Lexical Semantic Change Discovery

Sinan Kurtyigit[♠] **Maike Park**[♡] **Dominik Schlechtweg**[♠]
Jonas Kuhn[♠] **Sabine Schulte im Walde**[♠]

[♠]Institute for Natural Language Processing, University of Stuttgart
[♡]Leibniz Institute for the German Language, Mannheim
sinan.kurtyigit@gmail.com, park@ids-mannheim.de
{schlecdk,jonas.kuhn,schulte}@ims.uni-stuttgart.de

Abstract

While there is a large amount of research in the field of Lexical Semantic Change Detection, only few approaches go beyond a standard benchmark evaluation of existing models. In this paper, we propose a shift of focus from change detection to change discovery, i.e., discovering novel word senses over time from the full corpus vocabulary. By heavily fine-tuning a type-based and a token-based approach on recently published German data, we demonstrate that both models can successfully be applied to discover new words undergoing meaning change. Furthermore, we provide an almost fully automated framework for both evaluation and discovery.

1 Introduction

There has been considerable progress in Lexical Semantic Change Detection (LSCD) in recent years (Kutuzov et al., 2018; Tahmasebi et al., 2018; Hengchen et al., 2021), with milestones such as the first approaches using neural language models (Kim et al., 2014; Kulkarni et al., 2015), the introduction of Orthogonal Procrustes alignment (Kulkarni et al., 2015; Hamilton et al., 2016), detecting sources of noise (Dubossarsky et al., 2017, 2019), the formulation of continuous models (Frermann and Lapata, 2016; Rosenfeld and Erk, 2018; Tsakalidis and Liakata, 2020), the first uses of contextualized embeddings (Hu et al., 2019; Giulianelli et al., 2020), the development of solid annotation and evaluation frameworks (Schlechtweg et al., 2018, 2019; Shoemark et al., 2019) and shared tasks (Basile et al., 2020; Schlechtweg et al., 2020).

However, only a very limited amount of work applies the methods to **discover novel instances of semantic change** and to evaluate the usefulness of such discovered senses for external fields. That is, the majority of research focuses on the introduction of novel LSCD models, and on analyzing and evaluating existing models. Up to now, these preferences for development and analysis vs. application represented a well-motivated choice, because the quality of state-of-the-art models had not been established yet, and because no tuning and testing data were available. But with recent advances in evaluation (Basile et al., 2020; Schlechtweg et al., 2020; Kutuzov and Pivovarova, 2021), the field now owns standard corpora and tuning data for different languages. Furthermore, we have gained experience regarding the interaction of model parameters and modelling task (such as binary vs. graded semantic change). This enables the field to more confidently apply models to discover previously unknown semantic changes. Such discoveries may be useful in a range of fields (Hengchen et al., 2019; Jatowt et al., 2021), among which historical semantics and lexicography represent obvious choices (Ljubešić, 2020).

In this paper, we tune the most successful models from SemEval-2020 Task 1 (Schlechtweg et al., 2020) on the German task data set in order to obtain high-quality discovery predictions for novel semantic changes. We validate the model predictions in a standardized human annotation procedure and visualize the annotations in an intuitive way supporting further analysis of the semantic structure relating word usages. In this way, we automatically detect previously described semantic changes and at the same time discover novel instances of semantic change which had not been indexed in standard historical dictionaries before. Our approach is largely automated, by relying on unsupervized language models and a publicly available annotation system requiring only a small set of judgments from annotators. We further evaluate the usability of the approach from a lexicographer's viewpoint and show how intuitive visualizations of human-annotated data can benefit dictionary makers.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6985–6998
August 1–6, 2021.

2 Related Work

State-of-the-art semantic change detection models are Vector Space Models (VSMs) (Schlechtweg et al., 2020). These can be divided into type-based (static) (Turney and Pantel, 2010) and token-based (contextualized) (Schütze, 1998) approaches. For our study, we use both a static and a contextualized model. As mentioned above, previous work mostly focuses on creating data sets or developing, evaluating and analyzing models. A common approach for evaluation is to annotate target words selected from dictionaries in specific corpora (Tahmasebi and Risse, 2017; Schlechtweg et al., 2018; Perrone et al., 2019; Basile et al., 2020; Rodina and Kutuzov, 2020; Schlechtweg et al., 2020). Contrary to this, our goal is to find 'undiscovered' changing words and validate the predictions of our models by human annotators. Few studies focus on this task. Kim et al. (2014), Hamilton et al. (2016), Basile et al. (2016), Basile and Mcgillivray (2018), Takamura et al. (2017) and Tsakalidis et al. (2019) evaluate their approaches by validating the top ranked words through author intuitions or known historical data. The only approaches applying a systematic annotation process are Gulordava and Baroni (2011) and Cook et al. (2013). Gulordava and Baroni ask human annotators to rate 100 randomly sampled words on a 4-point scale from 0 (no change) to 3 (changed significantly), however without relating this to a data set. Cook et al. work closely with a professional lexicographer to inspect 20 lemmas predicted by their models plus 10 randomly selected ones. Gulordava and Baroni and Cook et al. evaluate their predictions on the (macro) lemma level. We, however, annotate our predictions on the (micro) usage level, enabling us to better control the criteria for annotation and their inter-subjectivity. In this way, we are also able to build clusters of usages with the same sense and to visualise the annotated data in an intuitive way. The annotation process is designed to not only improve the quality of the annotations, but also lessen the burden on the annotators. We additionally seek the opinion of a professional lexicographer to assess the usefulness of the predictions outside the field of LSCD.

In contrast to previous work, we obtain model predictions by fine-tuning static and contextualized embeddings on high-quality data sets (Schlechtweg et al., 2020) that were not available before. We provide a highly automated general framework for evaluating models and predicting changing words on all kinds of corpora.

3 Data

We use the German data set provided by the SemEval-2020 shared task (Schlechtweg et al., 2020, 2021). The data set contains a diachronic corpus pair for two time periods to be compared, a set of carefully selected target words as well as binary and graded gold data for semantic change evaluation and fine-tuning purposes.

Corpora The DTA corpus (Deutsches Textarchiv, 2017) and a combination of the BZ (Berliner Zeitung, 2018) and ND (Neues Deutschland, 2018) corpora are used. DTA contains texts from different genres spanning the 16th–20th centuries. BZ and ND are newspaper corpora jointly spanning 1945–1993. Schlechtweg et al. (2020) extract two time specific corpora C_1 (DTA, 1800–1899) and C_2 (BZ+ND 1946–1990) and provide raw and lemmatized versions.

Target Words A list of 48 target words, consisting of 32 nouns, 14 verbs and 2 adjectives is provided. These are controlled for word frequency to minimize model biases that may lead to artificially high performance (Dubossarsky et al., 2017; Schlechtweg and Schulte im Walde, 2020).

4 Models

Type-based models generate a single vector for each word from a pre-defined vocabulary. In contrast, token-based models generate one vector for each usage of a word. While the former do not take into account that most words have multiple senses, the latter are able to capture this particular aspect and are thus presumably more suited for the task of LSCD (Martinc et al., 2020). Even though contextualized approaches have indeed significantly outperformed static approaches in several NLP tasks over the past years (Ethayarajh, 2019), the field of LSCD is still dominated by type-based models (Schlechtweg et al., 2020). Kutuzov and Giulianelli (2020) yet show that the performance of token-based models (especially ELMo) can be increased by fine-tuning on the target corpora. Laicher et al. (2020, 2021) drastically improve the performance of BERT by reducing the influence of target word morphology. In this paper, we compare both families of approaches for change discovery.

4.1 Type-based approach

Most type-based approaches in LSCD combine three sub-systems: (i) creating semantic word representations, (ii) aligning them across corpora, and (iii) measuring differences between the aligned representations (Schlechtweg et al., 2019). Motivated by its wide usage and high performance among participants in SemEval-2020 (Schlechtweg et al., 2020) and DIACR-Ita (Basile et al., 2020), we use the Skip-gram with Negative Sampling model (SGNS, Mikolov et al., 2013a,b) to create static word embeddings. SGNS is a shallow neural language model trained on pairs of word co-occurrences extracted from a corpus with a symmetric window. The optimized parameters can be interpreted as a semantic vector space that contains the word vectors for all words in the vocabulary. In our case, we obtain two separately trained vector spaces, one for each subcorpus (C_1 and C_2). Following standard practice, both spaces are length-normalized, mean-centered (Artetxe et al., 2016; Schlechtweg et al., 2019) and then aligned by applying Orthogonal Procrustes (OP), because columns from different vector spaces may not correspond to the same coordinate axes (Hamilton et al., 2016). The change between two time-specific embeddings is measured by calculating their Cosine Distance (CD) (Salton and McGill, 1983). The strength of SGNS+OP+CD has been shown in two recent shared tasks with this sub-system combination ranking among the best submissions (Arefyev and Zhikov, 2020; Kaiser et al., 2020b; Pömsl and Lyapin, 2020; Pražák et al., 2020).

4.2 Token-based approach

Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019) is a transformer-based neural language model designed to find contextualized representations for text by analyzing left and right contexts. The base version processes text in 12 different layers. In each layer, a contextualized token vector representation is created for every word. A layer, or a combination of multiple layers (we use the average), then serves as a representation for a token. For every target word we extract usages (i.e., sentences in which the word appears) by randomly sub-sampling up to 100 sentences from both subcorpora C_1 and C_2.[1] These are then fed into BERT to create contextualized embeddings, resulting in two sets of up to 100 contextualized vectors for both time periods. To measure the change between these sets we use two different approaches: (i) We calculate the Average Pairwise Distance (APD). The idea is to randomly pick a number of vectors from both sets and measure their mutual distances (Schlechtweg et al., 2018; Kutuzov and Giulianelli, 2020). The change score corresponds to the mean average distance of all comparisons. (ii) We average both vector sets and measure the Cosine Distance (COS) between the two resulting mean vectors (Kutuzov and Giulianelli, 2020).

[1] We sub-sample as some words appear in 10,000 or more sentences.

5 Discovery

SemEval-2020 Task 1 consists of two subtasks: (i) binary classification: for a set of target words, decide whether (or not) the words lost or gained sense(s) between C_1 and C_2, and (ii) graded ranking: rank a set of target words according to their degree of LSC between C_1 and C_2. These require to detect semantic change in a small pre-selected set of target words. Instead, we are interested in the discovery of changing words from the full vocabulary of the corpus. We define the task of **lexical semantic change discovery** as follows.

> Given a diachronic corpus pair C_1 and C_2, decide for the intersection of their vocabularies which words lost or gained sense(s) between C_1 and C_2.

This task can also be seen as a special case of SemEval's Subtask 1 where the target words equal the intersection of the corpus vocabularies. Note, however, that discovery introduces additional difficulties for models, e.g. because a large number of predictions is required and the target words are not preselected, balanced or cleaned. Yet, discovery is an important task, with applications such as lexicography where dictionary makers aim to cover the full vocabulary of a language.

5.1 Approach

We start the discovery process by generating optimized graded value predictions using high-performing parameter configurations following previous work and fine-tuning. Afterwards, we infer binary scores with a thresholding technique (see below). We then tune the threshold to find the best-performing type- and token-based approach

4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated

Table 1: DURel relatedness scale (Schlechtweg et al., 2018).

for binary classification. These are used to generate two sets of predictions.[2]

Evaluation metrics We evaluate the graded rankings in Subtask 2 by computing Spearman's rank-order correlation coefficient ρ. For the binary classification subtask we compute precision, recall and $F_{0.5}$. The latter puts a stronger focus on precision than recall because our human evaluation cannot be automated, so we decided to weigh quality (precision) higher than quantity (recall).

Parameter tuning Solving Subtask 2 is straightforward, since both the type-based and token-based approaches output distances between representations for C_1 and C_2 for every target word. Like many approaches in SemEval-2020 Task 1 and DIACR-Ita we use thresholding to binarise these values. The idea is to define a threshold parameter, where all ranked words with a distance greater or equal to this threshold are labeled as changing words.

For cases where no tuning data is available, Kaiser et al. (2020b) propose to choose the threshold according to the population of CDs of all words in the corpus. Kaiser et al. set the threshold to $\mu + \sigma$, where μ is the mean and σ is the standard deviation of the population. We slightly modify this approach by changing the threshold to $\mu + t * \sigma$. In this way, we introduce an additional parameter t, which we tune on the SemEval-2020 test data. We test different values ranging from -2 to 2 in steps of 0.1.

Population Since SGNS generates type-based vectors for every word in the vocabulary, measuring the distances for the full vocabulary comes with low additional computational effort. Unfortunately, this is much more difficult for BERT. Creating up to 100 vectors for every word in the vocabulary drastically increases the computational burden. We choose a population of 500 words for our work allowing us to test multiple parameter configurations.[3] We sample words from different frequency areas to have predictions not only for low-frequency words. For this, we first compute the frequency range (highest frequency – lowest frequency) of the vocabulary. This range is then split into 5 areas of equal frequency width. Random samples from these areas are taken based on how many words they contain. For example: if the lowest frequency area contains 50% of all words from the vocabulary, then $0.5 * 500 = 250$ random samples are taken from this area. The SemEval-2020 target words are excluded from this sampling process. The resulting population is used to create predictions for both models.

Filtering The predictions contain proper names, foreign language and lemmatization errors, which we aim to filter out, as such cases are usually not considered as semantic changes. We only allow nouns, verbs and adjectives to pass. Words where over 10% of the usages are either non-German or contain more than 25% punctuation are filtered out as well.

6 Annotation

The model predictions are validated by human annotation. For this, we apply the SemEval-2020 Task 1 procedure, as described in Schlechtweg et al. (2020). Annotators are asked to judge the semantic relatedness of pairs of word usages, such as the two usages of *Aufkommen* in (1) and (2), on the scale in Table 1.

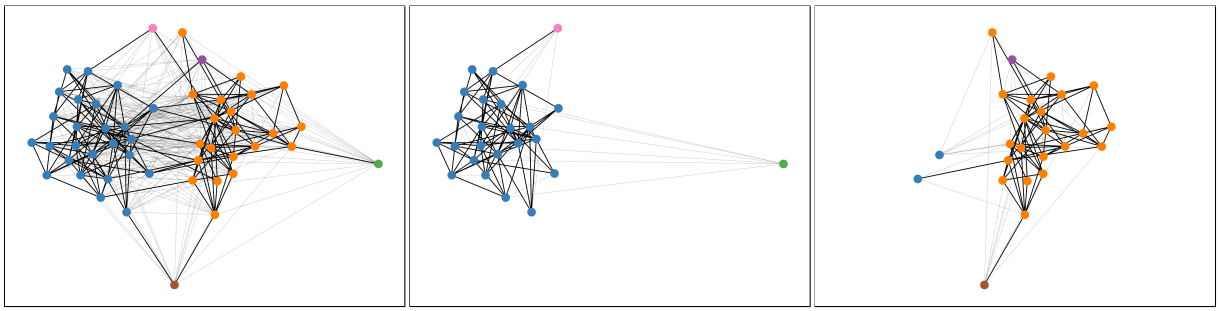
(1) Es ist richtig, dass mit dem **Aufkommen** der Manufaktur im Unterschied zum Handwerk sich Spuren der Kinderexploitation zeigen.
'*It is true that with the **emergence** of the manufactory, in contrast to the handicraft, traces of child labor are showing.*'

(2) Sie wissen, daß wir für das Vieh mehr Futter aus eigenem **Aufkommen** brauchen.
'*They know that we need more feed from our own **production** for the cattle.*'

The annotated data of a word is represented in a Word Usage Graph (WUG), where vertices represent word usages, and weights on edges represent

[2]Find the code used for each step of the prediction process at https://github.com/seinan9/LSCDiscovery.

[3]In a practical setting where predictions have to be generated only once, a much larger number may be chosen. Also, possibilities to scale up BERT performance can be applied (Montariol et al., 2021).



full C_1 C_2

Figure 1: Word Usage Graph of German *Aufkommen* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right). **black**/gray lines indicate **high**/low edge weights.

the (median) semantic relatedness judgment of a pair of usages such as (1) and (2). The final WUGs are clustered with a variation of correlation clustering (Bansal et al., 2004; Schlechtweg et al., 2020) (see Figure 1, left) and split into two subgraphs representing nodes from subcorpora C_1 and C_2, respectively (middle and right). Clusters are then interpreted as word senses and changes in clusters over time as lexical semantic change.

In contrast to Schlechtweg et al. we use the openly available DURel interface for annotation and visualization.[4] This also implies a change in sampling procedure, as the system currently implements only random sampling of use pairs (without SemEval-style optimization). For each target word we sample $|U_1| = |U_2| = 25$ usages (sentences) per subcorpus (C_1, C_2) and upload these to the DURel system, which presents use pairs to annotators in randomized order. We recruit eight German native speakers with university level education as annotators. Five have a background in linguistics, two in German studies, and one has an additional professional background in lexicography. Similar to Schlechtweg et al., we ensure the robustness of the obtained clusterings by continuing the annotation of a target word until all multi-clusters (clusters with more than one usage) in its WUG are connected by at least one judgment. We finally label a target word as changed (binary) if it gained or lost a cluster over time. For instance, *Aufkommen* in Figure 1 is labeled as change as it gains the orange cluster from C_1 to C_2. Following Schlechtweg et al. (2020) we use k and n as lower frequency thresholds to avoid that small random fluctuations in sense frequencies caused by sampling variability or annotation error be misclassified as change. As proposed in Schlechtweg and Schulte im Walde (submitted) for comparability across sample sizes we set $k = 1 \leq 0.01 * |U_i| \leq 3$ and $n = 3 \leq 0.1 * |U_i| \leq 5$, where $|U_i|$ is the number of usages from the respective time period (after removing incomprehensible usages from the graphs). This results in $k = 1$ and $n = 3$ for all target words.

Find an overview over the final set of WUGs in Table 2. We reach a comparably high inter-annotator agreement (Krippendorf's $\alpha = .58$).[5]

7 Results

We now describe the results of the tuning and discovery procedures.

7.1 Tuning

SGNS is commonly used (Schlechtweg et al., 2020) and also highly optimized (Kaiser et al., 2020a,b, 2021), so it is difficult to further increase the performance. We thus rely on the work of Kaiser et al. (2020a) and test their parameter configurations on the German SemEval-2020 data set.[6] We obtain three slightly different parameter configurations (see Table 3 for more details), yielding competitive $\rho = .690$, $\rho = .710$ and $\rho = .710$, respectively.

In order to improve the performance of BERT, we test different layer combinations, preprocessings and semantic change measures. Following Laicher et al. (2020, 2021), we are able to drastically increase the performance of BERT

[4] https://www.ims.uni-stuttgart.de/data/durel-tool.

[5] We provide WUGs as Python NetworkX graphs, descriptive statistics, inferred clusterings, change values and interactive visualizations for all target words and the respective code at https://www.ims.uni-stuttgart.de/data/wugs.

[6] All configurations use $w = 10$, $d = 300$, $e = 5$ and a minimum frequency count of 39.

Data set	n	N/V/A	|U|	AN	JUD	AV	SPR	KRI	UNC	LOSS	LSC_B	LSC_G
SemEval	48	32/14/2	178	8	37k	2	.59	.53	0	.12	.35	.31
Predictions	75	39/16/20	49	8	24k	1	.64	.58	0	.26	.48	.40

Table 2: Overview target words. n = no. of target words, N/V/A = no. of nouns/verbs/adjectives, $|U|$ = avg. no. of usages per word, AN = no. of annotators, JUD = total no. of judged usage pairs, AV = avg. no. of judgments per usage pair, SPR = weighted mean of pairwise Spearman, KRI = Krippendorff's α, UNC = avg. no. of uncompared multi-cluster combinations, LOSS = avg. of normalized clustering loss * 10, $LSC_{B/G}$ = mean binary/graded change score.

on the German SemEval-2020 data. In a preprocessing step, we replace the target word in every usage by its lemma. In combination with layer 12+1, both APD and COS perform competitively well on Subtask 2 ($\rho = .690$ and $\rho = .738$).

After applying thresholding as described in Section 5 we obtain $F_{0.5}$-scores for a large range of thresholds. SGNS achieves peak $F_{0.5}$-scores of .692, .738 and .685, respectively (see Table 3). Interestingly, the optimal threshold is at $t = 1.0$ in all three cases. This corresponds to the threshold used in Kaiser et al. (2020b). While the peak $F_{0.5}$ of BERT+APD is marginally worse (.598 at $t = -0.2$), BERT+COS is able to outperform the best SGNS configuration with a peak of .741 at $t = 0.1$.

In order to obtain an estimate on the sampling variability that is caused by sampling only up to 100 usages per word for BERT+APD and BERT+COS (see Section 4.2), we repeat the whole procedure 9 times and estimate mean and standard deviation of performance on the tuning data. In the beginning of every run the usages are randomly sampled from the corpora. We observe a mean ρ of .657 for BERT+APD and .743 for BERT+COS with a standard deviation of .015 and .012, respectively, as well as a mean $F_{0.5}$ of .576 for BERT+APD and .684 for BERT+COS with a standard deviation of .013 and .038, respectively. This shows that the variability caused by sub-sampling word usages is negligible.

7.2 Discovery

We use the top-performing configurations (see Table 3) to generate two sets of large-scale predictions. While we use the lemmatized corpora for SGNS, in BERT's case we choose the raw corpora with lemmatized target words instead. The latter choice is motivated by the previously described performance increases. After the filtering as described in Section 6, we obtain 27 and 75 words labeled as changing, respectively. We further sample 30 targets from the second set of predictions to obtain a feasible number for annotation. We call the first set SGNS targets and the second one BERT targets, with an overlap of 7 targets. Additionally, we randomly sample 30 words from the population (with an overlap of 5 with the SGNS and BERT targets) in order to have an indication of what the change distribution underlying the corpora is. We call these baseline (BL) targets. This baseline will help us to put the results of the predictions in context and to find out whether the predictions of the two models can be explained by pure randomness. Following the annotation process, binary gold data is generated for all three target sets, in order to validate the quality of the predictions.

The evaluation of the predictions is presented in Table 3. We achieve a $F_{0.5}$-score of .714 for SGNS and .620 for BERT. Out of the 27 words predicted by the SGNS model, 18 (67 %) were actually labeled as changing words by the human annotators. In comparison, only 17 out of the 30 (57 %) BERT predictions were annotated as such. The performance of SGNS for prediction (SGNS targets) is even higher than on the tuning data (SemEval targets). In contrast, BERT's performance for prediction drops strongly in comparison to the performance on the tuning data (.741 vs. .620). This reproduces previous results and confirms that (off-the-shelf) BERT generalises poorly for LSCD and does not transfer well between data sets (Laicher et al., 2020). If we compare these results to the baseline, we can see that both models perform much better than the random baseline ($F_{0.5}$ of .349). Only 10 out of the 30 (30 %) randomly sampled words are annotated as changing. This indicates, that the performance of SGNS and BERT is likely not a cause of randomness. Both models considerably increase the chance of finding changing words compared to a random model.

Figure 2 shows the detailed $F_{0.5}$ developments

	parameters	t	tuning				predictions			
			ρ	$F_{0.5}$	P	R	ρ	$F_{0.5}$	P	R
SGNS	$k = 1, s = .005$	1.0	.690	.692	.750	.529				
	$\mathbf{k = 5, s = .001}$	1.0	.710	**.738**	.818	.529	.295	**.714**	.667	1.0
	$k = 5$, $s =$ None	1.0	.710	.685	.714	.588				
BERT	APD	-0.2	.673	.598	.560	.824				
	COS	0.1	.738	**.741**	.706	.788	.482	.620	.567	1.0
BL	random sampling							.349	.300	1.0

Table 3: Performance (Spearman ρ, $F_{0.5}$-measure, precision P and recall R) of different approaches on tuning data (SemEval targets) and performance of best type- and token-based approach on respective predictions with optimal tuning threshold t, as well as the performance of a randomly sampled baseline.

across different thresholds on the SemEval targets and the predicted words. Increasing the threshold on the predicted words improves the $F_{0.5}$ for both the type-based and token-based approach. A new high-score of .783 at $t = 1.3$ is achievable for SGNS. While BERT's performance also increases to a peak of .714 at $t = 1.0$, it is still lower than in the tuning phase.

7.3 Analysis

For further insights into sources of errors, we take a close look at the false positives, their WUGs and the underlying usages. Most of the wrong predictions can be grouped into one out of two error sources (cf. Kutuzov, 2020, pp. 175–182).

Context change The first category includes words where the context in the usages shifts between time periods, while the meaning stays the same. The WUG of *Angriffswaffe* ('offensive weapon') (see Figure 5 in Appendix A) shows a single cluster for both C_1 and C_2. In the first time period *Angriffswaffe* is used to refer to a hand weapon (such as 'sword', 'spear'). In the second period, however, the context changes to nuclear weaponry. We can see a clear contextual shift, while the meaning did not change. In this case both models are tricked by the change of context. Further false positives in this category are the SGNS targets *Ächtung* ('ostracism') and *aussterben* ('to die out') and the COS targets *Königreich* ('kingdom') and *Waffenruhe* ('ceasefire').

Context variety Words that can be used in a large variety of contexts form the second group of false positives. SGNS falsely predicts *neunjährig* as a changing word. We take a closer look at its WUG (see Figure 6 in Appendix A). We observe that there is only one and the same cluster in both time periods, and the meaning of the target does not change, even though a large variety of contexts exists in both C_1 and C_2. For example: 'which bears oats at **nine years** fertilization', 'courageously, a **nine-year-old** Spaniard did something' and 'after nine years of work'. Both models are misguided by this large context variety. Examples include the SGNS targets *neunjährig* ('9-year-old') and *vorjährig* ('of the previous year') and the COS targets *bemerken* ('to notice') and *durchdenken* ('to think through').

8 Lexicographical evaluation

We now evaluate the usefulness of the proposed semantic change discovery procedure including the annotation system and WUG visualization from a lexicographer's viewpoint. The advantage of our approach lies in providing lexicographers and dictionary makers the choice to take a look into predictions they consider promising with respect to their research objective (disambiguation of word senses, detection of novel senses, detection of archaisms, describing senses in regard to specific discourses etc.) and the type of dictionary. Visualized predictions for target words may be analyzed in regard to single senses, clusters of senses, the semantic proximity of sense clusters and a stylized representation of frequency. Random sampling of usages also offers the opportunity to judge underrepresented senses in a sample that might be infrequent in a corpus or during a specific period of time (although currently a high number of overall annotations would be required in order to do so). Most importantly, the use of a variable number of human annotators has the potential to ensure a more objective analysis of large amounts of corpus

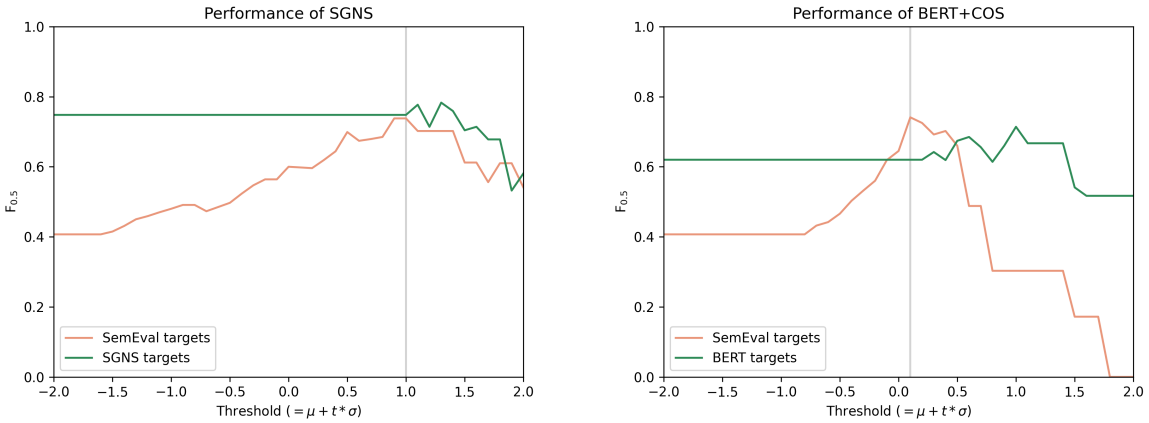


Figure 2: $F_{0.5}$ performance on SemEval targets (orange) and respective predictions (green) across different thresholds. Left: SGNS. Right: BERT+COS. Gray vertical line indicates optimal performance on SemEval targets.

data. In order to evaluate the potential of the approach for assisting lexicographers with extending dictionaries, we analyze statistical measures and predictions of the models provided for the two sets of predictions (SGNS, BERT) and compare them to existing dictionary contents.

We consider overall inter-annotator agreement ($\alpha >= .5$) and annotated binary change label to select 21 target words for lexicographical analysis. In this way, we exclude unclear cases and non-changing words. The target words are analyzed by inspecting cluster visualizations of WUGs (such as in Figure 1) and comparing them to entries in general and specialized dictionaries in order to determine:

- whether a candidate novel sense is already included in one of the reference dictionaries,
- whether a candidate novel sense is included in one of the two reference dictionaries that are consulted for C_1 (covering the period between 1800–1899) and C_2 (covering the period between 1946–1990), indicating the rise of a novel sense, the archaization of older senses or a change in frequency.

Three dictionaries are consulted throughout the analysis: (i) the Dictionary of the German language (DWB) by Jacob und Wilhelm Grimm (digitized version of the 1st print published between 1854–1961), (ii) the Dictionary of Contemporary German (WGD), published between 1964–1977, now curated and digitized by the DWDS and (iii) the Duden online dictionary of German language (DUDEN), reflecting usage of Contemporary German up until today.[7] Additionally, lemma entries in the Wiktionary online dictionary (Wiktionary) are consulted to verify genuinely novel senses described in Section 8.1.

8.1 Records of novel senses

In the case of 17 target words, all senses identified by the system are included in at least one of the three dictionaries consulted for the analysis. In the four remaining cases, at least one novel sense of a word is neither paraphrased nor given as an example of semantically related senses in the dictionaries:

einbinden Reference to the integration or embedding of details on a topic, event, person in respect to a chronological order within written text or visual presentation (e.g. for an exhibition on an author) is judged as a novel sense in close semantic proximity to the old sense 'to bind sth. into sth.', e.g. flowers into a bundle of flowers. *einbinden* is also used in technical contexts, meaning 'to (physically) implement parts of a construction or machine into their intended slots'.

niederschlagen In cases where the verb *niederschlagen* co-occurs with the verb particle *auf* and the noun *Flügel*, the verb refers to a bird's action of repeatedly moving its wings up and down in order to fly.

regelrecht Used as an adverb, *regelrecht* may refer to something being the usual outcome that ought

[7] Only the fully-digitized version of the DWB's first print was consulted for this evaluation, since a revised version has not been completed yet and is only available for lemmas starting with letters a–f.

to be expected due to scientific principles, with an emphasis on the actual result of an action (such as the dyeing of fiber of a piece of clothing following the bleaching process), whereas senses included in dictionaries for general language emphasize either the intended accordance with a rule or something usually happening (the latter being colloquial use).

Zehner (see Figure 3 in Appendix A) The meaning 'a winning sequence of numbers in the national lottery', predicted to have risen as a novel sense between C_1 and C_2, is not included in any of the reference dictionaries.

In most of these cases, senses identified as novel reflect metaphoric use, indicating that definitions in existing dictionary entries may need to be broadened, or example sentences would have to be added. Some of the senses described in this section might be included in specialized dictionaries, e.g. technical usage of *einbinden*.

8.2 Records of changes

For 12 target words, semantic change predicted by the models (innovative, reductive or a salient change of frequency of a sense) correlates with the addition or non-inclusion of senses in dictionary entries consulted for the respective period of time (DWB for C_1, WGD for C_2). It should be noted though, that lemma lists of the two dictionaries might be lacking lemmas in the headword list, and lemma entries might be lacking paraphrases or examples of senses of the lemma, simply because corpus-based lexicography was not available at the time of their first print and revisions of the dictionaries are currently work in progress.

Additionally, we consult a dictionary for Early New High German (FHD) in order to check whether discovered novel senses existed at an earlier stage and may be discovered due to low frequency or sampling error. In two cases, discovered novel senses that are not included in the DWB (for C_1) are found to be included in the FHD.

Interestingly, one sense paraphrased for *Ausrufung* ('a loud wording, a shout') is included in neither of the two dictionaries consulted to judge senses from C_1 and C_2, but in the FHD (earlier) and DUDEN (as of now). These findings suggest that it might be reasonable to use more than two reference corpora. This would also alleviate the corpus bias stemming from idiosyncratic data sampling procedures.

9 Conclusion

We used two state-of-the-art approaches to LSC detection in combination with a recently published high-quality data set to automatically discover semantic changes in a German diachronic corpus pair. While both approaches were able to discover various semantic changes with above-random probability, some of them previously undescribed in etymological dictionaries, the type-based approach showed a clearly better performance.

We validated model predictions by an optimized human annotation process yielding high inter-annotator agreement and providing convenient ways of visualization. In addition, we evaluated the full discovery process from a lexicographer's point of view and conclude that we obtained high-quality predictions, useful visualizations and previously unreported changes. On the other hand, we discovered some issues with respect to the reliability of predictions for semantic change and number and composition of reference corpora that are going to be dealt with in the future. The results of the analyses endorse that our approach might aid lexicographers with extending and altering existing dictionary entries.

Acknowledgments

We thank the three reviewers for their insightful feedback and Pedro González Bascoy for setting up the DURel annotation tool. Dominik Schlechtweg was supported by the Konrad Adenauer Foundation and the CRETA center funded by the German Ministry for Education and Research (BMBF) during the conduct of this study.

References

Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, digitalized edition curated by the Wörterbuchnetz at the Trier Center for Digital Humanities. https://www.woerterbuchnetz.de/DWB. Accessed: 2021-01-07.

Frühneuhochdeutsches Wörterbuch. https://fwb-online.de. Accessed: 2021-01-07.

Wörterbuch der deutschen Gegenwartssprache 1964–1977, curated and provided by the Digital Dictionary of German Language. https://www.dwds.de/d/wb-wdg. Accessed: 2021-01-07.

Nikolay Arefyev and Vasily Zhikov. 2020. BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change

Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine Learning*, 56(1-3):89–113.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Pierpaolo Basile, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. *Diachronic Analysis of the Italian Language exploiting Google Ngram*, pages 56–60.

Pierpaolo Basile and Barbara Mcgillivray. 2018. *Exploiting the Web for Semantic Change Detection*, pages 194–208.

Berliner Zeitung. Diachronic newspaper corpus published by Staatsbibliothek zu Berlin [online]. 2018.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word senses. In *Proceedings of eLex 2013*, pages 49–65.

Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften [online]. 2017.

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

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1156, Copenhagen, Denmark.

DUDEN. Duden online. www.duden.de. Accessed: 2021-02-01.

DWDS. Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. https://www.dwds.de/. Accessed: 02.02.2021.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71, Stroudsburg, PA, USA.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany.

Simon Hengchen, Ruben Ros, and Jani Marjanen. 2019. A data-driven approach to the changing vocabulary of the 'nation' in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In *Proceedings of the Digital Humanities (DH) conference 2019*, Utrecht, The Netherlands.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for Computational Lexical Semantic Change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors,

Computational Approaches to Semantic Change, volume Language Variation, chapter 11. Language Science Press, Berlin.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Adam Jatowt, Nina Tahmasebi, and Lars Borin. 2021. Computational approaches to lexical semantic change: Visualization systems and novel applications. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, Language Variation, chapter 10. Language Science Press, Berlin.

Jens Kaiser, Sinan Kurtyigit, Serge Kotchourko, and Dominik Schlechtweg. 2021. Effects of Pre- and Post-Processing on type-based Embeddings in Lexical Semantic Change Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jens Kaiser, Dominik Schlechtweg, Sean Papay, and Sabine Schulte im Walde. 2020a. IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020b. OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org. Winning Submission!

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *LTCSS@ACL*, pages 61–65. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web, WWW*, pages 625–635, Florence, Italy.

Andrey Kutuzov. 2020. Distributional word embeddings in modeling diachronic semantic change.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

Severin Laicher, Gioia Baldissin, Enrique Castaneda, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

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 Student Research Workshop at the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Nikola Ljubešić. 2020. "deep lexicography" – fad or opportunity? "duboka leksikografija" – pomodnost ili prilika? *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46:839–852.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 343—-349, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Neues Deutschland. Diachronic newspaper corpus published by Staatsbibliothek zu Berlin [online]. 2018.

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.

Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Ondřej Pražák, Pavel Přibáň, and Stephen Taylor. 2020. UWB @ DIACR-Ita: Lexical Semantic Change Detection with CCA and Orthogonal Transformation. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Julia Rodina and Andrey Kutuzov. 2020. Rusemshift: a dataset of historical lexical semantic change in russian. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. Association for Computational Linguistics.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 474–484, New Orleans, Louisiana.

Gerard Salton and Michael J McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, New York.

Dominik Schlechtweg, Anna Hätty, Marco del Tredici, and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. 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 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Dominik Schlechtweg and Sabine Schulte im Walde. submitted. Clustering Word Usage Graphs: A Flexible Framework to Measure Changes in Contextual Word Meaning.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages.

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

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv e-prints*.

Nina Tahmasebi and Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 741–749, Varna, Bulgaria.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2017. Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1195–1204, Valencia, Spain. Association for Computational Linguistics.

Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. 2019. Mining the UK web archive for semantic change detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1212–1221, Varna, Bulgaria. INCOMA Ltd.

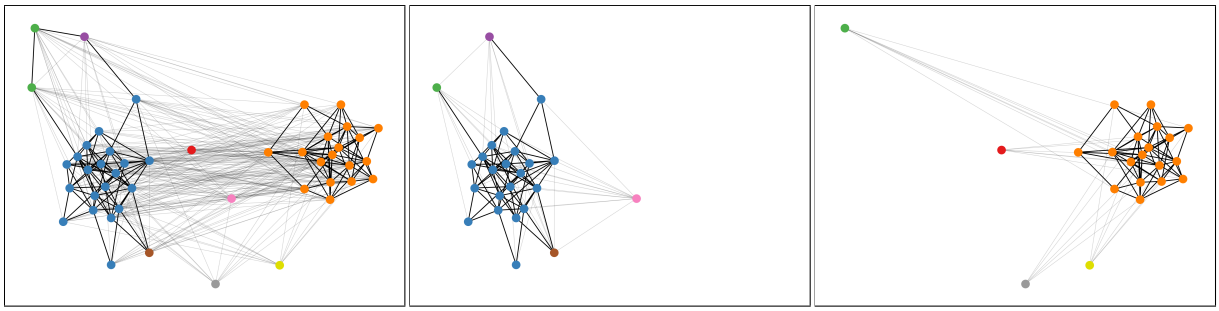
Adam Tsakalidis and Maria Liakata. 2020. Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8485–8497, Online. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188.

Wiktionary. Wiktionary, das freie Wörterbuch. https://de.wiktionary.org. Accessed: 2021-01-07.

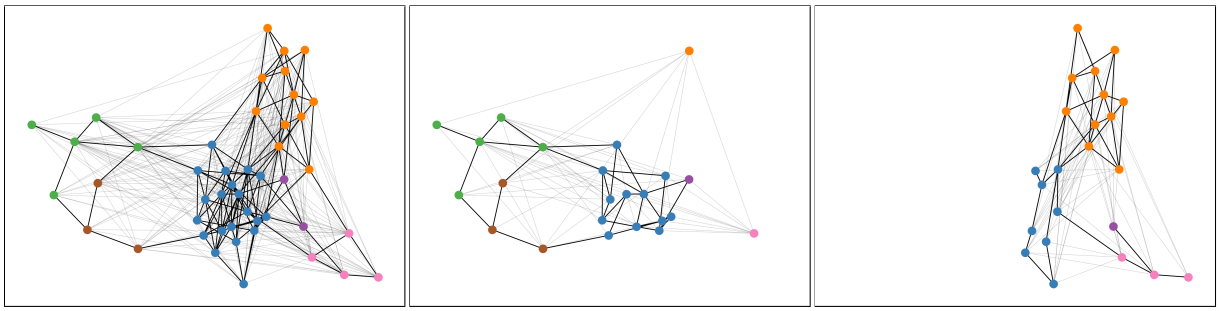
A Additional plots

Please find additional plots of Word Usage Graphs in Figures 3–6.



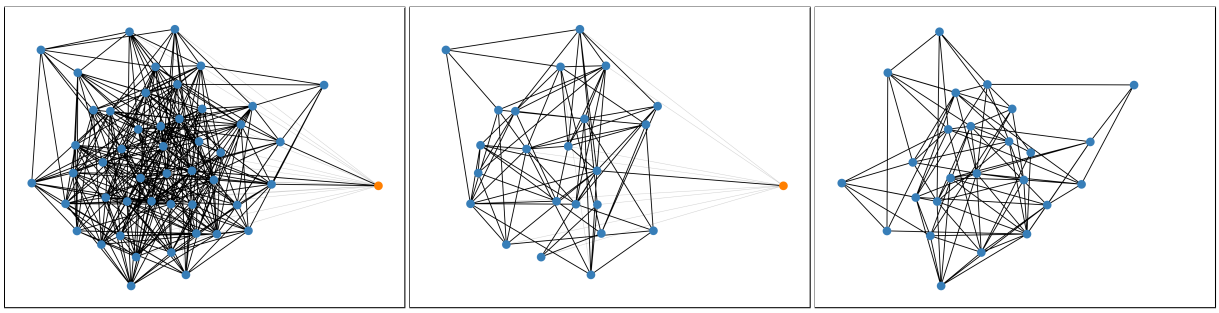
full C_1 C_2

Figure 3: Word Usage Graph of German *Zehner* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).



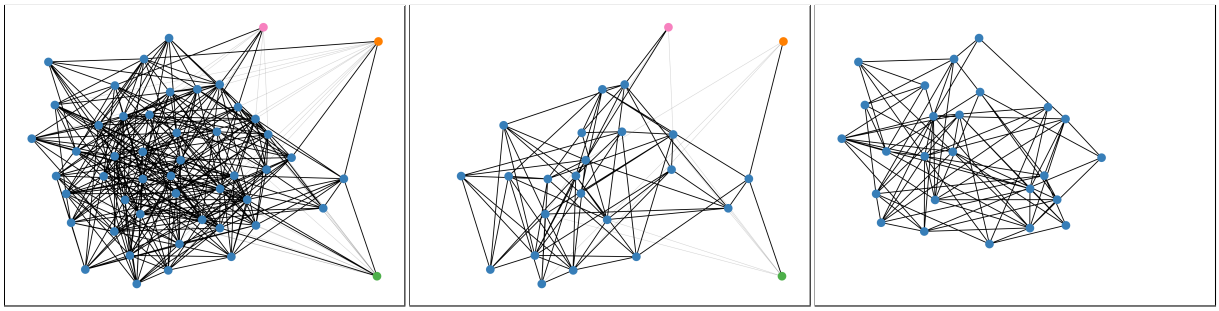
full C_1 C_2

Figure 4: Word Usage Graph of German *Lager* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).



full C_1 C_2

Figure 5: Word Usage Graph of German *Anriffswaffe* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).



full C_1 C_2

Figure 6: Word Usage Graph of German *neunjährig* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).