IMS at SemEval-2020 Task 1:
How low can you go? Dimensionality in Lexical Semantic Change Detection

Jens Kaiser, Dominik Schlechtweg,[*] **Sean Papay, Sabine Schulte im Walde**
Institute for Natural Language Processing, University of Stuttgart
{jens.kaiser,schlecdk,sean.papay,schulte}@ims.uni-stuttgart.de

Abstract

We present the results of our system for SemEval-2020 Task 1 that exploits a commonly used lexical semantic change detection model based on Skip-Gram with Negative Sampling. Our system focuses on Vector Initialization (VI) alignment, compares VI to the currently top-ranking models for Subtask 2 and demonstrates that these can be outperformed if we optimize VI dimensionality. We demonstrate that differences in performance can largely be attributed to model-specific sources of noise, and we reveal a strong relationship between dimensionality and frequency-induced noise in VI alignment. Our results suggest that lexical semantic change models integrating vector space alignment should pay more attention to the role of the dimensionality parameter.

1 Introduction

Lexical Semantic Change (LSC) Detection has drawn increasing attention in recent years (Tahmasebi et al., 2018; Kutuzov et al., 2018). SemEval-2020 Task 1 provides a multi-lingual evaluation framework to compare the variety of proposed model architectures (Schlechtweg et al., 2020). An important component of high-performance LSC detection models is an alignment method to make semantic vector spaces comparable across time. In this paper we focus on a particular alignment method for type embeddings, Vector Initialization (VI), and how its performance interacts with vector dimensionality. We compare VI to two further state-of-the-art alignment methods, Orthogonal Procrustes (OP) and Word Injection (WI), which have shown high performance in previous studies (Hamilton et al., 2016b; Schlechtweg et al., 2019; Dubossarsky et al., 2019) and are also used in the top-ranking systems for Subtask 2. A systematic comparison of performance across dimensionalities d reveals that the optimal d of the models on the SemEval test data is lower than in standard choices, and that VI's performance strongly depends on d, showing large drops for high dimensionalities. We demonstrate that this effect is correlated with the amount of frequency noise picked up by VI, i.e., the degree to which cosine distances between vectors reflect frequency differences between words rather than semantic differences. If properly tuned regarding dimensionalities and noise, VI outperforms OP and WI as alignment method.

2 Related Work

The semantic representations we test fall into the large body of work on distributional semantic vector space models (Turney and Pantel, 2010) and represent specific instances of type-based word embeddings (Mikolov et al., 2013a). The need for vector space alignment in LSC detection is shared with bilingual lexicon induction (Ruder et al., 2019) and term extraction (Hätty et al., 2020) where corpus-specific semantic representations need to be mapped to common coordinate axes.

Alignment techniques introduce varying levels of noise (Dubossarsky et al., 2019), and the noise level (the signal-to-noise-ratio) determines the optimal dimensionality of word embeddings (Yin and Shen, 2018). We regard information contained within the semantic representation capturing anything but semantic relations between words as noise (e.g. word frequency). Sources of noise include the corpora, the

---

representation method and alignment techniques. Consequently, a specific semantic representation learning algorithm (such as Skip-Gram with Negative Sampling) may have a different optimal dimensionality depending on the alignment technique it relies on. Up to now, previous research on LSC detection has not paid much attention to this relationship between dimensionality and noise: models with different susceptibilities to noise have typically been tested without varying the dimensionality (Hamilton et al., 2016b; Dubossarsky et al., 2019; Schlechtweg et al., 2019; Shoemark et al., 2019).

3   System overview

Most models in LSC detection 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). Semantic representations can either be token-based, keeping one representation (e.g. a vector) per word use (Hu et al., 2019, e.g.), or type-based, collapsing information from different uses into one representation (Hamilton et al., 2016b, e.g.). The alignment step is needed mostly for vector space models, which might otherwise introduce arbitrary orthogonal transformations to the vector spaces they produce (Hamilton et al., 2016b).

Our system focuses on the type-based Skip-gram Negative Sampling (SGNS) with Vector Initialisation alignment (VI) and Cosine Distance (CD). We chose this method due to its surprisingly good performance with $d = 5$ in a student shared-task project (Ahmad et al., 2020)[1] and compare it to two variations of the alignment method, (i) Orthogonal Procrustes (OP) and (ii) Word Injection (WI). For our experiments we use the code provided by Schlechtweg et al. (2019).[2]

3.1   Semantic Representation

SGNS is a shallow neural network trained on pairs of word co-occurrences extracted from a corpus with a symmetric window. It represents each word w and each context c as a d-dimensional vector to solve

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w), \tag{1}$$

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; Mikolov et al., 2013b; Goldberg and Levy, 2014). The optimized parameters θ are v_{w_i} and v_{c_i} for $i \in 1, ..., d$. 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). After training, each word w is represented by its word vector v_w. To keep our results comparable to previous research (Hamilton et al., 2016b; Schlechtweg et al., 2019) we chose common settings for most of the hyper-parameters. We decided on a symmetrical context window of size 10, initial learning rate α of 0.025, number of negative samples $k = 5$ and no sub-sampling. Depending on corpus size we trained the model for either 5 (German, Swedish) or 30 epochs e (English, Latin).[3] As we focus on the effect of dimensionality, each experiment was performed for each $d \in \{5, 10, 25, 50, 80, 150, 200, 250, 300, 350, 500, 750, 1000\}$. Prior to the shared task application we validated all models with these hyper-parameters on the German DURel dataset (Schlechtweg et al., 2018).

3.2   Alignment

Vector Initialisation.   In VI we first train the SGNS model on one corpus and then use these vectors to initialize the vectors for training on the second corpus (Kim et al., 2014). The motivation of this procedure is that if a word is used in similar contexts in both corpora, the second training step will not change the initial word vector much, while more different contexts will lead to a greater change of the vector. SGNS represents each word by two vectors, a word vector and a context vector. The former is modified when a word occurs as target w in a target-context pair (w, c), while the latter is modified when it occurs as context c. While Schlechtweg et al. (2019) only initialize the word vectors on the first model and context

---

[1]These referenced results were achieved by us after experimenting with extreme and unusual parameter choices.
[2]https://github.com/Garrafao/LSCDetection
[3]We tried alternative numbers of epochs with mixed results.

vectors randomly, we also initialize context vectors on the first model, as done by Ahmad et al. (2020). In this way, we expect to introduce considerably less noise to the vectors in the second corpus.

Orthogonal Procrustes. SGNS is trained on each corpus separately, resulting in matrices A and B. To align them we follow Hamilton et al. (2016b) and calculate an orthogonally-constrained matrix W^*:

$$W^* = \underset{W \in O(d)}{\arg\min} \|BW - A\|_F \tag{2}$$

where the i-th row in matrices A and B correspond to the same word. Using W^* we get the aligned matrices $A^{OP} = A$ and $B^{OP} = BW^*$. Prior to this alignment step we length-normalize and mean-center both matrices (Artetxe et al., 2017; Schlechtweg et al., 2019).

Word Injection. The sentences of both corpora are shuffled into one joint corpus, but all occurrences of target words are substituted by the target word concatenated with a tag indicating the corpus it originated from (Ferrari et al., 2017; Schlechtweg et al., 2019). This leads to the creation of two vectors for each target word in one vector space, while non-target words receive only one vector encoding information from both corpora. This is very similar to Temporal Referencing (TR) (Dubossarsky et al., 2019), the difference being that with TR, the target-context pairs used for training never contain tagged target words as contexts but rather the genuine (untagged) words.

3.3 Measures

To quantify semantic change on the aligned vector representations, we use two vector similarity measures. For Subtask 1 we apply Local Neighborhood Distance (LND), as it showed superior performance to CD for binary change detection in Schlechtweg and Schulte im Walde (2020). LND is based on second-order cosine similarity and measures to which extent $\vec{x}$ and $\vec{y}$'s distances to a union of their k nearest neighbors differ (Hamilton et al., 2016a). Similar to Hamilton et al. (2016a) we chose $k = 25$. We split the LND scores into two equally sized groups; the group containing the high values was labelled as 1 (change). For Subtask 2 we use a simple cosine distance (Salton and McGill, 1983).

4 Experimental setup

SemEval-2020 Task 1 comprises a binary classification task (Subtask 1) and a ranking task (Subtask 2) on data from four languages: English, German, Latin and Swedish (Schlechtweg et al., 2020). Subtask 1 asks participants to decide which target words lost or gained senses between corpora from two time periods t_1 and t_2, and which ones did not. Subtask 2 asks participants to rank a set of target words according to their degree of LSC (change in sense frequency distribution) between t_1 and t_2. The tasks are different in that it is possible for a word to show a high degree of LSC in Subtask 2, while not gaining or losing a sense in Subtask 1 (or vice versa). For example the German word *abgebrüht* is used to describe (1) the process of cooking food in water and (2) an emotionally insensitive person. Both senses are present across time periods t_1 and t_2, but sense 1 dominates period t_1 while sense 2 dominates t_2. Thus, the number of senses has not changed, but the word has undergone significant semantic change.

The four languages have a list of 31 to 48 target words, each annotated with values for Subtasks 1 and 2. Performance of a model is measured by accuracy and Spearman's rank-order correlation coefficient. For each of the four languages two corpora (corpus$_1$, corpus$_2$) are provided by the organizers, containing sentences from different time periods. These corpora show strong differences in terms of size, time-period and genre (see Appendix A), posing a very heterogeneous, challenging setting for evaluation and parameter tuning.[4]

5 Results

Table 1 lists the evaluation phase scores of the top three contenders for both subtasks as well as our system. During this phase, submission scores and leaderboards were hidden. At the end of the evaluation

---

[4]For pre-processing details see Appendix B.

Phase	Team	Model	Subtask 1 (accuracy)					Subtask 2 (spearman ρ)				
			AVG	English	German	Latin	Swedish	AVG	English	German	Latin	Swedish
Evaluation	1. UWB	SGNS+CA	**.69**	.62	.75	.70	.68					
	2. Life-Lang.	fastText	.69	.70	.75	.55	.74					
	3. Jia. & Jin.	SGNS+WI	.67	.65	.73	.70	.58					
	1. UG Stud.	SGNS+OP						**.53**	.42	.73	.41	.55
	2. Jia. & Jin.	SGNS+WI						.52	.33	.72	.44	.59
	3. cs2020	SGNS+OP						.50	.38	.70	.40	.54
	13./8. IMS *d / e*	SGNS+VI	.60	.54 *5 / 30*	.69 *5 / 30*	.55 *5 / 30*	.61 *5 / 30*	.37	.30 *5 / 50*	.66 *5 / 50*	.10 *5 / 50*	.43 *5 / 50*
Post-Evaluation	IMS *dim / epoch*	SGNS+VI	**.73**	.75 *10 / 30*	.74 *50 / 5*	.67 *80 / 30*	.74 *300 / 10*	**.58**	**.46** *10 / 30*	**.78** *50 / 5*	.39 *80 / 30*	**.67** *300 / 10*
		SGNS+OP	.65	.62 *350 / 30*	.80 *300 / 5*	.50 *10 / 30*	.68 *500 / 5*	.56	.44 *350 / 30*	.73 *300 / 5*	**.41** *10 / 30*	.64 *500 /5*
		SGNS+WI	.67	.62 *10 / 30*	.71 *50 / 5*	.65 *10 / 30*	.68 *500 / 5*	.54	.36 *10 / 30*	.76 *50 / 5*	.41 *10 / 30*	.61 *500 / 5*

Table 1: Comparing our model to the top 3 of both subtasks in evaluation phase. Our results are annotated with dimensionality d and training epochs e at each entry. CA = Canonical Correlation Analysis.

phase the best submission (out of a maximum of ten) was put on the leaderboard for both subtasks. We only submitted results for VI with d/e of 5/30, 3/50 and 8/30. The very low choice for d is motivated by the results found in Ahmad et al. (2020), where VI has high performance with $d = 5$. With this exceptionally low d we scored 13th in Subtask 1 and 8th in Subtask 2 out of 33 teams. Our methodology for Subtask 1 has much room for improvement but we decided to point our attention towards Subtask 2 during post-evaluation. The three best teams for Subtask 2 used models based on OP and TR/WI. Their average scores are very similar and ahead of ours. The models seem to consistently get the best performances on German, then Swedish, followed by English and Latin. With this limited picture VI alignment seems to be inferior to OP and WI. But after tuning each model to its optimal d, performances are barely distinguishable (see Table 1, Post-Evaluation). This is most prominent for Latin. We attribute the low score (0.10) during the Evaluation phase to the Latin data-set being very challenging due to its size and heterogeneity in combination with high discrepancies between the number of vector updates in the second training step (see below). Performance significantly improved after switching training order (see Appendix C).

5.1 Analysis

We now tested the performance of VI, OP and WI on the four languages with varying dimensionality. Training epochs were adjusted to compensate differences in corpus size: German and Swedish were trained for 5 epochs, Latin and English for 30 epochs. We realized that VI is very sensitive to training order (see Appendix C): this was very prominent with Swedish and Latin, as corpus$_1$ and corpus$_2$ show big differences in size. To get comparable results, we switched training order for these two languages, which means instead of first training on corpus$_1$ and then corpus$_2$, we train on corpus$_2$ and then corpus$_1$.

As shown in Figure 1 all models were able to achieve high correlations for German and Swedish (0.6–0.7). On English and Latin correlation was lower (0.3–0.4), which is probably related to corpus size, and performances are barely distinguishable, see Figure 1 bottom. For German, VI and WI show a clear global peak in performance at $d = 50$, while for other languages we find several local peaks. In German, highest performance is obtained by VI, closely followed by WI, while in Swedish OP is best. In all languages performance of OP is very consistent across different $d > 25$, and it is the most robust model in high d. WI also shows robust performance in high d, but is mostly outperformed by OP. In all languages, we see a steep drop-off in performance of VI with higher d.

What is each model's optimal dimensionality? For VI and WI the optimal d in our models is much lower than the common choice of 200–300 (Hamilton et al., 2016b; Dubossarsky et al., 2019; Schlechtweg et al., 2019; Shoemark et al., 2019).[5] In Swedish, which has the largest training corpus, optimal ds are higher, suggesting a possible correspondence between corpus size and optimal d.

Across all corpora, OP tends to have higher optimal values than VI and WI. This may be because orthogonal alignment works better in high dimensions, as there are more degrees of freedom to rotate the
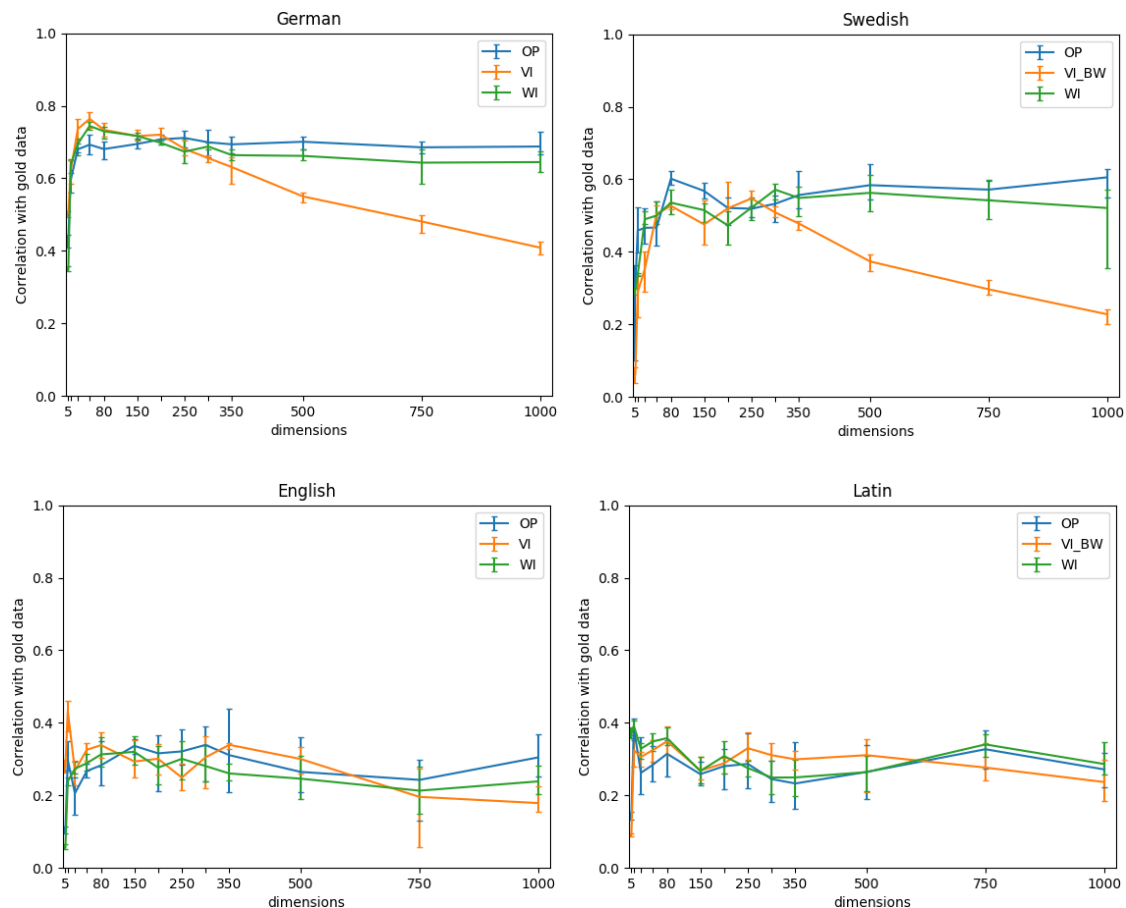
---

[5]Find an overview in Table 1 Post-Evaluation

Figure 1: Comparing all models with varying dimensionality. The line refers to the mean of 5 runs; error bars show the max/min values.'_BW' indicates switched training order.

vectors. In Swedish, all methods show two local maxima instead of a global one. This behavior becomes more pronounced with increasing numbers of training epochs (see Figure 2). This could be explained by the two Swedish corpora having different optimal ds (due to their different sizes and homogeneity).

Why does VI's performance drop in high d? We tested several hypotheses to explain the drop. We first speculated that the embeddings for corpus$_2$ generally drift away from their initial state, as seen in (Kim et al., 2014) where even the word with the least amount of measured semantic change had cosine distance scores of almost 0.1. This drift could be repaired by post-hoc alignment. However, additional OP alignment did not eliminate the drop. We then tested the hypothesis that the number of training updates influences cosine distances (Schlechtweg et al., 2019). Hence, we calculated the correlation between the predicted ranking of target words (according to cosine distance) and their frequency (reflecting the number of training updates) in the second corpus, and compared this correlation across d (see Figure 2, top). There is a clear frequency bias for VI becoming stronger with increasing d. Thus, the predicted cosine distances do not reflect LSC but rather frequency, leading to poor performance. The bias correlates negatively with performance as comparing the top to the bottom figures. Interestingly, the number of training epochs also has a strong effect on this bias: An increasing e reduces the bias for low d and consequently drastically improves performance (see Figure 2, cf. top and bottom). We were not able to find an explanation for the cause of this behaviour.

The extend of the frequency bias is determined by several parameters, the main one being word frequency in the second corpus. We experimented with modified corpora wherein the frequency of all target words was fixed to 200, and less frequent target words were ignored. This completely removed the frequency bias. However, if frequency differences amongst target words exist, as is the case with most data sets, the noise induced by those differences may be exaggerated by dimensionality or reduced by
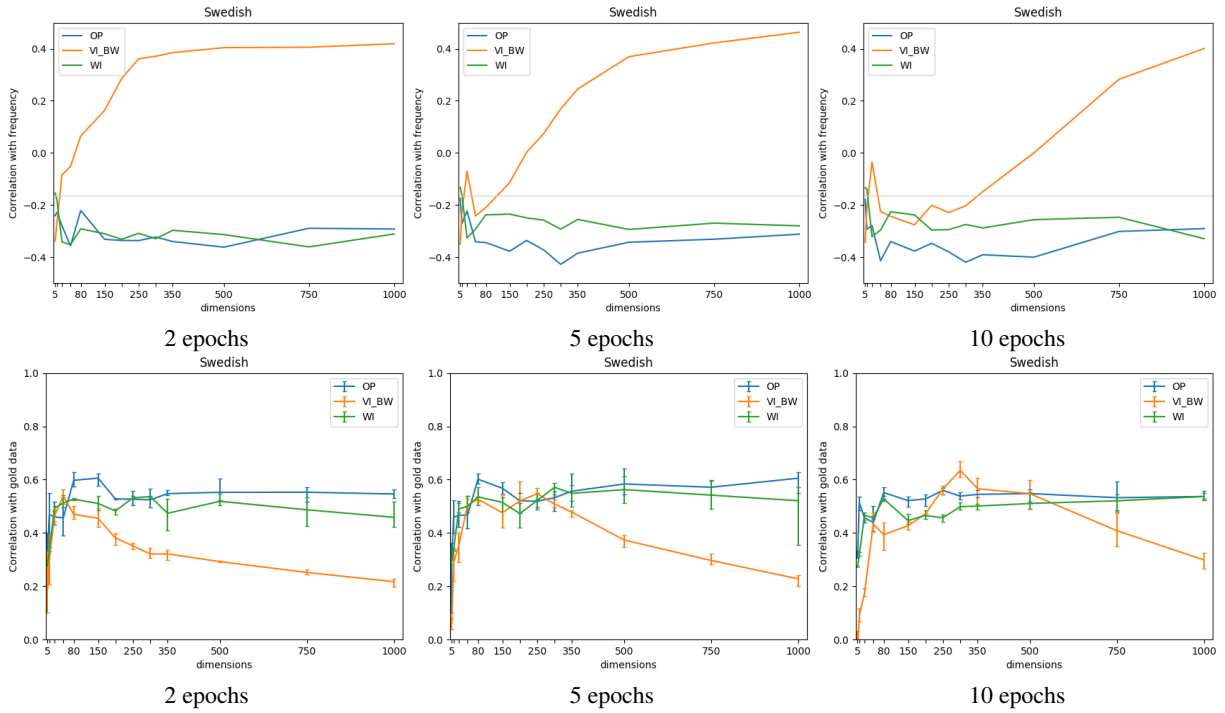
Figure 2: **Top** figures show correlations between CD and frequency in Swedish corpus$_2$ across d, with increasing numbers of epochs. Gray lines indicate true correlation in gold data. **Bottom** figures show performance on Swedish data for Subtask 2.

number of training epochs. Understanding and explaining the frequency bias is outside of the scope of this work, but will be part of future work.

6 Conclusion

Our shared task system investigated Vector Initialization (VI) alignment in a commonly used LSC detection model based on Skip-Gram with Negative Sampling, while focusing on the role of vector-space dimensionality. Our results suggest that LSC detection models integrating vector-space alignment should pay more attention to model-specific characteristics and the dimensionality parameter in particular. Current state-of-the-art models are often dominated by applying OP and WI to alignment, as a wide variety of reasonable parameters yield good results, whereas VI is very susceptible to parameters like training order, dimensionality and epochs. However we demonstrate that VI is able to outperform OP and WI alignment if tuned properly. Due to time limitations we could not fully explore the effects of epochs on VI, which have proven to play a significant role for dimensionality-dependent performance. Future work will include a closer look at the connection between dimensionality, frequency noise and training epochs.

Acknowledgments

References

Adnan Ahmad, Kiflom Desta, Fabian Lang, and Dominik Schlechtweg. 2020. Shared task: Lexical semantic change detection in German (Student Project Report). *arXiv:2001.07786*.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, Marseille, France. European Language Resources Association.

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. Association for Computational Linguistics.

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

Mark Davies. 2012. Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English. *Corpora*, 7(2):121–157.

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

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.

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.

Yoav Goldberg and Omer Levy. 2014. Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722*.

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, Austin, Texas. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Anna Hätty, Dominik Schlechtweg, Michael Dorna, and Sabine Schulte im Walde. 2020. Predicting degrees of technicality in automatic terminology extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

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.

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.

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.

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.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, Tübingen. Narr.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *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 *Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe, Nevada, USA.

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

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

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

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. In A. Ravignani, C. Barbieri, M. Martins, M. Flaherty, Y. Jadoul, E. Lattenkamp, H. Little, K. Mudd, and T. Verhoef, editors, *The Evolution of Language: Proceedings of the 13th International Conference*, Brussels, Belgium.

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

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.

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*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Språkbanken, Downloaded in 2019. *The Kubhist Corpus*. Department of Swedish, University of Gothenburg.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *arXiv:1811.06278*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 887–898. Curran Associates, Inc.

A   Details on Corpora

Table 2 shows basic corpus statistics.

	t_1	t_2	tokens$_1$	tokens$_2$	types$_1$	types$_2$	TTR$_1$	TTR$_2$
English	CCOHA 1810–1860	CCOHA 1960–2010	6.5M	6.7M	87k	150k	13.38	22.38
German	DTA 1800–1899	BZ+ND 1946–1990	70.2M	72.3M	1.0M	2.3M	14.25	31.81
Latin	LatinISE -200–0	LatinISE 0–2000	1.7M	9.4M	65k	253k	38.24	26.91
Swedish	Kubhist 1790–1830	Kubhist 1895–1903	71.0M	110.0M	3.4M	1.9M	47.88	17.27

Table 2: Corpus statistics. The SemEval corpora are samples from CCOHA (Davies, 2012; Alatrash et al., 2020), DTA (Deutsches Textarchiv, 2017), BZ (Berliner Zeitung, 2018), ND (Neues Deutschland, 2018), LatinISE (McGillivray and Kilgarriff, 2013) and KubHist (Språkbanken, Downloaded in 2019). TTR = Type-Token ratio (number of types / number of tokes * 1000).

B   Pre-Processing Details

Table 3 shows our thresholds for infrequent words that we removed from the corpora.

	corpus$_1$				**corpus$_2$**			
	English	German	Latin	Swedish	English	German	Latin	Swedish
threshold	4	39	1	42	4	39	6	65

Table 3: Corpus-specific frequency thresholds.

C   Influence of training order

We switched the training order for Swedish and Latin due to their size differences across corpora. The switch led to noticeable increases in performance for both languages, see Figure 3.
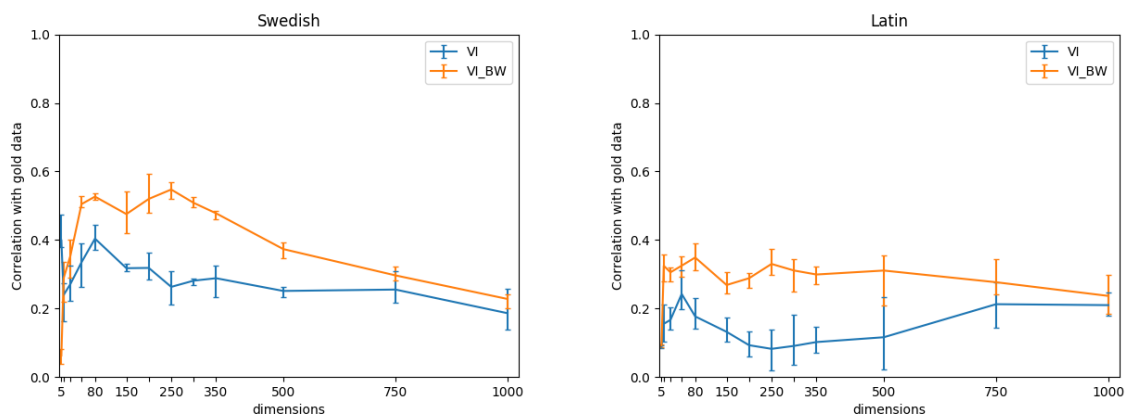

Figure 3: Comparing performances on Swedish and Latin when changing the training order. VI: normal order, VI_BW: backwards order