2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

doi 10.26089/NumMet.v27r329

Automatic systems with word sense induction for lexical semantic change detection

Denis V. Kokosinskii
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics,
Moscow, Russia
ORCID: 0000-0001-5642-8725, e-mail: deniskokossskappa@gmail.com

Nikolay V. Arefyev
University of Oslo, Department of Informatics,
Oslo, Norway
ORCID: 0000-0003-1867-9405, e-mail: nikolare@uio.no

Abstract: Lexical Semantic Change Detection (LSCD) is the task of identifying words that have changed their meaning over time. The semantic change can be formalized using different mathematical models. The Jensen–Shannon Distance model quantifies semantic change at the level of word senses, while the COMPARE model works directly at the level of individual word usages. Automatic LSCD systems reflect this split: some of them explicitly induce word senses by means of clustering, while others operate at the level of individual word usages without resorting to clustering. We argue that the automatic LSCD systems with word sense induction are preferable due to their interpretability, since they answer not only the question of the degree to which a word's meaning has changed, but also the question of which word senses caused the lexical semantic change. The main counter-argument against such systems was their inferior quality. Recent studies found that the usage-level systems outperform the systems with word sense induction, even on benchmarks that themselves rely on induced senses. In this paper, we propose an automatic LSCD system with word sense induction that surpasses the existing usage-level systems in quality. We systematically evaluate each component of the systems on benchmarks for eight languages. It is shown that specialized compact models ($\sim$ 350 million parameters) achieve the quality on par with human annotators in the semantic similarity estimation problem (the Word-in-Context subtask). A new metric for evaluating clustering quality is proposed, which makes it possible to find a more optimal system configuration and thereby eliminate the quality gap. As a result, we build a sense-level LSCD system that combines interpretability, quality, and computational efficiency.

Keywords: lexical semantic change detection, mathematical modeling in linguistics, natural language processing, clustering.

For citation: D. V. Kokosinskii, N. V. Arefyev, "Automatic systems with word sense induction for lexical semantic change detection," Numerical Methods and Programming. **27** (3), 450–469 (2026). doi 10.26089/NumMet.v27r329.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

Автоматические системы с выводом значений слов в задаче обнаружения лексико-семантического сдвига

Д. В. Кокосинский
Московский Государственный Университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики,
Москва, Российская Федерация
ORCID: 0000-0001-5642-8725, e-mail: deniskokossskappa@gmail.com

Н. В. Арефьев
Университет Осло, факультет информатики,
Осло, Норвегия
ORCID: 0000-0003-1867-9405, e-mail: nikolare@uio.no

Аннотация: Задача обнаружения лексико-семантического сдвига заключается в поиске слов, изменивших свое значение со временем. Лексико-семантический сдвиг может быть формализован с помощью различных математических моделей. Модель Jensen–Shannon Distance представляет сдвиг как изменение на уровне значений слова, тогда как модель COMPARE работает на уровне отдельных словоупотреблений. Автоматические системы обнаружения семантического сдвига отражают это разделение: некоторые из них явно моделируют значения слов посредством кластеризации, тогда как другие работают на уровне отдельных словоупотреблений, не прибегая к кластеризации. Мы полагаем, что системы с моделированием значений являются предпочтительными благодаря интерпретируемости, поскольку они отвечают не только на вопрос о степени изменения значения слова, но и на вопрос о том, какие именно значения слова обусловливают лексико-семантический сдвиг. Основным аргументом против систем с моделированием значений было их более низкое качество. Недавние исследования показали, что системы, работающие на уровне словоупотреблений, превосходят системы, моделирующие значения слов, даже на бенчмарках, которые сами основаны на выводе значений. В данной работе предлагается система для обнаружения лексико-семантического сдвига с моделированием значений слов, превосходящая по качеству существующие системы. Систематически оценивается каждый компонент систем обнаружения лексико-семантического сдвига на бенчмарках для восьми языков. Демонстрируется, что специализированные компактные модели ($\sim$ 350 млн параметров) достигают качества, сопоставимого с качеством разметки человека, в задаче оценки семантической близости (также известной как задача “Слово в контексте”). Предлагается новая метрика для оценки качества кластеризации, которая позволяет найти более оптимальную конфигурацию системы обнаружения лексико-семантического сдвига и устранить разрыв в качестве. В результате предлагается моделирующая значения слов система обнаружения лексико-семантического сдвига, которая сочетает интерпретируемость, качество и вычислительную эффективность.

Ключевые слова: обнаружение семантического сдвига, компьютерное моделирование в лингвистике, обработка естественного языка, кластеризация.

Для цитирования: Кокосинский Д.В., Арефьев Н.В. Автоматические системы с выводом значений слов в задаче обнаружения лексико-семантического сдвига // Вычислительные методы и программирование. 2026. **27**, № 3. 450–469. doi 10.26089/NumMet.v27r329.

1. Introduction. Lexical semantic change is the process of word meanings evolving over time [1]. For example, the word *plane* in the 20th century underwent semantic change from primarily being used as a concept in geometry to also becoming a synonym for aircraft. Lexical Semantic Change Detection (LSCD) [2–4] is the task of identifying when semantic change occurs. A well-known example of LSCD is the “Word of the Year” published by, among others, the Oxford Dictionary[1]. To reveal the “Word of the Year”, the lexicographers

[1] https://corp.oup.com/word-of-the-year

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

manually "analyze data and trends to identify new and emerging words and expressions". LSCD is relevant not only in analyzing trends in modern language, but also in studying how the language has changed in the past, i.e. in the field of historical linguistics [1]. A more formal definition of the LSCD task is provided in [5]: given a list of target words and two corpora from different time periods, it is required to rank the words by the degree of semantic change. Notably, the concept of the "degree of semantic change" between two time periods for a particular word is vague and not well defined. Therefore, several distinct mathematical models that quantify semantic change were proposed.

The model that has recently been widely adopted is the Diachronic Word Usage Graph (DWUG) framework [5]. Instead of quantifying semantic change directly for a target word, it defines a multistep procedure. Assume we are given a target word (e.g. *plane*) and two time periods we are interested in (e.g. the 19th and the 20th centuries). In the first 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 called word usages. Next, annotators make judgments about semantic proximity for pairs of word usages. In other words, they assess how close the meanings of the target word are in a particular pair of usages. For instance, the usage pair "... parallel to the ground *plane* ..." and "... the two *planes* onto which we project ..." has high semantic proximity, whereas "... parallel to the ground *plane* ..." and "... lower compartments of a *plane* ..." has low semantic proximity. After collecting the usage pair annotations, they are represented in a graph, called a DWUG. An illustrative example of a DWUG can be found in Figure 1. The nodes of this graph are the target word usages and the edges are the average scores provided by annotators for each particular pair. Further, the resulting graph is clustered in an effort to determine distinct word senses. Finally, the authors define semantic change as a change in the frequency of usage of particular word senses. In more mathematical terms, the larger the Jensen–Shannon distance between the frequency distributions of word senses in the two time periods, the stronger the semantic change. Due to the use of the Jensen–Shannon distance as a measure of semantic change, we call this approach the JSD model of semantic change.

An alternative semantic change formalization is the COMPARE model from the widely used DURel framework [6]. It quantifies semantic change as the average semantic similarity in cross-period usage pairs (pairs consisting of one word usage from an earlier period and one word usage from a later period). Thus, the COMPARE model assumes that the more semantically distant the cross-period usages are, the stronger the semantic change for that word. The full set of cross-period pairs is called the COMPARE group. Notably,

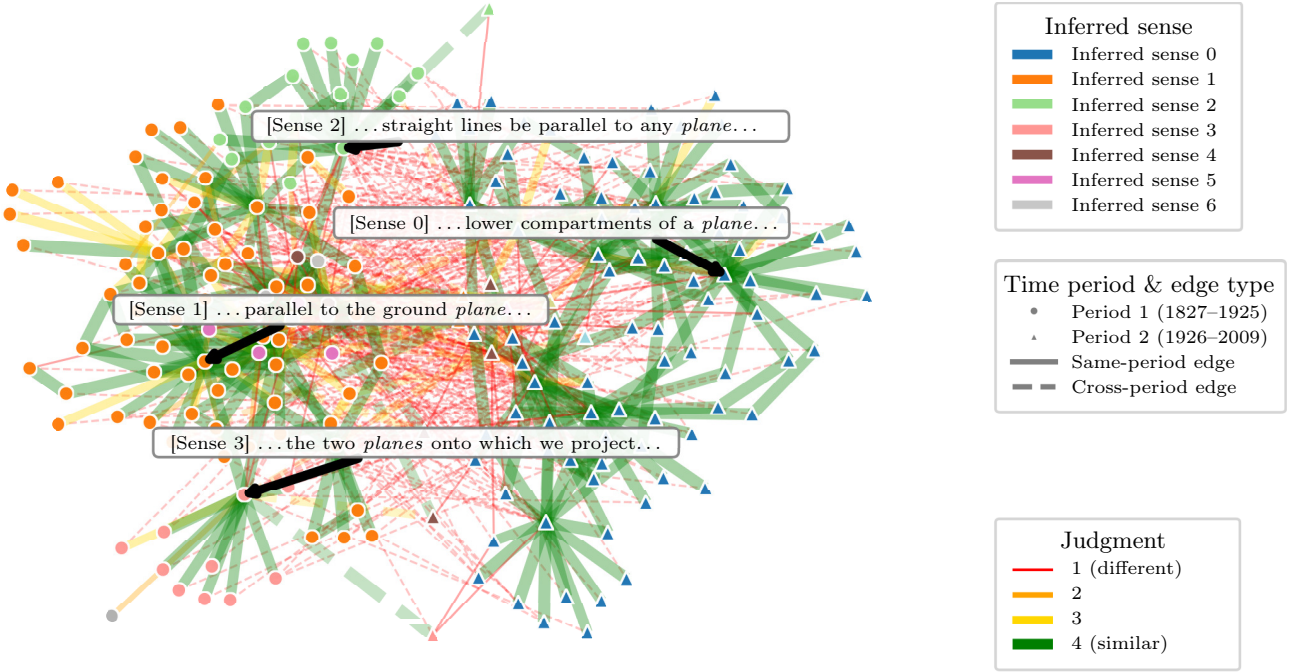

Figure 1. DWUG for the target word *plane* [5], 200 usages, 701 usage pairs annotations. Word usages are sampled from two time periods: 1827–1925 and 1926–2009. Nodes of the graph are word usages, edges are manual judgments of semantic proximity. Senses are inferred by clustering the graph

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

the two models of semantic change do not contradict each other, since significant shifts in sense distributions are often accompanied by low semantic similarity in cross-period usage pairs [6]. Nevertheless, the JSD model answers not only whether semantic change has occurred, but also how the meaning of the word has changed. The JSD model also relies more strongly on linguistic and cognitive studies [5], specifically on Blank's linguistic studies on historic semantic change [1].

Another perspective on the JSD and COMPARE models lies in examining the intermediate tasks associated with them. Both models first solve the Word-in-Context (WiC) task, which is itself a well-established task in lexical semantics [7]. In WiC one needs to estimate semantic similarity in usage pairs. More specifically, considering two examples of how a word is used in some context, one has to determine whether the meanings of the word in these two examples are similar or not. The COMPARE LSCD model aggregates the WiC outputs directly, while JSD also involves solving another well-established task in lexical semantics — Word Sense Induction (WSI) [8]. WSI is essentially a clustering task, where a set of word usages needs to be clustered in accordance with the senses of the target word. The intermediate tasks of the JSD pipeline are illustrated in Figure 2.

Recent automatic systems for LSCD can be divided into two groups, which reflect the distinction between the COMPARE and JSD models. WSI-based systems [9, 10] follow the JSD approach: they first estimate pairwise semantic similarity using a WiC model (the Word-in-Context subtask), then proceed to cluster usages into senses (the Word Sense Induction subtask), and finally compute the Jensen–Shannon distance between the resulting cluster distributions to assign each word a single semantic change score. Average Pairwise Distance (APD) systems [11, 12] implement the COMPARE approach: they use the WiC model to estimate pairwise similarities at the usage level and aggregate them directly without clustering.

To summarize, APD systems and COMPARE-based benchmarks share the underlying COMPARE model of semantic change, estimating semantic change directly from the WiC outputs. On the contrary, WSI-based LSCD systems and JSD-based benchmarks implement the JSD model of semantic change, constructing a DWUG and inducing word senses as their intermediate steps.

APD systems have achieved remarkable success and constitute the current state of the art, outperforming WSI-based systems [2] even on the JSD-based benchmarks. The observation seems counterintuitive, since both the WSI-based systems and the JSD-based benchmarks share the same underlying mathematical model, which one might expect to be advantageous. This raises a central question: why does following the same mathematical model of the underlying phenomena as the dataset creators fail to yield better results?

In this paper, we investigate this apparent contradiction through a separate evaluation of each component in WSI-based systems and APD systems. In short, the main contributions of this work are:

1. We perform a systematic quality evaluation of the WiC and WSI subtasks using reliable intermediate data from the LSCD benchmarks;
2. We show that in the WiC subtask on diachronic datasets, the specialized compact models (neural networks with $\sim$ 350 million parameters) achieve the quality of human annotators, producing results closer to the average annotator judgment than those of most individual annotators;

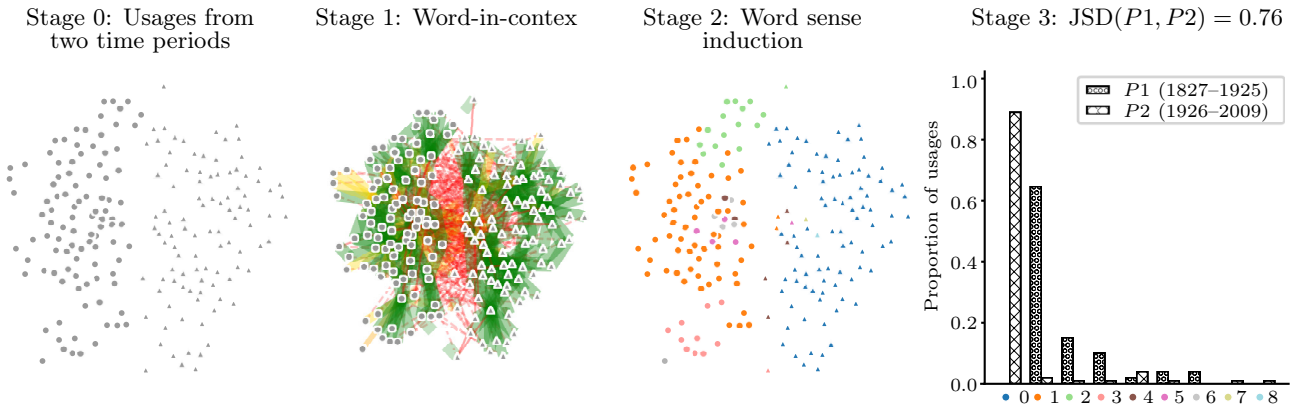

Figure 2. The subtasks involved in the JSD model of semantic change: Word-in-Context, Word Sense Induction and the semantic change score computation. The notations here are the same as in Figure 1

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

3. For the WSI subtask, we propose the Annotation Rand Index (AnnotRI) — a novel clustering metric that is evaluated directly based on manual pairwise annotations rather than potentially unreliable semi-automatically inferred senses. With the help of AnnotRI, we find a clustering configuration that outperforms previous approaches.

As a result of our thorough evaluation, we propose a fully automatic WSI-based LSCD system, which combines interpretability inherent to the JSD model, quality surpassing APD, and computational efficiency.

2. Related work. As was mentioned above, LSCD is an established task in lexical semantics. There is a wide variety of existing datasets, benchmarks, and automatic systems. In this section we describe them in more detail.

2.1. Datasets. Our experiments rely on existing LSCD benchmarks. They can be divided into two types depending on their underlying mathematical model: JSD-based [5, 13–15] and COMPARE-based [16]. In both cases, the DURel annotation framework [6] is used as the first step in creating the gold data (the reference human annotations that are treated as the ground truth when evaluating automatic systems) for the benchmark. In this framework, annotators provide semantic proximity judgments on a four-level scale for pairs of target word usages sampled from two corpora. These four levels are designed to capture varying degrees of semantic proximity: score 1 corresponds to "Unrelated," 2 — to "Distantly Related," 3 — to "Closely Related," and 4 — to "Identical". For instance, the annotation for the target word *plane* in the pair of usages "... parallel to the ground *plane* ..." and "... the two *planes* onto which we project ..." is "Identical", while the pair "... parallel to the ground *plane* ..." and "... lower compartments of a *plane* ..." is "Unrelated", and the pair "the bead *plane*, for sticking beads" and "... parallel to the ground *plane* ..." is "Linked". The source for the word usages in the case of DWUGs is a diachronic corpus, which contains texts from two non-overlapping periods of time. The usages in a pair may belong to the same period or to two different ones. In the latter case, the pairs are referred to as cross-period usage pairs.

Although the JSD-based and the COMPARE-based benchmarks share the initial step, then they differ. The difference lies in how they introduce a scalar measure of semantic change from the DURel human usage pair annotations. In the case of JSD-based benchmarks, the usages and annotations are represented as nodes and weighted edges in a DWUG. The DWUG is clustered using correlation clustering [17, 18], and the resulting clusters are interpreted as word senses (inferred senses). The degree of semantic change for a target word is then quantified as the Jensen–Shannon distance between the corresponding frequency distributions of word usages across the inferred senses in the two time periods. Let P_1 and P_2 denote these frequency distributions in the first and second time periods. The JSD score is defined as

$$\mathrm{JSD}(P_1, P_2) = \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_1 \| M) + \frac{1}{2} D_{\mathrm{KL}}(P_2 \| M)},$$

where $M = \frac{1}{2}(P_1 + P_2)$ is the mixture distribution and $D_{\mathrm{KL}}(P\|Q) = \sum\limits_i P_i \ln \frac{P_i}{Q_i}$ is the Kullback–Leibler divergence.

The COMPARE-based benchmarks offer a simpler alternative that does not require clustering. Semantic change is measured directly as the average semantic similarity over all cross-period usage pairs. Let $E_c = \{(u_i, u_j) : u_i \in T_1, u_j \in T_2\}$ stand for the set of cross-period usage pairs, where T_1 and T_2 are the sets of usages from the first and second time periods. The COMPARE score is defined as

$$\mathrm{COMPARE} = \frac{1}{|E_c|} \sum_{(u_i,u_j)\in E_c} \bar{s}_{ij},$$

where the summation runs over all pairs $(u_i, u_j) \in E_c$, $|E_c|$ denotes the number of such pairs and $\bar{s}_{ij}$ is the average of the semantic proximity judgments assigned to the pair (u_i, u_j) by the annotators, i.e. $\bar{s}_{ij} = \frac{1}{n_{ij}} \sum\limits_{k=1}^{n_{ij}} s_{ij}^{(k)}$. Here $s_{ij}^{(k)}$ is the judgment of the k-th of the n_{ij} annotators who annotated this pair. In the COMPARE model, low values of $\bar{s}_{ij}$ are assumed to indicate the word has undergone a substantial semantic change.

JSD-based benchmarks were created for English (hereafter, for brevity, we denote it by en), German (de), and Swedish (sv) [17], Spanish (es) [13], Norwegian (three time periods, two datasets: no12, no23) [14], and Chinese (zh) [15]. On the other hand, three COMPARE-based datasets are available in Russian due to the work

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

[16] (three time periods, three datasets: ru12, ru23, ru13). We use all these benchmarks in our experiments. Other related resources for studying lexical semantic change include the SURel gold standard for incorporating meaning shifts into term extraction [19], the Lexical Semantic Change Discovery task [20], and the RuDSI graph-based word sense induction dataset for Russian [21].

2.2. Evaluation of LSCD systems. Our work extends the existing evaluation of the LSCD systems on the JSD-based and the COMPARE-based datasets conducted in [2]. It was demonstrated that APD systems (realizing the COMPARE model) outperform WSI-based systems (realizing the JSD model) across benchmarks, including the JSD-based benchmarks. This finding is counterintuitive. Indeed, WSI-based systems are a more natural fit for JSD-based benchmarks, since they rely on the same mathematical model of semantic change as the benchmark creators. We aim to resolve this contradiction by systematically evaluating each component in WSI-based systems to identify potential bottlenecks.

In addition to the end-to-end LSCD evaluation, a computational annotation setup was introduced in [2] to separately evaluate the WiC, WSI, and LSCD components of WSI-based systems. We use this setup as a basis for our evaluation. We identify several limitations in the evaluation setup [2] and address them: 1) it fails to compare WiC models directly with the established state-of-the-art models (DeepMistake and GlossReader) from the previous shared tasks, misrepresenting the current state-of-the-art (see Section 3); 2) the WiC evaluation treats the task as a ranking problem, failing to assess models' ability to distinguish the specific semantic variations (see Section 3); 3) the DWUG inferred senses are treated as gold labels (i.e. as the ground-truth sense clustering) in WSI evaluation, though they are not explicitly annotated by humans but are rather inferred using an automated procedure (see Section 4); 4) gold labels are implicitly incorporated at prediction time, as only edges with manual annotations are included in the dataset, which results in an unrealistic prediction scenario (see Section 4); 5) the WSI evaluation does not include a comparison of correlation clustering with alternative approaches and tuning of important hyperparameters (see Section 4).

2.3. Fully automatic LSCD systems. Recent fully automatic LSCD systems [9–12, 22, 23] follow procedures similar to those used to create the benchmarks, but replace human annotator judgments with automatically inferred ones. In other words, an automatic WiC model replaces annotators in estimating semantic similarity in usage pairs. Recent WiC models predominantly leverage transformer encoders, including BERT [24] used in [25] and XLM-RoBERTa [26] employed in [27]. Specialized WiC models, such as GlossReader [28], DeepMistake [29], and XL-Lexeme [30], are built on XLM-RoBERTa [26] with task-specific fine-tuning. These models differ in architecture. For example, GlossReader and XL-Lexeme are dual encoders that produce a fixed-dimensional embedding for each usage independently, requiring $O(n)$ model forward passes for n usages of a target word. Pairwise similarities are then computed via cosine distance (XL-Lexeme) or L1 distance (GlossReader) between embeddings. DeepMistake, in contrast, is a cross-encoder that processes both usages jointly, requiring $O(n^2)$ forward passes. This distinction has a significant impact on the computational cost of the pipeline, which we discuss in Section 8. We provide a more detailed description of these specialized models in Appendix 1.

Despite sharing the first step of the pipeline (WiC), the LSCD systems differ in how they derive the final semantic change score from the similarity estimates. In WSI-based systems [9, 10, 22], usages of a target word are clustered based on WiC similarity scores. These clusters can be interpreted as discrete senses of the target word. The Jensen–Shannon distance between the distributions of usages across predicted clusters in two time periods is the final semantic score for the target word. The interpretability of the underlying JSD model carries over to the automatic system, since they enable one to track the evolution of specific word senses over time. Average Pairwise Distance systems [11, 12, 23] implement the COMPARE model by aggregating WiC similarity scores for cross-period usage pairs directly, without clustering or modeling specific word senses. Fully automatic approaches to semantic change modeling have also been studied in recent shared tasks, such as AXOLOTL-24 [31, 32].

3. Word-in-Context subtask. Our first step toward improving WSI-based LSCD systems is a thorough inspection of the initial pipeline stage, namely the WiC stage. This stage involves estimating semantic similarity in pairs of word usages and is the component shared by both WSI-based and APD systems. We make several additions to the WiC evaluation procedure and test whether the existing WiC models are able to distinguish the degrees of semantic relation. We also evaluate how well the WiC models perform compared to individual human annotators.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

3.1. WiC evaluation setup. WiC evaluation is complicated by the guidelines used for manual annotation of usage pairs. The DURel annotation guidelines [17] use a four-level scale (see Section 2.1). To evaluate WiC models, ranking/ordinal classification metrics [33] or binary classification metrics are typically used. For instance, the first option was chosen in [2], where Spearman's correlation was used to compare the full list of model scores for all usage pairs with the full list of manual annotations. Thus, the models are tested in the task of ranking word usage pairs by their semantic proximity. However, WiC is frequently formulated as a binary classification task [7, 34], and the most widely used WiC training datasets [35–37] therefore employ binary labels. In other words, the training data of the models contains only "Unrelated" and "Identical" labels, rather than the four labels used in the DURel guidelines. Since these models are probabilistic binary classifiers, the predicted probability of the positive class can be used as a similarity score for each usage pair. However, the alignment of this score with the actual degrees of semantic proximity was not the training objective. Therefore, we aim to evaluate how well the WiC models' predictions are consistent with the four-level scale.

We adapt the standard binary classification Average Precision (AP) to evaluate models at intermediate points of the DURel scale.[2] In our WiC setting, each usage pair (u_i, u_j) is associated with a model score r_{ij} (the semantic similarity predicted by the WiC model for this pair) and a binarized gold label $y_{ij}^{(\tau)} = \mathbb{I}[\bar{s}_{ij} \geqslant \tau]$ (the reference label derived from the human annotations), where $\bar{s}_{ij}$ is the mean human judgment on the four-level DURel scale, $\tau \in \{1.5, 2.5, 3.5\}$ is the threshold for annotations, and $\mathbb{I}[\cdot]$ is the indicator function, which equals 1 if the condition in the brackets holds and 0 otherwise. For a fixed τ, AP is computed by sorting all pairs in descending order of r_{ij} and traversing this ranking from top to bottom. For a given rank k, let $P(k)$ stand for the precision among the top-k pairs and let $\Delta R(k)$ denote the corresponding recall increment. Then,

$$\mathrm{AP} = \sum_k P(k)\,\Delta R(k).$$

According to this formulation, AP measures how strongly the model concentrates binarized pairs with $y_{ij}^{(\tau)} = 1$ near the top of the ranking for r_{ij}.

We employ three gold-label binarization thresholds $\tau \in \{1.5, 2.5, 3.5\}$ and denote the resulting metrics as AP@1.5, AP@2.5, and AP@3.5. These thresholds correspond to the boundaries between the four annotation levels (see Section 2.1) and together probe whether WiC systems can distinguish every intermediate level of the scale. For example, AP@3.5 tests how well the model distinguishes "Identical" (4) from the rest (1, 2 and 3). We also include AP_1vs4, defined as the average precision computed only on clear-cut pairs for which all annotators unanimously assigned a score of either 1 or 4.[3] Unlike AP@1.5, AP@2.5, and AP@3.5, this metric avoids ambiguity from intermediate annotations and evaluates models according to the binary classification setting they were trained for.

In our experiments we include the following state-of-the-art WiC models: GlossReader [28], XL-Lexeme [30], and models from the DeepMistake family, which differ in training languages. We evaluate four DeepMistake variants: DM-MCL+ru+es and DM-MCL→es from [38], and DM-MCL→ru and DM-en from [29]. We also include the method from [27] based on XLM-RoBERTa-large [26] to estimate the effect of specialized fine-tuning for lexical similarity. We evaluate WiC models using AP@1.5, AP@2.5, AP@3.5, and AP_1vs4, as well as Spearman rank correlation (Spearman's ρ) following the original setup of [10].

3.2. Results of WiC evaluation. The results of our WiC evaluation for XL-Lexeme, DeepMistake, and GlossReader are presented in Figure 3. A full comparison with XLM-RoBERTa is provided in Appendix 2. Among specialized models, GlossReader and DeepMistake variants generally outperform XL-Lexeme on most benchmarks and metrics, surpassing the previously reported state-of-the-art XL-Lexeme [2] (the only exception is Chinese, where XL-Lexeme remains superior).

The specialized WiC models demonstrate strong cross-lingual transfer. Although GlossReader, originally a multilingual model, is fine-tuned only on English data, it performs well in all languages. Language-specific fine-tuning (DM-MCL→es, DM-MCL→ru) leads to further improvements. DM-MCL→es outperforms other models by +0.03 in Spearman's ρ on Spanish, for DM-MCL→ru the margin is +0.05 on the Russian benchmarks. However, the general effect of specialized WiC fine-tuning compared to the base XLM-RoBERTa-large is more substantial (+0.35 in Spearman's ρ).

[2] We use AP rather than F-measure or accuracy as in previous shared tasks, because the desired precision-recall trade-off is unclear.

[3] This retains from 70% to 90% of pairs, depending on the dataset.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

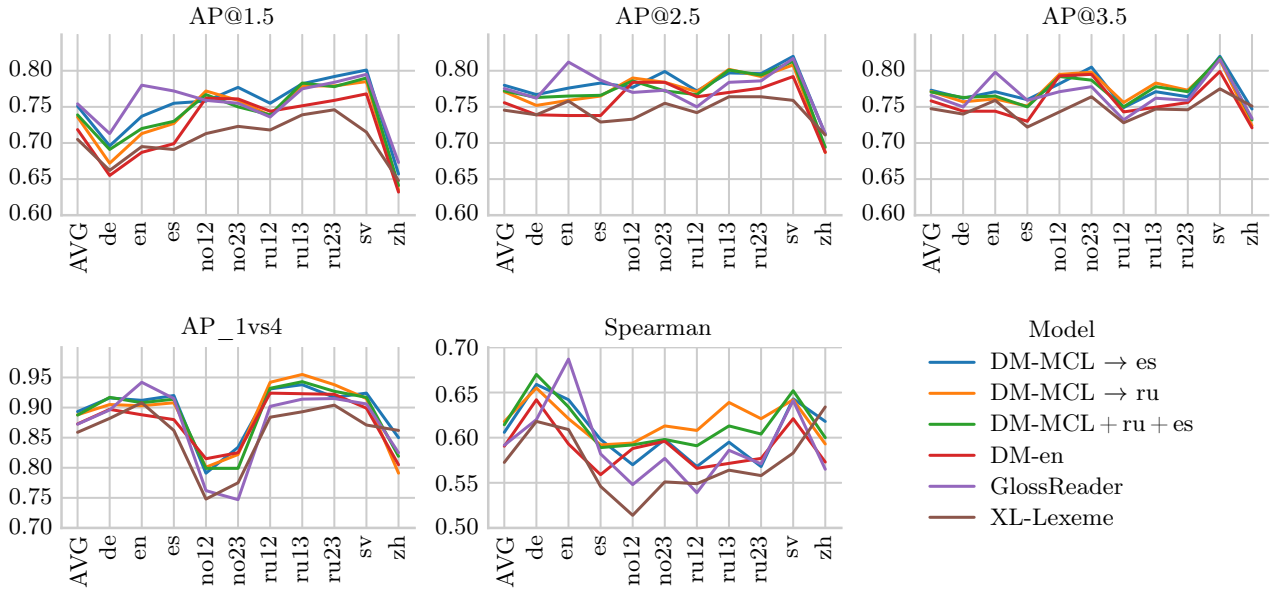

Figure 3. Evaluation results of specialized WiC models. Full results, including non-specialized models, are presented in Appendix 2

Our multi-threshold evaluation reveals that models trained for binary classification can also distinguish more nuanced semantic relations better than the general-purpose XLM-RoBERTa. On average, the models' AP@3.5 and AP@2.5 are similar, while AP@1.5 is slightly lower. This shows that predicted probabilities of the positive class actually align with the four-level annotation scale quite well. Specialized models achieve even higher AP_1vs4 on clear-cut pairs, reaching 0.95 in English and Russian, which shows that there is not much room for improvement for WiC models in the binary classification formulation of WiC. Given these high scores, we hypothesize that WiC models can perform on par with human annotators.

3.3. WiC models as annotators. To test the hypothesis, we conduct a side-by-side comparison of the best-performing models with individual annotators. Let us assume we select some human annotator X[4]. For annotator X, gold labels in our evaluation are defined as the average scores from all other annotators on pairs, where X provided judgments. We can then use AP@1.5, AP@2.5, and AP@3.5 to evaluate how well annotator X agrees with the average annotations. Further, we test the WiC model predictions with the same gold labels and the same metrics. We declare a winner (the model or annotator X), which achieves the higher score. This setup allows us to check whether the WiC model is closer to the average judgment ("the wisdom of the crowd") than an individual annotator. For each benchmark (e.g., en) and metric (e.g., AP@2.5), we repeat this process for every annotator and compute the final win percentage of the WiC model relative to the annotators.

The win percentages for each benchmark are presented in Table 1. The models perform especially well on AP@3.5, where only one annotator in no12 outperforms

Table 1. Side-by-side win percentage of best-performing models against annotators. For example, 37% for DM-MCL+ru+es on de with AP@1.5 means that DM-MCL+ru+es scored higher AP@1.5 than 37% of annotators of de

		Model wins, %		
	Model	AP@1.5	AP@2.5	AP@3.5
de	DM-MCL+ru+es	37	100	100
en	GlossReader	88	100	100
es	DM-MCL→es	100	91	100
no12	DM-MCL→ru	0	33	66
no23	DM-MCL→ru	0	33	100
ru12	DM-MCL→ru	100	100	100
ru13	DM-MCL→ru	100	100	100
ru23	DM-MCL→ru	100	100	100
sv	DM-MCL+ru+es	57	71	100
zh	XL-Lexeme	50	75	100

[4]Datasets ru12, ru13, and ru23 rely on crowdsourcing for annotation [16], where each annotator was assigned only a small subset of pairs, which precludes identification of specific annotators. Therefore, in ru12, ru13, and ru23, annotator X in our comparisons is actually a group of annotators with non-intersecting subsets.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

them (no12 has three annotators, yielding a 66% win rate). Over all three metrics, the models outperform every individual annotator in 16 out of 30 metric-benchmark pairs. In addition, in 28 out of 30 pairs, they outperform at least one annotator. Based on these observations, we conclude that specialized WiC models perform at least on par with individual human annotators.

4. Word Sense Induction subtask. Now we investigate the next stage of the WSI-based system — Word Sense Induction. The clustering step is a component unique to WSI-based systems: it transforms pairwise similarity scores into discrete clusters, which can be interpreted as senses of a word. We propose both a new evaluation metric and an improvement to existing clustering approaches.

4.1. WSI evaluation setup. One of the key challenges in WSI evaluation for DWUG datasets is the reliability of gold labels. Unlike WiC evaluation, where models are compared directly with manual annotations, the inferred senses in DWUGs are not manually annotated but are instead themselves created using an automatic procedure: correlation clustering is performed on the annotation graph.

To address this issue, we propose an alternative WSI evaluation approach for DWUGs. Instead of comparing predicted clusters with inferred senses, we introduce a new metric — Annotation Rand Index (AnnotRI) inspired by the Rand Index [39]). Unlike the Rand Index, which compares complete clusterings, AnnotRI is evaluated only on the subset of usage pairs containing annotations. Formally, let E_a denote the set of usage pairs with manual annotations, U be the set of all usages of the target word, $c : U \to \mathbb{N}$ be the predicted cluster assignment function, and $\bar{s}_{ij}$ be the average annotation for pair (u_i, u_j). For a binarization threshold τ, AnnotRI is defined as

$$\text{AnnotRI@}\tau = \frac{1}{|E_a|} \sum_{(u_i,u_j)\in E_a} \mathbb{I}[\mathbb{I}[c(u_i) = c(u_j)] = \mathbb{I}[\bar{s}_{ij} \geqslant \tau]]\,,$$

where $\mathbb{I}[\cdot]$ is the indicator function. AnnotRI measures the proportion of annotated pairs for which clustering decisions (i.e. whether two usages are assigned to the same or different clusters) agree with binarized human annotation. Thus, AnnotRI does not rely on inferred senses and instead uses manual pairwise annotations directly. While standard clustering metrics assess the agreement between the predicted clustering and gold clusters, AnnotRI evaluates the agreement between the predicted clustering and the underlying pairwise annotation. This possibility is relevant not only for some predicted clustering, but also for the DWUG inferred senses themselves. They employ a potentially lossy automatic clustering of annotations, and AnnotRI allows one to estimate how much noise the clustering process introduces.

By varying the threshold τ, AnnotRI evaluates clustering quality at different levels of sense granularity using the same gold data, which is useful because word-meaning boundaries are often not sharply defined [40]. We set the thresholds at 1.5, 2.5, and 3.5, which correspond to the four grades of the pairwise usage annotation of DWUG. Thus, we use three separate metrics: AnnotRI@1.5, AnnotRI@2.5, and AnnotRI@3.5. One can argue that AnnotRI@2.5 is the most important one for the final end-to-end LSCD evaluation, since it was employed internally in clustering for the JSD-based LSCD benchmarks. The choice of the 2.5 threshold is motivated by historic linguistic studies [5].

To contextualize the AnnotRI results of WSI systems, we employ two random baselines where all usages are assigned random cluster indices. The first baseline (Random Uniform) assumes equal cluster sizes and draws cluster labels from a uniform distribution. For each target word, the number of clusters is assigned uniformly from the range 1 to 9. The second baseline (Random Geom. $p = 0.8$) assumes significantly different cluster sizes and uses a skewed geometric distribution with parameter $p = 0.8$, which more realistically reflects how word senses are typically distributed in language [41].

4.2. Clustering methods. In our WSI evaluation, we base the WSI systems on WiC similarity scores provided by XL-Lexeme, GlossReader and DeepMistake[5]. We employ a total of four clustering approaches: affinity propagation, agglomerative clustering, "correlation clustering on gold pairs" and "correlation clustering with optimized θ". The first two were applied in the existing WSI systems [9, 42]. Following the original WSI system [9], we do not explicitly set any parameters in affinity propagation. For agglomerative clustering, we select the number of clusters based on Silhouette score, as in the original method [42].

We also include the clustering approach used in the computational annotation setup [2]. In this setup, only graph edges with human judgments are provided as input to the clustering algorithm. Therefore, we refer

[5]Here and further we will report joint results for DeepMistake models, where we use DM-MCL→es for Spanish, DM-MCL→ru for Russian and DM-MCL+ru+es for other languages.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

to this approach as "correlation clustering on gold pairs." Although the judgments themselves are replaced with WiC model predictions, filtering graph edges this way effectively leaks gold information. For instance, pairs for annotation were partly selected using gold cluster information in English, German, and Swedish DWUGs [5], e.g., by sampling edges between inferred clusters or increasing the number of annotated edges for words that were more difficult to cluster. Therefore, we argue that "correlation clustering on gold pairs" constitutes an unrealistic evaluation scenario.

We implement an alternative correlation clustering setup that addresses two key issues. First, we eliminate gold-information leakage by feeding all pairwise similarities predicted by the WiC model to the clustering algorithm, not just those corresponding to gold usage pairs. In other words, we build a full $n \times n$ distance matrix for n usages. Second, we tune the threshold parameter θ in correlation clustering based on AnnotRI on DWUGs, unlike "correlation clustering on gold pairs," where θ was simply set to the average of all similarity scores. The threshold parameter is crucial for clustering quality. Formally, correlation clustering [18] aims to find a partition C of the usage set U that minimizes the total disagreement

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

where r_{ij} is the similarity score predicted by the WiC model and $c(u)$ is the cluster assignment. The first term penalizes placing similar usages (with $r_{ij} \geqslant \theta$) into different clusters, and the second penalizes placing dissimilar usages into the same cluster. Adjusting θ allows us to make predicted clusters more coarse-grained or more fine-grained.

We employ a supervised procedure to select θ. To avoid using test data in this process, we perform cross-validation across benchmarks. For a given benchmark (e.g., en), we consider nine quantiles $(0.1, 0.2, \ldots, 0.9)$ of the distribution of r_{ij} over all (u_i, u_j) pairs as candidate θ values, and then select the value that is optimal for the majority of the other benchmarks. As shown in Appendix 3, the optimal θ value transfers well across benchmarks. In 22 out of 30 model-benchmark pairs, the most common θ is the optimal one, which shows that the optimal θ threshold is robust. As a result, our cross-validation procedure selects the same threshold for all benchmarks. To distinguish this modified approach from correlation clustering on gold pairs, we refer to it as "correlation clustering with optimized θ".

4.3. Results of WSI evaluation. The WSI evaluation results averaged over all benchmarks are presented in Figure 4, while the full per-benchmark results are shown in Figure 5 (Appendix 4). Correlation clustering with optimized θ shows superior results across different WiC backbones and metrics. Specifically, when comparing the two correlation clustering setups, we observe a clear benefit from optimizing value θ for the Adjusted Rand Index (ARI), AnnotRI@1.5, and AnnotRI@2.5. At the same time, average AnnotRI@3.5 across all benchmarks is almost identical for the two setups.

As expected, the DWUG inferred senses achieve higher scores on AnnotRI@2.5 than on AnnotRI@1.5 and AnnotRI@3.5, since the threshold of 2.5 was employed in their creation. Average AnnotRI@2.5 across benchmarks for inferred senses is 0.92, which implies that in 8% of manually annotated usage pairs the clustering contradicts the manual annotations. This number can serve as a rough estimate of the amount of noise generated by the automatic clustering.

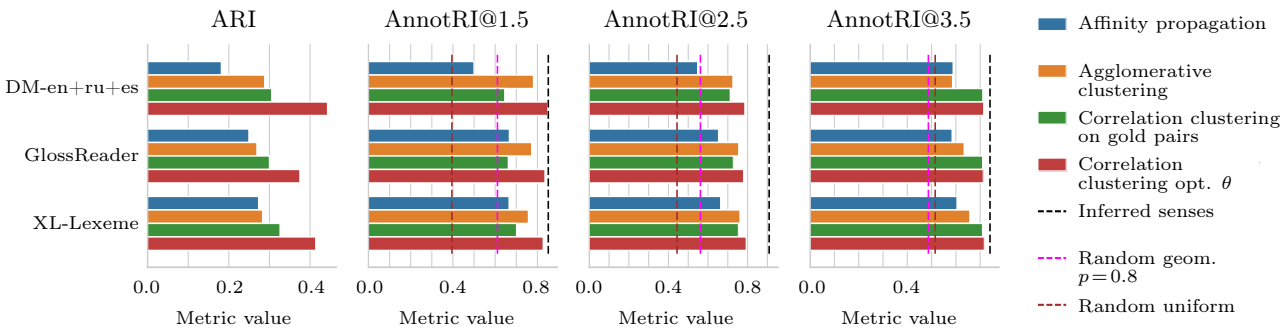

Figure 4. WSI evaluation results averaged over all benchmarks. The full per-benchmark breakdown is presented in Figure 5 (Appendix 4)

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

Despite the introduced noise, the inferred senses still outperform the best automatic system by approximately 0.15 AnnotRI@2.5. Therefore, although pairwise WiC scores agree with human pairwise annotations in about the same way that human annotators agree with one another (Section 3), the clusters obtained from these WiC scores are still less consistent with human pairwise annotations than the clusters derived from the human annotations themselves. Based on this observation, we argue that there is more room for future improvement in clustering than in WiC.

5. End-to-end LSCD evaluation. After analyzing the individual components of LSCD systems (WiC and WSI), we perform the end-to-end LSCD evaluation[6] of full systems. We include the best-performing WiC models: XL-Lexeme, GlossReader, and DeepMistake. Our final WSI-based approach is CC+JSD (Correlation Clustering + Jensen–Shannon distance). It uses correlation clustering with the optimized θ threshold (Section 4), as this approach demonstrates superior WSI results across the three WiC backbones. Following the JSD model of semantic change, we use the Jensen–Shannon distance between the frequency distributions of usages across predicted clusters in the two time periods as a measure of semantic change. We also report the results of APD systems (implementing the COMPARE model) with the same WiC models; the evaluation of APD on gold usage pairs is provided in Appendix 5. The evaluation results are presented in Table 2.

Table 2. LSCD evaluation results. Spearman rank correlation is used as a metric. AP stands for Affinity Propagation, CC for Correlation Clustering with optimized θ, CC(GP) for Correlation Clustering on Gold Pairs. Symbol * denotes the methods from [2]. Evaluation setup ** is flawed, see Section 4

		JSD-based datasets								COMPARE-based datasets			
WiC model	**LSCD**	**en**	**de**	**es**	**no12**	**no23**	**sv**	**zh**	**AVG**	**ru12**	**ru13**	**ru23**	**AVG**
WSI-based methods													
XL-Lexeme	AP+JSD*	0.49	0.50	0.12	0.30	0.38	0.12	0.31	0.32	0.11	0.13	0.05	0.10
XL-Lexeme	WiDiD*	0.65	0.68	0.52	0.56	0.46	0.47	0.56	0.56	0.18	0.36	0.35	0.30
XL-Lexeme	CC(GP)+JSD**	0.80	0.80	0.65	0.62	0.57	0.72	0.71	0.69	-	-	-	-
XL-Lexeme	CC+JSD	0.72	**0.84**	0.65	0.55	0.59	0.85	**0.77**	0.71	0.62	**0.76**	0.66	0.68
DeepMistake	CC+JSD	0.79	**0.84**	**0.72**	0.57	0.67	**0.86**	0.70	**0.74**	**0.66**	0.66	0.64	0.65
GlossReader	CC+JSD	**0.89**	0.76	0.68	**0.64**	**0.70**	0.83	0.68	**0.74**	**0.66**	**0.76**	**0.74**	**0.72**
APD													
XL-Lexeme	APD*	0.83	**0.82**	0.59	0.63	0.65	**0.82**	**0.71**	**0.72**	0.79	**0.85**	0.81	0.82
XL-Lexeme	APD (reproduced)	0.83	**0.82**	0.59	0.63	0.65	**0.82**	**0.71**	**0.72**	0.79	**0.85**	0.81	0.82
DeepMistake	APD	0.75	0.79	**0.67**	**0.64**	**0.70**	0.72	0.69	0.71	**0.83**	**0.85**	**0.83**	**0.84**
GlossReader	APD	**0.85**	0.75	0.66	0.60	0.63	0.79	0.70	0.71	0.76	0.80	0.78	0.78

On the JSD-based benchmark datasets, CC+JSD substantially outperforms previous WSI-based configurations and outperforms APD as well. Specifically, DeepMistake CC+JSD and GlossReader CC+JSD are the best methods on average for DWUG datasets. Overall, XL-Lexeme CC+JSD performs better than XL-Lexeme CC(GP)+JSD (+0.02 in Spearman's ρ), validating the clustering improvement. The use of alternative WiC models (GlossReader and DeepMistake) leads to further quality improvements (+0.03 in Spearman's ρ).

In en, de, es, sv, and zh benchmarks, CC+JSD systems outperform the best APD configurations. These results eliminate the quality gap identified in [2]. When the clustering step is properly configured, WSI-based systems surpass APD on JSD-based benchmarks. In contrast, on COMPARE-based datasets (ru12, ru13, ru23), where gold LSCD scores are derived using the COMPARE model, APD systems substantially outperform WSI-based systems (0.84 vs. 0.72 for the average Spearman correlation coefficient for the best configurations). This yields a symmetric pattern, where each system type achieves its best performance on datasets implementing the same mathematical model. Therefore, we state that the primary counter-argument against WSI-based systems, namely, their weaker end-to-end LSCD results, is no longer valid. At the same time, as noted in Section 1, WSI-based systems are more practically useful because they provide explicit sense-level information.

[6]We make only minor adjustments to the LSCD evaluation setup from [2], namely, contexts exceeding 512 tokens are truncated, but are not filtered out, and we do not filter out usages not assigned to any inferred sense. Unlike the computational annotation setup [2], we also do not provide information about which usage pairs were manually annotated.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

6. Error analysis. To supplement the aggregate Spearman correlation results reported in Table 2, we conduct a qualitative error analysis of our best WSI-based configuration (GlossReader CC+JSD) on the English DWUG benchmark. We rank the target words by the predicted JSD score and compare the top-five predictions with the gold ranking. The words at the top of the ranking are of the highest interest, as they are the most likely candidates for semantic change. The results are presented in Table 3. Four of the five words predicted to change the most (*plane*, *prop*, *chef*, *graft*) belong to the top-five in the gold ranking, and the fifth word (*tip*) is gold rank 6. The missing gold rank 2 is the word *bar*, which has 72 inferred senses in the benchmark. Therefore, we do not treat this case as a failure of the automatic system, but rather an issue with the gold data.

Table 3. Top-5 target words by predicted semantic change (GlossReader CC+JSD on the English DWUG benchmark) compared with the gold ranking

Target word	Predicted rank	Gold rank
plane	1	1
graft	2	5
prop	3	3
chef	4	4
tip	5	6

For each of the five top-predicted words we consulted the Oxford English Dictionary (OED) [43] and the English DWUG data [17] to verify that the predicted change reflects a real diachronic shift. As for the word *plane*, the OED dates the first English attestation of the concept of "airplane" to 1908 (a clipping of "aeroplane"), after the flight of the Wright brothers in 1903. As for the word *graft*, the OED records the "unlawful gain obtained through political influence" sense as American colloquial slang first attested in 1859 and came into mainstream usage in the late 19th century. The theatrical meaning of the word *prop* ("object used on stage", short for "property") was attested earlier (1841), but we assume it became more widespread in the 20th century due to its use in cinema and television.

In the case of the word *chef*, the change is less a single novel meaning than a narrowing of the lexeme. In the 19th century, *chef* appears primarily in unassimilated French borrowings, such as "chef d'œuvre" (masterpiece), "chef de brigade", "chef d'escadron", and the Louisiana toponym "Chef Menteur". This pattern is consistent with the OED entries for these borrowed phrases. In the 20th century these usages have mostly disappeared and the bare noun *chef* has acquired the meaning of "head cook". As for the word *tip*, the rising sense is the slang verb meaning "to give information or a hint" (as in "to tip off"), which refers to the British slang of the late 19th century. Thus, in all five cases the external lexicographic evidence generally confirms that the predicted changes correspond to genuine lexical semantic shifts.

7. Interpretability. We stated in the Introduction that the WSI-based systems are preferable to APD systems because each predicted cluster can be inspected individually. To illustrate this concretely, we show that the clusters produced by our GlossReader CC+JSD configuration can be read directly as candidate senses, including potentially novel ones. For each target word and each predicted cluster, we determine two simple quantities: the cluster size and the fraction of its usages that belong to the later time period (p_{late}). A large cluster with p_{late} close to one is direct evidence that the system has identified a sense that is characteristic of the later period. Similarly, a large cluster with p_{late} close to zero indicates a potential candidate for disappearance.

In theory, only the senses with $p_{\text{late}} = 1$ can be considered genuinely novel, but we also need to take the limitations of the system into account. The clustering is performed jointly on usages from both periods, without using the temporal label of each usage. Therefore, instead of finding senses with $p_{\text{late}} = 1$, we compare p_{late} of some cluster with p_{late} of the other clusters to identify potential novel senses. We also filter out clusters with a small number of usages, as we notice that they tend to be erroneous.

We apply this inspection to three of the five[7] top-predicted words: *plane*, *graft*, and *chef*. For each word we list the predicted clusters containing at least 10 usages (there are two hundred usages for each target word in the dataset) together with a short descriptive label and a representative usage snippet chosen from the actual DWUG data. The results are presented in Table 4.

For the word *plane*, the system produces one large cluster with a strong late lean (cluster 1, size 83, $p_{\text{late}} = 0.76$) in which most of the usages refer to an "aircraft". The complementary cluster 0 (size 93, $p_{\text{late}} = 0.25$) collects together the older geometric, mechanical, and woodworking senses. However, despite this arguably unjustified merging of senses in cluster 0, a user examining this output can recover the appearance of a new meaning in cluster 1 based solely on the high p_{late} value.

[7]We inspected only three words because the process requires manually investigating a large number of usages to assign a fitting label for a cluster.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

Table 4. Predicted clusters for *plane*, *graft*, and *chef*: only the eight clusters that contain at least 10 usages each are shown. The fraction of usages of the cluster coming from the 20th century is denoted as p_{late}. Bold labels mark novel-sense candidates

Target word	Cluster	Size	p_{late}	Descriptive label	Example usage
plane	1	83	0.76	**aircraft**	*"She'd been a fool to take it with her onto the* ***plane****"* (2005)
	0	93	0.25	geometry / mechanics	*"the height of the* ***plane****: the base"* (1827)
graft	0	79	0.82	**bribery**	*"****graft*** *had been his stock-in-trade"* (2009)
	1	50	0.18	horticultural grafting	*"Supposing the tree to be two or three years old from the* ***graft*** *or bud"* (1839)
	2	44	0.25	horticultural grafting	*"When the* ***graft*** *is set, it is to be bound with bass-string"* (1839)
chef	0	109	0.75	**head cook**	*"Trained as a pastry* ***chef****, Rachel Thebault opened"* (2009)
	1	13	0.31	French borrowings	*"in the direction of* ***Chef*** *Menteur"* (1819)
	2	13	0.31	French borrowings	*"****chef*** *d'oeuvres of the Dutch artists"* (1827)

For the word *graft*, the same inspection recovers both directions of the shift: cluster 0 is large, with a strong late lean, combining the political-corruption sense with later related uses, while clusters 1 and 2 are large, strongly early-leaning, and together covering the original horticultural meaning. Again, despite that clusters 1 and 2 arguably need to be merged, a user reading the table is naturally led to the correct interpretation that a new meaning has appeared.

As for the word *chef*, cluster 0 is large and late-leaning. It collects the modern "cook" sense. The two smaller early-leaning clusters collect the older French-borrowed uses of the lexeme ("chef d'œuvre", "Chef Menteur", "chef de brigade"). This case is informative because the "cook" sense was already present in the 19th century. The change lies in its increasing prevalence. The p_{late} values still correctly indicate which cluster corresponds to the rising dominant sense, and the distinction between cluster 0 and clusters 1–2 reflects the narrowing of the lexeme away from unassimilated French phrases.

To conclude, although the individual senses are often not captured well enough (supporting our claim made in Section 4 that clustering might be further improved), the system outputs can be used to identify clear-cut cases of appearing novel senses or other noticeable changes in sense frequency.

8. Computational complexity of LSCD pipeline. After demonstrating that WSI-based systems implementing the JSD model can achieve state-of-the-art quality, we now analyze the computational cost of the full pipeline to guide deployment for larger corpora. We frame the pipeline as a sequence of computational stages and characterize the complexity of each. For a target word with n usages extracted from two time periods, the pipeline consists of the following stages.

- Stage 1: Word-in-Context. For each target word, the WiC model computes pairwise similarity scores to construct the similarity matrix $S \in \mathbb{R}^{n\times n}$. This stage dominates the computational cost of the pipeline because it involves neural network inference. The cost depends on model architecture:
  - Dual encoders (GlossReader, XL-Lexeme) compute an embedding $\boldsymbol{e}_i \in \mathbb{R}^d$ for each usage independently, which requires $O(n)$ forward passes. Pairwise cosine similarities are then computed as $s_{ij} = \dfrac{\boldsymbol{e}_i \cdot \boldsymbol{e}_j}{\|\boldsymbol{e}_i\|\|\boldsymbol{e}_j\|}$ in $O(n^2 d)$ time, where the per-operation cost is negligible compared with neural inference. In practice, running GlossReader on the en benchmark took about 50 minutes on a RTX 3080 GPU.
  - Cross-encoders (DeepMistake) process both usages jointly, which requires $O(n^2)$ forward passes through the model. In practice, running DeepMistake on the en benchmark took about 10 hours on a RTX 3080 GPU, roughly 11 times slower than GlossReader.
- Stage 2: Clustering. Given the similarity matrix S, at the clustering stage the n usages are divided into groups corresponding to word senses. The correlation clustering problem (minimizing the total disagree-

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

ment between edge weights and cluster assignments) is NP-hard in general [18]. We use the implementation from [6], which employs a simulated annealing heuristic with multiprocessing with complexity $O(s \cdot T \cdot n^2)$ per target word, where s is the maximum number of clusters (we set the upper bound to 10) and T is the number of optimization iterations (we set this parameter to 20). The alternative algorithms we evaluate possess the following asymptotic complexity: agglomerative clustering has $O(n^2 \log n)$ complexity and affinity propagation has $O(n^2 \cdot T')$ complexity. Here T' is the number of affinity propagation iterations.

Despite the fact that performing Stage 1 requires $O(n)$ operations for dual encoders (compared to $O(n^2)$ of Stage 2), the clustering stage is in practice significantly cheaper than Stage 1. The constant factor C_1 in the $O(n)$ complexity of Stage 1 has the lower bound equal to the number of parameters in the WiC model (each model parameter is used at least once in model inference), which amounts to $3.5 \cdot 10^8$ in the case of XL-Lexeme and GlossReader. These constants dominate over n (the number of target word usages) for any reasonable dataset, thereby making Stage 1 the predominant component of the total computational cost.

- Stage 3: Change score computation. Calculation of the Jensen–Shannon distance between cluster frequency distributions in the two time periods requires $O(k)$ operations, where k is the number of clusters. This stage has negligible cost.

In summary, the computational bottleneck of the LSCD pipeline is the WiC stage. For practical deployment in larger corpora, dual-encoder models provide the best trade-off between quality and efficiency: they achieve quality comparable to cross-encoders (as shown in our WiC, WSI and LSCD evaluations) while requiring only $O(n)$ rather than $O(n^2)$ neural network forward passes per target word. Taking the above discussion into account, we conclude that GlossReader CC+JSD is the optimal default choice for practical deployment, since it implements the interpretable JSD model and combines state-of-the-art quality (Table 2) with computational efficiency.

9. Conclusion. In this work, we studied WSI-based LSCD systems, which implement the JSD mathematical model of semantic change.

First, we demonstrated that specialized WiC models achieve quality at the level of individual annotators, substantially outperforming general-purpose XLM-RoBERTa. We also showed that GlossReader and DeepMistake, which had not been thoroughly evaluated in LSCD before, surpass the previously reported state-of-the-art XL-Lexeme. For the WSI subtask, we introduced the Annotation Rand Index (AnnotRI), a clustering metric that is evaluated directly on reliable manual pairwise annotations and therefore provides a more reliable assessment than standard clustering metrics based on inferred senses. Using AnnotRI for cross-validated hyperparameter selection in correlation clustering, we substantially improved the clustering step.

Improved clustering and better WiC models enable WSI-based systems to achieve state-of-the-art results on the JSD-based benchmarks. Despite the fact that the automatic clusters are still often noisy, we were already able to automatically detect the manually verified emergence/disappearance of word's senses. For practical deployment, we finally recommend GlossReader CC+JSD as the optimal configuration, since it combines interpretability of the JSD mathematical model with state-of-the-art quality and best computational efficiency.

References

1. A. Blank, *Prinzipien des Lexikalischen Bedeutungswandels am Beispiel der Romanischen Sprachen* (Niemeyer, Tübingen, 1997). doi 10.1515/9783110931600.
2. F. Periti and N. Tahmasebi, "A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, June 16–21, 2024, Vol. 1: Long Papers* (Association for Computational Linguistics, 2024), pp. 4262–4282. doi 10.18653/v1/2024.naacl-long.240.
3. F. Periti and S. Montanelli, "Lexical Semantic Change through Large Language Models: a Survey," ACM Computing Surveys **56** (11), 1–38 (2024). doi 10.1145/3672393.
4. A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: a survey," in *Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 20–26, 2018* (Association for Computational Linguistics, 2018), pp. 1384–1397. https://aclanthology.org/C18-1117/. Cited July 17, 2026.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

5. D. Schlechtweg, B. McGillivray, S. Hengchen, et al., "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Online, December 12–13, 2020* (International Committee for Computational Linguistics, 2020), pp. 1–23. doi 10.18653/v1/2020.semeval-1.1.
6. D. Schlechtweg, S. Schulte im Walde, and S. Eckmann, "Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, June 1–6, 2018, Vol. 2: Short Papers* (Association for Computational Linguistics, 2018), pp. 169–174. doi 10.18653/v1/N18-2027.
7. M. T. Pilehvar and J. Camacho-Collados, "WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, June 2–7, Vol. 1: Long and Short Papers* (Association for Computational Linguistics, 2019), pp. 1267–1273. doi 10.18653/v1/N19-1128.
8. J. H. Lau, P. Cook, D. McCarthy, et al., "Word Sense Induction for Novel Sense Detection," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23–27, 2012* (Association for Computational Linguistics, 2012), pp. 591–601. https://aclanthology.org/E12-1060/. Cited July 17, 2026.
9. M. Martinc, S. Montariol, E. Zosa, and L. Pivovarova, "Capturing Evolution in Word Usage: Just Add More Clusters?" in *WWW'20: Companion Proceedings of the Web Conference, Taipei, Taiwan, April 20–24, 2020* (ACM, New York, 2020), pp. 343–349. doi 10.1145/3366424.3382186.
10. F. Periti, A. Ferrara, S. Montanelli, and M. Ruskov, "What is Done is Done: an Incremental Approach to Semantic Shift Detection," in *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland, May 26–27, 2022* (Association for Computational Linguistics, 2022), pp. 33–43. doi 10.18653/v1/2022.lchange-1.4.
11. A. Kutuzov and M. Giulianelli, "UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Online, December 12–13, 2020* (International Committee for Computational Linguistics, 2020), pp. 126–134. doi 10.18653/v1/2020.semeval-1.14.
12. M. Giulianelli, M. Del Tredici, and R. Fernández, "Analysing Lexical Semantic Change with Contextualised Word Representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 5–10, 2020* (Association for Computational Linguistics, 2020), pp. 3960–3973. doi 10.18653/v1/2020.acl-main.365.
13. F. D. Zamora-Reina, F. Bravo-Marquez, and D. Schlechtweg, "LSCDiscovery: A shared task on semantic change discovery and detection in Spanish," in *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland, May 26–27, 2022* (Association for Computational Linguistics, 2022), pp. 149–164. doi 10.18653/v1/2022.lchange-1.16.
14. A. Kutuzov, S. Touileb, P. Mæhlum, et al., "NorDiaChange: Diachronic Semantic Change Dataset for Norwegian," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, June 20–25, 2022* (European Language Resources Association, 2022), pp. 2563–2572. https://aclanthology.org/2022.lrec-1.274/. Cited July 18, 2026.
15. J. Chen, E. Chersoni, D. Schlechtweg, et al., "ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection," in *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, Singapore, December 6, 2023* (Association for Computational Linguistics, 2023), pp. 93–99. doi 10.18653/v1/2023.lchange-1.10.
16. A. Kutuzov and L. Pivovarova, "Three-part diachronic semantic change dataset for Russian," in *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change, Online, August 6, 2021* (Association for Computational Linguistics, 2021), pp. 7–13. doi 10.18653/v1/2021.lchange-1.2.
17. D. Schlechtweg, N. Tahmasebi, S. Hengchen, et al., "DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, Online, November 7–11, 2021* (Association for Computational Linguistics, 2021), pp. 7079–7091. doi 10.18653/v1/2021.emnlp-main.567.
18. N. Bansal, A. Blum, and S. Chawla, "Correlation Clustering," Machine Learning **56** (1), 89–113 (2004). doi 10.1023/B:MACH.0000033116.57574.95.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

19. A. Hätty, D. Schlechtweg, and S. Schulte im Walde, "SURel: A Gold Standard for Incorporating Meaning Shifts into Term Extraction," in *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics, June 2019, Minneapolis, Minnesota, USA* (Association for Computational Linguistics, 2019), pp. 1–8. doi 10.18653/v1/S19-1001.
20. S. Kurtyigit, M. Park, D. Schlechtweg, J. Kuhn, and S. Schulte im Walde, "Lexical Semantic Change Discovery," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Association for Computational Linguistics, Online, 2021), pp. 6985–6998. doi 10.18653/v1/2021.acl-long.543.
21. A. Aksenova, E. Gavrishina, E. Rykov, and A. Kutuzov, "RuDSI: Graph-based Word Sense Induction Dataset for Russian," in *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, October 2022, Gyeongju, Republic of Korea* (Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022), pp. 77–88. https://aclanthology.org/2022.textgraphs-1.9. Cited July 29, 2026.
22. M. Martinc, P. K. Novak, and S. Pollak, "Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift," in *Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, May 11–16, 2020* (European Language Resources Association, 2020), pp. 4811–4819. https://aclanthology.org/2020.lrec-1.592/. Cited July 17, 2026.
23. C. Beck, "DiaSense at SemEval-2020 Task 1: Modeling Sense Change via Pre-trained BERT embeddings," in *Proceedings of the Fourteenth International Workshop on Semantic Evaluation, Barcelona, Online, December 12–13, 2020* (Association for Computational Linguistics, 2020), pp. 50–58. doi 10.18653/v1/2020.semeval-1.4.
24. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, June 2–7, 2019, Vol. 1: Long and Short Papers* (Association for Computational Linguistics, 2019), pp. 4171–4186. doi 10.18653/v1/N19-1423.
25. S. Laicher, S. Kurtyigit, D. Schlechtweg, et al., "Explaining and Improving BERT Performance on Lexical Semantic Change Detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Online, April 19–23, 2021* (Association for Computational Linguistics, 2021), pp. 192–202. doi 10.18653/v1/2021.eacl-srw.25.
26. A. Conneau, K. Khandelwal, N. Goyal, et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 5–10, 2020* (Association for Computational Linguistics, 2020), pp. 8440–8451. doi 10.18653/v1/2020.acl-main.747.
27. M. Giulianelli, A. Kutuzov, and L. Pivovarova, "Do Not Fire the Linguist: Grammatical Profiles Help Language Models Detect Semantic Change," in *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland, May 26–27, 2022* (Association for Computational Linguistics, 2022), pp. 54–67. doi 10.18653/v1/2022.lchange-1.6.
28. M. Rachinskiy and N. Arefyev, "GlossReader at SemEval-2021 Task 2: Reading Definitions Improves Contextualized Word Embeddings," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, August 5–6, 2021* (Association for Computational Linguistics, 2021), pp. 756–762. doi 10.18653/v1/2021.semeval-1.100.
29. N. Arefyev, M. Fedoseev, V. Protasov, et al., "DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model," in *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii* (RGGU, Moscow, 2021), pp. 16–30. https://dialogue-conf.org/media/5491/arefyevnplusetal133.pdf. Cited July 29, 2026.
30. P. Cassotti, L. Siciliani, M. DeGemmis, et al., "XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, July 9–14, 2023, Vol. 2: Short Papers* (Association for Computational Linguistics, 2023), pp. 1577–1585. doi 10.18653/v1/2023.acl-short.135.
31. M. Fedorova, T. Mickus, N. Partanen, et al., "AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling," in *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, Bangkok, Thailand, August 15, 2024* (Association for Computational Linguistics, 2024), pp. 72–91. doi 10.18653/v1/2024.lchange-1.8.
32. D. Kokosinskii, M. Kuklin, and N. Arefyev, "Deep-change at AXOLOTL-24: Orchestrating WSD and WSI Models for Semantic Change Modeling," in *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, Bangkok, Thailand, August 15, 2024* (Association for Computational Linguistics, 2024), pp. 168–179. doi 10.18653/v1/2024.lchange-1.16.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

33. D. Schlechtweg, T. Choppa, W. Zhao, and M. Roth, "CoMeDi Shared Task: Median Judgment Classification & Mean Disagreement Ranking with Ordinal Word-in-Context Judgments," in *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation, Abu Dhabi, UAE, January 19, 2025* (International Committee on Computational Linguistics, 2025), pp. 33–47. https://aclanthology.org/2025.comedi-1.4/. Cited July 19, 2026.
34. A. Raganato, T. Pasini, J. Camacho-Collados, and M. T. Pilehvar, "XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, November 16–20, 2020* (Association for Computational Linguistics, 2020), pp. 7193–7206. doi 10.18653/v1/2020.emnlp-main.584.
35. F. Martelli, N. Kalach, G. Tola, and R. Navigli, "SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC)," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, August 5–6, 2021* (Association for Computational Linguistics, 2021), pp. 24–36. doi 10.18653/v1/2021.semeval-1.3.
36. Q. Liu, E. M. Ponti, D. McCarthy, et al., "AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, Online, November 7–11, 2021* (Association for Computational Linguistics, 2021), pp. 7151–7162. doi 10.18653/v1/2021.emnlp-main.571.
37. G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker, "A Semantic Concordance," in *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21–24, 1993* https://aclanthology.org/H93-1061/. Cited July 19, 2026.
38. D. Homskiy and N. Arefyev, "DeepMistake at LSCDiscovery: Can a Multilingual Word-in-Context Model Replace Human Annotators?" in *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland, May 26–27, 2022* (Association for Computational Linguistics, 2022), pp. 173–179. doi 10.18653/v1/2022.lchange-1.18.
39. W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," Journal of the American Statistical Association **66** (336), 846–850 (1971). doi 10.1080/01621459.1971.10482356.
40. K. Erk, D. McCarthy, and N. Gaylord, "Measuring Word Meaning in Context," Computational Linguistics **39** (3), 511–554 (2013). doi 10.1162/coli_a_00142.
41. A. Kilgarriff, "How Dominant Is the Commonest Sense of a Word?" in *International Conference on Text, Speech and Dialogue Lecture Notes in Computer Science. Vol. 3206.* (Springer, Berlin Heidelberg, 2004), pp. 103–111. doi 10.1007/978-3-540-30120-2_14.
42. D. Kokosinskii and N. Arefyev, "Multilingual Substitution-based Word Sense Induction," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, May 20–25, 2024* (ELRA and ICCL, Torino, 2024), pp. 11859–11872. https://aclanthology.org/2024.lrec-main.1035/. Cited July 19, 2026.
43. *Oxford English Dictionary* (Oxford University Press, Oxford, online edition, 2026). https://www.oed.com. Cited July 17, 2026.
44. T. Blevins and L. Zettlemoyer, "Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July, 2020* (Association for Computational Linguistics, Online, 2020), pp. 1006–1017. doi 10.18653/v1/2020.acl-main.95.
45. T. Pasini, A. Raganato, and R. Navigli, "XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation," in *Proceedings of the AAAI Conference on Artificial Intelligence, February 2-9, 2021, Online* **35** (15), 13648–13656 (2021). doi 10.1609/aaai.v35i15.17609.
46. A. Grattafiori, A. Dubey, A. Jauhri, et al., "The Llama 3 Herd of Models," arXiv:2407.21783 (2024). doi 10.48550/arXiv.2407.21783.

Received	*Accepted*	*Published*
March 13, 2026	*May 28, 2026*	*July 31, 2026*

Information about the authors

Denis V. Kokosinskii — Ph.D. student; Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Leninskie Gory, 1, building 4, 119991, Moscow, Russia.

Nikolay V. Arefyev — Ph.D., Postdoctoral Research Fellow; University of Oslo, Department of Informatics, Boks 1072 Blindern, NO-0316 OSLO, Norway, NO-0316, Oslo, Norway.

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

Appendix 1
Detailed description of WiC models

GlossReader [28] is an XLM-RoBERTa-based multilingual extension of BEM [44], a joint encoder trained to map an ambiguous target word to the correct sense description. Although it was trained on the English dataset SemCor [37], GlossReader demonstrated high results at LSCDiscovery [13] (an LSCD competition on Spanish DWUG).

DeepMistake [29, 38] is a cross-encoder specifically trained for the WiC task. In other words, it predicts whether two word usages represent the same meaning or different ones. DeepMistake MCL→RSS and MCL+RSS+DWUG_{es}+XLWSD_{es} were the best-performing models at the RuShiftEval [16] and LSCDiscovery [13] competitions, respectively. The first one was trained on a multilingual mix of MCL-WiC [35], the development (dev) sets of Russian and Spanish DWUGs, and the Spanish part of XLWSD [45]. The latter model was trained on MCL-WiC [35] and then fine-tuned only on the dev set of Russian DWUG. For brevity, we refer to the first one as DM-MCL+ru+es and the second as DM-MCL→ru.

Another lexical semantic model, XL-Lexeme [30], was already evaluated in [2] on DWUGs and demonstrated state-of-the-art results. XL-Lexeme was trained on a mix of multilingual datasets, namely on MCL-WiC [35] and AM2iCo [36].

Appendix 2
WiC on DWUGs

The full comparison of WiC models on DWUGs is presented in Table 5.

Table 5. WiC evaluation on DWUGs

	model	AVG	de	en	es	no12	no23	ru12	ru13	ru23	sv	zh
AP@1.5	DM-MCL→es	0.751	0.696	0.737	0.755	0.758	**0.777**	**0.755**	0.782	**0.792**	**0.801**	0.657
	DM-MCL→ru	0.736	0.672	0.713	0.727	**0.772**	0.759	0.738	0.779	0.779	0.785	0.635
	DM-MCL+ru+es	0.739	0.691	0.720	0.730	0.767	0.750	0.739	**0.783**	0.778	0.790	0.641
	DM-en	0.719	0.655	0.687	0.699	0.762	0.761	0.744	0.744	0.759	0.768	0.632
	GlossReader	**0.754**	**0.713**	**0.780**	**0.772**	0.759	0.755	0.736	0.775	0.784	0.795	**0.673**
	XL-Lexeme	0.705	0.662	0.695	0.691	0.713	0.723	0.718	0.739	0.746	0.715	0.648
	XLM-RoBERTa	0.582	0.568	0.586	0.578	0.653	0.622	0.522	0.562	0.560	0.589	0.582
AP@2.5	DM-MCL→es	**0.780**	**0.767**	0.776	0.783	0.777	**0.799**	**0.772**	0.797	**0.796**	**0.820**	0.712
	DM-MCL→ru	0.771	0.752	0.759	0.765	**0.790**	0.784	0.771	**0.802**	0.792	0.808	0.687
	DM-MCL+ru+es	0.772	0.763	0.765	0.766	0.786	0.772	0.767	0.801	0.794	0.813	0.694
	DM-en	0.756	0.739	0.738	0.738	0.784	0.784	0.764	0.764	0.776	0.792	0.687
	GlossReader	0.775	0.762	**0.812**	**0.787**	0.770	0.773	0.750	0.784	0.786	0.817	**0.713**
	XL-Lexeme	0.745	0.739	0.758	0.729	0.733	0.755	0.742	0.764	0.764	0.759	0.711
	XLM-RoBERTa	0.596	0.593	0.611	0.596	0.664	0.629	0.528	0.568	0.574	0.599	0.599
AP@3.5	DM-MCL→es	**0.773**	0.762	0.771	**0.760**	0.782	**0.805**	0.749	0.771	0.764	**0.820**	0.747
	DM-MCL→ru	0.771	0.757	0.761	0.751	**0.795**	0.798	**0.756**	**0.783**	**0.773**	0.815	0.724
	DM-MCL+ru+es	0.771	**0.763**	0.765	0.750	0.792	0.787	0.750	0.778	0.771	0.818	0.732
	DM-en	0.758	0.744	0.744	0.730	0.793	0.795	0.743	0.743	0.756	0.799	0.721
	GlossReader	0.766	0.750	**0.798**	0.759	0.771	0.778	0.732	0.762	0.759	0.816	0.734
	XL-Lexeme	0.747	0.740	0.759	0.722	0.743	0.764	0.728	0.747	0.746	0.775	**0.751**
	XLM-RoBERTa	0.601	0.599	0.616	0.600	0.668	0.633	0.539	0.571	0.577	0.603	0.607
AP_1vs4	DM-MCL→es	**0.893**	0.916	0.912	**0.920**	0.791	**0.834**	0.931	0.938	0.917	**0.924**	0.850
	DM-MCL→ru	0.888	0.905	0.903	0.908	0.801	0.822	**0.942**	**0.955**	**0.938**	0.915	0.791
	DM-MCL+ru+es	0.887	**0.917**	0.908	0.914	0.799	0.799	0.932	0.943	0.927	0.917	0.819
	DM-en	0.873	0.897	0.888	0.880	**0.815**	0.825	0.924	0.924	0.922	0.899	0.805
	GlossReader	0.872	0.896	**0.942**	0.915	0.762	0.747	0.902	0.914	0.915	0.906	0.825
	XL-Lexeme	0.859	0.882	0.908	0.862	0.748	0.775	0.884	0.893	0.904	0.871	**0.862**
	XLM-RoBERTa	0.624	0.642	0.672	0.660	0.648	0.574	0.562	0.634	0.618	0.617	0.616
SpearmanR	DM-MCL→es	0.606	0.659	0.642	**0.598**	0.570	0.598	0.568	0.595	0.568	0.643	0.618
	DM-MCL→ru	**0.618**	0.655	0.621	0.592	0.594	**0.613**	**0.608**	**0.639**	**0.621**	0.642	0.593
	DM-MCL+ru+es	0.614	**0.670**	0.634	0.589	0.592	0.598	0.591	0.613	0.604	**0.652**	0.600
	DM-en	0.591	0.642	0.593	0.559	0.588	0.596	0.566	0.566	0.577	0.621	0.573
	GlossReader	0.592	0.620	**0.687**	0.582	0.548	0.577	0.539	0.586	0.571	0.641	0.565
	XL-Lexeme	0.573	0.618	0.609	0.546	0.514	0.551	0.549	0.564	0.558	0.583	**0.634**
	XLM-RoBERTa	0.251	0.270	0.293	0.261	0.370	0.312	0.119	0.178	0.186	0.223	0.301

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

Appendix 3
Optimal threshold in correlation clustering

Table 6 presents optimal thresholds in correlation clustering for DeepMistake, GlossReader and XL-Lexeme.

Table 6. Optimal threshold in correlation clustering according to different metrics for WiC models

Metric	Model	de	en	es	no12	no23	ru12	ru13	ru23	sv	zh	mode
ARI	DeepMistake	0.479	0.479	0.487	0.487	0.474	n./a.	n./a.	n./a.	0.487	0.487	0.487
	GlossReader	0.259	0.178	0.082	0.178	0.178	n./a.	n./a.	n./a.	0.082	0.178	0.178
	XL-Lexeme	0.577	0.577	0.470	0.577	0.500	n./a.	n./a.	n./a.	0.577	0.350	0.577
AnnotRI	DeepMistake	0.479	0.487	0.487	0.487	0.487	0.487	0.487	0.487	0.487	0.487	0.487
	GlossReader	0.259	0.178	0.178	0.082	0.178	0.259	0.259	0.259	0.259	0.259	0.259
	XL-Lexeme	0.470	0.577	0.577	0.676	0.500	0.577	0.577	0.577	0.577	0.577	0.577

Appendix 4
WSI on DWUGs

The full per-benchmark breakdown of the WSI evaluation is presented in Figure 5.

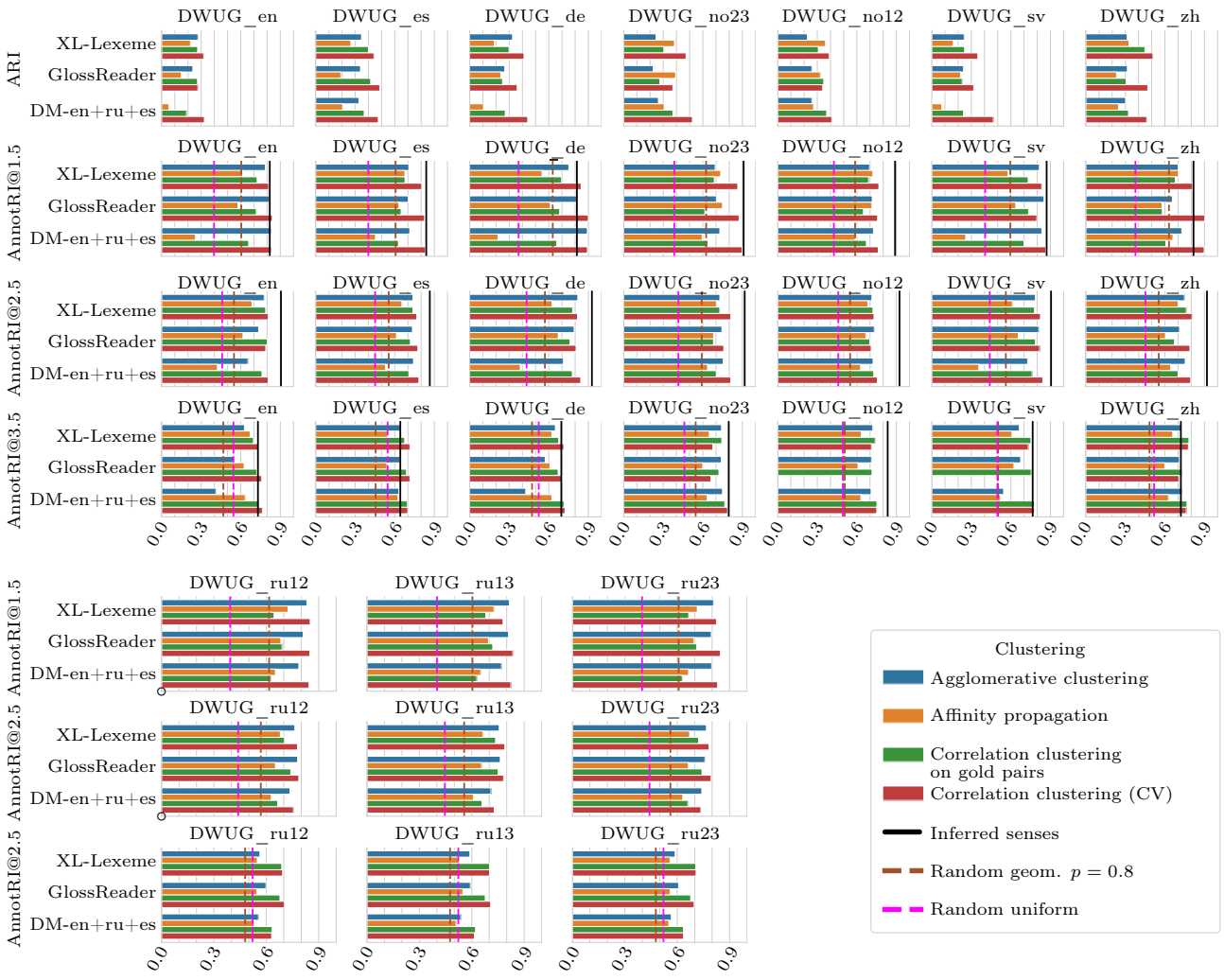

Figure 5. WSI evaluation: top — DWUGs (with inferred senses available for evaluation using ARI), bottom — COMPARE-based datasets (ru12, ru13, ru23; no inferred senses)

2026, **27** (3), 450–469. doi 10.26089/NumMet.v27r329

Appendix 5
Evaluation of APD on Gold Pairs

The LSCD evaluation of APD on gold usage pairs (pairs with human judgments) for different WiC models is presented in Table 7. Here llama3.1-8b and llama3.1-70b denote the Llama 3 models [46], and XLM-RoBERTa-large stands for the base XLM-RoBERTa model without task-specific fine-tuning.

Table 7. LSCD evaluation of APD (COMPARE model) on DWUG and COMPARE-based datasets

benchmark	de	en	es	no12	no23	ru12	ru13	ru23	sv	zh
DM-MCL→es	0.744	0.788	0.681	0.720	0.628	0.821	0.833	0.813	0.820	0.685
DM-MCL→ru	0.722	0.766	0.648	0.786	0.584	0.839	0.858	0.837	0.786	0.683
DM-MCL+ru+es	0.730	0.762	0.640	0.744	0.567	0.838	0.846	0.822	0.811	0.635
DM-en	0.677	0.752	0.469	0.760	0.556	0.815	n./a.	0.783	0.757	0.711
GlossReader	0.721	0.836	0.660	0.615	0.594	0.762	0.808	0.792	0.810	0.698
XL-Lexeme	0.746	0.833	0.589	0.645	0.632	0.780	0.851	0.824	0.831	0.701
XLM-RoBERTa-large	0.229	0.373	0.251	0.503	0.333	0.168	0.330	0.306	0.349	0.415
llama3.1-8b	0.373	0.391	0.300	0.369	0.135	0.518	0.533	0.469	0.577	0.519
llama3.1-70b	0.737	0.748	0.514	0.701	0.569	0.837	0.837	0.827	0.871	0.757