Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2025"

April 23–25, 2025

Interpretable approach to detecting semantic changes based on generated definitions

Tatarinov Maksim
HSE University
Nizhny Novgorod, Russia
tatarinovst0@gmail.com

Demidovsky Aleksandr
HSE University
Nizhny Novgorod, Russia
monadv@yandex.ru

Abstract

This paper investigates definition modeling as an approach to semantic change detection, which offers the advantage of providing human-readable explanations, unlike traditional embedding-based approaches that lack interpretability. Definition modeling leverages large language models to generate dictionary-like definitions based on target words and their contextual usages. Despite its potential, practical evaluations of this method remain scarce. In this study, FRED-T5 was fine-tuned using the Small Academic Dictionary for the task of definition modeling. Both quantitative and qualitative assessments of definition modeling's effectiveness in detecting semantic shifts within the Russian language were conducted. The approach achieved a Spearman's rank correlation coefficient of 0.815 on the Rushifteval task, demonstrating strong alignment with expert annotations and ranking among the leading solutions. For interpretability, a visualization algorithm was proposed that displays semantic changes over time. In the qualitative evaluation, our system successfully replicated manual linguistic analysis of 20 Russian words that had undergone semantic shifts. Analysis of the generated meanings and their temporal frequencies showed that this approach could be valuable for historical linguists and lexicographers.

Keywords: Semantic change, definition modeling, definition generation
DOI: 10.28995/2075-7182-2025-23-XX-XX

Интерпретируемый подход к детектированию семантических изменений слов на основе генерируемых определений

Татаринов Максим
НИУ ВШЭ
Нижний Новгород, Россия
tatarinovst0@gmail.com

Демидовский Александр
НИУ ВШЭ
Нижний Новгород, Россия
monadv@yandex.ru

Аннотация

В данной работе исследуется моделирование определений как подход к обнаружению семантических изменений, который имеет преимущество в виде понятных для человека объяснений, в отличие от традиционных подходов на основе векторных представлений, страдающих от недостатка интерпретируемости. Моделирование определений использует большую языковую модель для генерации словарных определений на основе целевых слов и их контекста. Несмотря на потенциал, практико-ориентированные оценки этого метода остаются ограниченными. В данном исследовании FRED-T5 была дообучена с помощью Малого академического словаря на задаче моделирования определений. Были проведены как количественные, так и качественные оценки эффективности моделирования определений в обнаружении семантических сдвигов в рамках русского языка. Подход достиг коэффициента ранговой корреляции Спирмена 0,815 в задаче Rushifteval, что демонстрирует сильное соответствие экспертным аннотациям, находясь среди лидирующих решений. Для интерпретируемости был предложен алгоритм визуализации, который отображает семантические изменения во времени. В качественной оценке наша система успешно воспроизвела ручной лингвистический анализ 20 русских слов, имевших семантическими сдвиги. Анализ сгенерированных значений и их временных частот показал, что этот подход может быть востребован для исторических лингвистов и лексикографов.

Ключевые слова: Семантические изменения, моделирование определений, генерация определений

1 Introduction

Static and contextual embeddings excel at capturing semantic relationships for detecting semantic change, but lack human-readable word descriptions. Advancements in recent research involve definition generation with language models, which offer more illustrative descriptions (Giulianelli et al., 2023; Fedorova et al., 2024). It could aid historical linguists and lexicographers in creating dictionaries and language history studies, such as Dobrushina and Daniel' (2018). However, the practical evaluation of this approach remains limited.

The primary objective of this study is to assess the effectiveness of language models in detecting semantic changes in words through the generation of definitions. It would use both quantitative metrics from a shared task and qualitative analysis by reproducing a linguistic analysis of words known to have undergone semantic shifts.

The paper is organized as follows: Section 2 reviews semantic change detection methods, evaluation methods for classifying errors in generated definitions and a strategy to acquire correct ones for comparison. Section 3 describes the proposed methodology. Section 4 presents the results and discusses their implications.

2 Related Work

2.1 Approaches to Semantic Change Detection

Semantic change is understood as change in the polysemy of a word over time. Although most solutions provide a quantitative measure of semantic change, such as a score or distance between vectors, to determine the extent of change, recently, a step towards a more explainable approach has been taken (Giulianelli et al., 2023; Fedorova et al., 2024).

There have been multiple approaches to semantic change detection:

Static Embeddings. Static embeddings provide a fixed representation of a word for the entire corpus. In the Shiftry (Kutuzov et al., 2020), Word2Vec (Mikolov et al., 2013) was utilized to examine semantic shifts by dividing the corpus by years to generate distinct word vectors for each period.

They need extensive data for stable representations, fail to differentiate multiple meanings of a word, and independently trained models produce incompatible vector spaces requiring alignment.

Contextual Embeddings. Contextual models such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) generate different embeddings for a word depending on its context. Rachinskiy and Arefyev (2021) fine-tuned the XLM-R model to generate embeddings aligned with dictionary definitions. Arefyev et al. (2021) trained XLM-R on a large multilingual dataset and RuSemShift data.

The GlossReader approach showed limitations with culturally specific words and depended on predefined senses for visualization, while the DeepMistake method lacked visualization capabilities.

Definition Modeling. Definition modeling takes a target word with a usage example to generate a human-readable word definition based on context, akin to a dictionary entry (Giulianelli et al., 2023), unlike previous embedding approaches which produce abstract vector representations that are difficult to interpret.

Table 1: Example of Definition Modeling

Example Usage	He started to sleep poorly at night, waking up with a persistent headache.
Target Word	night
Generated Definition	The part of the day from sunset to sunrise.

Giulianelli et al. (2023) proposed using generated definitions as semantic embeddings for words, enabling semantic change detection. Fedorova et al. (2024) researched definition modeling for the task of semantic change detection finding it successful.

The main limitation of the approaches employing embeddings is their non-interpretability. The best case is the DeepMistake, whose visualization is limited to predetermined senses.

As for definition modeling, qualitative evaluation in Fedorova et al. (2024) is limited, as they leave ”in-practice” evaluation for future research. Also, they used an unsupervised approach for an evaluation, while the proposed approach involved fine-tuning the vectorizer.

2.2 Classification of Errors in Generated Definitions

Studies by Huang et al. (2021) and Noraset et al. (2017) have proposed classifications for errors in generated definitions. Their work identified the following types:

Table 2: Types of Errors in Generated Definitions

Type	Russian Example	English Example (Translation)
Over-specification	кофе – горячий, горький напиток из жареных бразильских зерен	coffee – a hot, bitter beverage made from roasted Brazilian beans
Under-specification	капитан – член команды.	captain – team member
Self-referential	самосознание – состояние, при котором у человека присутствует самосознание	self-awareness – a state in which a person has self-awareness
Wrong Part of Speech	стекло – переместиться вниз, сбежать (о жидкости)	glass/spilt – to move down, escape (of a liquid)
Opposite Meaning	внутрь – ненаправленный в центр	inward – non-directed to the center
Close Semantics	машина – устройство с автоматическими функциями	machine – a device with automatic functions
Redundancy or Excessive Use of Generic Phrases	спутник – тот, кто совершает путь, путь вместе с кем-л.	companion – one who makes a journey, journey together with someone
Incorrectness	первый – следующий после всех остальных в списке предметов	first – next after all other items in the list
Correct	винодельня – заведение, помещение для изготовления вина	vineyard – establishment, premises for wine production

2.3 Acquiring correct definitions

Sternin and Rudakova (2017) outlines a method of generalizing dictionary definitions for determining correct semantic description of words, emphasizing the integration of diverse dictionary definitions to capture the full meaning. This procedure involves compiling all available definitions, differentiating meanings based on denotative principles, and synthesizing a unified semantic structure, with the final step organizing meanings from core to peripheral, accompanied by usage examples.

3 Proposed Approach

3.1 Fine-tuning LLM

A generative large language model M is trained on a dataset $D = \{(w_i, c_i, d_i)\}_{i=1}^{N}$, where each tuple contains a word w, its context c, and a corresponding definition d. The model learns to generate an accurate definition $\hat{d} = M(w, c)$ by minimizing the cross-entropy loss between its predicted token probabilities and the reference definitions:

$$L(M) = \sum_{i=1}^{N} \text{loss}(M(w_i, s_i), d_i), \tag{1}$$

3.2 Testing

Intrinsic evaluation is conducted using a test subset D_{test} of the dataset D to assess the quality of generated definitions $\hat{d}_j = M(w_j, c_j)$ compared to reference definitions d_j using string similarity metrics, defined as:

$$\text{metric} = \frac{1}{M}\sum_{j=1}^{P} \text{similarity}(\hat{d}_j, d_j) \quad (2)$$

where similarity measures the match between definitions, ranging from 0 (no similarity) to 1 (identical).

Extrinsic evaluation assesses the model's performance on a semantic change detection task with test set $S = \{(w_k, g_{k,(t_i,t_j)})\}_{k=1}^{Q}$, where w_k represents a target word, $g_{k,(t_i,t_j)}$ its gold semantic change score for the transition between periods t_i and t_j, and Q is the number of words in the test set.

For each word w_k in the test set, a set of usage contexts $U_{k,t} = \{u_{k,t,1}, u_{k,t,2}, \ldots, u_{k,t,n}\}$ is sampled from each time period $t \in \{t_1, t_2, t_3\}$ of the diachronic Russian National Corpus (Savchuk et al., 2024), where n is 100 or all if fewer available, in a similar way to Arefyev et al. (2021). For each period transition, the usages are paired, and definitions $\hat{d}_{k1}, \hat{d}_{k2}$ are generated by the model for each pair.

These definitions are then vectorized $\vec{d}_{k1}, \vec{d}_{k2}$ using a vectorizer V. The distance between the vectorized definitions $dist(\vec{d}_{k1}, \vec{d}_{k2})$ is calculated and converted to scores ranging from 1 (senses unrelated) to 4 (identical).

The mean values of the ratings for each word are compared with the gold scores from the task using Spearman's rank correlation.

3.3 Visualization

To illustrate semantic changes over time, generated definitions are transformed into vector representations using a vectorizer V.

A clustering algorithm C is then applied to group similar definitions.

For each cluster K_j, a prototypical definition $\hat{d}_{\text{proto}}$ is selected, which is defined as original definition whose vector $\vec{d}_{\text{proto}}$ is the closest to the center of the cluster (centroid).

Let $\vec{c_j}$ be the centroid of cluster K_j:

$$\vec{d}_{\text{proto},j} = \arg\min_{\vec{d} \in K_j} dist(\vec{d}, \vec{c_j}) \quad (3)$$

where $dist$ is a distance metric.

Bar charts are then created to display the frequency of different meanings over time.

3.4 Qualitative Analysis

A qualitative assessment begins with the selection of words known to have undergone semantic shifts based on existing linguistic research. Usage examples for these words are obtained from different time periods using a diachronic corpus. The trained model is applied to generate definitions for each word usage. The obtained definitions are compared with information from semantic descriptions of words, written based on Sternin and Rudakova (2017) method of generalizing dictionary definitions, and classified according to the error types in Table 2. Finally, changes in the frequency of meanings over time provided by the visualizations are examined and compared with historical usage data.

4 Results and Discussion

4.1 Model

FRED-T5-1.7B was chosen due to its performance in processing the Russian language (Zmitrovich et al., 2024). At the time of selection, it was the top performer on the RussianSuperGLUE benchmark (Shavrina et al., 2020), with a score of 0.762.

4.2 Training Data

FRED-T5-1.7B was trained on a dataset derived from "Small academic dictionary" (MAS) (Evgenyeva, 1981 1984).

The dataset was cleaned to remove usage labels, entries without usage examples or without informative definitions, such as *Состояние по знач. глаг. линять* [*State by the meaning of the verb "to shed"*], and those that provided grammatical rather than lexical information, such as *наречие к причастию приглашающий* [*Adverb to the participle "inviting"*]. The resulting dataset of 122,350 entries was partitioned into training, development, and test sets with a 90%/5%/5% split.

Each entry was formatted and began with the word "Контекст" ["Context"] followed by a usage example, then the phrase "Определение слова" ["Word definition"], and the word itself.[1]

4.3 Evaluation Data

The *RuShiftEval* competition's test set (Kutuzov and Pivovarova, 2021) was utilized for evaluation. The task focuses on detecting semantic changes in Russian nouns across three historical transitions: RuShiftEval-1 (Pre-Soviet:Soviet), RuShiftEval-2 (Soviet:Post-Soviet), and RuShiftEval-3 (Pre-Soviet:Post-Soviet). The competition provided a test set of gold change scores for 99 Russian nouns corresponding to the transitions.

4.4 String Similarity Metrics in Model Testing

BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERT-F1 (Zhang* et al., 2020) metrics from the *evaluate* library (Hugging Face, 2023) were employed for the definitions generated using the test part of the MAS dataset. BLEU measures n-gram overlap between texts, ROUGE-L focuses on the longest common subsequence, and BERT-F1 leverages contextual embeddings for semantic similarity. The evaluation results[2] are presented in Table 3.

Table 3: Fine-tuning Results of FRED-T5-1.7B on the MAS Dataset

Metric	**Value**
BLEU	11.02
ROUGE-L	29.36
BERT-F1	75.22

Low BLEU and ROUGE-L scores indicate that the model generates definitions differently from the test set, although high BERT-F1 scores imply semantic similarity.

At this stage, self-referential errors were fixed by excluding tokens related to the target word from being sampled in the model's output.

4.5 Rushifteval Testing

The paraphrase-multilingual-mpnet-base-v2 model (Transformers, 2023), additionally fine-tuned on RuSemShift, a similar dataset (Rodina and Kutuzov, 2020), was used to vectorize definitions. The distances between the definitions were calculated using the cosine distance. Results were compared against approaches from the Rushifteval task, as shown in Table 4.

Table 4: Algorithm Results Compared to Rushifteval Teams

Team	**Average**	**Word Representation Type**	**Model Used**
DeepMistake (post-competition)	**0.850**	**Contextual Emb.**	**XLM-R**
Proposed Approach	**0.815**	**Generated Definitions**	**FRED-T5-1.7B**
GlossReader	0.802	Contextual Emb.	XLM-R
DeepMistake	0.791	Contextual Emb.	XLM-R
vanyatko	0.720	Contextual Emb.	RuBERT
Other 10 Teams	0.457-0.178	…	…

[1]A special denoiser token <LM>, dedicated to the task of text continuation, was utilized.
[2]Out of 100, higher is better.

The proposed approach outperforms most entries in the Rushifteval competition.

Table 5: Comparison with definition generation approaches

Method	RuShiftEval-1	RuShiftEval-2	RuShiftEval-3	Base Model
Proposed Approach without vectorizer fine-tuning	**0.722**	**0.763**	**0.749**	**FRED-T5-1.7B**
Fedorova et al. (2024)	0.488	0.462	0.504	MT0-XL

A shown in Table 5, the proposed approach significantly outperforms the results of Fedorova et al. (2024). The vectorizer fine-tuning step was omitted to ensure that the results are directly comparable.

It could be noted that Fedorova et al. (2024) appears to retain unhelpful definitions in the training data, unlike proposed approach in 4.2, possibly resulting in their model reproducing non-informative patterns and the lower performance of their approach.

4.6 Visualization

Generated word vectors were clustered using the DBSCAN algorithm. Each cluster is represented by a prototypical definition closest to its centroid. DBSCAN parameters (`eps` and `min_samples`) are manually tuned by incremental adjustment to ensure the formation of cohesive clusters. Then, the temporal distribution of these meanings is displayed using bar charts, as shown in Figure 1.

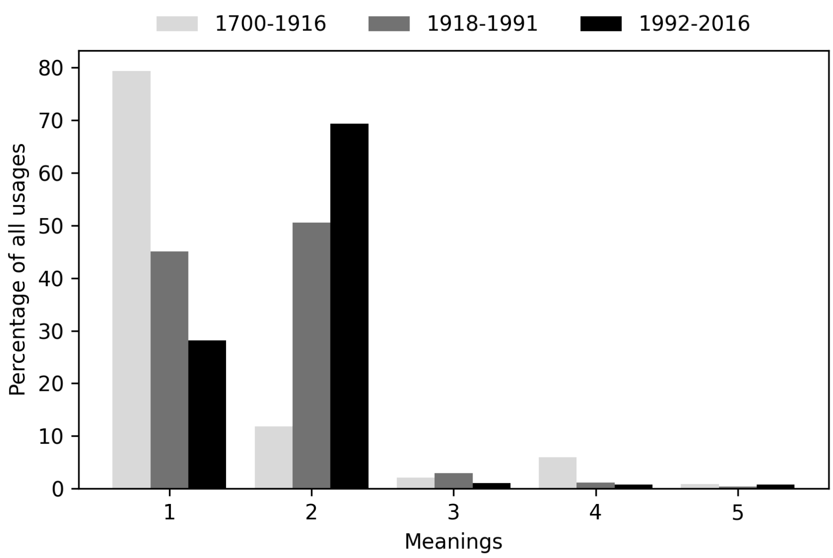

Figure 1: Semantic Shift of the Word *машина [machine/car]* (Parameters: eps=0.14, min_samples=5)

Meanings for *машина [machine/car]*:

1. A device or instrument for a specific task.
2. An automobile or vehicle.
3. An aircraft or helicopter.
4. A mechanically or thoughtlessly acting person.
5. A system of institutions or organizations.

4.7 Qualitative Analysis

For deeper examination, 20 words exhibiting semantic shifts from *Two Centuries in Twenty Words* (Dobrushina and Daniel', 2018) were selected: *знатный [noble]*, *кануть [to disappear]*, *классный [classy/cool]*, *мама [mom]*, *машина [machine/car]*, *молодец [young man/attaboy]*, *пакет* *[bag/package]*, *передовой [advanced]*, *пионер [pioneer]*, *пожалуй [perhaps]*, *пока [until/bye]*, *привет [hello]*, *пружина [spring]*, *публика [public]*, *свалка [landfill/fight]*, *сволочь [bastard]*, *стиль [style]*, *тётка [aunt]*, *тройка [three/a set of three]*, *червяк [worm]*. The usages were extracted from the diachronic sub-corpus of Russian National Corpus (Savchuk et al., 2024).

For each word, 300 instances were randomly sampled for each period of the corpus (pre-Soviet, Soviet, post-Soviet). The model generated definitions for each occurrence, followed by the creation of corresponding visualizations.

Next, the semantics of each word based on multiple dictionaries were described following Sternin and Rudakova (2017). To ensure comprehensive meaning descriptions, we synthesized information from 3 modern Russian dictionaries: *Big Explanatory Dictionary* (Kuznetsov, 1998), *Dmitriev's Explanatory Dictionary of the Russian Language* (Dmitriev, 2003), and *Ozhegov and Shvedova's Explanatory Dictionary*, in addition to *Two Centuries in Twenty Words*. Usage labels were omitted since the model wasn't trained to generate them.

The manually obtained semantic descriptions were compared with those in the visualization, and changes in their usage across periods for meanings corresponding to those in *Two Centuries in Twenty Words* were analyzed.

4.8 Qualitative Analysis of Generated Definitions

As a result of generalizing dictionary definitions, 121 meanings were compiled for 20 words. A total of 83 definitions were obtained using the proposed approach. Thus, excluding 5 incorrect definitions, 64.4% of the meanings were identified.

Table 6: Types of Definitions and Their Counts

Type of Definition	**Count**	**Percentage**
Correct	57	68.67%
Close	10	12.04%
Incorrect	5	6.02%
Insufficiently Specific	3	3.61%
Redundancy or Excessive Use of General Phrases	4	4.81%
Close, Redundancy or Excessive Use of General Phrases	1	1.20%
Overly Specific	3	3.61%
Self-reference	0	0.00%
Opposite Meaning	0	0.00%
Incorrect Part of Speech	0	0.00%

As shown in Table 6, the majority of definitions are correct without any errors or shortcomings (68.67%).

Common issues include close or incorrect meanings, such as defining *червяк [worm]* as an adult insect or describing *пожалуй [perhaps]* as a conjunction. Redundancy is present, exemplified by the repetitive "chaotic" in the definition of *свалка [landfill/fight]* ('Беспорядочная, беспорядочная схватка'), possibly due to the abundance of synonymous expressions in the training dataset, a common method in lexicology. Additionally, some definitions lack specificity, such as describing *мама [mom]* simply as 'a tender address to a woman.' These problems may arise from the model's limited world knowledge.

Another issue is insufficient context, leading to ambiguity in distinguishing meanings, as seen with *пионер [pioneer]* in *Pioneers listen to this and admire it [Пионеры слушают это и восхищаются]*.

4.9 Statistical Analysis of Semantic Shifts

For most of the words, the visualizations partially or fully align with the data from *Two Centuries in Twenty Words*, except for the word *пока [until/bye]*, where the visualization results contradict the study's findings. Overall, main meaning changes consistent with the book's data were identified in 12 out of 20 words. Additionally, changes partially aligned in 4 other words.

One of the best visualizations was created for the word *пакет [bag/package]*. 7 definitions were identified correctly, 4 of which appear only in the post-Soviet period.

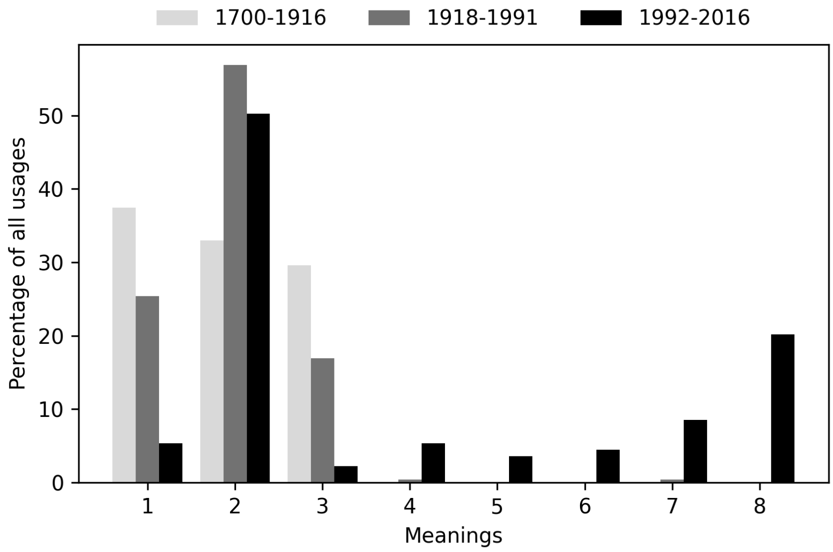

Figure 2: Semantic Shift of the Word *пакет [bag/package]* (Parameters: eps=0.11, min_samples=8)

Meanings for *пакет [bag/package]*:

1. A letter, parcel, etc., in such a form.
2. A paper or fabric pouch for storing, transporting, etc.
3. A letter, parcel, etc., sealed in such an envelope.
4. A collection of homogeneous, related objects, phenomena, etc.
5. A collection of software tools united by a certain criterion.
6. A part of something belonging to someone under certain conditions. (marked as incorrect)
7. A collection of homogeneous objects, documents, etc.
8. A collection of shares of a joint-stock company.

A comprehensive analysis is not feasible for *публика [public]* and *кануть [to disappear]*, because *Two Centuries in Twenty Words* does not provide sufficient usage frequency diagrams for their meanings.

Similarly, for *сволочь [bastard]*, only 2 out of 4 meanings were detected by the proposed approach (*употребляется как бранное слово [used as a swear word]* and *о подлом, гнусном человеке [referring to a vile, despicable person]*), both falling under 'Индивидуальное оскорбление [Individual insult]' in the book.

Conclusion

The study demonstrated the effectiveness of definition modeling in detecting and visualizing semantic shifts in the Russian language. A FRED-T5-1.7B model, fine-tuned on the MAS dictionary, was used to generate context-based word definitions. The model demonstrated high BERTScore similarity metrics on the test set, performed among the top solutions on the Rushifteval shared task and outperformed the results of Fedorova et al. (2024). A visualization algorithm was developed to represent semantic changes over time, allowing for reproducing a manual effort of studying semantic changes for a set of 20 words. Qualitative analysis of the results revealed that 68.67% of generated definitions were fully correct, with main meaning changes accurately detected in 12 out of 18 words available for analysis and partial alignment in 4 others. This shows that the approach could aid historical linguists and lexicographers in linguistic studies.

The findings can be applied to assess the extent of semantic shifts in lexemes, providing visualizations and definitions for each identified meaning.

Future research directions might include incorporating multiple dictionaries as training data or utilizing more advanced LLMs.

The code for this project and the model are available on GitHub: `https://github.com/tatarinovst2/work-definition-modeling`

References

Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov, Alexander Panchenko, Daniil Homskiy, and Adis Davletov. 2021. Deepmistake: Which senses are hard to distinguish for a word-in-context model. // *Computational Linguistics and Intellectual Technologies*, volume 20, P 16–30, 06.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // Jill Burstein, Christy Doran, and Thamar Solorio, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

D.V. Dmitriev. 2003. *Tolkovy slovar' russkogo yazyka: Ok. 2000 slovar. st., svyshe 12000 znacheniy [Explanatory Dictionary of the Russian Language: About 2000 Dictionary Entries, Over 12000 Meanings]*. Slovari Akademii Rossiyskoy. Astrel' [i dr.], Moskva. GUP IPK Ulyan. Dom pechati.

N.R. Dobrushina and M.A. Daniel'. 2018. *Dva veka v dvadtsati slovakh [Two Centuries in Twenty Words]*. Izdatel'skiy dom Vysshey shkoly ekonomiki, Moskva, 2 edition.

A.P. Evgenyeva. 1981-1984. *Slovar' russkogo yazyka: V 4-kh t. [Dictionary of the Russian Language: In 4 Volumes]*. Russkiy yazyk, Moskva, 4-e izd., ispr. i dop [4th ed., corrected and supplemented] edition. V 4-kh tomakh [In 4 volumes].

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. // Lun-Wei Ku, Andre Martins, and Vivek Srikumar, *Findings of the Association for Computational Linguistics: ACL 2024*, P 5712–5724, Bangkok, Thailand, August. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. // Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 3130–3148, Toronto, Canada, July. Association for Computational Linguistics.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. // Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 2499–2509, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Hugging Face. 2023. Evaluate. `https://github.com/huggingface/evaluate`. Retrieved November 15, 2023.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. // *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*, P 533–545.

Andrei Kutuzov, V. Fomin, V. Mikhailov, and Julia Rodina. 2020. Shiftry: Web service for diachronic analysis of russian news. // *Computational Linguistics and Intellectual Technologies*, volume 19, P 500–516, 01.

S.A. Kuznetsov. 1998. *Bol'shoy tolkovy slovar' russkogo yazyka: A-Ya [Large Explanatory Dictionary of the Russian Language: A-Ya]*. Norint, SPb. RAN. Inst. lingv. issled. Sost., gl. red. kand. filol. nauk S.A. Kuznetsov.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. // *Text Summarization Branches Out*, P 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. // *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: learning to define word embeddings in natural language. // *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, P 3259–3266. AAAI Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. // Pierre Isabelle, Eugene Charniak, and Dekang Lin, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. // Marilyn Walker, Heng Ji, and Amanda Stent, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2021. Zero-shot cross-lingual transfer of a gloss language model for semantic change detection. // *Computational Linguistics and Intellectual Technologies*, volume 20, P 578–586, 06.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. // Donia Scott, Nuria Bel, and Chengqing Zong, *Proceedings of the 28th International Conference on Computational Linguistics*, P 1037–1047, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

S.O. Savchuk, T.A. Arkhangelskiy, A.A. Bonch-Osmolovskaya, O.V. Donina, Yu.N. Kuznetsova, O.N. Lyashevskaya, B.V. Orekhov, and M.V. Podryadchikova. 2024. Natsionalny korpus russkogo yazyka 2.0: novye vozmozhnosti i perspektivy razvitiya. *Voprosy Yazykoznaniya*, 2:7–34.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. // Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, P 1–23, Barcelona (online), December. International Committee for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A russian language understanding evaluation benchmark. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

I.A. Sternin and A.V. Rudakova. 2017. *Slovarnye definicii i semanticheskiy analiz [Dictionary Definitions and Semantic Analysis]*. Istoki, Voronezh.

Sentence Transformers. 2023. paraphrase-multilingual-mpnet-base-v2. `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`. Retrieved April 19, 2024.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. // *International Conference on Learning Representations*.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. // Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, P 507–524, Torino, Italia, May. ELRA and ICCL.