Analyzing Semantic Change through Lexical Replacements

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

Pierluigi Cassotti*
University of Gothenburg
Renströmsgatan 6
40530 Gothenburg, Sweden
pierluigi.cassotti@gu.se

Haim Dubossarsky
Queen Mary University of London
Mile End Road
E1 4NS London, England
h.dubossarsky@qmul.ac.uk

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

Abstract

Modern language models are capable of contextualizing words based on their surrounding context. However, this capability is often compromised due to semantic change that leads to words being used in new, unexpected contexts not encountered during pre-training. In this paper, we model *semantic change* by studying the effect of unexpected contexts introduced by *lexical replacements*. We propose a *replacement schema* where a target word is substituted with lexical replacements of varying relatedness, thus simulating different kinds of semantic change. Furthermore, we leverage the replacement schema as a basis for a novel *interpretable* model for semantic change. We are also the first to evaluate the use of LLaMa for semantic change detection.

1 Introduction

The major advancement that novel Language Models (LMs) have brought is the ability to dynamically generate contextualized representations (i.e., embeddings) based on specific usage context. When words are used in contexts similar to those encountered during training, LMs can easily differentiate, in a computational way, between word meanings. Like in the case of *rock* in the sentences *sitting on a rock* and *listening to rock*.

However, when an existing word in our vocabulary gains a new meaning through semantic change, LMs' ability to differentiate that meaning can be affected. This stems from the fact that semantic change is evidenced through new contexts that were previously unknown for the word. Sometimes, the new meaning is novel to the dictionary, for example, the metaphorical Web-meaning of *surfing*. Other times, the meaning is already in existence and get the word as a new referent. This is, for example, the case for *happy*. It used to mean exclusively `to be lucky` and then gained the meaning of `happiness`. In an inverse process, the word *gay* lost its meaning of `happiness` and began to refer exclusively to `homosexuality`. One can think of this process of *semantic change* to be a *lexical replacement* of the word *happy* into the context of *gay*, like in the following sentence[1]

The heart is sportive, light, and **gay**,
life seems a long glad summer's day[2]

When using LMs, the representation of a word w is based on (i) the pre-trained knowledge that the model has about w given its position in the context, and (ii) the context c in which w is used. Thus, when this replacement happens, LMs experience a *tension* between the existing sense/s of *happy* (which do not include `happiness`) and the meaning of the new context (which does indicate `happiness`). Due to semantic change, LMs do not known the relationship between the new context c and the replacement word r. As a consequence, the representation of r (i.e., *happy* in the sense `to be lucky`) and the representation of c (i.e., the context of *gay* in the sense of `happiness`) pull in different directions challenging the LMs' ability to contextualize (Ethayarajh, 2019).

The heart is sportive, light, and happy,
life seems a long glad summer's day

The tension increases as the gap between the data used for training the model, and the data on which the model is applied grows larger. Indeed, the LMs we use serve as the lens through which we view the studied texts: if our texts are contemporary with the pre-training, the gap is likely to be minimal. If, however, we intend to study historical or other out-of-domain corpora through LMs trained on modern text, this gap can be arbitrarily large and have major

* These authors contributed equally

[1] We are aware that happy gained the meaning of happiness several hundred years before gay lost its sense of happiness, and only use the example for illustrative purposes.

[2] Manchester Times, Wednesday 03 May 1854, found via https://discovery.nationalarchives.gov.uk.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4495–4510
August 11-16, 2024

effects on follow-up studies. Thus, using LMs for modeling relationships beyond their pre-trained knowledge will likely result in an underestimation of semantic change.

Our contributions: In this paper, we propose a replacement schema to study the tension experienced by LMs when words undergo semantic change. Such schema involves replacing a word w in the context c with a replacement r to analyze how the representation of r differs from the original representation of w.

Given a word w, our experiments systematically show that LMs (i.e., BERT, mBERT, XLM-R) experience a tension between the pre-trained knowledge of w and the new context of a gained meaning. This tension differs across linguistic relations, namely synonymy, antonymy, and hypernymy.

We then use the introduced schema for detecting semantic change. Our experiments show that, when random replacements are used to simulate *synthetic* semantic change, the use of a clustering algorithm (i.e, Affinity Propagation) falls short to differentiate meanings and detect such change. Furthermore, we use the replacement schema to introduce a new *interpretable* model for semantic change detection, while being comparable with state-of-the-art for English.

Finally, to further investigate the tension of LMs and semantic change, we compare the use of a predefined set of *replacements* with word *substitutes* generated by LMs (i.e., BERT, LLaMa 2). Our experiments show that smaller LMs are less able to provide substitutes that handle changing contexts, while LLaMa 2 significantly outperform the LMs. Notably, to the best of our knowledge, this represents the first experiment in the current literature to use LLaMa 2 to model semantic change and only the third paper to use any generative model (Periti et al., 2024; Periti and Tahmasebi, 2024).

2 Related Work

For this paper, relevant work pertains both to contextualization of modern LMs and the field of lexical semantic change.

Modern *contextualized* LMs leverage the Transformer architecture to capture the semantics of words (Vaswani et al., 2017). Their success in solving NLP tasks has prompted numerous studies to explore the nature and characteristics of their *contextualization* ability. Ethayarajh (2019); Coenen et al. (2019); Cai et al. (2021); Jawahar et al. (2019) shed light on the geometry of the embedding space. Serrano and Smith (2019); Bai et al. (2021); Guan et al. (2020) investigate the interpretability of the attention mechanism. Yenicelik et al. (2020); Garí Soler and Apidianaki (2021); Kalinowski and An (2021); Haber and Poesio (2021) examine the clusterability of word representations. Abdou et al. (2022); Hessel and Schofield (2021); Mickus et al. (2020); Wang et al. (2021) analyze the impact of word position in the embeddings generation. Coenen et al. (2019); Levine et al. (2020); Pedinotti and Lenci (2020) study how word meaning are represented in the embedding space.

Recent work have focused on a related aspect, namely adapting LMs to improve their *temporal* contextualization. This challenge has been addressed across various applications such as named entity recognition (Rijhwani and Preotiuc-Pietro, 2020), fake news detection (Hu et al., 2023), text summarization (Cheang et al., 2023), and lexical semantic change (Su et al., 2022; Rosin et al., 2022; Rosin and Radinsky, 2022). Nonetheless, while temporal domain adaptation can improve performance across various tasks, Agarwal and Nenkova (2022) demonstrated that temporal contextualization may not always be a concern. In our work we complement existing research by using lexical replacements as a proxy to analyze how language models contextualize words that have undergone semantic change.

Lexical Semantic Change (LSC) is the task of automatically identifying words that change their meaning over time (Schlechtweg et al., 2020). The task is recently gaining more and more attention, acting as evaluation task to assess the capability of LMs in capturing word meanings across diachronic corpora (Montanelli and Periti, 2023; Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021). Current benchmarks consist of a diachronic corpus spanning two time periods t_1 and t_2 and a reference set of target words T annotated with a degree of semantic change between t_1 and t_2 (Zamora-Reina et al., 2022; Kutuzov and Pivovarova, 2021; Schlechtweg et al., 2020; Basile et al., 2020).[3] The main goal is to rank the target words in T according to their degree of semantic change.

Currently, contextualized embeddings represents

[3] Kutuzov and Pivovarova, 2021 introduced a benchmark encompassing two time intervals. However, these intervals have been treated independently, leading to their consideration as two distinct sub-benchmarks over a single time interval.

the state-of-the-art solution for addressing LSC. Approaches to LSC relying on contextualized embeddings are typically distinguished into two main categories: *form-based* approaches and *sense-based* approaches (Montanelli and Periti, 2023). The former captures semantic change by solely relying on similarities among raw embeddings without depending on sense disambiguation and representation. Given a word w, a common strategy involves averaging all the embeddings of w from t_1 and all the embeddings of w from t_2, and modeling the change as the cosine similarity of the average representations **(PRT)** (Martinc et al., 2020a). The latter generally use clustering algorithm like Affinity Propagation to identify senses and subsequently model the change as divergence of cluster distributions **(JSD)** (Martinc et al., 2020b).

However, when form-based approaches are employed, interpreting the detected change is not supported as they do not model each individual sense of a word. In contrast, when sense-based approaches are employed, they typically represent clusters of word usages rather than word meanings (Kutuzov et al., 2022). As a result, although a new powerful LMs has recently been introduced for modeling the semantics of words (Cassotti et al., 2023), the *interpretation* of which meaning of a word has changed and in which way, remains an open challenge.

In this regard, our work is related to the novel substitute-based approaches to LSC, which interpret word meaning by generating substitutes of words in context (Cuscito et al., 2024; Card, 2023; Kudisov and Arefyev, 2022; Arefyev and Zhikov, 2020). On one hand, word substitutes represents relevant keywords to aid the interpretation of senses. On the other hand, the generation process can only provide substitutes according to training data, and as we show in Section 5.3, LMs lack the knowledge to adapt to semantic changes.

To this end, we propose a novel *interpretable* approach based on a pre-defined set of lexical *replacements* rather than generated *substitutions*.

3 Methodology

In our experiments, we leverage a replacement schema to investigate the tension experienced by pre-trained LMs due to semantic change. This involves analyzing the variations in embedding representations when a target replacement is introduced. For instance, by replacing a target like *cat* with a replacement like *chair* in a specific context like:

$$\text{The } \underset{\text{target}}{cat} \leftarrow \underset{\text{replacement}}{chair} \text{ was purring loudly .}$$

3.1 The replacement schema

We use WordNet to generate different classes of replacements for a specific word (Fellbaum, 1998), which correspond to a varying degree of plausibility (i.e. suitability of a specific replacement) between the target word and its replacement. Thus, we hypothesize that each class is associated with a different impact on contextualization. Each class of replacements also has diachronic relevance, as the synchronic, semantic relation can be considered to have a parallel in semantic change (de Sá et al., 2024; Wegmann et al., 2020). To ensure accurate linguistic replacements, we maintain part of speech (PoS) agreement with the target words; e.g., *nouns* are replaced with *nouns*.

- **synonyms** (e.g. *sadness* ← *unhappiness*) are used to evaluate the stability in contextualization; that is, we hypothesize similar embeddings between target and replacement words. Indeed, synonyms are considered equally likely alternatives in LM's pre-trained knowledge. On the diachronic level, they emulate the absence of any semantic change of the replacement word;
- **antonyms** (e.g. *hot* ← *cold*) are used to evaluate a light change in contextualization; that is, we hypothesize slightly less similar embeddings between target and replacement words. Indeed, antonyms are sometimes equally plausible alternatives, for example: "I *love/hate* you". Other times they are likely to surprise the model. For example: "I burned my tongue because the coffee was too *hot/cold*". On the diachronic level, they emulate a contronym change. A contronym change occurs when a word's new meaning is the opposite of its original meaning (e.g. *sanction* in English) of the replacement word;
- **hypernyms** (e.g. *animal* ← *bird*) are used similarly to `antonyms`. However, on the diachronic level, they emulate a broadening semantic change of the replacement word;
- **random** words (e.g. *sadness* ← *eld*) are used to evaluate a change in contextualization. If LMs place high importance on the context, then the replacement should receive a similar

representation to the target word. Otherwise, if LMs heavily rely on its pre-trained knowledge, the replacement will exhibit dissimilarity to the target word despite the identical context, as well as dissimilarity to the typical replacement representations. On the diachronic level, `random` emulates the presence of strong semantic change of the replacement word, that is, the emergence of a homonymic sense.

3.2 Data

To avoid introducing noise into our experiments resulting from the conflation of senses, we replace words with contextually appropriate replacements based on the intended sense of the word within a specific sentence (e.g, *stone* and *music* for *sitting on a rock* and *listening to rock*, respectively). We therefore leverage the SemCor dataset (Miller et al., 1993), still the largest and most commonly used sense-annotated corpus for English. To select candidate replacements, we consider different PoS tags, namely *verbs*, *nouns*, *adjectives* and *adverbs*, and semantic classes, namely *synonyms*, *hypernyms* and *antonyms*. We randomly sample a set of synsets for each PoS tag occurring in SemCor, and for a specific synset, we extract a subset of contexts (i.e., sentences) where a word is annotated with that synset. We sample a maximum of 10 sentences per synset to prevent oversampling of high-frequency synsets. We control for the position of the replaced target has in the sentence, and the length of the sentence, to confirm that these aspects will not bias our experiments differently across PoS. For each sentence, we generate the *synonym* and *antonym* replacements for all PoS, and *hypernym* replacements only for nouns and verbs because WordNet lacks hypernym information for other PoS (see Table 1).

PoS	N. target words	Avg. N. of sampled senteces per target word	N. examples
noun	360	3.55	1277
verb	433	3.45	1494
adjective	393	3.39	1334
adverb	158	3.46	546

Table 1: Data statistics over PoS, sampled from SemCor.

Experimental setup We begin by studying the tension that occurs as a consequence of replacement focusing on the word contextualization in Section 4. Next, we the use of replacements as a proxy for semantic change in Section 5. In our experiments, we use monolingual BERT[4], mBERT[5], and XLM-R[6]. Our code and data are available at https://github.com/ChangeIsKey/asc-lr/.

4 Tension caused by semantic change

We analyze the tension experienced by LMs by comparing the embedding of a target word w in the original sentence c to the embedding of the replacement word r in the same sentence c. To perform this comparison, we rely on the cosine distance between the embeddings of w and r. We refer to this as the *self-embedding distance* (SED).

Concretely, if w and r are split into multiple sub-words by the model, we calculate the average embeddings of the corresponding sub-words. This approach ensures the preservation of the same number of tokens in the original and synthetic sentences and enables accurate distance calculations.

The less plausible the relationship between the context c and the replacement word r for LMs, the higher the SED, leading them to rely on the pre-trained knowledge of r to contextualize r in context. When there is a large mismatch between the meanings of the replacement word r and the context c, as is the case with the random replacement, then the SED is the highest.

4.1 Self-embedding distance

For each pair of original and synthetic sentences, we computed SED across each layer. We then analyzed the average SED for each class of replacement and PoS across the layers of LMs. It is know that contextualized embeddings experience an anisotropic nature, that is, the embeddings occupy a increasingly narrow cone within the vector space (Ethayarajh, 2019). This means that embeddings, and thus SED scores across layers, are not comparable. To address this issue and thus compare SED both across layers and PoS, we use a layer-specific normalization factor.

Specifically, for normalization, we randomly sampled an additional set of 3864 sentences independent from the sets in Table 1. For each sentence, we randomly choose a target word and replace it with a `random` replacement regardless of the PoS agreement. Then, for each layer, we computed the average SED over this set of replacements. We use the resulting SED scores as a normalization factor for each layer that represents an upper bound approximation. Thus, for each layer, the same normal-

[4] *bert-base-uncased* [5] *bert-base-multilingual-cased*
[6] *xlm-roberta-base*

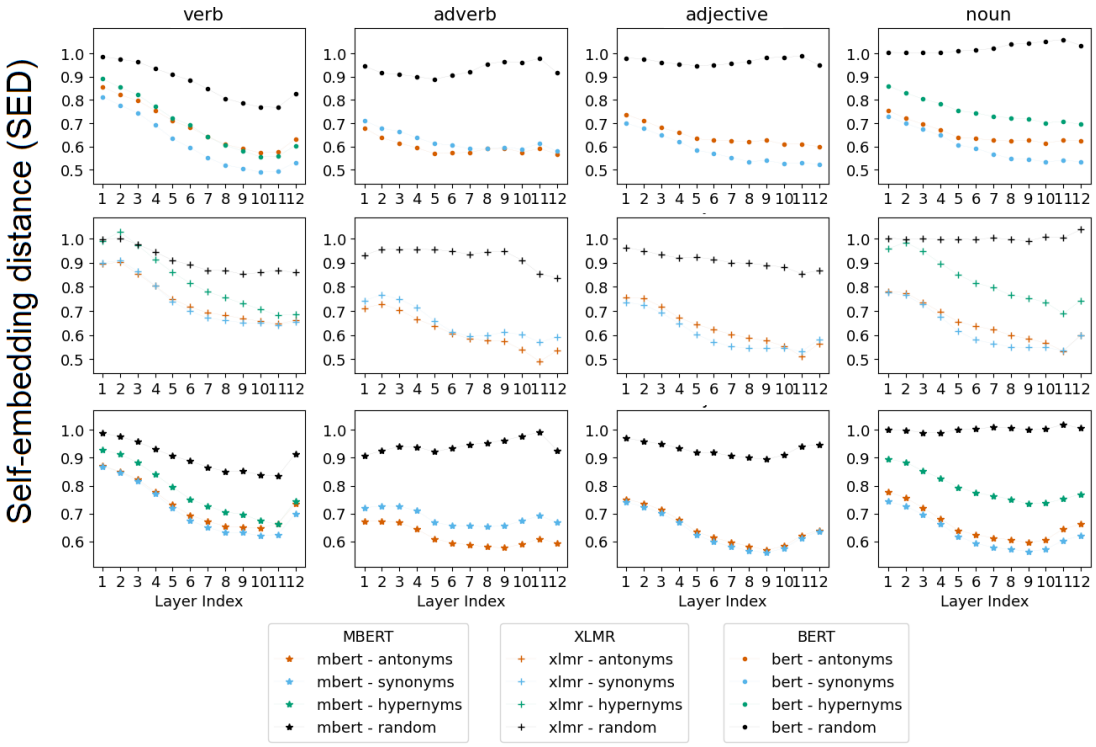

Figure 1: Average SED over layers.

ization factor is used across all PoS and semantic class of replacement. This way, the normalization cannot influence the discrepancies among different classes for a specific layer but serves to make the scores in different layers somewhat comparable.[7]

Like Ethayarajh (2019), we observe that the contextualization increases across layers as the SED decreases, the context thus has a larger effect in determining the representation of a word in the higher layers. For `adverbs`, `adjectives` and `nouns` the synonym and antonym classes are associated with an SED of around 0.6–0.8 in the first layer. The SED then decreases to between 0.5–0.6. For `adverb` the synonym and antonym class remain similar also in the later layers, while for `adjectives` and `nouns` we find that the synonyms have lower SED than do antonyms. For `nouns`, the hypernym class has consistently higher SED than synonyms and antonyms, despite being a more general concept where the subconcept of the target word should be contained (e.g., *fruit* as a hypernym of *banana*). This aligns with the recent findings of Hanna and Mareček (2021), suggesting that BERT's understanding of noun hypernyms is limited.

The SED score for `random` is fairly stable across all layers, meaning that when a word gains a completely novel sense, LMs fall short in contextualizing beyond the pre-trained knowledge it has of the word. That is, the representation of the random word does not mimic the representation of the target word that it replaces. The context thus has little or no effect in determining the representation of the replacement word.

For `verbs`, we note a higher SED for antonyms and synonyms in comparison to other PoS, comparable to the `noun` hypernyms, starting around 0.9. However, they all drop to 0.6–0.7 by the last layers. Additionally, there is a narrower gap between the SED for the random class and those for antonyms, synonyms, hypernyms. These observations suggest that, in the earlier layers, the contextualization of verbs is less pronounced for `verbs` and that the model relies more on pre-trained knowledge.

All in all, our results suggest that models exhibit varying tension for different PoS, and for different linguistic relationships between the target and the replacement word. Conversely, we interpret these findings in the following way: there is a low degree of contextualization, and thus a high degree of tension, when there is no relationship between the word and its replacement.

[7] We have tested with different normalization factors – e.g., replacing a word with a special token ("`[REPL]`") outside the LMs vocabulary – and found that the conclusions remain.

5 Semantic change

We argue that our findings in Section 4 regarding the tension between a word and its context has important implications when pre-trained LMs are used for modeling semantic change as we will show in this section.

5.1 LCS through synthetic dataset

Form-based approaches can still detect this semantic change to a certain degree (as an estimate of model confusion), despite using contextualized word embeddings that are not correctly capturing a word's meaning in a novel context. However, *sense-based* approaches fall short in accurately detecting the same change. This is because *sense-based* approaches require modeling meanings outside the model's pre-trained knowledge before detecting the change. Since these meanings cannot be adequately modeled when semantic change has occurred, the performance of *sense-based* approaches is reduced compared to that of *form-based* approaches.

We further tested these implications in the LSC task by comparing PRT (based on *averaging* contextualized embeddings) and JSD (based on *clustering* contextualized embeddings) on an artificial diachronic corpus spanning two time periods (see details in Appendix A). Essentially, we introduced random replacements in C_2 with varying probabilities to emulate different degrees of change for a set of 46 target words. Subsequently, we compared the Spearman Correlation between the scores obtained with PRT and JSD with the artificially graded score of emulated semantic change. Results using BERT are presented in Figure 2 (see Appendix A for additional results). Our hypothesis is that while PRT can predict changes to a fairly high degree, JSD falls short because it can only correctly model the meanings that BERT is already aware of.

As shown in the figure, using PRT, we can model artificial semantic changes already from layer 3. This is not the case for JSD, where we observe statistically significant correlations for only a few layers. However, the significance of performance for JSD is an artifact of BERT embeddings and does not authentically represent the simulated change. We verify this by examining the modeled clusters. While, in general, the number of clusters of AP is large (Periti et al., 2022; Martinc et al., 2020b), representing *sense nodules*[8] rather than word meanings (Kutuzov et al., 2022), we find that the injected confusion in the model due to the `random` replacements results in a very low number of clusters (typically 2, maximum of 4). We report similar results in Appendix for other languages (i.e. German, Swedish, Spanish)

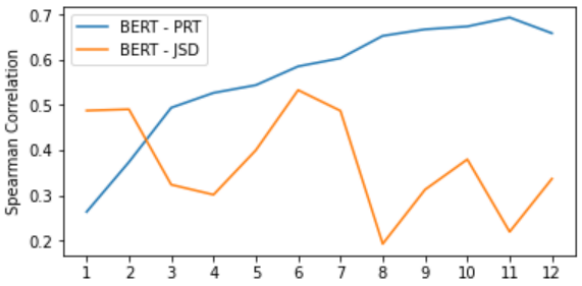

Figure 2: Spearman Correlation over layers for artificial semantic change.

5.2 LSC through replacements

As a further contribution of this paper, we propose a novel supervised approach to Graded Change Detection building upon the replacement schema. Our approach leverages a curated set of word replacements from WordNet and Wiktionary.

We denote $T = \{w_1, w_2, ..., w_N\}$ as the set of target words. For each target word, we extract a set of possible replacements $\rho(w_i) = \{r_1, r_2, ..., r_M\}$, resulting in $N \cdot M$ replacement pairs. The set of replacements is obtained by considering the lemmas of synonyms and hypernyms associated with the target word w_i in WordNet and words extracted from the Wiktionary page corresponding to the target word. For each target word w_i, we sample up to 200 sentences from each period that remain stable regardless of the replacement word r_j. For each replacement pair (w_i, r_j), we denote the set of sentences for a time period $t \in \{1, 2\}$ as $S^t(w_i, r_j)$.

For each sentence $s \in S^t(w_i, r_j)$ we measure the self-embedding distance $sed(s)$of the target and replacement word. The average self-embedding distance of a target-replacement pair is defined as

$$awd^t(w_i, r_j) = \frac{1}{|S^t(w_i, r_j)|} \sum_{s \in S^t(w_i, r_j)} sed(s)$$

The absolute difference in awd over time is denoted $\text{TD}(w_i, r_j)$. Finally, we rank the replacements $\rho(w_i)$ according to their degree of time difference:

$$R(\rho(w_i)) = \{r_1, r_2, ..., r_M | \text{ TD}(w_i, r_{i+1},) \leq \text{TD}(w_i, r_i)\}$$

and we compute a semantic change score lsc_w as

[8] Lumps of meaning with greater stability under contextual changes (Cruse, 2000)

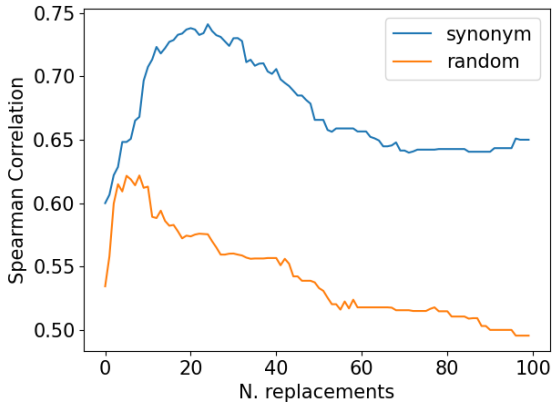

Figure 3: Top-k replacement vs Spearman Correlation.

	Model	Spearman Correlation
	Rosin and Radinsky	0.629
	Kutuzov and Giulianelli	0.605
	Laicher et al.	0.571
	Periti et al.	0.512
	Cassotti et al. (XL-LEXEME)	0.757
Synonym Replacement	Replacement Min. Corr.	0.600
	Replacement Max. Corr.	0.741
	Replacement Avg. Corr.	0.674
Random Replacement	Replacement Min. Corr.	0.495
	Replacement Max. Corr.	0.622
	Replacement Avg. Corr.	0.542

Table 2: Spearman Correlation on SemEval-2020 Task 1 (Eng)

the average TD considering the top k replacements:

$$lsc_w = \frac{1}{k} \sum_{r \in R(\rho(w_i))_k} \text{TD}(w_i, r)$$

We evaluate our approach on the SemEval-2020 Task 1, Subtask 2 dataset for English. We compute the Spearman Correlation between the graded score reported in the gold truth and the lsc scores. Figure 3 reports the correlation computed for different values of k. The highest correlation of 0.741 is achieved when considering the first 22 replacements, while the lowest correlation of 0.600 is obtained using only the first replacement (see Table 2). Interestingly, the minimum correlation obtained using the replacements is competitive with SOTA results. Moreover, on average, the correlation is higher than the SOTA model's performance. The replacements are reported in Table 5.

By replacing the target words with different semantically related words, we generate contextual variations that enable the detection of semantic shifts. In the case of words like *record* (attainment, track record $\longrightarrow$ evidence, document) and *land* (real estate, real property $\longrightarrow$ realm, country) that have undergone semantic change through narrowing and generalisation, respectively, linguistically aware replacements can provide valuable insights. The replacement process generates a list of replacements that can be used as labels for the types of semantic change observed. By associating each replacement with a specific semantic category or change type, it becomes possible to analyze and quantify the semantic shifts experienced by words over time. The method can also be combined with a priori clustering to get changes specific to a sense.

Random replacements Here, we focus on the results using randomly selected words with the same PoS as the target word, i.e. `random` replacement as introduced in Section 3. This approach generates a list of replacement words contextually unrelated to the target word. Some interesting patterns emerge when these results are compared with those obtained using synonym replacement. In the case of semantic change detection, the use of synonyms can provide more contextually relevant replacements, as they share semantic relationships with the target word. However, using random replacements can still yield reasonable results, as evidenced by an average correlation of 0.542. These results is in line with the finding of Section 5.

In this approach, although random replacements tend to perform worse than synonym replacements, they have one distinct advantage: they do not rely on external lexical resources and are thus suitable for unsupervised scenarios. While synonym replacements can improve contextualization and semantic relevance, they are not always readily available or reliable for languages with limited linguistic resources. In such cases, random replacements can still provide reasonable results and serve as a practical, resource-efficient approach for tasks where synonym information is scarce or unavailable.

In Section 4.1, when using SemCor, we effectively account for the nuances of different word senses, thereby improving the contextualization and semantic relevance of synonym replacements. This approach is more targeted as synonyms are selected based on their association with a particular sense, leading to higher quality contextualization in the context of that sense. As a result, synonym replacements are more finely tuned to the specific meaning of the target word, reducing noise and improving correlation with semantic change labels.

5.3 LSC through substitutions

Finally, we assess the use of lexical substitutes generated by LMs for LSC. By asking LMs' to generate substitutions, we probe them for their information about the target word given the context. Similar to Card (2023); Arefyev and Zhikov (2020), we use monolingual BERT. We additionally compared the use of a larger, generative model such as LLaMa 2 7B (Touvron et al., 2023)[9].

For BERT, we use the masking strategy, meaning that we mask a target word with the special token and generate possible substitutes. For LLaMa 2, we fine-tune the model to enable it to predict the target word. Specifically, we fine-tune LLaMa 2 by inputting the original sentence, adding two asterisks at the beginning and end of the target word. Following the sentence we provide the list of substitutes found in ALaSCA (Lacerra et al., 2021), the largest existing dataset for lexical substitution:

During the siege, George Robertson had appointed Shuja ul-Mulk, who was a **bright** boy only 12 years old and the youngest surviving son of Aman ul-Mulk, as the ruler of chitral. **|answer|** *intelligent* **|s|** *clever* **|s|** *smart* **|end|**

where **|answer|**, **|s|**, and **|end|** are added as special tokens in the model. For efficiency reasons, we train the model using the QLoRA paradigm[10]. We fine-tuned for one epoch using a learning rate of 2e-4, and set the LoRA configuration with a rank of 8 and an alpha of 16.

The data used for the evaluations is the same in Section 5.2. In Table 6 we report an example of the generated substitutions. To calculate the degree of semantic change, we consider all uses of a word in time periods t_1 and t_2. We consider the substitutes generated for each usage and calculate the distance between all possible pairs of uses between t_1 and t_2. To calculate the distance, we use the Jaccard Distance between the sets of generated substitutes. Lastly, the Jaccard distances are averaged, and we use the average as a score for LSC. In Table 3 we show the result on the SemEval 2020 Task 1 - Subtask 2 (other comparable results in Table 2). Our results for BERT are somewhat comparable with SOTA results, while being lower to those obtained through lexical replacements, likely because the replacements are of higher quality when found using WordNet, while the substitutions are generated by the model with its limited knowledge of the context. In contrast, our results for LLaMa 2 are even higher than the results obtained with lexical replacements achieving comparable performance to the one obtained with XL-LEXEME. We attributed this higher performance to the fact that both LLaMa and XL-LEXEME have been fine-tuned on generating lexical substitutes and WiC task, respectively which, rather than using all of the model's pretrained knowledge, forces the model to focus on the semantic aspect specifically.

Model	Spearman Correlation
Arefyev and Zhikov (2020)	0.299
Card (2023)	0.547
LLaMa 2 7B	**0.731**
BERT	*0.450*

Table 3: Spearman Corr. on SemEval-2020 Task 1 (EN)

6 Discussions and Conclusions

In this paper, we study semantic change using lexical replacement. From the point of view of the replaced word, a semantic change takes place as the word gains contexts which it has not encountered previously. When the replacement is closely related to the target word, for example by synonymy, the novelty of the context for the replacement word should be low. However, novelty will increase as the relation between the target and replacement becomes more distant. We are assuming that the replacements based on synchronic relations will offer insights into semantic change diachronically.

To test this hypothesis, we used self-embedding distance (SED) when the context stays the same, using all layers of BERT, mBERT and XLM-R across four part of speech. Not surprising, we found that the self-embedding distance is smallest for synonym replacements and highest for the random replacements. And like Ethayarajh (2019), we found that more contextualization happens across the last layers. For the different models, we also find slightly different behaviors. However, consistently, adverbs and adjective have lower SED scores than verbs and nouns. We show that hypernymy is a more distant relation for LMs than antonymy and synonymy

We then employ replacements for measuring the degree of semantic change. For this, we generate synonym replacements using WordNet, for each word in the English portion of the SemEval-2020

[9] *meta-llama/Llama-2-7b-hf*

[10] Quantization and Low-Ranking Adaptation (Dettmers et al., 2023)

Task 1 benchmark. We assume that if a word has not experienced semantic change, the SED between the replacements and the target word are similar across time. If however, a word has experienced semantic change resulting in context changes, SED scores will be different over time as the replacements will be more distantly related to the contexts. This method offers a novel *interpretable* semantic change detection. Finally, we ask the LMs themselves to generate substitutions for a target word in the English SemEval data.

7 Limitations

A potential limitation of our study lies in the use of the replacement schema in conjunction with lexical replacements generated from WordNet: inherent limitations of WordNet, such as potential gaps, inaccuracies, or ambiguities in the semantic relationships may influence our analysis. WordNet also limits the data sources from which we can draw sentences, since we need a corpus with sense annotations corresponding to a lexicon.

Furthermore, in our experiment, the lexical replacement process involves replacing a *word* occurrence in the original sentence with a related *lemma* extracted from WordNet. As a result, providing the model with synthetic sentences containing the lemma instead of the inflected word may influence the generation of word embeddings and the contextualization of every word in the sentences. However, we assume that this limitation equally affects every class we consider and all models. For example, while the lemma of a verb may reduce the third singular verb form, the plural forms of adjectives and nouns can also be simplified to singular lemma forms. Additionally, to mitigate these issues and ensure that all PoS are equally affected by the replacement procedure, we replaced both the target and replacement words with lemmas in the original and synthetic sentences, respectively. We did not analyze semantic change in Section 5 with respect to different PoS because there are no available LSC benchmarks with a substantial number of targets for different PoS, nor any sense-tagged benchmarks except for a small subset for German.

Finding the correct form of a replacement requires advanced morphological analysis and carries the risk of leading to errors. For now, we therefore opted to circumvent this by replacing targets and lemmas alike. Furthermore, we would like to highlight a relevant study by Laicher et al. (2021) that delves into the influence of various linguistic variables on the use of BERT embeddings for the LSC task. This research demonstrates that by reducing the influence of orthography through lemma usage, significant enhancements in BERT's performance were observed for German and Swedish, while maintaining comparable results for English. This underscores the potential benefits of lemma-based contextualization and that linguistic features like orthography can sometimes be minimized without substantial loss of performance.

We use LLaMa 2 only for our last experiment. This stems from the difficulty to generate contexualized representation of a single word in context in LLMs. We also do not exhaustively test LLMs as this lies outside the scope of the paper, while requiring a lot more resources. Instead we use one open LLM to test the knowledge of a model when trained on significantly more data.

Acknowledgements

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.

We would also like to thank the anonymous reviewers for their helpful comments, and Stefano De Pascale, David Alfter, and Dirk Geeraerts for providing valuable feedback on the preliminary draft of this work, as well as for engaging in early discussions that contributed to the development of this research.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does matter and shuffled language models know it. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language pro-

cessing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

Nikolay Arefyev and Vasily Zhikov. 2020. BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change Detection. In *Proc. of SemEval*, pages 171–179, Barcelona (online). ICCL.

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why Attentions May Not Be Interpretable? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 25–34, New York, NY, USA. Association for Computing Machinery.

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 *Proc. of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, Online. CEUR-WS.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the Contextual Embedding Space: Clusters and Manifolds. In *International Conference on Learning Representations*.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association 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.

Chi Cheang, Hou Chan, Derek Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia Chao. 2023. Can LMs Generalize to Future Data? An Empirical Analysis on Text Summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16205–16217, Singapore. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and Measuring the Geometry of BERT. In *Proc. of NeurIPS*, Red Hook, NY, USA. Curran Associates Inc.

D. Alan Cruse. 2000. Aspects of the microstructure of word meanings. In Yael Ravin and Claudia Leacock, editors, *Polysemy: Theoretical and Computational Approaches*, pages 30–51. Oxford University Press.

Miriam Cuscito, Alfio Ferrara, and Martin Ruskov. 2024. How BERT Speaks Shakespearean English? Evaluating Historical Bias in Contextual Language Models.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proc. of EMNLP-IJCNLP*, pages 55–65, Hong Kong, China. ACL.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Yue Guan, Jingwen Leng, Chao Li, Quan Chen, and Minyi Guo. 2020. How Far Does BERT Look At: Distance-based Clustering and Analysis of BERT's Attention. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3853–3860, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Janosch Haber and Massimo Poesio. 2021. Patterns of Polysemy and Homonymy in Contextualised

Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Hanna and David Mareček. 2021. Analyzing BERT's Knowledge of Hypernymy via Prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 275–282, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? Implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online. Association for Computational Linguistics.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 116–125, Toronto, Canada. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proc. of ACL*, pages 3651–3657, Florence, Italy. ACL.

Alexander Kalinowski and Yuan An. 2021. Exploring Sentence Embedding Structures for Semantic Relation Extraction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Artem Kudisov and Nikolay Arefyev. 2022. BOS at LSCDiscovery: Lexical Substitution for Interpretable Lexical Semantic Change Detection. In *Proc. of LChange*, Dublin, Ireland. ACL.

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 and Lidia Pivovarova. 2021. RuShiftEval: A Shared Task on Semantic Shift Detection for Russian. In *Proc. of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*, 20, (online). Redkollegija sbornika.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. In *Northern European Journal of Language Technology, Volume 8*.

Caterina Lacerra, Tommaso Pasini, Rocco Tripodi, and Roberto Navigli. 2021. ALaSca: an Automated approach for Large-Scale Lexical Substitution. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3836–3842. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving Some Sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

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?*, page 343–349. Association for Computing Machinery (ACM), New York, NY, USA.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

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

Paolo Pedinotti and Alessandro Lenci. 2020. Don't Invite BERT to Drink a Bottle: Modeling the Interpretation of Metonymies Using BERT and Distributional Representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6831–6837, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 and Nina Tahmasebi. 2024. A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.

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, New York, NY, USA. Association for Computing Machinery.

Guy D. Rosin and Kira Radinsky. 2022. Temporal Attention for Language Models. In *Findings of the Association for Computational Linguistics (NAACL 2022)*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics (ACL).

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

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. 2022. Improving Temporal Generalization of Pre-trained Language Models with Lexical Semantic Change. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection.

Nina Tahmasebi and Haim Dubossarsky. 2023. Computational modeling of semantic change.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei,

Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proc. of NeurIPS*, pages 5998–6008.

Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. 2021. On Position Embeddings in BERT. In *International Conference on Learning Representations*.

Anna Wegmann, Florian Lemmerich, and Markus Strohmaier. 2020. Detecting Different Forms of Semantic Shift in Word Embeddings via Paradigmatic and Syntagmatic Association Changes. In *The Semantic Web – ISWC 2020*, pages 619–635, Cham. Springer International Publishing.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT Capture Semantics? A Closer Look at Polysemous Words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

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

Appendix

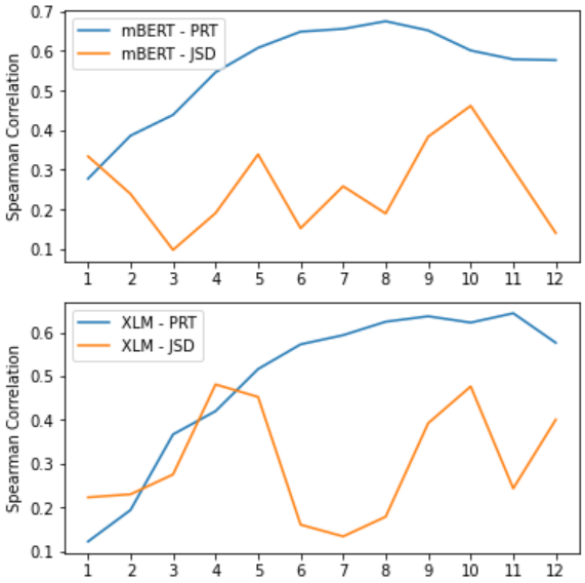

Figure 4: PRT and JSD performance on the artificial LSC dataset

A `Random` Lexical Semantic Change

A.1 Artificial diachronic corpus

We generated an artificial diachronic corpus for LSC by utilising the SemEval and LSCDiscovery benchmakrs for LSC in DWUG format[11] (see Table 4). Instead of incorporating data from both time periods, T_1 and T_2, we discarded information from the first time period as it is more likely to contain word meanings outside the pre-trained knowledge of the models under examination. We created two distinct artificial sub-corpora, C_1 and C_2, by randomly sampling occurrences from the data of the second time period T_2. The DWUG English dataset contains data for 46 target words.

For each target t, we considered all sentences where another target $t1$, with $t1 \neq t$, appeared as potential candidates to emulate instances of semantic change. We simulated a change instance through a *random* replacement, that is by replacing t in the sentence where $t1$ occurred – i.e., $t1 \leftarrow t$. We sample a varying number of sentences and perform replacements for each target, thereby emulating a varying degree of semantic change.

[11] English: https://zenodo.org/records/5796878, German: https://zenodo.org/records/5796871, Swedish: https://zenodo.org/records/5090648, Spanish: https://zenodo.org/records/6433667

References	Benchmark	# targets
Schlechtweg et al., 2020	DWUG-English	46
Schlechtweg et al., 2020	DWUG-German	50
Schlechtweg et al., 2020	DWUG-Swedish	44
Zamora-Reina et al., 2022	DWUG-Spanish	100

Table 4: References and number of targets for each consider artificial corpus

B Lexical Semantic Change

Word	Time span	(Ranked) Farthest replacements	lsc_w (k=1)
attack	T1	**physical**, degeneration, blast, crime, disease, death, condition, plane, affliction, birthday attack	-0.036
	T2	**approach**, force, onslaught, assault, exploit, challenge, commencement, aim, worth, signal	0.059
bit	T1	**nominative case**, accusative case, cryptography, information theory, bdsm, time,point, binary digit, sociologic, sublative	-0.018
	T2	**saddlery**, chard, illative case, iron, bevelled, tack, small, gun, cut, elative case	0.067
circle	T1	**wicca**, circumlocution, encircle, astronomy, tavern, semicircle, around, logic, go,wand	0.002
	T2	**pitch**, place, graduated, figure, disk, territorial, enforce, worship, line, bagginess	0.064
edge	T1	**brink**, cricket, instrument, margin, polytope, side, edge computing, verge, demarcation line, demarcation	-0.015
	T2	**data**, production, climax, division, superiority, organization, sharpness, graph, win, geometry	0.047
graft	T1	**lesion**, bribery, felony, politics, bribe, corruption, autoplasty, surgery, nautical, illicit	-0.047
	T2	**branch**, stock, tree, fruit, shoot, join, cut, graft the forked tree, stem, portion	0.103
head	T1	**headland**, head word, capitulum, syntactic, pedagogue, fluid dynamics, hip hop, headway, pedagog, word	0.004
	T2	**leader**, organs, implement, top, tail, foreland, chief, bolt, axe, forefront	0.084
land	T1	**real estate**, real property, surface, property,build, physical object, Edwin Herbert Land, electronics, landing, first person	-0.032
	T2	**realm**, country, kingdom, province, domain, people, homeland, territory, nation, region	0.076
lass	T1	**sweetheart**, girl, missy, woman, yorkshire, lassem, lasst, lassie, loss, miss	0.014
	T2	**fille**, dative case, jeune fille, loose, lasses, unattached, young lady, young woman, north east england, past participle	0.099
plane	T1	**airplane**, aeroplane, pt boat, heavier-than-air craft, glide , boat, lycaenidae, lift, bow, hand tool	-0.197
	T2	**geometry**, point, shape, surface, flat, degree, form, range, anatomy, smooth	0.205
player	T1	**media player**, idler, soul, thespian, person, individual, trifler, performer, somebody, histrion	-0.065
	T2	**contestant**, performing artist, actor, musician, musical instrument, music, gamer, theater, player piano, play the field	0.042
prop	T1	**props**, airscrew, astronautics, actor, airplane propeller, seashell, stagecraft, stage, property, art	-0.042
	T2	**around**, rugby, imperative mood, about, singular, scrum, ignition, roughly, ballot, manually	0.088
rag	T1	**ragtime**, nominative case, accusative case, rag week, terminative case, inflectional, sublative, piece of material, tag, sanitary napkin	-0.049
	T2	**clothes**, exhaustion, university, society, silk, ragged, journalism, haze, ranking, torment	0.071
record	T1	**attainment**, track record, achievement, accomplishment, struct, number, intransitive, record book, criminal record, disc	-0.036
	T2	**evidence**, document, information, audio, recollection, storage medium, memory, electronic, sound recording, data	0.089
stab	T1	**thread**, staccato, feeling, nominative case, sheet, chord, bacterial, culture, twinge, sensation	-0.046
	T2	**wound**, tool, knife thrust, weapon, plaster, criticism, wire, pierce, thrust, try	0.029
thump	T1	**clunk**, throb, clump, thud, pound, thumping, rhythmic, sound, blow, hit	-0.036
	T2	**muffled**, hit, blow, sound, rhythmic, thumping, pound, thud, clump, throb	0.033
tip	T1	**gratuity**, first person, forty, bloke, singular, overturn, stringed instrument, unbalanced, taxi driver, sated	-0.031
	T2	**brush**, tap, strike, gift, tram, flex, tumble, heap, full, hint	0.070

Table 5: Words annotated as changed in SemEval 2020 Task 1: Binary Subtask and retrieved farthest replacements for each time span.

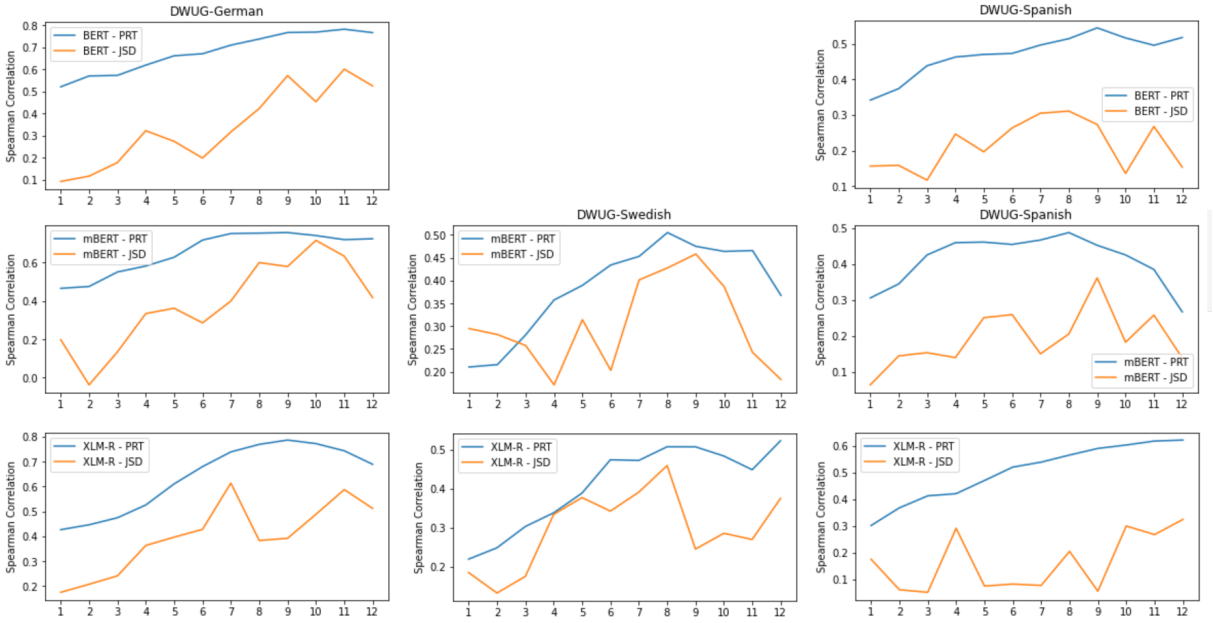

Figure 5: PRT and JSD performance on the artificial LSC dataset

	T1	**T2**
	remember that it be only such line as be nearer the ground **plane** than the eye that be draw under the horizon line	as his **plane** cross north carolina and head south over the atlantic it pick up a small convoy of escort military craft that try to make radio contact but fail
BERT	there, be, where, here, and	planes, over, out, boats, aircraft
LLaMa 2	level,surface,flat plane,horizontal plane	aircraft,airplane,jet,plane model,propeller-driven vehicle

Table 6: Generated substitutions for usages of **plane** extracted by SemEval 2020 Task 1 English.