PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

UDC 004.8:81'322:81-115
10.25209/2079-3316-2026-17-2-103-146

Selecting and Controlling Sense Granularity for Lexical Semantic Change Detection

Denis Vladislavovich **Kokosinskii**[1]✉, Dominik **Schlechtweg**[2], Nikolay Viktorovich **Arefyev**[3]

[1] Lomonosov Moscow State University, Moscow, Russia
[2] University of Stuttgart, Germany
[3] University of Oslo, Norway
[1]✉*deniskokossskappa@gmail.com*

Abstract. Studying how word meanings change is a long standing problem in linguistics. We present an automatic approach that groups usages of a word into clusters corresponding to word senses and measures how the usage frequency in those senses changes between historical periods. The method follows the established procedures used to create recent human-annotated language resources (Diachronic Word Usage Graphs) and lets users adjust how coarse- or fine-grained the senses should be. We also introduce a novel metric that allows to reliably evaluate the quality of the clusters, specifically tailored for the Diachronic Word Usage Graphs.

Across multiple languages, the approach performs on par with, and often better than, existing alternatives while providing clear, interpretable outputs that reveal which word senses contribute to the semantic change. (*Linked article texts in English and in Russian*).

Key words and phrases: lexical semantic change detection, diachronic word usage graphs, word sense induction

2020 *Mathematics Subject Classification*: 91F20; 68T50

For citation: Denis V. Kokosinskii, Dominik Schlechtweg, Nikolay V. Arefyev. *Selecting and Controlling Sense Granularity for Lexical Semantic Change Detection*. Program Systems: Theory and Applications, 2026, **17**:2(71), pp. 103–146. (*In English, in Russian*). https://psta.psiras.ru/read/psta2026_2_103-146.pdf

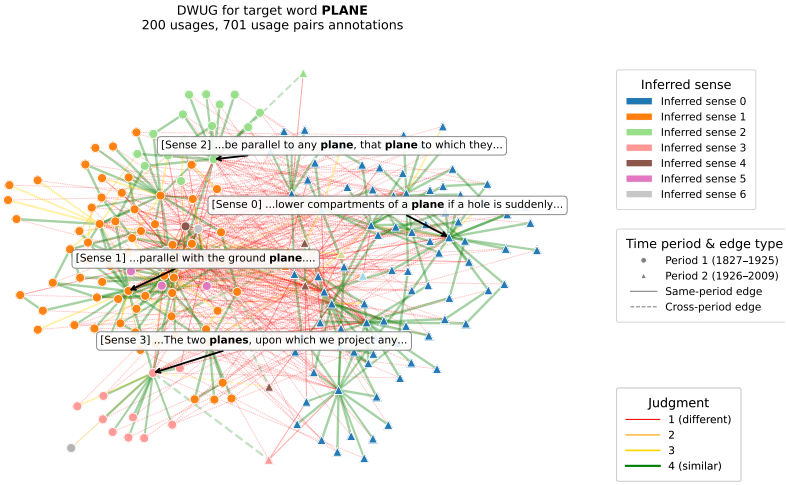


Figure 1. Diachronic Word Usage Graph for the target word *plane* for two time periods. Nodes are word usages, edges are human judgements of semantic proximity in word usage pairs.

Introduction

Lexical Semantic Change Detection (LSCD; see [1–3]) involves automatically identifying words that change meaning over time. Schlechtweg et al. [4] formulates the LSCD task as follows: given a list of *target words* and two sets of *word usages* from two time periods, the goal is to rank the words by the magnitude of their semantic change.

Datasets for LSCD are typically annotated using the Diachronic Word Usage Graph (DWUG) framework [5], illustrated on Figure 1. Assume we are given a target word (e.g. *plane*) and two time periods of interest (e.g. XIX century and XX century). As a preliminary step, the examples of sentences or paragraphs with the target word are collected from two corpora of real texts spanning the desired time periods; these examples are referred to as *word usages*. Then, human annotators assess how close the meanings of a target word are in individual pairs of usages. The resulting set of pairwise annotations gets represented in a graph, called the Diachronic Word Usage Graph, where nodes are word usages and edges are annotations.

In the next step, the resulting graph gets clustered in an effort to determine distinct word senses. Finally, the authors semantic change is defined as change in sizes of clusters/senses. In more mathematical

terms, the higher the Jensen-Shannon Distance between the frequency distributions of word senses in two time periods, the stronger the semantic change.

Fully automatic LSCD systems vary in how they model semantic change. Some rely only on pairwise semantic proximity between usages, while others produce fully automatic inferred senses via clustering. Thus, LSCD systems internally solve other well-established tasks in the field of lexical semantics: Word-in-Context and Word Sense Induction.

Word-in-Context (WiC) is a task of estimating semantic proximity in a word usage pair. E.g. a usage pair "...parallel with the ground *plane*..." and "...lower compartments of a *plane*..." has low semantic proximity, and "...parallel with the ground *plane*..." and "...the two *planes*, upon which we project..." has high semantic proximity. Some LSCD systems also internally solve the Word Sense Induction (WSI) task. WSI is a clustering task, where, given a set of word usages for some target word, the goal is to group those usages into discrete clusters; the clusters are to align with word senses.

Recent LSCD systems can be broadly categorized in WSI-based and WiC-only systems. WSI-based LSCD systems have an advantage in interpretability, since they estimate semantic change directly from changes in inferred senses, while the internal mechanics of WiC-only systems are less interpretable. WiC-only systems have achieved state-of-the-art results [6–8], outperforming WSI-based alternatives [9–11].

A contradictory situation occurred in the field: WSI-based LSCD systems that perform clustering and thus more closely mirror the DWUG annotation pipeline underperform relative to WiC-only systems that implement a completely different approach to modeling semantic change [2,12]. In this paper, our aim is to address this contradiction and present a fully automatic LSCD system that closely follows the well-established DWUG framework, providing a clear and interpretable way to automatically model semantic change.

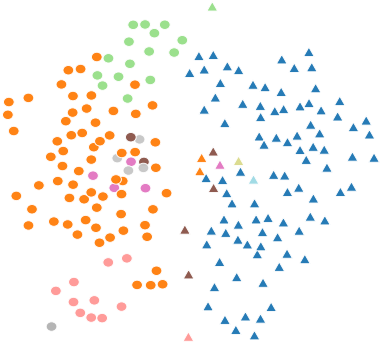
To summarize, our two main contributions are as follows:

(1) A LSCD system that explicitly infers word senses and achieves state-of-the-art performance. Our system offers controllable sense granularity, which is important because the sense boundaries are often fuzzy [13].

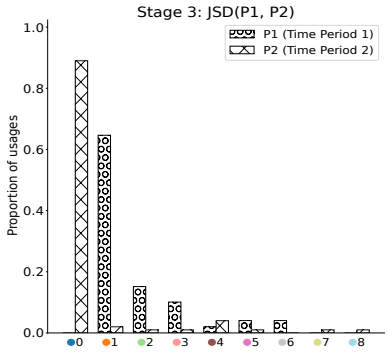
(2) A novel AnnotRI metric for the WSI task, better suited for DWUGs. It uses only reliable manually annotated data as gold labels and enables evaluation at varying levels of granularity of word senses.

(*a*) Stage 0: Usages from two time periods

(*b*) Stage 1: Word-in-Context



(*c*) Stage 2: Word Sense Induction



(*d*) Stage 3: JSD(P1, P2) semantic change score

Figure 2. Steps of the DWUG pipeline for quantifying semantic change

1. Related work

1.1. Diachronic Word Usage Graphs

We will first describe in more detail how the semantic change is quantified in the DWUG framework. It is a three-step procedure, illustrated on Figure 2.

The first step (Figure 2*a*) involves human annotators, who rate the semantic proximity of usage pairs sampled from two time periods on a four-grade scale reflecting Blank's degrees of semantic proximity: homonymy, polysemy, context variance, and identity [14].

It is essentialy Word-in-Context task, solved by human annotators. Usages become nodes, and the human judgments define edge weights. In the second step (Figure 2*b*), usages from both periods are jointly clustered using correlation clustering [**15**]; the resulting clusters are interpreted as word senses.

Thus, Word Sense Induction is solved in the DWUG framework semi-automatically: the manual human judgements are clustered with an automatic method. In the third and final step (Figure 2*c*), the ground-truth semantic change score for each target word is calculated as the Jensen-Shannon distance between the sense distributions across the two periods.

Based on the use of Jensen-Shannon distance, we refer to the resulting semantic change score as JSD (Figure 2*d*).

In summary, existing DWUGs include manually annotated usage pairs, *silver*[1] *inferred senses* obtained by clustering the annotated graph, and a final semantic change score for each target word. Originally, DWUGs were created [**16**] for English (en), German (de) and Swedish (sv), and subsequently for Spanish (es; [**12**]), Norwegian (three time periods, two datasets: DWUG-no12, DWUG-no23; [**17**]), and Chinese (zh; [**18**])[2].

Kutuzov and Pivovarova [**19**] use an alternative COMPARE [**20**] semantic change score for Russian DWUG-ru12, DWUG-ru23, and DWUG-ru13. For each target word, COMPARE is the average of annotator judgments in the «compare» group, consisting of all usage pairs with the first usage from the older corpus and the second from the newer one (the dashed lines on Figure 1 and Figure 2). COMPARE assumes that predominantly low judgments in this group indicate substantial semantic change. This procedure does not involve clustering and thus does not produce inferred senses. In other words, COMPARE only involves the Word-in-Context subtask, and no Word Sense Induction. The intermediate steps in the COMPARE annotation procedure are therefore less interpretable than those in JSD, as they do not reveal which senses appeared, disappeared, or changed in frequency over time. The JSD score, however, has its own downside: it relies on automatic clustering, which may introduce additional errors. Despite these differences, COMPARE and JSD have been shown to correlate strongly [**5**].

[1] We use the word silver to highlight the semi-automatic nature (manual pairwise annotation with automatic clustering) of inferred senses in DWUG.

[2] For exact versions of the datasets we use in our evaluation refer to Appendix 8

1.2. Lexical Semantic Change Detection methods

Automatic Lexical Semantic Change Detection methods can be broadly categorized into WSI-based and WiC-only Average Pairwise Distance (APD) methods, loosely mirroring the JSD and COMPARE annotation procedures. WSI-based methods [9–11, 21] employ clustering to implicitly infer senses, thus addressing WSI as an intermediate step. They then calculate the Jensen-Shannon distance between cluster distributions across periods. In contrast, APD methods [22–24] quantify semantic change by averaging model-predicted semantic similarity scores in word usage pairs, similarly to COMPARE.

Both the WSI-based and APD approaches rely on a WiC model that produces similarity scores for pairs of usages, the step that is manually annotated in DWUGs. WiC is a well-established task in lexical semantics on its own with a number of automatic systems to choose from. For instance, Laicher et al. [25] use BERT [26] and Giulianelli et al. [27] use XLM-Roberta [28] as WiC models, surpassing non-transformer approaches without any fine-tuning. Specialized WiC models, namely GlossReader [8], DeepMistake [7], and XL-Lexeme [6], further improve quality through task-specific fine-tuning. More recently, [2, 29, 30] demonstrate the capabilities of general-purpose LLMs as WiC models, reporting 1–3% improvements over aforementioned specialized models in several languages. However, we leave LLMs as WiC models out of scope due to their substantially higher inference cost[3] relative to smaller specialized models.

A recent unified evaluation on DWUGs [2] finds that APD outperforms WSI-based methods on all DWUG benchmarks, including those with JSD gold labels. This is counterintuitive, since WSI mirrors the JSD pipeline more closely than APD. In addition to end-to-end LSCD evaluation, Periti and Tahmasebi [2] separately evaluate the WiC and WSI components, but their WSI evaluation has limitations:

(i) The WiC predictions are only done for graph edges that had manual annotation, implicitly incorporating gold information at prediction time (see Section 2).

[3] For example, [30] show Llama-3.3-70b to be the best performing LLM for LSCD. Llama-3.3-70b has 70 billion parameters, while XL-Lexeme has only 0.35 billion, so the inference cost is ≈200 times higher for Llama.

(*ii*) Silver inferred senses are treated as gold labels in WSI evaluation, although they are not explicitly annotated by humans but inferred using an automated procedure (see Section 2).

(*iii*) No systematic hyperparameter selection is used for correlation clustering, leading to heuristic suboptimal choices that can underrate strong models (see Section 4.1).

2. Word Sense Induction evaluation on DWUGs

We aim to build a WSI-based system, we thus need a way to reliably evaluate WSI quality in our diachronic setup. One of the key challenges in using DWUGs for WSI is the reliability of gold clustering. The inferred senses in the DWUGs are not manually annotated but are created through an semi-automatic procedure. The reliability is further undermined in some DWUGs by the sparsity of edge annotation, which means only a small proportion of usage pairs is annotated (e.g., ≈10% of all possible edges in DWUG-en). Although correlation clustering can handle missing edges, such clustering is inherently less reliable than clustering on a fully annotated graph.

To address the reliability issue of silver inferred senses, we propose an alternative WSI evaluation for DWUGs. In addition to comparing predicted clusters with silver inferred senses using Adjusted Rand Index (ARI), we introduce AnnotRI (Annotation Rand Index, inspired by the Rand Index; [**31**]). Unlike ARI, which compares complete clusterings, AnnotRI evaluates directly on manually annotated usage pairs. It measures the percentage of usage pairs in which the system's clustering decision (the same cluster or different clusters for a given usage pair) matches the binarized human annotation. In other words, AnnotRI is the binary accuracy of a WSI system on a set of human-annotated usage pairs. AnnotRI thus bypasses inferred senses and relies directly on manual judgments. By varying the binarization threshold for human annotations along the four-grade scale (homonymy, polysemy, context variance, and identity) AnnotRI can assess clustering quality at different sense granularities using the same gold data; we report AnnotRI@1.5, AnnotRI@2.5, and AnnotRI@3.5. This is helpful because the boundaries between the senses of the word are often not well defined [**13**].

3. WSI-based LSCD system

We propose a fully automatic LSCD system that follows the DWUG framework, which we denote CC+JSD (Correlation Clustering+Jensen-Shannon Distance). First, a specialized WiC model produces pairwise semantic proximity of *all* usage pairs of the target word in both time periods. Notably, our system works with any WiC model, but requires some model-specific parameter selection in the clustering step (see Section 4.1). In our evaluation, we include specialized WiC models: GlossReader [8], XL-Lexeme [6], and DeepMistake [32][4]. Having obtained WiC predictions for usage pairs, we use them as the edge weights to build a DWUG.

In the second step, we cluster the DWUG using Correlation Clustering [15], which we denote CC for short. Formally, CC seeks a partition C of the usage set U that minimizes the total disagreement:

$$\mathcal{L}(C,\theta) = \sum_{\substack{(u_i,u_j):\, r_{ij}\geq\theta,\\ c(u_i)\neq c(u_j)}} r_{ij} \quad + \sum_{\substack{(u_i,u_j):\, r_{ij}<\theta,\\ c(u_i)=c(u_j)}} |r_{ij}|, \tag{1}$$

where

- r_{ij} is the similarity score predicted by the WiC model,
- $c(u)$ is the cluster assignment and
- θ is the threshold parameter.

The first term penalizes placing similar usages (with $r_{ij} \geq \theta$) into different clusters, and the second penalizes placing dissimilar usages into the same cluster.

We chose correlation clustering specifically over other clustering algorithms to mimic the DWUG annotation pipeline as closely as possible. The original reasons described in [16] mostly translate to our fully-automatic setting:

(*i*) CC finds the optimal number of clusters on its own.

(*ii*) It is more robust to errors, as it uses the global information on the graph. That is, a wrong judgment can be outweighed by the correct ones.

[4] GlossReader and DeepMistake have multiple publicly available checkpoints, we select the ones with the best performance on previously held shared tasks. See Appendix 9.

(*iii*) It directly optimizes an intuitive criterion — the sum of cluster disagreements, i.e., the sum of negative edge weights within clusters plus the sum of positive edge weights across clusters.

We place special emphasis on tuning the *threshold* parameter in CC (see Section 4.1). The threshold determines which edge weights are treated as positive versus negative in the optimization objective of CC and is crucial for cluster quality. Adjusting it controls cluster granularity (coarser vs. finer). At the same time, our novel metric, AnnotRI (see Section 2), allows us to fine-tune the granularity of the predicted clusters to align with four levels of semantic proximity.

Finally, having clustered the usages, our CC+JSD method collects the two frequency distributions of usages in clusters for two time periods. We compute the Jensen-Shannon distance between those two distributions as the final semantic change score for a given target word. The LSCD benchmarks require to rank a list of target words, we assign the semantic change score to each of them and sort the list accordingly.

Previously, Periti and Tahmasebi [2] reported the evaluation of a similar system in their computational annotation setup. However, in their setup a WiC model only reannotates the manually annotated edges, hence the name "computational annotation". In other words, it skips the edges that were missing in the original DWUG annotation, effectively leaking the DWUG structure at prediction time. We instead use WiC predictions for all possible edges, presenting a more realistic setup for automatic LSCD systems.

4. Evaluation results

4.1. Word Sense Induction

We first study how the CC threshold affects the WSI results on DWUG benchmarks. We use specialized WiC backbones: XL-Lexeme [6], GlossReader [8], and DeepMistake [7]. For each model, we collect the distribution of similarity scores across all DWUGs[5] and use ten

[5] Score ranges depend on the WiC model. We therefore use a fixed grid for a given model across DWUGs to generalize to new datasets and languages.

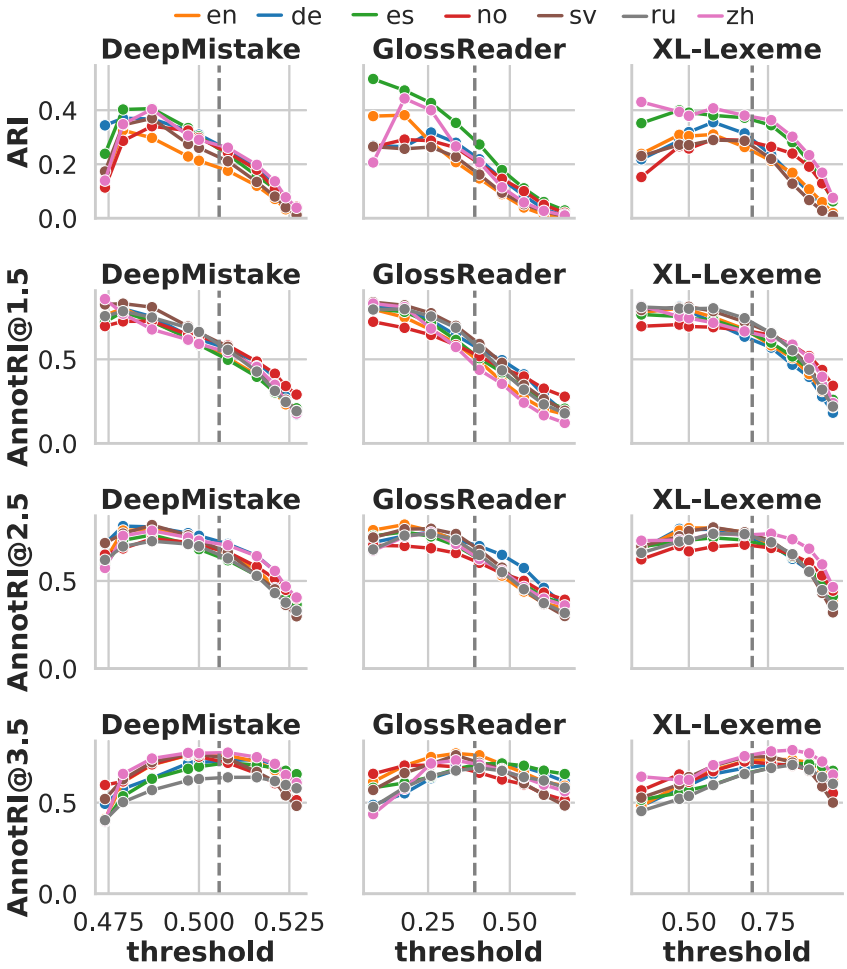


Figure 3. The WSI results for correlation clustering depending on the threshold. The grey vertical line represents the average of similarity scores across all benchmarks for each model, the CC threshold used in [2].

quantiles of this distribution as candidate CC thresholds. We report ARI (against silver inferred senses), as well as AnnotRI@1.5, AnnotRI@2.5, and AnnotRI@3.5. The results are presented in Figure 3.

We observe consistent trends across benchmarks for each model. Typically, the optimal CC threshold increases with the AnnotRI threshold: similarity quantiles of 0.1–0.2 yield the highest AnnotRI@1.5, quantiles of 0.2–0.3 are optimal for AnnotRI@2.5, and quantiles of 0.4–0.5 are optimal for AnnotRI@3.5. In particular, setting the threshold at the average distance (as in [2]) produces relatively fine-grained clusters, which align best with AnnotRI@3.5. Consequently, this threshold is not optimal for ARI. As expected, ARI better aligns with AnnotRI@2.5, as the threshold of 2.5 has been used in the creation of silver inferred senses, treated as gold labels in the ARI calculation.

For our end-to-end LSCD evaluation, we set the threshold based on AnnotRI@2.5. More specifically, to select the CC threshold without data leakage, we perform a 10-fold cross-validation across benchmarks (DWUG-en, DWUG-es, etc.): for each held-out benchmark we select the threshold that maximizes the average AnnotRI@2.5 on the remaining nine[6]. We also test the same setup with ARI instead of AnnotRI@2.5 for DeepMistake (see Appendix 10) and find that optimizing AnnotRI@2.5 is beneficial for the end-to-end LSCD metric on six out of seven JSD-based benchmarks, validating the usefulness of AnnotRI.

4.2. End-to-end LSCD evaluation

Having thoroughly studied the WSI substep, we return to the Lexical Semantic Change Detection evaluation for our CC+JSD system (Correlation Clustering+Jensen-Shannon Distance). The WiC model is a parameter of this system. CC+JSD first collects WiC predictions for all usage pairs to automatically build a DWUG, then runs Correlation Clustering to infer senses and finally computes the cluster-level Jensen-Shannon Distance between the two time perids to produce a single semantic change score for a target word. For a selected WiC model (XL-Lexeme, DeepMistake or GlossReader) we use cross-validation of the CC threshold with AnnotRI@2.5 as the target metric, as described in Section 4.1.

We evaluate the LSCD systems on DWUG benchmarks with the standard LSCD metric: Spearman rank correlation between predicted and silver semantic change scores for the provided list of target words. We

[6]This yields an unbiased estimate for a new language/dataset outside the test benchmarks.

also include the best WSI-based systems from [2]: XL-Lexeme Affinity Propagation (AP+JSD) and What-is-Done-is-Done (WiDiD). We also report APD results with the XL-Lexeme, DeepMistake, and GlossReader backbones in two versions: APD_{all}, which follows [2] and computes APD on all usage pairs for a target word and APD_{cmp}, which follows [32] and computes APD only inside the "compare" group (see Section 1.1).

The evaluation results are presented in Table 1. Tool CC+JSD establishes the new state-of-the-art on all JSD-based benchmarks except for DWUG-es[7], outperforming both the previous WSI-based approaches and APD. We also see that in some of the JSD-based benchmarks (DWUG-en, DWUG-de, DWIG-sv) the Spearman rank correlation with the gold semantic changes scores is .891, thus approaching optimal performance.

Overall, we see methods that more closely adhere to the gold data creation process achieve higher quality: CC+JSD often outperforms APD on JSD-based benchmarks, while the opposite holds for COMPARE-based DWUG-ru. We thus resolve the contradiction observed in previous work, where WSI-based systems performed worse than APD, despite more closely following the annotation procedure for JSD-based DWUG benchmarks. We mainly attribute the weaker performance of previous WSI-based approaches to the misalignment between the predicted clusters and the silver inferred senses in the clustering step, which we addressed in Section 4.1.

5. Interpretability and Error Analysis

As mentioned previously, WSI-based methods allow us to track how specific senses of a word change over time. We demonstrate this interpretability by examining the results of GlossReader CC+JSD on DWUG-en.

We assign sense descriptions (cluster labels) by inspecting a sample of a dozen usages from each inferred sense and selecting the best-fitting description from the Oxford Dictionary for the given word.[8] The first thing we observe is the abundance of usages with incorrect clustering.

[7] DWUG-es underpinned the SemEval 2021 shared task, where DeepMistake DM-MCL→es+APD won. We attribute the strong DWUG-es APD results to careful task-specific tuning of the DeepMistake pipeline.

[8] For large-scale sense description annotation, automatic methods can be used; e.g., [30, 33].

TABLE 1. LSCD evaluation results on DWUGs. Spearman rank correlation is the metric. COMPARE is used as gold labels for Russian, JSD for other DWUGs. The star (*) denotes reproduced results from [2]. Best results for WSI-based systems and APD systems are in **bold** font, the best overall is **underlined**.

WiC Model	LSCD	en	de	es	no_{12}	no_{23}	sv	zh	AVG
WSI-based Methods									
XL-Lexeme	AP+JSD*	.49	.50	.12	.30	.38	.12	.31	.32
XL-Lexeme	WiDiD*	.65	.68	.52	.56	.46	.47	.56	.56
XL-Lexeme	CC+JSD	.72	.84	.65	.55	.59	.85	**.77**	.71
DeepMistake	CC+JSD	.81	**.91**	**.67**	**.85**	.61	**.90**	.76	**.79**
GlossReader	CC+JSD	**.891**	.76	.65	.64	**.70**	.83	.68	.74
APD									
XL-Lexeme	APD^*_{all}	**.886**	.84	.66	.66	.64	.81	**.73**	.75
XL-Lexeme	APD_{cmp}	.87	**.84**	.56	.65	**.66**	**.82**	**.73**	**.73**
DeepMistake	APD_{cmp}	.87	**.87**	**.72**	**.83**	.63	.77	.67	.77
GlossReader	APD_{cmp}	.87	.78	.70	.62	.65	.79	.70	.73

WiC Model	LSCD	ru_{12}	ru_{13}	ru_{23}	AVG
WSI-based Methods					
XL-Lexeme	AP+JSD*	.11	.13	.05	.10
XL-Lexeme	WiDiD*	.18	.36	.35	.30
XL-Lexeme	CC+JSD	.62	**.76**	.66	.68
DeepMistake	CC+JSD	**.70**	.75	**.81**	**.75**
GlossReader	CC+JSD	.66	**.76**	.74	.72
APD					
XL-Lexeme	APD^*_{all}	.80	**.86**	.82	**.826**
XL-Lexeme	APD_{cmp}	.79	**.85**	.81	.82
DeepMistake	APD_{cmp}	**.83**	**.84**	**.83**	**.833**
GlossReader	APD_{cmp}	.78	.81	.78	.79

In almost every cluster, we assign the label for the most common sense (determined manually) inside the cluster, and ignoring the usages with errors. Seeing so many errors is surprising, given the high overall LSCD metric of 0.891 in the setup we analyze. Therefore, we argue that there is still much room for quality improvement at the usage level.

Table 2. Top-3 words with the highest semantic change score from DWUG-en predicted by GlossReader CC+JSD. Clusters with five or fewer examples in either time period were filtered out. Each word has a total of 100 usages in each time period. The XIX period includes usages from years 1810–1860 and the XX period includes years 1960–2010.

		Time Period		
Word	**Sense Description**	**XIX**	**XX**	**Δ**
plane	geometric	71	23	-48
	aircraft	20	63	+43
graft	political bribery	15	67	+52
	agricultural technique	41	7	-34
	medical transplant	34	12	-22
prop	support (abstract)	24	21	-3
	support (physical)	29	10	-19
	theatrical/film props	12	20	+8

Table 2 shows how the distribution of usages in the predicted clusters changes between the two time periods for three words with the highest predicted degree of semantic change.

Despite usage-level errors, the system is still able to identify sense-level semantic change. For example, the system captures the semantic change for the highest-ranking word *plane*[9], shifting from being a predominantly geometric term to becoming used to describe a type of aircraft.

However, the same example also highlights a limitation: 20 usages from 1810–1860 were incorrectly[10] assigned to a large "aircraft" cluster, even though the concept did not exist in that period. As a result, the system may struggle to detect the true emergence or disappearance of a sense, sometimes interpreting it as a frequency shift within an existing sense.

[9] which is also the highest-ranking word according to gold labels

[10] we inspected them all manually

TABLE 3. Top-3 words with the highest difference in predicted semantic change score between CC thresholds optimized for AnnotRI@1.5 and AnnotRI@3.5. For each lemma there is a total of 200 usages. Clusters with 10 usages or less are excluded.

Lemma	CC threshold optimized for AnnotRI@1.5		AnnotRI@3.5	
prop	physical	127	physical	45
			construction	31
			theatrical/film	31
	abstract	65	abstract	45
stab	all	187	stabbing act	124
			attempt	24
			attempt	10
			pain	13
multitude	all	200	large group	101
			variety	98

Another issue is the abundance of tiny clusters in the predictions, often with only a single usage[11]. After manual inspection, we conclude that these are predominantly clustering artifacts rather than genuine niche senses.

To conclude, while the system demonstrates good LSCD performance and captures overall semantic change, there is still room for improvement on the sense level and even more so on the usage level.

We also examine the clusters with different granularity. We take three words where the predicted JSD scores with a lower CC threshold (optimized for AnnotRI@1.5) and the higher CC threshold (optimized for AnnotRI@3.5) differ the most. The summary is presented in Table 3.

The largest difference occurs for the word *prop*, where the lower threshold reduces the JSD from 0.72 to 0.12. Basically, every "physical" sense of the word gets merged into a single cluster when using the lower threshold. We also see that for two of the three words, the lower CC threshold yields an "all" cluster with almost all usages, which is arguably suitable for words *stab* and *multitude*.

[11] A similar issue was also reported by the authors of DWUGs when using correlation clustering [**16**].

While analyzing the predictions with the higher CC threshold, we were frequently unable to distinguish between senses; for the word *stab* the two "attempt" clusters contained usages with a very similar sense. In general, although the CC threshold optimized for AnnotRI@3.5 aligns better with the Oxford Dictionary, it also frequently introduces many small clusters indistinguishable (at least for us) from the other clusters. We hypothesize that per-word selection of the CC threshold may address this issue, which we leave for future work.

6. Conclusion

We proposed a practical, fully automatic WSI-based LSCD system that follows the DWUG annotation paradigm end-to-end: WiC similarities for all usage pairs, CC with a tuned threshold, and JSD over predicted senses. We also introduced AnnotRI, which evaluates WSI directly against manual pair annotations and naturally supports different sense granularities. Guided by AnnotRI, careful tuning of the CC threshold yields strong and often state-of-the-art LSCD results on JSD-based benchmarks while retaining the interpretability of sense-based models. Our analysis shows that, despite errors for individual word usages, WSI delivers actionable insights by tracking which senses drive semantic change.

7. Limitations

- Our WiC backbones exclude large LLMs due to cost and latency; results may improve with stronger yet efficient models.
- CC jointly clusters usages from both periods, which may limit detection of truly novel/vanishing senses.
- DWUGs are sparsely annotated in some languages; AnnotRI uses only annotated pairs and may reflect annotator coverage biases.
- Computing all-pairs WiC similarities is quadratic in the number of usages; scaling to larger DWUGs requires approximation or pruning.
- Performance varies with gold definitions (JSD vs. COMPARE); conclusions for one supervision signal may not generalize to the other.

TABLE 4. LSCD evaluation using Spearman rank correlation between predicted and gold semantic change scores for DeepMistake CC+JSD using AnnotRI@2.5 and ARI.

WiC Model	WSI metric	en	de	es	no12	no23	sv	zh
DeepMistake	AnnotRI@2.5	.81	**.91**	**.67**	**.85**	**.61**	**.90**	**.76**
DeepMistake	ARI	**.85**	.71	.57	.69	.59	.87	.74

8. Benchmark versions

DWUG-en 2.0.1, DWUG-de 2.3.0, DWUG-es (LSCDiscovery) 3.0.0, DWUG-no (NorDiaChange) 1.0.1, DWUG-ru (RuShiftEval 2021) 2.0.1, DWUG-sv 2.0.1, DWUG-zh (ChiWUG) 1.0.0

9. Model versions

XL-Lexeme, [6]. We use cosine distance between unnormalized embeddings, as the original paper suggests.

GlossReader, part of the winning solution [34] at [12]. We use L1 distance between L1-normalized embeddings, as the original paper suggests.

DeepMistake, denoted as MCL+RSS+DWUG-es+XL-WSD in [32].

10. Using AnnotRI and ARI for threshold selection

Table 4 presents LSCD results when employing AnnotRI@2.5 and ARI as metrics optimized in cross-validation to select CC threshold.

References

[1] Montanelli S., Periti F.. "A survey on contextualised semantic shift detection", *ACM Computing Surveys*, **56**:11 (2024), id. 282, 38 pp. arXiv 2304.01666 ↑104

[2] Periti F., Tahmasebi N.. "A systematic comparison of contextualized word embeddings for lexical semantic change", *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. V. 1: *Long papers*, NAACL 2024 (Mexico City, Mexico, June 16–21, 2024), ACL, 2024, ISBN 979-8-89176-120-9, pp. 4262–4282. arXiv 2402.12011 ↑104, 105, 108, 111, 112, 113, 114, 115

[3] Kutuzov A., Ovrelid L., Szymanski T., Velldal E.. "Diachronic word embeddings and semantic shifts: A survey", *Proceedings of the 27th International Conference on Computational Linguistics*, COLING-2018 (Santa Fe, New Mexico, USA, August 20–26, 2018), ACL, 2018, ISBN 978-1-5108-6906-6, pp. 1384–1397. URL arXiv 1806.03537 doi ↑104

[4] Schlechtweg D., McGillivray B., Hengchen S., Dubossarsky H., Tahmasebi N.. "SemEval-2020 task 1: Unsupervised lexical semantic change detection", *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval2020 (Barcelona, Spain (Online), December 12–13, 2020), 2020, ISBN 9781713828389, pp. 1–23. URL doi arXiv 2007.11464 ↑104

[5] Schlechtweg D.. *Human and computational measurement of lexical semantic change*, Thesis Dr. phil., Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart, 2023, 227 pp. doi ↑104, 107

[6] Cassotti P., Siciliani L., DeGemmis M., Semeraro G., Basile P.. "XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE", *Conference: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. V. 2: *Short papers* (Toronto, Canada, July 9–14, 2023), ACL, 2023, ISBN 978-1-959429-70-8, pp. 1577–1585. doi URL ↑105, 108, 110, 111, 119

[7] Arefyev N., Fedoseev M., Protasov V., Homskiy D., Davletov A., Panchenko A.. "DeepMistake: which senses are hard to distinguish for a word-in-Context model" (Moscow, 2021), Computational Linguistics and Intellectual Technologies, vol. **20**, Papers from the Annual International Conference "Dialogue", 2021, ISBN 978-5-7281-3032-1, pp. 16–30. doi URL ↑105, 108, 111

[8] Rachinskiy M., Arefyev N.. "GlossReader at SemEval-2021 task 2: Reading definitions improves contextualized word embeddings", *Conference: Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval-2021 (Bangkok, Thailand, August 5–6, 2021), ACL, 2021, ISBN 978-1-954085-70-1, pp. 756–762. doi URL ↑105, 108, 110, 111

[9] Martinc M., Montariol S., Zosa E., Pivovarova L.. Capturing evolution in word usage: Just add more clusters? *Companion Proceedings of the Web Conference 2020*, WWW'20 Companion (Taipei, Taiwan, April 20–24, 2020), ACM, 2020, ISBN 978-1-4503-7024-0, pp. 343–349. arXiv 2001.06629 doi URL ↑105, 108

[10] Martinc M., Kralj Novak P., Pollak S.. "Leveraging contextual embeddings for detecting diachronic semantic shift", *Proceedings of the Twelfth Language Resources and Evaluation Conference*, LREC 2020 (Marseille, France, May 11–16, 2020), ELRA, 2020, ISBN 979-10-95546-34-4, pp. 4811–4819. arXiv 1912.01072 doi ↑105, 108

[11] Periti F., Ferrara A., Montanelli S., Ruskov M.. "What is done is done: An incremental approach to semantic shift detection", *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022), ACL, 2022, ISBN 978-1-955917-42-1, pp. 33–43. doi URL URL ↑105, 108

[12] Zamora-Reina F. D., Bravo-Marquez F., Schlechtweg D.. "LSCDiscovery: A shared task on semantic change discovery and detection in Spanish", *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022), ACL, 2022, ISBN 978-1-955917-42-1, pp. 149–164. arXiv 2205.06691 ↑105, 107, 119

[13] Erk K., McCarthy D., Gaylord N.. "Measuring word meaning in context", *Computational Linguistics*, **39**:3 (2013), pp. 511–554. ↑105, 109

[14] Blank A.. "Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change", *Historical Semantics and Cognition*, eds. A. Blank, P. Koch, Mouton de Gruyter, Berlin–New York, 1999, ISBN 9783110804195, pp. 61–90. ↑106

[15] Bansal N., Blum A., Chawla S.. "Correlation clustering", *Machine Learning*, **56**:1 (2004), pp. 89–113. ↑107, 110

[16] Schlechtweg D., Tahmasebi N., Hengchen S., Dubossarsky H., McGillivray B.. "DWUG: A large resource of diachronic word usage graphs in four languages", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2021 (Punta Cana, Dominican Republic and Online, November 7–11, 2021), ACL, 2021, ISBN 978-1-7138-9162-8, pp. 7079–7091. arXiv 2104.08540 % ↑107, 110, 117

[17] Kutuzov A., Touileb S., Maehlum P., Enstad T., Wittemann A.. "NorDiaChange: Diachronic semantic change dataset for Norwegian", *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, LREC 2022 (Marseille, France, June 20–25, 2022), ELRA, 2022, ISBN 979-10-95546-72-6, pp. 2563–2572. arXiv 2201.05123 ↑107

[18] Chen J., Chersoni E., Schlechtweg D., Prokic J., Huang C.-R.. "ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection", *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, Singapore, 6 December 2023 (LChange 2023), ACL, 2023, ISBN 978-1-7138-8601-3, pp. 93–99. ↑107

[19] Kutuzov A., Pivovarova L.. "Three-part diachronic semantic change dataset for Russian", *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, LChange 2021 (Online, August 6, 2021), ACL, 2021, ISBN 978-1-954085-60-2, pp. 7–13. arXiv 2106.08294 ↑107

[20] Schlechtweg D., Schulte im Walde S., Eckmann S.. "Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. V. 2: *Short papers*, NAACL-HLT 2018 (New Orleans, Louisiana, USA, June 1–6, 2018), ACL, 2018, ISBN 978-1-948087-29-2, pp. 169–174. arXiv 1804.06517 ↑107

[21] Schlechtweg D., Zamora-Reina F. D., Bravo-Marquez F., Arefyev N.. "Sense through time: diachronic word sense annotations for word sense induction and lexical semantic change detection", *Language Resources and Evaluation*, **59** (2025), pp. 1431–1465. ↑108

[22] Kutuzov A., Giulianelli M.. "UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval@COLING 2020 (Barcelona, online, December 12–13, 2020), ICCL, 2020, ISBN 978-1-952148-31-6, pp. 126–134. arXiv 2005.00050 ↑108

[23] Giulianelli M., Del Tredici M., Fernández R.. "Analysing lexical semantic change with contextualised word representations", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020 (Online, July 5–10, 2020), ACL, 2020, ISBN 978-1-952148-25-5, pp. 3960–3973. arXiv 2004.14118 ↑108

[24] Beck C.. "DiaSense at SemEval-2020 task 1: Modeling sense change via pre-trained BERT embeddings", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval@COLING 2020 (Barcelona, online, December 12–13, 2020), ICCL, 2020, ISBN 978-1-952148-31-6, pp. 50–58. ↑108

[25] Laicher S., Kurtyigit S., Schlechtweg D., Kuhn J., Schulte im Walde S.. "Explaining and improving BERT performance on lexical semantic change detection", *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL 2021 (Online, April 19–23, 2021), ACL, 2021, ISBN 978-1-954085-04-6, pp. 192–202. arXiv 2103.07259 ↑108

[26] Devlin J., Chang M., Lee K., Toutanova K.. "BERT: pre-training of deep bidirectional transformers for language understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. V. 1: *Long and Short Papers*, NAACL-HLT 2019 (Minneapolis, MN, USA, June 2–7, 2019), ACL, 2019, ISBN 978-1-950737-13-0, pp. 4171–4186. arXiv 1810.04805 ↑108

[27] Giulianelli M., Kutuzov A., Pivovarova L.. "Do not fire the linguist: Grammatical profiles help language models detect semantic change", *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022), ACL, 2022, ISBN 978-1-955917-42-1, pp. 54–67. ↑108

[28] Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzman F., Grave E., Ott M., Zettlemoyer L., Stoyanov V.. "Unsupervised cross-lingual representation learning at scale", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020 (Online, July 5–10, 2020), ACL, 2020, ISBN 978-1-952148-25-5, pp. 8440–8451. arXiv 1911.02116 ↑108

[29] Ren Z., Caputo A., Jones G.. “A few-shot learning approach for lexical semantic change detection using GPT-4”, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, LChange@ACL 2024 (Bangkok, Thailand, August 15, 2024), ACL, 2024, ISBN 979-8-89176-138-4, pp. 187–192. ↑108

[30] Hagen T.. “Lexical semantic change annotation with large language models”, *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, LaTeCH-CLfL 2025 (Albuquerque, New Mexico, USA, May 4, 2025), ACL, 2025, ISBN 979-8-3313-1921-2, pp. 172–178. ↑108, 114

[31] Rand W. M.. “Objective criteria for the evaluation of clustering methods”, *Journal of the American Statistical Association*, **66**:336 (1971), pp. 846–850. ↑109

[32] Homskiy D., Arefyev N.. DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022), ACL, 2022, ISBN 978-1-955917-42-1, pp. 173–179. ↑110, 114, 119

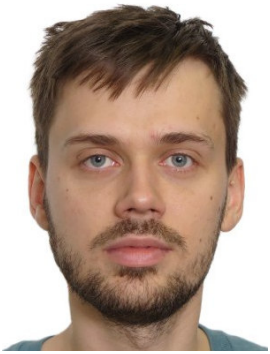
[33] Fedorova M., Kutuzov A., Scherrer Y.. “Definition generation for lexical semantic change detection”, *Findings of the Association for Computational Linguistics*, ACL 2024 (Bangkok, Thailand and virtual meeting, August 11–16, 2024), ACL, 2024, ISBN 979-8-89176-099-8, pp. 5712–5724. arXiv 2406.14167 ↑114

[34] Rachinskiy M., Arefyev N.. “GlossReader at LSCDiscovery: Train to select a proper gloss in English — discover lexical semantic change in Spanish”, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022), ACL, 2022, ISBN 978-1-955917-42-1, pp. 198–203. ↑119

Received 18.02.2026;
approved after reviewing 23.04.2026;
accepted for publication 22.05.2026;
published online 07.06.2026.

Recommended by *PhD. A. N. Vinogradov*

Information about the authors:



Denis Vladislavovich Kokosinskii

Phd Student. Faculty of Mathematics and Cybernetics. Research Interests: Lexical Semantics

0000-0001-5642-8725

e-mail: *deniskokossskappa@gmail.com*

Dominik Schlechtweg

Postdoc. Institute for Natural Language Processing (IMS). Lexical Semantic Change Detection, Computational Semantic Modeling. Research Interests: Graded Lexical Semantic Change Detection, Datasets for computer linguistics

0000-0002-0685-2576

e-mail: *dominik.schlechtweg@ims.uni-stuttgart.de*

Nikolay Viktorovich Arefyev

Postdoctoral Research Fellow in the Language Technology Group at the Department of Informatics. Research Interests: Neural networks in lexical semantics, natural language processing, machine learning

0000-0003-1867-9405

e-mail: *nikolare@uio.no*

The authors contributed equally to this article.

The authors declare no conflicts of interests.

RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.8:81'322:81-115
10.25209/2079-3316-2026-17-2-103-146

Выбор и настройка гранулярности значений в задаче выявления лексико-семантических изменений

Денис Владиславович **Кокосинский**[1], Доминик **Шлехтвег**[2], Николай Викторович **Арефьев**[3]

[1] Московский государственный университет имени М. В. Ломоносова, Москва, Россия
[2] Университет Штутгарта, Штутгарт, Германия
[3] Университет Осло, Осло, Норвегия
[1]✉ *deniskokossskappa@gmail.com*

Аннотация. Изучение того, как меняются значения слов, является давней задачей лингвистики. Мы представляем автоматический подход, который группирует употребления слова в кластеры, соответствующие его значениям, и измеряет, как частота употребления этих значений меняется между историческими периодами. Метод следует устоявшимся процедурам, используемым для создания современных вручную аннотированных языковых ресурсов (Diachronic Word Usage Graphs), и позволяет пользователям настраивать степень укрупнения или детализации значений. Мы также вводим новую метрику, специально адаптированную для диахронических графов словоупотреблений и позволяющую надежно оценивать качество кластеров.

На данных нескольких языков предложенный подход показывает результаты на уровне существующих альтернатив, а часто и лучше их, сохраняя при этом ясные и интерпретируемые выходные данные, которые показывают, какие значения слова вносят вклад в семантическое изменение. (*Связанные тексты статьи на английском и на русском языках*)

Ключевые слова и фразы: выявление лексико-семантических изменений, диахронические графы словоупотреблений, индукция значений слов

Для цитирования: Кокосинский Д. В., Шлехтвег Д., Арефьев Н. В. *Выбор и настройка гранулярности значений в задаче выявления лексико-семантических изменений* // Программные системы: теория и приложения. 2026. **Т. 17**. № 2(71). С. 103–146. (*Англ.+русс.*) https://psta.psiras.ru/read/psta2026_2_103-146.pdf

Введение

Выявление лексико-семантических изменений (Lexical Semantic Change Detection, LSCD; см. [1–3]) состоит в автоматическом определении слов, значение которых меняется со временем. Schlechtweg et al. [4] формулируют задачу LSCD следующим образом: дана совокупность *целевых слов* и два набора *словоупотреблений* из двух временных периодов; требуется ранжировать слова по величине их семантического изменения.

Наборы данных для LSCD обычно аннотируются в рамках Diachronic Word Usage Graph (DWUG) [5], как показано на рисунке 1.

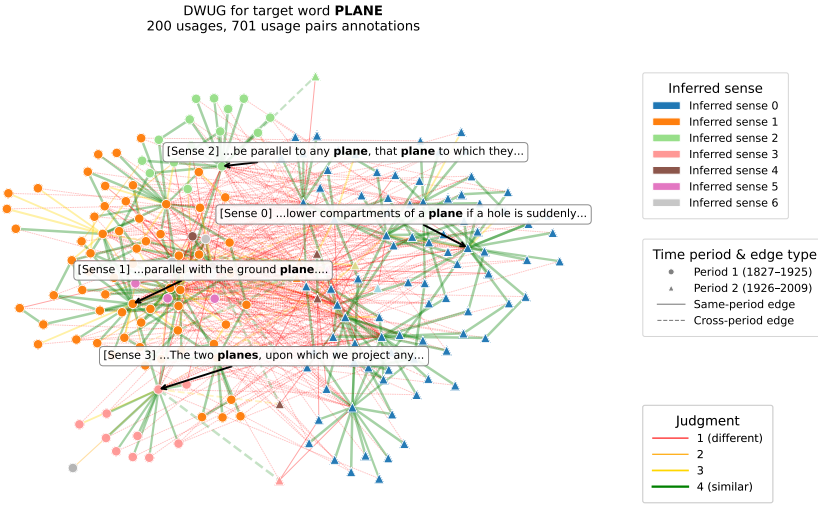


Рисунок 1. Диахронический граф словоупотреблений для целевого слова *plane* в двух временных периодах. Узлы соответствуют словоупотреблениям, ребра – экспертным оценкам семантической близости в парах словоупотреблений.

Предположим, что задано целевое слово (например, *plane*) и два интересующих временных периода (например, XIX и XX века). На предварительном этапе из двух корпусов реальных текстов, охватывающих нужные периоды, собираются примеры предложений или абзацев с целевым словом; такие примеры называются *словоупотреблениями*. Затем эксперты оценивают, насколько близки значения целевого слова в отдельных парах употреблений. Полученный набор попарных аннотаций представляется в виде графа, называемого диахроническим графом словоупотреблений: узлы графа соответствуют словоупотреблениям, а ребра – аннотациям.

На следующем шаге полученный граф кластеризуется, чтобы выделить отдельные значения слова. Наконец, семантическое изменение определяется как изменение размеров кластеров/значений. В более

формальных терминах: чем больше расстояние Дженсена–Шеннона между распределениями частот значений слова в двух временных периодах, тем сильнее семантическое изменение.

Полностью автоматические системы LSCD различаются тем, как они моделируют семантическое изменение. Одни опираются только на попарную семантическую близость между употреблениями, тогда как другие получают полностью автоматически выведенные значения с помощью кластеризации. Тем самым системы LSCD внутри себя решают и другие хорошо известные задачи лексической семантики: Word-in-Context и Word Sense Induction.

Word-in-Context (WiC) – это задача оценки семантической близости в паре словоупотреблений. Например, пара употреблений «...parallel with the ground *plane*...» и «...lower compartments of a *plane*...» имеет низкую семантическую близость, а пара «...parallel with the ground *plane*...» и «...the two *planes*, upon which we project...» – высокую. Некоторые системы LSCD также внутри себя решают задачу индукции значений слов (Word Sense Induction, WSI). WSI – это задача кластеризации: по набору словоупотреблений некоторого целевого слова нужно сгруппировать эти употребления в дискретные кластеры, соответствующие значениям слова.

Современные системы LSCD можно в широком смысле разделить на WSI-основанные и WiC-only-системы. WSI-основанные системы LSCD обладают преимуществом интерпретируемости, поскольку оценивают семантическое изменение непосредственно по изменениям выведенных значений, тогда как внутренний механизм WiC-only-систем менее интерпретируем. WiC-only-системы достигли результатов уровня state-of-the-art [6–8], превосходя WSI-основанные альтернативы [9–11].

В области сложилась противоречивая ситуация: WSI-основанные системы LSCD, которые выполняют кластеризацию и поэтому ближе воспроизводят процедуру аннотирования DWUG, уступают WiC-only-системам, реализующим совершенно иной подход к моделированию семантического изменения [2, 12]. В этой статье мы стремимся устранить это противоречие и представить полностью автоматическую систему LSCD, которая близко следует устоявшейся схеме DWUG и дает ясный, интерпретируемый способ автоматического моделирования семантического изменения.

Итак, два основных вклада нашей работы таковы:

(*1*) Система LSCD, которая явно выводит значения слов и достигает качества уровня state-of-the-art. Наша система поддерживает управляемую гранулярность значений, что важно, поскольку границы между значениями часто размыты [13].

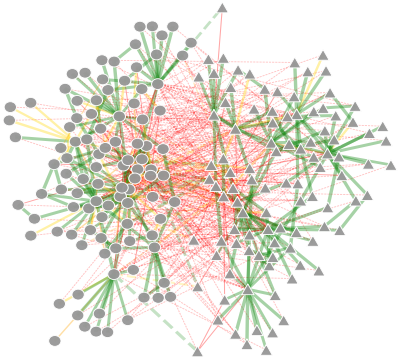
(*2*) Новая метрика AnnotRI для задачи WSI, лучше подходящая для DWUG. Она использует в качестве эталонных меток только надежные вручную аннотированные данные и позволяет оценивать качество при разных уровнях гранулярности значений слов.

1. Связанные работы

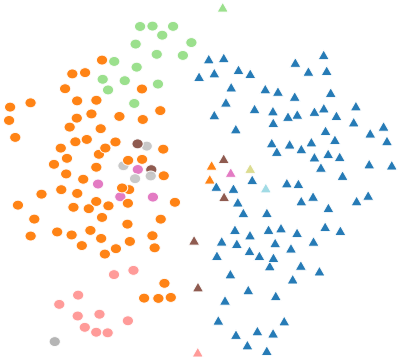
1.1. Диахронические графы словоупотреблений

Сначала подробнее опишем, как семантическое изменение количественно оценивается в рамках DWUG. Это трехэтапная процедура, показанная на рисунке 2.

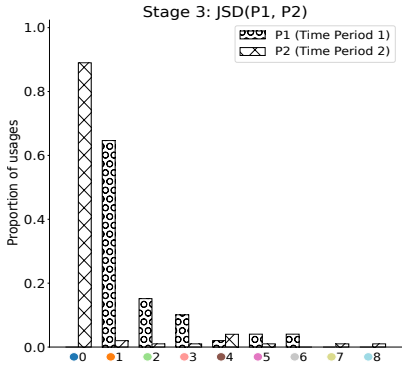
(*а*) Stage 0: Использование за два временных периода



(*б*) Stage 1: Word-in-Context



(*в*) Stage 2: Word Sense Induction



(*г*) Stage 3: Оценка семантического изменения JSD(P1, P2)

Рисунок 2. Этапы DWUG-процедуры для количественной оценки семантического изменения

На первом этапе (рисунок 2*б*) участвуют эксперты: они оценивают семантическую близость пар употреблений, выбранных из двух временных периодов, по четырехбалльной шкале, отражающей степени семантической близости по Blank: омонимия, полисемия, контекстная вариативность и тождество [14].

По сути, это задача Word-in-Context, решаемая экспертами. Употребления становятся узлами, а экспертные суждения задают веса ребер. На втором этапе (рисунок 2*б*) употребления из обоих периодов совместно кластеризуются с помощью корреляционной кластеризации [**15**]; полученные кластеры интерпретируются как значения слова.

Таким образом, в рамках DWUG задача Word Sense Induction решается полуавтоматически: ручные экспертные оценки кластеризуются автоматическим методом. На третьем, заключительном этапе (рисунок 2*в*) истинная оценка семантического изменения для каждого целевого слова вычисляется как расстояние Дженсена–Шеннона между распределениями значений по двум периодам.

Поскольку используется расстояние Дженсена–Шеннона, далее мы называем итоговую оценку семантического изменения JSD (рисунок 2*г*).

Итак, существующие DWUG включают вручную аннотированные пары употреблений, *серебряные*[1] *выведенные значения*, полученные кластеризацией аннотированного графа, и итоговую оценку семантического изменения для каждого целевого слова. Первоначально DWUG были созданы [**16**] для английского (en), немецкого (de) и шведского (sv), а затем для испанского (es; [**12**]), норвежского (три временных периода, два набора данных: DWUG-no12, DWUG-no23; [**17**]) и китайского (zh; [**18**])[2].

Kutuzov and Pivovarova [**19**] используют альтернативную оценку семантического изменения COMPARE [**20**] для русских наборов DWUG-ru12, DWUG-ru23 и DWUG-ru13. Для каждого целевого слова COMPARE определяется как среднее экспертных оценок в группе «compare», состоящей из всех пар употреблений, где первое употребление взято из более старого корпуса, а второе – из более нового (пунктирные линии на рисунках 1 и 2). COMPARE предполагает, что преимущественно низкие оценки в этой группе указывают на значительное семантическое изменение. Эта процедура не включает кластеризацию и поэтому не порождает выведенных значений. Иными словами, COMPARE включает только подзадачу Word-in-Context, но не Word Sense Induction. Промежуточные этапы процедуры аннотирования COMPARE поэтому менее интерпретируемы, чем в JSD: они не показывают, какие значения появились, исчезли или изменили частоту со временем. Однако у JSD есть собственный недостаток: эта оценка опирается на автоматическую кластеризацию, которая может вносить дополнительные ошибки. Несмотря на эти различия, было показано, что COMPARE и JSD сильно коррелируют [**5**].

[1] Мы используем слово «серебряные», чтобы подчеркнуть полуавтоматический характер выведенных значений в DWUG: ручная попарная аннотация сочетается с автоматической кластеризацией.

[2] Точные версии наборов данных, использованных в нашей оценке, приведены в приложении 8

1.2. Методы выявления лексико-семантических изменений

Автоматические методы выявления лексико-семантических изменений можно в широком смысле разделить на WSI-основанные методы и WiC-only-методы среднего попарного расстояния (Average Pairwise Distance, APD), что приблизительно соответствует процедурам аннотирования JSD и COMPARE. WSI-основанные методы [9–11, 21] используют кластеризацию для неявного вывода значений, тем самым решая WSI как промежуточную задачу. Затем они вычисляют расстояние Дженсена–Шеннона между распределениями кластеров по периодам. В отличие от них, APD-методы [22–24] количественно оценивают семантическое изменение, усредняя предсказанные моделью оценки семантической близости в парах словоупотреблений, аналогично COMPARE.

Как WSI-основанные, так и APD-подходы опираются на WiC-модель, которая выдает оценки сходства для пар употреблений, то есть автоматизирует тот шаг, который в DWUG аннотируется вручную. WiC сам по себе является хорошо изученной задачей лексической семантики, для которой существует ряд автоматических систем. Например, Laicher et al. [25] используют BERT [26], а Giulianelli et al. [27] используют XLM-Roberta [28] как WiC-модели, превосходя нетрансформерные подходы без какого-либо дообучения.

Специализированные WiC-модели, а именно GlossReader [8], DeepMistake [7] и XL-Lexeme [6], дополнительно повышают качество благодаря дообучению под конкретную задачу. В более недавних работах [2, 29, 30] показаны возможности LLM общего назначения в роли WiC-моделей: сообщается об улучшениях на 1–3% относительно указанных специализированных моделей в нескольких языках.

Однако мы оставляем LLM как WiC-модели за рамками данной работы из-за существенно более высокой стоимости инференса[3] по сравнению с меньшими специализированными моделями.

Недавняя унифицированная оценка на DWUG [2] показывает, что APD превосходит WSI-основанные методы на всех DWUG-бенчмарках, включая те, где эталонными метками служат JSD. Это противоречит интуиции, поскольку WSI ближе воспроизводит JSD-процедуру, чем APD. Помимо end-to-end-оценки LSCD, Periti and Tahmasebi [2] отдельно оценивают компоненты WiC и WSI, но их WSI-оценка имеет ограничения:

(*i*) WiC-предсказания выполняются только для ребер графа, которые имели ручную аннотацию, что неявно включает золотую информацию на этапе предсказания (см. раздел 2).

[3] Например, [30] показывают, что Llama-3.3-70b является лучшей LLM для LSCD. Llama-3.3-70b содержит 70 миллиардов параметров, тогда как XL-Lexeme – только 0.35 миллиарда; следовательно, стоимость инференса для Llama примерно в ≈200 раз выше.

(*ii*) Серебряные выведенные значения рассматриваются как золотые метки при WSI-оценке, хотя они не аннотированы людьми явно, а получены автоматической процедурой (см. раздел 2).

(*iii*) Не используется систематический подбор гиперпараметров для корреляционной кластеризации, что ведет к эвристическим неоптимальным выборам и может занижать качество сильных моделей (см. раздел 4.1).

2. Оценка Word Sense Induction на DWUG

Мы стремимся построить WSI-основанную систему, поэтому нам нужен способ надежно оценивать качество WSI в нашей диахронической постановке. Одна из ключевых проблем использования DWUG для WSI – надежность золотой кластеризации. Выведенные значения в DWUG не аннотируются вручную, а создаются с помощью полуавтоматической процедуры.

В некоторых DWUG надежность дополнительно снижается разреженностью аннотации ребер: аннотирована лишь малая доля пар употреблений (например, ≈10% всех возможных ребер в DWUG-en). Хотя корреляционная кластеризация способна работать с отсутствующими ребрами, такая кластеризация по своей природе менее надежна, чем кластеризация на полностью аннотированном графе.

Чтобы решить проблему надежности серебряных выведенных значений, мы предлагаем альтернативную оценку WSI для DWUG. В дополнение к сравнению предсказанных кластеров с серебряными выведенными значениями с помощью Adjusted Rand Index (ARI), мы вводим AnnotRI (Annotation Rand Index, по аналогии с Rand Index; [**31**]). В отличие от ARI, который сравнивает полные кластеризации, AnnotRI оценивает качество непосредственно на вручную аннотированных парах употреблений. Она измеряет долю пар употреблений, в которых кластеризационное решение системы (один кластер или разные кластеры для данной пары) совпадает с бинаризованной экспертной аннотацией. Иными словами, AnnotRI – это бинарная accuracy WSI-системы на множестве вручную аннотированных пар употреблений. Тем самым AnnotRI обходит выведенные значения и опирается напрямую на ручные суждения. Меняя порог бинаризации экспертных оценок на четырехбалльной шкале (омонимия, полисемия, контекстная вариативность и тождество), AnnotRI может оценивать качество кластеризации при разных гранулярностях значений на одних и тех же золотых данных; мы приводим AnnotRI@1.5, AnnotRI@2.5 и AnnotRI@3.5. Это полезно, поскольку границы между значениями слова часто нечетко определены [**13**].

3. WSI-основанная LSCD-система

Мы предлагаем полностью автоматическую LSCD-систему, следующую DWUG-подходу; далее обозначаем ее CC+JSD (Correlation Clustering+Jensen-Shannon Distance). Сначала специализированная WiC-модель предсказывает попарную семантическую близость для *всех* пар употреблений целевого слова в обоих временных периодах. Важно, что наша система работает с любой WiC-моделью, но требует некоторого подбора специфичных для модели параметров на этапе кластеризации (см. раздел 4.1). В нашей оценке мы включаем специализированные WiC-модели: GlossReader [8], XL-Lexeme [6] и DeepMistake [32][4]. Получив WiC-предсказания для пар употреблений, мы используем их как веса ребер для построения DWUG.

На втором шаге мы кластеризуем DWUG с помощью корреляционной кластеризации [15], далее кратко CC. Формально CC ищет разбиение C множества употреблений U, минимизирующее суммарное несогласие:

$$\mathcal{L}(C,\theta) = \sum_{\substack{(u_i,u_j):\, r_{ij}\geq\theta,\\ c(u_i)\neq c(u_j)}} r_{ij} \quad + \sum_{\substack{(u_i,u_j):\, r_{ij}<\theta,\\ c(u_i)=c(u_j)}} |r_{ij}|, \tag{1}$$

где

r_{ij} – оценка сходства, предсказанная WiC-моделью,
$c(u)$ – назначение употребления кластеру, а
θ – пороговый параметр.

Первое слагаемое штрафует помещение похожих употреблений (с $r_{ij} \geq \theta$) в разные кластеры, а второе штрафует помещение непохожих употреблений в один кластер.

Мы выбрали именно корреляционную кластеризацию, а не другие алгоритмы кластеризации, чтобы как можно ближе воспроизвести DWUG-процедуру аннотирования. Исходные причины, описанные в [16], в основном переносятся и на нашу полностью автоматическую постановку:

(*i*) CC самостоятельно находит оптимальное число кластеров.
(*ii*) Она более устойчива к ошибкам, поскольку использует глобальную информацию о графе: неверное суждение может быть перевешено корректными.

[4] У GlossReader и DeepMistake есть несколько публично доступных чекпойнтов; мы выбираем те, которые показали лучшее качество на ранее проведенных shared tasks. См. приложение 9.

(*iii*) Она напрямую оптимизирует интуитивный критерий — сумму кластерных несогласий, то есть сумму отрицательных весов ребер внутри кластеров и положительных весов ребер между кластерами.

Особое внимание мы уделяем настройке *порогового* параметра в CC (см. раздел 4.1). Порог определяет, какие веса ребер считаются положительными, а какие отрицательными в целевой функции CC, и критически важен для качества кластеров. Его настройка управляет гранулярностью кластеров (более укрупненные или более детальные). При этом наша новая метрика AnnotRI (см. раздел 2) позволяет тонко настраивать гранулярность предсказанных кластеров в соответствии с четырьмя уровнями семантической близости.

Наконец, после кластеризации употреблений наш метод CC+JSD собирает два распределения частот употреблений по кластерам для двух временных периодов. Мы вычисляем расстояние Дженсена–Шеннона между этими двумя распределениями как итоговую оценку семантического изменения для данного целевого слова. Поскольку LSCD-бенчмарки требуют ранжировать список целевых слов, мы назначаем каждому слову оценку семантического изменения и сортируем список по этой оценке.

Ранее Periti and Tahmasebi [2] сообщали об оценке похожей системы в постановке вычислительного аннотирования. Однако в их постановке WiC-модель переаннотирует только вручную размеченные ребра, отсюда название «computational annotation». Иными словами, она пропускает ребра, отсутствовавшие в исходной DWUG-аннотации, что фактически приводит к утечке структуры DWUG на этапе предсказания. Мы, напротив, используем WiC-предсказания для всех возможных ребер и тем самым рассматриваем более реалистичную постановку для автоматических LSCD-систем.

4. Результаты оценки

4.1. Word Sense Induction

Сначала мы изучаем, как CC-порог влияет на результаты WSI на DWUG-бенчмарках. Мы используем специализированные WiC-модели как основу: XL-Lexeme [6], GlossReader [8] и DeepMistake [7]. Для каждой модели мы собираем распределение оценок сходства по всем DWUG[5] и

[5] Диапазоны оценок зависят от WiC-модели. Поэтому для данной модели мы используем фиксированную сетку по всем DWUG, чтобы обобщаться на новые наборы

используем десять квантилей этого распределения как кандидаты для CC-порога. Мы приводим ARI (относительно серебряных выведенных значений), а также AnnotRI@1.5, AnnotRI@2.5 и AnnotRI@3.5. Результаты представлены на рисунке 3.

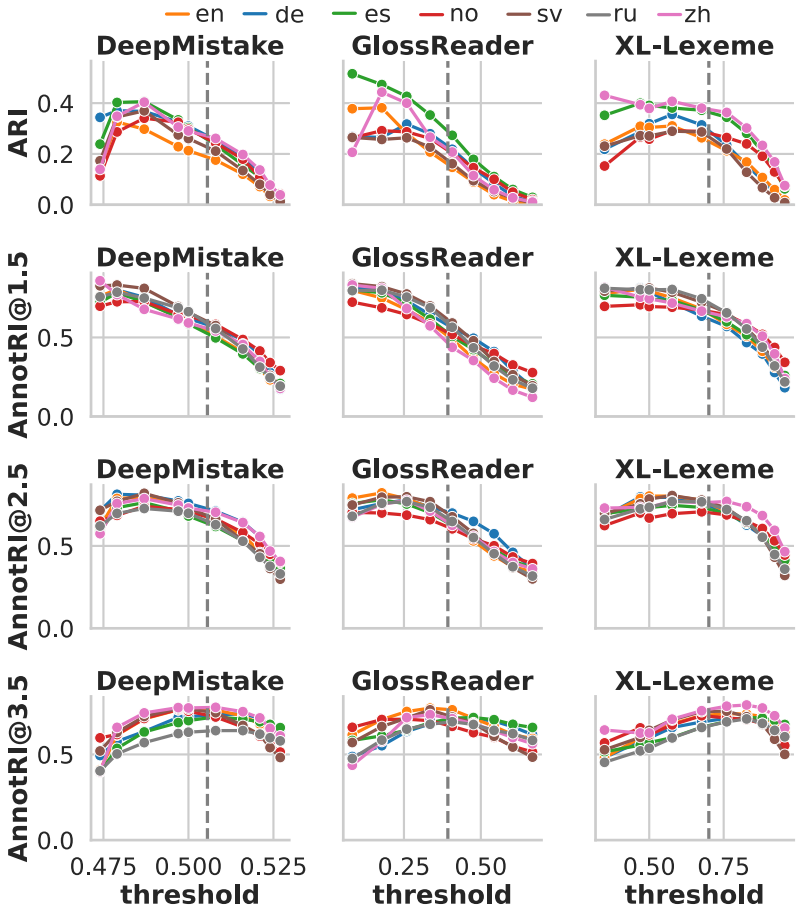


Рисунок 3. Результаты WSI для корреляционной кластеризации в зависимости от порога; серая вертикальная линия обозначает среднее значение оценок сходства по всем бенчмаркам для каждой модели, это CC-порог, использованный в [2]

данных и языки.

Мы наблюдаем устойчивые тенденции по бенчмаркам для каждой модели. Обычно оптимальный CC-порог растет вместе с порогом AnnotRI: квантили сходства 0.1–0.2 дают наилучший AnnotRI@1.5, квантили 0.2–0.3 оптимальны для AnnotRI@2.5, а квантили 0.4–0.5 оптимальны для AnnotRI@3.5. В частности, установка порога на среднем расстоянии (как в [2]) дает относительно мелкозернистые кластеры, которые лучше всего соответствуют AnnotRI@3.5. Следовательно, этот порог не оптимален для ARI. Как и ожидалось, ARI лучше согласуется с AnnotRI@2.5, поскольку порог 2.5 использовался при создании серебряных выведенных значений, рассматриваемых как золотые метки при расчете ARI.

Для нашей end-to-end-оценки LSCD мы задаем порог на основе AnnotRI@2.5. Более конкретно, чтобы выбрать CC-порог без утечки данных, мы выполняем 10-кратную кросс-валидацию по бенчмаркам (DWUG-en, DWUG-es и т. д.): для каждого удержанного бенчмарка выбираем порог, максимизирующий средний AnnotRI@2.5 на оставшихся девяти[6]. Мы также проверяем ту же постановку с ARI вместо AnnotRI@2.5 для DeepMistake (см. приложение 10) и обнаруживаем, что оптимизация AnnotRI@2.5 полезна для итоговой LSCD-метрики на шести из семи JSD-основанных бенчмарков, что подтверждает полезность AnnotRI.

4.2. End-to-end-оценка LSCD

Подробно изучив WSI-подшаг, мы возвращаемся к оценке выявления лексико-семантических изменений для нашей системы CC+JSD (Correlation Clustering+Jensen-Shannon Distance). WiC-модель является параметром этой системы. CC+JSD сначала собирает WiC-предсказания для всех пар употреблений, чтобы автоматически построить DWUG, затем запускает корреляционную кластеризацию для вывода значений и, наконец, вычисляет кластерное расстояние Дженсена–Шеннона между двумя временными периодами, получая единую оценку семантического изменения для целевого слова. Для выбранной WiC-модели (XL-Lexeme, DeepMistake или GlossReader) мы используем кросс-валидацию CC-порога с AnnotRI@2.5 в качестве целевой метрики, как описано в разделе 4.1.

Мы оцениваем LSCD-системы на DWUG-бенчмарках стандартной LSCD-метрикой: ранговой корреляцией Спирмена между предсказанными и серебряными оценками семантического изменения для заданного списка целевых слов. Мы также включаем лучшие WSI-основанные системы из

[6] Это дает несмещенную оценку для нового языка/набора данных за пределами тестовых бенчмарков.

[2]: XL-Lexeme Affinity Propagation (AP+JSD) и What-is-Done-is-Done (WiDiD). Кроме того, мы приводим результаты APD с основами XL-Lexeme, DeepMistake и GlossReader в двух вариантах: APD_{all}, который следует [2] и вычисляет APD по всем парам употреблений целевого слова, и APD_{cmp}, который следует [32] и вычисляет APD только внутри группы «compare» (см. раздел 1.1).

Результаты оценки представлены в таблице 1. Инструмент CC+JSD

Таблица 1. Результаты LSCD-оценки на DWUG. Метрика – ранговая корреляция Спирмена. Для русского языка в качестве золотых меток используется COMPARE, для остальных DWUG – JSD. Звездочка (*) обозначает воспроизведенные результаты из [2]. Лучшие результаты для WSI-основанных систем и APD-систем выделены **жирным**, лучший результат в целом – **подчеркиванием**.

WiC-модель	LSCD	en	de	es	no_{12}	no_{23}	sv	zh	AVG
WSI-основанные методы									
XL-Lexeme	AP+JSD*	.49	.50	.12	.30	.38	.12	.31	.32
XL-Lexeme	WiDiD*	.65	.68	.52	.56	.46	.47	.56	.56
XL-Lexeme	CC+JSD	.72	.84	.65	.55	.59	.85	**.77**	.71
DeepMistake	CC+JSD	.81	**.91**	**.67**	**.85**	.61	**.90**	.76	**.79**
GlossReader	CC+JSD	**.891**	.76	.65	.64	**.70**	.83	.68	.74
APD									
XL-Lexeme	APD^*_{all}	**.886**	.84	.66	.66	.64	.81	**.73**	.75
XL-Lexeme	APD_{cmp}	.87	**.84**	.56	.65	**.66**	**.82**	**.73**	**.73**
DeepMistake	APD_{cmp}	.87	**.87**	**.72**	**.83**	.63	.77	.67	.77
GlossReader	APD_{cmp}	.87	.78	.70	.62	.65	.79	.70	.73

WiC-модель	LSCD	ru_{12}	ru_{13}	ru_{23}	AVG
WSI-основанные методы					
XL-Lexeme	AP+JSD*	.11	.13	.05	.10
XL-Lexeme	WiDiD*	.18	.36	.35	.30
XL-Lexeme	CC+JSD	.62	**.76**	.66	.68
DeepMistake	CC+JSD	**.70**	.75	**.81**	**.75**
GlossReader	CC+JSD	.66	**.76**	.74	.72
APD					
XL-Lexeme	APD^*_{all}	.80	**.86**	.82	**.826**
XL-Lexeme	APD_{cmp}	.79	**.85**	.81	.82
DeepMistake	APD_{cmp}	**.83**	**.84**	**.83**	**.833**
GlossReader	APD_{cmp}	.78	.81	.78	.79

выводит на новый уровень на всех JSD-основанных бенчмарках, кроме DWUG-es[7], превосходя как прежние WSI-основанные подходы, так и APD. Мы также видим, что на некоторых JSD-основанных бенчмарках (DWUG-en, DWUG-de, DWIG-sv) ранговая корреляция Спирмена с золотыми оценками семантических изменений достигает .891, то есть приближается к оптимальному качеству.

В целом мы видим, что методы, более близко следующие процессу создания золотых данных, достигают более высокого качества: CC+JSD часто превосходит APD на JSD-основанных бенчмарках, тогда как для COMPARE-основанного DWUG-ru наблюдается обратная картина. Тем самым мы разрешаем противоречие, отмеченное в предыдущих работах: WSI-основанные системы работали хуже APD, хотя ближе следовали процедуре аннотирования для JSD-основанных DWUG-бенчмарков. Мы в основном связываем более слабое качество прежних WSI-основанных подходов с несоответствием между предсказанными кластерами и серебряными выведенными значениями на этапе кластеризации; эту проблему мы рассматривали в разделе 4.1.

5. Интерпретируемость и анализ ошибок

Как упоминалось выше, WSI-основанные методы позволяют отслеживать, как конкретные значения слова меняются со временем. Мы демонстрируем эту интерпретируемость, анализируя результаты GlossReader CC+JSD на DWUG-en.

Мы назначаем описания значений (метки кластеров), просматривая выборку примерно из дюжины употреблений из каждого выведенного значения и выбирая наиболее подходящее описание из Oxford Dictionary для данного слова.[8] Первое, что мы наблюдаем, – большое количество употреблений с неверной кластеризацией. Почти в каждом кластере мы назначаем метку наиболее частого значения (определенного вручную) внутри кластера, игнорируя ошибочные употребления. Такое количество ошибок удивительно с учетом высокой итоговой LSCD-метрики 0.891 в анализируемой постановке. Поэтому мы считаем, что на уровне отдельных употреблений остается значительный простор для улучшения качества.

[7] DWUG-es лежал в основе shared task SemEval 2021, где победил DeepMistake DM-MCL→es+APD. Мы связываем сильные результаты APD на DWUG-es с тщательной task-specific-настройкой DeepMistake-пайплайна.

[8] Для крупномасштабной разметки описаний значений можно использовать автоматические методы; например, [30, 33].

В таблице 2 показано, как распределение употреблений по предсказанным кластерам меняется между двумя временными периодами для трех слов с наибольшей предсказанной степенью семантического изменения.

Таблица 2. Три слова с наибольшей оценкой семантического изменения из DWUG-en по предсказаниям GlossReader CC+JSD. Кластеры с пятью или меньшим числом примеров в любом из временных периодов были отфильтрованы. Для каждого слова всего имеется по 100 употреблений в каждом периоде. Период XIX века включает употребления за 1810–1860 годы, а период XX века – за 1960–2010 годы.

		Временной период		
Слово	**Описание значения**	**XIX**	**XX**	**Δ**
plane	геометрическое	71	23	-48
	летательный аппарат	20	63	+43
graft	политическая взятка	15	67	+52
	сельскохозяйственный прием	41	7	-34
	медицинская трансплантация	34	12	-22
prop	опора (абстрактная)	24	21	-3
	опора (физическая)	29	10	-19
	театральный/киношный реквизит	12	20	+8

Несмотря на ошибки на уровне употреблений, система все же способна выявлять семантические изменения на уровне значений. Например, система фиксирует семантическое изменение для слова *plane*[9], стоящего на первом месте в ранжировании: от преимущественно геометрического термина оно переходит к употреблению для обозначения типа летательного аппарата.

Однако тот же пример показывает и ограничение: 20 употреблений из 1810–1860 годов были ошибочно[10] отнесены к большому кластеру «летательный аппарат», хотя такой концепт в тот период еще не существовал. В результате системе может быть трудно обнаруживать истинное появление или исчезновение значения; иногда она интерпретирует его как сдвиг частоты внутри уже существующего значения.

[9] которое также является словом с наибольшим изменением согласно золотым меткам

[10] мы вручную просмотрели их все

Еще одна проблема – изобилие очень маленьких кластеров в предсказаниях, часто состоящих всего из одного употребления[11]. После ручной проверки мы заключаем, что это преимущественно артефакты кластеризации, а не настоящие редкие значения.

Итак, хотя система демонстрирует хорошее качество LSCD и улавливает общее семантическое изменение, остается пространство для улучшения на уровне значений и тем более на уровне отдельных употреблений.

Мы также анализируем кластеры с разной гранулярностью. Мы берем три слова, для которых предсказанные JSD-оценки при более низком CC-пороге (оптимизированном для AnnotRI@1.5) и более высоком CC-пороге (оптимизированном для AnnotRI@3.5) различаются сильнее всего. Сводка приведена в таблице 3.

Таблица 3. Три слова с наибольшей разницей в предсказанной оценке семантического изменения между CC-порогами, оптимизированными для AnnotRI@1.5 и AnnotRI@3.5. Для каждой леммы всего имеется 200 употреблений. Кластеры с 10 или меньшим числом употреблений исключены.

	CC-порог оптимизирован для			
Лемма	**AnnotRI@1.5**		**AnnotRI@3.5**	
prop	физическое	127	физическое	45
			строительство	31
			театр/кино	31
	абстрактное	65	абстрактное	45
stab	все	187	акт удара ножом	124
			попытка	24
			попытка	10
			боль	13
multitude	все	200	большая группа	101
			разнообразие	98

Наибольшая разница наблюдается для слова *prop*: более низкий порог снижает JSD с 0.72 до 0.12. Фактически все «физические» значения слова сливаются в один кластер при использовании более низкого порога. Мы также видим, что для двух из трех слов более низкий CC-порог дает кластер «все», включающий почти все употребления; это, вероятно, уместно для слов *stab* и *multitude*.

[11] Похожая проблема также отмечалась авторами DWUG при использовании корреляционной кластеризации [16].

При анализе предсказаний с более высоким CC-порогом мы часто не могли различить значения; для слова *stab* два кластера «попытка» содержали употребления с очень похожим значением. В целом, хотя CC-порог, оптимизированный для AnnotRI@3.5, лучше согласуется с Oxford Dictionary, он также часто вводит множество малых кластеров, неотличимых (по крайней мере для нас) от других кластеров. Мы предполагаем, что выбор CC-порога отдельно для каждого слова может решить эту проблему; это остается направлением будущей работы.

6. Заключение

Мы предложили практическую, полностью автоматическую WSI-основанную LSCD-систему, которая следует DWUG-парадигме аннотирования от начала до конца: WiC-сходства для всех пар употреблений, CC с настроенным порогом и JSD по предсказанным значениям. Мы также ввели AnnotRI, которая оценивает WSI напрямую относительно ручных попарных аннотаций и естественно поддерживает разные гранулярности значений. Управляемая AnnotRI тщательная настройка CC-порога дает сильные, часто state-of-the-art-результаты LSCD на JSD-основанных бенчмарках, сохраняя интерпретируемость моделей, основанных на значениях. Наш анализ показывает, что, несмотря на ошибки для отдельных словоупотреблений, WSI дает практически полезные выводы, позволяя отслеживать, какие значения определяют семантическое изменение.

7. Ограничения

- Наши WiC-основы исключают большие LLM из-за стоимости и задержки; результаты могут улучшиться при использовании более сильных, но при этом эффективных моделей.
- CC совместно кластеризует употребления из обоих периодов, что может ограничивать выявление действительно новых или исчезающих значений.
- DWUG в некоторых языках разреженно аннотированы; AnnotRI использует только аннотированные пары и может отражать смещения экспертного покрытия.
- Вычисление WiC-сходств для всех пар квадратично по числу употреблений; масштабирование на более крупные DWUG требует аппроксимации или отсечения.
- Качество зависит от определения золотых меток (JSD или COMPARE); выводы для одного сигнала супервизии могут не переноситься на другой.

8. Версии бенчмарков

DWUG-en 2.0.1, DWUG-de 2.3.0, DWUG-es (LSCDiscovery) 3.0.0, DWUG-no (NorDiaChange) 1.0.1, DWUG-ru (RuShiftEval 2021) 2.0.1, DWUG-sv 2.0.1, DWUG-zh (ChiWUG) 1.0.0

9. Версии моделей

XL-Lexeme [6]. Мы используем косинусное расстояние между ненормализованными эмбеддингами, как рекомендуется в исходной статье.

GlossReader, часть победившего решения [34] на [12]. Мы используем L1-расстояние между L1-нормализованными эмбеддингами, как рекомендуется в исходной статье.

DeepMistake, обозначенный как MCL+RSS+DWUG-es+XL-WSD в [32].

10. Использование AnnotRI и ARI для выбора порога

Таблица 4. LSCD-оценка с использованием ранговой корреляции Спирмена между предсказанными и золотыми оценками семантического изменения для DeepMistake CC+JSD при использовании AnnotRI@2.5 и ARI.

WiC-модель	WSI-метрика	en	de	es	no12	no23	sv	zh
DeepMistake	AnnotRI@2.5	.81	**.91**	**.67**	**.85**	**.61**	**.90**	**.76**
DeepMistake	ARI	**.85**	.71	.57	.69	.59	.87	.74

В таблице 4 приведены LSCD-результаты при использовании AnnotRI@2.5 и ARI как метрик, оптимизируемых в кросс-валидации для выбора CC-порога.

Список использованных источников

[1] S. Montanelli, F. Periti *A survey on contextualised semantic shift detection* // ACM Computing Surveys.– 2024.– Vol. **56**.– No. 11.– id. 282.– 38 pp. doi arXiv 2304.01666 ↑126

[2] F. Periti, N. Tahmasebi *A systematic comparison of contextualized word embeddings for lexical semantic change* // *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.– V. 1: *Long papers*, NAACL 2024 (Mexico City, Mexico, June 16–21, 2024).– ACL.– 2024.– ISBN 979-8-89176-120-9.– Pp. 4262–4282. doi arXiv 2402.12011 ↑126, 127, 130, 133, 134, 135, 136

[3] A. Kutuzov, L. Ovrelid, T. Szymanski, E. Velldal *Diachronic word embeddings and semantic shifts: A survey* // *Proceedings of the 27th International Conference on Computational Linguistics*, COLING-2018 (Santa Fe, New Mexico, USA, August 20–26, 2018).– ACL.– 2018.– ISBN 978-1-5108-6906-6.– Pp. 1384–1397. URL arXiv 1806.03537 doi ↑126

[4] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi *SemEval-2020 task 1: Unsupervised lexical semantic change detection* // *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval2020 (Barcelona, Spain (Online), December 12–13, 2020).– 2020.– ISBN 9781713828389.– Pp. 1–23. URL doi arXiv 2007.11464 ↑126

[5] D. Schlechtweg *Human and computational measurement of lexical semantic change*, Thesis Dr. phil..– Institut f̈ur Maschinelle Sprachverarbeitung der Universiẗat Stuttgart.– 2023.– 227 pp. doi ↑126, 129

[6] P. Cassotti, L. Siciliani, M. DeGemmis, G. Semeraro, P. Basile *XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE* // *Conference: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.– V. 2: *Short papers* (Toronto, Canada, July 9–14, 2023).– ACL.– 2023.– ISBN 978-1-959429-70-8.– Pp. 1577–1585. doi URL ↑127, 130, 132, 133, 141

[7] N. Arefyev, M. Fedoseev, V. Protasov, D. Homskiy, A. Davletov, A. Panchenko *DeepMistake: which senses are hard to distinguish for a word-in-Context model* (Moscow, 2021), Computational Linguistics and Intellectual Technologies.– vol. **20**, Papers from the Annual International Conference "Dialogue".– 2021.– ISBN 978-5-7281-3032-1.– Pp. 16–30. doi URL ↑127, 130, 133

[8] M. Rachinskiy, N. Arefyev *GlossReader at SemEval-2021 task 2: Reading definitions improves contextualized word embeddings* // *Conference: Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval-2021 (Bangkok, Thailand, August 5–6, 2021).– ACL.– 2021.– ISBN 978-1-954085-70-1.– Pp. 756–762. doi URL ↑127, 130, 132, 133

[9] M. Martinc, S. Montariol, E. Zosa, L. Pivovarova *Capturing evolution in word usage: Just add more clusters?Companion Proceedings of the Web Conference 2020*, WWW'20 Companion (Taipei, Taiwan, April 20–24, 2020).– ACM.– 2020.– ISBN 978-1-4503-7024-0.– Pp. 343–349. arXiv 2001.06629 doi URL ↑127, 130

[10] M. Martinc, P. Kralj Novak, S. Pollak *Leveraging contextual embeddings for detecting diachronic semantic shift* // *Proceedings of the Twelfth Language Resources and Evaluation Conference*, LREC 2020 (Marseille, France, May 11–16, 2020).– ELRA.– 2020.– ISBN 979-10-95546-34-4.– Pp. 4811–4819. arXiv 1912.01072 doi ↑127, 130

[11] F. Periti, A. Ferrara, S. Montanelli, M. Ruskov *What is done is done: An incremental approach to semantic shift detection* // *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022).– ACL.– 2022.– ISBN 978-1-955917-42-1.– Pp. 33–43. doi URL URL ↑127, 130

[12] F. D. Zamora-Reina, F. Bravo-Marquez, D. Schlechtweg *LSCDiscovery: A shared task on semantic change discovery and detection in Spanish* // *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022).– ACL.– 2022.– ISBN 978-1-955917-42-1.– Pp. 149–164. arXiv 2205.06691 doi URL ↑127, 129, 141

[13] K. Erk, D. McCarthy, N. Gaylord *Measuring word meaning in context* // Computational Linguistics.– 2013.– Vol. **39**.– No. 3.– Pp. 511–554. doi ↑127, 131

[14] A. Blank *Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change* // *Historical Semantics and Cognition*, eds. A. Blank, P. Koch, Berlin–New York: Mouton de Gruyter.– 1999.– ISBN 9783110804195.– Pp. 61–90. doi ↑128

[15] N. Bansal, A. Blum, S. Chawla *Correlation clustering* // Machine Learning.– 2004.– Vol. **56**.– No. 1.– Pp. 89–113. doi ↑129, 132

[16] D. Schlechtweg, N. Tahmasebi, S. Hengchen, H. Dubossarsky, B. McGillivray *DWUG: A large resource of diachronic word usage graphs in four languages* // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2021 (Punta Cana, Dominican Republic and Online, November 7–11, 2021).– ACL.– 2021.– ISBN 978-1-7138-9162-8.– Pp. 7079–7091. arXiv 2104.08540 % doi doi URL ↑129, 132, 139

[17] A. Kutuzov, S. Touileb, P. Maehlum, T. Enstad, A. Wittemann *NorDiaChange: Diachronic semantic change dataset for Norwegian* // *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, LREC 2022 (Marseille, France, June 20–25, 2022).– ELRA.– 2022.– ISBN 979-10-95546-72-6.– Pp. 2563–2572. arXiv 2201.05123 doi URL ↑129

[18] J. Chen, E. Chersoni, D. Schlechtweg, J. Prokic, C.-R. Huang *ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection* // *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, Singapore, 6 December 2023 (LChange 2023).– ACL.– 2023.– ISBN 978-1-7138-8601-3.– Pp. 93–99. doi URL ↑129

[19] A. Kutuzov, L. Pivovarova *Three-part diachronic semantic change dataset for Russian* // *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, LChange 2021 (Online, August 6, 2021).– ACL.– 2021.– ISBN 978-1-954085-60-2.– Pp. 7–13. arXiv 2106.08294 doi URL ↑129

[20] D. Schlechtweg, S. Schulte im Walde, S. Eckmann *Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change* // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.– V. 2: *Short papers*, NAACL-HLT 2018 (New Orleans, Louisiana, USA, June 1–6, 2018).– ACL.– 2018.– ISBN 978-1-948087-29-2.– Pp. 169–174. URL doi arXiv 1804.06517 ↑129

[21] D. Schlechtweg, F. D. Zamora-Reina, F. Bravo-Marquez, N. Arefyev *Sense through time: diachronic word sense annotations for word sense induction and lexical semantic change detection* // Language Resources and Evaluation.– 2025.– Vol. **59**.– Pp. 1431–1465. ↑130

[22] A. Kutuzov, M. Giulianelli *UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection* // *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval@COLING 2020 (Barcelona, online, December 12–13, 2020).– ICCL.– 2020.– ISBN 978-1-952148-31-6.– Pp. 126–134. arXiv 2005.00050 ↑130

[23] M. Giulianelli, M. Del Tredici, R. Fernández *Analysing lexical semantic change with contextualised word representations* // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020 (Online, July 5–10, 2020).– ACL.– 2020.– ISBN 978-1-952148-25-5.– Pp. 3960–3973. arXiv 2004.14118 ↑130

[24] C. Beck *DiaSense at SemEval-2020 task 1: Modeling sense change via pre-trained BERT embeddings* // *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval@COLING 2020 (Barcelona, online, December 12–13, 2020).– ICCL.– 2020.– ISBN 978-1-952148-31-6.– Pp. 50–58. ↑130

[25] S. Laicher, S. Kurtyigit, D. Schlechtweg, J. Kuhn, S. Schulte im Walde *Explaining and improving BERT performance on lexical semantic change detection* // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL 2021 (Online, April 19–23, 2021).– ACL.– 2021.– ISBN 978-1-954085-04-6.– Pp. 192–202. arXiv 2103.07259 ↑130

[26] J. Devlin, M. Chang, K. Lee, K. Toutanova *BERT: pre-training of deep bidirectional transformers for language understanding* // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.– V. 1: *Long and Short Papers*, NAACL-HLT 2019 (Minneapolis, MN, USA, June 2–7, 2019).– ACL.– 2019.– ISBN 978-1-950737-13-0.– Pp. 4171–4186. arXiv 1810.04805 % ↑130

[27] M. Giulianelli, A. Kutuzov, L. Pivovarova *Do not fire the linguist: Grammatical profiles help language models detect semantic change* // *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022).– ACL.– 2022.– ISBN 978-1-955917-42-1.– Pp. 54–67. ↑130

[28] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov *Unsupervised cross-lingual representation learning at scale* // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020 (Online, July 5–10, 2020).– ACL.– 2020.– ISBN 978-1-952148-25-5.– Pp. 8440–8451. arXiv 1911.02116 ↑130

[29] Z. Ren, A. Caputo, G. Jones *A few-shot learning approach for lexical semantic change detection using GPT-4 // Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, LChange@ACL 2024 (Bangkok, Thailand, August 15, 2024).– ACL.– 2024.– ISBN 979-8-89176-138-4.– Pp. 187–192. ↑130

[30] T. Hagen *Lexical semantic change annotation with large language models // Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, LaTeCH-CLfL 2025 (Albuquerque, New Mexico, USA, May 4, 2025).– ACL.– 2025.– ISBN 979-8-3313-1921-2.– Pp. 172–178. ↑130, 137

[31] W. M. Rand *Objective criteria for the evaluation of clustering methods* // Journal of the American Statistical Association.– 1971.– Vol. **66**.– No. 336.– Pp. 846–850. ↑131

[32] D. Homskiy, N. Arefyev *DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators?Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022).– ACL.– 2022.– ISBN 978-1-955917-42-1.– Pp. 173–179. ↑132, 136, 141

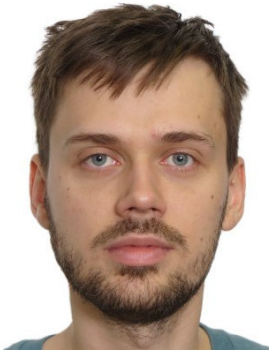
[33] M. Fedorova, A. Kutuzov, Y. Scherrer *Definition generation for lexical semantic change detection // Findings of the Association for Computational Linguistics*, ACL 2024 (Bangkok, Thailand and virtual meeting, August 11–16, 2024).– ACL.– 2024.– ISBN 979-8-89176-099-8.– Pp. 5712–5724. arXiv:2406.14167 ↑137

[34] M. Rachinskiy, N. Arefyev *GlossReader at LSCDiscovery: Train to select a proper gloss in English — discover lexical semantic change in Spanish // Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Dublin, Ireland, May 26–27, 2022).– ACL.– 2022.– ISBN 978-1-955917-42-1.– Pp. 198–203. ↑141

Поступила в редакцию 18.02.2026;
одобрена после рецензирования 23.04.2026;
принята к публикации 22.05.2026;
опубликована онлайн 07.06.2026.

Рекомендовал к публикации *к.ф.-м.н. А. Н. Виноградов*

Информация об авторах:



Денис Владиславович Кокосинский

Аспирант. Факультет вычислительной математики и кибернетики. Научные интерсы: Лексическая семантика

0000-0001-5642-8725

e-mail: deniskokossskappa@gmail.com

Доминик Шлехтвег

Постдок. Институт обработки естественного языка (IMS). Основания компьютерной лингвистики. Научные интересы: Обнаружение лексико-семантического сдвига, Вычислительное моделирование семантики

0000-0002-0685-2576

e-mail: dominik.schlechtweg@ims.uni-stuttgart.de

Николай Викторович Арефьев

Постдок-исследователь в группе языковых технологий (LTG) Департамента информатики. Научные интересы: Обнаружение лексико-семантического сдвига, наборы данных в компьютерной лингвистике

0000-0003-1867-9405

e-mail: nikolare@uio.no

Авторы внесли равный вклад в подготовку публикации.

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*