Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

SHIFTRY: WEB SERVICE FOR DIACHRONIC ANALYSIS OF RUSSIAN NEWS

Kutuzov A. (andreku@ifi.uio.no)
University of Oslo, Oslo, Norway

Fomin V. (wadimiusz@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

Mikhailov V. (Mikhaylov.V.Nikola@sberbank.ru)
National Research University Higher School of Economics; Sberbank, Moscow, Russia

Rodina J. (julia.rodina97@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

We present the ShiftRy web service. It helps to analyze temporal changes in the usage of words in news texts from Russian mass media. For that, we employ diachronic word embedding models trained on large Russian news corpora from 2010 up to 2019. The users can explore the usage history of any given query word, or browse the lists of words ranked by the degree of their semantic drift in any couple of years. Visualizations of the words' trajectories through time are provided. Importantly, users can obtain corpus examples with the query word before and after the semantic shift (if any). The aim of ShiftRy is to ease the task of studying word history on short-term time spans, and the influence of social and political events on word usage. The service will be updated with new data yearly.

Key words: diachronic word embeddings, semantic shifts, web service, Russian, news

DOI: 10.28995/2075-7182-2020-19-500-516

SHIFTRY: ВЕБ-СЕРВИС ДЛЯ АНАЛИЗА ДИАХРОНИЧЕСКИХ ИЗМЕНЕНИЙ СЛОВ В НОВОСТНЫХ ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

Кутузов А. (andreku@ifi.uio.no)
Университет Осло, Осло, Норвегия

Фомин В. (wadimiusz@gmail.com)
Высшая Школа Экономики, Москва, Россия

Михайлов В. (Mikhaylov.V.Nikola@sberbank.ru)
Высшая Школа Экономики; Сбербанк, Москва, Россия

Родина Ю. (julia.rodina97@gmail.com)
Высшая Школа Экономики, Москва, Россия

В статье представлен веб-сервис ShiftRy, который помогает анализировать диахронические изменения в употреблении слов в российских новостных текстах. Мы используем диахронические дистрибутивно-семантические векторные модели, обученные на большом русскоязычном новостном корпусе, который включает в себя тексты за временной период с 2010 по 2019 годы. Пользователь веб-сервиса имеет возможность не только исследовать семантическую историю любого заданного слова, но и проанализировать списки изменившихся слов, упорядоченных по степени семантических изменений между двумя рассматриваемыми годами. Кроме того, ShiftRy предоставляет визуализации семантических «траекторий» движения слов во времени и генерирует корпусные примеры употребления заданного слова до и после семантического сдвига, если он имел место. Исследователи истории языка могут использовать ShiftRy для облегчения анализа динамики словоупотребления на коротких промежутках времени, а также для изучения влияния социально-политических событий на семантику русской лексики. Мы планируем обновлять ShiftRy ежегодно.

Ключевые слова: диахронические эмбеддинги, семантический сдвиг, веб-сервис, русский язык, корпус новостных текстов

1. Introduction

In this paper, we describe *ShiftRy*: the web service aimed to help the analysis of temporal shifts in word usage in Russian news texts[1]. It leverages several existing methods of semantic shift detection using diachronic word embedding models.

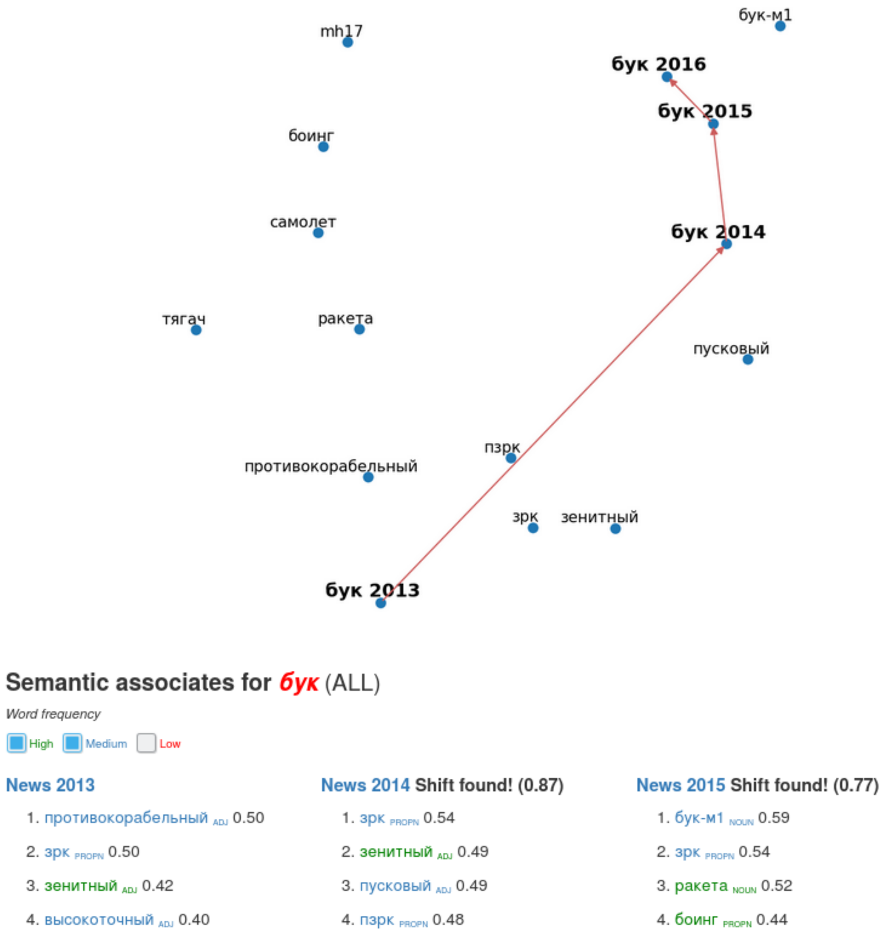
Words change their meaning over time. These processes can be triggered either by linguistic or by social and cultural causes [6]. In the latter case, it is often not a discrete lexicographic shift (acquiring a new sense, losing an old one or changing an existing one), but rather a change in contextual usage, an attitude, or associations bound to an object in public opinion. A typical example is a country name which can acquire new associations after a military campaign starts in this country. Such changes are often short-term and can be traced on the time periods spanning across years or even months (unlike 'linguistic' semantic shifts which usually occur on the scale of decades or centuries). This is the most common type of semantic change found in our material: Russian news texts.

News texts are deeply interconnected with social, cultural and technological dynamics of human society. Additionally, time-annotated news corpora are relatively easy to obtain. Thus, applying diachronic computational semantic methods to this type of texts is interesting both for evaluating the corresponding algorithms and for getting insights about changes in society reflected through the lens of media.

ShiftRy performs this kind of analysis on Russian news texts published in the time span between 2010 and 2019 (in the future, it will be updated yearly). Russian material is attractive for our purpose, since it includes both state-sponsored media always supporting the stance of the authorities, and relatively free media presenting alternative points of view (see more on that in **Section 3**). Additionally, it provides numerous examples of word usage drift, caused by political and social events in the last decades, both outside and inside Russia.

The intended type of *ShiftRy* usage is asking it about a particular query word. The service produces the word's semantic trajectory across the whole 2010–2019 time span, or across any sub-span determined by the user. It also tries to guess whether any substantial semantic shifts had happened in the course of these years (see more about the classifier performing this job in **subsection 5.1**). For deeper analysis, the user can inspect the lists of the nearest neighbors (semantically similar words) for the query word in the distributional representations trained on different time periods. Additionally, we provide samples of sentences from the corresponding corpora containing the query word, so that its real contexts can be examined. These examples are mostly intended to be used with the words which had undergone a semantic shift. For this reason, they are sampled in a way which maximizes the semantic difference between examples from different time bins (see **subsection 5.3**).

[1] https://shiftry.rusvectores.org



Semantic associates for *бук* (ALL)

Word frequency

☑ High ☑ Medium □ Low

News 2013	News 2014 Shift found! (0.87)	News 2015 Shift found! (0.77)
1. противокорабельный ADJ 0.50	1. зрк PROPN 0.54	1. бук-м1 NOUN 0.59
2. зрк PROPN 0.50	2. зенитный ADJ 0.49	2. зрк PROPN 0.54
3. зенитный ADJ 0.42	3. пусковый ADJ 0.49	3. ракета NOUN 0.52
4. высокоточный ADJ 0.40	4. пзрк PROPN 0.48	4. боинг PROPN 0.44

Figure 1: *ShiftRy* interface with the semantic trajectory and time-specific nearest neighbours of the Russian proper name 'Бук' 'Buk' (missile system) from 2013 to 2016

Figure 1 shows a screenshot of *ShiftRy* interface, featuring the analysis of the Russian proper name 'Бук' denoting the 'Buk' missile system (or a type of trees, but this sense is rare in news texts). Note how its typical neighbours drift from other missile systems to the words 'боинг' ('Boeing') and 'MH-17'. This reflects the downing of the Malaysia Airlines Boeing Flight MH-17 in 2014, allegedly by Russia-sponsored Eastern Ukraine insurgents using this particular missile system. Note that this is a case of cultural context variation: the word did not acquire a new lexicographic sense. However, it had significantly changed its typical contexts and the meaning behind it had undoubtedly shifted for Russian speakers. Thus, we believe this to be an example of diachronic semantic change.

Our contributions in this paper are:

1. We provide a publicly available web service which allows to visually explore temporal changes in word usage, based on yearly Russian news texts (from 2010 to 2019).
2. A modification of Procrustes alignment for word embedding models is described, allowing to preserve full models' vocabularies, not only their intersection.
3. Diachronic word embedding models trained on the Russian news corpora and aligned using the aforementioned technique are released.
4. We describe a simple novel method of sampling sentences from corpora to better illustrate lexical semantic change.

The rest of the paper is organized as follows. In **Section 2**, we put our project in the context of the previous related work. **Section 3** describes the news corpora we employed for training word embedding models, which are in turn presented in **Section 4**, along with the explanation of our alignment method. **Section 5** provides a detailed account of *ShiftRy* features, including the word trajectory visualizations, the classifier and the corpus examples sampling. Finally, in **Section 6**, we conclude and account for the future work.

2. Previous Work

Recent years saw a growing amount of research aimed at automatically tracing diachronic semantic shifts and detecting their nature. This increase in popularity arguably started with the paper [6], following the advent of dense word embeddings [16], although some research was published even before that. Because of space limit, we do not list here all the relevant publications. Instead, we refer the reader to the surveys in [12] and [21], and to the proceedings of the first Workshop on Computational Approaches to Historical Language Change [20].

Among recent research on semantic shifts in Russian, one can mention [2], who extensively described 20 hand-picked examples of words having changed their meaning across two centuries, and [4], who evaluated several embedding-based semantic shift detection approaches on both long-term and short-term time spans. *ShiftRy* mostly uses the Procrustes alignment approach [5], which outperformed its competitors both for Russian in [4] and for English in [18].

Note that, in principle, semantic change modeling problem can also be cast in the context of lexical databases, like WordNet [22] for English or RuWordNet [13] for Russian. In this case, a semantic shift means a new synset (or one of the existing synsets) is attached to a word or, vice versa, detached from it: it is then a matter of a word acquiring or losing senses. This approach arguably can return more interpretable results, but, unfortunately, it unavoidably requires expensive human annotation. Thus, any manually curated lexical database is always inherently limited. In comparison, the distributional approach taken by *ShiftRy* is completely unsupervised, and the only text annotation it requires is the time of creation. Thus, it can in theory handle shifts occurring to any words of interest.

Russian news texts were studied in [3], who discovered that in state-sponsored Russian media, the United States are regularly mentioned more frequently the month directly following an economic downturn in Russia. This represents an interesting analysis of propagandist agenda-setting and the frames employed in these texts. However, unlike our work, [3] did not focus on meaning or usage changes, and limited themselves to only one media source.

The web service closest to *ShiftRy* is arguably *JeSeMe*[2] [7]. It allows to trace diachronic semantic trajectory of query words across several English and German corpora, mostly with time spans of 2 or 3 centuries. Instead, we deal with Russian language, and time spans of years, making it possible to study small cultural changes, including those influenced by political events.

The *ShiftRy* code is based on the *WebVectors* framework [11], which we substantially extended and adapted to the task of semantic change analysis. More specifically, we added new visualizations of semantic trajectories, and integrated semantic shift classifier and corpus examples sampling.

3. Collection of data

To collect training corpora, we crawled texts published from 2010 to 2019 from the Russian news web sites listed in **Table 1**.

Table 1: Training corpora sources and stances

Nr.	Title	URL	Stance
1	Fontanka.ru	https://www.fontanka.ru/	Opposition
2	Gazeta.ru	https://www.gazeta.ru/	Loyal
3	Interfax	https://www.interfax.ru/	Neutral
4	Izvestia	https://iz.ru/	Loyal
5	KP	https://www.kp.ru/	Loyal
6	Lenta.ru	https://lenta.ru/	Mixed
7	Novaya Gazeta	https://novayagazeta.ru/	Opposition
8	N + 1	https://nplus1.ru/	Scientific
9	RBC	https://www.rbc.ru/	Neutral
10	The Village	https://www.the-village.ru/	Opposition

The selection of the source web sites does not claim to cover the whole media space of Russian news (that would be impossible within the natural constraints of an academic paper). However, it is balanced with regards to the sources' political stance. Sources 2, 4 and 5 as a rule are loyal to the point of view of the Russian authorities, while sources 1, 7 and 10 regularly publish critical opinions (source 7 is generally considered to be one of the most influential opposition newspapers). Source 6 belonged more to the 'opposition' category until March 2014, when its top staff was forcefully changed to include managers loyal to the authorities. Since that time,

[2] http://jeseme.org

source 6 falls into the 'loyal' category. Sources 3 and 9 position themselves as a neutral news-wire service and a business analytic media correspondingly. Finally, source 8 publishes scientific news. The stances are listed in the right column of **Table 1**.

The collected corpora were tokenized, lemmatized and PoS-tagged using UDPipe 2.3 [19]. After removing functional words, the full corpus for 10 years contains about 156 million tokens, with yearly sub-corpora sizes varying from 9 million (2014) to 20 million (2015) tokens.

4. Word Embedding Models

Continuous Bag-of-Words (CBOW) word embedding models [16] were trained on each of the yearly news corpora, with vector size 300, symmetric context window size 5, and no down-sampling, following the hyper-parameters from [4].

One important change compared to the previous work was that each yearly model was initialized with word vectors trained on the full Russian National Corpus (RNC). The RNC size is about 250 million word tokens, and it is very well balanced with respect to text genres. The reason for this pre-training was that our yearly news corpora are comparatively small, and thus the models trained solely on them might end up in a sub-optimal state. We hypothesized that initializing them with quality vectors trained on a representative Russian corpus would provide a stabler foundation for further learning year-specific embeddings.

The RNC model was trained for 10 epochs with the vocabulary size of 50,000 (most frequent) words. Since news texts do not constitute the majority of the RNC, this vocabulary certainly lacks many proper names and toponyms frequent in news pieces. This is why when updating the RNC model with the co-occurrence data from the yearly news corpora, we expanded the vocabulary, increasing its maximum size from 50,000 to 100,000 words. It means that if a new (not present in the RNC model) word occurred in a yearly corpus with the frequency 10 or more, it was added to the yearly model vocabulary. As a result, the sizes of the yearly models' vocabularies vary from about 60 thousand to about 95 thousand words. This operation can be thought of as a sort of 'fine-tuning' the original RNC model, mixing its general world knowledge with the knowledge about recent political and social events from the yearly news corpora. To make the news texts influence the resulting models more, we performed this 'fine-tuning' in 20 epochs, instead of 10.

We found out that the RNC pre-training greatly benefited the diachronic models. **Table 2** shows the intrinsic performance scores of our embedding models trained on yearly news corpora (for comparison, we also report the performance of the original RNC model). For evaluation, we used the SimLex965 semantic similarity test set [10] and the Russian translation of the Google Analogies test set [15]. Without the pre-training, the results on Google analogies were pretty much the same as in the table. However, the accuracies on SimLex965 (that is, ranking word pairs by their semantic similarity) dropped almost twice: the average score was about 0.16. The reason is that SimLex965 (unlike Google Analogies) does not contain proper names and toponyms. At the same time, proper names are heavily over-represented in news texts. This why the models trained solely on them perform good enough in the Google

Analogies evaluation, but fail with SimLex965. RNC pre-training allowed us to avoid this and to make the resulting models more 'aware' of general world knowledge.

Table 2: Intrinsic evaluation scores of the yearly models in comparison to a single model trained on the RNC; the right column shows the number of words not shared by **all** models

Model	SimLex965	Analogies	Unique words
RNC	0.35	0.14	
2010	0.33	0.12	18,001
2011	0.33	0.12	17,953
2012	0.33	0.12	17,258
2013	0.33	0.12	17,240
2014	0.33	0.11	11,334
2015	0.31	0.11	19,849
2016	0.31	0.11	18,996
2017	0.32	0.12	18,666
2018	0.32	0.11	14,005
2019	0.31	0.11	13,928

4.1. Model alignment

Our yearly models are trained independently (except the shared RNC initialisation), and thus must be aligned to make it possible to directly measure cosine similarity between word vectors produced by them [12]. For that, we used orthogonal Procrustes analysis [5], which became a tool of choice for diachronic semantic change modeling with word embeddings since [6]. We transformed all the yearly models' vector spaces to match the vector space of the **2017** model (its vocabulary was the largest of all, so we chose it as our 'baseline' model). All the methods described below in **Section 5** deal with these aligned models.

The novel part of our alignment procedure was that we did not remove any words from the models' vocabularies, even if they were not shared across the two models being aligned. Since singular value decomposition (SVD) in Procrustes analysis requires two matrices of the same shape, applying this method typically results in excluding all the words not present in the intersection of the models' vocabularies. However, with our 10 time bins, intersecting vocabularies of all models would result in excluding too many words (arguably, most time-specific and thus most interesting). Hence, we conduct alignment in three stages:

1. For the baseline 2017 model M_{2017} and the model to align M_i, the matrices S_{2017} and S_i are generated, containing vectors for the words occurring in both models' vocabularies. Their dimensionalities are identical.
2. The standard Procrustes alignment with SVD is applied to S_{2017} and S_i, resulting in the optimal orthogonal projection matrix O.

3. The original embedding matrix E_i from the model M_i is multiplied by O, thus projecting all M_i vectors into the vector space of M_{2017}. Note that the dimensionality of E_i can be (and most often is) different from the dimensionality of the original 2017 matrix S_{2017}, because of the differences in their vocabularies. However, all the words from M_i are projected into the M_{2017} vector space, including those missing from the M_{2017} vocabulary.

This modification allowed us to keep the unique vocabularies for yearly models. The right column of **Table 2** lists the number of unique words for each models: that is, the words from the model vocabulary missing from the intersection of all vocabularies. This gives the intuition of how many (potentially crucial) words would be lost if we stuck to the classical Procrustes alignment workflow.

5. Features of *ShiftRy* as a web service

Primary way of user interaction with *ShiftRy* is by entering word queries. The system will generate a visualization of the semantic drift trajectory for the query word across all the featured years (currently from 2010 to 2019) or across the user-specified time span. It will also apply a pre-trained classifier (described in **subsection 5.1**) to make educated guesses about whether the query word experienced sharp semantic shifts between any two consequent years. Users can use this classifier for a particular pair of years in the 'Shift classifier' tab.

To support deeper analysis of word usage change, *ShiftRy* also provides lists of the nearest neighbors of the query word for the query years. The 'Synchronic visualizations' tab allows to generate synchronic 2-dimensional PCA or t-SNE projections of any set of word vectors for any particular year.

Finally, the 'Shift lists' tab takes a pair of years as an input, and produces a ranked list of words, the usage of which changed most when comparing these two years. Here, the usage change is defined as a simple cosine distance between word vectors in each of the two (already aligned) models. In this way, a user can get insights about words most likely to have undergone a usage shift in a given period of time.

5.1. Shift classifier

The shift classifier predicts whether or not a query word has experienced a semantic shift within a given pair of years. It is a logistic regression trained on the 'Micro' dataset described in [4].

The 'Micro' dataset contains 280 Russian adjectives, together with year pairs, for instance: ['экстремистский' 'extremist (adjective)', 2007 → 2008]. The dataset covers the years from 2000 to 2014. Three independent informants annotated each word with one of the three labels: the meaning of the word did not change at all (0), somewhat changed (1), or experienced a significant change (2). The gold truth consists of their averaged scores.

Since the ternary granularity is not needed for *ShiftRy*, we binarized the target labels by mapping 0 → 0, 1 → 0 and 2 → 1. This was done to prioritize precision and minimize false positives: we want the classifier to capture only significant shifts. The classifier uses four features, each of which is a similarity score between embeddings of the query word in two yearly models:

1. Cosine similarity between the query word vectors.
2. Similarity score produced by the Global Anchors algorithm [23]. It first creates a shared vocabulary V by intersecting the vocabularies of 2 models under analysis. Then, for each model, a vector of cosine similarities between the embedding of the query word and the embeddings of all words in V is produced. The resulting semantic change score is the cosine similarity between these 2 vectors (also known as second-order similarity).
3. Jaccard similarity [8] between the sets of 50 nearest neighbors of the query word in in each embedding model.
4. Kendall's τ [9] on the intersection of the ranked lists of 50 nearest neighbors; it tries to determine whether there are any significant changes in the neighbours ranking.

We evaluated the classifier on the 'Micro' training set using cross-validation with 5 folds. **Table 3** shows the results comparing classifiers employing all four features and cosine similarity only. Note that these are scores on the 'shifted' class, which is the most important for us. As one can see, the 4-feature classifier consistently outperforms the single-feature one. Thus, we use the 4-feature classifier. Note that the 'Micro' dataset contains only adjectives, while *ShiftRy* works with all parts of speech: for this reason, the real reliability of the *ShiftRy* predictions is arguably somewhat lower. However, until another manually annotated test set of short-term Russian semantic shifts appears, we believe the results in **Table 3** to be a good approximation of the *ShiftRy* performance.

Table 3: Semantic shift classifier scores on the 'shifted' class from the 'Micro' dataset

Features	F^1	Precision	Recall
All features	**27.8%**	**18.0%**	**80.0%**
Cosine similarity only	20.1%	13.3%	50.2%

5.2. Word trajectory visualization

Projection of high-dimensional embedding data into a low-dimensional space using dimensionality reduction techniques is a well-established method of visualizing word embeddings. It can be used for diachronic change analysis as well.

ShiftRy employs T-distributed Stochastic Neighbor Embedding (t-SNE) [14], a nonlinear dimensionality reduction technique, to plot semantic trajectory of a specific word via several embedding models trained on different time bins. These visualizations make it easier to explore the usage drift of a query word and to trace how it changed its typical neighbours through time. Our visualization procedure is similar to the one used in [6] and [17].

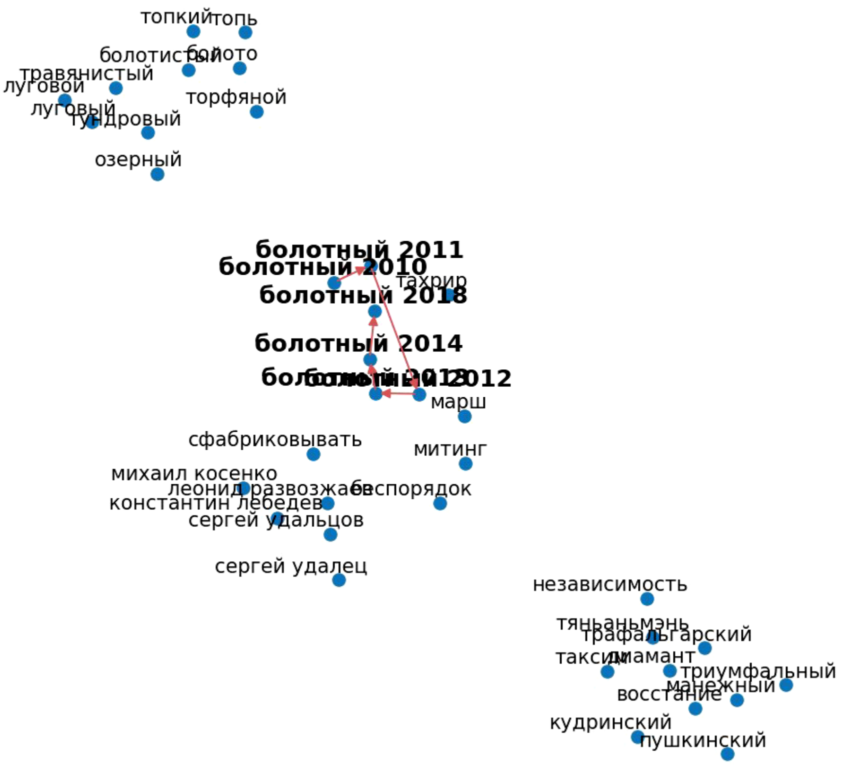


Figure 2: Alterations in the meaning of the word ‘болотный’ ’swamp’: from ‘торфяной’ ’peat’ or ‘топкий’ ’marshy’ in 2010 to ‘марш’ ’political demonstration’ in 2012 and to ‘митинг’ ’rally’, as well as the names of the prosecuted in 2014. In 2018, the word meaning is slowly shifting back to its original sense.

Given a query word *w* and a list of query years *L*, we generate *N*, which is a union of 6 nearest neighbors of *w* in every model in *L*. The corresponding embeddings of *w* are inferred from each model. For the words from *N*, we try to infer their embeddings from the most recent model in *L*. In the case the most recent model misses this word from the vocabulary, we try models back in time one by one, until we find a model which has this word in its vocabulary. Then, the embedding is taken from this model. Thus, the procedure provides the most recent embeddings of the nearest neighbours, while preserving time-specific embeddings of the query word.

After that, the standard t-SNE 2-dimensional projection is generated for both the query word time-specific vectors (labeled with the corresponding year) and the neighbours’ vectors. On the resulting plot, the semantic change trajectory of the query word is shown with red arrows which sequentially connect the considered years from the past to the most recent one.

Figure 2 shows the trajectory of the word ‘болотный’ ’swamp’ shifting from its primary sense ‘торфяной’ ’peat’ or ‘топкий’ ’marshy’ to ‘марш’ ’political demonstration’, caused by the media coverage of large protest rallies at the Bolotnaya (literally ‘Swampy’) Square in Moscow in 2012. These rallies resulted in mass arrests of its participants and organizers, with the ‘Bolotnaya case’ being in court trials throughout 2013 and 2014. This is why the 2014 point on the plot is close to the cluster of the names of the arrested and to the words ‘митинг’ ’rally’ and ‘сфабриковывать’ ’to fabricate the case’. Note also a separate cluster with the names of other famous squares where protest events took place both in Russia and throughout the world (e.g. ‘манежный’ ‘Manezhnyy’ and ‘трафальгарский’ ‘Trafalgar’), together with the word ‘независимость’ ’independence’. In 2018, the meaning of the target word is slowly shifting back. Notably, our shift classifier does not find semantic shifts for this word after 2013 (for any consequent pair of years). A possible reason for this is that while the shift was rapid and strong in 2012 (due to political events), the ‘return’ of the meaning back to the original state after 2014 was slow and gradual.

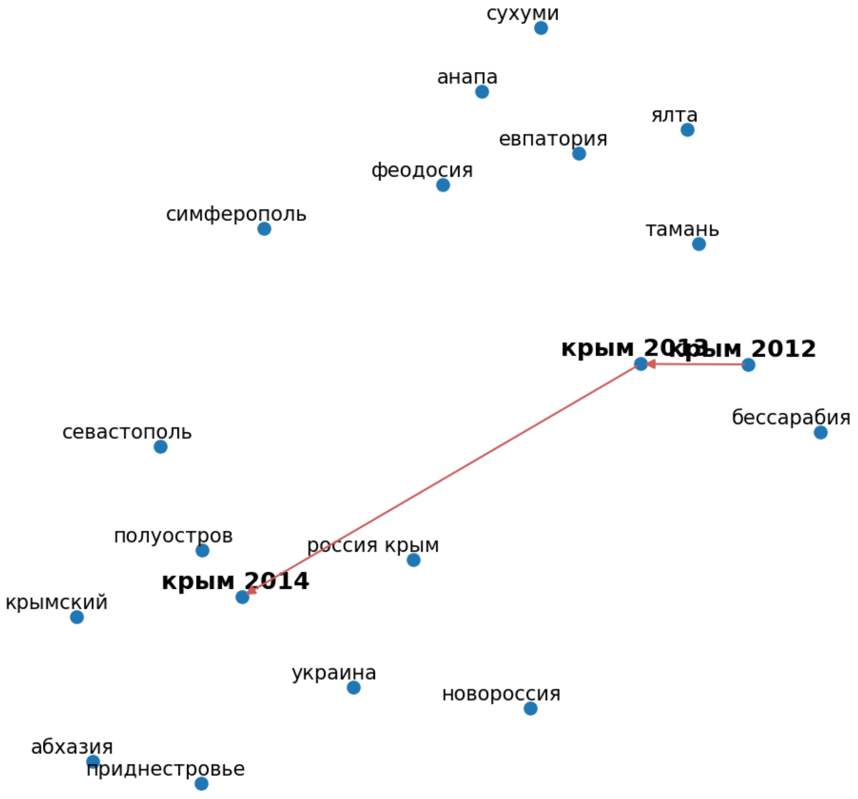


Figure 3: Alterations in the meaning of the word ‘Крым’ ’the Crimea’: from ‘Ялта’ ’Yalta’ and ‘Бессарабия’ ’Bessarabia’ in 2012 to the ‘Россия Крым’ ’Russia Crimea’ bigram in 2014

Another example is shown on Figure 3. Here, the semantic trajectory of the word 'Крым' 'the Crimea' moves from the cluster of Black Sea resort town names in 2012 and 2013 towards the words 'Россия' 'Russia', 'Украина' 'Ukraine' and 'Новороссия' 'Novorossiya' in 2014. This is, of course, caused by the Russian annexation of the Crimea, and the resulting territory conflict with the Ukraine.

Currently, the plots are static and non-interactive, but we plan to remedy that by making terms on them clickable in order to get lexical information and/or to explore the diachronic trajectories of neighbours. Another possible issue is that t-SNE is inherently stochastic, and the 2-dimensional projections can linearly change (rotate or scale) with the selected years, which in some cases can make it harder to explore the word's history. This is, however, an engineering problem, which we plan to solve in further release versions.

5.3. Sampling examples from corpora

Given a query word *w* and a pair of years *x* and *y*, *ShiftRy* can produce lists of sentences containing *w* in both corresponding corpora, while making sure to sample the sentences so that *x* sentences are as far away semantically from *y* sentences as possible. This allows to manually check whether a usage shift is actually taking place or not.

To provide the user with the most representative examples of the word's usage we first sample all the sentences that contain the query word from the text corpora of two chosen years. Then, we calculate the vectors of these sentences by averaging the embeddings of their words (recall that all the models are already aligned). Two sentences from different years with the highest cosine distance between their sentence vector are chosen as the first context pair. For each of them, we find 4 most similar sentences from the same corpora. This allows to get examples of the most specific contexts for each year, especially in the cases when the query word really shifted its meaning.

Below, we give an example with the word 'пожар' 'fire' and the years 2018 and 2019:

Old contexts (2018):

- В ночь перед пожаром мужчина со своей знакомой гостил у 70-летней погибшей. (The night before the fire, a man and his friend were visiting a 70-year-old friend.)
- Родители и четырехлетняя сестра погибли в пожаре. (Parents and a four-year-old sister died in a fire.)
- Сотрудница центра «Медицина катастроф» из Улан-Удэ Светлана Пежемская спасла из пожара 12 человек — родственников и соседей. (Svetlana Pezhemskaya, an employee of the Center for Medicine of Catastrophes from Ulan-Ude, rescued 12 people—relatives and neighbors—from the fire.)
- Незадолго до пожара мужчина грозился спалить весь дом и бросался в прохожих разными предметами. (Shortly before the fire, the man threatened to burn down the whole house and threw various objects at passers-by.)

- У кемеровчанина Игоря Вострикова в пожаре погибли трое детей, жена и сестра. (Kemerovo resident Igor Vostrikov lost three children in a fire, his wife and sister.)

New contexts (2019):

- В ведомстве полагают, что переход на новую модель организации позволит улучшить систему управления, оптимизировать излишние управленческие и обеспечивающие структуры подразделений, повысить эффективность реагирования на чрезвычайные ситуации и пожары. (The agency believes that the transition to a new model of organization will improve the management system, optimize redundant management and support structures of units, improve the efficiency of emergency response and fires.)
- Военная авиация привлечена для ликвидации крупного пожара близ города Новотроицк Оренбургской области, угрожающего населенным пунктам. (Military aviation has been involved in the elimination of a major fire near the town of Novotroitsk in the Orenburg region, which threatens populated areas.)
- Основными причинами пожаров стали нехватка финансирование, отсутствие техники для тушения, неукомплектованность подразделений наземной и авиационной охраны в регионах и низкое качество противопожарных мероприятий, которые проводят арендаторы. (The main causes of the fires were lack of funding, lack of equipment for firefighting, understaffing of ground and aviation security units in the regions and poor quality of firefighting activities carried out by tenants.)
- Секретарь генсовета «Единой России» Андрей Турчак предлагает пересмотреть подходы к борьбе с лесными пожарами, заявил он РБК: «Существующая система борьбы с лесными пожарами менялась, исходя из необходимости оптимизировать расходы на лесоохрану». (Andrei Turchak, secretary of the United Russia general council, suggests revising approaches to forest firefighting, he said: “The existing system of forest firefighting has been changing, based on the need to optimize the cost of forest protection”.)
- Медведев также поручил подготовить предложения и представить в правительство, чтобы усилить группировку по борьбе с пожарами и помочь тем регионам, которые находятся в наиболее сложной ситуации: «Субъекты не справляются с тушением лесных пожаров, значит, необходимо усилить роль федерального центра в системе мониторинга и тушения пожаров», — пояснил Турчак. (Medvedev also instructed to prepare proposals and submit to the government to strengthen the group of firefighters and help those regions that are in the most difficult situation: “The subjects are not able to cope with extinguishing forest fires, so it is necessary to strengthen the role of the federal center in the system of monitoring and extinguishing fires,”—said Turchak.)

Examples from 2018 describe fires in buildings and generally within cities. At the same time, examples from 2019 are mostly related to the unprecedented large-scale Siberian forest fires which were widely covered in Russian media in 2019. Thus, they reflect a significant shift in this word usage.

6. Conclusion

In this paper, we described *ShiftRy*, which is a web service to explore the dynamics of word usage in Russian news texts from 2010 to 2019. It employs several existing methods of semantic change modeling with diachronic word embeddings to generate visualizations of temporal semantic trajectories for any user-entered word, and to provide guesses on whether this word experienced a sharp semantic shift within any pair of years. For deeper exploration of semantic shifts, *ShiftRy* also features the possibility to intellectually sample maximally different corpus examples for the query word from two different time periods. The aim of *ShiftRy* is to ease the task of studying word history on short-term time spans, and the influence of social and political events on word usage. Thus, it can be used for studies in digital humanities, as well as in historical linguistics.

Under the hood, the system runs 10 word embedding models trained on the corresponding yearly sub-corpora and aligned using a novel modification of the Procrustes analysis, which preserves time-specific vocabularies. The underlying diachronic word embeddings and the source code of the web service are available online (https://shiftry.rusvectores.org), making it easy to apply the same approach to other sources, for example, social media. For this, one would have only to crawl the relevant textual data and train the corresponding diachronic word embeddings.

We plan to maintain *ShiftRy* and update it with new data yearly, possibly also extending the covered time span backwards in the XX century. Another point for future work is to improve the semantic trajectory visualizations by making them more interactive and more robust with regards to t-SNE initialization randomness, and by trying other dimensionality reduction techniques such as UMAP [1]. Finally, we plan to apply contextualized embedding architectures both for semantic shift detection and for sampling of corpus examples with sentences containing one and the same word in different senses.

References

1. *Becht, E. et al.:* Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology. 37, 1, 38 (2019).
2. *Daniel, M., Dobrushina, N.:* Two centuries in twenty words (in Russian). NRU HSE (2016).
3. *Field, A. et al.:* Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 3570–3580 Association for Computational Linguistics, Brussels, Belgium (2018).

4. *Fomin, V. et al.:* Tracing cultural diachronic semantic shifts in Russian using word embeddings: Test sets and baselines. Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. 203–218 (2019).
5. *Gower, J. C. et al.:* Procrustes problems. Oxford University Press on Demand (2004).
6. *Hamilton, W. et al.:* 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). pp. 1489–1501 Association for Computational Linguistics, Berlin, Germany (2016).
7. *Hellrich, J. et al.:* JeSemE: Interleaving semantics and emotions in a web service for the exploration of language change phenomena. In: Proceedings of the 27th international conference on computational linguistics: System demonstrations. pp. 10–14 Association for Computational Linguistics, Santa Fe, New Mexico (2018).
8. *Jaccard, P.:* Distribution de la flore alpine: Dans le bassin des dranses et dans quelques régions voisines. Rouge (1901).
9. *Kendall, M. G.:* Rank correlation methods. Griffin (1948).
10. *Kutuzov, A., Kunilovskaya, M.:* Size vs. Structure in training corpora for word embedding models: Araneum Russicum Maximum and Russian National Corpus. In: Aalst, W. M. van der et al. (eds.) Analysis of images, social networks and texts. pp. 47–58 Springer International Publishing, Cham (2017).
11. *Kutuzov, A., Kuzmenko, E.:* WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).
12. *Kutuzov, A. et al.:* Diachronic word embeddings and semantic shifts: A survey. In: Proceedings of the 27th international conference on computational linguistics. pp. 1384–1397 Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018).
13. *Loukachevitch, N., Lashevich, G.:* Comparing two thesaurus representations for Russian. In: Proceedings of global wordnet conference gwc. pp. 35–44 (2018).
14. *Maaten, L. van der, Hinton, G.:* Visualizing data using t-SNE. Journal of Machine Learning Research. 9, 2579–2605 (2008).
15. *Mikolov, T. et al.:* Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. (2013).
16. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 26, 3111–3119 (2013).
17. *Rodina, J. et al.:* Measuring diachronic evolution of evaluative adjectives with word embeddings: The case for English, Norwegian, and Russian. In: Proceedings of the 1st international workshop on computational approaches to historical language change. pp. 202–209 Association for Computational Linguistics, Florence, Italy (2019).

18. *Shoemark, P. et al.:* Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). pp. 66–76 Association for Computational Linguistics, Hong Kong, China (2019).
19. *Straka, M., Straková, J.:* 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. pp. 88–99 Association for Computational Linguistics, Vancouver, Canada (2017).
20. *Tahmasebi, N. et al.:* Proceedings of the 1st international workshop on computational approaches to historical language change. Association for Computational Linguistics, Florence, Italy (2019).
21. *Tang, X.:* A state-of-the-art of semantic change computation. Natural Language Engineering. 24, 5, 649–676 (2018).
22. *University, P.:* About WordNet, (2010).
23. *Yin, Z. et al.:* The global anchor method for quantifying linguistic shifts and domain adaptation. In: Advances in neural information processing systems. pp. 9433–9444 (2018).