



Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities: LLM4DH

Izroček projekta / Project Deliverable

Nova metodologija za diahrono analizo z velikimi jezikovnimi modeli /

Novel methodology for diachronic analysis using LLMs

Št. izročka / Deliverable No.	4.2
Datum oddaje / Delivery date	30. 9. 2026 (Verzija / Version 1.0)
Vrsta / Nature	R – Poročilo / Report
Nivo diseminacije / Dissemination level	PU – Javno / Public
Delovni sklop / Work package	WP4: Advanced Technologies for Digital Humanities
Naloga / Task	T4.2: LLMs for diachronic analysis
Avtor, Vodilni partner / Author, Lead partner	Marko Pranjić (IJS)
Sodelujoči partner / Contributing partner	Senja Pollak (IJS)

Vrsta / Nature: R = Report, P = Prototype, D = Demonstrator, O = Other
Diseminacija / Dissemination: PU = Public; CO = Confidential, for consortium members and ARIS

Informacijska tabela projekta / Project information card	
Akronim / Acronym	LLM4DH
Naslov / Title	Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities
Spletna stran / Website	https://www.cjvt.si/llm4dh/
Št. projekta / Project No.	GC-0002
Razpis / Call	Javni razpis za (so)financiranje Gravitacije, št. 5100-70/2024-6 z dne 7. 5. 2024, in spremembe javnega razpisa, št. 5100-70/2024-11 z dne 29. 5. 2024.
Koordinator / Coordinator	Marko Robnik-Šikonja
Trajanje / Duration	oktober 2024–september 2027, 36 mesecev / 36 months
Povzetek projekta	Jezikovne tehnologije igrajo vse pomembnejšo vlogo na številnih področjih znanosti. Z bliskovitim napredkom velikih jezikovnih modelov se ti pozicionirajo kot osrednja tehnologija umetne inteligence z daljnosežnim vplivom ne le na znanstveno raziskovanje, ampak tudi na širši družbeni razvoj in celotno družbo. Veliki jezikovni modeli kljub svojim velikim zmožnostim izkazujejo vrsto pomanjkljivosti, kot so nekonsistentni in napačni odgovori, velike računske zahteve, šibko poznavanje jezikov z manj viri, neprilagojenost za nekatere pomembne naloge in domene ter slabosti v razumevanju družbe, etike in človekovih potreb. Projekt bo tehnologijo velikih jezikovnih modelov izboljšal z različnimi znanji in podatki, jim zmanjšal računsko zahtevnost pri govornih tehnologijah in jim omogočil dostop do zanesljivih zunanjih informacij. Izboljšave v temeljnih zmožnostih velikih jezikovnih modelov bodo vodile do prebojnih dosežkov na več področjih digitalne humanistike: v jezikoslovju, leksikografiji, izobraževanju, analizi družbenih pojavov in sodobne zgodovine, političnih vedah, folkloristiki in pravu.
Project abstract	Language technologies are increasingly important in many areas of science. With the rapid advances in large-scale language models, they are positioning themselves as a core AI technology with far-reaching impact on scientific research, wider societal development, and society. Despite their great capabilities, large language models exhibit several shortcomings, such as inconsistent and incorrect answers, high computational demands, weak knowledge of less-resourced languages, poor performance in some important tasks and domains, and weaknesses in understanding society, ethics, and human needs. The project will improve the technology of large language models by enriching them with diverse knowledge and data, using them to improve the performance of speech technologies, and giving them access to reliable external information. Improvements in the core capabilities of large language models will lead to breakthroughs in several areas of the digital humanities: linguistics, lexicography, education, analysis of social phenomena and contemporary history, political sciences, folkloristics, and law.
Vodilni partner projekta / Project lead partner	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko (UL FRI)

Informacijska tabela projekta / Project information card (continued)	
Sodelujoči partnerji / Participating partners	Institut Jožef Stefan (IJS) Inštitut za novejšo zgodovino (INZ) Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja (IRRIS) Inštitut za kriminologijo pri Pravni fakulteti v Ljubljani (IK) Univerza v Ljubljani, Filozofska fakulteta (UL FF) Univerza v Ljubljani, Fakulteta za elektrotehniko (UL FE) Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko (UM FERi)
Opis naloge T4.2 / Description of Task T4.2	Razvijamo metode za avtomatizirano odkrivanje semantičnih sprememb in diahrono analizo besednih rab z uporabo velikih jezikovnih modelov. Metode povezujejo vrednotenje na slovenskem zlatem standardu z razložljivo analizo obsežnih korpusov. We develop methods for automated semantic change detection and diachronic analysis of word usages using large language models. The methods connect evaluation on a Slovene gold standard with interpretable analysis of large corpora.

Zgodovina dokumenta / Document history

Datum / Date	Verzija	Avtor / Author	Komentar / Comment
30. 9. 2026	1.0	Marko Pranjić, Senja Pollak	Izroček naloge T4.2 / Deliverable for Task T4.2.

Okrajšave in akronimi / Abbreviations and acronyms

APD	Average Pairwise Distance
ARIS	Slovenian Research and Innovation Agency
COMPARE	Mean cross-period usage-relatedness rating
DH	Digital humanities
DURel	Diachronic Usage Relatedness
IJS	Jožef Stefan Institute
INZ	Institute of Contemporary History
LLM / VJM	Large language model / veliki jezikovni model
MRL	Matryoshka Representation Learning
OCR	Optical character recognition
OT	Optimal transport
PCA	Principal component analysis
WP	Work package

Kazalo vsebine / Table of contents

Executive Summary	7
1 Introduction	8
2 Gold-standard dataset for Slovene	8
2.1 Annotation task and change scores	9
3 From usage judgments to corpus timelines	9
3.1 Benchmarking generative LLMs	9
3.1.1 Models	9
3.1.2 Experimental setup	9
3.1.3 Evaluation	10
3.1.4 Results	10
3.1.5 Instruction-based embedding analysis	12
3.1.6 Discussion and limitations	13
3.2 APD metric analysis and optimal transport based semantic change detection . .	13
3.2.1 Formal analysis of the APD metric	14
3.2.2 Regularized optimal transport	14
3.2.3 Results	14
3.3 Scaling word usage analysis to the corpus level	15
3.3.1 Decoupling context encoding from target selection	16
3.3.2 Compact usage representations	16
3.3.3 Results	16
3.3.4 Diachronic change diagnostic	18
3.4 Building timelines of word usage clusters	18
3.4.1 Corpus preparation and occurrence extraction	18
3.4.2 Two-stage clustering of usages	19
3.4.3 Constructing and interpreting timelines	20
4 Use cases	20
4.1 Poverty in SPeriodika	20
4.2 Global warming in Slovenian news	22
4.2.1 Corpus preparation and clustering results	22
4.2.2 Context families and temporal patterns	23
4.2.3 Source checks and interpretive limits	25
5 Availability	26

6 Conclusion	26
References	26

Executive Summary

This deliverable presents the methodology developed in Task T4.2 for detecting semantic change and examining word usages over time. The aim is to support interpretable analysis of large historical corpora, including subtle differences in how words are used.

We evaluate eleven prompted language models on a Slovene gold standard. The strongest models produce informative word rankings, but their agreement with human judgments on individual usage pairs remains below agreement among human annotators. General-purpose generative and encoder models represent many aspects of context, so variation among word-occurrence vectors can reflect more than differences in meaning. Our formal analysis establishes that dispersion can inflate semantic-change scores even when usage distributions remain unchanged across periods. The encoder evaluation shows the practical impact of this effect, motivating representations trained for usage similarity. We focus on encoders because of their lower computational cost at corpus scale. Regularized optimal transport reduces this bias and improves results with SloBERTa. Using XL-LEXEME, an encoder trained to compare word usages, produces a larger observed improvement than changing the distance measure alone.

To support corpus-scale analysis, we develop representations that encode multiple target words in a sentence with a single encoder pass. Compact 128–256-dimensional vectors retain evaluation quality while reducing storage and similarity-computation costs. We combine these representations with occurrence extraction, near-duplicate detection, and two-stage clustering. Community detection identifies tightly connected usage groups; hierarchical merging of their centroids brings related groups together into broader contexts that can be traced through time.

An earlier prototype applied to poverty-related discourse in sPeriodika revealed limitations that informed the revised methodology. Its timelines show contrasting patterns for religious and worker-related contexts of *ubog*. A smaller multilingual news corpus on “global warming” serves as a pilot for the revised approach before further analysis of sPeriodika. Our approach produces timelines of recurring usage contexts and links them to dated source passages, allowing researchers to examine and interpret changes in their distribution. The analysis can be restricted to selected years or publishers to investigate when and where particular uses occur. These outputs connect automated usage grouping with historical analysis of the source texts.

1. Introduction

LLM4DH develops language technologies for research across the digital humanities. Within its work package on advanced technologies for digital humanities (WP4), Task T4.2 addresses diachronic analysis: how to identify and explain changes in the uses and meanings of words across time. The project description identifies a gap between semantic-change methods tested on small gold-standard datasets and the nuanced, interpretable results needed for humanities research.¹

In this deliverable, we first introduce a gold standard for Slovene on which we evaluate several approaches (Section 2). Next, we benchmark eleven open-weight generative LLMs on a Slovene gold standard by prompting them to rate pairs of word usages (Section 3.1). We find that the strongest models provide useful judgments, with pair-level results broadly comparable to those of the non-fine-tuned SloBERTa encoder (Uičar and Robnik-Šikonja, 2021). XL-LEXEME, an encoder trained to compare word usages, achieves the highest pair-level correlation. After aggregation by word, the strongest prompted models and XL-LEXEME are broadly comparable. These results, together with the lower computational cost of encoding target occurrences than generating substitutes for each one, led us to focus the remaining work on encoder-based methods for large corpora.

Second, we show that average pairwise distance (APD), a common measure of semantic change, can assign a high change score to a word whose usages are diverse but stable over time. General-purpose generative and encoder models represent many aspects of context, so variation among word-occurrence vectors can reflect more than differences in meaning. Our formal analysis establishes how this variation affects APD, and our encoder evaluation shows its practical impact. We propose and evaluate a semantic-change measure based on regularized optimal transport to reduce this bias (Section 3.2). The stronger results of XL-LEXEME further support training representations for usage similarity. Together with the computational advantage described above, these findings motivate our focus on fine-tuned encoders for corpus-scale analysis.

Third, we improve the efficiency of encoder-based analysis. XL-LEXEME requires a separate encoder pass for each target word occurrence, repeating work when several targets occur in the same text. We develop compact usage representations that encode multiple words in a sentence with a single encoder pass (Section 3.3).

Fourth, we build a method for grouping word usages and tracing the resulting context families through time (Section 3.4).

We demonstrate our methods in specific use cases (Section 4): on the Slovenian corpus SPeriodika and on a multilingual news pilot on “global warming”.

Section 5 lists the dataset, code, and other project artifacts. Finally, in Section 6, we conclude this report.

2. Gold-standard dataset for Slovene

We evaluate several of our approaches on the Slovene gold-standard dataset introduced in our earlier work (Pranjić et al., 2025). The dataset covers two temporal slices of Gigafida 2.0, the reference corpus of written Slovene: texts from 1990 to 1997 (approximately 70 million words) and from 2018 (approximately 80 million words) (Krek et al., 2020). Candidate words with suspected semantic change were proposed by three automatic change detection systems, verified by two native speakers, and complemented with filler words matched in part of speech

¹LLM4DH project description, Section 21.1.2 (Task T4.2).

and frequency, yielding a 105-word list that spans changed and stable vocabulary. It includes words with new or shifting uses (e.g., *evro* ‘euro’, *portal* ‘web portal’, *robot*) and stable vocabulary (e.g., *molekula* ‘molecule’, *vrag* ‘devil’). For each word, 30 usages per period were randomly paired across periods, producing 3,150 usage pairs.

2.1. Annotation task and change scores

Three annotators rated each pair for the semantic relatedness of the target word across the two usages. The four-point ordinal scale follows the DUREL framework for lexical semantic change (Schlechtweg et al., 2018): (1) unrelated meanings (homonymy), (2) distantly related meanings (metaphor), (3) closely related meanings (polysemy or metonymy), and (4) identical meaning. A score of 0 marks a missing judgment. We use the human annotations as the reference standard for evaluation.

The annotators agreed exactly on 62.1% of the decisions and reached an overall ordinal Krippendorff’s α of 0.720. The COMPARE metric averages the relatedness ratings for each word: lower scores indicate greater semantic change. Scores range from 1.18 (*burka*, ‘burqa’/‘farce’) to 4.0 (*pastor*).

After annotation, one word (*zenit*) was removed because a preprocessing artifact matched the soccer club name *Zenit* to the common noun. The published release therefore contains 104 words and 3,120 pairs. The prompted LLM benchmark uses the full 105-word annotation file for pair-level evaluation; its word-level evaluation excludes *zenit* to match the published gold standard.

3. From usage judgments to corpus timelines

3.1. Benchmarking generative LLMs

We evaluate how well current open-weight LLMs reproduce human judgments of graded semantic relatedness between two usages of the same word, one sampled from 1990–1997 and one from 2018 (Section 2). The benchmark targets the annotation step of diachronic semantic-change analysis: reliable usage-level judgments are a prerequisite for word-level change detection.

3.1.1. Models

We benchmark eleven open-weight instruction-tuned models from five families, spanning 2B to 31B parameters: the Slovenian-adapted GaMS family (GaMS-2B, GaMS-9B, GaMS3-12B, GaMS-27B, and the Nemotron-tuned GaMS-27B-Nemotron variant), Qwen3 (Qwen3-14B with and without thinking mode, and Qwen3-30B-A3B), GLM-4.7-Flash, GPT-OSS-20B, and Gemma-3-27B-it. The selection supports three comparisons: language-adapted versus general-purpose models, model scale, and the effect of explicit reasoning (thinking) mode.

3.1.2. Experimental setup

We use the cross-period usage pairs described in Section 2, presenting the same pairs to each model. Each model receives a Slovenian prompt with the target word, the two usages, the four-point rubric, decision instructions, and four worked examples. We extract a score from each response; unparsable answers and *cannot decide* responses are treated as missing. GLM-4.7-Flash has valid responses for 2,439 of the 3,150 pairs, while every other model covers more than 99% of the dataset.

3.1.3. Evaluation

We compute pairwise Krippendorff's α with the ordinal distance between every pair of annotators: the three humans and the eleven LLMs (Krippendorff, 2018). α equals 1 for perfect agreement and 0 for agreement at chance level; negative values indicate systematic disagreement. The three human–human pairs provide a reference for the difficulty of the task.

The LLM benchmark and encoder study use the same Slovene target words and source usages. For an exploratory word-level check, we average each prompted model's valid 1–4 ratings on the annotated pairs within each word and rank the 104 words in the published gold standard by that mean. We exclude score 0 and missing or unparseable responses, and omit *zenit*, as in the published dataset. Spearman's ρ compares this ranking with the human COMPARE ranking. Unlike pairwise α , it tests whether a model orders words by degree of change after ratings are aggregated. COMPARE is itself the mean human rating on these annotated pairs. The encoder study instead computes APD from usage embeddings across periods, so the two word-level procedures are not matched evaluations of pair prediction.

3.1.4. Results

Table 2 reports pairwise agreement and word-level ranking correlation; Figure 1 shows the full agreement matrix.

Human–human agreement ranges from 0.688 to 0.740 (mean 0.719), which confirms a reliable but difficult annotation task. On its evaluated items, GLM-4.7-Flash has the highest observed LLM agreement, with a mean of 0.540, followed by the two Qwen3-14B variants (0.503 and 0.510). Three observations follow.

First, performance varies substantially across model families and is not monotonic in parameter count. These are comparisons among differently trained models, not controlled tests of parameter count or language adaptation. At the 27–31B scale, the general-purpose GLM-4.7-Flash (0.540) and Gemma-3-27B-it (0.376) and the Nemotron-tuned GaMS-27B-Nemotron (0.403) all outperform the base GaMS-27B (0.103) by a wide margin. Within the GaMS family, agreement is negative for the 2B and 9B models, 0.187 for GaMS3-12B, and 0.103 for the base 27B model, while the Nemotron-tuned 27B variant reaches 0.403.

Second, thinking mode gives little observed benefit in this comparison. The Qwen3-14B variants with and without thinking mode agree with each other at $\alpha = 0.741$, the highest LLM–LLM pair in the matrix, and reach nearly identical agreement with the human annotators (0.510 versus 0.503).

Third, even the strongest observed LLM–human agreement stays below the human range (0.540 versus 0.688–0.740). The result motivates human review of LLM-generated labels in a semi-automatic pipeline. Because GLM-4.7-Flash covers only 2,439 of 3,150 items, its lead over models evaluated on nearly all items needs a common-item check. On GLM's valid items for the published 104-word set, its mean α remains 0.540; the Qwen3-14B variants reach 0.522 and 0.526. Thus the observed lead persists but is small on this subset.

At word level, the highest prompted-model point estimate is $\rho = 0.806$ for GaMS-27B-Nemotron; GLM-4.7-Flash and Qwen3-14B without thinking each reach 0.775. GaMS-27B-Nemotron illustrates why this result needs care: despite its highest word-level ρ , its mean pair-level α is only 0.403. Averaging roughly 30 ratings per word can preserve the ranking of words even when many individual usage judgments disagree with humans. The lower pair-level agreement gives much less support for using GaMS's ratings to distinguish individual usages. Its high word-level ρ does not establish that it can split them into coherent sense groups for corpus analysis. XL-LEXEME with APD reaches $\rho = 0.768$ against the same 104-word COMPARE ranking (Section 3.2.3). These numbers describe different scoring procedures: prompted ratings

Table 2: Agreement with human usage judgments and correlation with human word rankings. H1–H3 give ordinal Krippendorff’s α between each model and each human annotator; Mean α averages these three comparisons. Word ρ is Spearman correlation between the model’s mean valid rating per word and the human COMPARE ranking for 104 words. Pair-level agreement uses available judgments from the 105-word annotation file. Models are ordered by Mean α ; the human–human rows provide a reference.

Annotator	H1	H2	H3	Mean α	Word ρ
<i>Human–human pairs</i>					
Human H2	0.688				
Human H3	0.740	0.730			
Human mean (3 pairs)				0.719	
GLM-4.7-Flash	0.527	0.542	0.550	0.540	0.775
Qwen3-14B (thinking)	0.476	0.531	0.523	0.510	0.752
Qwen3-14B (no thinking)	0.463	0.525	0.522	0.503	0.775
Qwen3-30B-A3B	0.372	0.470	0.422	0.421	0.727
GPT-OSS-20B	0.434	0.373	0.431	0.413	0.704
GaMS-27B-Nemotron	0.345	0.448	0.415	0.403	0.806
Gemma-3-27B-it	0.302	0.458	0.367	0.376	0.749
GaMS3-12B	0.138	0.235	0.188	0.187	0.408
GaMS-27B	0.032	0.157	0.119	0.103	0.680
GaMS-9B	−0.161	−0.045	−0.103	−0.103	0.110
GaMS-2B	−0.209	−0.159	−0.162	−0.177	−0.018

are averaged over the exact annotated pairs used to build COMPARE, whereas APD uses distances among usage embeddings across periods. The numerical proximity does not establish equivalent performance, and the benchmark does not show that prompting is preferable to the fine-tuned encoder. Missing GLM responses are spread across all 104 words, with at least 11 valid pairs per word. Recomputing the human word ranking on just GLM’s valid pairs gives $\rho = 0.994$ against the full ranking, suggesting that this subset retains the gold-standard order; it does not prove that the missing responses are random.

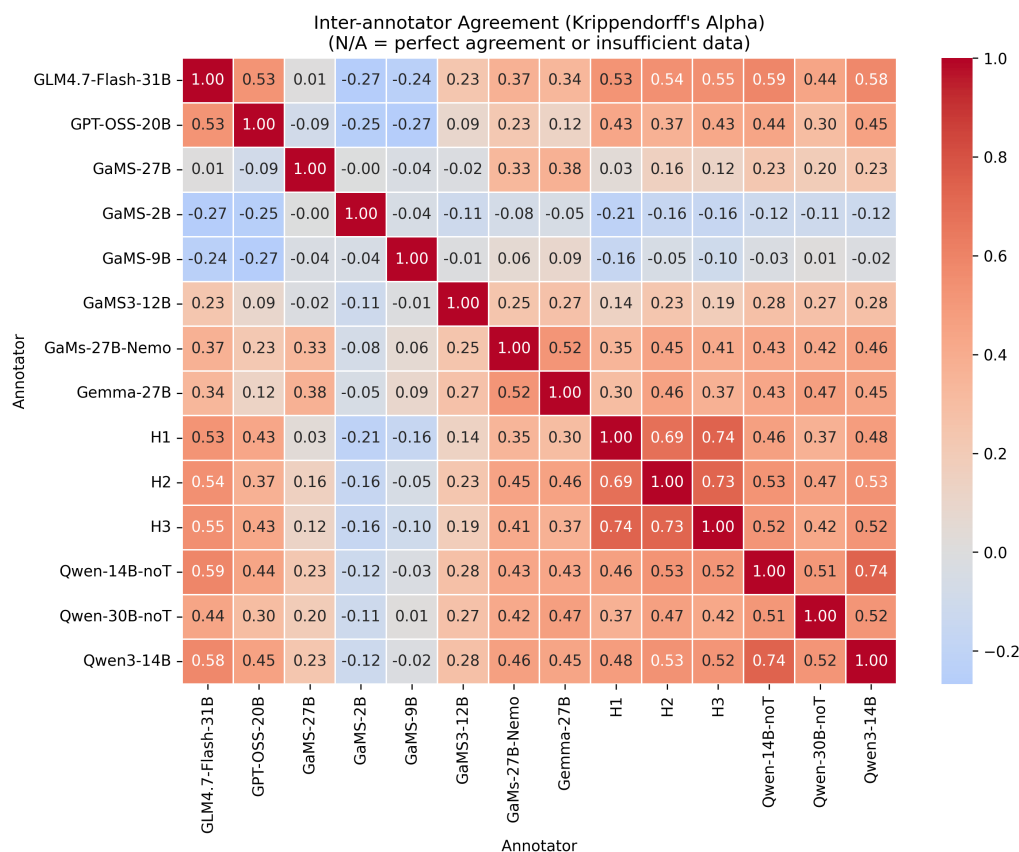


Figure 1: Agreement among three human annotators (H1–H3) and eleven prompted language models. Each cell gives ordinal Krippendorff's α for the annotators named on its row and column, using their shared valid judgments. Higher values indicate stronger agreement; zero indicates chance agreement and negative values indicate disagreement beyond chance. The diagonal shows self-comparisons. Model labels are abbreviated; “noT” denotes prompting without thinking mode.

3.1.5. Instruction-based embedding analysis

In a separate embedding experiment, we test whether instructions help isolate the meaning of a word in context. We encode each usage with Qwen3-Embedding-0.6B, which accepts a natural-language task instruction. We locate the target word in each sentence with a Slovenian lemmatizer, mean-pool its token embeddings, and compare usages by cosine distance. The evidence below is a small exploratory probe, not a corpus-wide validation of this embedding method.

We ran a sweep of 15 instruction variants on examples divided into three sense groups (Animal, Device, Person). For each instruction we compare the average cosine similarity within a sense group, which measures sense consistency, with the average similarity across groups, which measures sense confusion. Table 3 shows six representative variants. The example list and its selection protocol are not documented in this report, so these values should guide further testing rather than support a general claim about instruction quality.

Two results follow. First, the instruction that asks for the specific sense in context produces the largest gap between within-group and across-group similarity (0.925 versus 0.624). Second, instructions that explicitly demand sense separation perform worse on these examples than the plain instruction about meaning in context: the *pushed apart* variant raises across-group similarity to 0.655 and reduces the gap to 0.252. Across-group similarity for the Animal–Person

pair stays between 0.76 and 0.81 under every instruction, consistent with semantic overlap between the two groups. The differences among instructions on this small set did not show that changing the instruction would reliably separate senses in a larger corpus.

Table 3: Six representative instructions from the 15-instruction sweep with Qwen3-Embedding-0.6B. Within and Across give mean cosine similarity between word-occurrence vectors from the same sense group and from different groups, respectively, using the Animal, Device, and Person examples. Gap is Within minus Across: a larger gap indicates stronger separation of the groups in this probe.

Instruction	Within	Across	Gap
“Represent the specific sense of the words as they are used in the context of the sentence.”	0.925	0.624	0.301
“Represent text focusing on literal word senses, distinguishing homonyms, metonyms, and metaphors.”	0.914	0.624	0.290
“Represent text to capture the stable, dictionary-style meanings of its words, minimizing the influence of specific context.”	0.905	0.617	0.288
“Represent each usage so that different senses of the same word are pushed apart, even at the cost of reducing similarity within the same sense.”	0.907	0.655	0.252
“Represent the semantic concept of the words in the text.”	0.889	0.617	0.273
“Represent the word usage so that different senses of the same word are clearly distinguished in embedding space.”	0.886	0.673	0.213

3.1.6. Discussion and limitations

The benchmark shows that prompted models can provide graded usage judgments, although their pair-level agreement remains below human agreement. Their word-level ratings also track the human ranking, but the evaluation does not establish a head-to-head advantage over an encoder. Performance varies by model family; increasing size within the tested GaMS models does not guarantee improvement, and the two tested Qwen3-14B modes perform similarly. The instruction probe suggests that a plain context-sense instruction is more useful than explicitly demanding separation of senses on these examples. Since the probe did not establish a reliable gain from instruction wording, we did not pursue further instruction sweeps. Instead, we focused on representations trained to distinguish word usages and on methods for comparing them across time. The dispersion analysis in Section 3.2.1 gives a more specific reason to improve representations and how they are compared.

The study covers one language and a four-point ordinal scale. Responses without a parseable 1–4 score, including *cannot decide*, are missing from that model’s evaluated set; this can exclude genuinely hard items. The GLM-4.7-Flash run covers 77% of the dataset. Prompt variants were compared informally during development rather than in a controlled sensitivity study. These limits make the benchmark a diagnostic of the tested systems, not a general verdict on all prompting strategies.

3.2. APD metric analysis and optimal transport based semantic change detection

We report the distance-measure analysis in our paper “Tracking Semantic Change in Slovene: A Novel Dataset and Optimal Transport-Based Distance” (Pranjić et al., 2026), accepted in *PeerJ*.

In that work, we formally analyze APD and propose a semantic change measure based on regularized optimal transport.

3.2.1. Formal analysis of the APD metric

We analyze the Average Pairwise Distance (APD), a strong baseline that averages the distances between all cross-period pairs of contextual usage vectors. With cosine distance, APD reduces algebraically to $1 - \bar{p} \cdot \bar{q}$, where each barred vector is the average of individually normalized usage vectors in one period. This is a dot-product expression, not the cosine distance between the two averages: averaging generally makes their norms smaller than one. This formulation has two consequences. First, APD can be computed in linear rather than quadratic time. Second, it exposes a dispersion-driven bias: the centroid norm shrinks as usage vectors spread around the mean direction, so words with diverse but stable usage (for example, stable polysemy) receive inflated change scores. Controlled experiments with synthetic von Mises–Fisher distributions confirm the prediction: APD conflates within-period dispersion with distributional difference and is blind to structural rotation of a distribution.

This dispersion need not represent a new sense. A word may retain the same mix of meanings in both periods while its usages occupy a broad area of the embedding space; ordinary contextual, syntactic, or orthographic variation can also spread vectors that refer to the same meaning. APD then assigns a high cross-period distance even when the period distributions are unchanged. The analysis tests whether this is visible beyond the synthetic example: an APD variant that pools usages across both periods, discarding the temporal contrast, still correlates with the human change ranking at $\rho = 0.406$. That result indicates that time-independent usage diversity carries a substantial part of the raw APD signal. The quartile analysis below identifies the resulting ranking error more directly.

3.2.2. Regularized optimal transport

To address the bias, the study compares the empirical distributions of usage embeddings with the cost of optimal transport, using cosine distance between individual usages as the transport cost and entropic regularization for tractability. The regularization parameter interpolates between a strict point-to-point matching of usages ($\lambda \rightarrow 0$) and the uniform averaging of APD ($\lambda \rightarrow \infty$), where the transport plan becomes uniform and the cost reduces exactly to APD. Compared with clustering-based pipelines, the approach avoids discretizing the embedding distribution, and it distinguishes distributions whose overlap is negligible, a regime in which the Jensen–Shannon distance saturates.

3.2.3. Results

With SloBERTa embeddings, ranked against the COMPARE gold standard by Spearman's ρ , APD is the strongest baseline (0.594), ahead of clustering-based methods (JSD 0.458, Wasserstein distance 0.523) and static embedding baselines (up to 0.527). Regularized optimal transport improves over APD, with the best result at $\lambda = 0.1$ ($\rho = 0.648$) and stable performance across regularization values. A word-level analysis explains the gain: the rank error of APD relative to COMPARE grows monotonically with word dispersion (from -15.9 to $+18.4$ rank positions across dispersion quartiles), and optimal transport reduces this dispersion-related deviation. Validation with XL-LEXEME, a model fine-tuned for word-in-context comparison (Cassotti et al., 2023), shows a different pattern: APD reaches 0.768 on the Slovene COMPARE ranking and 0.872 on an English benchmark based on annotated sense distributions. The two scores use different reference measures. Regularized transport matches or slightly exceeds APD only at higher regularization values. The relative performance of the two measures therefore depends on the geometry of the embedding space. The dataset and code are publicly available (Section 5).

The formal APD analysis establishes that dispersion can inflate change scores even when the usage distributions are unchanged across periods. General-purpose models represent semantic, syntactic, and other contextual information; variation among their word-occurrence

vectors can therefore extend beyond differences in the target word's meaning. This concern applies to representations from both generative and encoder models. Our empirical encoder analysis shows that dispersion has a practical effect on the resulting word rankings, motivating representations trained specifically for usage similarity.

The encoder comparison supports this direction. On the same Slovene word-level ranking task, replacing SloBERTa with the usage-comparison model XL-LEXEME raises APD from $\rho = 0.594$ to 0.768; with XL-LEXEME, regularized OT reaches 0.769. The encoder choice has a larger observed effect than switching from APD to OT in these experiments. OT helps correct a specific dispersion bias in the SloBERTa space, while a representation trained to distinguish usages makes the simpler APD measure highly competitive. This is consistent with a useful explanation: word-in-context fine-tuning can make semantic relatedness more prominent relative to incidental contextual variation, so APD's sensitivity to vector spread becomes less damaging. We did not directly measure whether XL-LEXEME has lower dispersion than SloBERTa, and the models also differ in architecture and training data; this remains an interpretation of the result, not a controlled causal finding. Even with that limit, the much larger encoder effect motivates improving the representation before adding more complex comparison machinery. Figure 2 shows both the dispersion pattern and the encoder comparison. These ρ values assess rankings of words and are not directly comparable with the annotator α values in Section 3.1.

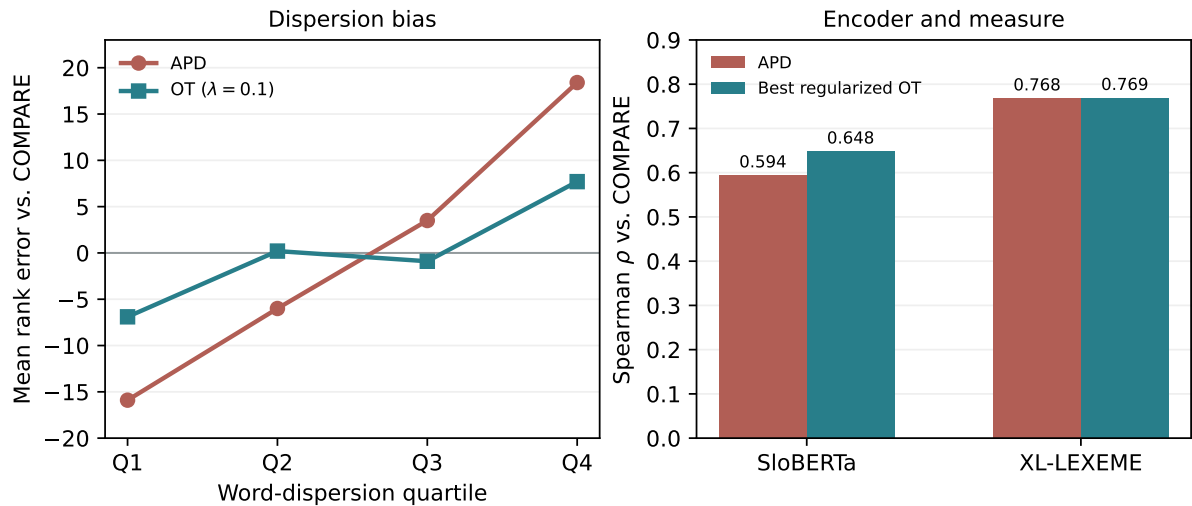


Figure 2: Dispersion bias and encoder performance on the Slovene gold standard. Left: mean rank error relative to the human COMPARE ranking for SloBERTa, with words divided into dispersion quartiles (Q1: lowest; Q4: highest). Positive values indicate overestimated change, negative values underestimated change, and zero agreement with the reference rank. Regularized OT reduces the dispersion-related trend seen with APD. Right: Spearman correlation with COMPARE for APD and the best regularized OT setting ($\lambda = 0.1$ for SloBERTa; $\lambda = 2.0$ for XL-LEXEME). Higher bars indicate closer agreement with the human word ranking. Results are from our accepted PeerJ paper.

3.3. Scaling word usage analysis to the corpus level

The preceding results favor encoders trained to compare word usages. Applying them to a large corpus, however, is costly: XL-LEXEME (Cassotti et al., 2023) and XL-DURel (Yadav and Schlechtweg, 2025) mark each target before encoding, which requires a separate encoder pass for every target occurrence. They also store a full-width vector for each usage. We address the cost of encoding and storage in the manuscript “Compact Contextual Word Usage Representations via Decoupling Context Encoding from Target Selection”, currently in preparation.

We evaluate the representations on CoMeDi (Schlechtweg et al., 2025), a benchmark of more than 71,000 usage pairs from seven languages. It uses DUREL-style graded relatedness judgments, as does the Slovene benchmark in Section 3.1. The train, development, and test splits contain 47,833, 8,287, and 15,332 pairs, respectively, with lemmas separated across splits. Each pair supplies the target spans and a median 1–4 relatedness judgment. The primary evaluation is Spearman correlation between predicted similarity and those judgments. We calculate it within each lemma and average it across lemmas and languages. This tests whether a representation distinguishes usages of the *same* word. We also report ordinal Krippendorff's α and a language-pooled Spearman score for comparison with XL-DUREL.

3.3.1. Decoupling context encoding from target selection

We encode the unmodified text once and then average the contextual states of the tokens in each target span. Target spans are known in this experiment. For long contexts, we select overlapping 128-token windows from a grid whose boundaries do not depend on the target. Several targets in one window can thus share a single encoder pass. We compare three configurations: marked input with sequence-wide pooling (MARK), marked input with target-span pooling (MARK+SPAN), and unmodified input with target-span pooling (SPAN). Comparing MARK with MARK+SPAN isolates the effect of target-span pooling; SPAN allows encoded windows to be reused across targets.

3.3.2. Compact usage representations

We compare four ways to shorten usage vectors from XLM-R-large (Conneau et al., 2020) fine-tuned with the CoSENT ranking objective: direct training, Matryoshka Representation Learning (MRL; Kusupati et al., 2022), PCA projection, and truncation. Direct training produces one model per width. MRL trains several nested prefixes in one model, allowing a shorter vector to be selected after training. PCA projects full-width vectors; truncation keeps their first coordinates. These methods reduce storage and comparison costs, but they do not reduce the encoder computation for one window.

3.3.3. Results

At 128 and 256 dimensions, MARK+SPAN improves mean Spearman correlation over MARK by about 0.011 on development and 0.009 on test data. The bootstrap intervals exclude zero, supporting a benefit from target-span pooling. The intervals for the difference between SPAN and MARK+SPAN include zero. This comparison also cannot isolate the effect of removing target markers because the two configurations use different window selection rules.

Figure 3 shows how development-set correlation changes with vector width. At 128–256 dimensions, the representations match or exceed their full-width counterparts. An 8-dimensional MRL prefix retains 85% of the 256-dimensional score. Compared with full-width vectors, 128–256-dimensional vectors use one-quarter to one-eighth of the storage and speed up exact similarity computation by up to 7.78x on CPU and 3.24x on GPU. The selected 128–256-dimensional configurations also exceed the published CoMeDi-only point estimates of XL-DUREL on both reported metrics (Spearman's ρ 0.605–0.617 versus 0.600, ordinal Krippendorff's α 0.638–0.643 versus 0.622). MRL also allows several vector widths from one model; direct training needs a separate model for each width.

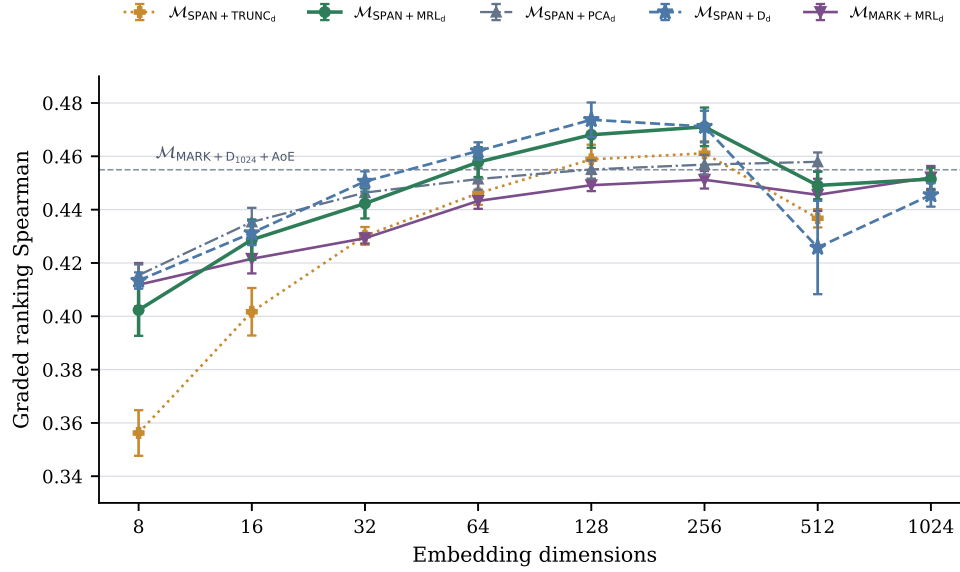


Figure 3: Usage-similarity quality on the CoMeDi development set as vector width changes. Spearman correlation is computed within each target lemma, then averaged across lemmas and languages. Points show means over five runs; error bars show one standard deviation. SPAN pools the target word’s subtokens from unmarked text; MARK uses target markers. TRUNC keeps the first vector coordinates, PCA projects full-width vectors, MRL trains nested vector prefixes, and D trains a separate model at each width d . The horizontal dashed line is the 1,024-dimensional marked reference trained with the angle-optimized embedding (AoE) objective. Adapted from the scaling manuscript in preparation.

With one target per window, SPAN reaches 0.93 times the extraction throughput of MARK. When two, four, or eight targets share a window, reusing its encoding speeds up end-to-end extraction by 1.72x, 3.23x, or 6.05x, respectively (Figure 4). This test used 128 development contexts on an NVIDIA A100 with encoder batch size 32. Corpus windows with fewer shared targets will yield smaller gains. For one million float32 vectors, storage falls from 3.815 GiB at 1,024 dimensions to 0.954 GiB at 256 and 0.477 GiB at 128. Encoder reuse reduces extraction cost; shorter vectors reduce storage and the cost of similarity comparisons.

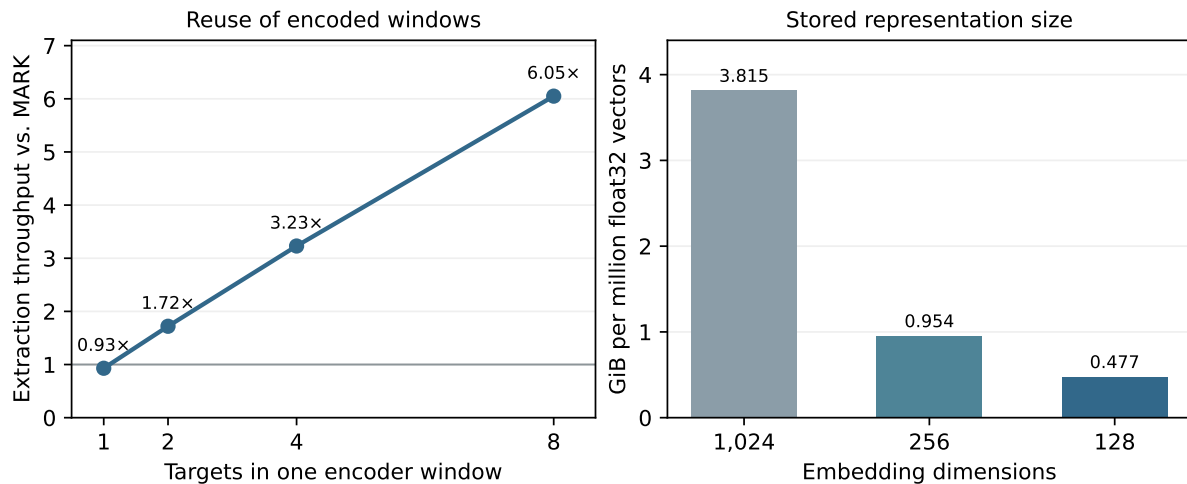


Figure 4: Encoding reuse and compact vectors reduce different costs. Left: the ratio of SPAN to MARK extraction throughput as the number of target occurrences in one window increases. SPAN reuses one encoding; MARK encodes each target separately. A ratio of 1 indicates equal throughput. Measurements use 128 development contexts on an NVIDIA A100, with batch size 32. Right: storage in GiB for one million float32 word-occurrence vectors at 1,024, 256, and 128 dimensions; labels give the calculated storage sizes.

3.3.4. Diachronic change diagnostic

As a downstream check, we rank lemmas by semantic change using the average cosine distance between cross-period usage pairs, as in the APD analysis in Section 3.2.1. Across 93 lemmas in six languages, the compact span-pooled models raise the macro Spearman correlation with graded change scores from 0.583 for an XL-DURel-style reference to 0.675–0.680. This result is exploratory because the benchmark samples cross-period pairs rather than covering every corpus occurrence.

Reusable windows reduce repeated encoding; compact vectors reduce storage and comparison cost. Together, they support analysis of many targets and usages in large corpora. The next section describes how we group and trace those usages through time.

3.4. Building timelines of word usage clusters

The approach organizes the steps from dated documents to target occurrences, reusable embeddings, context families, and timelines. Each step remains queryable or inspectable, so the analysis can be rerun for another target or a selected part of the corpus. Section 4 presents applications of the approach.

3.4.1. Corpus preparation and occurrence extraction

The approach first cleans the news text, preserving offsets into each article body, and stores the articles with their dates. Explicit target rules map inflected and multilingual forms to one *concept* identifier. We extract target occurrences from article bodies. In exploratory analyses, embeddings of target occurrences in titles consistently clustered together, separately from those in article bodies. This suggests that the short, compressed phrasing of headlines can influence the representations enough to organize clusters by text type rather than by the target's use. Article bodies generally provide fuller syntactic and semantic context for the target word, making them better suited to grouping usages by meaning in this analysis. Restricting occurrence extraction to article bodies reduces this source of variation and makes the resulting groups easier

to interpret. Titles can still supply context for reading an article, but title occurrences are excluded from clustering.

The cleaned articles, their dates and publication metadata, and the extracted target usages are kept in a queryable DuckDB SQL database. This makes the same corpus useful beyond one all-news timeline. A researcher can select a range of years or windows around an event, restrict the analysis to one publisher, or compare publishers while controlling for language, country, outlet type, or topical section where those metadata are available. Adding the geographic location of each publisher would also allow researchers to compare when and where a context family appears across regional outlets, for example whether a climate framing appears earlier in one region than another. Publisher location would identify where an outlet is based, not where the events it reports occurred. The selected occurrences can then be grouped and plotted on their own, or compared with another selection. The database also supports repeating the analysis for other target words and phrases without treating each case study as a new corpus collection.

Deduplication matters for these sources. News collections can contain repeated advertisements, syndicated stories, boilerplate, and quotations. The procedure identifies repeated text without classifying the source of each duplicate. Repetition can itself be historically informative, but counting every copy as independent evidence of a distinct use would overrepresent that context in the clusters and create peaks driven by republication. The approach therefore tracks distinct contextual evidence rather than circulation frequency. The deduplication step first removes articles with identical title and body, keeping the earliest dated copy. It then compares occurrences of the same target concept before embedding: a window of ten words on each side of the target is extracted, with the target string excluded. The two sides are normalized for case, Unicode and common OCR artifacts such as soft hyphens and line-break hyphenation. Token-bigram and character-fivegram MinHash signatures select candidate repeats; a RapidFuzz character-similarity ratio of at least 80% confirms a match. The candidate search is approximate, so the procedure can miss some repeated passages. Only the earliest matching occurrence remains canonical and receives an embedding; later matches retain a link to it but are excluded from the timeline analysis.

For each retained occurrence, the approach selects a window of at most 128 model tokens around the target. It encodes each selected window once, including when several targets appear in the same window. The final-layer vectors of subtokens overlapping the target span are averaged and L_2 -normalized. This is the reusable span-pooling approach evaluated in Section 3.3: one article window can yield representations for multiple target occurrences without re-marking and re-encoding the text for each one.

3.4.2. Two-stage clustering of usages

The approach clusters the occurrence vectors across the *whole* observation period, rather than fitting separate clusters to each year. This gives a shared set of groups whose counts can be compared through time. Before clustering, it removes the overall vector mean and the means of frequent target forms, then applies partial PCA whitening and renormalizes. This step reduces variation caused by surface form and dominant embedding directions. Clustering has two stages. First, it connects occurrences that appear in each other's 26 nearest neighbours by cosine similarity. Repeated Leiden community detection (Traag et al., 2019) on this reciprocal-neighbour graph, with consensus edge weights, finds locally coherent microcommunities along with isolated and tiny fragments. This stage preserves tight local relationships, but graph connectivity can split examples of one broader context into separate communities.

Second, it averages the occurrence vectors within each microcommunity and normalizes the resulting centroid. For communities with multiple members, averaging reduces variation from individual sentences and provides a common unit for comparing groups. This motivates the merge

step. Singletons remain individual vectors. The approach computes cosine distances between the microcommunity centroids and applies average-linkage hierarchical clustering. The hierarchy merges close small communities, but stops before a merge would combine two well-supported cores (each containing at least 2% of sampled occurrences). Thus nearby microcommunities can become a broader context family even when Leiden did not place them in the same graph community. The merged label is assigned back to every member occurrence, and representative passages nearest each final group's centroid support manual interpretation. We call these groups "context families": they describe recurring uses in this corpus, not necessarily distinct dictionary senses.

3.4.3. Constructing and interpreting timelines

The approach assigns each retained embedded occurrence to one group fitted across the whole observation period and counts the dated assignments for each calendar day. Missing days receive zero counts. Daily counts can be smoothed to make temporal patterns easier to inspect. Annual shares divide each group's yearly count by the number of all retained target occurrences in that year. Groups are ordered by total size; if only the largest groups are displayed, their shares need not sum to 100%.

Separately from plotting, the approach quantifies how much the context-family mix depends on time. It bins occurrences by quarter, builds a time-by-family count table with Laplace smoothing, and measures normalized mutual information between time and group. It compares this value with 100 shuffled-label null runs and gates the ranking score on sample size, active periods, and diversity of supported groups. The resulting score ranks candidate concepts for review; the plotted counts and shares describe the temporal distribution of usages. A peak is interpreted by inspecting its dated passages and comparing them with external events, not by the score alone.

4. Use cases

We describe two applications at different stages of methodological development. The poverty study used an earlier prototype, while the smaller climate-change corpus served as a pilot for developing and testing the revised methodology before further analysis of the larger historical corpus.

4.1. Poverty in SPeriodika

We applied an earlier, work-in-progress version of the approach to poverty-related discourse in sPeriodika 1.0, a corpus of Slovene historical periodicals published in 1771–1914 (Dobranič et al., 2023). The analysis examined words such as *ubog*, *revež*, *sirota*, and *siromak*, grouping their usages and visualizing the groups through time. We restricted the working collection to text marked as having "good" OCR quality in the corpus metadata. This collection contains approximately 30 million sentences, with most of the material concentrated in the nineteenth and early twentieth centuries. Its size, remaining OCR noise, and historical spelling variation make it a challenging setting for usage analysis. The presented timelines focus on approximately 1845–1915, a period with substantial corpus coverage.

This prototype used a standard SentenceBERT model (Reimers and Gurevych, 2019), `paraphrase-multilingual-MiniLM-L12-v2`² without fine-tuning for word-in-context similarity. Each representation was a contextual vector for a target word occurrence, obtained by pooling the model's subtoken vectors for that word. It included target occurrences in both titles and article bodies. Preprocessing removed identical sentences, but did not detect near-duplicate passages.

²<https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

In OCR-derived text, copies of the same original passage often differ because of minor recognition errors and therefore survive exact matching. Clustering consisted of community detection alone, without the subsequent hierarchical merging of community centroids described in Section 3.4.2.

Precomputed representations covered approximately 300 million words, allowing us to analyze additional targets without encoding the corpus again. Similarity search connected usages in a graph, and community detection identified groups of related contexts. KeyBERT (Grootendorst, 2021) extracted keywords and short phrases from each group to support interpretation.

The groups for *ubog* ('poor') distinguished contexts involving religious institutions and devotion, peasants, workers, charitable support, and animals. The curves for the religious groups peak earlier in the observation period and fall thereafter, while the worker-related curve rises toward the end of the period. Figure 5 illustrates these contrasting patterns alongside a group involving animals. Groups for *revež* ('a poor person') included contexts of misfortune, peasants and taxation, church support, and railway accidents. These examples show that poverty-related vocabulary also appears in expressions of sympathy and misfortune, beyond material deprivation. The keyword summaries and timelines offered starting points for examining these contexts in their source passages.

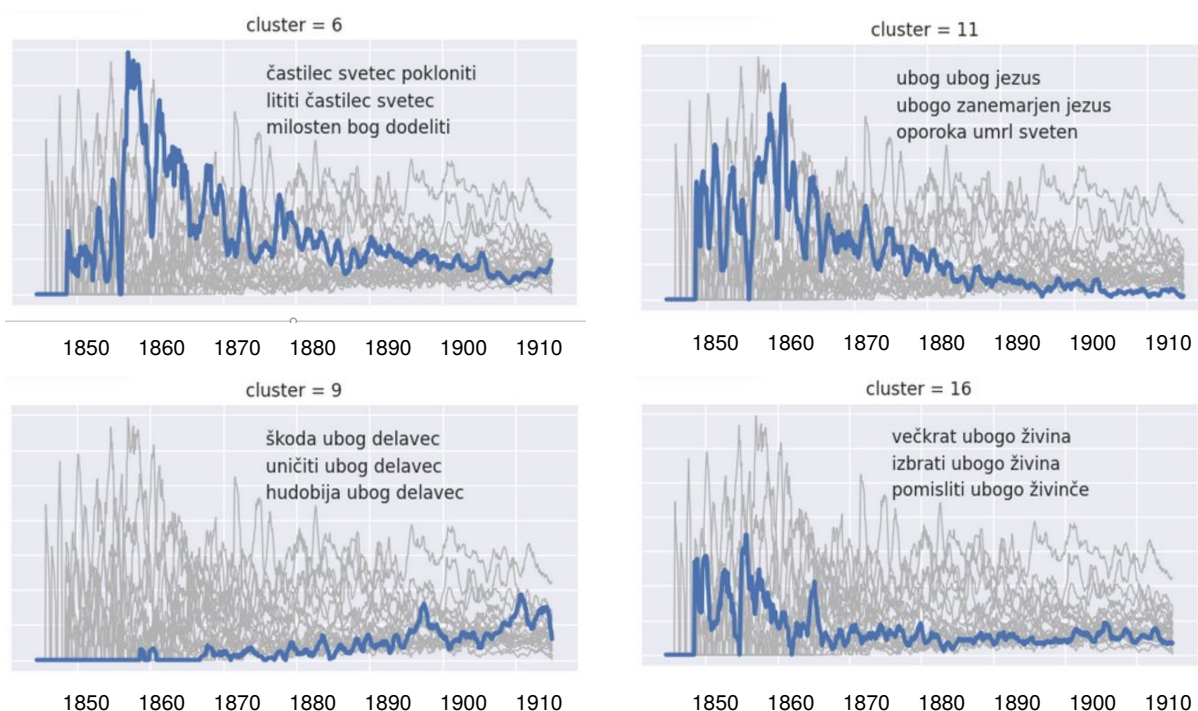


Figure 5: Selected usage groups for *ubog* in the SPeriodika prototype. Each panel highlights one group in blue against the other groups in grey, with automatically extracted keyword phrases. The upper panels show religious contexts (groups 6 and 11); the lower panels show workers (group 9) and animals (group 16). Time runs from left to right across the study period, approximately 1845–1915. The religious-group curves peak earlier and fall, while the worker-related curve rises later. The panels are reproduced from presentation screenshots, with year labels restored from the original timeline. They illustrate temporal patterns; the vertical scale is omitted.

The poverty study helped us identify limitations of the prototype and informed the subsequent development of the method. Repeated passages with minor OCR differences could contribute multiple counts, and title occurrences formed separate clusters. Community detection produced more than a hundred clusters for a target word, leaving related contexts in separate small groups and making their timelines difficult to interpret together. This fragmentation motivated the second clustering stage, which compares community centroids and merges related groups. Together with

the representation analysis in Section 3.3, these observations motivated near-duplicate detection, extraction of occurrences from article bodies, representations trained for usage similarity, and hierarchical merging of community centroids. The climate-change pilot below examines the revised approach before its further application to SPeriodika.

4.2. Global warming in Slovenian news

As a pilot of the revised corpus-scale method, we examine the contexts in which “global warming” and its equivalents appear in news from 2005–2024. The collection contains 44,909 dated articles after exact article deduplication; 30,879 have a retained target occurrence. The analyzed set contains 41,169 occurrence embeddings after context deduplication (Section 3.4.1). Metadata for these occurrences identifies 16,432 as Slovene, 16,192 as Serbian (including script and regional variants), 3,482 as Croatian, 3,123 as Bosnian, and 1,865 as Macedonian; 75 carry other language tags. The study tests whether the approach can produce an interpretable diachronic map over this multilingual collection. It is one application of the method, not a gold-standard evaluation of semantic change.

4.2.1. Corpus preparation and clustering results

The pilot uses a multilingual contextual encoder built on XLM-R-large and the preparation and clustering procedures in Section 3.4.

After exact duplicate articles were removed, the use case indexed 57,973 target occurrences. The context-level procedure marked 16,804 (29%) as near duplicates, leaving 41,169 retained occurrences for embedding and analysis (Figure 6). The effect grows with the volume of collected news, making this step consequential for the later timelines.

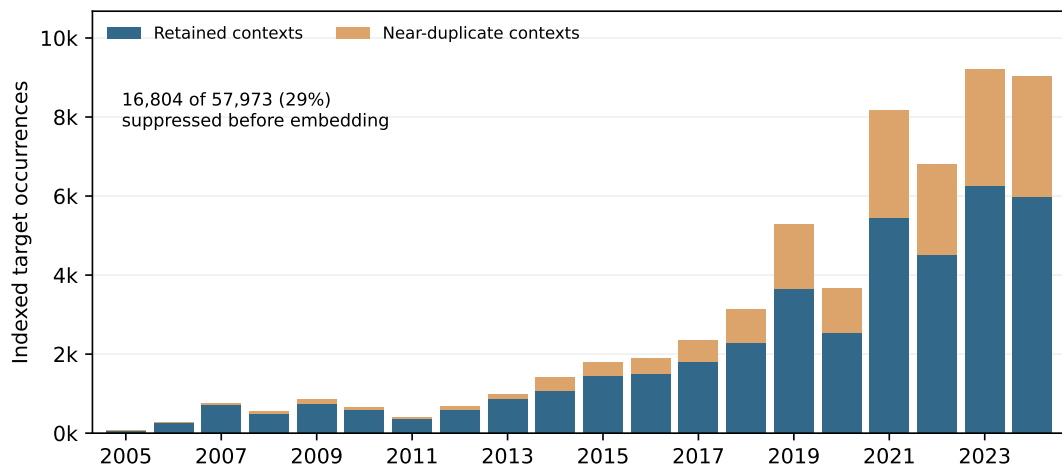


Figure 6: Effect of context deduplication on “global warming” occurrences in the collected news corpus. Each stacked bar shows the number of target occurrences in one year, after identical articles were removed. Blue shows occurrences retained for embedding and clustering; orange shows occurrences excluded because their surrounding text closely matches an earlier occurrence. Across all years, 16,804 of 57,973 occurrences (29%) were excluded, leaving 41,169 for analysis.

In the “global warming” use case, Leiden produced 389 communities, including 279 singletons; centroid-based merging reduced them to 36 groups (Figure 7). The ten largest are displayed in the report. Some broader groups only become apparent after the second stage: the second-largest group combines 67 Leiden communities, and the seventh-largest combines 77. Figure 7

demonstrates how much the two-stage procedure consolidates the graph communities. The pilot does not independently test centroid stability. Counts alone do not show that the merged groups are semantically coherent; we examine representative passages and a small exploratory passage sample.

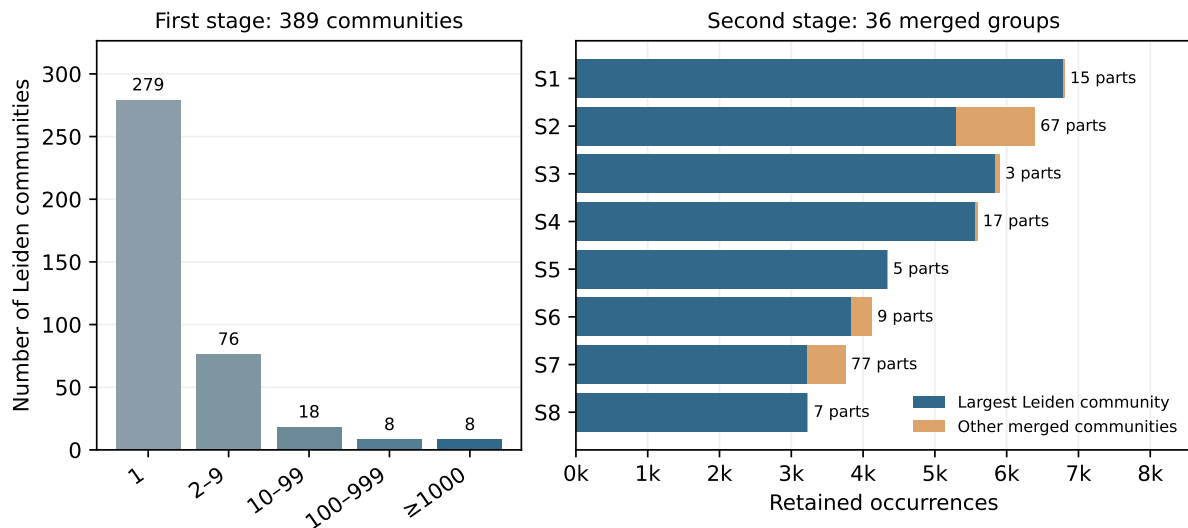


Figure 7: How two-stage clustering groups the 41,169 retained “global warming” occurrences. Left: the number of Leiden communities in each size range, measured in occurrences. Of 389 communities, 279 contain only one occurrence. Numbers above the bars give community counts. Right: the eight largest groups after hierarchical merging reduces the 389 communities to 36 groups. Each bar’s length gives the number of occurrences in the final group: blue shows those from its largest original community, and orange shows those from the other communities merged into it. The “parts” label gives the total number of original communities in that group; for example, S7 combines 77 communities.

4.2.2. Context families and temporal patterns

The analysis displays ten groups. Eight large groups account for most analyzed occurrences. Inspection of passages near their centroids suggests themes of consequences for ecosystems and health (S1, 16.5%), general responses (S2, 15.5%), scientific explanation and debate (S3, 14.3%), emissions and technical mitigation (S4, 13.6%), wider social and ideological discourse (S5, 10.5%), political action (S6, 10.0%), Paris Agreement targets and temperature thresholds (S7, 9.1%), and extreme-weather contexts (S8, 7.9%). These are working descriptions, not exclusive sense labels. The two small remaining displayed groups are ambiguous residual contexts (S9, 0.5%) and noncanonical uses such as titles or wordplay (S10, 0.3%). Shares are rounded and describe this analyzed corpus; smaller, undisplayed groups account for the remainder.

The temporal view makes the grouping useful for historical interpretation. Among the eight recurring families, S7’s share of retained target occurrences rises from 6.3% of 4,739 occurrences in 2015–2017 to 11.5% of 16,743 in 2022–2024, while S3’s share falls from 16.9% to 11.9%. These values were recalculated from the saved cluster labels and dated group assignments. Figure 8 shows annual shares and the period comparison. All major groups remain present: the result is a changing mix of assigned groups within the collected articles rather than a single abrupt replacement of meaning. These shares control for the total number of retained target occurrences in each year, but not for changes in the publishers or languages represented in the collection.

Representative passages make three labels concrete. A September 2024 S1 article connects loss of sea ice with hungry polar bears; an August 2021 S3 article discusses IPCC forecasts and climate science; and a November 2022 S7 article discusses whether the 1.5-degree target remains attainable. These are paraphrases of passages selected near the group centroids in the pilot analysis. Figure 9 shows the S7 timeline, using a centered 28-day rolling mean of daily occurrence counts and weekly averages of the smoothed values. The timeline has peaks near dated climate policy events. Their alignment is useful for reading the group, but does not show that the events caused the peaks.

We also drew a blinded sample of 32 passages: two from each group's largest Leiden community and two from its other merged communities, for S1–S8. In an exploratory first pass, the primary theme assigned from a short passage agreed with the stored group label for 11 of 32 items. The disagreements include overlapping themes and fragments with too little context, but also show that the group names do not reliably describe every member. The sample is too small and too deliberately balanced to estimate corpus-wide accuracy. The timelines below therefore track algorithmic group assignments, which should not be read as clean counts of mutually exclusive discourse frames.

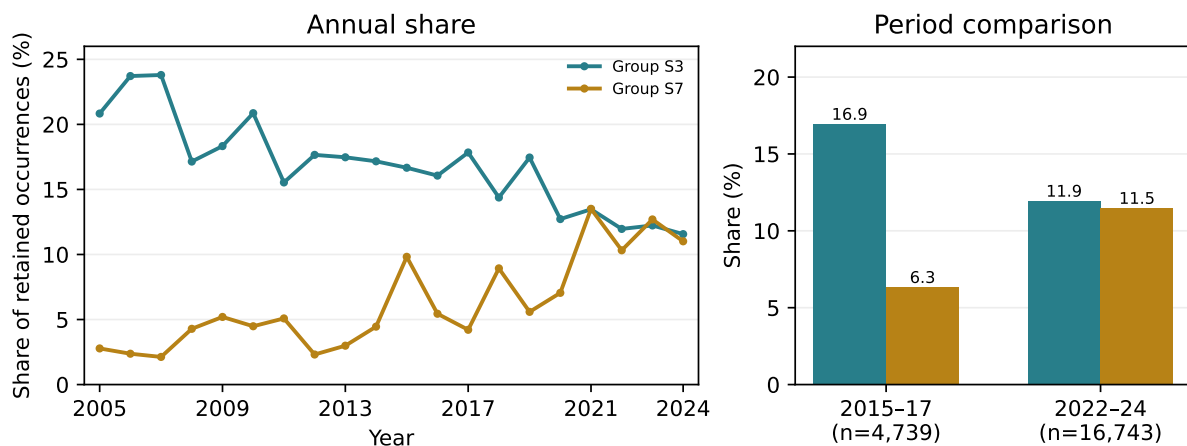


Figure 8: Percentage of retained “global warming” occurrences assigned to S3 (science and debate; teal) and S7 (policy targets; ochre). Left: annual percentages, using all retained target occurrences in each year as the denominator. Right: percentages computed from counts pooled within 2015–2017 and 2022–2024. Here n is the total number of retained target occurrences in each period, across all groups. These percentages account for varying numbers of target occurrences, but not changes in the publishers or languages represented in the corpus.

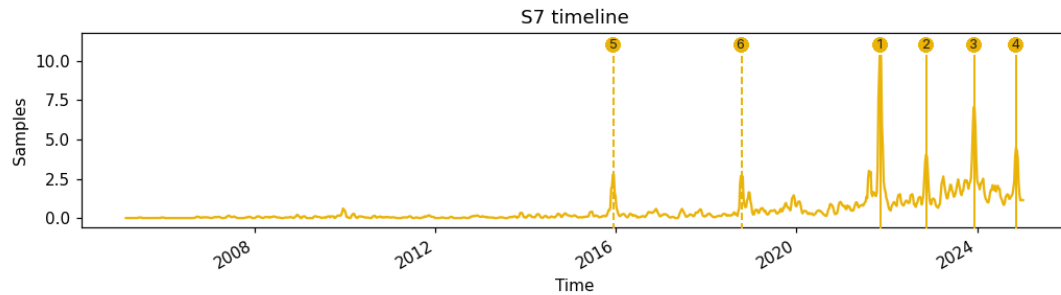


Figure 9: Timeline of retained “global warming” occurrences assigned to S7, a group often concerning climate-policy targets and temperature thresholds. The curve shows weekly averages of daily occurrence counts smoothed with a centered 28-day rolling mean; values are not normalized by corpus size. Numbered lines mark selected dates: (1) COP26, 8 November 2021; (2) COP27, 14 November 2022; (3) COP28, 4 December 2023; (4) COP29, 11 November 2024; (5) Paris Agreement adoption, 12 December 2015; and (6) the IPCC report on 1.5 °C warming, 8 October 2018. These dates were selected after inspecting the timeline to help interpret its peaks.

4.2.3. Source checks and interpretive limits

The same group assignments can be compared within source subsets. Table 4 repeats the 2015–2017 versus 2022–2024 comparison by language and for publishers with at least ten retained target occurrences in each period. S7 rises in every listed language, though the early Bosnian and Croatian samples are small. S3 falls in four languages and is nearly unchanged in Croatian. The shared-publisher subset shows the same directions. Holding the early-period language mix fixed gives S7 6.3% to 11.3% and S3 16.9% to 11.9%. If each publisher’s late share is weighted by its early-period occurrence count, holding the publisher mix fixed, S7 rises from 6.2% to 9.5% and S3 falls from 18.4% to 11.7%. The shifts are smaller than the unadjusted S7 comparison but are not explained away by the changing publisher mix in this panel. Resampling these 75 publishers 3,000 times gives a 95% interval of +1.5 to +4.9 percentage points for S7’s fixed-mix change and –10.1 to –3.7 points for S3’s. These intervals describe variation across the included publishers, not uncertainty from clustering or corpus selection.

Table 4: Percentage of retained “global warming” occurrences assigned to S3 (science and debate) and S7 (policy targets), by source subset. The two n columns give all retained target occurrences in each period; each percentage cell shows 2015–2017 → 2022–2024. Serbian includes its script and regional metadata variants. The final row restricts the sample to 75 publishers with at least ten target occurrences in each period; its pooled shares can still reflect changes in the relative volume of those publishers.

Subset	n_{15-17}	n_{22-24}	S3 (%)	S7 (%)
All sources	4,739	16,743	16.9 → 11.9	6.3 → 11.5
Slovene	2,028	4,820	15.0 → 11.7	5.5 → 10.5
Serbian	1,910	7,208	18.7 → 11.8	5.9 → 11.9
Croatian	249	1,727	15.3 → 15.6	10.4 → 11.5
Bosnian	106	2,012	13.2 → 10.0	10.4 → 12.0
Macedonian	442	919	19.9 → 12.0	8.4 → 11.5
Shared publishers	2,778	5,042	18.4 → 11.2	6.2 → 10.5

News volume, source coverage, language mix, and OCR quality vary over time, so these checks cannot establish a change in population prevalence. Since repeated contexts are assigned to their earliest copy, later republication does not contribute a new count; the curves must not be

read as raw phrase frequency. Approximate matching can also merge contexts with small but meaningful differences. Group labels are working interpretations of representative passages; the exploratory sample above shows substantial overlap, S2 remains broad, and S9 has no clear single interpretation. The top-ten group shares omit smaller clusters, and the S7 time series uses unadjusted counts. The ranking score for this concept is 0.0136, classified as low; the measured normalized mutual information between quarter and group is 0.0178. Thus the selected peaks and share changes do not imply a large corpus-wide semantic shift. The pilot shows that the system can recover dated, inspectable groups of contexts; interpreting the group timelines requires further validation.

5. Availability

The peer-reviewed study “Tracking Semantic Change in Slovene: A Novel Dataset and Optimal Transport-Based Distance” is accepted in PeerJ.

The manuscript “Compact Contextual Word Usage Representations via Decoupling Context Encoding from Target Selection” is in preparation.

The annotated dataset is published via CLARIN.SI at <http://hdl.handle.net/11356/1651>. The code of the PeerJ study is available at <https://github.com/sharpsy/slovene-OT-semchange>. Code for the remaining experiments is currently available only to project partners for internal use. Public release of the code and results is expected upon publication of the paper describing the methodology and its evaluation.

6. Conclusion

T4.2 began with a prompting-centered plan and developed into a methodology for reliable, scalable analysis of contextual word usage. The LLM benchmark shows that prompted judgments carry semantic information but remain below human agreement on the tested Slovene pairs. Their aggregated word rankings are informative, but the available comparison does not establish that prompting matches the fine-tuned encoder. The Slovene gold standard and APD analysis expose a more specific obstacle: dispersion of contextual vectors can look like semantic change. Regularized OT partly corrects that bias for SloBERTa, while usage-aware XL-LEXEME representations produce a larger observed gain than changing the distance measure alone. The scaling study then shows how to extract reusable target-span representations and retain quality with 128–256-dimensional vectors.

The corpus-scale approach connects those representations to dated, inspectable context families: it keeps each usage linked to its source in a queryable database, suppresses repeated contextual evidence, groups usages through microcommunity detection and centroid merging, and constructs timelines from whole-period assignments. The multilingual “global warming” study pilots the revised method ahead of further analysis of the larger corpus on poverty-related discourse, where an earlier prototype was already applied. The climate pilot’s groups separate scientific, political, technical, and impact contexts imperfectly, and their assigned shares change over the observed period. The main directions persist within languages and a shared-publisher panel, but group labels and corpus coverage limit the interpretation. Together, the work joins reliability testing, dispersion-aware comparison, efficient representation, and contextual interpretation. The methodology will be further evaluated in the use cases in WP5.

References

- Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual lexical semantic change. In *Proceedings of ACL*, pages 1577–1585. <https://aclanthology.org/2023.acl-short.135/>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451. <https://aclanthology.org/2020.acl-main.747/>.
- Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2023. Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0. CLARIN.SI language resource repository. <http://hdl.handle.net/11356/1881>.
- Maarten Grootendorst. 2021. KeyBERT. Software. Zenodo. <https://doi.org/10.5281/zenodo.4461265>.
- Simon Krek et al. 2020. Gigafida 2.0: The reference corpus of written standard Slovene. In *Proceedings of LREC*, pages 3340–3345.
- Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35. <https://arxiv.org/abs/2205.13147>.
- Marko Pranjić, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2025. Tracking semantic change in Slovene: A novel dataset and optimal transport-based distance. arXiv:2402.16596v2. <https://arxiv.org/abs/2402.16596v2>.
- Marko Pranjić, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2026. Tracking semantic change in Slovene: A novel dataset and optimal transport-based distance. Accepted in *PeerJ*; publication details pending.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992. <https://aclanthology.org/D19-1410/>.
- Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DUREl): A framework for the annotation of lexical semantic change. In *Proceedings of NAACL*, pages 169–174. <https://aclanthology.org/N18-2027/>.
- Dominik Schlechtweg, Tejaswi Chopra, Wei Zhao, and Michael Roth. 2025. CoMeDi shared task: Median judgment classification and mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of Context and Meaning*, pages 33–47. <https://aclanthology.org/2025.comedi-1.4/>.
- Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. 2019. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9:5233. <https://doi.org/10.1038/s41598-019-41695-z>.
- Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. In *Proceedings of Information Society 2021*, volume C: Data Mining and Data Warehouses.

Sachin Yadav and Dominik Schlechtweg. 2025. XL-DURel: Finetuning sentence transformers for ordinal word-in-context classification. In *Findings of IJCNLP-AACL*, pages 338–351. <https://aclanthology.org/2025.findings-ijcnlp.19/>.