# Definition generation for lexical semantic change detection

Anonymous ACL submission

## Abstract

We use contextualized word definitions generated by large language model as semantic representations in the task of diachronic lexical semantic change detection (LSCD). In short, generated definitions are used as 'senses', and the change score of a target word is retrieved by comparing their distributions in two time periods under comparison. On the material of five datasets and three languages, we show that generated definitions are indeed specific and general enough to convey a signal sufficient to rank sets of words by the degree of their semantic change over time. Our approach is on par with or outperforms prior non-supervised sense-based LSCD methods. At the same time, it preserves interpretability and allows to inspect the reasons behind a specific shift in terms of discrete definitions-as-senses. This is another step in the direction of explainable semantic change modeling.

## 1 Introduction and related work

Lexical semantic change detection (LSCD) methods up to now have mostly been based on token embeddings produced by large language models. While efficient, when measured on the existing benchmarks like diachronic word usage graphs (Schlechtweg et al., 2021), these methods are largely non-interpretable and produce rather abstract 'change scores'. On the other hand, historical linguistics usually deals with semantic change in terms of discrete and interpretable **senses** being lost or gained (or changing their frequency).

Recently, a number of works were published which made an attempt to bridge this gap. In particular, Tang et al. (2023) proposed a sense distribution based LSCD method. Basically, they perform word sense disambiguation (WSD) on every occurrence of a target word in two diachronic corpora, using pre-trained sense embeddings (based on WordNet and BabelNet). Once all the occurrences are assigned a sense, the sense frequency distributions are compared between two time periods to quantify the semantic change. This approach preserves the possibility to interpret these shifts, e.g., by analyzing which sense is 'responsible' for the shift.

We argue that while such methods constitute a significant advance for LSCD, they are inherently limited by their reliance on a pre-defined sense inventory. Even the best ontologies like BabelNet can miss important senses, especially when dealing with chronologically recent text data. For many languages, good ontologies simply do not exist.

Thus, we propose to replace retrieving a fitting sense for a given target word usage from an external ontology by *generating a dictionary-like contextualized definition for this specific occurrence*, using a large language model (LLM). These definitions serve as semantic representations of target word usages in the LSCD pipeline. The usage of generated definitions as semantic representations in LSCD was first proposed by Giulianelli et al. (2023), but they did not conduct comprehensive empirical evaluations for semantic change detection *per se*. In this paper, we fill in this gap and actually test definitions as representations on the existing diachronic semantic change benchmarks. We show that our method yields competitive results, often outperforming Tang et al. (2023), without relying on any manually created lexical database, but at the same time preserves interpretability via human-readable definitions of senses.

The contributions of this paper are as follows:

1. Contextualized definitions generated by LLMs can be used to rank words by the degree of their diachronic semantic change, with competitive performance.

2. Using *definition* embeddings with classical LSCD methods (APD and PRT) gives better

results than using contextualized *token* embeddings as in prior work. However, this approach makes it less convenient to interpret and analyze semantic shifts.

3. Using generated definitions as *text strings* (with some merging based on their form) yields slightly lower results in comparison, but allows inspecting the nature of a semantic shift: e.g., what senses appeared or disappeared, or changed their frequency significantly.

All the software and models will be made available under permissive licenses upon paper acceptance.

## 2 Data

We experiment on English, Norwegian and Russian benchmarks, since for these languages we had easy access to resources for fine-tuning definition generators. However, scaling to other languages is comparatively easy and requires only a small dataset of contextualized definitions (see §3).

To evaluate the performance of a semantic change detection system, we used existing LSCD datasets (diachronic corpora and gold scores for the target words): the English part of the SemEval'20 Task 1[1] (Schlechtweg et al., 2020), NorDiaChange (Kutuzov et al., 2022a) for Norwegian, and RuShiftEval (Kutuzov and Pivovarova, 2021b) for Russian. NorDiaChange actually contains two datasets and RuShiftEval contains three datasets, with different time period pairs under comparison (for Norwegian, the sets of target words are also different). The Russian datasets feature the highest number of target words (99, as compared to 37 in English and Norwegian datasets).

Note that SemEval'20 Task 1 included two subtasks: binary classification of words (changed or not changed) and ranking the words by the degree of their change. In this work, we focus only on the ranking task: 1) because the Russian dataset does not include binary labels, and 2) because even in the English and Norwegian datasets the binary labels are in many ways derivatives of the change scores.[2]

It is also important to note that the RuShiftEval dataset was used in a shared task of the same name (Kutuzov and Pivovarova, 2021a). However, the scores in its leaderboard or in Cassotti et al. (2023) are not directly comparable to the scores in this work, since in the shared task, the dataset was split into the development and test parts, so that the participants were able to tune their systems on the development set. In this paper, we focus on unsupervised approaches, aiming to avoid the necessity of tuning hyperparameters and leaving this for future work.

### 2.1 Preprocessing

We use the lemmatized versions of the SemEval-2020 English corpora when reproducing Tang et al. (2023)'s Lesk baseline. No preprocessing of the Norwegian and Russian corpora has been done, except for lower-casing when running the Lesk baselines (see the details in the section 4) and taking lemmas of the target words into account when sampling usage examples for both Lesk and definition generation methods. Since frequent words may have more than 100 000 usages in the Norwegian and Russian corpora, we sampled randomly no more than 1000 usages for each target word from every diachronic corpus.

This resulted in sub-corpora of total 58 000 usages for English, 47 000 for Norwegian-1, 51 000 for Norwegian-2, and 164 000, 183 000 and 168 000 for Russian-1, Russian-2 and Russian-3 correspondingly.[3]

## 3 Definition generation methods

Our general pipeline of generating definitions from an LLM ('DefGen') is similar to Giulianelli et al. (2023). The definition generation models were fine-tuned on **WordNet** (Ishiwatari et al., 2019), **Oxford** (Gadetsky et al., 2018) and **CoDWoE** (Mickus et al., 2022) for English, **CoDWoE** for Russian and **Bokmålsordboka**[4] for Norwegian. All CoDWoE datasets originally come from Wiktionary so it is straightforward to extend this method to any major language. As a prompt for the LLM, we used the original example usage with the question 'What is the definition of TARGETWORD?' (in English, Norwegian or Russian) added at the end.

The differences in comparison to Giulianelli et al. (2023) are as follows:

[1] https://www.ims.uni-stuttgart.de/en/research/resources/corpora/sem-eval-ulscd/

[2] In contrast to the English and Norwegian datasets which contain *change* scores, the Russian datasets contain *similarity* scores. The obtained correlations are thus negative. We flip the sign when reporting these numbers to improve readability.

[3] We use only examples no longer than 350 subword tokens in all our experiments.

[4] https://ordbokene.no/

| Strategy | BLEU | RougeL | BertScore |
|---|---|---|---|
| Greedy decoding | 7.384 / 6.237 / 5.113 | 0.223 / 0.198 / 0.130 | 0.860 / 0.735 / 0.700 |
| Repetition penalty (1.2) | 6.401 / 6.026 / 5.599 | 0.204 / 0.200 / 0.145 | 0.856 / 0.737 / 0.706 |
| Multinomial sampling | 6.745 / 6.037 / 4.853 | 0.200 / 0.198 / 0.122 | 0.855 / 0.736 / 0.697 |
| Beam search (5 beams) | 7.052 / **7.523** / **5.863** | 0.219 / **0.246** / **0.154** | 0.860 / 0.747 / 0.709 |
| Diverse beam search | **7.651** / 7.356 / 5.713 | **0.225** / 0.243 / 0.150 | **0.862** / **0.750** / **0.710** |

Table 1: Performance of English / Norwegian / Russian definition generation with different generation strategies.

1. As the base model, we used `mT0` (Muennighoff et al., 2023), which is essentially a multilingual version of `Flan-T5`, the model used by (Giulianelli et al., 2023). The fine-tuned models are extensively described and evaluated in another work currently under review, but we provide important details in Appendix A.

2. We additionally conducted a series of experiments with different generation strategies. Giulianelli et al. (2023) used only basic greedy decoding, while we experimented with alternative strategies such as multinomial sampling, beam search, and diverse beam search (Vijayakumar et al., 2018).

Note that these experiments do not deal with LSCD – they only evaluate the capability of the models to generate definitions similar to the gold ones. The results of our experiments for English, Russian and Norwegian, evaluated with BLEU, RougeL and BertScore, are shown in Table 1. We used the default implementations of these metrics from the `Evaluate` library[5] with the only change of using whitespace tokenizer in RougeL for all languages (instead of the default one aimed at English). For English and Russian, we evaluated on the CoDWoE trial sets (about 200 instances each); for Norwegian, we used our own test set of about 7000 instances.[6]

The performance scores are consistent across all three languages: the default mode of greedy decoding turned to be a hard-to-beat baseline. However, using beam search with 5 beams (or its diverse version with diversity penalty of 0.5) does outperform greedy decoding according to all three metrics.

In the experiments below, we use definitions generated with all three approaches: greedy decoding, beam search and diverse beam search, to explore to what extent the definition generation performance translates to LSCD performance.

## 4 Semantic change detection with generated definitions

### 4.1 Baselines

Table 2 shows the results from previous studies that we use as our baselines, as well as the best results of our definition-based systems.

The scores of **XLM-R token embeddings** with APD, PRT or APD/PRT (AM) aggregation methods are taken from Giulianelli et al. (2022). Although this approach is not interpretable, it yields state-of-the-art scores for unsupervised LSCD.

As an interpretable baseline, Tang et al. (2023) used the **Lesk** WSD algorithm (Lesk, 1986) with WordNet definitions (this method is called 'NLTK' in the Table 1 of their paper). Their result for English, as well as our extensions to Norwegian and Russian, are shown in the lower part of Table 2. We were able to reproduce their Lesk results with only small fluctuations[7]. Since no open WordNet-like databases exist for Norwegian or Russian[8], we used the aforementioned Bokmålsordboka and CoDWoE/Wiktionary as sources of Norwegian and Russian sense definitions.

We also experimented with adding part-of-speech information to the Lesk algorithm (that is, restricting Lesk WSD search to only the synsets corresponding to the desired part-of-speech of the target word). The English SemEval'20 dataset explicitly specifies parts-of-speech for the target words, while the Norwegian and Russian datasets contain

[5] https://huggingface.co/docs/evaluate/

[6] Generating definitions with our models for 1000 example usages takes about 2 minutes on an NVIDIA A100 GPU.

[7] Probably due to the fact that we used top 1 sense in all our experiments, while Tang et al. (2023) experimented with top k highest ranked senses on a held-out set and found k = 2 to perform best. However, we focus on unsupervised approaches to the task and leave hyperparameter tuning on development sets for future work.

[8] The Open Multilingual WordNet allows searching for words in other languages than English, but the synset definitions remain in English.

| Method | English | Norwegian-1 | Norwegian-2 | Russian-1 | Russian-2 | Russian-3 |
|---|---|---|---|---|---|---|
| **Non-interpretable methods:** | | | | | | |
| XLM-R token embeddings | 0.514♢ | 0.394♢ | 0.387♢ | 0.376♢ | **0.480**♢ | 0.457♢ |
| Definition embeddings (ours) (See Table 3 for details) | **0.637** | **0.496** | **0.565** | **0.488** | 0.462 | **0.504** |
| **Interpretable methods:** | | | | | | |
| Lesk without PoS | 0.423♣ | 0.178 | 0.500 | 0.294 | 0.279 | 0.286 |
| Lesk with PoS | 0.587 | 0.150 | 0.474 | | | |
| ARES sense embeddings | 0.529♣ | — | — | — | — | — |
| LMMS sense embeddings | 0.589♣ | — | — | — | — | — |
| Definitions as senses (ours) (See Table 4 for details) | 0.605 | 0.362 | 0.260 | 0.391 | 0.431 | 0.491 |

Table 2: Summary of our results and baselines (Spearman's $\rho$ for graded LSCD). Figures marked with ♢ are taken from Giulianelli et al. (2022); AM (arithmetic mean of APD and PRT) is called APD-PRT in their paper. Figures marked with ♣ are taken from Tang et al. (2023); Lesk is called NLTK in their paper. Numbers without a symbol are our own results.

nouns only. The two variants of Lesk yield identical results on the Russian datasets since the target words are not PoS-ambiguous. For English and Norwegian-2, Lesk even outperforms the XLM-R token embeddings and comes close to our approach based on definition embeddings.

However, Tang et al. (2023)'s main results are based on **ARES and LMMS sense embeddings**. Unfortunately, these embeddings are not publicly available anymore due to link rot, and thus we can only quote the performance scores from Tang et al. (2023). The LMMS download link[9] leads to a private file storage, and the ARES embeddings are also not available anymore at the provided URL. We contacted the authors but got no answer by the time of writing.

### 4.2 Using definition embeddings

Generated definitions can be easily vectorized by using any sentence embedding model. We embedded the generated definitions for every target word usage with DistilRoBERTa[10]. After that, it becomes possible to use the standard LSCD methods like PRT (prototype embeddings), APD (average pairwise distance), and their arithmetic mean (AM) (Kutuzov et al., 2022b). The only difference to the standard setup is that instead of *token* embeddings, we feed contextualized *definition* embeddings into the algorithm.

The intuition here is that by measuring the average or pairwise distances between definitions of one and the same target word in two historical corpora, one can quantify the degree of semantic change for this word between two time periods. As can be seen in in Table 3, this is indeed the case. Our definition embeddings outperform the contextualized XLM-R token embeddings from Giulianelli et al. (2022) on five of the six evaluated datasets.

Note in this context that using token embeddings from a masked LM requires the knowledge of the exact position of the target token in the input sentence (with additional issues in case of the target word being split into multiple sub-words). In our approach, adding the 'What is the definition of X?' prompt to the input sentence is completely decoupled from the location of X within the sentence.

The decoding strategy does not seem to make a significant difference in terms of LSCD performance. Greedy decoding is a reasonable default choice despite its slightly lower scores in Table 1.

On English, the APD method on definition embeddings also outperforms the best sense-embedding-based approaches from Tang et al. (2023) by a large margin (see Table 2). Note, however, that using definition embeddings in this case still yields a non-interpretable result: we do not know what exact senses are responsible for a high degree of semantic change. For this reason, we propose to use the generated definitions directly in the next section.

[9] https://github.com/danlou/LMMS

[10] https://huggingface.co/sentence-transformers/all-distilroberta-v1

| | English | | | Norwegian-1 | | | Norwegian-2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | APD | PRT | AM | APD | PRT | AM | APD | PRT | AM |
| **Token emb.** | 0.514 | 0.320 | 0.457 | 0.389 | 0.378 | 0.394 | 0.387 | 0.270 | 0.325 |
| **Greedy** (ours) | 0.633 | 0.331 | 0.580 | 0.416 | 0.368 | **0.496** | **0.565** | 0.413 | 0.558 |
| **Beam** (ours) | **0.637** | 0.355 | 0.601 | 0.317 | 0.411 | 0.434 | 0.478 | 0.452 | 0.479 |
| **Diverse** (ours) | 0.613 | 0.359 | 0.591 | 0.335 | 0.364 | 0.444 | 0.508 | 0.470 | 0.523 |
| | **Russian-1** | | | **Russian-2** | | | **Russian-3** | | |
| | APD | PRT | AM | APD | PRT | AM | APD | PRT | AM |
| **Token emb.** | 0.372 | 0.294 | 0.376 | **0.480** | 0.313 | 0.374 | 0.457 | 0.313 | 0.384 |
| **Greedy** (ours) | 0.464 | 0.406 | **0.488** | 0.453 | 0.430 | 0.462 | 0.489 | **0.504** | 0.494 |
| **Beam** (ours) | 0.381 | 0.387 | 0.401 | 0.400 | 0.451 | 0.411 | 0.386 | 0.439 | 0.413 |
| **Diverse** (ours) | 0.396 | 0.457 | 0.433 | 0.405 | 0.449 | 0.417 | 0.414 | 0.476 | 0.436 |

Table 3: LSCD performance (Spearman's $\rho$) with definition embeddings obtained with different decoding strategies (greedy decoding, beam search and diverse beam search). For comparison, *Token emb.* presents the results by Giulianelli et al. (2022) with contextualized XLM-R token embeddings. AM is the arithmetic mean of APD and PRT.

### 4.3 Merging definitions together

The definitions generated by a DefGen system can be used directly for LSCD. In this case, each unique definition is considered a separate word sense, and the sense distributions of the two time periods can be compared in the same way as in Tang et al. (2023). This approach is straightforward and already results in competitive performance (see the "No merging" section in Table 4[11]).

However, it obviously suffers from too granular senses. As an example, for almost 1000 occurrences of the word '*plane*' in the SemEval'20 English dataset, more than 200 unique definitions were generated, most only with one occurrence. This list includes definitions obviously belonging to one and the same sense: for example, '*An aircraft, especially one designed for military use*' and '*An aircraft, especially a military aircraft*'. This leads to noise and – even worse – to reduced interpretability. It is easy to observe that definitions belonging to the same sense are often similar in their surface form. Thus, in this subsection, we describe our experiments with merging similar definitions together.

Any decision about what word usages belong to one sense is inherently arbitrary (Kilgarriff, 1997). The same applies to definitions: in order to decide whether two definitions represent one and the same sense, one has to find a way to quantify their similarity. In order to preserve interpretability, we decided to use surface string similarity metrics (as opposed to, e.g., cosine similarity between definition embeddings).

We remind again that the top part of Table 4 shows the performance scores on our datasets with no merging involved: every unique definition is considered to be a separate sense on its own and we simply compare the distribution of these 'senses' across two time periods. In addition to that, we introduce two merging strategies which we dub 'minimalist' and 'full-fledged' merging. The intuition behind them is that one replaces some of the generated definitions for a target word with another *similar* definition generated for the same target word, thus reducing the total number of unique definitions per word and making it closer to a realistic number of senses.

First, we filter out punctuation marks from all definitions. Second, every time period (out of two) is processed *separately*[12] in the following way. For every target word, we sort the generated definitions by their frequency and loop over them, starting from the top (most frequent) ones, representing the

[11]Table 4 reports results after using two different distance metrics: cosine and Jensen-Shannon divergence (JSD). JSD is superior in most cases, but not always.

[12]We also tried *joint* processing of both time periods to make the resulting definitions-as-senses more comparable. However, it consistently resulted in worse performance: the most probable reason being that it makes the sense distributions too close to each other, eliminating meaningful differences. It also can bias the predictions if one time period is represented by a larger corpus than another.

| | English | | Norwegian-1 | | Norwegian-2 | | Russian-1 | | Russian-2 | | Russian-3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cosine | JS | Cosine | JS | Cosine | JS | Cosine | JS | Cosine | JS | Cosine | JS |
| **No merging:** | | | | | | | | | | | | |
| **Greedy** | 0.461 | 0.405 | 0.303 | 0.332 | 0.211 | 0.232 | 0.299 | 0.390 | 0.337 | 0.427 | 0.383 | 0.469 |
| **Beam** | 0.457 | 0.476 | 0.268 | 0.238 | 0.216 | 0.201 | 0.304 | 0.368 | 0.297 | 0.403 | 0.317 | 0.417 |
| **Diverse** | 0.449 | 0.382 | 0.241 | 0.280 | 0.069 | 0.164 | 0.301 | 0.345 | 0.310 | 0.389 | 0.348 | 0.421 |
| **Minimalist merging:** | | | | | | | | | | | | |
| **Greedy** | 0.564 | **0.605** | 0.251 | 0.280 | 0.192 | 0.197 | 0.271 | **0.391** | 0.233 | **0.431** | 0.325 | **0.491** |
| **Beam** | 0.510 | 0.463 | 0.297 | 0.240 | 0.112 | 0.189 | 0.298 | 0.366 | 0.252 | 0.383 | 0.301 | 0.409 |
| **Diverse** | 0.478 | 0.430 | 0.325 | 0.296 | 0.162 | 0.215 | 0.265 | 0.354 | 0.268 | 0.406 | 0.287 | 0.443 |
| **Full-fledged merging:** | | | | | | | | | | | | |
| **Greedy** | 0.439 | 0.409 | 0.261 | **0.362** | 0.193 | **0.260** | 0.286 | **0.391** | 0.250 | 0.416 | 0.360 | 0.476 |
| **Beam** | 0.492 | 0.489 | 0.265 | 0.215 | 0.186 | 0.226 | 0.304 | 0.360 | 0.250 | 0.347 | 0.327 | 0.420 |
| **Diverse** | 0.312 | 0.375 | 0.209 | 0.315 | 0.202 | 0.221 | 0.236 | 0.301 | 0.217 | 0.379 | 0.262 | 0.411 |
| **Threshold** | 50 | | 10 | | 10 | | 10 | | 10 | | 10 | |

Table 4: LSCD performance (Spearman's $\rho$) with generated definitions and different generation and merging strategies. Results are reported with two distance metrics: cosine similarity and Jensen-Shannon divergence. *Threshold* refers to the Levenshtein edit distance threshold used for merging definitions.

dominant senses of the word. For every step in this loop (let's designate it as 'hub definition'), we loop again over the remaining definitions, calculating the edit distance[13] between them and the hub definition. If the edit distance is lower than the predefined threshold, the current definition is replaced with the hub definition (we assume they belong to one sense). With the 'minimal strategy', only the first (most frequent) definition can be the hub (and have other definitions replaced by it), the loop is stopped after it is compared to all other definitions. With the 'full-fledged' strategy, the loop continues, and other (less frequent) definitions also get a chance to become hubs, if they were not subsumed by another definition before. The 'full-fledged' strategy naturally results in even stronger reduction on the number of unique senses (see Figure 1). The datasets with replaced definitions are used for LSCD in the usual way.

The value of the edit distance threshold is a hyperparameter. In theory, one can tune it on a designated development set, but in this study, we tried to avoid the supervised setup. Thus, after studying the data, we only tested two intuitively sensible threshold values of 10 and 50. It turned out that the value of 10 is optimal for Norwegian and Russian, while the value of 50 (more merging) is optimal for English (changes in the number of senses after merging are more obvious for English on Figure 1 exactly because of the more aggressive Levenshtein threshold).

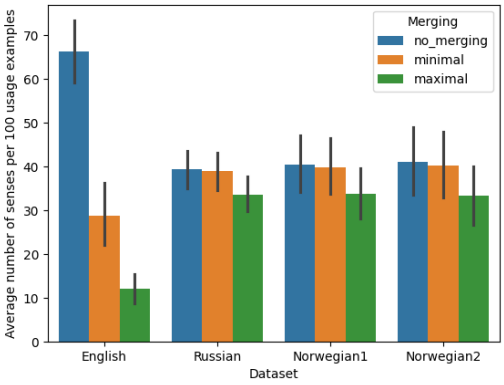


Figure 1: Average number of senses per 100 usages before and after merging, calculated across all datasets for each language.

As shown in Table 4, merging definitions indeed improves the performance in the LSCD ranking task for three languages and five benchmarking datasets, in comparison to no merging at all. It still does not reach the level of APD on definition embeddings (Table 3), but outperforms the best sense embedding approach from Tang et al. (2023). For two Russian datasets out of three, our merged definitions-as-senses outperform previous best unsupervised results (Giulianelli et al., 2022).

Curiously, only for Norwegian the best perfor-

[13]We use Levenshtein distance (Levenshtein, 1966); other edit distance formulations are possible. In addition, only definitions containing 4 words and more could serve as hubs, to prevent merging together very short definitions with low edit distances.

mance is achieved using the full-fledged merging strategy (and is still comparatively low); for English and Russian, the minimalist strategy (only merging with the dominant sense) gives the best scores. Thus, merging should be done cautiously: merging too much can degrade the performance. Another interesting finding is that greedy decoding again turned out to be the best generation strategy for LSCD. We hypothesize that using beam search often results in too diverse a set of definitions, which prevents efficient comparisons between their diachronic distributions.

Overall, using surface forms of generated definitions directly is outperformed by using vectorized definitions and APD or PRT methods for semantic change detection (see 4.2), which is consistent with findings in Giulianelli et al. (2023). They found that cosine similarity between vectorized definitions better approximates human similarity judgments than surface form similarity metrics like edit distance or BLEU. However, using definitions in their textual form has a clear advantage of being interpretable. In the next section, we illustrate how semantic change can be explained and analyzed by comparing the generated (and merged) definitions.

## 5 Qualitative analysis

When predicting semantic change on the basis of the distribution of senses (or definitions-as-senses), it becomes possible to analyze and interpret this change, by simply looking at the distribution of entries (senses) which contribute most to the difference.

Let's consider the top performing set of English definitions (generated with greedy decoding and merged in the minimalist approach with the edit distance threshold of 50). For the word '**ball**' in the SemEval'20 time periods, the JSD metric yields a high change score of 0.83. After looking at the list of top frequent definitions-as-senses for this word, it becomes clear that its dominant sense has changed: while in time period 1 (19th century), more than 82% of all usage examples were given the definition *'A spherical object especially one that is round in shape'* with *'A party'* being the next most frequent sense, in time period 2 (20th century), 80% of '**ball**' usages were defined as *'The object hit in a game'*, with similar definitions following this one in the top-frequent list. This is a clear evidence of the 'dancing party' sense for the word 'ball' becoming obsolete in the 20th century, with the sports-related sense taking the dominant position.

For the noun '**attack**', the system predicts a medium change degree of 0.34. Again, it is straightforward to find the reasons. While in both periods the dominant sense is the same (*'An instance of military action against an enemy'*), in the time period 2 its ratio drops down from 87% to 80%, and we observe the appearance of a new rare but not unique sense of *'An instance of sudden violent activity of a bodily organ or system especially the heart'*. This is a linguistically plausible explanation of a semantic shift, much more useful to a lexicographer than a raw change score. See Appendix B for more examples in English, Norwegian and Russian.

## 6 Conclusion

In this paper, we showed how contextualized dictionary-like definitions generated by a fine-tuned large language model can be used for the practical downstream task of semantic change detection (in particular, ranking words by the degree of their diachronic semantic change).

Following Giulianelli et al. (2023), we treat generated definitions as semantic representations of the target words. These definitions (and their frequency distributions) can be used 'as is', using Tang et al. (2023)'s method, or after embedding them in a dense vector space using any available sentence embedding model. The second method yields results which are empirically better (considering existing benchmarks for three different languages), but the first method makes it much easier to interpret and explain semantic change, since it operates directly on generated definitions in their textual forms.

We consider this study a small step towards more explainable semantic change modeling, which can be closer to linguistically plausible discrete 'senses', while still retaining empirical performance. In the future, we plan to explore to what extent it is possible to improve our results by tuning hyperparameters on development sets (where available). Another direction for future research is using more advanced string distance metrics like weighted Levenshtein distance, Longest Common Subsequence Ratio, or Word Mover's Distance (Kusner et al., 2015), in the hope that it will allow to handle more nuanced similarities and dissimilarities between generated definitions.

## Ethical impact

For fine-tuning our definition generators, we used only open and publicly available datasets, mostly dictionaries. However, some of them (especially Wiktionary) are crowd-sourced, and thus can (and do) contain inappropriate phrases. In addition, the foundational `mT0` language model on which we base our pipeline, was trained among other data on web-crawled texts, also far from being clean. Thus, generated definitions are not guaranteed to be free from swearing, discriminative passages and other inappropriate content.

## Limitations

This work is limited to only three languages (English, Norwegian and Russian), while the standard SemEval'20 'LSCD suite' contains four languages (English, German, Latin, Swedish). Also, we did not experiment with hyperparameter tuning or different ways of training definition generators. It should also be noted that Spearman rank correlation can be non-accurate for samples the size of LSCD benchmarks: we use it to preserve compatibility and comparability with prior work.

Finally, we have not yet empirically evaluated how useful in practice the definition-based explanations of semantic change will be for historical linguists and lexicographers (although what we see after manual inspection of the system predictions is promising).

## References


Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. 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.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. 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 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Andrey Kutuzov and Lidia Pivovarova. 2021a. RuShiftEval: a shared task on semantic shift detection for Russian. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Andrey Kutuzov and Lidia Pivovarova. 2021b. 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, 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.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a

pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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, 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.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. *AAAI*.

## A Definition generator models

Contextual definitions in this work are created with fine-tuned large language models, using the method proposed by Giulianelli et al. (2023): an encoder-decoder language model is fine-tuned on a dataset of target word usages and the corresponding definitions. Then, definitions are conditionally generated for every example in the test set. Giulianelli et al. (2023) used `Flan-T5` (Chung et al., 2022) as the underlying language model. However, it was trained predominantly on English and lacks the capability to properly encode or generate texts in languages with significantly different writing systems (especially true for Russian, but Norwegian characters 'å', 'ø' and 'æ' are also not processed by the `Flan-T5` tokenizer). Because of that, in this work we are using `mT0` (Muennighoff et al., 2023) which is essentially a multilingual version of `Flan-T5`. For all the experiments, we employ the `mT0-xl` version of the model[14], 3.7B parameters in size. Fine-tuning was done in a standard text-to-text setup, for every language (English, Norwegian, Russian) separately, so that in the end we had three language-specific models.

## B Examples of merged definitions

Tables 5, 6 and 7 show examples of most frequent definitions-as-senses for some of the target words in our English, Norwegian and Russian benchmarks. All the definitions are generated with the best strategy for the specific dataset (see Table 4).

[14] https://huggingface.co/bigscience/mT0-xl

| | **Period 1** (1810-1860) | **Period 2** (1960-2010) |
|---|---|---|
| *circle* JS=0.07 | *To move in a circular course* (99%) | *To move in a circular course* (98%) |
| | *To move in a circular course.* (1%) | *To move in a circular course.* (<1%) |
| | | *To move around something especially so as to make it appear to move around* (<1%) |
| *risk* JS=0.44 | *The probability of a negative outcome to a decision or event* (59%) | *The probability of a negative outcome to a decision or event* (63%) |
| | *The probability of a negative outcome to a decision or event the chance of a negative outcome to a decision or event* (8%) | *The probability of a negative outcome to a decision or event the chance of a negative outcome to a decision or event* (3%) |
| | *A venture undertaken without regard to possible loss or injury especially if significant* (3%) | *A venture undertaken without regard to possible loss or injury especially if significant* (3%) |
| *ball* JS=0.83 | *A spherical object especially one that is round in shape* (82%) | *The object hit in a game* (80%) |
| | *A party* (6%) | *The object used in various sports especially in soccer tennis basketball etc* (<1%) |
| | *A wedding* (<1%) | *The object used in various sports especially in soccer basketball and other games which is thrown or kicked* (<1%) |

Table 5: The three most frequent definitions per period for three English words: *circle* (low predicted change rate), *risk* (medium predicted change rate), and *ball* (high predicted change rate). Parentheses indicate the relative frequency of each definition among all samples of the period.

| | **Period 1** (1980-1990) | **Period 2** (2012-2019) |
|---|---|---|
| *oppvarming* 'heating, warm-up' JS=0.19 | *det å varme opp* 'the action of heating' (98%) | *det å varme opp* 'the action of heating'(91%) |
| | *i regnskap* 'in accounting' (<1%) | *det å varme opp jordoverflaten* 'the action of warming the Earth surface' (1%) |
| | *i statistikk* 'in statistics' (<1%) | *i fotball* 'in football' (<1%) |
| *bank* 'bank' JS=0.64 | *institusjon som tar imot innskudd av penger og gir lån* (13%) | *institusjon som tar imot innskudd av penger og gir lån* (14%) |
| | 'institution that accepts money deposits and gives loans' | |
| | *institusjon som tar imot innskudd og utfører pengetransaksjonstjenester* (6%) | *institusjon som tar imot innskudd og utfører pengetransaksjonstjenester* (6%) |
| | 'institution that accepts deposits and provides financial transaction services' | |
| | *institusjon som tar imot innskudd av penger og driver pengetransaksjonsvirksomhet* (5%) | *institusjon som tar imot innskudd av penger og driver pengetransaksjonsvirksomhet* (4%) |
| | 'institution that accepts deposits of money and conducts financial transaction business' | |
| *kode* 'code' JS=0.81 | *i i sammensetninger* 'i in compounds' (4%) | *i i bestemt form* 'i in the definite form' (4%) |
| | *i i bestemt form* 'i in the definite form' (3%) | *mønster oppskrift på hvordan noe skal lykkes* 'pattern, recipe for how something succeeds' (1%) |
| | *i statistikk* 'in statistics' (3%) | *i overført betydning mønster mønstergyldighet* 'in a figurative sense pattern, pattern validity' (1%) |

Table 6: The three most frequent definitions per period for three words of the Norwegian-2 dataset: *oppvarming* (low predicted change rate), *bank* (medium predicted change rate), and *kode* (high predicted change rate). Parentheses indicate the relative frequency of each definition among all samples of the period.

| | **Period 1** (before 1917) | **Period 2** (after 1991) |
|---|---|---|
| цензура *'censorship'* JS=0.09 | система государственного надзоря за печатью и средствами массовой информации (99%) | система государственного надзоря за печатью и средствами массовой информации (99%) |
| | *system of state control over printing and mass media* | |
| | истор. государственный орган, осуществляющий цензуру *historically, a state body conducting censorship* (<1%) | государственная система государственного надзоря за печатью и средствами массовой информации *a state system of controlling printing and mass media* (<1%) |
| | контроль, надзор за печатью и средствами массовой информации *control and monitoring of printing and mass media* (<1%) | сокр. от цензурная служба, государственная организация, осуществляющая цензуру *abbrev. censoring body, state organ conducting censorship* (<1%) |
| огонь *'fire'* JS=0.66 | источник огня, источник света *source of fire or light* (7%) | действие по значению глагола стрелять *nominal form of the verb 'to fire'* (3%) |
| | источник света, источник тепла, дыма и т.п. *source of light, warmth, smoke etc* (2%) | источник света, источник освещения *source of light, of illumination* (2.5%) |
| | перен. страсть, пыл *metaphoric. passion or rage* (2%) | воен. стрельба из огнестрельного оружия *metaphoric. gunfire* (1%) |
| линейка *'line, ruler'* JS=0.80 | измерительный инструмент в виде прямой пластинки с нанесёнными на неё делениями для измерения длины и расстояния *a measuring tool looking like a straight plane with marks to measure length and distance* (2.6%) | перен. совокупность однородных предметов, изделий, продуктов и т. п. *metaphoric. a batch of similar items, goods, products* (3.5%) |
| | устар. длинные узкие сани *archaic. long narrow sledges* (1%) | измерительный инструмент в виде прямой пластинки с нанесёнными на неё делениями для определения длины линии *a measuring tool looking like a straight plane with marks to measure the length of a line* (1.4%) |
| | измерительный инструмент в виде прямой линии с нанесёнными на неё делениями *a measuring tool looking like a straight plane with marks* (1%) | измерительный инструмент в виде прямой пластинки с нанесёнными на неё делениями для измерения длины и ширины *a measuring tool looking like a straight plane with marks to measure length and width* (1%) |

Table 7: The three most frequent definitions per period for three words from the Russian-1 dataset: 'цензура' (low predicted change rate), 'огонь' (medium predicted change rate), and 'линейка' (high predicted change rate). Parentheses indicate the relative frequency of each definition among all samples of the period.