Bad Company—
Neighborhoods in Neural Embedding Spaces Considered Harmful

Johannes Hellrich
Graduate School "The Romantic Model. Variation—Scope—Relevance"
Friedrich Schiller University Jena
Jena, Germany
johannes.hellrich@uni-jena.de

Udo Hahn
Jena University Language & Information Engineering (JULIE) Lab
Friedrich Schiller University Jena
Jena, Germany
http://www.julielab.de

Abstract

We assess the reliability and accuracy of (neural) word embeddings for both modern and historical English and German. Our research provides deeper insights into the empirically justified choice of optimal training methods and parameters. The overall low reliability we observe, nevertheless, casts doubt on the suitability of word neighborhoods in embedding spaces as a basis for qualitative conclusions on synchronic and diachronic lexico-semantic matters, an issue currently high up in the agenda of Digital Humanities.

1 Introduction

Distributional methods applied to large-sized, often temporally stratified corpora have markedly enhanced the methodological repertoire of both synchronic and diachronic computational linguistics and are getting more and more popular in the Digital Humanities (see Section 2.2). However, using such quantitative data as a basis for qualitative, empirically-grounded theories requires that measurements should not only be accurate, but also reliable. Only under such a guarantee, quantitative data can be assembled from different experiments as a foundation for trustful theories.

Measuring word similarity by word neighborhoods in embedding space can be used to detect diachronic shifts or domain specific usage, by training word embeddings on suited corpora and comparing these representations. Additionally, lexical items near in the embedding space to the lexical item under scrutiny can be considered as approximating its meaning at a given point in time or in a specific domain. These two lines of research converge in prior work to show, e.g., the increasing association of the lexical item *'gay'* with the meaning dimension of homosexuality (Kim et al., 2014; Kulkarni et al., 2015). Neural word embeddings (Mikolov et al., 2013) are probably the most influential among all embedding types (see Section 2.1). Yet, we gathered evidence that the inherent randomness involved in their generation affects the reliability of word neighborhood judgments and demonstrate how this hampers qualitative conclusions based on such models.

Our investigation was performed on both historical (for the time span of 1900 to 1904) and contemporary texts (for the time span of 2005 to 2009) in two languages, English and German. It is thus a continuation of prior work, in which we investigated historical English texts only (Hellrich and Hahn, 2016a), and also influenced by the design decisions of Kim et al. (2014) and Kulkarni et al. (2015) which were the first to use word embeddings in diachronic studies. Our results cast doubt on the reproducibility of such experiments where neighborhoods between words in embedding space are taken as a computationally valid indicator for properly capturing lexical meaning (and, consequently, meaning shifts).

2 Related Work

2.1 Word Embeddings

Word embeddings, i.e., low (several hundred) dimensional vector word representations encoding both semantic and syntactic information, are currently one of the most influential methods in computational

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2785–2796, Osaka, Japan, December 11-17 2016.

linguistics. The word2vec family of algorithms, developed from heavily trimmed artificial neural networks, is a widely used and robust way to generate such embeddings (Mikolov et al., 2013; Levy et al., 2015). Its skip-gram variant predicts plausible contexts for a given word, whereas the alternative continuous bag-of-words variant tries to predict words from contexts; we focus on the former as it is generally reported to be superior (see e.g., Levy et al. (2015)). There are two strategies for managing the huge number of potential contexts a word can appear in. Skip-gram hierarchical softmax (SGHS) uses a binary tree to more efficiently represent the vocabulary, whereas skip-gram negative sampling (SGNS) updates only a limited number of word vectors during each training step. SGNS is preferred in general, yet SGHS showed slight benefits in some reliability scenarios in our prior investigations (Hellrich and Hahn, 2016a).

There are two sources of randomness involved in the training of neural word embeddings: First, the random initialization of all word vectors before any examples are processed. Second, the order in which these examples are processed. Both can be replaced by deterministic alternatives,[1] yet this would simply replace a random distortion with a fixed one, thus providing faux reliability only useful for testing purposes. A range of other word embedding algorithms was inspired by word2vec, either trying to avoid the opaqueness stemming from its neural network heritage (GloVe; still using random initialization, see Pennington et al. (2014)) or adding capabilities, like using syntactic information during training (Levy and Goldberg, 2014) or modeling multiple word senses (Bartunov et al., 2016; Panchenko, 2016). Levy et al. (2015) created SVD_{PPMI}, a variant of the classical pointwise mutual information co-occurrence metric (see e.g., Manning and Schütze (1999, pp.178–183)), by transferring pre-processing steps and hyper-parameters uncovered by the development of these algorithms, and reported similar or slightly better performance than SGNS on evaluation tasks. It is conceptually not affected by reliability problems, as there is no random initialization or relevant processing order.

Word embeddings capture both syntactic and semantic information (and arguably also social biases, see Bolukbasi et al. (2016)) in vector form and can thus be evaluated by their ability to calculate the similarity of two words and perform analogy-based reasoning; there exist several other evaluation methods and more test sets than discussed here, see e.g., Baroni et al. (2014). Mikolov et al. (2013) provide an analogy test set for measuring performance as the percentage of correctly calculated analogies for test cases such as the frequently cited *'king'–'queen'* example (see Section 3). Word similarity is evaluated by calculating Spearman's rank coefficient between embedding-derived predictions and a gold standard of human word similarity judgments. Finkelstein et al. (2002) developed a widely used test set with 353 English word pairs,[2] a similar resource for German with 350 word pairs was provided by Zesch and Gurevych (2006).[3] Recent work cautions that performance on such tasks is not always predictive for performance in down-stream applications (Batchkarov et al., 2016).

2.2 Diachronic Application

Word embeddings can be used rather directly for tracking semantic changes, namely by measuring the similarity of word representations generated for one word at different points in time—words which underwent semantic shifts will be dissimilar with themselves. These models must either be trained in a continuous manner where the model for each time span is initialized with its predecessor (Kim et al., 2014; Hellrich and Hahn, 2016b), or a mapping between models for different points in time must be calculated (Kulkarni et al., 2015; Hamilton et al., 2016). The first approach cannot be performed in parallel and is thus rather time-consuming, if texts are not subsampled. We nevertheless discourage using samples instead of full corpora, as we observed extremely low reliability values between different samples (Hellrich and Hahn, 2016a). Word embeddings can also be used in diachronic studies without any kind of mapping to track clusters of similar words over time and, thus, model the evolution of topics (Kenter et al., 2015) or compare neighborhoods in embedding space for preselected words (Jo, 2016). Besides temporal variations, word embeddings can also used to analyze geographic ones, e.g., the distinction between US American and British English variants (Kulkarni et al., 2016). Most of these studies were

[1] In fact, in some implementations, yet not in ours, vectors are initialized via a deterministic process.

[2] www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

[3] www.ukp.tu-darmstadt.de/data/semantic-relatedness/german-relatedness-datasets/

performed with algorithms from the `word2vec` family, respectively `GloVe` in Jo (2016), and are thus likely to be affected by the same systematic reliability problems on which we focus here. Only Hamilton et al. (2016) used SVD_{PPMI} in some of their very recent experiments and showed it to be adequate for exploring historical semantics.

The Google Books Ngram corpus (GBN; Michel et al. (2011), Lin et al. (2012)) is used in most of the studies we already mentioned, including our current study and its predecessor (Hellrich and Hahn, 2016a). It contains about 6% of all books published between 1500 and 2009 in the form of n-grams (up to pentagrams), together with their frequency for each year. This corpus has often been criticized for its opaque sampling strategy, as its constituent books remain unknown and can be shown to form an unbalanced collection (Pechenick et al., 2015). GBN is multilingual, with its English part being subdivided into regional segments (British, US) and topic categories (general language and fiction texts). Diachronic research focuses on the English Fiction part, with the exception of some work relating to German data (Hellrich and Hahn, 2016b).

3 Evaluation Methods

Reliability, in this study, is judged by training three identically parametrized models for each experiment and by comparing the n next neighbors (by cosine distance) for each word modeled by the experiments with a variant of the Jaccard coefficient (Manning and Schütze, 1999, p.299). The 3-dimensional array $W_{i,j,k}$ contains words ordered by closeness (i) for a word in question (j) according to an experiment (k). The reliability r for a specific value of n ($r@n$) is defined as the magnitude of the intersection of similar words produced by all three experiments with a rank of n or lower, averaged over all t words modeled by these experiments and normalized by n, which is the maximally achievable score for this value of n:

$$r@n := \frac{1}{t * n} \sum_{j=1}^{t} || \bigcap_{k=1}^{3} \{W_{1 \leq i \leq n,j,k}\} || \tag{1}$$

Accuracy, in this study, is measured considering two different approaches—analogy and similarity. The *analogy* approach uses the English test set developed by Mikolov et al. (2013) by calculating the percentage of correct analogies made by a `word2vec` model. It contains groups of four words connected via the analogy relation '::' and the similarity relation '$\sim$', as exemplified by the expression *'king'* $\sim$ *'queen'* :: *'man'* $\sim$ *'woman'*. The *similarity* approach covers both English and German by calculating Spearman's rank correlation coefficient between the similarity judgments made by a `word2vec` model for a word pair (e.g., *'bread'* and *'butter'*) and the human judgment thereof (Finkelstein et al., 2002; Zesch and Gurevych, 2006). Pairs containing words not modeled for the time span in question, such as the at that time non-existent *'FBI'* in the early 20th century, are simply ignored. All three test sets are based on contemporary language and current world knowledge and might thus not fully match the requirements for historical texts, yet are also used for these due to the lack of a suitable alternative. Accuracy values were calculated independently for each of the three identically parametrized models and subsequently averaged, but resulting deviations were negligible.

4 Experimental Set-up

4.1 Corpus

Our experiments[4] were performed on the German part and the English Fiction part of the GBN; the latter is known to be less unbalanced than the general English part (Pechenick et al., 2015). Both corpus splits differ in size and contain mainly contemporary texts (from the past fifty years), as is evident from Figure 1; note the logarithmic axis and the negative impact of both World Wars on book production. Following Kulkarni et al. (2015), we trained our models on all 5-grams occurring during five consecutive years for the two time spans,[5] 1900–1904 and 2005–2009; the number of 5-grams[6] for each time span

[4]Code used in experiments available from `https://github.com/hellrich/coling2016`

[5]This is due to computational demands, e.g., using 8 parallel processes on a server with Intel Xeon E5649@2.53Ghz processors five days were necessary to complete each of ten training epochs for SGNS with 2005–2009 English Fiction data.

[6]Note that we treat 5-grams with k occurrences during the same time span as k different 5-grams.

is listed in Table 1. The two languages share a similar number of 5-grams for 1900–1904, yet not for 2005–2009. 5-grams from both corpus parts were lower cased for training. The German part was not only taken as is, but also orthographically normalized using the CAB service (Jurish, 2013).[7] We incorporated this step because major changes in German orthography occurred during the 20th century, an issue that could hamper diachronic comparisons, e.g., archaic *'Gemüth'* (in English: "mind, emotional disposition") became modern *'Gemüt'*. Table 1 shows the resulting reduction in the number of types, bringing the morphologically richer German to levels below English (yet this reduction is in line with the respective corpus sizes).

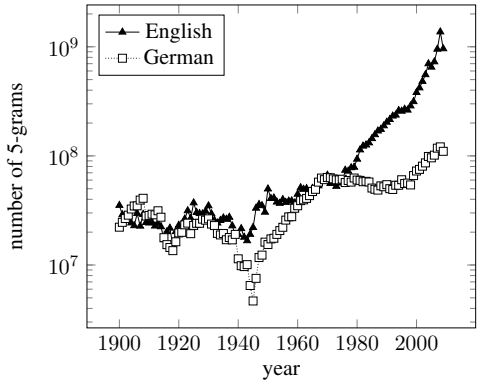

Figure 1: Number of 5-grams per year (on the logarithmic y-axis) in the English Fiction part and the German part of the GOOGLE BOOKS NGRAM corpus.

Language	Time Span	5-grams	Types
English	1900–1904	143M	80k
English	2005–2009	4,658M	216k
German	1900–1904	135M	111k
Normalized German			72k
German	2005–2009	546M	243k
Normalized German			179k

Table 1: Number of 5-grams and lemma types contained in the English Fiction part and the German part of the GOOGLE BOOKS NGRAM corpus for the two time spans used in our experiments.

4.2 Training

We used the PYTHON-based GENSIM[8] implementation of `word2vec` to independently train word embeddings for each time span with 200 dimensions, a context window of 4 (limited by the 5-gram size), a minimum frequency of 10, and 10^{-5} as the threshold for downsampling frequent words. We processed the full subcorpora for each time span, due to the extremely low reliability values between samples we observed in previous investigations (Hellrich and Hahn, 2016a). We tested both SGNS with 5 noise words and SGHS training strategies and trained for 10 iterations, saving the resulting embeddings after each epoch. During each epoch the learning rate was decreased from 0.025 to 0.0001. The averaged cosine values between word embeddings before and after an epoch are used as a convergence measure c (Kim et al., 2014; Kulkarni et al., 2015). It is defined for a vocabulary with n words and a matrix W containing word embedding vectors (normalized to length 1) for words i from training epochs e and e-1:

$$c := \frac{1}{n} \sum_{i=1}^{n} W_{i,e} \cdot W_{i,e-1} \tag{2}$$

We also define Δc, the change of c during subsequent epochs e-1, as another convergence criterion:

$$\Delta c := c_e - c_{e-1} \tag{3}$$

5 Results

Table 2 shows the performance of the systems trained according to the settings described in Section 4.2, as measured by similarity accuracy and top-1 reliability (see below for other cut-offs). We make the following observations:

[7] `www.deutschestextarchiv.de/demo/cab/`
[8] `www.radimrehurek.com/gensim/`

1. Both accuracy and reliability are higher for SGNS than for SGHS for all tested combinations of languages and time spans, if 10 training epochs are used.
2. If only one training epoch is used—as in many other experimental set-ups reported in the literature—there is only little difference in accuracy between SGNS and SGHS, but SGHS is clearly better in terms of reliability.
3. Accuracy is higher for 2005–2009 than for the 1900–1904 interval, with the exception of non-normalized German (which can most likely be explained by the temporal currency of the test sets).
4. Normalization of German data slightly decreases reliability, yet increases accuracy.

Training Scenario			Top-1 Reliability			Similarity Accuracy		
Language	Time Span	Embeddings	1 Epoch	5 Epochs	10 Epochs	1 Epoch	5 Epochs	10 Epochs
English Fiction	1900–1904	SGNS	0.11	0.33	0.40	0.45	0.51	0.51
		SGHS	0.23	0.33	0.33	0.46	0.45	0.45
	2005–2009	SGNS	0.36	0.54	0.57	0.58	0.58	0.57
		SGHS	0.36	0.39	0.38	0.55	0.52	0.52
German	1900–1904	SGNS	0.20	0.47	0.54	0.45	0.56	0.56
		SGHS	0.34	0.43	0.42	0.48	0.49	0.47
	2005–2009	SGNS	0.31	0.50	0.53	0.51	0.54	0.54
		SGHS	0.34	0.38	0.36	0.49	0.48	0.47
Normalized German	1900–1904	SGNS	0.19	0.45	0.52	0.47	0.55	0.57
		SGHS	0.32	0.42	0.42	0.47	0.48	0.48
	2005–2009	SGNS	0.30	0.48	0.52	0.54	0.59	0.60
		SGHS	0.33	0.37	0.36	0.51	0.52	0.52

Table 2: Accuracy and reliability among top-1 words for threefold repetition of different training scenarios after completing 1, 5 and 10 training epochs, respectively.

We also measured analogy accuracy for the English Fiction data sets, and observed no negative effect of multiple training epochs, yet a more pronounced gap between training methods, e.g., 36% of all analogies were correct for SGNS and only 27% for SGHS after one epoch on 1900–1904 data.

In the following, we further explore system performance as influenced, e.g., by word frequency, word ambiguity and the number of training epochs. For German, we focus on the normalized version due to the overall similar performance and suitability for further applications.

Influence of Neighborhood Size. Reliability at different top-n cut-offs is very similar for all languages and time spans under scrutiny, confirming previous observations in Hellrich and Hahn (2016a) and strengthening the suggestion to use only top-1 reliability for evaluation. Figure 2 illustrates this phenomenon with an SGNS trained on 1900–1904 English Fiction data. We assume this to be connected with the general decrease in `word2vec` embedding utility for high values of n already observed by Schnabel et al. (2015).

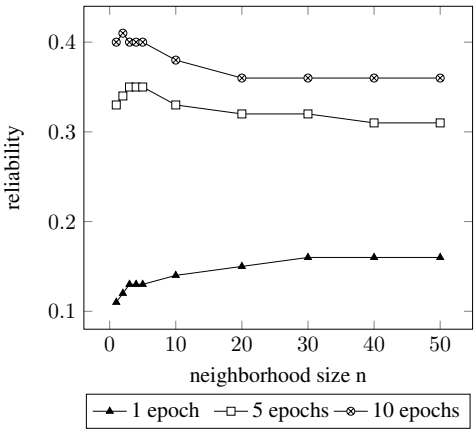

Figure 2: Effect of neighborhood size parameter n in reliability calculation for SGNS embeddings trained on 1900–1904 English Fiction data.

Influence of Word Frequency. Figures 3 and 4 depict the influence of word frequency (as percentile ranks) for English, as well as orthographically normalized German. Negative sampling is overall more reliable, especially for words with low or medium frequency. Word frequency has a less pronounced effect on reliability for German and negative sampling is again preferable, especially for low or medium frequency words. The 21 English

words reported to have undergone traceable semantic changes in prior work[9] are all frequent with percentiles between 89 and 99—for such high-frequency words hierarchical softmax performs similarly or even slightly better. The relatively low reliability for medium-frequency English words, as compared to German ones, could be caused by a peculiar pattern of word co-occurrences, illustrated in Figures 5 and 6 for 1900–1904 English Fiction, respectively normalized German. Medium-frequency English words have fewer co-occurrences with low-frequency words than German ones, which might result in a lack of specific contexts for these words during training and thus hamper embedding quality.

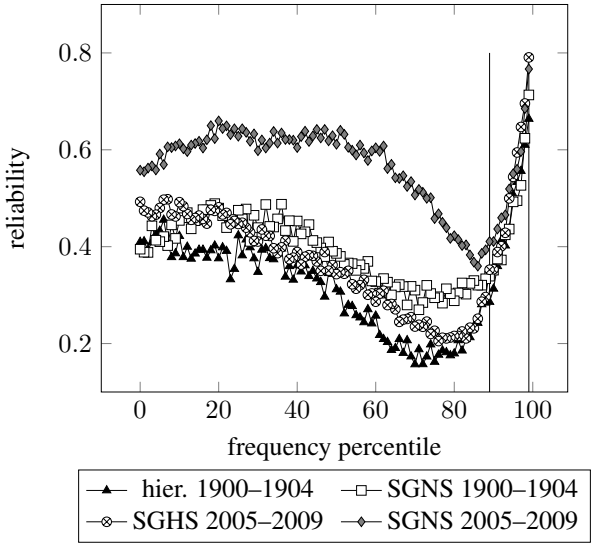

Figure 3: Influence of frequency percentile on reliability for models trained for 10 epochs on English Fiction data from 1900–1904 and 2005–2009. Words reported to have changed their semantics during the 20th century fall into the frequency range marked by the vertical lines.

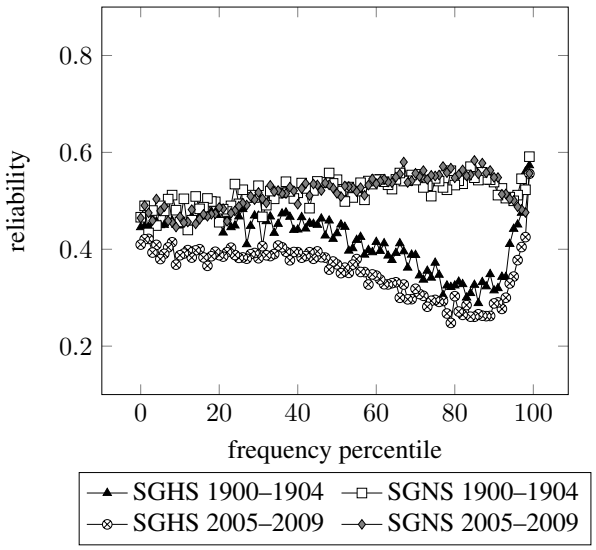

Figure 4: Influence of frequency percentile on reliability for models trained for 10 epochs on orthographically normalized German data from 1900–1904 and 2005–2009.

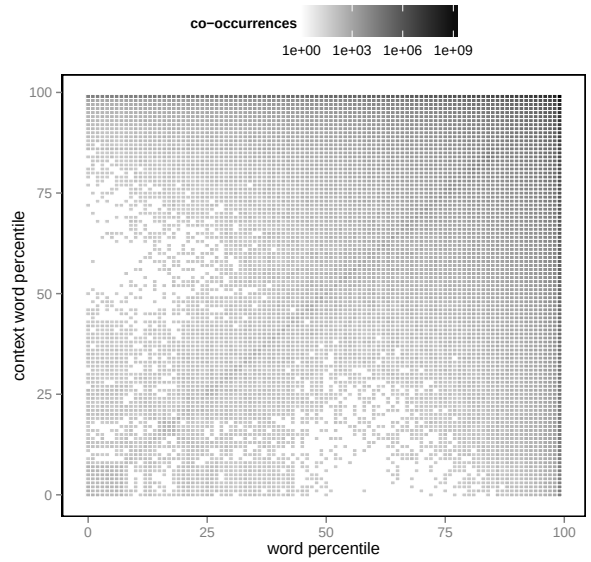

Figure 5: Number of co-occurrences (indicated by shade; only values above mode) between words and context words per frequency percentile for English Fiction 1900–1904 data.

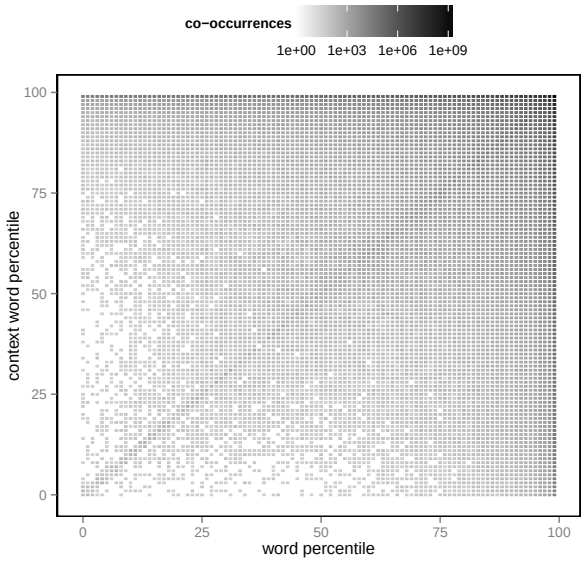

Figure 6: Number of co-occurrences (indicated by shade; only values above mode) between words and context words per frequency percentile for normalized German 1900–1904 data.

[9]Kulkarni et al. (2015) compiled the following list based on prior work (Wijaya and Yeniterzi, 2011; Gulordava and Baroni, 2011; Jatowt and Duh, 2014; Kim et al., 2014): *card*, *sleep*, *parent*, *address*, *gay*, *mouse*, *king*, *checked*, *check*, *actually*, *supposed*, *guess*, *cell*, *headed*, *ass*, *mail*, *toilet*, *cock*, *bloody*, *nice* and *guy*.

Influence of Word Ambiguity. Entries in lexical databases, such as WORDNET[10] (Fellbaum, 1998) and its German counterpart GERMANET[11] (Lemnitzer and Kunze, 2002), can be employed to approximate the effect of word ambiguity on reliability. The number of synsets a word belongs to (i.e., the number of its senses) seems to be positively correlated with top-1 reliability for English, as shown in Figure 7, whereas orthographically normalized German is less affected by ambiguity as Figure 8 reveals. This counter-intuitive effect for English seems to be caused by the low ambiguity of infrequent words—results become more uniform, if analysis is limited to high frequency words (e.g., 90th frequency percentile or higher).

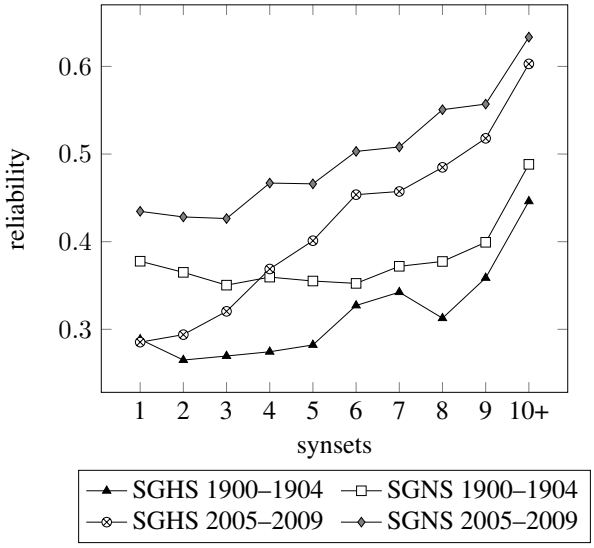

Figure 7: Influence of ambiguity (measured by the number of WORDNET synsets) on top-1 reliability for models trained for 10 epochs on English Fiction data from 1900–1904 and 2005–2009.

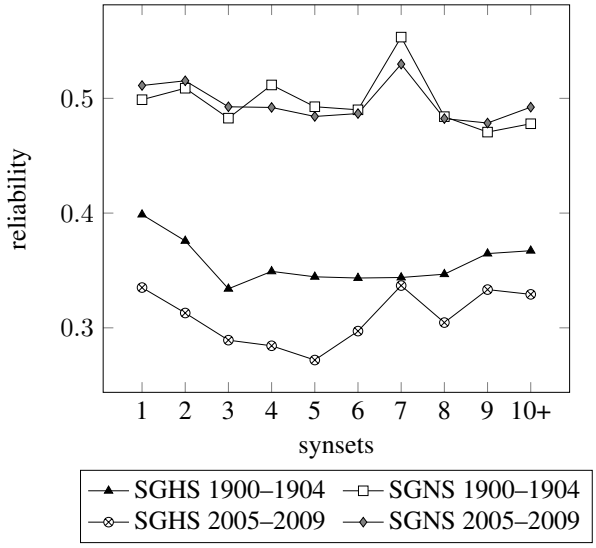

Figure 8: Influence of ambiguity (measured by the number of GERMANET synsets) on top-1 reliability for models trained for 10 epochs on orthographically normalized German data from 1900–1904 and 2005–2009.

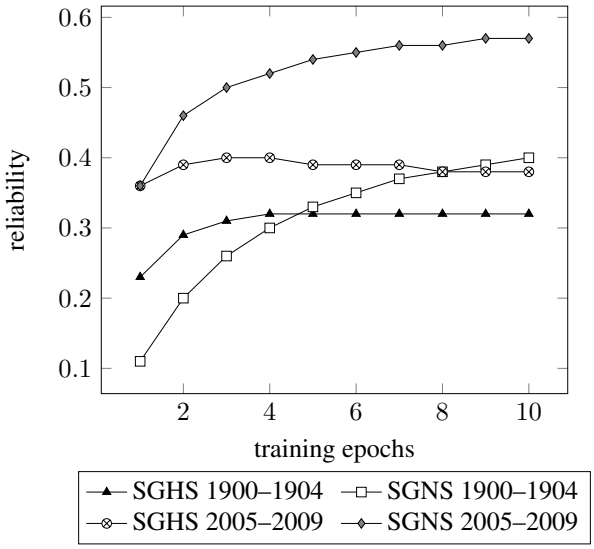

Figure 9: Top-1 reliability as influenced by the number of training epochs, for English Fiction data relative to the 1900–1904 and 2005–2009 time slices.

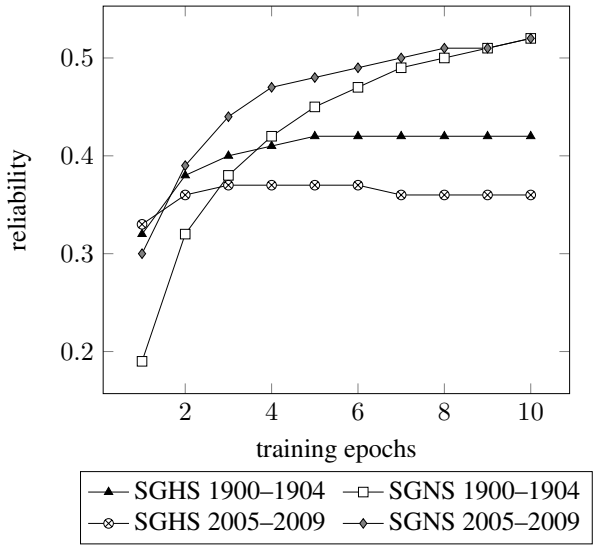

Figure 10: Top-1 reliability as influenced by the number of training epochs, for orthographically normalized German data relative to the 1900–1904 and 2005–2009 time slices.

Influence of the Number of Training Epochs. Model reliability and accuracy depend on the number of training epochs, as shown in Figures 9 and 10 for English and normalized German, respectively. For

[10]We used WORDNET 3.0 and the API provided by the Natural Language Toolkit (NLTK): `www.nltk.org`

[11]We used GERMANET 11.0 and the PYGERMANET API: `https://pypi.python.org/pypi/pygermanet`

both languages and time spans negative sampling outperforms hierarchical softmax, if training lasts for a sufficient number of epochs. The number of necessary epochs for negative sampling to become superior seems to be linked to both language and corpus size, as it is lower for 2005–2009 than for 1900–1904 data. While reliability continues to increase for each subsequent epoch under negative sampling, there are clear diminishing returns and even regression under hierarchical softmax.

To test for potential overfitting effects, we analyzed similarity accuracy as influenced by the number of training epochs (some values are already given in Table 2). Figures 11 and 12 show the results for English and orthographically normalized German, respectively. Note that accuracy is assessed on a test set for modern-day language, and can thus not be considered a fully valid yardstick. Accuracy behaves similar to reliability, as under the negative sampling condition it clearly profits from multiple training epochs. This effect is more pronounced for smaller corpora; the biggest corpus (i.e., English Fiction 2005–2009) shows a slight regression in accuracy after more than 5 training epochs.

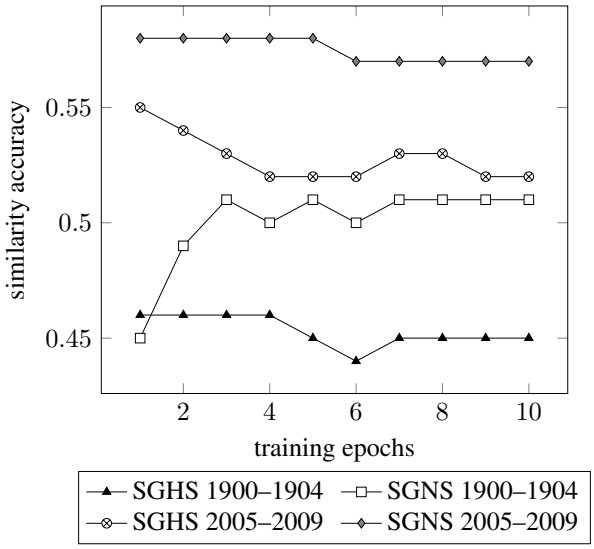

Figure 11: Similarity accuracy as influenced by the number of training epochs for English Fiction data data relative to the 1900–1904 and 2005–2009 time slices. Error bars are not displayed on purpose due to constant values for each training method.

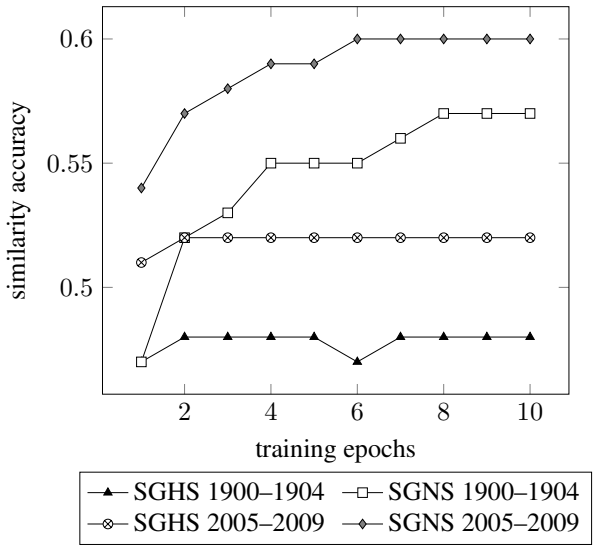

Figure 12: Similarity accuracy as influenced by the number of training epochs for orthographically normalized German data relative to the 1900–1904 and 2005–2009 time slices. Error bars are not displayed on purpose due to constant values for each training method.

Conclusions. Both reliability and accuracy point towards negative sampling with 4 to 6 training epochs (6 being better for smaller and 4 being better for larger corpora) as the optimal training regime for all tested combinations of languages and time spans (implicitly, this is also a test on largely varying corpus sizes, see Table 1). Such a training scheme yields models with high reliability without losses in accuracy (that would indicate overfitting). Figure 13 shows Δc, i.e., the difference of the convergence measure c (Equations (2) and (3) averaged over all three models) between subsequent epochs, for both German and English data from the intervals 1900–1904 and 2005–2009. Few changes occur after 4–6 epochs, which could be alternatively expressed as a Δc of about 0.003. The convergence criterion proposed by Kulkarni et al. (2015), i.e., $c = 0.9999$, was never reached (this observation might be explained by Kulkarni et al.'s decision not to reset the learning rate for each training epoch, as was done by us and Kim et al. (2014)).

SVD_{PPMI}, which are conceptually not bothered by the reliability problems we discussed here, were not a good fit for the hyperparameters we adopted from Kulkarni et al. (2015). Hamilton et al. (2016) reports similarity accuracy superior to SGNS, whereas for our set-up results in pretests were about 10 percent points worse than skip-gram embeddings, e.g., only 0.35 for 1900–1904 English Fiction.

Finally, to want to illustrate how this reliability problem affects qualitative conclusions. In Table 3 we provide some examples in which three negative sampling models for 1900–1904 English Fiction did not agree on the closest neighbor for words in question (mostly drawn from the list in Footnote 9). The most

inconsistent word neighborhoods are provided for *'romantic'* which is connected to *'lazzaroni'*,[12] *'fanciful'* and *'melancholies'*. This holds despite the high frequency (94th percentile) and moderate ambiguity (5 synsets) of the target item *'romantic'*.

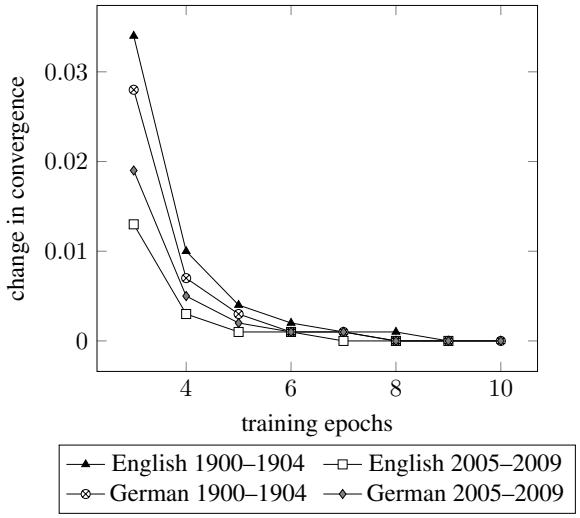

Figure 13: Change of averaged convergence measurement c between each epoch and its predecessor for models of orthographically normalized German and English Fiction trained with negative sampling on the 1900–1904 and 2005–2009 time slices. Values for epochs 1 and 2 would be one magnitude higher and are thus not displayed.

Word	Disputed Closest Neighbor
romantic	lazzaroni, fanciful, melancholies
parent	child, child, mother
mouse	mice, rat, cat
checked	checking, check, checking
check	cheque, checked, cheque
guess	reckon, reckon, suppose
headed	headedness, haired, haired
ass	atheist, fool, fool
toilet	ironing, dressing, dressing
cock	cocks, arty, hen
bloody	mistyken, mistyken, wreaks
nice	stunner, fine, fine

Table 3: A sample list of target lexical items for which three identically parametrized systems (trained with negative sampling on 1900–1904 English Fiction data) disagreed on the closest neighbor. Examples are mostly drawn from the list of the twenty-one aforementioned words (see Footnote 9) that were claimed to have undergone changes during the 20th century.

6 Discussion

Our investigation into the accuracy and reliability of skip-gram word embeddings shows even the most reliable systems too often provide inconsistent word neighborhoods. This carries unwarranted potential for erroneous conclusions on a word's semantic evolution as was shown, e.g., for the lexical item *'romantic'* and English Fiction texts from the 1900–1904 time slice. We are thus skeptical about using word neighborhoods in skip-gram embedding space to adequately capture natural languages' lexical semantics (for English and German, at least). While we found some mitigation strategies, i.e., training for multiple epochs or using our convergence criterion of $\Delta c \lesssim 0.003$, we assume SVD_{PPMI} to be conceptually superior. Future work might try to provide general guidelines for proper hyperparameter selection for SVD_{PPMI}, especially regarding complete temporal slices of the GBN (Hamilton et al. (2016) used samples). Alternatively, training several identically parametrized SGNS/SGHS models and combining them into an ensemble might constitute an easy way to reduce the reliability problems we described, yet at the price of exorbitant computational costs.

Acknowledgments

This research was conducted within the Graduate School *"The Romantic Model. Variation – Scope – Relevance"* (`http://www.modellromantik.uni-jena.de/?lang=en`) supported by grant GRK 2041/1 from the *Deutsche Forschungsgemeinschaft (DFG)*.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. *Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors. In Daniel Marcu, Kristina Toutanova, and Hua Wu, editors, *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA, June 22-27, 2014*, volume 1: Long Papers, pages 238–247, Stroudsburg/PA. Association for Computational Linguistics (ACL).

[12] A historical group of lower-class persons from Naples ("lazzarone, n", 2016).

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry P. Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In Arthur Gretton and Christian C. Robert, editors, *AISTATS 2016 — Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. Cadiz, Spain, May 7-11, 2016*, number 51 in JMLR Workshop and Conference Proceedings, pages 130–138.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David J. Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In Omer Levy, Felix Hill, Anna Korhonen, Kyunghyun Cho, Roi Reichart, Yoav Goldberg, and Antoine Bordes, editors, *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP @ ACL 2016. Berlin, Germany, August 12, 2016*, pages 7–12, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and reducing stereotypes in word embeddings. In James Faghmous, Rayid Ghani, Matt Ghee, Gideon S. Mann, Aleksandra Mojsilovic, and Kush R. Varshney, editors, *Proceedings of the Workshop on #Data4Good: Machine Learning in Social Good Applications @ ICML 2016. New York City, NY, USA, June 24, 2016*, pages 41–45.

Christiane Fellbaum, editor. 1998. WORDNET*: An Electronic Lexical Database*. MIT Press, Cambridge/MA; London/England.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the GOOGLE BOOKS NGRAM corpus. In Sebastian Padó and Yves Peirsman, editors, *GEMS 2011 — Proceedings of the Workshop on GEometrical Models of Natural Language Semantics @ EMNLP 2011. Edinburgh, UK, July 31, 2011*, pages 67–71, Stroudsburg/PA. Association for Computational Linguistics (ACL).

William L. Hamilton, Jure Leskovec, and Daniel Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In Antal van den Bosch, Katrin Erk, and Noah A. Smith, editors, *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*, volume 1: Long Papers, pages 1489–1501, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Johannes Hellrich and Udo Hahn. 2016a. An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In Beatrice Alex and Nils Reiter, editors, *LaTeCH 2016 — Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities @ ACL 2016, Berlin, Germany, August 11, 2016*, pages 111–117, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Johannes Hellrich and Udo Hahn. 2016b. Measuring the dynamics of lexico-semantic change since the German Romantic period. In *Digital Humanities 2016 — Conference Abstracts of the 2016 Conference of the Alliance of Digital Humanities Organizations (ADHO). 'Digital Identities: The Past and the Future'. Kraków, Poland, 11-16 July 2016*, pages 545–547.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *JCDL '14 — Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. London, U.K., September 8-12, 2014*, pages 229–238, Piscataway/NJ. IEEE Computer Society Press.

Eun Seo Jo. 2016. Diplomatic history by data. Understanding Cold War foreign policy ideology using networks and NLP. In Maciej Eder and Jan Rybicki, editors, *Digital Humanities 2016 — Conference Abstracts of the 2016 Conference of the Alliance of Digital Humanities Organizations (ADHO). 'Digital Identities: The Past and the Future'. Kraków, Poland, 11-16 July 2016*, pages 582–585.

Bryan Jurish. 2013. Canonicalizing the Deutsches Textarchiv. In Ingelore Hafemann, editor, *Proceedings of Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens "Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften. Berlin, Germany, December 12-13, 2011*, number 4 in Thesaurus Linguae Aegyptiae, pages 235–244. Berlin-Brandenburgische Akademie der Wissenschaften.

Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In James Bailey and Alistair Moffat, editors, *CIKM '15 — Proceedings of the 24th ACM International Conference on Information and Knowledge Management. Melbourne, Australia, October 19-23, 2015*, pages 1191–1200, New York/NY, USA. Association for Computing Machinery (ACM).

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In Cristian Danescu-Niculescu-Mizil, Jacob Eisenstein, Kathleen R. McKeown, and Noah A. Smith, editors, *Proceedings of the Workshop on Language Technologies and Computational Social Science @ ACL 2014. Baltimore, Maryland, USA, June 26, 2014*, pages 61–65, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, editors, *WWW '15 — Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, May 18-22, 2015*, volume Technical Papers, pages 625–635, New York/NY. Association for Computing Machinery (ACM).

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. In Krishna P. Gummadi, Markus Strohmaier, Eric Gilbert, Michael Macy, and Claudia Wagner, editors, *ICWSM-16 — Proceedings of the 10th International AAAI Conference on Web and Social Media. Cologne, Germany, May 17-20, 2016*, pages 615–618, Palo Alto/CA. Association for the Advancement of Artificial Intelligence (AAAI), AAAI Press.

"lazzarone, n". 2016. In *OED Online*. Oxford University Press. `http://www.oed.com/view/Entry/106565` (accessed June 16, 2016).

Lothar Lemnitzer and Claudia Kunze. 2002. GERMANET: Representation, visualization, application. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Angel Martin Municio, Daniel Tapias, and Antonio Zampolli, editors, *LREC 2002 — Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas, Canary Islands, Spain, 27 May - June 2, 2002*, pages 1485–1491, Paris. European Language Resources Association (ELRA).

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In Daniel Marcu, Kristina Toutanova, and Hua Wu, editors, *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA, June 22-27, 2014*, volume 2: Short Papers, pages 302–308, Stroudsburg/PA. Association for Computational Linguistics (ACL).

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.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. In Min Zhang, editor, *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, July 10, 2012*, volume System Demonstrations, pages 169–174, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge/MA.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January.

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

Alexander Panchenko. 2016. Best of both worlds: Making word sense embeddings interpretable. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan E. J. M. Odijk, and Stelios Piperidis, editors, *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 2649–2655, Paris. European Language Resources Association (ELRA-ELDA).

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the GOOGLE BOOKS corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One*, 10(10):e0137041, October.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GLOVE: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25-29, 2014*, pages 1532–1543, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In Lluís Márquez, Chris Callison-Burch, and Jian Su, editors, *EMNLP 2015 — Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 17-21 September 2015*, pages 298–307, Red Hook/N.Y. Association for Computational Linguistics (ACL).

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In Sergej Sizov, Stefan Siersdorfer, Philipp Sorg, and Thomas Gottron, editors, *DETECT '11 — Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web @ CIKM 2011. Glasgow, U.K., October 24, 2011*, pages 35–40, New York/N.Y. Association for Computing Machinery (ACM).

Torsten Zesch and Iryna Gurevych. 2006. Automatically creating datasets for measures of semantic relatedness. In John Nerbonne and Erhard W. Hinrichs, editors, *Proceedings of the Workshop on Linguistic Distances @ COLING-ACL 2006. Sydney, Australia, 23 July 2006*, pages 16–24, Stroudsburg/PA. Association for Computational Linguistics (ACL).