A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change

Francesco Periti
University of Milan
Via Celoria, 18
20133 Milano, Italy
francesco.periti@unimi.it

Nina Tahmasebi
University of Gothenburg
Renströmsgatan 6
40530 Göteborg, Sweden
nina.tahmasebi@gu.se

Abstract

Contextualized embeddings are the preferred tool for modeling Lexical Semantic Change (LSC). Current evaluations typically focus on a specific task known as Graded Change Detection (GCD). However, performance comparison across work are often misleading due to their reliance on diverse settings. In this paper, we evaluate state-of-the-art models and approaches for GCD under equal conditions. We further break the LSC problem into Word-in-Context (WiC) and Word Sense Induction (WSI) tasks, and compare models across these different levels. Our evaluation is performed across different languages on eight available benchmarks for LSC, and shows that (i) APD outperforms other approaches for GCD; (ii) XL-LEXEME outperforms other contextualized models for WiC, WSI, and GCD, while being comparable to GPT-4; (iii) there is a clear need for improving the modeling of word meanings, as well as focus on *how*, *when*, and *why* these meanings change, rather than solely focusing on the extent of semantic change.

1 Introduction

Lexical Semantic Change (LSC) is the problem of automatically identifying words that change their meaning over time (Montanelli and Periti, 2023; Tahmasebi et al., 2021; Kutuzov et al., 2018; Tang, 2018). The interest in this problem has been significantly fueled by the advent of word embeddings and modern language models. After more than a decade of ad hoc evaluation, a new evaluation framework was recently introduced, aimed at assessing and comparing the performance of different models and approaches (Schlechtweg et al., 2020). This framework was adopted to create benchmarks in different languages. Each benchmark includes a diachronic corpus spanning two time periods, along with a list of target words and tasks aimed at detecting word meaning change over time. The most popular task, known as `Graded Change Detection` (GCD), consists of ranking a list of target words based on their degree of change.

The initial excitement for word embeddings prompted researchers and practitioners to solve the GCD task by using static embedding models (Schlechtweg et al., 2020; Shoemark et al., 2019). However, the shift towards more advanced Transformer architectures has established the use of contextualized embedding models as the preferred tool for addressing GCD (Montanelli and Periti, 2023; Kutuzov et al., 2022b). On one hand, these models distinguish the different meanings of a word by contextualizing each occurrence with a different embedding. On the other hand, the generation and processing of contextualized embeddings across entire corpora pose scalability challenges, both in terms of time and memory consumption (Periti et al., 2022; Montariol et al., 2021). Different strategies have been adopted to tackle these challenges, leading to a proliferation of evaluations across diverse settings (e.g., limited samples of benchmarks) and conditions (e.g., pre-trained vs. fine-tuned models). As a result, these evaluations on GCD hinder a fair comparison among the performance of different models and approaches, thereby deviating from the original goal of the framework.

Moreover, while the GCD task is attracting more and more evaluations, it addresses only a partial complexity inherent to the established framework. Notably, the framework includes three distinct aspects also discussed in Schlechtweg et al. (2024):

- **i)** **semantic proximity judgments** of word *in-context*,
- **ii)** **word sense induction** based on proximity judgments,
- **iii)** **quantification of semantic change** from induced senses.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4262–4282

As a matter of fact, when contextualized embedding models are used to address GCD, cosine similarities among word embeddings serve as surrogate for **(i)**, without evaluation focused on this aspect. Additionally, most approaches to GCD, pass from **(i)** to **(iii)**, sidestepping the intermediate aspect **(ii)**. That is, they quantify semantic change as overall proximity variation, without inducing word senses. Consequently, while these approaches can be evaluated through GDC, they preclude the interpretation of which meaning(s) have changed.

We argue that **(i)** and **(ii)** are equally relevant aspects as **(iii)**, constituting a fundamental aspect of the LSC problem. Their evaluation can provide valuable insights into the current state of LSC modeling, while offering a broader perspective on contextualized embedding models in Natural Language Processing (NLP).[1]

Original contribution of our work

- We systematically evaluate and compare various models and approaches for GCD under equal settings and conditions. Our evaluation for GCD spans eight different languages. Importantly, we perform the first evaluation over Chinese and the second evaluation for Norwegian within the existing literature. Our results show superior performance of the recent state-of-the-art model for GCD, namely XL-LEXEME, over various approaches.
- We are the first to evaluate contextualized embedding models for **(i)** and **(ii)** within the existing literature. Our evaluation of **(i)** and **(ii)** relies on two well-known tasks in NLP, namely `Word-in-Context` (WiC), and `Word Sense Induction` (WSI). Importantly, we evaluate various models as *computational annotators*.
- We compare GPT-4 to contextualized models through the WiC, WSI, and GCD tasks. Our evaluation reveals that GPT-4 obtains comparable performance to XL-LEXEME. In contrast to the limited accessibility[2] and high associated cost[3] of GPT-4, XL-LEXEME is a considerably smaller, open-source model. Thus, we argue that the use of GPT-4 is not justified for modeling the LSC problem.

[1]Software is available at https://github.com/FrancescoPeriti/CSSDetection.

[2]https://platform.openai.com/docs/guides/rate-limits

[3]https://openai.com/pricing

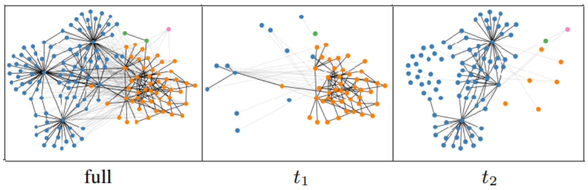

Figure 1: DWUG for the German word *Eintagsfliege*. Nodes represent word usages. Edges represent the relatedness between usages. Colors indicate clusters (senses) inferred from the full graph (Laicher et al., 2021).

2 Background and related work

The established LSC framework adheres to the novel annotation paradigm for word senses and encompasses **(i-iii)** (Schlechtweg et al., 2021). **(i)** Human annotators provide semantic proximity judgments for pairs of word usages *sampled* from a diachronic corpus spanning two time periods. **(ii)** Word usages and judgments are represented as nodes and edges in a weighted, *diachronic* graph, known as Diachronic Word Usage Graph (DWUG). This graph is then clustered with a graph clustering algorithm and the resulting clusters are interpreted as word senses (see Figure 1), thus sidestepping the need for explicit word sense definitions. Finally, **(iii)** given a word, a ground truth score of semantic change is computed by comparing the probability distributions of clusters in different time periods, e.g., a cluster with most of its usages from one time period indicates a substantial semantic change.

Originally, the framework was proposed in a shared task at SemEval-2020, including benchmarks for four languages, namely English (EN), German (DE), Swedish (SW), and Latin (LA) (Schlechtweg et al., 2020). Benchmarks for Italian (Basile et al., 2020), Russian (RU) (Kutuzov and Pivovarova, 2021b,c), Spanish (SP) (Zamora-Reina et al., 2022b), Norwegian (NO) (Kutuzov et al., 2022a), and Chinese (ZH) (Chen et al., 2023a, 2022) have recently been introduced. Each benchmark[4] consists of a diachronic corpus and a set of target words over which the human annotation was conducted. The evaluation over a benchmark is typically conducted through the GCD task where the goal is to rank the targets by degree of semantic change across the corpus. The Spearman correlation between *predicted* and *ground truth* scores is used to evaluate models and approaches.

[4]See https://github.com/ChangeIsKey/LSCDBenchmark for a comprehensive overview of available benchmarks

2.1 Approaches to Graded Change Detection

GCD is typically addressed using two kinds of approaches for modeling word meanings: *form-* and *sense*-based (Montanelli and Periti, 2023; Giulianelli et al., 2020). The former capture signals of change by analysing how the dominant meaning, or the degree of polysemy of a word, changes over time (e.g., Giulianelli et al., 2020; Martinc et al., 2020a). The latter cluster word usages according to their meanings and then estimate the semantic change of a word by comparing the cluster distribution of its usages over time (e.g., Periti et al., 2023; Martinc et al., 2020b). Form- and sense-based approaches can be further distinguished into *supervised*, which leverage external knowledge (e.g., dictionaries, Rachinskiy and Arefyev, 2022) or other forms of supervision (e.g., Word-in-Context datasets, Cassotti et al., 2023), and *unsupervised*, which rely solely on the knowledge encoded in pre-trained models (e.g., Aida and Bollegala, 2023).

2.2 Comparison of approaches

Models and approaches for GCD have been evaluated under different settings and conditions. For example, some studies utilized the *entire* diachronic corpus to estimate the change of each target (e.g., Periti et al., 2022), while others relied on smaller *samples* (e.g., Rodina et al., 2021), or solely on the annotated word usages (e.g., Laicher et al., 2021). Also, different versions of the ground truth are used (e.g., Schlechtweg et al., 2022a). In the current literature, some studies fine-tune the models on the corpus (e.g., Rosin et al., 2022), while others directly use pre-trained models (e.g., Kudisov and Arefyev, 2022). Performance comparison are conducted across different models such as BERT (e.g., Laicher et al., 2021), mBERT (e.g., Beck, 2020), and XLM-R (e.g., Giulianelli et al., 2022). However, even when the same model is employed, different layer aggregations are used, such as concatenating the output of the last four encoder layers (e.g., Kanjirangat et al., 2020), or summing the output of all the encoder layers (e.g., Giulianelli et al., 2022). Moreover, sense-based approaches are compared with different clustering algorithms such as Affinity Propagation (e.g., Martinc et al., 2020b), A Posteriori affinity Propagation (e.g., Periti et al., 2022), and K-Means (e.g., Montariol et al., 2021).

As a results, comparing Spearman correlation across different evaluations is often **misleading**.

2.3 Current modeling of LSC

Current modeling of LSC overlooks the procedure **(i-iii)** used to generate the ground truth. Mostly, only **(iii)** is evaluated by relying on form-based approaches. However, these approaches capture only the *degree* of semantic change, preventing its interpretation. Sense-based approaches could fill this gap by explaining *how* and *what* has changed, but currently suffer from lower performance on **(iii)** and are therefore less pursued. As a results, it is not clear which meanings these models and approaches are capturing. There is thus a need to carefully evaluate their ability in both **(i)** and **(ii)**.

Thus far, this evaluation is missing. To the best of our knowledge, only Laicher et al. (2021) evaluate **(ii)** through the WSI task. This evaluation needs to be extended beyond a single model, using the same procedure used to generate the ground truth.

A systematic comparison under equal settings and conditions is necessary to evaluate different models and approaches. Thus, we first evaluate standard form- and sense-based approaches to provide a fair performance comparison on GCD across eight languages. We then assess different models as *computational annotators* by evaluating them on **(i-iii)** through WiC, WSI, and GCD. Aligning with Karjus (2023), if computational models perform close to human-level, their usage would represent an unprecedented opportunity to scale up semantic change studies in the humanities and social sciences.

3 Evaluation setup

We consider benchmarks for eight different languages: EN, LA, DE, SV, ES, RU, NO, and ZH (see Table 6). For each benchmark, we evaluate four different models: BERT (Devlin et al., 2019), mBERT, XLM-R (Conneau et al., 2020), and XL-LEXEME (Cassotti et al., 2023). BERT is a monolingual model, mBERT, XLM-R, and XL-LEXEME are multilingual models. BERT, mBERT, and XLM-R are pre-trained masked language models. XL-LEXEME is a fine-tuned XLM-R model that leverages the SBERT architecture to solve the WiC task, thus serving as a WiC pre-trained model.

Aligning with the *unsupervised* nature of the LSC framework, we compare pre-trained models without performing additional fine-tuning (see Table 7). For each model and each target word in a benchmark, we collect contextualized embeddings for all its word usages in both time periods.

Specifically, we generate the sets of embeddings $\Phi^1 = \{a_1, ..., a_n\}$ and $\Phi^2 = \{b_1, ..., b_m\}$ for the word usages associated to time periods t_1 and t_2, respectively.

3.1 Standard Graded Change Detection

We compare the use of different models with four standard approaches to GCD, specifically two form-based and two sense-based. Similar to Laicher et al. (2021), we consider the raw data originally used to derive ground truth scores, instead of considering the associated corpora. This ensures an accurate evaluation under a controlled setting.

3.2 Computational annotators

We assess different models as computational annotators by using cosine similarities between embeddings as a surrogate of human judgments. In our evaluation, we consider word usage pairs where human judgments are available, instead of considering all potential usage pairs (as in Section 3.1). Specifically, we adhere to the framework **(i-iii)** and evaluate different models through the WiC, WSI, and GCD tasks.

Inspired by Periti et al. (2024); Laskar et al. (2023); Kocoń et al. (2023); Karjus (2023), we evaluate GPT-4 and compare its use to contextualized models. However, the limited accessibility and high associated cost constraint our extension only to the EN benchmark.

4 Comparing approaches for GCD

We evaluate different approaches for GCD using Spearman correlation (Spearman, 1904) between computational predictions and ground truth scores. Specifically, we process the embeddings of each target using the following approaches.

4.1 Form-based approaches

In the most recent survey on LSC by Montanelli and Periti (2023), it was observed that cosine distance over word prototype (PRT) and the average pairwise distance (APD) consistently demonstrated superior performance compared to alternative approaches. Thus, we employ these approaches:

PRT computes the degree of change of a word w as the cosine distance between the average embeddings μ_1 and μ_2 (also know as *prototype* embeddings) of w in the time periods t_1 and t_2 (Martinc et al., 2020a; Kutuzov and Giulianelli, 2020). Formally, given a word w, we compute its degree of change by computing:

$$\text{PRT}(\Phi^1, \Phi^2) = 1 - cosine(\mu_1, \mu_2) \quad (1)$$

The intuition behind PRT is that a prototype embedding encodes the dominant meaning of a word, and as such, the semantic change is computed as a shift in the dominant meaning over time.

APD computes the degree of change of a word w as the average pairwise distance between the word embeddings in Φ^1 and Φ^2 (Giulianelli et al., 2020; Kutuzov and Giulianelli, 2020). Formally, given a word w, we compute its degree of change, where d is cosine distance, as follows:

$$\text{APD}(\Phi^1, \Phi^2) = \frac{1}{|\Phi^1||\Phi^2|} \cdot \sum_{a \in \Phi^1,\, b \in \Phi^2} d(a, b) \quad (2)$$

The intuition behind APD is that different word embeddings encode the polysemy of a word, and as such, the semantic change is computed as a shift in the word's degree of polysemy.

4.2 Sense-based approaches

We choose two state-of-the-art sense-based approaches (Montanelli and Periti, 2023).The first utilizes the unsupervised clustering algorithm Affinity Propagation (AP, Frey and Dueck, 2007) combined with the Jensen Shannon divergence (JSD, Lin, 1991). Additionally, we employ the evolutionary extension of Affinity Propagation, called A Posteriori affinity Propagation (APP, Castano et al., 2024), combined with the average pairwise distances between sense prototypes (APDP). This approach is called WiDiD (Periti et al., 2022).

AP+JSD leverages the AP clustering to distinguish the different contextual usages of a given word w. Specifically, the embeddings Φ^1, and Φ^2 are *collectively* clustered to generate clusters comprising embeddings from both time periods (i.e., t_1 and t_2), or embeddings exclusive from a time period (i.e., t_1 or t_2). The semantic change of w is computed as the JSD between the probability distributions p_1 and p_2 of clusters in time periods t_1 and t_2. These distributions represent the relative number of embeddings from Φ^1 and Φ^2 grouped in each cluster, respectively (Martinc et al., 2020b,c). Formally, the degree of semantic change is:

$$\text{JSD}(p_1, p_2) = \frac{1}{2} \left(KL(p_1||M) + KL(p_2||M) \right) \quad (3)$$

where KL stands for Kullback-Leibler divergence and $M = \frac{(p^1+p^2)}{2}$. The intuition behind AP+JSD is that different clusters encode nuanced word meanings, and as such, the semantic change is computed as an overall measure of the differences in the prominence of each sense over time.

WiDiD leverages the APP clustering to distinguish the usages of a given word w. Specifically, the embeddings Φ^1, and Φ^2 are *individually* clustered to generate incremental clusters of embeddings that evolve with each clustering iteration. The semantic change of w is computed as the average pairwise distances between the *sense prototypes* Ψ^1 and Ψ^2 of w in the time periods t_1 and t_2, where Ψ^1 and Ψ^2 are the set of embeddings obtained by averaging the embeddings Φ^1 and Φ^2 in each cluster, respectively (Periti et al., 2023; Kashleva et al., 2022). Formally, given a word w, the degree of semantic change is computed as follows:[5]

$$\text{APDP}(\Phi^1, \Phi^2) = \text{APD}(\Psi^1, \Psi^2) \qquad (4)$$

The intuition behind WiDiD is similar to AP+JSD. However, while the latter considers change as the difference between the amount of probability for a sense over time, WiDiD is similar to APD in computing the shift in prototypical word meanings.

4.3 Evaluation results - Table 1

We present the results of our evaluation in Table 1 for both form- and sense-based approaches. For the sake of comparison, we include state-of-the-art (SOTA) results in Table 5.[6] As a general remark, we note instances where our results surpass SOTA (e.g., XL-LEXEME+APD for EN). We attribute this to the controlled setting established in our experiments. We note also instances where our results are lower than SOTA (e.g., BERT+APD for SV). This discrepancy may be influenced by various factors such as *different versions* of the benchmarks (e.g., 37 vs 46 targets for EN in DWUG version 2.0.1, Schlechtweg et al., 2020). Additionally, *variations in text pre-processing* can play a beneficial role. For instance, Laicher et al. (2021) demonstrate the effectiveness of lemmatization to mitigate word form biases, while Martinc et al. (2020c) suggest that filtering Named Entities can help models avoid inflating semantic change. Moreover, some studies *fine-tune or utilize different embedding layers*, whereas we adhere to the standard, generally adopted procedures without fine-tuning, considering embeddings generated from the last (i.e., 12$^{\text{th}}$) layer of the models. Finally, there are sometimes significantly different results reported by different studies under similar conditions. For instance, Zhou et al. (2023) achieve a correlation of .706 using pre-trained BERT and APD, whereas others typically report correlations ranging between .400 and .600 (e.g., .489, Keidar et al., 2022; .514, Giulianelli et al., 2020; .546, Kutuzov and Giulianelli, 2020; .571, Laicher et al., 2021). This disparity cannot currently be explained.

Languages. We obtain strong correlations with all benchmarks but LA. Our results show a *weighted average* correlation of **.751** when employing XL-LEXEME + APD. In this calculation, we assign weights based on the number of targets in each benchmark, considering larger sets more reliable than smaller ones. For LA, it can be argued that the models were not directly tailored or fine-tuned for Latin. However, XL-LEXEME demonstrates optimal performance in GCD in SV and medium performance in SP and NO without specific training on either (Cassotti et al., 2023). This leads us to consider that the quality of the LA benchmark potentially is lower than other benchmarks, as it was developed using a different procedure (Schlechtweg et al., 2020).

Form-based vs Sense-based. We note that form-based approaches significantly outperform sense-based approaches. Our results consistently highlight APD as the most effective approach, regardless of the skewness in the distribution of judgments, as previously argued by Kutuzov and Giulianelli (2020). In addition, WiDiD consistently demonstrate superior performance over AP+JSD. This can be attributed to the use of i) an evolutionary clustering algorithm, which enables to consider the time dimension of text in a dynamic way; or, alternatively ii) APD over sense-prototypes, as APD has demonstrated high effectiveness.

Our **leaderboard** is as follows: APD, PRT, WiDiD, AP+JSD. Although form-based approaches exhibit superior effectiveness, they fall short in capturing word meanings and interpreting detected semantic changes. In contrast, although sense-based approaches theoretically facilitate such modeling and interpretation, they obtain poor results in GCD,

[5]Following Periti et al. (2023), we use the Canberra distance instead of the cosine distance

[6]Our comparison includes results from different benchmarks using the same approaches. However, some benchmarks might have been assessed using other approaches.

			EN	LA	DE	SV	ES	RU			NO		ZH	Avg$_w$
			C_1-C_2	C_1-C_2	C_1-C_2	C_1-C_2	C_1-C_2	C_1-C_2	C_2-C_3	C_1-C_3	C_1-C_2	C_2-C_3	C_1-C_2	C_i-C_j
form-based	APD	BERT	.563	-	.271	.270	.335	.518	.482	.416	.441	.466	.656	.449
		mBERT	.363	.102	.398	.389	.341	.368	.345	.386	.279	.488	.689	.371
		XLM-R	.444	.151	.264	.257	.386	.290	.287	.318	.195	.379	.500	.316
		XL-LEXEME	**.886***	**.231**	**.839***	**.812***	**.665***	**.796***	**.820***	**.863***	**.659**	**.640***	**.731***	**.751***
		SOTA: sup.	*.757*	*-.056*	*.877*	*.754*	*n.a.*	*.799*	*.833*	*.842*	*.757*	*.757*	*n.a.*	
		SOTA: uns.	*.706*	*.443*	*.731*	*.602*	*n.a.*	*.372*	*.480*	*.457*	*.389*	*.387*	*n.a.*	
	PRT	BERT	.457	-	.422	.158	.413	.400	.374	.347	.507	.444	**.712**	.406
		mBERT	.270	.380	.436	.193	.543	.391	.356	.423	.219	.438	.524	.395
		XLM-R	.411	.424	.369	.020	.505	.321	.443	.405	.387	.149	.558	.381
		XL-LEXEME	**.676**	**.506***	**.824**	**.696**	**.632**	**.704**	**.750**	**.727**	**.764***	**.519**	.699	**.693**
		SOTA: sup.	*.531*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	
		SOTA: uns.	*.467*	*.561*	*.755*	*.392*	*n.a.*	*.294*	*313*	*313*	*.378*	*.270*	*n.a.*	
sense-based	AP+JSD	BERT	.289	-	.469	- .090	.225	.069	.279	.094	**.314**	.011	.165	.179
		mBERT	.181	.277	.280	.023	.067	.017	.086	- .116	.035	- .090	.465	.077
		XLM-R	.278	**.398**	.224	-.076	.224	- .068	**.209**	**.130**	- .100	.030	.448	.142
		XL-LEXEME	**.493**	.033	**.499**	**.118**	**.392**	**.106**	.053	.117	.297	**.381**	**.308**	**.223**
		SOTA: sup.	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	
		SOTA: uns.	*.436*	*.481*	*.583*	*.343*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	
	WiDiD	BERT	.385	-	.355	.106	.383	.135	.102	.243	.233	.087	.533	.239
		mBERT	.323	- .039	.312	.195	.343	- .068	.160	.142	.241	.290	.338	.181
		XLM-R	.564	- .064	.499	.129	.459	**.268**	.216	.342	.226	.349	.382	.314
		XL-LEXEME	**.652**	**.236**	**.677**	**.475**	**.522**	.178	**.354**	**.364**	**.561**	**.457**	**.563**	**.422**
		SOTA: sup.	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	
		SOTA: uns.	*.651*	*-.096*	*.527*	*.499*	*.544*	*.273*	*.393*	*.407*	*n.a.*	*n.a.*	*n.a.*	

Table 1: **Evaluation of standard approaches to GCD** in terms of Spearman correlation. Top score for each approach and benchmark in **bold**. The top score of each benchmark is marked with an asterisk (*). We include state-of-the-art performance achieved by *supervised* (sup.) and *unsupervised* (uns.) approaches in *italic*. Avg is the weighted average score based on the number of targets in each benchmark. Results not available denoted as n.a.

raising concerns about their reliability and whether they capture meaningful patterns or produce noisy aggregation. We will investigate this in Section 5.

Supervised vs Unsupervised. We note that the use of supervision significantly improves the modeling of semantic change for both form- and sense-based approaches. While Cassotti et al. (2023) have previously evaluated XL-LEXEME + APD, we extend the evaluation to sense-based approaches, demonstrating that *supervision* enhances the performance of AP+JSD and WiDiD.

Models. We note that the use of XL-LEXEME significantly improves the modeling of LSC compared to standard BERT, mBERT, and XLM-R. However, we observe a pattern in performance, indicating that on average, BERT performs better than mBERT, which, in turn, performs better than XLM-R for form-based approaches. This suggests that the use of XLM-R models is not more effective than BERT models for LSC, confirming the medium-low correlation coefficients obtained by Giulianelli et al. (2022) using XLM-R.

Layers. As different works employ different embedding layers, we repeat our evaluation by considering embeddings generated by each layer of BERT, mBERT, and XLM-R (see Appendix C). Our evaluation aligns with recent findings on other downstream tasks (Ma et al., 2019; Reif et al., 2019; Liang and Shi, 2023) and shows that using early layers consistently results in higher performance. For example, we note a correlation of .747 for ZH by using layer 4, compared to .656 obtained by using the last layer of BERT. On average, and in line with Periti and Dubossarsky (2023), we find that the best results for each language are obtained by leveraging embeddings from layers 8 – 10.

Furthermore, since previous studies aggregated outputs from different layers, we also use aggregated embeddings extracted from different layers through sum and concatenation (see Appendix C). Specifically, our evaluation covers all possible layer combinations with lengths of 2 (e.g., layers 1 and 2), 3 (e.g., layers 6, 7, and 8), and 4 (e.g., layers 9, 10, 11, 12). We find no improvement in aggregating the output of the last four layers for addressing GCD. By employing alternative layer combinations, we obtain higher correlation compared to both the last layer and the last four layers. For instance, for EN, using the sum of layers 2, 4, 5, and 8 for APD+BERT, or the concatenation of layers 4, 5, 6, and 11 for WiDiD+BERT, results in correlation of .692 and .760, respectively; compared to .563 (APD) and .385 (WiDiD) by using the last BERT layer. However, no combination consistently emerges as the optimal choice across various benchmarks or models. Instead, we observe that using a middle layer, such as layer 8, tends to be advantageous across benchmarks and models compared to the last layer or the aggregation of the last

		EN	DE	SV	ES	RU			NO		ZH	Avg$_w$
		$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_3$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_2$	$C_i - C_j$
WiC	BERT	.503	.350	.221	.319	.314	.344	.350	.429	.406	.516	.358
	mBERT	.332	.344	.284	.289	.280	.273	.293	.283	.333	.413	.301
	XLM-R	.352	.289	.255	.288	.212	.250	.251	.317	.261	.392	.272
	XL-LEXEME	**.626**	**.628**	**.631**	**.547**	**.549**	**.558**	**.564**	**.484**	**.521**	**.630**	**.568**
	GPT-4.0	.606	-	-	-	-	-	-	-	-	-	-
	Agreement	*.633*	*.666*	*.672*	*.531*	*.531*	*.567*	*.564*	*.761*	*.667*	*.602*	*.593*
WSI	BERT	.136 / .700	.047 / .662	.023 / .596	.189 / .695	- / -	- / -	- / -	.251 / .771	.247 / .758	.279 / .759	.166 / .702
	mBERT	.067 / .644	.054 / .679	.024 / .648	.228 / .700	- / -	- / -	- / -	.241 / .759	.159 / .753	.172 / .713	.146 / .696
	XLM-R	.068 / .737	.024 / .725	.031 / .680	.164 / .755	- / -	- / -	- / -	.179 / .775	.183 / .715	.279 / .806	.133 / .743
	XL-LEXEME	.273 / .834	**.300 / .788**	**.249 / .766**	**.400 / .820**	- / -	- / -	- / -	**.337 / .806**	**.304 / .808**	**.448 / .836**	**.339 / .810**
	GPT-4.0	**.340 / .877**	- / -	- / -	- / -	- / -	- / -	- / -	- / -	- / -	- / -	- / -
GCD	BERT	.425	.116	.148	.284	.487	.452	.469	.571	.521	**.808**	.422
	mBERT	.120	.205	.234	.394	.372	.325	.408	.290	.454	.737	.357
	XLM-R	.219	.069	.143	.464	.284	.301	.375	.395	.345	.557	.324
	XL-LEXEME	.801	**.799**	**.721**	**.655**	**.780**	**.824**	**.851**	**.620**	**.567**	.716	**.754**
	GPT-4.0	**.818**	-	-	-	-	-	-	-	-	-	-

Table 2: **Evaluation of contextualized models as computational annotators**: Spearman correlation for WiC and GCD, Adjusted Random Index and Purity (ARI / PUR) for WSI. Top score for each approach and benchmark is highlighted in **bold**. Avg is a weighted average based on the number of targets in each benchmark test set. For the sake of comparison, we report the Krippendorff's α score for inter-human annotator *agreement* in WiC (*italic*).

four layers (see Figure 2 and 3).

5 Computational annotation

We evaluate different models on reproducing human judgments **(i)**, the inferred word senses **(ii)**, and the resulting change scores **(iii)**.

We leverage models as annotators, hence the term *computational annotator*, using the same procedure employed for benchmark construction (Schlechtweg, 2023; Schlechtweg et al., 2021, 2020; Schlechtweg and Schulte im Walde, 2020; Schlechtweg et al., 2018). However, we cannot evaluate LA as the benchmark was developed differently nor **(ii)** for the RU benchmark since no word senses were provided (Kutuzov and Pivovarova, 2021b,c).

5.1 (i) - Word-in-Context

Given a benchmark, a word usage pair is associated with two contexts, c_1 and c_2, along with the average judgment of multiple annotators (see Example A). We thus use the cosine similarity between the embeddings of w in the contexts c_1 and c_2 as computational proximity judgement.

Our evaluation is grounded in the Word-in-Context (WiC) task (Loureiro et al., 2022; Raganato et al., 2020; Pilehvar and Camacho-Collados, 2019). In contrast to the original WiC definition, our WiC evaluation aligns with the continuous framework introduced by Armendariz et al. (2020) in the Graded Word Similarity in Context task. Specifically, we evaluate the quality of computational predictions by computing the Spearman correlation with human judgments.

5.2 (ii) - Word Sense Induction

We first create a DWUG using the computational annotations in Section 5.1. Then, we derive sense clusters through a variation of correlation clustering (Bansal et al., 2004) on the DWUG.

Our evaluation is grounded in the Word Sense Induction (WSI) task (Aksenova et al., 2022; Manandhar et al., 2010; Agirre and Soroa, 2007). We evaluate the quality of clusters from computationally annotated DWUGs against clusters from human-annotated DWUGs. Specifically, we use Adjusted Rand Index (ARI, Hubert and Arabie, 1985) and Purity (PUR, Manning, 2009) as metrics to quantify the cluster agreement. ARI comprehensively evaluates the similarity among clustering result. However, it may yield low scores when a clustering result contains numerous small, yet coherent clusters. This does not necessarily indicate poor clustering quality, especially when the clusters are semantically meaningful. PUR assigns each cluster to the class that is most frequent in the cluster, measuring the accuracy of this assignment by counting the relative number of correctly assigned elements.

5.3 (iii) - Graded Change Detection

Given a word w, we split its DWUG into two subgraphs representing nodes from the two time periods (see Figure 1) and quantify the semantic change of w by computing the $\sqrt{JSD}$ between the two time-specific cluster distributions. In contrast, for RU, we adhere to the RuShiftEval procedure and quantify semantic change through the application of the COMPARE metric that directly measures the mean relatedness of annotated word usage pairs as semantic change scores (Schlechtweg et al., 2018). Our evaluation is based on the GCD task and thus use Spearman correlation as evaluation metric between predicted ranking and ground truth rankings.

5.4 Evaluation results – Table 2

(i) - Word-in-Context Our evaluation reveals that pre-trained models such as BERT, mBERT, and XLM-R demonstrate a low average correlation with human judgments (.358, .301, .272). In contrast, XL-LEXEME and GPT-4 emerge as powerful solutions for scaling up and aiding human annotations. For EN, they obtain a moderately strong correlation (.626, .606) with human judgments, only marginally lower than the Krippendorf α human agreement (.633). In particular, XL-LEXEME slightly outperforms a considerably larger model like GPT-4 in terms of parameters, at a considerable lower cost. In contrast to previous cross-lingual evaluation (Conneau et al., 2020) and in line with the finding in Table 1, mBERT consistently outperforms XLM-R. However, our results highlight the advantageous use of monolingual BERT models over the multilingual ones, for assessing **(i)** - WiC.

We consider the WiC evaluation to be the most valuable as it involves a direct comparison between computational predictions and human judgments.

(ii) - Word Sense Induction Our evaluation indicates that moderate performance in **(i)**-WiC leads to moderately *low* performance in inferring word sense. We obtain low ARI scores across all models and benchmarks, with XL-LEXEME and GPT-4 exhibiting the highest values. Specifically, GPT-4 outperforms XL-LEXEME (with .340 compared to .273) in ARI for EN. However, we highlight that even such low scores represent a moderately *high* result, given an inter-annotator agreement of .633.

XL-LEXEME consistently demonstrates high PUR scores across all benchmarks, while other models yield slightly lower PUR scores, suggesting that some word sense patterns are captured when using contextualized models. Previous studies highlight that contextualized models tend to produce a large number of clusters (Martinc et al., 2020b; Periti et al., 2022), thereby influencing PUR scores. Therefore, it is crucial to interpret PUR in conjunction with ARI.

(iii) - Graded Change Detection As for GCD, we obtain average results for BERT, mBERT, XLM-R, and XL-LEXEME equal to .422, .357, .324, .754, respectively. These results are consistent with those presented in Table 1, when compared to form-based approaches (.316 – .751). We observe that employing more word usage pairs, as in Table 1, proves beneficial for certain benchmarks in the GCD tasks (e.g., XL-LEXEME+APD for EN and DE). However, we note that these results for **(ii)** - WSI are significantly higher to those obtained by sense-based approaches (.077 – .422). This can likely be attributed the fact that here we are using the same clustering algorithm that was used for obtaining the ground truth clusters, or to the fact that the clustering algorithm is more able to capture nuanced word meaning than AP and APP. In contrast, for RU, following the RuShiftEval procedure does not improve the performance and results between Table 1 and 2 are somewhat comparable.

6 Concluding remarks

We have performed a first-ever evaluation of models and approaches for modeling LSC under equal settings and conditions, over eight different languages. First, we evaluated different models combined with standard approaches to the popular GCD task. In particular, we consider BERT, mBERT, XLM-R, XL-LEXEME as pre-trained models, APD and PRT as form-based approaches, and AP+JSD and WiDiD as sense-based approaches. We find that the XL-LEXEME consistently outperforms other models across all approaches, and thus should be used as the defacto standard. We also find that form-based approaches significantly outperform sense-based approaches, with APD as the best approach for GCD. Among the sense-based approaches, we find that *evolutionary* clustering is advantageous in contrast to static clustering and should be a focus of future work. We additionally extended the evaluation to includes the WiC and WSI tasks, both inherently crucial to solve the complex task of LSC. We compare GPT-4 to the previous models and find that GPT-4 and XL-LEXEME both perform close to human-level while the other models obtain only low-moderate performance. Due to the costs associated with using GPT-4, it is not affordable to evaluate it on the remaining languages. Since XL-LEXEME obtains results close those of GPT-4, even beating it for the WiC task, we argue that XL-LEXEME can be used for LSC tasks as a affordable, scalable solution.

All in all, considering the current state of the LSC modeling, we argue that **only obtaining state-**

of-the-art performance on GCD does not solve the LSC problem, as there is a clear need to **distinguish the different senses of a word and how these evolve over time** (Periti et al., 2023, 2022; Castano et al., 2024). GCD maintains relevance for identifying words that have changed across multiple time periods in need of further *sense-based* modeling. GCD also serves to quantify the change on the level of vocabulary. In conclusion, we offer a first comparable evaluation of contextualized word embeddings for LSC and establish clear settings that should be used for future comparison and evaluation. With this work, we want to raise awareness of the current trend of the community in modeling only the GCD task. Our aim is to shift the focus from merely assessing *how much* to *how* (de Sá et al., 2024), *when*, and *why*, prompting the development of both *unsupervised* and *supervised* approaches for addressing the full spectrum of LSC.

7 Limitations

There are limitations we had to consider in the making of this paper. Firstly, we could not evaluate GPT-4 across all languages due to both price and API limitations. This means that while the results are comparable with XL-LEXEME for EN, we do not know how GPT-4 will behave for the other languages. Although we are aware of open source solution such as LLaMA, our initial experiments, revealed that its performance does not match that of GPT-4. As LLaMA still necessitates expensive research infrastructure, we chose to focus only GPT-4. Our decision to use GPT-4 over the cheaper GPT-3 is based on recent studies showing conflicting results across different tasks. Notably, Karjus (2023) reported high scores for GPT-4 in the GCD task. However, Periti et al. (2024); Laskar et al. (2023); Kocoń et al. (2023) reported low scores for the WiC task when employing GPT-3. As a result, we opted for GPT-4 to ensure relevance and accuracy in our evaluations.

Since the instruction-tuning datasets of OpenAI models are unknown, the datasets used for evaluation may or may not be part of the instruction-tuning training data of OpenAI. We also acknowledge that OpenAI continues to train and release new models, which could potentially affect the reproducibility of our results, as well as invalidate future evaluations (Balloccu et al., 2024).

In this paper, we evaluate different contextualized models utilizing the popular Transformers library for deep learning maintained by Hugging Face (Wolf et al., 2020). We specifically excluded the evaluation of a BERT model for Latin, opting instead to focus on mBERT, XLM-R, and XL-LEXEME. At the beginning of our evaluation, we were not aware of any experiments using Latin BERT models to address GCD, nor were we aware of an open BERT version for Latin on the Hugging Face platform. As we have only recently become aware of novel BERT models that are exclusively trained and fine-tuned for Latin (Riemenschneider and Frank, 2023; Lendvai and Wick, 2022), we plan to further test and utilize these models in our future work.

To make a fair comparison between different contextualized models, we employed the same procedure across all benchmarks and languages. However, different languages have different structures and hence different requirements. It would be equally fair to have different processing of the different benchmarks (e.g., lemmatization for German, Laicher et al., 2021). We opted to reduce the number of open variables to be able to make this first evaluation. Future work could optimize each language and then compare model performance.

Lastly, the models compared in this study, despite sharing similar architectures, tokenize text sequences differently based on their reference vocabulary. Consequently, a word may be split into different subtokens by one model and represented as a single token by another. Additionally, when contexts exceed the maximum input size, different models may truncate them at various points. Adhering to standard procedures in the field of LSC, we use the average embeddings of sub-words when a word is split into multiple sub-words. However, the impact of different truncation methods was not evaluated.

Acknowledgments

This work has in part been funded by the project Towards Computational Lexical Semantic Change Detection supported by the Swedish Research Council (2019–2022; contract 2018-01184), and in part by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2023. Swap and Predict – Predicting the Semantic Changes in Words across Corpora by Context Swapping. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7753–7772, Singapore. Association for Computational Linguistics.

Anna Aksenova, Ekaterina Gavrishina, Elisei Rykov, and Andrey Kutuzov. 2022. RuDSI: Graph-based Word Sense Induction Dataset for Russian. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 77–88, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 Task 3: Graded Word Similarity in Context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine learning*, 56:89–113.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita@ EVALITA2020: Overview of the EVALITA2020 DiachronicLexical Semantics (DIACR-Ita) Task. In *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, Online. CEUR-WS.

Christin Beck. 2020. DiaSense at SemEval-2020 Task 1: Modeling Sense Change via Pre-trained BERT Embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 50–58, Barcelona (online). International Committee for Computational Linguistics.

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 the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Silvana Castano, Alfio Ferrara, Stefano Montanelli, and Francesco Periti. 2024. Incremental Affinity Propagation based on Cluster Consolidation and Stratification.

Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. 2022. Lexicon of Changes: Towards the Evaluation of Diachronic Semantic Shift in Chinese. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023a. ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023b. ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in Characterization of Semantic Change.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Brendan J Frey and Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. *science*, 315(5814):972–976.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. 2022. Do Not Fire the Linguist: Grammatical Profiles Help Language Models Detect Semantic Change. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 54–67, Dublin, Ireland. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *Journal of classification*, 2:193–218.

Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221, Barcelona (online). International Committee for Computational Linguistics.

Andres Karjus. 2023. Machine-assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence.

Kseniia Kashleva, Alexander Shein, Elizaveta Tukhtina, and Svetlana Vydrina. 2022. HSE at LSCDiscovery in Spanish: Clustering and Profiling for Lexical Semantic Change Discovery. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 193–197, Dublin, Ireland. Association for Computational Linguistics.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of All Trades, Master of None. *Information Fusion*, 99:101861.

Artem Kudisov and Nikolay Arefyev. 2022. BOS at LSCDiscovery: Lexical Substitution for Interpretable Lexical Semantic Change Detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 165–172, Dublin, Ireland. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: a Survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021a. RuShiftEval.

Andrey Kutuzov and Lidia Pivovarova. 2021b. RuShiftEval: A Shared Task on Semantic Shift Detection for Russian. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*, 20, (online). RSUH.

Andrey Kutuzov and Lidia Pivovarova. 2021c. Three-part Diachronic Semantic Change Dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022a. NorDiaChange: Diachronic Semantic Change Dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Ranveig Enstad, and Alexandra Wittemann. 2021. NorDiaChange.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022b. Contextualized Embeddings for Semantic Change Detection: Lessons Learned. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A Systematic Study and

Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Piroska Lendvai and Claudia Wick. 2022. Finetuning Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 37–41, Taipei, Taiwan. Association for Computational Linguistics.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large Language Models Understand and Can be Enhanced by Emotional Stimuli.

Meng Liang and Yao Shi. 2023. Named Entity Recognition Method Based on BERT-whitening and Dynamic Fusion Model. In *2023 5th International Conference on Natural Language Processing (ICNLP)*, pages 191–197.

Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. TempoWiC: An Evaluation Benchmark for Detecting Meaning Shift in Social Media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3353–3359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal Text Representation from BERT: An Empirical Study.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 Task 14: Word Sense Induction & Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

Christopher D Manning. 2009. *An Introduction to Information Retrieval*. Cambridge university press.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020a. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Capturing Evolution in Word Usage: Just Add More Clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, Taipei, Taiwan. Association for Computing Machinery.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020c. Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings Not Always Better than Static for Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73, Barcelona (online). International Committee for Computational Linguistics.

Barbara McGillivray, Dominik Schlechtweg, Haim Dubossarsky, Nina Tahmasebi, and Simon Hengchen. 2021. DWUG LA: Diachronic Word Usage Graphs for Latin.

Stefano Montanelli and Francesco Periti. 2023. A Survey on Contextualised Semantic Shift Detection.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and Interpretable Semantic Change Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Francesco Periti and Haim Dubossarsky. 2023. The Time-Embedding Travelers@WiC-ITA. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy. CEUR.org.

Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024. (Chat)GPT v BERT Dawn of Justice for Semantic Change Detection. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 420–436, St. Julian's, Malta. Association for Computational Linguistics.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is Done is Done: an Incremental Approach to Semantic Shift Detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43, Dublin, Ireland. Association for Computational Linguistics.

Francesco Periti, Sergio Picascia, Stefano Montanelli, Alfio Ferrara, and Nina Tahmasebi. 2023. Studying Word Meaning Evolution through Incremental Semantic Shift Detection: A Case Study of Italian Parliamentary Speeches.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and

Context-Dependent Word Representations. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 180–186, Barcelona (online). International Committee for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2022. GlossReader at LSCDiscovery: Train to Select a Proper Gloss in English – Discover Lexical Semantic Change in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 198–203, Dublin, Ireland. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and Measuring the Geometry of BERT. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Frederick Riemenschneider and Anette Frank. 2023. Exploring Large Language Models for Classical Philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Julia Rodina, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. 2021. ELMo and BERT in Semantic Change Detection for Russian. In *Analysis of Images, Social Networks and Texts*, pages 175–186, Cham. Springer International Publishing.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time Masking for Temporal Language Models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 833–841, Virtual Event, AZ, USA. Association for Computing Machinery.

Dominik Schlechtweg. 2023. *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart.

Dominik Schlechtweg, Haim Dubossarsky, Simon Hengchen, Barbara McGillivray, and Nina Tahmasebi. 2022a. DWUG EN: Diachronic Word Usage Graphs for English.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2022b. DWUG DE: Diachronic Word Usage Graphs for German.

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating Lexical Semantic Change from Sense-Annotated Data. In *Proceedings of the 13th International Conference on the Evolution of Language (EvoLang13)*, Brussels, Belgium.

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 the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dominik Schlechtweg, Shafqat Mumtaz Virk, and Nikolay Arefyev. 2024. The lscd benchmark: a testbed for diachronic word meaning tasks.

Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte im Walde. 2023. The DURel Annotation Tool: Human and Computational Measurement of Semantic Proximity, Sense Clusters and Semantic Change.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Charles Spearman. 1904. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:88–103.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. *Survey of Computational Approaches to Lexical Semantic Change Detection*, pages 1–91. Language Science Press, Berlin.

Nina Tahmasebi, Simon Hengchen, Dominik Schlechtweg, Barbara McGillivray, and Haim Dubossarsky. 2022. DWUG SV: Diachronic Word Usage Graphs for Swedish.

Xuri Tang. 2018. A State-of-the-art of Semantic Change Computation. *Natural Language Engineering*, 24(5):649–676.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022a. Dwug es: Diachronic word usage graphs for spanish.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022b. LSCDiscovery: A Shared Task on Semantic Change Discovery and Detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

Jinan Zhou and Jiaxin Li. 2020. TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Referencing. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 222–231, Barcelona (online). International Committee for Computational Linguistics.

Wei Zhou, Nina Tahmasebi, and Haim Dubossarsky. 2023. The Finer They Get: Combining Fine-Tuned Models For Better Semantic Change Detection. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 518–528, Tórshavn, Faroe Islands. University of Tartu Library.

Appendix

A Semantic proximity

As an example, consider the following word usage pair $\langle w, c_1, c_2 \rangle$ extracted by the English benchmark for the word w =*plane*.

- c_1: But we are most familiar with the exhibitions of gravity in bodies descending inclined **planes**, as in the avalanche and the cataract.
- c_2: Over the next several years, he said, the Coast Guard will get 60 more people, two new 270-foot vessels and al twin-engine **planes**.

Following the DURel relatedness scale (see Table 3), the pair is annotated with an average judgment of **1** by human annotators.

↑	4:	Identical
	3:	Closely related
	2:	Distantly related
	1:	Unrelated

Table 3: The DURel relatedness scale used in Schlechtweg et al. (2023); Schlechtweg (2023); Schlechtweg et al. (2021, 2020); Schlechtweg and Schulte im Walde (2020); Schlechtweg et al. (2018)

B State-of-the-art for Graded Change Detection

In Table 5, we report the current top scores for GCD in the state-of-the-art with a reference to the paper from where the result is taken. Notably, we report results for different benchmarks using four different approaches evaluated in this paper. However, some benchmarks might have been assessed using other approaches that are excluded from this table.

C Graded Change Detection across layers

In Table 4, we report correlation scores for GCD across benchmarks. Specifically, we report results for BERT, mBERT, and XLM-R (separated by slash, i.e. "/") by utilizing all layers of the models (1-12), *individually*.

In Figure 2 and 3, we report correlation scores distribution for GCD obtained by using all possible layer *combinations* of length 2 (e.g., Layer 1 and 2), length 3 (e.g., Layer 10, 11, 12), and length 4 (e.g., Layer 1, 10, 11, 12) for BERT, mBERT, and XLM-R.

For the sake of comparison, we report in Table 8 the overall top score for GCD obtained using BERT, mBERT, and XLM-R. Specifically, we present results for the optimal combination and the outcome obtained by summing the last four layers, separated by a slash. Additionally, we include the standard result obtained using the last layer individually.

D Benchmarks

In Table 6, we report the benchmarks used in this work. Specifically, for each benchmark, we report time periods, diachronic corpus composition, number of targets, and benchmark versions.

E BERT, mBERT, XLM-R, XL-LEXEME

In Table 7, we report the BERT, mBERT, XLM-R, and XL-LEXEME models employed in our evaluation. All the models are base versions with 12 encoder layers and can be accessed on `huggingface.co`.

F BERT, mBERT, XLM-R, XL-LEXEME

In Table 7, we report the BERT, mBERT, XLM-R, and XL-LEXEME models employed in our evaluation. All the models are base versions with 12 encoder layers and can be accessed on `huggingface.co`.

G GPT-4 evaluation

We evaluate GPT-4 as computational annotator by relying on computational proximity judgments gathered through the following method.

Model initialization. We initialized the model with the following prompt (guideline):

```
Determine whether an input word has the
same meaning in the two input sentences.
Answer with 'Same', 'Related', 'Linked',
or 'Distinct'.  This is very important
to my career.
```

Notably, we combine and refine two different prompts used in previous works. We drew inspiration from the prompt utilized by Karjus (2023) to assess GPT-4 in addressing the Graded Change Detection task. Additionally, we drew inspiration from the prompt utilized by Li et al. (2023), called *EmotionPrompt*, which combines the original prompt with emotional stimuli to enhance the performance of Large Language Models.

Model template. For each word usage pair, we used the following prompt:

```
Determine whether [Target word] has
the same meaning in the following
sentences.  Do they refer to roughly
the Same, different but closely Related,
distant/figuratively Linked or unrelated
Distinct word meanings?
Sentence 1:  [Context 1]
Sentence 2:  [Context 2]
```

Notably, drawing inspiration from the OpenAI documentation[7] and the prompts utilized in previous work for the Word-in-Context task (Kocoń et al., 2023; Laskar et al., 2023), we structured our prompt in a format that facilitates parsing and comprehension. For each usage pair $\langle w, c_1, c_2 \rangle$ of a word w, we substitute [Target word] with the actual target w and [Context 1] and [Context 2] with c_1 and c_2, respectively.

We prompt GPT-4 without providing any message history. This means that, for each usage pair $\langle w, c_1, c_2 \rangle$, we re-initialize the model with the initial prompt (guideline) and subsequently prompt the model to gather a semantic proximity judgment for the pair $\langle w, c_1, c_2 \rangle$. This approach ensures that the model relies solely on its pre-trained knowledge, preventing potential biases stemming from previously prompted pairs.

[7] platform.openai.com/docs/guides/prompt-engineering

			EN	**LA**	**DE**	**SV**	**ES**	**RU**			**NO**		**ZH**	**Avg$_w$**
			$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_3$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_2$	$C_i - C_j$
form-based	APD	1	.358 / .278 / .064	- / **.153** / .073	.144 / .218 / .270	.213 / .132 / .134	.167 / .104 / .003	.335 / .204 / .258	.281 / .204 / .308	.261 / .214 / .253	.160 / .143 / .145	.234 / .219 / .203	.340 / -.100 / -.222	.255 / .171 / .166
		2	.464 / .346 / .229	- / .119 / .006	.155 / .208 / .319	.255 / .129 / .234	.255 / .164 / .076	.374 / .198 / .245	.309 / .188 / .283	.303 / .218 / .236	.199 / .155 / .153	.288 / .213 / .235	.540 / .263 / .338	.312 / .198 / .216
		3	.574 / .389 / .314	- / .047 / -.025	.164 / .232 / .301	.295 / .189 / .289	.307 / .212 / .139	.427 / .215 / .238	.370 / .218 / .292	.360 / .242 / .241	.290 / .170 / .171	.371 / .223 / .243	.594 / .464 / .540	.371 / .232 / .244
		4	.628 / .410 / .400	- / .022 / -.010	.176 / .241 / .326	.307 / .254 / .286	.394 / .276 / .184	.492 / .257 / .287	.427 / .247 / .346	.431 / .280 / .288	.364 / .168 / .143	.463 / .322 / .264	**.747** / .613 / .615	.438 / .275 / .284
		5	**.684** / .412 / .452	- / -.028 / .043	.237 / .344 / .414	**.305** / .321 / .351	.450 / .345 / .279	.519 / .295 / .374	.465 / .275 / .453	.456 / .318 / .373	.396 / .192 / .165	.497 / .364 / .330	.720 / .662 / .600	.471 / .315 / .36
		6	.667 / .395 / .438	- / -.005 / .061	.309 / .397 / .471	.242 / .352 / .424	.468 / .361 / .277	.516 / .338 / .438	.463 / .305 / **.503**	.467 / .347 / **.432**	.400 / .180 / .172	.532 / .374 / .367	.667 / .661 / .629	.473 / .338 / **.398**
		7	.614 / **.419** / .395	- / -.009 / .073	.335 / .434 / .471	.237 / .404 / .441	**.479** / .364 / .280	.549 / .402 / **.439**	**.495** / .379 / .473	.523 / .429 / .430	.429 / .262 / .191	**.547** / .437 / .375	.645 / **.725** / .618	.494 / .390 / .393
		8	.642 / .408 / .426	- / .023 / .043	.389 / **.481** / .474	.248 / .455 / .456	.438 / **.430** / .297	**.566** / **.427** / .430	**.495** / **.400** / .466	.531 / **.451** / .427	.416 / **.291** / .197	.529 / **.499** / .373	.654 / .715 / .638	**.497** / **.421** / .396
		9	.600 / .406 / .460	- / .044 / -.047	**.427** / .423 / **.479**	.250 / **.463** / .468	.399 / .413 / .352	.539 / .382 / .401	.479 / .364 / .419	**.534** / .405 / .404	.429 / .257 / .190	.525 / .462 / .394	.667 / .670 / **.646**	.486 / .391 / .388
		10	.530 / .348 / .511	- / .008 / -.082	.354 / .333 / .433	.275 / .414 / .497	.282 / .331 / .407	.515 / .362 / .369	.461 / .313 / .405	.523 / .379 / .402	.418 / .226 / .191	.531 / .425 / .411	.625 / .656 / .613	.450 / .346 / .387
		11	.554 / .305 / **.548**	- / .023 / -.069	.275 / .315 / .409	.267 / .309 / **.500**	.257 / .265 / **.444**	.439 / .333 / .361	.393 / .256 / .394	.461 / .330 / .401	.378 / .196 / **.215**	.530 / .403 / **.432**	.604 / .628 / .601	.405 / .303 / .392
		12	.563 / .363 / .444	- / .102 / **.151**	.271 / .398 / .264	.270 / .389 / .257	.335 / .341 / .386	.518 / .368 / .290	.482 / .345 / .287	.476 / .386 / .318	**.441** / .279 / .195	.466 / .488 / .379	.656 / .689 / .500	.449 / .371 / .316
	PRT	1	.295 / .195 / .221	- / .289 / .303	.133 / .162 / .122	.215 / .001 / .045	.303 / .295 / .190	.263 / .271 / .220	.206 / .149 / .305	.159 / .169 / .144	.032 / -.005 / .028	.161 / .168 / .039	.383 / .017 / -.139	.220 / .178 / .165
		2	.409 / .271 / .382	- / .286 / .263	.217 / .198 / .125	.274 / .006 / .066	.407 / .397 / .328	.304 / .279 / .216	.261 / .139 / .352	.196 / .161 / .153	.122 / -.020 / .092	.349 / .215 / -.020	.582 / .192 / .140	.302 / .209 / .216
		3	.436 / .295 / .453	- / .277 / .271	.267 / .230 / .141	**.301** / .012 / .078	.438 / .424 / .364	.338 / .311 / .203	.305 / .191 / .405	.251 / .195 / .162	.250 / .042 / .111	.365 / .294 / .005	.676 / .397 / .424	.348 / .253 / .253
		4	.467 / .290 / .487	- / .255 / .297	.297 / .285 / .204	.280 / .017 / .087	.455 / .446 / .388	.398 / .329 / .246	.346 / .235 / .433	.306 / .250 / .234	.378 / .019 / .102	.408 / .303 / .075	.691 / .525 / .544	.389 / .283 / .296
		5	.494 / .315 / .476	- / .232 / .322	.343 / .384 / .294	.233 / .060 / .129	.455 / .495 / .439	.399 / .364 / .323	.395 / .327 / .509	.331 / .313 / .323	.440 / .096 / .137	.466 / .367 / .189	.651 / .551 / .531	.408 / .337 / .357
		6	.516 / .353 / .447	- / .257 / .350	.379 / .421 / .357	.206 / .082 / .171	.451 / .524 / .449	.391 / .359 / .365	.390 / .374 / .519	.331 / .365 / .384	.449 / .104 / .181	.471 / .330 / .232	.637 / .556 / .475	.408 / .362 / .383
		7	.529 / **.383** / .462	- / .304 / .349	.400 / .437 / .385	.178 / .008 / .184	.466 / .498 / .453	**.411** / .379 / .358	**.426** / **.447** / .510	**.380** / .413 / .384	**.511** / .161 / .192	.501 / .371 / .236	.641 / .613 / .549	**.433** / .389 / .390
		8	.539 / **.383** / .464	- / .292 / .359	.398 / .468 / .402	.197 / .081 / .196	.453 / .514 / .463	.404 / **.393** / .375	.410 / .421 / **.531**	**.380** / .411 / .396	.449 / .227 / .292	.493 / .389 / **.246**	.664 / **.619** / .575	.426 / **.400** / .409
		9	**.549** / .358 / .437	- / .311 / .319	.390 / **.469** / .477	.201 / .096 / .247	**.476** / .501 / .503	.375 / .353 / .382	.402 / .404 / .471	.353 / .384 / .401	.481 / **.243** / .351	.485 / .380 / .239	.671 / .606 / **.646**	.422 / .385 / .418
		10	.511 / .355 / .481	- / .280 / .329	.380 / .454 / .486	.193 / **.133** / .223	.417 / .482 / .538	.349 / .376 / .409	.379 / .382 / .447	.335 / .366 / .431	.482 / .212 / .373	.481 / .398 / .263	.626 / .583 / .619	.396 / .378 / .431
		11	.452 / .342 / **.501**	- / .298 / .308	.412 / .430 / **.507**	.169 / .076 / **.245**	.422 / .489 / **.540**	.319 / .344 / **.412**	.317 / .335 / .439	.303 / .321 / **.438**	.448 / .197 / .360	**.503** / .365 / .214	.602 / .550 / .620	.371 / .350 / **.432**
		12	.457 / .270 / .411	- / **.380** / **.424**	**.422** / .436 / .369	.158 / .193 / .020	.413 / **.543** / .505	.400 / .391 / .321	.374 / .356 / .443	.347 / **.423** / .405	.507 / .219 / **.387**	.444 / **.438** / .149	**.712** / .524 / .558	.406 / .395 / .381
sense-based	AP	1	.129 / .220 / .032	- / -.011 / **.409**	-.108 / -.087 / -.040	-.121 / -.021 / -.244	.168 / .233 / .172	.050 / -.001 / -.154	.132 / .108 / .060	.098 / -.143 / .023	-.104 / -.237 / -.019	-.048 / .021 / -.239	.118 / -.179 / .110	.060 / .011 / .012
		2	.288 / .079 / -.128	- / .008 / .215	.113 / -.131 / -.017	-.138 / -.141 / -.244	.104 / .109 / .140	-.127 / -.154 / -.036	.038 / .110 / .073	.096 / -.109 / -.025	.031 / -.230 / -.025	-.039 / .104 / .028	.301 / -.058 / -.048	.052 / -.030 / .006
		3	.267 / .161 / .016	- / -.012 / .218	.007 / -.043 / .120	-.201 / -.117 / -.177	.161 / .142 / .063	-.006 / .007 / -.019	-.002 / .058 / .129	.027 / -.130 / -.020	-.118 / .016 / -.060	-.051 / -.011 / .124	.189 / .221 / -.143	.033 / .021 / .028
		4	.353 / .330 / .087	- / -.106 / .253	-.041 / .088 / .054	-.213 / -.131 / -.172	.263 / .195 / **.266**	.093 / -.159 / -.042	.045 / .096 / .104	.168 / -.076 / .050	-.281 / -.123 / -.016	.257 / -.282 / .020	.360 / .322 / -.047	.113 / .014 / .064
		5	**.432** / .221 / .322	- / -.024 / .281	.143 / .235 / .196	-.015 / -.083 / -.125	.247 / .319 / .162	.072 / -.085 / -.035	.169 / .014 / .140	.081 / -.019 / .025	-.318 / -.027 / .033	.323 / .143 / .149	.251 / **.689** / .343	.140 / .097 / .112
		6	.431 / .208 / .330	- / -.000 / .286	.243 / .372 / .280	-.129 / -.040 / -.070	.363 / .251 / .002	-.049 / -.111 / -.094	.173 / .093 / .176	.091 / .035 / .291	-.192 / -.076 / .031	**.440** / .206 / .131	**.458** / .342 / .280	.166 / .099 / .132
		7	.144 / .362 / .321	- / -.044 / .233	.284 / **.443** / .387	-.070 / -.031 / -.155	**.406** / .301 / .216	.082 / -.069 / **.067**	.288 / **.235** / .084	**.190** / **.158** / .131	-.257 / -.114 / -.051	.115 / .140 / -.130	.292 / .226 / .344	.183 / .153 / .131
		8	.228 / **.418** / .175	- / -.101 / .260	.417 / .353 / .393	**.124** / .114 / -.082	.384 / **.401** / .031	.058 / -.014 / -.073	.128 / .230 / .211	.088 / .137 / .228	-.165 / -.114 / -.109	-.029 / **.469** / **.256**	.113 / .231 / .045	.148 / **.192** / .117
		9	.424 / .357 / .311	- / .120 / .153	.339 / .322 / .361	.054 / .010 / -.195	.270 / .296 / .157	.038 / .013 / -.081	.072 / .149 / .232	.098 / .055 / .011	-.016 / .005 / **.045**	.092 / .198 / .031	.423 / .404 / .245	.157 / .158 / .104
		10	.233 / .317 / .289	- / .124 / .381	.393 / .328 / .334	-.023 / .061 / -.210	.294 / .201 / .151	**.126** / **.108** / .044	.116 / .169 / .240	.187 / .082 / .194	.151 / -.127 / -.041	.168 / .271 / .101	.430 / .291 / .436	**.197** / .158 / **.169**
		11	.148 / .338 / **.374**	- / .132 / .266	.465 / .275 / **.435**	-.057 / **.175** / **.133**	.351 / .310 / .039	-.004 / .034 / -.069	.068 / .141 / **.279**	.157 / .113 / **.262**	.021 / -.232 / -.211	.090 / .146 / .062	.322 / .223 / .243	.151 / .151 / .158
		12	.289 / .181 / .278	- / **.277** / .398	**.469** / .280 / .224	-.090 / .023 / -.076	.225 / .067 / .224	.069 / .017 / -.068	**.279** / .086 / .209	.094 / -.116 / .130	**.314** / **.035** / -.100	.011 / -.090 / .030	.165 / .465 / **.448**	.179 / .077 / .142
	WiDiD	1	.253 / .301 / .278	- / .028 / -.048	.147 / .204 / .219	.120 / .052 / -.062	.132 / .051 / -.015	.159 / .047 / .125	.108 / .073 / .197	.090 / -.036 / .051	.356 / .150 / .090	.120 / .127 / .154	.122 / .026 / .160	.146 / .074 / .103
		2	.434 / .261 / .065	- / .018 / -.130	.106 / .143 / .292	-.041 / .015 / -.118	.103 / .105 / .110	.209 / -.046 / .274	.076 / .180 / .060	.212 / -.038 / -.008	.285 / -.030 / .085	.161 / .103 / .214	.371 / -.013 / .063	.175 / .060 / .094
		3	.423 / .268 / .147	- / .026 / .019	.115 / .120 / .474	.198 / .029 / .106	.228 / .108 / .118	**.251** / -.073 / **.345**	.091 / .113 / .184	.233 / .077 / .153	.229 / -.102 / .074	.239 / .064 / .204	.256 / .114 / .349	.216 / .065 / .203
		4	**.611** / .228 / .448	- / .030 / .108	.126 / .067 / .424	.176 / -.130 / .312	.292 / .175 / .221	.091 / -.039 / .332	.010 / .041 / .307	.157 / -.053 / .059	.242 / .038 / .002	.340 / .152 / .062	.388 / .279 / **.417**	.200 / .054 / .244
		5	.527 / .078 / .393	- / -.020 / -.037	.190 / .173 / **.509**	.151 / -.074 / .300	.356 / .295 / .310	-.034 / .023 / .259	.071 / .076 / .314	.205 / .137 / .202	**.297** / .100 / .023	.380 / .156 / .316	.524 / .193 / .217	.218 / .112 / .265
		6	.458 / .250 / .625	- / -.030 / -.050	.293 / .294 / .433	.211 / .148 / .335	.382 / .387 / .346	.094 / .063 / .184	.141 / .066 / .210	.182 / **.288** / .264	.261 / -.080 / .215	.428 / .295 / .102	.446 / .271 / .335	.252 / .185 / .269
		7	**.305** / .328 / .475	- / **.139** / .106	.235 / .253 / .514	**.295** / .198 / **.414**	.382 / .318 / .324	.017 / .032 / .292	.203 / **.285** / .152	.216 / .188 / **.458**	.244 / .119 / .247	.397 / .195 / -.034	.338 / .298 / .293	.237 / .211 / .304
		8	.449 / .312 / .411	- / .091 / .038	.344 / .341 / .565	.071 / **.354** / .321	.340 / .371 / .395	.000 / -.008 / .105	.284 / .260 / .243	.025 / .203 / .267	.221 / .226 / .262	.449 / **.428** / .155	.475 / .325 / .286	.224 / **.242** / .271
		9	.544 / **.509** / .567	- / -.066 / .104	.353 / .299 / .573	.184 / .319 / .203	.324 / **.450** / .372	-.002 / .075 / .108	.083 / .076 / .171	.205 / .205 / .388	.183 / .063 / .174	.390 / .118 / .149	.404 / .347 / .328	.222 / .212 / .280
		10	.396 / .301 / .587	- / -.024 / **.187**	.315 / **.407** / .477	.145 / .233 / .148	.306 / .388 / **.471**	.011 / .087 / .270	**.302** / .090 / .308	.060 / .172 / .328	.155 / .179 / .234	**.488** / .175 / .275	.428 / **.355** / .383	.224 / .204 / .339
		11	.299 / .218 / **.627**	- / -.064 / -.111	.258 / .381 / .486	.172 / .128 / .343	**.424** / .432 / .464	.134 / **.152** / .220	.234 / .120 / **.334**	.185 / .087 / .312	.218 / .195 / **.345**	.296 / .291 / **.438**	**.539** / .277 / .372	**.260** / .199 / **.345**
		12	.385 / .323 / .564	- / -.039 / -.064	**.355** / .312 / .499	.106 / .195 / .129	.383 / .343 / .459	.135 / -.068 / .268	.102 / .160 / .216	**.243** / .142 / .342	.233 / **.241** / .226	.087 / .290 / .349	.533 / .338 / .382	.239 / .181 / .314

Table 4: **Comprehensive evaluation of standard approaches to GCD** by using the layers 1-12 of **BERT / mBERT / XLM-R**. Top score for each approach, model, and benchmark in **bold**. Avg is the weighted average score based on the number of targets in each benchmark.

		EN	LA	DE	SV	ES	RU			NO		ZH
		$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_3$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_2$
form-based	**APD**	XL-L. : .757 Cassotti et al. ― BERT: .706 Zhou et al.	XL-L. : -.056 Cassotti et al. ― mBERT: .443 Pömsl and Lyapin	XL-L. : .877 Cassotti et al. ― BERT: .731 Laicher et al.	XL-L. .754 Cassotti et al. ― BERT: .602 Laicher et al.	n.a. ― n.a.	XL-L. : .799 Cassotti et al. ― XLM-R: .372 Giulianelli et al.	XL-L. : .833 Cassotti et al. ― XLM-R: .480 Giulianelli et al.	XL-L. : .842 Cassotti et al. ― XLM-R: .457 Giulianelli et al.	XL-L. : .757 Cassotti et al. ― XLM-R: .389 Giulianelli et al.	XL-L. : .757 Cassotti et al. ― XLM-R: .387 Giulianelli et al.	n.a. ― n.a.
	PRT	BERT: .531 Zhou et al. ― BERT: .467 Rosin et al.	n.a. ― mBERT: .561 Kutuzov and Giulianelli	n.a. ― BERT: .755 Laicher et al.	n.a. ― BERT: .392 Zhou and Li	n.a. ― n.a.	n.a. ― XLM-R: .294 Giulianelli et al.	n.a. ― XLM-R: .313 Giulianelli et al.	n.a. ― XLM-R: .313 Giulianelli et al.	n.a. ― XLM-R: .378 Giulianelli et al.	n.a. ― XLM-R: .270 Giulianelli et al.	n.a. ― n.a.
sense-based	AP+JSD	n.a. ― BERT: .436 Martinc et al.	n.a. ― mBERT: .481 Martinc et al.	n.a. ― BERT: .583 Montariol et al.	n.a. ― BERT: .343 Martinc et al.	n.a. ― n.a.	n.a. ― n.a.	n.a. ― n.a.	n.a. ― n.a.	n.a. ― n.a.	n.a. ― n.a.	n.a. ― n.a.
	WiDiD	n.a. ― BERT: .651 Periti et al.	n.a. ― XLM-R: - .096 Periti et al.	n.a. ― XLM-R: .527 Periti et al.	n.a. ― XLM-R: .499 Periti et al.	n.a. ― BERT: .544 Periti et al.	n.a. ― mBERT: .273 Periti et al.	n.a. ― mBERT: .393 Periti et al.	n.a. ― mBERT: .407 Periti et al.	n.a. ― n.a.	n.a. ― n.a.	n.a. ― n.a.

Table 5: **State-of-the-art performance for GCD**: Top Spearman correlations obtained across benchmarks by form- and sense-based approaches. For each approach, we report correlation for both *supervised* (above the line) and *unsupervised* (below the line) settings.

	EN	LA	DE	SV	ES	RU			NO		ZH
	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_3$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_2$
Time periods	C_1: 1810 – 1860 C_2: 1960 – 2010	C_1: 200 – 0 C_2: 0 – 2000	C_1: 1800 – 1899 C_2: 1946 – 1990	C_1: 1790 – 1830 C_2: 1895 – 1903	C_1: 1810 – 1906 C_2: 1994 – 2020	C_1: 1700 – 1916 C_2: 1918 – 1990	C_2: 1918 – 1990 C_3: 1992 –2016	C_1: 1700 – 1916 C_3: 1992 –2016	C_1: 1929 –1965 C_2: 1970 – 2013	C_1: 1980 – 1990 C_2: 2012 – 2019	C_1: 1954 – 1978 C_2: 1979 – 2003
Diachronic Corpus	C_1: CCOHA C_2: CCOHA	C_1: LatinISE C_2: LatinISE	C_1: DTA C_2: BZ+ND	C_1: Kubhist C_2: Kubhist	C_1: PG C_2: TED2013, NC MultiUN Europarl	C_1: RNC C_2: RNC C_3: RNC	C_1: RNC C_2: RNC C_3: RNC	C_1: RNC C_2: RNC C_3: RNC	C_1: NBdigital C_2: NBdigital	C_1: NBdigital C_2: NAK	C_1: People's Daily C_2: People's Daily
# targets	46	40	50	44	100	111	111	111	40	40	40
Benchmark version	version 2.0.1 Schlechtweg et al.	version 1 McGillivray et al.	version 2.3.0 Schlechtweg et al.	version 2.0.1 Tahmasebi et al.	version 4.0.0 Zamora-Reina et al.	version 1 Kutuzov and Pivovarova			version 1 Kutuzov et al.		version 1 Chen et al.

Table 6: **LSC benchmark for Graded Change Detection**. Overview of time periods, diachronic corpus composition, number of targets, and benchmark versions used in this study.

	BERT	**mBERT**	**XLM-R**	**XL-LEXEME**
English	`bert-base-uncased`	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`
Latin	-	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`
German	`bert-base-german-cased`	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`
Swedish	`af-ai-center/bert-base-swedish-uncased`	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`
Spanish	`dccuchile/bert-base-spanish-wwm-uncased`	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`
Russian	`DeepPavlov/rubert-base-cased`	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`
Norwegian	`NbAiLab/nb-bert-base`	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`
Chinese	`bert-base-chinese`	`bert-base-multilingual-cased`	`xlm-roberta-base`	`pierluigic/xl-lexeme`

Table 7: BERT, mBERT, XLM-R, and XL-LEXEME models employed in our evaluation. All models are available at `huggingface.co`.

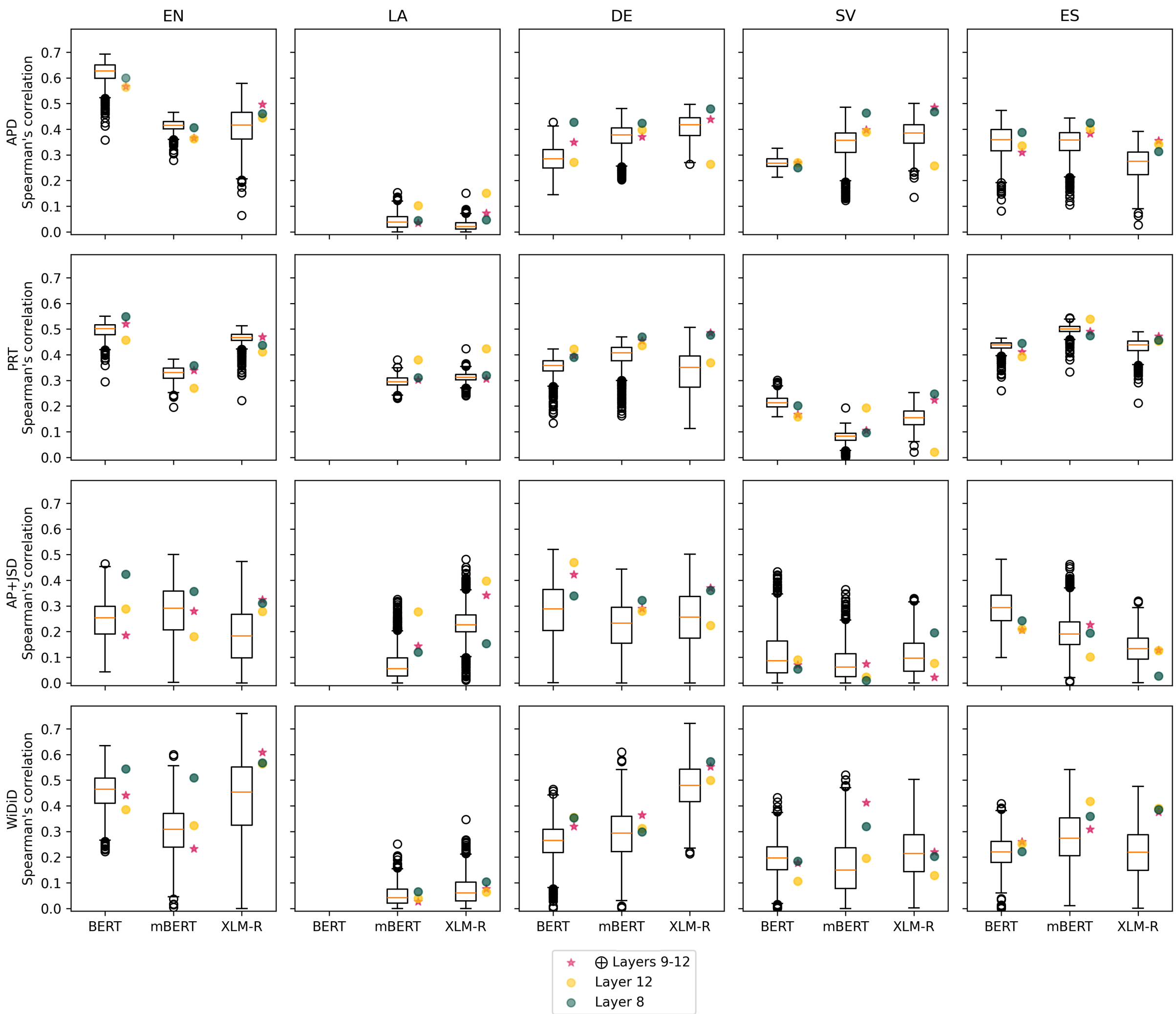

Figure 2: **Score distribution for GCD** obtained by using all possible layer combinations of length 2 (e.g., Layer 1 and 2), length 3 (e.g., Layer 10, 11, 12), and length 4 (e.g., Layer 1, 10, 11, 12) for BERT, mBERT, and XLM-R. The y-axis represents the Spearman correlation. We highlight the performance for GCD obtained using Layer 8, Layer 12, and the sum of the last 4 layers (i.e., $\bigoplus$ 9-12).

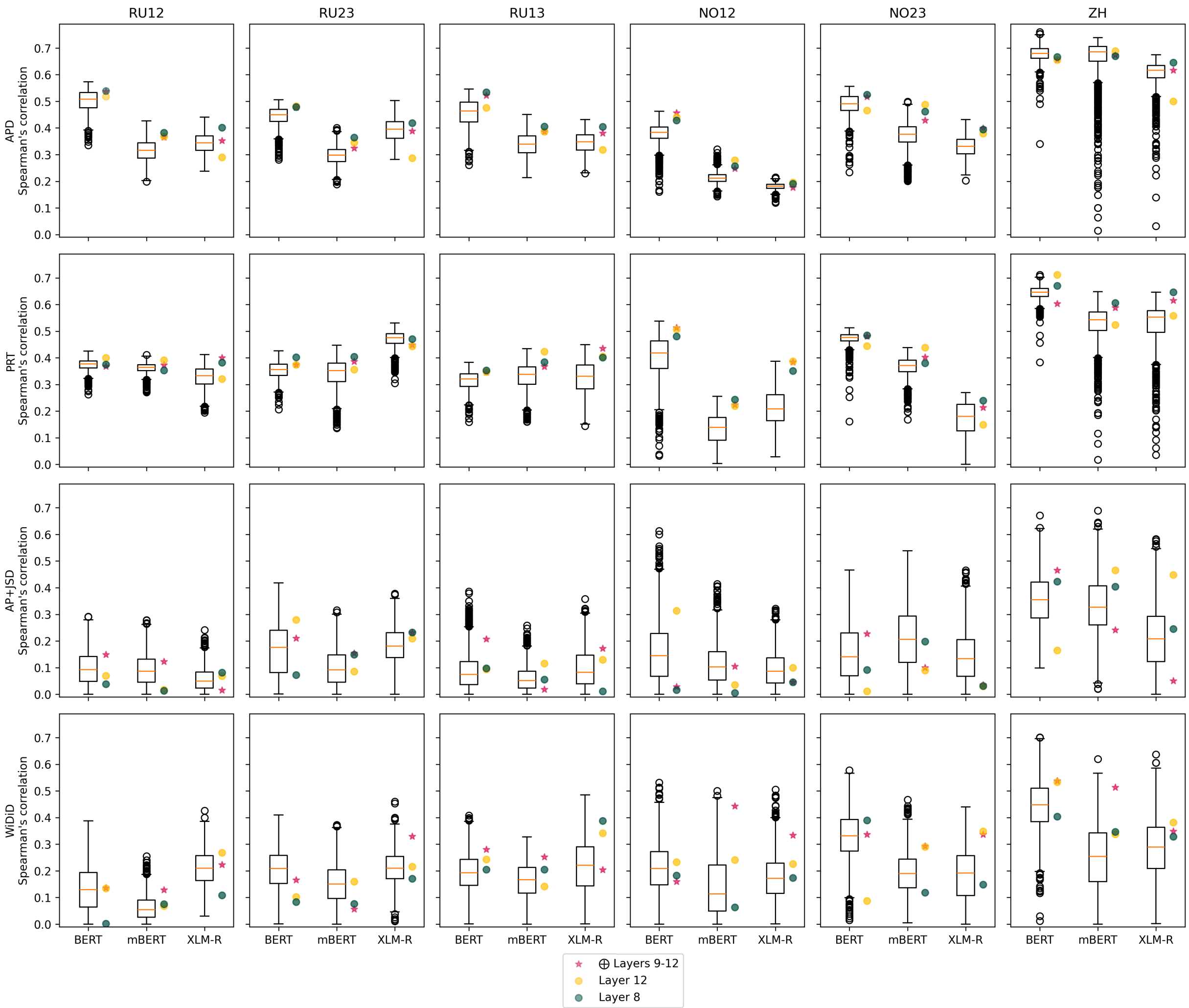

Figure 3: **Score distribution for GCD** obtained by using all possible layer combinations of length 2 (e.g., Layer 1 and 2), length 3 (e.g., Layer 10, 11, 12), and length 4 (e.g., Layer 1, 10, 11, 12) for BERT, mBERT, and XLM-R. The y-axis represents the Spearman correlation. We highlight the performance for GCD obtained using Layer 8, Layer 12, and the sum of the last 4 layers (i.e., $\bigoplus$ 9-12).

]

			EN	LA	DE	SV	ES	RU			NO		ZH
			$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_3$	$C_1 - C_2$	$C_2 - C_3$	$C_1 - C_2$
form-based	APD	BERT	**.692** / .566 *(.563)*	/	.412 / .349 *(.271)*	.325 / .272 *(.270)*	**.488** / .310 *(.335)*	**.573** / .537 *(.518)*	**.506** / .477 *(.482)*	**.546** / .522 *(.476)*	**.463** / .457 *(.441)*	**.556** / .521 *(.466)*	**.760** / .658 *(.656)*
		mBERT	.466 / .365 *(.363)*	.136 / .034 *(.102)*	.468 / .370 *(.398)*	.486 / .398 *(.389)*	.423 / .351 *(.341)*	.419 / .365 *(.368)*	.393 / .324 *(.345)*	.443 / .386 *(.386)*	.320 / .248 *(.279)*	.496 / .429 *(.488)*	.739 / .674 *(.689)*
		XLM-R	.579 / .518 *(.444)*	.080 / -.072 ***(.151)***	**.496** / .438 *(.264)*	**.496** / **.496** *(.257)*	.443 / .398 *(.386)*	.441 / .368 *(.290)*	.491 / .404 *(.287)*	.432 / .397 *(.318)*	.215 / .180 *(.195)*	.421 / .418 *(.379)*	.675 / .627 *(.500)*
	PRT	BERT	**.550** / .520 *(.457)*	/	.421 / .397 *(.422)*	**.293** / .170 *(.158)*	.478 / .441 *(.413)*	**.425** / .368 *(.400)*	.418 / .374 *(.374)*	.383 / .346 *(.347)*	**.538** / .513 *(.507)*	**.513** / .481 *(.444)*	.706 / .649 *(**.712**)*
		mBERT	.382 / .339 *(.270)*	.352 / .305 *(.380)*	.467 / .454 *(.436)*	.132 / .105 *(.193)*	**.555** / .514 *(.543)*	.411 / .373 *(.391)*	.442 / .386 *(.356)*	.434 / .367 *(.423)*	.256 / .228 *(.219)*	.432 / .405 *(.438)*	.648 / .588 *(.524)*
		XLM-R	.513 / .476 *(.411)*	.365 / .312 ***(.424)***	**.497** / .486 *(.369)*	.253 / .236 *(.020)*	.538 / .522 *(.505)*	.409 / .402 *(.320)*	**.530** / .453 *(.443)*	**.449** / .435 *(.405)*	.384 / .384 *(.387)*	.270 / .220 *(.149)*	.642 / .627 *(.558)*
sense-based	AP	BERT	.464 / .245 *(.289)*	/	**.520** / .435 *(.469)*	.201 / -.061 *(-.090)*	**.499** / .295 *(.225)*	.292 / .149 *(.069)*	**.418** / .216 *(.279)*	**.386** / .207 *(.094)*	**.329** / .028 *(.314)*	.466 / .227 *(.011)*	**.671** / .587 *(.165)*
		mBERT	**.501** / .313 *(.181)*	.326 / .179 *(.277)*	.428 / .329 *(.280)*	.193 / .090 *(.023)*	.484 / .259 *(.067)*	**.209** / .123 *(.017)*	.316 / .175 *(.086)*	.247 / .058 *(-.116)*	.194 / -.105 *(.035)*	**.539** / .275 *(-.090)*	.645 / .256 *(465)*
		XLM-R	.473 / .340 *(.278)*	**.482** / .398 *(.398)*	.502 / .370 *(.224)*	**.235** / .022 *(-.076)*	.307 / .170 *(.224)*	.162 / .012 *(-.068)*	.378 / .247 *(.209)*	.358 / .224 *(.130)*	.322 / .132 *(-.100)*	.465 / .035 *(.030)*	.583 / .135 *(.448)*
	WiDiD	BERT	.635 / .441 *(.385)*	/	.465 / .322 *(.355)*	.432 / .177 *(.106)*	.466 / .361 *(.383)*	.388 / .136 *(.135)*	.410 / .190 *(.102)*	.408 / .280 *(.243)*	**.531** / .160 *(.233)*	**.578** / .336 *(.087)*	**.701** / .537 *(.533)*
		mBERT	.600 / .317 *(.323)*	.252 / .055 *(-.039)*	.610 / .422 *(.312)*	**.521** / .413 *(.195)*	**.575** / .272 *(.343)*	.255 / .215 *(-.068)*	.373 / .056 *(.160)*	.327 / .252 *(.142)*	.500 / .459 *(.241)*	.467 / .292 *(.290)*	.620 / .513 *(.338)*
		XLM-R	**.760** / .663 *(.564)*	**.347** / -.077 *(-.064)*	**.721** / .557 *(.499)*	.503 / .220 *(.129)*	.526 / .437 *(.459)*	**.426** / .223 *(.268)*	**.460** / .352 *(.216)*	**.485** / .304 *(.342)*	.505 / .399 *(.226)*	.440 / .336 *(.349)*	.637 / .349 *(.382)*

Table 8: **Top score for GCD** obtained using BERT, mBERT, and XLM-R. We present results for the optimal combination and the outcome obtained by summing the last four layers, separated by a slash (i.e., best results **/** sum of last four layers). Additionally, for comparison purposes, we include the result obtained using the last layer individually *(enclosed in brackets).*. Top scores for approach and benchmark are highlighted in **bold**.