SBX-HY at RuShiftEval 2021: Доверяй, но проверяй

Simon Hengchen,[♠] Kate Viloria,[♠] Andrey Indukaev[♡]
[♠]University of Gothenburg, Sweden, simon.hengchen@gu.se; kateviloria@outlook.com
[♡]University of Helsinki, Finland, andrey.indukaev@helsinki.fi

Abstract

Research in computational lexical semantic change, due to the inherent nature of language change, has been notoriously difficult to evaluate. This led to the creation of many new exciting models that cannot be easily compared. In this system paper, we describe our submissions at RuShiftEval 2021 – one of the few recently shared tasks that enable researchers, through a standard evaluation set and control conditions, to systematically compare models and gain insights from previous work. We show that despite top results in similar tasks on other languages, Temporal Referencing does not seem to perform as well on Russian.

Keywords: semantic change, Russian, diachronic word embeddings
DOI: 10.28995/2075-7182-2021-20-1055-1062

SBX-HY at RuShiftEval 2021: Доверяй, но проверяй

Симон Хенгчен,[♠] Кейт Вилориа,[♠] Андрей Индукаев[♡]
[♠]Гётеборгский университет, Швеция, simon.hengchen@gu.se; kateviloria@outlook.com
[♡]Хельсинкский университет, Финляндия, andrey.indukaev@helsinki.fi

Аннотация

Исследования семантических сдвигов методами компьютерной лингвистики трудно оценивать, поскольку языку свойственно меняться на всех уровнях. Как следствие, создаются множество интересных моделей, но их сравнение предстает нетривиальной задачей. Эта статья описывает систему представленную нами на RuShiftEval 2021, одну из немногих недавних дорожек (shared task), которые позволяют исследователям, благодаря унифицированной системе оценки и единым данным ”золотого стандарта систематически сравнивать модели семантических сдвигов и расширять понимание разработанных методов. Мы показываем, что, несмотря на высокие результаты, полученные при решении аналогичных задачах на других языках, метод Temporal Referencing продемонстрировал низкую эффективность на русском языке.

Ключевые слова: семантический сдвиг, русский язык, диахронические векторные семантические модели

1 Introduction

The computational study of lexical semantic change (LSC) presents many challenges [4]. Evaluation is one of many hurdles this flourishing field faces (for overviews, we refer to [31, 6, 33]). Until recently, the majority of prior works evaluated their models on limited, ad-hoc test sets specifically created for the purpose of evaluating that very model—making thorough comparison between models nigh-impossible.[1]

The evaluation hurdle was somewhat alleviated in the last decade with works such as [5, 29, 26, 34, 25] who applied different methods for introducing *synthetic* change. This allows for a larger-scale evaluation of models on the data they are trained on (or fine-tuned with), thereby bypassing the need to rely on historical dictionaries compiled with different methodologies and on different data [4]. In 2020, the SemEval-2020 Task 1 on Unsupervised Detection of Lexical Semantic Change [28] laid the groundwork for the first comparative study of computational models of lexical semantic change in English, Swedish, German, and Latin, based on the DURel framework [27]. Follow-up tasks followed with DIACR-ita [23] on Italian, and this shared task, RuShiftEval, on Russian [21]. In this paper, we describe the hyperparameter search and submissions of our team, SBX-HY, at RuShiftEval 2021.

[1]This observation is particularly obvious in Table 3 of [31].

2 Task description

The task requires teams to rank a list of 99 target words by their degree of predicted semantic change. Every team can submit up to 10 predictions. Following previous tasks, the predictions are evaluated against the ground truth using the Spearman correlation ρ, and each team's best prediction is kept for the final ranking.

As mentioned in the introduction, this task follows the same principles as the two tasks that precede it, allowing for the comparison of models across languages. This task also has the advantage of using annotations created using the theoretically-motivated DURel annotation framework [27] which was the basis of the SemEval-2020 Task 1 on Unsupervised Detection of Lexical Semantic Change [28]—reinforcing comparison potential.

Compared to the previous tasks, RuShiftEval presents three advantages. First, instead of asking participants to estimate semantic change between two periods, RuShiftEval benefits from ground truth annotations for three periods: pre-Soviet, Soviet, post-Soviet. This is beneficial in two ways: it triples the number of predictions, as we move from "before $\rightarrow$ after" to "pre-Soviet $\rightarrow$ Soviet" AND "Soviet $\rightarrow$ post-Soviet" AND "pre-Soviet $\rightarrow$ post-Soviet", but also presents a perhaps more realistic view of semantic change. Theoretically, such a task set-up should also benefit "dynamic" models that share data across time bins, as alignment procedures have been shown to be extremely noisy [8, 34].

Second, RuShiftEval is based on the RNC corpus,[2] which was also the basis of RuSemShift [24], a large-scale manually-annotated test set for LSC in Russian. While RuSemShift only provides annotation scores for two distinct time periods ("pre-Soviet $\rightarrow$ Soviet" and "Soviet $\rightarrow$ post-Soviet") and not for the longer-term change ("pre-Soviet $\rightarrow$ post-Soviet"), this task is the first to provide training data for the task of LSC detection, at least for two periods.

Finally, RuShiftEval presents time slices that are based on historical periods reflecting deep changes in Russian society.[3] Previous work in (lexical) semantics, such as [1], do point out the relevance of real-world changes on semantic change: the changing reality of our conception of the world sometimes requires the shift of existing concepts. RuShiftEval, by splitting the corpus in three subsets corresponding to strong schisms in Russian society ("pre-Soviet", "Soviet", "post-Soviet"), provide, just like previous work (eg [16], who split their corpus according to socialist milestones), a more ecological view of semantic change.[4]

3 System overview

As described in previous work [28, 18], we consider an LSC system as a combination of:

- a semantic representation,
- a temporal alignment procedure,
- a change measure.

Semantic representations are, quite straightforwardly, representations of the word(s) at hand: vectors, clusters, etc. Temporal alignment procedures serve, if needed, to make representation comparable across time slices. Finally, the change measure allows to compare representations extracted from different time slices.

Until very recently, most work computationally tackling lexical semantic change made use of type embeddings.[5] Recent work employed token embeddings from large-scale language models, either pre-trained or trained from scratch on the data at hand [17, 22, 3, 20, 15, 11, to cite but a few]. Despite these models performing extremely well on a variety of NLP downstream tasks, they give relatively poor

[2] `https://ruscorpora.ru/new/corpora-intro.html`

[3] We do not make the assumption that only the people of Russia speak Russian, and that therefore the RNC is representative of all Russian languages. Nonetheless, the corpus is carefully crafted and, in the remainder of this paper, we use 'Russian' as 'the Russian language as described in the corpus.'

[4] The SemEval 2020 Task on ULSCD did make use of large shifts in society for one of its languages (Latin), by splitting the corpora into "pre-Christian" and "Christian" eras.

[5] We make the difference between "type" and "token" embeddings. Type embeddings are often referred to as "static" embeddings, a term that can lead to confusion in LSC as some models encode temporal information and are thus called "dynamic."

results in previous LSC shared tasks [2, 28, 23]. Following this observation, our team decided to employ type embeddings as they currently are better understood [9, 35].

More specifically, we use in this task Temporal Referencing [34, 'TR'], as it is shown to perform very well on the task on LSC: it is shown to be much less noisy than the then-state-of-the-art SGNS+OP[6] combination on both a synthetic and real-life task on English data by its authors. TR uses a re-labeling trick and manages to achieve the same goal as dynamic models (i.e.: avoiding alignment) while using static embeddings. As best put by Tahmasebi et al [32], TR consists in "training embeddings on a corpus as a whole, while relabeling target words during training with their time information, following work such as [12], [13], and [35]. A word w in a sentence c_1, c_2, w, c_3, c_4 from time t would be relabeled as c_1, c_2, w_t, c_3, c_4 only when w is a target word. This results in individual time-dependent embeddings for each target word but avoids alignment since they are all situated in the same space. The context embeddings are average embeddings across the whole corpus and thus suffer from bias towards time periods with more data." In the case of the RuShiftEval task, with similar data sizes between time bins, the data discrepancy often seen in diachronic corpora is not problematic.

The two main advantages of TR compared to, among others, SGNS+OP, is first its implicit alignment: given the fact that the entirety of the data is used for training but only target words are assigned a temporal label, there is no need for noise-creating alignment between time periods. Second, given the fact that TR uses the entirety of the data, there simply is 'more data' to build vectors, which often results in better vectors. At SemEval 2020 Task 1, team Jiaxin & Jinan [37] implement TR and rank 3rd and 2nd on subtasks 1 and 2 respectively, team DCC [36] rank 6th on the first subtask, and team Random [14] show, in the post-evaluation phase, excellent results with TR. Finally, given the fact that TR has been successfully used on Russian data in another context [19], namely a text-based sociological study of modernisation politics in Russia, this particular model seemed like an excellent choice.

As with most previous work, we use the cosine distance between two vectors as a measure for semantic change. Embeddings were trained on a lemmatised version of the corpus. We deviated from the task organisers in the choice of tool for the lemmatisation, which in our case was carried out through pymystem3[7] (a Python wrapper for Yandex' MyStem[8]), while the organisers' version of the corpus on which baselines models were trained was processed with UDPipe [30]. In the next section, we describe the choice of hyperparameters.

4 LSC as a supervised task: hyperparameter tuning

As mentioned in Section 2, RuShiftEval 2021 is the first task on LSC to provide high-quality training data. As a result, hyperparameter tuning – the problem of choosing the best possible hyperparameters for a machine learning algorithm – becomes possible. Using the available gold labels from RuSemShift [24], we train a variety of TR models and follow [18] in paying particular attention to dimensionality. The hyperparameters considered are the following:

- frequency threshold $\in [50, 70]$,
- window size $\in [2, 5, 7]$,
- vector dimensionality $\in [50, 75, 100, 200, 300]$,
- iterations $\in [5, 10, 25]$.

We train models for all combinations of these hyperparameters, as well as, given interesting preliminary results with low-dimensional vectors and large window sizes, two additional models:

- frequency threshold 50, window size 11, dimensionality 22, iterations 5
- frequency threshold 50, window size 11, dimensionality 50, iterations 5.

5 Results

We finished last at the end of the competition: unfortunately, we misread the instructions and ranked our prediction in the opposite order of what was required, leading to a negative correlation of -0.369. After

[6]Skipgram with negative sampling [7, 10] as a model, and Orthogonal Procrustes as an alignment method.
[7]`https://github.com/nlpub/pymystem3`
[8]`https://yandex.ru/dev/mystem/`

recalculation by the organisers, our team was officially ranked 11^{th}.

Further analysing our models in the post-evaluation phase, we see that our approach, despite benefiting from a hyperparameter search based on RuSemShift, yields averaged results on three subtasks ranging from 0.028 to 0.379, and beating the organiser's baseline with some difficulty. We present our fifteen best results, during the post-evaluation phase, in Table 1.[9] Obviously, evaluating the advantage of a hyperparameter search is complicated, as the aim of such an entreprise is to find the ideal combination of different hyperparameters – there is thus no "ground truth" to compare against. We thus look at models trained with oft-used hyperparameters, as well as the overall performance of all our models. Models trained with the usual default settings for SGNS for English (window size of 5, dimensionality of 300, 5 iterations) score 0.236 and 0.23 respectively for the frequency thresholds of 50 and 70 – well below the top scores, while Figure 1 shows a relatively wide variation of scores, with roughly 50% of all models being outperformed by at least 0.086 point. Both observations indicate that a hyperparameter search on training data is useful for this sort of task.

Table 1: Best 15 hyperparameter configurations during the post-evaluation phase. Average score is Spearman's ρ correlation with the ground truth for all three subtasks.

Freq. threshold	Window size	Dimensionality	Iterations	ρ
70	5	50	5	0.379
70	5	75	5	0.374
70	7	50	5	0.369
50	11	22	5	0.368
50	7	75	5	0.365
70	7	75	5	0.365
50	5	50	5	0.364
50	7	50	5	0.364
70	5	50	25	0.364
70	7	50	10	0.363
50	7	50	25	0.362
70	7	50	25	0.360
70	7	100	5	0.356
50	7	100	5	0.355
70	7	75	10	0.351

6 Conclusions

Our shared task system investigated the usefulness of a simple hyperparameter search on Russian, in a controlled setting. Despite the assumption based on several previous works that the particular method we utilise, Temporal Referencing, should produce good performance, our submissions – both during the evaluation and the post-evaluation phases – seem to indicate otherwise. Indeed, our system beats the baseline with some difficulty, and, even with post-evaluation scores, is vastly outperformed by other teams. Nonetheless, it does not mean that the method is unfit for such a task: it yields comparable absolute performance on Russian as it did on English at the SemEval 2020 Task 1 as per team Jiaxin & Jinan [37], where TR performed well and other teams – including variations of the high-performance system in this task – performed poorly.

While pure conjecture at this point,[10] it seems that this task setup might have been particularly difficult for type embeddings systems: indeed, other teams with similar setups (TR, SGNS) do not perform very well either. Whether the presence of tuning data better benefits token embeddings rather than type embeddings should also be investigated. We leave for future work a thorough analysis of the annotated data, of the semantic shift measure used in the dataset ("COMPARE" vs graph clustering), as well as an investigation of the effect of the lemmatiser on the quality of the embeddings.

[9] The predictions based on the models making up the top two results were not submitted to the competition website, as unfortunately the models did not finish training on time.

[10] This will need to be investigated in future work, once all system description papers are available.

	score
count	91.0
mean	0.2711318681318681
std	0.08079909361306925
min	0.028
25%	0.2045
50%	0.293
75%	0.338
max	0.379

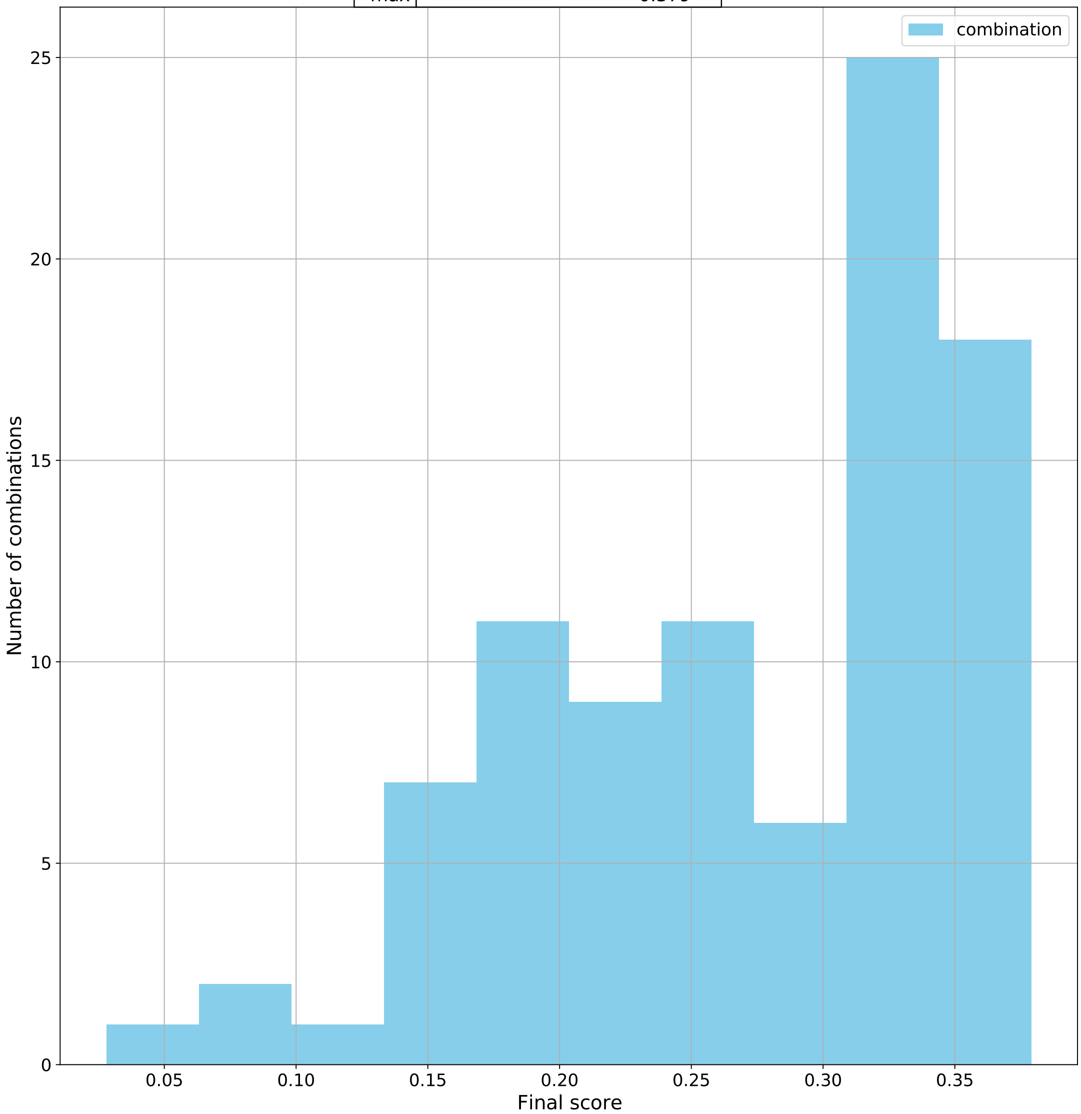

Figure 1: Distribution of scores

Acknowledgments

The authors would like to thank Pascal Hein, Jens Kaiser, Dominik Schlechtweg and Nina Tahmasebi for their insights, and the reviewers for their helpful comments. Additional thanks go to the task organisers. This work has been funded in part by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184), and *Nationella Språkbanken* (the Swedish National Language Bank), jointly funded by the Swedish Research Council (2018–2024; dnr 2017-00626) and its ten partner institutions.

References

[1] Blank Andreas. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change // Historical semantics and cognition / Ed. by Andreas Blank, Peter Koch. — Mouton de Gruyter, 1999. — Cognitive Linguistics Research no. 13. — P. 61–90.

[2] CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection / Severin Laicher, Gioia Baldissin, Enrique Castañeda et al. // arXiv preprint arXiv:2011.07247. — 2020.

[3] Capturing Evolution in Word Usage: Just Add More Clusters? / Matej Martinc, Syrielle Montariol, Elaine Zosa, Lidia Pivovarova // Companion of The 2020 Web Conference / Ed. by Amal El Fallah Seghrouchni, Gita Sukthankar, Tie-Yan Liu, Maarten van Steen. — Taipei : ACM / IW3C2, 2020. — P. 343–349. — Access mode: https://doi.org/10.1145/3366424.3382186.

[4] Challenges for Computational Lexical Semantic Change / Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, Haim Dubossarsky // Computational Approaches to Semantic Change / Ed. by Nina Tahmasebi, Lars Borin, Adam Jatowt et al. — Berlin : Language Science Press, 2021. — Language Variation.

[5] Cook Paul, Stevenson Susanne. Automatically Identifying Changes in the Semantic Orientation of Words // Proceedings of LREC 2010 / Ed. by N. C. C. Chair, K. Choukri, B. Maegaard et al. — Valletta : ELRA, 2010.

[6] Diachronic word embeddings and semantic shifts: A survey / Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, Erik Velldal // Proceedings of COLING 2018. — Santa Fe : ACL, 2018. — P. 1384–1397.

[7] Distributed Representations of Words and Phrases and their Compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in Neural Information Processing Systems / Ed. by C. J. C. Burges, L. Bottou, M. Welling et al. — Red Hook, NY : Curran Associates, Inc., 2013. — P. 3111–3119. — Access mode: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[8] Dubossarsky Haim, Weinshall Daphna, Grossman Eitan. Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models // Proceedings of EMNLP 2017. — Copenhagen : ACL, 2017. — P. 1136–1145. — Access mode: https://www.aclweb.org/anthology/D17-1118.

[9] Effects of Pre-and Post-Processing on type-based Embeddings in Lexical Semantic Change Detection / Jens Kaiser, Sinan Kurtyigit, Serge Kotchourko, Dominik Schlechtweg // arXiv preprint arXiv:2101.09368. — 2021.

[10] Efficient Estimation of Word Representations in Vector Space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean // ICLR Workshop. — 2013. — Access mode: https://arxiv.org/abs/1301.3781.

[11] Explaining and Improving BERT Performance on Lexical Semantic Change Detection / Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg et al. // arXiv preprint arXiv:2103.07259. — 2021.

[12] Ferrari Alessio, Donati Beatrice, Gnesi Stefania. Detecting Domain-specific Ambiguities: An NLP Approach based on Wikipedia Crawling and Word Embeddings // Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference Workshops. — 2017. — P. 393–399.

[13] Fišer Darja, Ljubešić Nikola. Distributional modelling for semantic shift detection // International Journal of Lexicography. — 2018. — Vol. 2018/advance articles, no. ecy011. — P. 1–21.

[14] GM-CTSC at SemEval-2020 Task 1: Gaussian Mixtures Cross Temporal Similarity Clustering / Pierluigi Cassotti, Annalina Caputo, Marco Polignano, Pierpaolo Basile // Proceedings of the 14th International Workshop on Semantic Evaluation. — Barcelona : ACL, 2020.

[15] Giulianelli Mario, Del Tredici Marco, Fernández Raquel. Analysing Lexical Semantic Change with Contextualised Word Representations // Proceedings of ACL 2020. — Online : ACL, 2020. — P. 3960–3973. — Access mode: `https://www.aclweb.org/anthology/2020.acl-main.365`.

[16] Hengchen Simon. When Does it Mean? Detecting Semantic Change in Historical Texts : Ph. D. thesis / Simon Hengchen ; Université libre de Bruxelles. — Brussels, 2017.

[17] Hu Renfen, Li Shen, Liang Shichen. Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View // Proceedings of ACL 2019. — Florence : ACL, 2019. — P. 3899–3908. — Access mode: `https://www.aclweb.org/anthology/P19-1379`.

[18] IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection / Jens Kaiser, Dominik Schlechtweg, Sean Papay, Sabine Schulte im Walde // Proceedings of the 14th International Workshop on Semantic Evaluation. — Barcelona : ACL, 2020.

[19] Indukaev Andrey. Studying Ideational Change in Russian Politics with Topic Models and Word Embeddings // Palgrave Handbook of Digital Russia Studies / Ed. by Daria Gritsenko, Mariëlle Wijermars, Mikhail Kopotev. — Basingstoke : Palgrave Macmillan, 2021. — P. 443–465.

[20] Kutuzov Andrey. Distributional word embeddings in modeling diachronic semantic change : Ph. D. thesis / Andrey Kutuzov ; University of Oslo. — 2020.

[21] Kutuzov Andrey, Pivovarova Lidia. RuShiftEval: a shared task on semantic shift detection for Russian // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2021.

[22] Martinc Matej, Kralj Novak Petra, Pollak Senja. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift // Proceedings of LREC 2020. — Marseille : ELRA, 2020. — P. 4811–4819. — Access mode: `https://www.aclweb.org/anthology/2020.lrec-1.592`.

[23] Overview of the EVALITA 2020 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) / Ed. by Valerio Basile, Danilo Croce, Maria Di Maro, Lucia C. Passaro. — Online : CEUR.org, 2020.

[24] 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`.

[25] Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings / Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen et al. // Proceedings of EMNLP-IJCNLP 2019. — Hong Kong : ACL, 2019. — P. 66–76.

[26] Rosenfeld Alex, Erk Katrin. Deep Neural Models of Semantic Shift // Proceedings of NAACL-HLT 2018: Volume 1 (Long Papers). — New Orleans : ACL, 2018. — P. 474–484. — Access mode: `https://www.aclweb.org/anthology/N18-1044`.

[27] Schlechtweg Dominik, Schulte im Walde Sabine, Eckmann Stefanie. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change // Proceedings of NAACL 2018. — ACL, 2018.

[28] SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection / Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen et al. // Proceedings of SemEval 2020. — Barcelona : ACL, 2020. — P. 1–23. — Access mode: `https://www.aclweb.org/anthology/2020.semeval-1.1`.

[29] Statistically Significant Detection of Linguistic Change / Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena // Proceedings of WWW 2015. — WWW '15. — Florence : ACM, 2015. — P. 625–635. — Access mode: `https://doi.org/10.1145/2736277.2741627`.

[30] Straka Milan, Straková Jana. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe // Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. — Vancouver, Canada : Association for Computational Linguistics,

2017. — August. — P. 88–99. — Access mode: http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

[31] Tahmasebi Nina, Borin Lars, Jatowt Adam. Survey of computational approaches to lexical semantic change // arXiv preprint arXiv:1811.06278. — 2018.

[32] Tahmasebi Nina, Borin Lars, Jatowt Adam. Survey of computational approaches to lexical semantic change detection // Computational Approaches to Semantic Change / Ed. by Nina Tahmasebi, Lars Borin, Adam Jatowt et al. — Berlin : Language Science Press, 2021. — Language Variation.

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

[34] Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change / Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg // Proceedings of ACL 2019. — Florence : ACL, 2019. — P. 457–470. — Access mode: https://www.aclweb.org/anthology/P19-1044.

[35] 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 // Proceedings of ACL 2019. — Florence : ACL, 2019. — P. 732–746.

[36] Zamora-Reina Frank D., Bravo-Marquez Felipe. DCC-Uchile at SemEval-2020 Task 1: Temporal Referencing Word Embeddings // Proceedings of the 14th International Workshop on Semantic Evaluation. — Barcelona : ACL, 2020.

[37] Zhou Jinan, Li Jiaxin. TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Reference // Proceedings of the 14th International Workshop on Semantic Evaluation. — Barcelona : ACL, 2020.