A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains

Dominik Schlechtweg[1], Anna Hätty[1,2], Marco del Tredici[3], Sabine Schulte im Walde[1]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Robert Bosch GmbH, Corporate Research
[3]Institute for Logic, Language and Computation, University of Amsterdam
{schlecdk,schulte}@ims.uni-stuttgart.de,
anna.haetty@de.bosch.com, m.deltredici@uva.nl

Abstract

We perform an interdisciplinary large-scale evaluation for detecting lexical semantic divergences in a diachronic and in a synchronic task: semantic sense changes across time, and semantic sense changes across domains. Our work addresses the superficialness and lack of comparison in assessing models of diachronic lexical change, by bringing together and extending benchmark models on a common state-of-the-art evaluation task. In addition, we demonstrate that the same evaluation task and modelling approaches can successfully be utilised for the synchronic detection of domain-specific sense divergences in the field of term extraction.

1 Introduction

Diachronic *Lexical Semantic Change (LSC)* detection, i.e., the automatic detection of word sense changes over time, is a flourishing new field within NLP (Frermann and Lapata, 2016; Hamilton et al., 2016b; Schlechtweg et al., 2017, i.a.).[1] Yet, it is hard to compare the performances of the various models, and optimal parameter choices remain unclear, because up to now most models have been compared on different evaluation tasks and data. Presently, we do not know which model performs best under which conditions, and if more complex model architectures gain performance benefits over simpler models. This situation hinders advances in the field and favors unfelicitous drawings of statistical laws of diachronic LSC (Dubossarsky et al., 2017).

In this study, we provide the first large-scale evaluation of an extensive number of approaches. Relying on an existing German LSC dataset we compare models regarding different combinations of semantic representations, alignment techniques and detection measures, while exploring various pre-processing and parameter settings. Furthermore, we introduce *Word Injection* to LSC, a modeling idea drawn from term extraction, that overcomes the problem of vector space alignment. Our comparison of state-of-the-art approaches identifies best models and optimal parameter settings, and it suggests modifications to existing models which consistently show superior performance.

Meanwhile, the detection of lexical sense divergences across time-specific corpora is not the only possible application of LSC detection models. In more general terms, they have the potential to detect sense divergences between corpora of any type, not necessarily time-specific ones. We acknowledge this observation and further explore a *synchronic LSC detection* task: identifying domain-specific changes of word senses in comparison to general-language usage, which is addressed, e.g., in term identification and automatic term extraction (Drouin, 2004; Pérez, 2016; Hätty and Schulte im Walde, 2018), and in determining social and dialectal language variations (Del Tredici and Fernández, 2017; Hovy and Purschke, 2018).[2]

For addressing the synchronic LSC task, we present a recent sense-specific term dataset (Hätty et al., 2019) that we created analogously to the existing diachronic dataset, and we show that the diachronic models can be successfully applied to the synchronic task as well. This two-fold evaluation assures robustness and reproducibility of our model comparisons under various conditions.

[1]An example for diachronic LSC is the German noun *Vorwort* (Paul, 2002), which was mainly used in the meaning of 'preposition' before ≈1800. Then *Vorwort* rapidly acquired a new meaning 'preface', which after 1850 has nearly exclusively been used.

[2]An example for domain-specific synchronic LSC is the German noun *Form*. In general-language use, *Form* means 'shape'/'form', while in the cooking domain the predominant meaning is the domain-specific 'baking tin'.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 732–746
Florence, Italy, July 28 - August 2, 2019.

2 Related Work

Diachronic LSC Detection. Existing approaches for diachronic LSC detection are mainly based on three types of meaning representations: (i) semantic vector spaces, (ii) topic distributions, and (iii) sense clusters. In (i), semantic vector spaces, each word is represented as two vectors reflecting its co-occurrence statistics at different periods of time (Gulordava and Baroni, 2011; Kim et al., 2014; Xu and Kemp, 2015; Eger and Mehler, 2016; Hamilton et al., 2016a,b; Hellrich and Hahn, 2016; Rosenfeld and Erk, 2018). LSC is typically measured by the cosine distance (or some alternative metric) between the two vectors, or by differences in contextual dispersion between the two vectors (Kisselew et al., 2016; Schlechtweg et al., 2017). (ii) Diachronic topic models infer a probability distribution for each word over different word senses (or topics), which are in turn modeled as a distribution over words (Wang and McCallum, 2006; Bamman and Crane, 2011; Wijaya and Yeniterzi, 2011; Lau et al., 2012; Mihalcea and Nastase, 2012; Cook et al., 2014; Frermann and Lapata, 2016). LSC of a word is measured by calculating a novelty score for its senses based on their frequency of use. (iii) Clustering models assign all uses of a word into sense clusters based on some contextual property (Mitra et al., 2015). Word sense clustering models are similar to topic models in that they map uses to senses. Accordingly, LSC of a word is measured similarly as in (ii). For an overview on diachronic LSC detection, see Tahmasebi et al. (2018).

Synchronic LSC Detection. We use the term synchronic LSC to refer to NLP research areas with a focus on how the meanings of words vary across domains or communities of speakers. Synchronic LSC per se is not widely researched; for meaning shifts across domains, there is strongly related research which is concerned with domain-specific word sense disambiguation (Maynard and Ananiadou, 1998; Chen and Al-Mubaid, 2006; Taghipour and Ng, 2015; Daille et al., 2016) or term ambiguity detection (Baldwin et al., 2013; Wang et al., 2013). The only notable work for explicitly measuring across domain meaning shifts is Ferrari et al. (2017), which is based on semantic vector spaces and cosine distance. Synchronic LSC across communities has been investigated as meaning variation in online communities, leveraging the large-scale data which has become available thanks to online social platforms (Del Tredici and Fernández, 2017; Rotabi et al., 2017).

Evaluation. Existing evaluation procedures for LSC detection can be distinguished into evaluation on (i) empirically observed data, and (ii) synthetic data or related tasks. (i) includes case studies of individual words (Sagi et al., 2009; Jatowt and Duh, 2014; Hamilton et al., 2016a), stand-alone comparison of a few hand-selected words (Wijaya and Yeniterzi, 2011; Hamilton et al., 2016b; Del Tredici and Fernández, 2017), comparison of hand-selected changing vs. semantically stable words (Lau et al., 2012; Cook et al., 2014), and post-hoc evaluation of the predictions of the presented models (Cook and Stevenson, 2010; Kulkarni et al., 2015; Del Tredici et al., 2016; Eger and Mehler, 2016; Ferrari et al., 2017). Schlechtweg et al. (2017) propose a small-scale annotation of diachronic metaphoric change.

Synthetic evaluation procedures (ii) include studies that simulate LSC (Cook and Stevenson, 2010; Kulkarni et al., 2015; Rosenfeld and Erk, 2018), evaluate sense assignments in WordNet (Mitra et al., 2015; Frermann and Lapata, 2016), identify text creation dates, (Mihalcea and Nastase, 2012; Frermann and Lapata, 2016), or predict the log-likelihood of textual data (Frermann and Lapata, 2016).

Overall, the various studies use different evaluation tasks and data, with little overlap. Most evaluation data has not been annotated. Models were rarely compared to previously suggested ones, especially if the models differed in meaning representations. Moreover, for the diachronic task, synthetic datasets are used which do not reflect actual diachronic changes.

3 Task and Data

Our study makes use of the evaluation framework proposed in Schlechtweg et al. (2018), where diachronic LSC detection is defined as a comparison between word uses in two time-specific corpora. We further applied the framework to create an analogous synchronic LSC dataset that compares word uses across general-language and domain-specific corpora. The common, meta-level task in our diachronic+synchronic setup is, given two corpora C_a and C_b, to rank the targets in the respective datasets according to their degree of relatedness between word uses in C_a and C_b.

	Times		Domains	
	DTA18	DTA19	SDEWAC	COOK
L_{ALL}	26M	40M	109M	1M
L/P	10M	16M	47M	0.6M

Table 1: Corpora and their approximate sizes.

3.1 Corpora

DTA (Deutsches Textarchiv, 2017) is a freely available lemmatized, POS-tagged and spelling-normalized diachronic corpus of German containing texts from the 16th to the 20th century.

COOK is a domain-specific corpus. We crawled cooking-related texts from several categories (recipes, ingredients, cookware and cooking techniques) from the German cooking recipes websites *kochwiki.de* and *Wikibooks Kochbuch*[3].

SDEWAC (Faaß and Eckart, 2013) is a cleaned version of the web-crawled corpus DEWAC (Baroni et al., 2009). We reduced SDEWAC to 1/8th of its original size by selecting every 8th sentence for our general-language corpus.

Table 1 summarizes the corpus sizes after applying pre-processing. See Appendix A for pre-processing details.

3.2 Datasets[4] and Evaluation

Diachronic Usage Relatedness (DURel). DURel is a gold standard for diachronic LSC consisting of 22 target words with varying degrees of LSC (Schlechtweg et al., 2018). Target words were chosen from a list of attested changes in a diachronic semantic dictionary (Paul, 2002), and for each target a random sample of use pairs from the DTA corpus was annotated for meaning relatedness of the uses on a scale from 1 (unrelated meanings) to 4 (identical meanings), both within and across the time periods 1750–1799 and 1850–1899. The annotation resulted in an average Spearman's $\rho = 0.66$ across five annotators and 1,320 use pairs. For our evaluation of diachronic meaning change we rely on the ranking of the target words according to their mean usage relatedness across the two time periods.

Synchronic Usage Relatedness (SURel). SURel is a recent gold standard for synchronic LSC (Hätty et al., 2019) using the same framework as in DURel. The 22 target words were chosen such as to exhibit different degrees of domain-specific meaning shifts, and use pairs were randomly selected from SDEWAC as general-language corpus and from COOK as domain-specific corpus. The annotation for usage relatedness across the corpora resulted in an average Spearman's $\rho = 0.88$ across four annotators and 1,320 use pairs. For our evaluation of synchronic meaning change we rely on the ranking of the target words according to their mean usage relatedness between general-language and domain-specific uses.

Evaluation. The gold LSC ranks in the DURel and SURel datasets are used to assess the correctness of model predictions by applying Spearman's rank-order correlation coefficient ρ as evaluation metric, as done in similar previous studies (Gulordava and Baroni, 2011; Schlechtweg et al., 2017; Schlechtweg and Schulte im Walde, 2018). As corpus data underlying the experiments we rely on the corpora from which the annotated use pairs were sampled: DTA documents from 1750–1799 as C_a and documents from 1850–1899 as C_b for the diachronic experiments, and the SDEWAC corpus as C_a and the COOK corpus as C_b for the synchronic experiments.

4 Meaning Representations[5]

Our models are based on two families of distributional meaning representations: semantic vector spaces (Section 4.1), and topic distributions (Section 4.2). All representations are bag-of-words-based, i.e. each word representation reflects a weighted bag of context words. The contexts of a target word w_i are the words surrounding it in an n-sized window: $w_{i-n}, ..., w_{i-1}, w_{i+1}, ..., w_{i+n}$.

4.1 Semantic Vector Spaces

A semantic vector space constructed from a corpus C with vocabulary V is a matrix M, where each row vector represents a word w in the vocabulary V reflecting its co-occurrence statistics (Turney and Pantel, 2010). We compare two state-of-the-art approaches to learn these vectors from co-occurrence data, (i) counting and (ii) predicting, and construct vector spaces for each time period and domain.

[3] de.wikibooks.org/wiki/Kochbuch

[4] The datasets are available in Appendix C and at https://github.com/Garrafao/LSCDetection.

[5] Find the hyperparameter settings in Appendix A. The scripts for vector space creation, alignment, measuring LSC and evaluation are available at https://github.com/Garrafao/LSCDetection.

4.1.1 Count-based Vector Spaces

In a count-based semantic vector space the matrix M is high-dimensional and sparse. The value of each matrix cell $M_{i,j}$ represents the number of co-occurrences of the word w_i and the context c_j, $\#(w_i, c_j)$. In line with Hamilton et al. (2016b) we apply a number of transformations to these raw co-occurrence matrices, as previous work has shown that this improves results on different tasks (Bullinaria and Levy, 2012; Levy et al., 2015).

Positive Pointwise Mutual Information (PPMI). In PPMI representations the co-occurrence counts in each matrix cell $M_{i,j}$ are weighted by the positive mutual information of target w_i and context c_j reflecting their degree of association. The values of the transformed matrix are

$$M_{i,j}^{\text{PPMI}} = \max\left\{\log\left(\frac{\#(w_i, c_j)\sum_c \#(c)^\alpha}{\#(w_i)\#(c_j)^\alpha}\right) - \log(k), 0\right\},$$

where $k > 1$ is a prior on the probability of observing an actual occurrence of (w_i, c_j) and $0 < \alpha < 1$ is a smoothing parameter reducing PPMI's bias towards rare words (Levy and Goldberg, 2014; Levy et al., 2015).

Singular Value Decomposition (SVD). Truncated SVD finds the optimal rank d factorization of matrix M with respect to L2 loss (Eckart and Young, 1936). We use truncated SVD to obtain low-dimensional approximations of the PPMI representations by factorizing M^{PPMI} into the product of the three matrices $U\Sigma V^\top$. We keep only the top d elements of Σ and obtain

$$M^{\text{SVD}} = U_d \Sigma_d^p,$$

where p is an eigenvalue weighting parameter (Levy et al., 2015). The ith row of M^{SVD} corresponds to w_i's d-dimensional representation.

Random Indexing (RI). RI is a dimensionality reduction technique based on the Johnson-Lindenstrauss lemma according to which points in a vector space can be mapped into a randomly selected subspace under approximate preservation of the distances between points, if the subspace has a sufficiently high dimensionality (Johnson and Lindenstrauss, 1984; Sahlgren, 2004). We reduce the dimensionality of a count-based matrix M by multiplying it with a random matrix R:

$$M^{\text{RI}} = MR^{|\mathcal{V}|\times d},$$

where the ith row of M^{RI} corresponds to w_i's d-dimensional semantic representation. The choice of the random vectors corresponding to the rows in R is important for RI. We follow previous work (Basile et al., 2015) and use sparse ternary random vectors with a small number s of randomly distributed -1s and $+1$s, all other elements set to 0, and we apply subsampling with a threshold t.

4.1.2 Predictive Vector Spaces

Skip-Gram with Negative Sampling (SGNS) differs from count-based techniques in that it directly represents each word $w \in V$ and each context $c \in V$ as a d-dimensional vector by implicitly factorizing $M = WC^\top$ when solving

$$\arg\max_\theta \sum_{(w,c)\in D} \log\sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log\sigma(-v_c \cdot v_w),$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$, D is the set of all observed word-context pairs and D' is the set of randomly generated negative samples (Mikolov et al., 2013a,b; Goldberg and Levy, 2014). The optimized parameters θ are $v_{c_i} = C_{i*}$ and $v_{w_i} = W_{i*}$ for $w, c \in V$, $i \in 1, ..., d$. D' is obtained by drawing k contexts from the empirical unigram distribution $P(c) = \frac{\#(c)}{|D|}$ for each observation of (w,c), cf. Levy et al. (2015). SGNS and PPMI representations are highly related in that the cells of the implicitly factorized matrix M are PPMI values shifted by the constant k (Levy and Goldberg, 2014). Hence, SGNS and PPMI share the hyperparameter k. The final SGNS matrix is given by

$$M^{\text{SGNS}} = W,$$

where the ith row of M^{SGNS} corresponds to w_i's d-dimensional semantic representation. As in RI we apply subsampling with a threshold t. SGNS with particular parameter configurations has shown to outperform transformed count-based techniques on a variety of tasks (Baroni et al., 2014; Levy et al., 2015).

4.1.3 Alignment

Column Intersection (CI). In order to make the matrices A and B from time periods $a < b$ (or domains a and b) comparable, they have to be aligned via a common coordinate axis. This is rather straightforward for count and PPMI representations, because their columns correspond to context words which often occur in both A and B (Hamilton et al., 2016b). In this case, the alignment for A and B is

$$A^{\text{CI}}_{*j} = A_{*j} \quad \text{for all } c_j \in V_a \cap V_b,$$
$$B^{\text{CI}}_{*j} = B_{*j} \quad \text{for all } c_j \in V_a \cap V_b,$$

where X_{*j} denotes the jth column of X.

Shared Random Vectors (SRV). RI offers an elegant way to align count-based vector spaces and reduce their dimensionality at the same time (Basile et al., 2015). Instead of multiplying count matrices A and B each by a separate random matrix R_A and R_B they may be multiplied both by the same random matrix R representing them in the same low-dimensional random space. Hence, A and B are aligned by

$$A^{\text{SVR}} = AR,$$
$$B^{\text{SVR}} = BR.$$

We follow Basile et al. and adopt a slight variation of this procedure: instead of multiplying both matrices by exactly the same random matrix (corresponding to an intersection of their columns) we first construct a shared random matrix and then multiply A and B by the respective sub-matrix.

Orthogonal Procrustes (OP). In the low-dimensional vector spaces produced by SVD, RI and SGNS the columns may represent different coordinate axes (orthogonal variants) and thus cannot directly be aligned to each other. Following Hamilton et al. (2016b) we apply OP analysis to solve this problem. We represent the dictionary as a binary matrix D, so that $D_{i,j} = 1$ if $w_i \in V_b$ (the ith word in the vocabulary at time b) corresponds to $w_j \in V_a$. The goal is then to find the optimal mapping matrix W^* such that the sum of squared Euclidean distances between B's mapping $B_{i*}W$ and A_{j*} for the dictionary entries $D_{i,j}$ is minimized:

$$W^* = \arg\min_W \sum_i \sum_j D_{i,j} \|B_{i*}W - A_{j*}\|^2.$$

Following standard practice we length-normalize and mean-center A and B in a pre-processing step (Artetxe et al., 2017), and constrain W to be orthogonal, which preserves distances within each time period. Under this constraint, minimizing the squared Euclidean distance becomes equivalent to maximizing the dot product when finding the optimal rotational alignment (Hamilton et al., 2016b; Artetxe et al., 2017). The optimal solution for this problem is then given by $W^* = UV^\top$, where $B^\top DA = U\Sigma V^\top$ is the SVD of $B^\top DA$. Hence, A and B are aligned by

$$A^{\text{OP}} = A,$$
$$B^{\text{OP}} = BW^*,$$

where A and B correspond to their preprocessed versions. We also experiment with two variants: OP$_-$ omits mean-centering (Hamilton et al., 2016b), which is potentially harmful as a better solution may be found after mean-centering. OP$_+$ corresponds to OP with additional pre- and post-processing steps and has been shown to improve performance in research on bilingual lexicon induction (Artetxe et al., 2018a,b). We apply all OP variants only to the low-dimensional matrices.

Vector Initialization (VI). In VI we first learn A^{VI} using standard SGNS and then initialize the SGNS model for learning B^{VI} on A^{VI} (Kim et al., 2014). The idea is that if a word is used in similar contexts in a and b, its vector will be updated only slightly, while more different contexts lead to a stronger update.

Word Injection (WI). Finally, we use the word injection approach by Ferrari et al. (2017) where target words are substituted by a placeholder in one corpus before learning semantic representations, and a single matrix M^{WI} is constructed for both corpora after mixing their sentences. The advantage of this approach is that all vector learning methods described above can be directly applied to the mixed corpus, and target vectors are constructed directly in the same space, so no post-hoc alignment is necessary.

4.2 Topic Distributions

Sense ChANge (SCAN). SCAN models LSC of word senses via smooth and gradual changes in associated topics (Frermann and Lapata, 2016). The semantic representation inferred for a target word w and time period t consists of a K-dimensional distribution over word senses ϕ^t and a V-dimensional distribution over the vocabulary $\psi^{t,k}$ for each word sense k, where K is a predefined number of senses for target word w. SCAN places parametrized logistic normal priors on ϕ^t and $\psi^{t,k}$ in order to encourage a smooth change of parameters, where the extent of change is controlled through the precision parameter K^ϕ, which is learned during training.

Although $\psi^{t,k}$ may change over time for word sense k, senses are intended to remain thematically consistent as controlled by word precision parameter K^{ψ}. This allows comparison of the topic distribution across time periods. For each target word w we infer a SCAN model for two time periods a and b and take ϕ_w^a and ϕ_w^b as the respective semantic representations.

5 LSC Detection Measures

LSC detection measures predict a degree of LSC from two time-specific semantic representations of a word w. They either capture the contextual similarity (Section 5.1) or changes in the contextual dispersion (Section 5.2) of w's representations.[6]

5.1 Similarity Measures

Cosine Distance (CD). CD is based on cosine similarity which measures the cosine of the angle between two non-zero vectors $\vec{x}, \vec{y}$ with equal magnitudes (Salton and McGill, 1983):

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\sqrt{\vec{x} \cdot \vec{x}}\sqrt{\vec{y} \cdot \vec{y}}}.$$

The cosine distance is then defined as

$$CD(\vec{x}, \vec{y}) = 1 - cos(\vec{x}, \vec{y}).$$

CD's prediction for a degree of LSC of w between time periods a and b is obtained by $CD(\vec{w}_a, \vec{w}_b)$.

Local Neighborhood Distance (LND). LND computes a second-order similarity for two non-zero vectors $\vec{x}, \vec{y}$ (Hamilton et al., 2016a). It measures the extent to which $\vec{x}$ and $\vec{y}$'s distances to their shared nearest neighbors differ. First the cosine similarity of $\vec{x}, \vec{y}$ with each vector in the union of the sets of their k nearest neighbors $N_k(\vec{x})$ and $N_k(\vec{y})$ is computed and represented as a vector s whose entries are given by

$$s(j) = \cos(\vec{x}, \vec{z}_j) \quad \forall \vec{z}_j \in N_k(\vec{x}) \cup N_k(\vec{y}).$$

LND is then computed as cosine distance between the two vectors:

$$LND(\vec{x}, \vec{y}) = CD(\vec{s_x}, \vec{s_y}).$$

LND does not require matrix alignment, because it measures the distances to the nearest neighbors in each space separately. It was claimed to capture changes in paradigmatic rather than syntagmatic relations between words (Hamilton et al., 2016a).

[6]Find an overview of which measure was applied to which representation type in Appendix A.

Jensen-Shannon Distance (JSD). JSD computes the distance between two probability distributions ϕ_x, ϕ_y of words w_x, w_y (Lin, 1991; Donoso and Sanchez, 2017). It is the symmetrized square root of the Kullback-Leibler divergence:

$$JSD(\phi_x||\phi_y) = \sqrt{\frac{D_{KL}(\phi_x||M) + D_{KL}(\phi_y||M)}{2}},$$

where $M = (\phi_x + \phi_y)/2$. JSD is high if ϕ_x and ϕ_y assign different probabilities to the same events.

5.2 Dispersion Measures

Frequency Difference (FD). The log-transformed relative frequency of a word w for a corpus C is defined by

$$F(w, C) = \log \frac{|w \in C|}{|C|}$$

FD of two words x and y in two corpora X and Y is then defined by the absolute difference in F:

$$FD(x, X, y, Y) = |F(x, X) - F(y, Y)|$$

FD's prediction for w's degree of LSC between time periods a and b with corpora C_a and C_b is computed as $FD(w, C_a, w, C_b)$ (parallel below).

Type Difference (TD). TD is similar to FD, but based on word vectors $\vec{w}$ for words w. The normalized log-transformed number of context types of a vector $\vec{w}$ in corpus C is defined by

$$T(\vec{w}, C) = \log \frac{\sum_{i=1} 1 \quad \text{if } \vec{w}_i \neq 0}{|C_T|},$$

where $|C_T|$ is the number of types in corpus C. The TD of two vectors $\vec{x}$ and $\vec{y}$ in two corpora X and Y is the absolute difference in T:

$$TD(\vec{x}, X, \vec{y}, Y) = |T(\vec{x}, X) - T(\vec{y}, Y)|.$$

Entropy Difference (HD). HD relies on vector entropy as suggested by Santus et al. (2014). The entropy of a non-zero word vector $\vec{w}$ is defined by

$$VH(\vec{w}) = -\sum_{i=1} \frac{\vec{w}_i}{\sum_{j=1} \vec{w}_j} \log \frac{\vec{w}_i}{\sum_{j=1} \vec{w}_j}.$$

VH is based on Shannon's entropy (Shannon, 1948), which measures the unpredictability of w's

Dataset	Preproc	Win	Space	Parameters	Align	Measure	Spearman m (h, l)
DURel	L_{ALL}	10	SGNS	k=1,t=None	OP	CD	**0.866** (0.914, 0.816)
	L_{ALL}	10	SGNS	k=5,t=None	OP	CD	0.857 (0.891, 0.830)
	L_{ALL}	5	SGNS	k=5,t=0.001	OP	CD	0.835 (0.872, 0.814)
	L_{ALL}	10	SGNS	k=5,t=0.001	OP	CD	0.826 (0.863, 0.768)
	L/P	2	SGNS	k=5,t=None	OP	CD	0.825 (0.826, 0.818)
SURel	L/P	2	SGNS	k=1,t=0.001	OP	CD	**0.851** (0.851, 0.851)
	L/P	2	SGNS	k=5,t=None	OP	CD	0.850 (0.850, 0.850)
	L/P	2	SGNS	k=5,t=0.001	OP	CD	0.834 (0.838, 0.828)
	L/P	2	SGNS	k=5,t=0.001	OP_	CD	0.831 (0.836, 0.817)
	L/P	2	SGNS	k=5,t=0.001	OP	CD	0.829 (0.832, 0.823)

Table 2: Best results of ρ scores (Win=Window Size, Preproc=Preprocessing, Align=Alignment, k=negative sampling, t=subsampling, Spearman m(h,l): mean, highest and lowest results).

co-occurrences (Schlechtweg et al., 2017). HD is defined as

$$HD(\vec{x}, \vec{y}) = |VH(\vec{x}) - VH(\vec{y})|.$$

We also experiment with differences in H between topic distributions ϕ_w^a, ϕ_w^b, which are computed in a similar fashion, and with normalizing VH by dividing it by $\log(VT(\vec{w}))$, its maximum value.

6 Results and Discussions

First of all, we observe that nearly all model predictions have a strong positive correlation with the gold rank. Table 2 presents the overall best results across models and parameters.[7] With $\rho = 0.87$ for diachronic LSC (DURel) and $\rho = 0.85$ for synchronic LSC (SURel), the models reach comparable and unexpectedly high performances on the two distinct datasets. The overall best-performing model is Skip-Gram with orthogonal alignment and cosine distance (SGNS+OP+CD). The model is robust in that it performs best on both datasets and produces very similar, sometimes the same results across different iterations.

Pre-processing and Parameters. Regarding pre-processing, the results are less consistent: L_{ALL} (all lemmas) dominates in the diachronic task, while L/P (lemma:pos of content words) dominates in the synchronic task. In addition, L/P pre-processing, which is already limited on content words, prefers shorter windows, while L_{ALL} (pre-processing where the complete sentence structure is maintained) prefers longer windows. Regarding the preference of L/P for SURel, we blame noise in the COOK corpus, which contains a lot of recipes listing ingredients and quantities with numerals and abbreviations, to presumably contribute little information about context words. For instance, COOK contains 4.6% numerals, while DTA only contains 1.2% numerals.

Looking at the influence of subsampling, we find that it does not improve the mean performance for Skip-Gram (SGNS) (with $\rho = 0.506$, without $\rho = 0.517$), but clearly for Random Indexing (RI) (with $\rho = 0.413$, without $\rho = 0.285$). Levy et al. (2015) found that SGNS prefers numerous negative samples ($k > 1$), which is confirmed here: mean ρ with $k = 1$ is 0.487, and mean ρ with $k = 5$ is 0.535.[8] This finding is also indicated in Table 2, where $k = 5$ dominates the 5 best results on both datasets; yet, $k = 1$ provides the overall best result on both datasets.

Semantic Representations. Table 3 shows the best and mean results for different semantic representations. SGNS is clearly the best vector space model, even though its mean performance does not outperform other representations as clearly as its best performance. Regarding count models, PPMI and SVD show the best results.

SCAN performs poorly, and its mean results indicate that it is rather unstable. This may be explained by the particular way in which SCAN constructs context windows: it ignores asymmetric windows, thus reducing the number of training instances considerably, in particular for large window sizes.

Alignments. The fact that our modification of Hamilton et al. (2016b) (SGNS+OP) performs best across datasets confirms our assumption that column-mean centering is an important preprocessing step in Orthogonal Procrustes analysis and

[7] For models with randomness we computed the average results of five iterations.

[8] For PPMI we observe the opposite preference, mean ρ with $k = 1$ is 0.549 and mean ρ with $k = 5$ is 0.439.

Dataset	Representation	best	mean
DURel	raw count	0.639	0.395
	PPMI	0.670	0.489
	SVD	0.728	0.498
	RI	0.601	0.374
	SGNS	**0.866**	**0.502**
	SCAN	0.327	0.156
SURel	raw count	0.599	0.120
	PPMI	0.791	0.500
	SVD	0.639	0.300
	RI	0.622	0.299
	SGNS	**0.851**	**0.520**
	SCAN	0.082	-0.244

Table 3: Best and mean ρ scores across similarity measures (CD, LND, JSD) on semantic representations.

should not be omitted.

Additionally, the mean performance in Table 4 shows that OP is generally more robust than its variants. OP_+ has the best mean performance on DURel, but performs poorly on SURel. Artetxe et al. (2018a) show that the additional pre- and post-processing steps of OP_+ can be harmful in certain conditions. We tested the influence of the different steps and identified the non-orthogonal whitening transformation as the main reason for a performance drop of ≈20%.

In order to see how important the alignment step is for the low-dimensional embeddings (SVD/RI/SGNS), we also tested the performance without alignment ('None' in Table 4). As expected, the mean performance drops considerably. However, it remains positive, which suggests that the spaces learned in the models are not random but rather slightly rotated variants.

Especially interesting is the comparison of Word Injection (WI) where one common vector space is learned against the OP-models where two separately learned vector spaces are aligned. Although WI avoids (post-hoc) alignment altogether, it is consistently outperformed by OP, which is shown in Table 4 for low-dimensional embeddings.[9] We found that OP profits from mean-centering in the pre-processing step: applying mean-centering to WI matrices improves the performance by 3% on WI+SGNS+CD.

The results for Vector Initialization (VI) are unexpectedly low (on DURel mean $\rho = -0.017$, on SURel mean $\rho = 0.082$). An essential parameter choice for VI is the number of training epochs

[9]We see the same tendency for WI against random indexing with a shared random space (SRV), but instead variable results for count and PPMI alignment (CI). This contradicts the findings in Dubossarsky et al. (2019), using, however, a different task and synthetic data.

Dataset	OP	OP_-	OP_+	WI	None
DURel	0.618	0.557	**0.621**	0.468	0.254
SURel	**0.590**	0.514	0.401	0.492	0.285

Table 4: Mean ρ scores for CD across the alignments. Applies only to RI, SVD and SGNS.

for the initialized model. We experimented with 20 epochs instead of 5, but could not improve the performance. This contradicts the results obtained by Hamilton et al. (2016b) who report a "negligible" impact of VI when compared to OP_-. We reckon that VI is strongly influenced by frequency. That is, the more frequent a word is in corpus C_b, the more its vector will be updated after initialization on C_a. Hence, VI predicts more change with higher frequency in C_b.

Detection Measures. Cosine distance (CD) dominates Local Neighborhood Distance (LND) on all vector space and alignment types (e.g., mean ρ on DURel with SGNS+OP is 0.723 for CD vs. 0.620 for LND) and hence should be generally preferred if alignment is possible. Otherwise LND or a variant of WI+CD should be used, as they show lower but robust results.[10] Dispersion measures in general exhibit a low performance, and previous positive results for them could not be reproduced (Schlechtweg et al., 2017). It is striking that, contrary to our expectation, dispersion measures on SURel show a strong negative correlation (max. $\rho = -0.79$). We suggest that this is due to frequency particularities of the dataset: SURel's gold LSC rank has a rather strong negative correlation with the targets' frequency rank in the COOK corpus ($\rho = -0.51$). Moreover, because COOK is magnitudes smaller than SDEWAC the normalized values computed in most dispersion measures in COOK are much higher. This gives them also a much higher weight in the final calculation of the absolute differences. Hence, the negative correlation in COOK propagates to the final results. This is supported by the fact that the only measure not normalized by corpus size (HD) has a positive correlation. As these findings show, the dispersion measures are strongly influenced by frequency and very sensitive to different corpus sizes.

Control Condition. As we saw, dispersion measures are sensitive to frequency. Similar obser-

[10]JSD was not included here, as it was only applied to SCAN and its performance thus strongly depends on the underlying meaning representation.

vations have been made for other LSC measures (Dubossarsky et al., 2017). In order to test for this influence within our datasets we follow Dubossarsky et al. (2017) in adding a control condition to the experiments for which sentences are randomly shuffled across corpora (time periods). For each target word we merge all sentences from the two corpora C_a and C_b containing it, shuffle them, split them again into two sets while holding their frequencies from the original corpora approximately stable and merge them again with the original corpora. This reduces the target words' mean degree of LSC between C_a and C_b significantly. Accordingly, the mean degree of LSC predicted by the models should reduce significantly if the models measure LSC (and not some other controlled property of the dataset such as frequency). We find that the mean prediction on a result sample (L/P, win=2) indeed reduces from 0.5 to 0.36 on DURel and from 0.53 to 0.44 on SURel. Moreover, shuffling should reduce the correlation of individual model predictions with the gold rank, as many items in the gold rank have a high degree of LSC, supposedly being canceled out by the shuffling and hence randomizing the ranking. Testing this on a result sample (SGNS+OP+CD, L/P, win=2, k=1, t=None), as shown in Table 5, we find that it holds for DURel with a drop from $\rho = 0.816$ (ORG) to 0.180 on the shuffled (SHF) corpora, but not for SURel where the correlation remains stable (0.767 vs. 0.763). We hypothesize that the latter may be due to SURel's frequency properties and find that downsampling all target words to approximately the same frequency in both corpora ($\approx$ 50) reduces the correlation (+DWN). However, there is still a rather high correlation left (0.576). Presumably, other factors play a role: (i) Time-shuffling may not totally randomize the rankings because words with a high change still end up having slightly different meaning distributions in the two corpora than words with no change at all. Combined with the fact that the SURel rank is less uniformly distributed than DURel this may lead to a rough preservation of the SURel rank after shuffling. (ii) For words with a strong change the shuffling creates two equally polysemous sets of word uses from two monosemous sets. The models may be sensitive to the different variances in these sets, and hence predict stronger change for more polysemous sets of uses. Overall, our findings demonstrate that much more work has to be done to understand the effects of time-shuffling as well as sensitivity effects of LSC detection models to frequency and polysemy.

Dataset	ORG	SHF	+DWN
DURel	**0.816**	0.180	0.372
SURel	**0.767**	0.763	0.576

Table 5: ρ for SGNS+OP+CD (L/P, win=2, k=1, t=None) before (ORG) and after time-shuffling (SHF) and downampling them to the same frequency (+DWN).

7 Conclusion

We carried out the first systematic comparison of a wide range of LSC detection models on two datasets which were reliably annotated for sense divergences across corpora. The diachronic and synchronic evaluation tasks we introduced were solved with impressively high performance and robustness. We introduced *Word Injection* to overcome the need of (post-hoc) alignment, but find that Orthogonal Procrustes yields a better performance across vector space types.

The overall best performing approach on both data suggests to learn vector representations for different time periods (or domains) with SGNS, to align them with an orthogonal mapping, and to measure change with cosine distance. We further improved the performance of the best approach with the application of mean-centering as an important pre-processing step for rotational vector space alignment.

Acknowledgments

The first author 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. We thank Haim Dubossarsky, Simon Hengchen, Andres Karjus, Barbara McGillivray, Cennet Oguz, Sascha Schlechtweg, Nina Tahmasebi and the three anonymous reviewers for their valuable comments. We further thank Michael Dorna and Bingqing Wang for their helpful advice. We also thank Lea Frermann for providing the code of SCAN and helping to set up the implementation.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798, Melbourne, Australia.

Tyler Baldwin, Yunyao Li, Bogdan Alexe, and Ioana R. Stanoi. 2013. Automatic term ambiguity detection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 804–809, Sofia, Bulgaria.

David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pages 1–10, New York, NY, USA.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A Systematic Comparison of Context-counting and Context-predicting Semantic Vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, MD, USA.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015. Temporal random indexing: A system for analysing word meaning over time. *Italian Journal of Computational Linguistics*, 1:55–68.

John Bullinaria and Joseph Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and svd. *Behavior research methods*, 44:890–907.

Ping Chen and Hisham Al-Mubaid. 2006. Context-based term disambiguation in biomedical literature. In *Proceedings of the International Conference of the Florida Artificial Intelligence Research Society*, pages 62–67.

Paul Cook, Jey H. Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1624–1635, Dublin, Ireland.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta.

Béatrice Daille, Evelyne Jacquey, Gaël Lejeune, Luis Melo, and Yannick Toussaint. 2016. Ambiguity diagnosis for terms in digital humanities. In *Language Resources and Evaluation Conference*.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *Proceedings of the 12th International Conference on Computational Semantics*, Montpellier, France.

Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. 2016. Tracing metaphors in time through self-distance in vector spaces. In *Proceedings of the 3rd Italian Conference on Computational Linguistics*.

Gonzalo Donoso and David Sanchez. 2017. Dialectometric analysis of language variation in twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 16–25, Valencia, Spain.

Patrick Drouin. 2004. Detection of domain specific terminology using corpora comparison. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

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*, Florence, Italy.

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.

Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – A corpus of parsable sentences from the web. In Iryna

Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg.

Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. 2017. Detecting domain-specific ambiguities: An NLP approach based on wikipedia crawling and word embeddings. In *Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference Workshops*, pages 393–399.

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

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

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. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121.

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

Anna Hätty, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, Minneapolis, MN, USA.

Anna Hätty and Sabine Schulte im Walde. 2018. A laypeople study on terminology identification across domains and task definitions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 321–326.

Johannes Hellrich and Udo Hahn. 2016. Bad company–Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of the International Conference on Computational Linguistics 2016*, pages 2785–2796, Osaka, Japan.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of Digital Libraries Conference*.

William B. Johnson and Joram Lindenstrauss. 1984. Extensions to Lipshitz mapping into Hilbert space. *Contemporary mathematics*, 26.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proccedings of the ACL Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Max Kisselew, Laura Rimell, Alexis Palmer, and Sebastian Pado. 2016. Predicting the direction of derivation in English conversion. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Berlin, Germany.

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*, pages 625–635, Florence, Italy.

Jey H. Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Stroudsburg, PA, USA.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2177–2185, Montreal, Canada.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151.

Diana Maynard and Sophia Ananiadou. 1998. Term sense disambiguation using a domain-specific thesaurus. In *Proceedings of 1st International Conference on Language Resources and Evaluation*, pages 681–687, Granada, Spain.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 259–263, Jeju Island, Korea.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119. Curran Associates, Inc.

Sunny Mitra, Ritwik Mitra, Suman K. Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

Hermann Paul. 2002. *Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau unseres Wortschatzes*, 10. edition. Niemeyer, Tübingen.

María José Marín Pérez. 2016. Measuring the degree of specialisation of sub-technical legal terms through corpus comparison. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 22(1):80–102.

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, LA, USA.

Rahmtin Rotabi, Cristian Danescu-Niculescu-Mizil, and Jon Kleinberg. 2017. Competition and selection among conventions. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1361–1370.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Stroudsburg, PA, USA.

Magnus Sahlgren. 2004. An introduction to random indexing. *Language*, pages 1–9.

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

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 38–42.

Dominik Schlechtweg, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 354–367, Vancouver, Canada.

Dominik Schlechtweg and Sabine Schulte im Walde. 2018. Distribution-based prediction of the degree of grammaticalization for German prepositions. In *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*.

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, LA, USA.

Claude E. Shannon. 1948. *A Mathematical Theory of Communication*. CSLI Publications, Stanford, CA.

Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–323.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv preprint arXiv:1811.06278*.

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

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, New York, NY, USA.

Yue Wang, Irene L. Manotas Gutiérrez, Kristina Winbladh, and Hui Fang. 2013. Automatic detection of ambiguous terminology for software requirements. In *Proceedings of the International Conference on Application of Natural Language to Information Systems*, pages 25–37.

Derry T. Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, pages 35–40, New York, NY, USA.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Pasadena, CA, USA.

A Pre-processing and Hyperparameter Details

Corpora. For our experiments we used the TCF-version of DTA released September 1, 2017.[11] For all corpora, we removed words below a frequency threshold t. For the smallest corpus COOK we set $t = 2$, and set the other thresholds in the same proportion to the corpus size. This led to $t = 25, 37, 97$ for DTA18, DTA19 and SDEWAC respectively. (Note that we excluded three targets from the DURel dataset and one target from the SURel dataset because they were below the frequency threshold.) We then created two versions:

- a version with minimal pre-processing, i.e., with punctuation removed and lemmatization (L_{ALL})
- a stronger preprocessed version with only content words. After punctuation removal, lemmatization and POS-tagging, only nouns, verbs and adjectives were retained in the form *lemma:POS* (L/P)

Context window. For all models we experimented with values $n = \{2, 5, 10\}$ as done in Levy et al. (2015). It is important to note that the extraction of context words differed between models, because of inherent parameter settings of the implementations. While our implementations of the count-based vectors have a stable window of size n, SGNS has a dynamic context window with maximal size n (cf. Levy et al., 2015) and SCAN has as stable window of size n, but ignores all occurrences of a target word where the number of context words on either side is smaller than n. This may affect the comparability of the different models, as especially the mechanism of SCAN can lead to very sparse representations on corpora with small sentence sizes, as e.g. the COOK corpus. Hence, this variable should be controlled in future experiments.

Vector Spaces. We followed previous work in setting further hyper-parameters (Hamilton et al., 2016b; Levy et al., 2015). We set the number of dimensions d for SVD, RI and SGNS to 300. We trained all SGNS with 5 epochs. For PPMI we set $\alpha = .75$ and experimented with $k = \{1, 5\}$ for PPMI and SGNS. For RI and SGNS we experimented with $t = \{none, .001\}$. For SVD we set $p = 0$. In line with Basile et al. (2015) we set $s = 2$ for RI and SRV. Note though that we had a lower d than Basile et al., who set $d = 500$.

SCAN. We experimented with $K = \{4, 8\}$. For further parameters we followed the settings chosen by Frermann and Lapata (2016): $K^{\psi} = 10$ (a high value forcing senses to remain thematically consistent across time). We set K^{ϕ} = 4, and the Gamma parameters $a = 7$ and $b = 3$. We used $1,000$ iterations for the Gibbs sampler and set the minimum amount of contexts for a target word per time period $min = 0$ and the maximum amount to $max = 2000$.

Measures. For LND we set $k = 25$ as recommended by Hamilton et al. (2016a). The normalization constants for FD, HD and TD were calculated on the full corpus with the respective preprocessing (without deleting words below a frequency threshold).

B Model Overview

Find an overview of all tested combinations of semantic representations, alignments and measures in Table 6.

C Datasets

Find the datasets with the target words and their annotated degree of LSC in Tables 7 and 8.

[11] http://www.deutschestextarchiv.de/download

Semantic Representation	Alignment					Measure					
	CI	SRV	OP	VI	WI	CD	LND	JSD	FD	TD	HD
raw count	x				x	x	x			x	x
PPMI	x				x	x	x				
SVD			x		x	x	x				
RI		x	x		x	x	x				
SGNS			x	x	x	x	x				
SCAN								x			(x)

Table 6: Combinations of semantic representation, alignment types and measures. (FD has been computed directly from the corpus.)

lexeme	POS	LSC	freq. C_a	freq. C_b
Vorwort	NN	-1.58	85	273
Donnerwetter	NN	-1.84	100	89
Presse	NN	-1.88	193	1519
Feine	NN	-1.93	112	84
Anstalt	NN	-2.07	425	911
Feder	NN	-2.14	1489	3022
billig	ADJ	-2.43	2073	1705
Motiv	NN	-2.66	104	2551
Anstellung	NN	-2.68	53	499
packen	VV	-2.74	279	1057
locker	ADJ	-2.84	454	769
technisch	ADJ	-2.89	25	2177
geharnischt	ADJ	-3.0	56	117
Zufall	NN	-3.11	2444	1618
Bilanz	NN	-3.2	51	58
englisch	ADJ	-3.34	1921	7280
Reichstag	NN	-3.45	609	1781
Museum	NN	-3.73	414	1827
Abend	NN	-3.79	4144	4372

Table 7: DURel dataset without *flott*, *Kinderstube* and *Steckenpferd*, which were excluded for low frequency. C_a=DTA18, C_b=DTA19. LSC denotes the inverse compare rank from (Schlechtweg et al., 2018), where high values mean high change.

lexeme	POS	LSC	freq. C_a	freq. C_b
Schnee	NN	-1.05	2228	53
Strudel	NN	-1.05	232	46
schlagen	VV	-1.1	14693	309
Gericht	NN	-1.15	13263	1071
Schuß	NN	-1.42	2153	117
Hamburger	NN	-1.53	5558	46
abschrecken	VV	-1.75	730	170
Form	NN	-2.25	36639	851
trennen	VV	-2.65	5771	170
Glas	NN	-2.7	3830	863
Blech	NN	-2.95	409	145
Prise	NN	-3.1	370	622
Paprika	NN	-3.33	377	453
Mandel	NN	-3.45	402	274
Messer	NN	-3.5	1774	925
Rum	NN	-3.55	244	181
Salz	NN	-3.74	3087	5806
Eiweiß	NN	-3.75	1075	3037
Schokolade	NN	-3.98	947	251
Gemüse	NN	-4.0	2696	1224
Schnittlauch	NN	-4.0	156	247

Table 8: SURel dataset without *Messerspitze*, which was excluded for low frequency. C_a=SDEWAC, C_b=COOK. LSC denotes the inverse compare rank from (Schlechtweg et al., 2018), where high values mean high change.