through widening processes: ‘tab’ (in graphical user interfaces) and ‘booby-trapping’ or ‘something hidden’ (often about illegal drugs cached by a distributor). Thus, low relatedness scores are observed across all possible pairs: the word is used differently in each time period.

2. The word ядро can mean either ‘cannonball’ or ‘core/kernel/nucleus’. It belongs to the type 2. In the Soviet period, the first sense almost disappeared (because artillery stopped using cannonballs in the 20th century), while the latter sense became more frequent. After this reduction, the meaning was stable, with no changes in the post-Soviet period.

3. The word тачка (‘wheelbarrow’) belongs to the type 3. It was stable until the end of the Soviet period, but in the post-Soviet times, тачка acquired a new colloquial sense of ‘car’, quite common even in written texts. This lead to divergence from both Soviet and pre-Soviet periods.

Semantic trajectory types could be visualized as time relatedness graphs; see Figure 1. Nodes of the graph are time periods, and edge widths represent the COMPARE score (see 3.2) for each pair of periods.[5] Thus, thicker edges denote stable meaning, while thinner and more transparent edges show a change. Each trajectory type has its own characteristic pattern of edge widths. For example, in the graph for тачка (the rightmost plot), the edges connecting the post-Soviet node to two other nodes are much thinner than the edge between the pre-Soviet and post-Soviet nodes. This signals a change in the post-Soviet times (trajectory type 3).

[5]Note that in most cases it is impossible to use nodes relative positions on the plot to reflect relatedness scores: one can’t change the length of an edge in a triangle without also changing the length of at least one other edge.

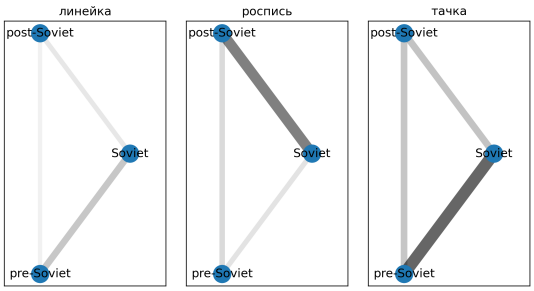

Figure 1: Time relatedness graphs for words belonging to different semantic trajectory types (from left to right): линейка (‘carriage/ruler/series of goods’) (1), роспись (‘list/painting’) (2), тачка (‘wheelbarrow/car’) (3).

Note that the annotation process and the definition of the COMPARE score itself do not guarantee perfect capturing of semantic changes. One example—made clear by the multi-period nature of *RuShiftEval* design—is the word радикал (‘radical’). Its relatedness scores are low across all time period pairs, suggesting that it experienced sequential changes similar to закладка. However, in fact, throughout all the times covered by *RuShiftEval*, this word had the same two persistent senses: political and chemical. Since their probabilities were almost equal, many randomly sampled sentence pairs contained the word радикал in two different senses, which led to low COMPARE scores. In this case, it stems from strong and persistent ambiguity of the word, not from diachronic semantic change. This limitation of the COMPARE metrics was already described in (Schlechtweg et al., 2018).

Another potential flaw is sampling variability. For annotation, we sampled 30 sentences with a target word from each time period for each comparison. Since our relatedness graph has three edges,

each word is represented with two samples. As it turned out, in some cases different samples can yield quite different picture of sense distributions.

Let us manually analyze the word полость ('cavity/hide to cover one's legs in an open cart'). Since horse-driven carts disappeared just a few years after the beginning of the Soviet period, one might expect the second sense to be lost in Soviet times and never to appear again. However, the relatedness between the Soviet and post-Soviet time periods (1.9) is even lower than between the pre-Soviet and Soviet periods (2.2), as if the word experienced another semantic shift. In fact, it is a random sampling artifact. In the 30 sentences from the Soviet period sampled for the 'pre-Soviet:Soviet' pair, only 4 used полость in this archaic sense. But in the 30 sentences *from the same period* sampled for the 'Soviet:post-Soviet' pair, this number grew to 10, 2.5 times more (mostly in fiction texts, where the plot is set in the pre-Soviet times). As a result, the Soviet usage pattern looks like it is different from the post-Soviet one, although in fact no shift has happened (as evident both from linguistic intuition of Russian speakers and from the Fisher exact test which in this case yields $p = 0.13$). The frequency of полость in the Soviet sub-corpus is about 600, so both samples together cover only 10% of the full concordance. Without manually annotating all six hundred occurrences, it is difficult to tell which sample is more representative of the real word usage in the Soviet times. It would be better to increase the sample size as much as possible: 30 is arguably already on the border.

4.1 Trajectory detection task?

The *RuShiftEval* dataset was used to evaluate the systems participating in a shared task on lexical semantic change detection for Russian (Kutuzov and Pivovarova, 2021). How good these submissions are in capturing the trajectory types described in the previous section? In this subsection, we describe a toy experiment to address this question.

For simplicity, we will use only 11 example words from Table 2 which appear in the *RuShiftEval* evaluation set (this excludes закладка, лох and спутник, since they appear in the development set only). Then a set of criteria is established for the system predictions, corresponding to each of the three trajectory types. We consider a system successful in capturing a word with the **trajectory 2** if the predicted relatedness score is higher for the 'Soviet:post-Soviet' pair than for other two pairs. For the words with the **trajectory 3**, the relatedness score for the 'pre-Soviet:Soviet' pair must be the highest among all pairs. For the words with the **trajectory 1**, the percentile ranks of the relatedness scores for all three sub-sets must be below 50 (admittedly, this is an *ad hoc* criterion, but it is used here just to give an example of how the task can be set up). Thus, at least for the trajectory types 2 and 3, this resembles a simple ranking task: not across target words within one period pair, but for one target word across three period pairs. At the same time, the trajectory type 1 (changes in every period) does not quite fit into this frame.

We compared the baseline system (which used static diachronic word embeddings and the local neighbors method from (Hamilton et al., 2016a)) and four best systems (employing contextualized language models: ELMo, BERT or XLM-R). The results are presented in Table 2. All of the best submissions captured the **trajectory 1** for all two target words, but the baseline method failed for центр (its percentile rank in *RuShiftEval*-1 is more than 60). For the **trajectory 3**, the top systems are considerably better than the baseline method. For example, according to the baseline method, полоса experienced its strongest change in the Soviet times, while in fact it was in the post-Soviet period. Only for the **trajectory 2**, the baseline is on par with the winners of the shared task.

This analysis is rather preliminary, but it shows that the systems performance in correctly detecting diachronic trajectories does to some extent correlate with their performance in the 'traditional' semantic change ranking (with binary datasets, like in the SemEval 2020 Shared Task 1). We believe that this can be an interesting sub-task within the larger field of semantic change detection, once more datasets like *RuShiftEval* are available and more formal definitions of 'capturing the trajectory successfully' are developed.

5 Conclusion

We presented *RuShiftEval*, a novel dataset of diachronic semantic changes in Russian nouns across three time periods, using the same set of target words for all comparisons. We also conducted a preliminary analysis of how *RuShiftEval* can be used in tracing diachronic semantic trajectories, and how current change detection systems for Russian deal with this potentially interesting task.

Acknowledgments

The annotation effort for *RuShiftEval* was supported by the Russian Science Foundation grant 20-18-00206. This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

References

Marina Baldanova and Irina Stepanova. 2016. Metaforizatsiya kak put' razvitiya semanticheskikh neologizmov v yazyke interneta (metaphorization as a way of developing semantic neologisms in the language of the internet). *In Russian*.

Nina Basko and Irina Andreeva. 2011. *Slovar' ustarevshey leksiki k proizvedeniyam russkoy klassiki (Dictionary of obsolete vocabulary for the works of Russian classics)*. In Russian.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. 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.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Anna Hätty, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Florence, Italy.

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*, page 95.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. RuShiftEval: a shared task on semantic shift detection for Russian. In print.

Elmi A. M. 2016. Izmeneniya znacheniya odnoznachnykh imen sushchestvitel'nykh, upotreblonnykh v povesti as pushkina "grobovshchik" (changes in the meaning of unambiguous nouns used in as pushkin's story "the undertaker"). *In Russian*.

Vladimir Novikov. 2016. *Dictionary of buzzwords. The linguistic picture of our time*. In Russian.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. *arXiv preprint arXiv:2104.08540*.

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. Association for Computational Linguistics.

Skljarevsky, editor. 1998. *Tolkovyy slovar' russkogo yazyka kontsa XX veka. YAzykovyye izmeneniya. (Explanatory dictionary of the Russian language at the end of the XX century. Language changes)*. In Russian.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*.

A Transliterations of Russian words mentioned in the article

WORD	TRANSLITERATION	TRANSLATION
бригада	brigada	brigade/gang/team
дядька	djadka	uncle/man/(male) tutor
живот	život	stomach/belly/life
закладка	zakladka	foundation/bookmark/hidden artifact
линейка	lineika	carriage/ruler/series of goods
лох	loh	salmon/silver-berry/easy victim
облако	oblako	cloud
полоса	polosa	tripe/ribbon/lane/runway
полость	polost	cavity/foot hide
радикал	radikal	radical
роспись	rospis	mural/signature/list
связка	svjazka	ligament/vocal cords/mutual connection
спутник	sputnik	fellow traveler/satellite/sputnik
ссылка	ssylka	exile/link
стол	stol	table/diet
тачка	tachka	wheelbarrow/car
формат	format	format
центр	tsentr	center
ядро	jadro	cannonball/core/nucleus