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.

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. 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.

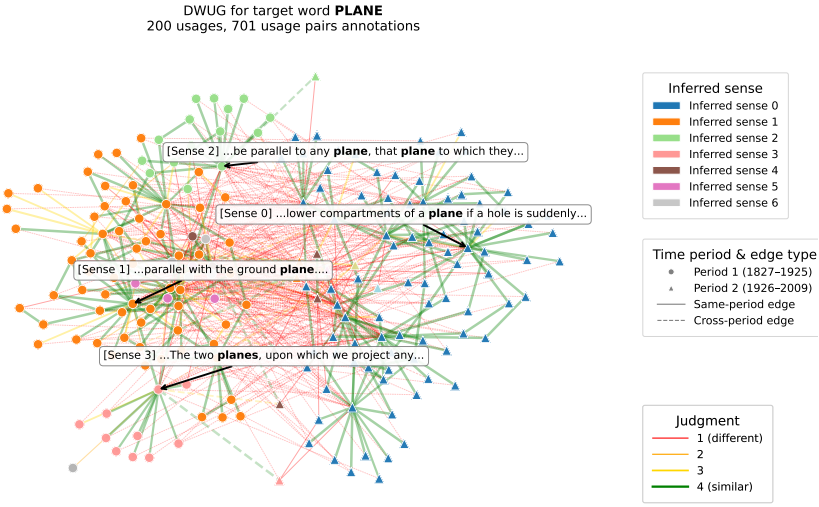


Рисунок 1. Диахронический граф словоупотреблений для целевого слова *plane* в двух временных периодах. Узлы соответствуют словоупотреблениям, ребра – экспертным оценкам семантической близости в парах словоупотреблений.

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].

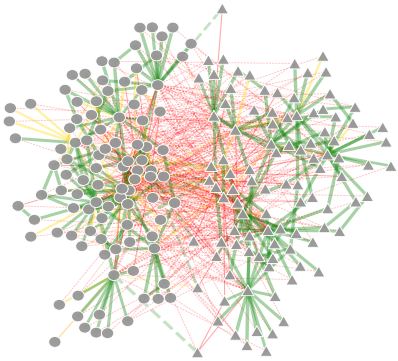
(*2*) Новая метрика AnnotRI для задачи WSI, лучше подходящая для DWUG. Она использует в качестве эталонных меток только надежные вручную аннотированные данные и позволяет оценивать качество при разных уровнях гранулярности значений слов.

1. Связанные работы

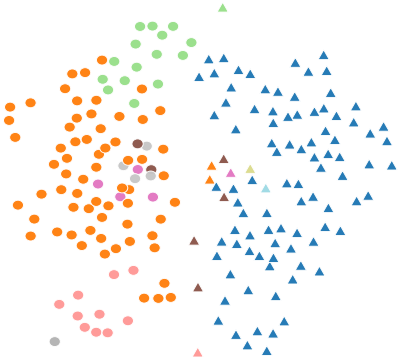
1.1. Диахронические графы словоупотреблений

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

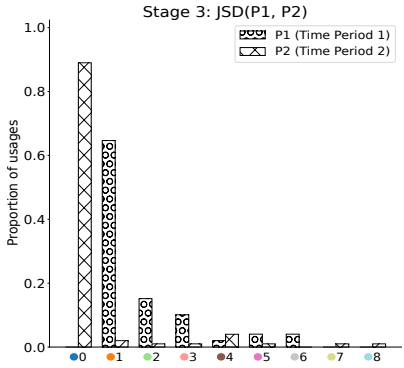
(*а*) Stage 0: Использование за два временных периода



(*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.

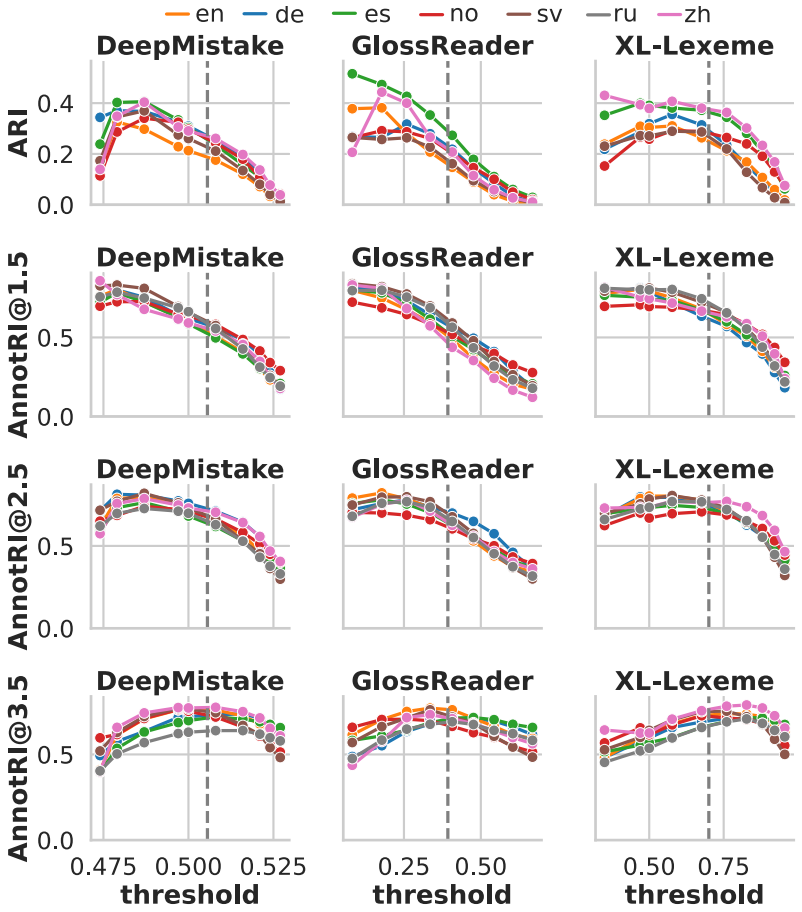


Рисунок 3. Результаты WSI для корреляционной кластеризации в зависимости от порога; серая вертикальная линия обозначает среднее значение оценок сходства по всем бенчмаркам для каждой модели, это CC-порог, использованный в [2]

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. arXiv 1806.03537 ↑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. 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. ↑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. ↑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. ↑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. ↑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 ↑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 ↑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. ↑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 ↑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. ↑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. ↑128

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

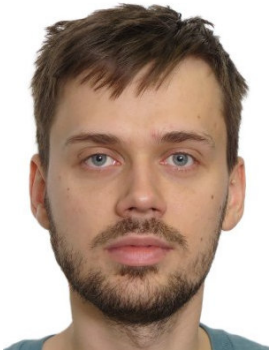
[33] M. Fedorova, A. Kutuzov, Y. Scherrer *Definition generation for lexical semantic change detection // Findings of the Association for Computational Linguistics*, ACL 2024 (Bangkok, Thailand and virtual meeting, August 11–16, 2024).– ACL.– 2024.– ISBN 979-8-89176-099-8.– Pp. 5712–5724. arXiv:2406.14167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