RuShiftEval: a shared task on semantic shift detection for Russian

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

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

Abstract

We present the first shared task on diachronic word meaning change detection for the Russian. The participating systems were provided with three sub-corpora of the Russian National Corpus — corresponding to pre-Soviet, Soviet and post-Soviet periods respectively — and a set of approximately one hundred Russian nouns. The task was to rank those nouns according to the degrees of their meaning change between periods.

Although RuShiftEval is in many respects similar to the previous tasks organized for other languages, we introduced several novel decisions that allow for using novel methods. First, our manually annotated semantic change dataset is split in more than two time periods. Second, this is the first shared task on word meaning change which provided a training set.

The shared task received submissions from 14 teams. The results of RuShiftEval show that a training set could be utilized for word meaning shift detection: the four top-performing systems trained or fine-tuned their methods on the training set. Results also suggest that using linguistic knowledge could improve performance on this task. Finally, this is the first time that contextualized embedding architectures (XLM-R, BERT and ELMo) clearly outperform their static counterparts in the semantic change detection task.

Keywords: semantic change detection, Russian, shared task
DOI: 10.28995/2075-7182-2021-20-533-545

RuShiftEval: соревнование по детектированию семантических сдвигов в русском языке

Пивоварова Лидия
Университет Хельсинки
Финляндия
lidia.pivovarova@helsinki.fi

Кутузов Андрей
Университет Осло
Норвегия
andreku@ifi.uio.no

Аннотация

Мы представляем первую дорожку по автоматическому определению изменения значений слов для русского языка. Участники дорожки получили три подкорпуса НКРЯ - досоветский, советский и постсоветский - и список из около ста русских существительных. Задача состояла в ранжировании этих слов по степени семантического сдвига между этими периодами.

Наша дорожка во многих отношениях похожа на предыдущие подобные соревнования, которые организовывались для других языков. Однако мы предложили несколько нововведений, которые позволили участникам протестировать новые подходы к этой задаче. Во-первых, мы опубликовали новый датасет, в котором данные разбиты более чем на два периода. Во-вторых, это первая дорожка по автоматическому определению семантических сдвигов, где участникам был предоставлен тренировочный набор данных.

Дорожка получила более сотни решений от четырнадцати участников. Результаты соревнования продемонстрировали полезность тренировочных данных для определения семантических сдвигов: четыре лучших результа были продемонстрированы моделями, которые тренировались или донастраивались на тренировочных данных. Результаты так же демонстрируют, что использование априорных лингвистических знаний или сложных языковых моделей улучшают показатели в этой задаче.

Ключевые слова: диахронические семантические сдвиги, детектирование семантических изменений

1 Introduction

Words change their semantics over time as a result of combination of various processes that affect language simultaneously. Automatic detection and measuring the degree of meaning change could accelerate research in the history of language and also support a number of text analysis tasks such as information retrieval or media monitoring.

The RuShiftEval shared task is aimed at the comparison of various methods for detection of word meaning shift from diachronic corpora. Recently, two shared tasks for semantic change detection were organized: SemEval Task 1 for English, German, Swedish and Latin [17], and DIACR-Ita for Italian [2]. RuShiftEval is the first attempt to organize such an event with Russian data.

In many aspects, we follow the practices established during the previous shared tasks. However, we introduced several novelties: first, we deal with *three* time periods, namely pre-Soviet, Soviet and post-Soviet; second, we provided the participants with a *training dataset*, thus allowing for using supervised methods.

The shared task is collocated with Dialogue 2021, the 27th International Conference on Computational Linguistics and Intellectual Technologies. The test and development datasets used in RuShiftEval are now publicly available, as well as the evaluation code and the baseline.[1]

2 Related work

Automatic detection of word meaning change is a fast developing research area. The majority of modern approaches utilize distributional *word embeddings* to detect changes in word context over time. Overview of various approaches for this task could be found in the recent surveys [18, 4, 22].

To perform numerical evaluation, the problem is most commonly formulated as following: an input is *a corpus* split into several (usually two) time periods and *a set of words*; the task is *to rank* these words according to the degree of meaning change they have undergone between the periods. The performance is measured by rank correlation between a produced ranking and the gold manually created ranking. Alternatively, the task could be cast as binary classification of words into changed and not-changed classes. In this case, evaluation is also done as comparison against manual annotation.

Thus, manually annotated datasets are key components for development of lexical semantic change models. Since word meaning shift is a *lexicon-level phenomenon*, annotation should take into account many word usages from each periods, making it a time-consuming task. The most recent DUReL framework solves this by annotating pairs of sentences and then computing an averaged metric that generalizes these annotations [16]. We follow this approach in our shared task.

The first shared task on word meaning change detection was organized in 2020 as a part of SemEval conference (SemEval 2020 Task 1). The shared task [17] provided datasets for four languages — English, German, Swedish, and Latin — with several dozens manually annotated words for each language. The task included two subtasks, described above: binary classification and ranking. More than twenty teams participated in it. One of the main results of SemEval 2020 Task 1 was that type-base (static) embeddings are more suitable for *unsupervised* semantic shift detection than more recent contextualized embeddings currently dominating almost all other NLP tasks. Another important observation is a high variety across corpora: a method that yields the best performance for one corpus may not be the best for another one. Another shared task was organized for Italian [2], where the task was binary classification, and the results largely replicated those from the SemEval.

Although RuShiftEval is the first shared task on word meaning change for Russian, semantic shift detection methods have been previously applied to this language, e.g. in [10, 20]. This research is accelerated by publishing of time-specific sub-corpora of the Russian National Corpus (RNC), consisting of sentences from the texts created in the pre-Soviet, Soviet and post-Soviet time periods. Together they cover nearly full RNC.[2] It is important to note that the RuShiftEval organizers are fully aware that 1)

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

[2] The sentence-shuffled version of the RNC split into 3 sub-corpora corresponding to the RuShiftEval time periods was made freely available specifically for this shared task (it is required to sign a license agreement to get access to the corpora): https://rusvectores.org/static/corpora/

the division of Russian language history into these particular periods is not the only possible option and the boundaries could be drawn differently; 2) the RNC itself is not fully representative of the history of Russian. However, some decisions had to be made with respect to the time bin boundaries; the division we chose is at least motivated with regards to historical events and yields sub-corpora of comparable sizes. In the same vein, no Russian corpus other than the RNC is available which is large enough, covers long enough time span, and provides the creation dates for the texts.

These diachronic sub-corpora of the RNC have previously been already used to create the *RuSemShift* dataset [14], which includes two subsets, each of 70 words, manually annotated and ranked according to their change from pre-Soviet to Soviet and from Soviet to post-Soviet times respectively. For the RuShiftEval data annotation, we used the same corpora and followed the same annotation procedure, so *RuSemShift* could be used as a training set by task participants. However, two parts of the *RuSemShift* dataset use different sets of words, while for the shared task we use the same list of words for *all three periods*, in principle allowing to study continuous word sense dynamic across time.

3 Task overview

The shared task focuses on three time periods, naturally stemming from the history of the Russian language and society. The boundary years of 1917 and 1991 were omitted from the annotation due to their transitioning nature:

1. pre-Soviet (1700-1916);
2. Soviet (1918-1990);
3. post-Soviet (1992-2016).

The RuShiftEval dataset consists of 111 Russian nouns (99 in the test set and 12 in the development set), manually annotated with the degrees of their meaning change in three time period pairs:

1. between pre-Soviet and Soviet periods (so called *RuSemShift1* score);
2. between Soviet and post-Soviet periods (so called *RuSemShift2* score);
3. between pre-Soviet and post-Soviet periods (so called *RuSemShift3* score).

We did not rely on any assumption on the dependencies of these three scores and annotated all pairs independently. Note that the resulting RuShiftEval dataset (about 30 000 human judgments in total) is described in more detail in a separate paper [9], so it is only briefly presented here. As per reviewers' suggestions, we provide the full list of target words with their change scores in the Appendix (although we strongly recommend to use the maintained version in our GitHub repository).

The annotation was conducted using crowd-sourcing (Yandex.Toloka platform). It followed the DuReL workflow described in [16]. An annotator had to read and score two sentences containing a target word and belonging to different time periods. The sentences were randomly sampled from the corresponding sub-corpora of the Russian National Corpus. The scores (from 1 to 4) grade semantic relatedness between the target word meanings in two sentences. The 1 score denotes 'the senses are unrelated', and the 4 score denotes 'the senses are identical'.

Then individual scores were accumulated into mean semantic relatedness between word usages from two different time periods; this measure is also known as COMPARE and was introduced in [16]. Basically, it reflects human judgments about such relatedness averaged across about 30 sentence pairs containing the target word. Thus, the lower is the score (the closer it is to 1), the stronger is the degree of semantic change. For each sentence pair, the score was in turn averaged across at least 3 human annotators.

As has been mentioned in Section 2, the *RuSemShift* dataset [14] could be used for training (or simply for sanity check in the *Practice* phase), and we encouraged participants to do this. To find out whether using training data actually helps semantic change detection was one of the purposes of the RuShiftEval shared task. We can now confirm that the answer is positive; see Section 5 for details.

We recommended the participants to use the RNC for their data-driven solutions, since this corpus has been used to annotate the data. They were free to employ any other linguistic sources, and some actually did; again, see Section 5. Submissions of the participants were processed, evaluated and ranked with the

help of Codalab platform.[3]

To help participants to start with the task, we also provided static word embeddings pre-trained on diachronic sub-corpora of the RNC, using the CBOW algorithm [11], with context window size 5 and vector size 300. Each model was published in two variants: trained on raw tokens and trained on lemmas with part of speech tags ('завод_NOUN', etc). These embeddings were used in the baseline solution, which was available as a part of the starting kit for the participants.

4 Evaluation workflow

The task was formulated as a ranking problem, similar to Subtask 2 of the SemEval 2020 Task 1 [17]: a set of Russian words should be ranked according to the strength of their meaning change. Thus, we did not make any binary decisions on whether a word has changed its meaning or not.

Importantly, it was one and the same set of words, for which the participants had to provide 3 semantic change scores per each word. The lower score meant a stronger change; the higher score meant a higher semantic similarity between word usages in different time periods, and thus a weaker change.

During the main *Evaluation* phase (February 22 - March 1, 2021), the participants were provided with a set of 99 target Russian words. For each word, they had to submit three non-negative values, corresponding to semantic change in the aforementioned time period pairs. These values were used to build 3 column-wise rankings: so called *RuSemShift1*, *RuSemShift2* and *RuSemShift3*. Since rank correlation was used as the evaluation metrics, the absolute numerical values of semantic change scores did not matter (only their relative ranks).

During the *Development* phase (February 1 - February 22, 2021), a small development set was provided (12 manually annotated Russian words), and the participants could submit their predictions to get a preliminary estimation of their system performance (no gold labels were openly published).

Before February 1, the shared task was in the *Practice* phase: the participants could submit predictions to the words from the *RuSemShift* test set [14]. This dataset was already public, so the true labels were known to everyone. This phase could be used to sanity check submission routines. There were only two time period pairs, each with its own set of words (this is how *RuSemShift* is built). We remind that in the *Development* and *Evaluation* phases, the participants had *one* set of words and *three* time period pairs.

Each participating team was able to submit up to 10 answers in the Evaluation phase, and up to 1000 answers in the Development phase. Submissions were evaluated using Spearman rank correlation between word ranking produced by a system and a gold ranking obtained in manual annotation. Thus, for each system we computed three correlations, for each of the time period pairs. The final ranking of the systems is based on averaging of the three scores.

5 Shared task results

In the Evaluation phase, we received submissions from 14 users (some of them in 4 different teams). Table 1 shows the performance of top submissions from each user or team (we give the name of the team by default or the name of the individual participant, if no team was associated with this submission). The teams are ranked by their average scores.

Some initial comments are due with regards to this table:

1. The baseline solution employed lemmatized diachronic embeddings trained on the Russian National Corpus[4] and the simple local neighborhood method from [5].
2. The differences between the first and the second best performing systems are not statistically significant according to the Fisher test; the differences between the second and the third systems are statistically significant at $p = 0.06$ for RuSemShift1 only. However, the differences between the top three systems and the rest of the submissions are all statistically significant.
3. Using median score instead of average score does not substantially change the ranking.

[3] https://competitions.codalab.org/competitions/28340

[4] These embeddings and diachronic corpora were available to all participants.

	Team	RuSemShift1	RuSemShift2	RuSemShift3	Mean	Type
1	**GlossReader**	0.781	**0.803**	**0.822**	**0.802**	token
2	**DeepMistake**	**0.798**	0.773	0.803	0.791	token
3	vanyatko	0.678	0.746	0.737	0.720	token
4	**aryzhova**	0.469	0.450	0.453	0.457	token
5	Discovery	0.455	0.410	0.494	0.453	token
6	**UWB**	0.362	0.354	0.533	0.417	type
7	dschlechtweg	0.419	0.373	0.383	0.392	type
8	jenskaiser	0.430	0.310	0.406	0.382	token
9	**SBX-HY**	0.388	0.281	0.439	0.369	type
	Baseline	0.314	0.302	0.381	0.332	type
10	svart	0.163	0.223	0.401	0.262	type
11	**BykovDmitrii**	0.274	0.202	0.307	0.261	token
12	fdzr	0.217	0.251	0.065	0.178	type

Table 1: Evaluation phase leaderboard (Spearman rank correlations). The Type column shows the type of the used distributional embeddings.

4. Bold names denote teams or individual participants who submitted papers with the description of their systems. For other participants, we rely on the contents of the 'Description' field in their Codalab submissions.
5. The DeepMistake team made several submissions of essentially the same system with varying hyperparameters; we show only the best one.
6. The SBX-HY team made a minor technical mistake, and their correlation scores were negative. Our opinion is that this does not undermine the developed system itself, so we show the absolute values in Table 1, and rank submissions accordingly.

5.1 Participating systems overview

Below, we give the descriptions of the participating systems. First, let us look at the submissions described in the submitted papers.

GlossReader [13] relied on the pretrained multilingual XLM-R language model [21]. On top of it, they trained a word sense disambiguation (WSD) system on English WSD datasets, using learned representations of sense definitions. Interestingly, this system shows excellent performance on Russian lexical semantic change data as well. Essentially, this participant reproduced the RuShiftEval annotation effort, replacing human judgments with the distances between XML-R contextualized embeddings of the target words. Additionally, a linear regression was trained on the *RuSemShift* dataset to convert vector distance values into relatedness scores (from 1 to 4).

DeepMistake [3] used the multilingual XLM-R as well, and also pre-trained on English WSD datasets, but without explicitly predicting senses. Similarly to **GlossReader**, they additionally fine-tuned this model on the *RuSemShift* using linear regression for mapping to relatedness scores.

aryzhova [15] tried both ruBERT [7] and ELMo contextualized embeddings.[5] Interestingly, in their experiments ELMo outperformed BERT. Note, however, that **aryzhova** system is different from **vanyatko** (described below) in that it does not fine-tune BERT or ELMo: instead, it calculates the average cosine similarity between target word embeddings (sometimes with the addition of the neighboring word

[5]The ELMo models for Russian were borrowed from the RusVectōrēs service.

tokens) in the sampled sentence pairs, reproducing the *APD* method from [8]. Another interesting experiment reported in the paper from this participant is using 'grammatical vectors' corresponding to the frequencies of 12 morphological forms of Russian nouns (6 cases and singular/plural forms). They report that the cosine similarities between such vectors calculated on different time bins improved the performance of relatedness score classifier (trained and evaluated on the *RuSemShift* dataset).

UWB [12] this team employed traditional 300-dimensional static word embedding (in particular, fastText). Orthogonal Procrustes and Canonical Correlation Analysis (CCA) were used for alignment, with CCA showing somewhat better results. The semantic change score was calculated as simple cosine similarity between word vectors across different time periods.

SBX-HY [6] again used static word embeddings, but in this case instead of post-training alignment, they relied on Temporal Referencing approach [19], successfully used for semantic change detection with other languages. In this approach, the target words are augmented with time period labels, and then one embedding model is trained on all available data. Hyper-parameters were selected based on the *RuSemShift* dataset. Interestingly, with the RuShiftEval data, Temporal Referencing barely managed to outperform the organizers' baseline, which is an interesting negative result.

BykovDmitrii [1] employed an interesting approach with lexical substitutes produced by the multilingual XLM-R as a masked language model. These substitutes were then clustered into senses and the divergence between clusters from different time periods was used as the semantic change score. This particular approach failed, but in the post-evaluation phase, the participant managed to significantly improve their result by skipping the clustering step and instead directly comparing bags of lexical substitutes (see more in their paper).

Now let us briefly describe the systems which did not submit papers, based on their descriptions in Codalab. **Vanyatko** employed the RuBERT model. They fine-tuned RuBERT with sentence pairs as inputs and relatedness scores (from 1 to 4) as outputs. Similar to **GlossReader** and **DeepMistake**, **vanyatko** tried to reproduce human annotation process. The **Discovery** team used BERT with ensemble of Average Pairwise Cosine Distance and Cosine Distance of averaged embeddings. **Dschlechtweg** trained regression on the labeled training examples from *RuSemShift* with SGNS embeddings. **Jenskaiser** also employed static SGNS embeddings and Temporal referencing or 'word injection (WI)'. They got results very similar to **SBX-HY**. Finally, **svart** used orthogonal Procrustes and cosine distances with the lemmatized word2vec embeddings provided by the organizers, and **fdzr** again relied on temporal referencing.

6 Discussion

We believe the results of the RuShiftEval are interesting for the lexical semantic change detection field in at least four aspects.

1 This is the first time the systems based on *contextualized embeddings* top the leaderboard. In both SemEval 2020 Task 1 [17] and DIACR-ITA [2], type embedding (or 'static' embedding) based architectures clearly won the rankings. But at the RuShiftEval, five top performing systems use pre-trained contextualized ('token-based') models: XLM-R, BERT and ELMo. In the previous work, the researchers in the field expressed doubts about the abilities of token embeddings with relation to semantic change detection. It seems that at least in the case of RuShiftEval, they are perfectly able to solve the task better than their static counterparts. However, the best performing teams introduced completely novel approaches to the problem, so the distinction between our results and results of the previous tasks lies in the difference between models rather than between embeddings themselves.

2 Surprisingly, the first and the second best submissions relied on the contextualized XLM-R model [21], which was not even specifically trained for processing Russian data. Its training corpus included texts in about 100 languages. Russian is well represented there but is far from being the largest in absolute size. The results of our shared task show that multilingual models like XLM-R can be very

successfully applied to semantic change detection for Russian (and arguably for many other languages): their transferability is extremely high.

Interestingly, at the SemEval 2020 Task 1, the attempts to use XLM-R did not end up very well: the system based on it ended up 7th in the Subtask 2 (closest to RuShiftEval), well below the type-based architectures. One of the reasons for this can be the next insightful outcome of RuShiftEval:

3 Using training data helps lexical semantic change detection. As already said, the *RuSemShift* dataset [14] was publicly available by the beginning of RuShiftEval, and the participants were free to use it as they saw fit. The annotation procedures were identical for *RuSemShift* and the shared task test sets. Thus, one of the aims of RuShiftEval was to find out whether using previously annotated data can improve the performance of semantic change ranking. As it turns out, it definitely can. Four top systems all train or fine-tune on *RuSemShift*. This was the first semantic change detection shared task to introduce such a setup. At the same time, using unsupervised methods with parameters fine tuning on the training set does not seem to be a productive strategy.

4 Finally, at least two participants (both in the top of the leaderboard) used explicit linguistic knowledge in addition to statistical distributional models. In particular, **GlossReader** (the winner of the task) fine-tuned their XLM-R model to select a definition (a gloss) from the WordNet, that is most appropriate for a particular target word occurrence [13]. Note that it was not the plain old classification: the model directly processed the definitions themselves as sequences of words. Another example is **aryzhova** who employed a linguistic intuition that semantic change is often linked to fluctuations in the frequency of different grammatical forms [15]. We believe using linguistic knowledge is an interesting direction for future development of the semantic change detection field.

It is important to note that the observations above are applicable only to the shared task setup used in RuShiftEval: that is, ranking words by the degree of semantic change estimated with the COMPARE measure calculated on human annotations conducted within the DUReL framework. Actually, many of the top-performing systems essentially reproduced the annotation process with large language models, which seems to be successful even though they could not know which particular sentences were sampled for manual annotation. With other evaluation setups, different approaches could be at the top. As an example, it is known that the COMPARE measure is much influenced by sense frequencies and can easily overlook changes occurring to rare senses — either their appearance or disappearance. If the systems were evaluated based on explicit senses they managed to detect, clustering-based approaches would arguably rank much higher.

7 Conclusion

In this paper, we summarized the outcome of RuShiftEval: the first shared task on lexical semantic change detection for Russian. The purpose of the shared task was twofold: first, to evaluate current state-of-the-art methods in semantic change detection on Russian data, and second, to explore the possibilities of *supervised* semantic change detection. This was ensured by the prior existence of *RuSemShift* dataset, annotated in exactly the same way as our testing data.

The results of the shared task show that training on existing semantic change data is indeed useful and can significantly boost evaluation scores. In absolute values, the correlations with human judgments achieved by the RuShiftEval participants are much higher than those demonstrated in the SemEval 2020 Task 1 across English, Latin, German and Swedish (the best system there yielded 0.527). Note that although *RuSemShift* (used as a training set) and RuShiftEval (used as a development and a test set) are annotated similarly, they are not splits of one and the same dataset. Thus, we believe this finding to be reliable and expect it to hold for other languages as well.

Another interesting outcome of RuShiftEval is the strong victory of contextualized (token-based) embedding architectures over static (type-based) ones. This is different from the results of previous shared tasks on semantic change detection, and we believe this means the community has finally learned how to properly use contextualized embeddings for this task. This is even more impressive considering the fact that the winning systems used the multilingual XLM-R instead of a Russian-specific model.

Despite these substantial findings, our shared task has just started to pave the way for studying approaches to automatic semantic change detection in Russian. Our evaluation setup (ranking by aggregated COMPARE score) cannot capture the entire spectrum of semantic change. This linguistic phenomenon is extremely complex, and we are hoping that future shared tasks will try to account for that.

Acknowledgments

The annotation effort for this shared task was supported by the Russian Science Foundation grant 20-18-00206. We are especially grateful to Valery Solovyev (Kazan Federal University). This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

References

[1] Arefyev Nikolay, Bykov Dmitrii. An Interpretable Approach to Lexical Semantic Change Detection with Lexical Substitution // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[2] DIACR-Ita@ EVALITA2020: Overview of the EVALITA2020 diachronic lexical semantics (DIACR-Ita) task / Pierpaolo Basile, Annalina Caputo, Tommaso Caselli et al. // Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org. — 2020.

[3] DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model / Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov et al. // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[4] Diachronic word embeddings and semantic shifts: a survey / Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, Erik Velldal // Proceedings of the 27th International Conference on Computational Linguistics. — 2018. — P. 1384–1397.

[5] Hamilton William L., Leskovec Jure, Jurafsky Dan. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. — Austin, Texas : Association for Computational Linguistics, 2016. — Nov. — P. 2116–2121. — Access mode: https://www.aclweb.org/anthology/D16-1229.

[6] Hengchen Simon, Viloria Kate, Indukaev Andrey. SBX-HY at RuShiftEval 2021: Doveriay, no proveriay // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[7] Kuratov Yury, Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2019. — Access mode: http://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf.

[8] Kutuzov Andrey, Giulianelli Mario. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection // Proceedings of the Fourteenth Workshop on Semantic Evaluation. — Barcelona (online) : International Committee for Computational Linguistics, 2020. — Dec. — P. 126–134. — Access mode: https://www.aclweb.org/anthology/2020.semeval-1.14.

[9] Kutuzov Andrey, Pivovarova Lidia. Three-part diachronic semantic change dataset for Russian // Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change. — online : Association for Computational Linguistics, 2021.

[10] Measuring Diachronic Evolution of Evaluative Adjectives with Word Embeddings: the Case for English, Norwegian, and Russian / Julia Rodina, Daria Bakshandaeva, Vadim Fomin et al. // Proceedings of the 1st International Workshop on Computational Approaches to Historical Language

Change. — Florence, Italy : Association for Computational Linguistics, 2019. — Aug. — P. 202–209. — Access mode: https://www.aclweb.org/anthology/W19-4725.

[11] Mikolov Tomas, Yih Wen-tau, Zweig Geoffrey. Linguistic Regularities in Continuous Space Word Representations // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Atlanta, Georgia : Association for Computational Linguistics, 2013. — Jun. — P. 746–751. — Access mode: https://www.aclweb.org/anthology/N13-1090.

[12] Priban Pavel, Pražák Ondřej, Taylor Stephen. UWB@RuShiftEval: Measuring Semantic Difference as per-word Variation in Aligned Semantic Spaces // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[13] Rachinskiy Maxim, Arefyev Nikolay. Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[14] Rodina Julia, Kutuzov Andrey. RuSemShift: a dataset of historical lexical semantic change in Russian // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 1037–1047. — Access mode: https://www.aclweb.org/anthology/2020.coling-main.90.

[15] Ryzhova Anastasiia, Ryzhova Daria, Sochenkov Ilya. Detection of Semantic Changes in Russian Nouns with Distributional Models and Grammatical Features // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[16] Schlechtweg Dominik, Schulte im Walde Sabine, Eckmann Stefanie. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). — New Orleans, Louisiana : Association for Computational Linguistics, 2018. — Jun. — P. 169–174. — Access mode: https://www.aclweb.org/anthology/N18-2027.

[17] SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection / Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen et al. // Proceedings of the Fourteenth Workshop on Semantic Evaluation. — Barcelona (online) : International Committee for Computational Linguistics, 2020. — Dec. — P. 1–23. — Access mode: https://www.aclweb.org/anthology/2020.semeval-1.1.

[18] Tang Xuri. A state-of-the-art of semantic change computation // Natural Language Engineering. — 2018. — Vol. 24, no. 5. — P. 649–676.

[19] Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change / Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Florence, Italy : Association for Computational Linguistics, 2019.

[20] Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines / Vadim Fomin, Daria Bakshandaeva, Julia Rodina, Andrey Kutuzov // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2019. — P. 203–218. — Access mode: http://www.dialog-21.ru/media/4598/fominvplusetal-116.pdf.

[21] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 8440–8451. — Access mode: https://www.aclweb.org/anthology/2020.acl-main.747.

[22] A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains / Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, Sabine Schulte im Walde // Pro-

ceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 732–746. — Access mode: https://www.aclweb.org/anthology/P19-1072.

A RuShiftEval gold datasets

1. 1-2: change from the pre-Soviet to Soviet times;
2. 2-3: change from the Soviet to the post-Soviet times;
3. 1-3: change from the pre-Soviet to the post-Soviet times.

DEVELOPMENT SET					
WORD	**TRANSLITERATION**	**TRANSLATION**	**CHANGE SCORE**		
			1-2	**2-3**	**1-3**
верховье	verhovje	upper reaches	3.68	3.74	3.87
возраст	vozrast	age	3.47	3.69	3.58
завод	zavod	factory/breeding farm	3.22	3.65	3.52
закладка	zakladka	foundation/bookmark/hidden artifact	1.93	1.74	1.74
земля	zemlja	earth/land/soil	2.83	2.8	2.28
лох	loh	salmon/silver-berry/easy victim	1.07	2.94	1.04
помощник	pomoštšnik	assistant	3.38	3.56	3.28
пролетарий	proletarij	proletarian	3.4	3.58	3.44
промышленность	promyšlennost’	industry	3.24	3.51	3.47
радикал	radikal	radical	1.42	1.68	2.01
спутник	sputnik	fellow traveler/satellite/sputnik	2.96	1.81	1.94
четверть	tšetvert	quarter	2.25	2.96	3.07

TEST SET					
WORD	**TRANSLITERATION**	**TRANSLATION**	**CHANGE SCORE**		
			1-2	**2-3**	**1-3**
авторитет	avtoritet	authority/prestige	3.23	2.95	2.84
амбиция	ambitsia	ambition	3.11	3.44	3.33
апостол	apostol	apostle/disciple	3.49	3.42	3.42
благодарность	blagodarnost’	gratitude/appreciation/thankfulness	3.23	3.56	3.65
блин	blin	pancake/damn	3.21	1.66	2.57
блондин	blondin	blonde (male)	3.94	3.92	3.95
брат	brat	brother	3.22	3.01	3.27
бригада	brigada	brigade/gang/team	2.8	2.71	3.08
веер	vejer	fan	2.55	2.43	2.44
век	vek	century/age	3.2	3.21	2.98
вызов	vyzov	call/challenge/summons	2.17	2.1	2.03
головка	golovka	(small) head	2.20	1.67	2.19
грех	greh	sin/fault	3.48	2.98	2.92
дух	duh	spirit/ghost/scent	2.32	1.63	1.88
дядька	djadka	uncle/man/(male) tutor	2.59	3.03	2.68
дядя	djadja	uncle/man	3.37	3.39	3.29
железо	železo	iron	2.2	2.56	2.40
жесть	žest	tin/horror	3.23	3.38	3.41
живот	život	stomach/belly/life	2.91	3.44	2.76
заблуждение	zabluždenije	delusion	3.5	3.62	3.55
издательство	izdatelstvo	publishing house	3.53	3.86	3.45
итальянец	italjanets	Italian	3.70	3.6	3.67
кабан	kaban	boar	3.6	3.32	3.30
карман	karman	pocket	3.46	3.47	3.56
крушение	krušenije	collapse	2.75	2.78	2.6
крыша	kryša	roof	3.57	3.0	2.82
кулиса	kulisa	wings	3.16	3.17	3.24
лечение	letsenije	cure	3.65	3.74	3.68
линейка	lineika	carriage/ruler/series of goods	1.87	1.37	1.22
лишение	lišenije	deprivation	2.94	2.07	2.33
локоть	lokot	elbow	3.27	3.41	3.73
любовник	ljubovnik	lover	3.45	3.71	3.65
любовь	ljubov	love	3.29	2.97	3.07
маньяк	manjak	maniac	3.08	3.01	3.11
монстр	monstr	monster	2.6	2.38	2.04
наволочка	navolotška	pillowcase	3.61	3.83	3.92
название	nazvanije	name/title	3.48	3.48	3.43

WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
наложение	naloženije	imposition	1.95	2.06	1.78
облако	oblako	cloud	3.17	3.0	3.16
обоснование	obosnovanije	grounds	3.74	3.5	3.58
огонь	ogon	fire	2.10	2.13	2.46
памятник	pamjatnik	monument	2.88	2.83	2.82
пафос	pafos	pathos	3.34	3.27	3.41
писк	pisk	squeak	3.21	3.0	2.53
план	plan	plan	2.67	2.27	2.54
поколение	pokolenie	generation	3.43	3.58	2.8
половинка	polovinka	half	2.51	2.75	2.62
полоса	polosa	stripe/ribbon/lane/runway	1.83	1.5	1.41
полость	polost	cavity/foot hide	2.23	1.88	2.56
полукруг	polukrug	semicircle	2.78	3.13	3.08
понедельник	ponedelnik	Monday	3.77	3.86	3.86
поставщик	postavštšik	supplier	3.56	3.44	3.25
поэзия	poezia	poetry	3.22	3.66	3.56
правда	pravda	truth/reality	3.13	2.94	2.96
предательство	predatelstvo	betrayal	3.67	3.48	3.8
прецедент	pretsedent	precedent	3.52	3.8	3.53
проникновение	proniknovenije	penetration	2.75	2.68	2.53
прорыв	proryv	breakthrough	2.08	2.05	2.05
путь	put’	way	2.41	2.04	2.3
размышление	razmyšlenije	reflection	3.52	3.55	3.62
ранец	ranets	backpack	3.6	3.53	3.38
расчет	rastšot	calculation/settlement	2.0	1.95	2.0
риторика	ritorika	rhetoric	3.06	2.95	2.93
роспись	rospis	mural/signature/list	1.43	2.98	1.57
сверстник	sverstnik	age-mate	3.86	3.86	3.82
связка	svjazka	ligament/vocal cords/mutual connection	2.33	1.96	1.77
собрат	sobrat	fellow	3.45	3.32	3.32
совершенство	soveršenstvo	perfection	2.95	3.16	3.08
советчик	sovettik	adviser	3.22	3.48	3.42
союзник	sojyznik	ally	3.66	3.47	3.75
список	spisok	list	3.28	3.31	3.05
ссылка	ssylka	exile/link	2.87	2.04	1.93
стена	stena	wall	3.1	3.16	3.32
стипендия	stipendia	scholarship	3.8	3.71	3.56

WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
стол	stol	table/diet	3.50	3.16	3.25
тачка	tachka	wheelbarrow/car	3.39	1.94	1.89
тупик	tupik	deadlock	3.17	2.83	3.14
увольнение	uvolnenie	furlough/layoff	3.21	3.53	3.32
углеводород	uglevodorod	hydrocarbon	3.68	3.31	3.2
удобство	udobstvo	convenience	2.43	2.42	2.51
уклад	uklad	setup	3.33	3.42	3.42
университет	universitet	university	3.54	3.7	3.72
установление	ustanovlenie	establishment	2.28	2.26	2.40
фаворит	favorit	favorite	3.15	2.53	2.84
формат	format	format	2.84	2.02	1.81
формула	formula	formula	2.81	2.26	2.57
хозяйка	hozjaika	hostess	3.25	3.22	3.42
хор	hor	choir	2.66	2.87	2.22
хрен	hren	horseradish/dick/old fart	1.8	2.26	1.6
цензура	tsenzura	censorship	3.49	3.46	3.45
центр	tsentr	center	2.14	1.83	1.87
цифра	tsifra	digit/number	2.96	2.87	3.19
частица	tšastitsa	part/particle	1.96	2.33	2.2
чек	tšek	check	2.37	1.95	2.65
штаб	štab	headquarters	3.63	3.38	3.5
эшелон	ešelon	echelon	2.92	2.28	2.33
юбилей	jubilei	anniversary/jubilee	3.68	3.7	3.78
ядро	jadro	cannonball/core/nucleus	1.55	1.91	1.47
ясли	jasli	nursery/manger	2.28	3.0	1.9