An Assessment of Experimental Protocols for Tracing Changes in Word Semantics Relative to Accuracy and Reliability

Johannes Hellrich
Research Training Group "The Romantic Model. Variation - Scope - Relevance"
Friedrich-Schiller-Universität Jena
Jena, Germany
johannes.hellrich@uni-jena.de

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

Abstract

Our research aims at tracking the semantic evolution of the lexicon over time. For this purpose, we investigated two well-known training protocols for neural language models in a synchronic experiment and encountered several problems relating to accuracy and reliability. We were able to identify critical parameters for improving the underlying protocols in order to generate more adequate diachronic language models.

1 Introduction

The lexicon can be considered the most dynamic part of all linguistic knowledge sources over time. There are two innovative change strategies typical for lexical systems: the creation of entirely new lexical items, commonly reflecting the emergence of novel ideas, technologies or artifacts, on the one hand, and, on the other hand, shifts in the meaning of already existing lexical items, a process which usually takes place over larger periods of time. Tracing semantic changes of the latter type is the main focus of our research.

Meaning shift has recently been investigated with emphasis on neural language models (Kim et al., 2014; Kulkarni et al., 2015). This work is based on the assumption that the measurement of semantic change patterns can be reduced to the measurement of lexical similarity between lexical items. Neural language models, originating from the `word2vec` algorithm (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c), are currently considered as state-of-the-art solutions for implementing this assumption (Schnabel et al., 2015). Within this approach, changes in similarity relations between lexical items at two different points of time are interpreted as a signal for meaning shift. Accordingly, lexical items which are very similar to the lexical item under scrutiny can be considered as approximating its meaning at a given point in time. Both techniques were already combined 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).

We here investigate the accuracy and reliability of such similarity judgments derived from different training protocols dependent on word frequency, word ambiguity and the number of training epochs (i.e., iterations over all training material). *Accuracy* renders a judgment of the overall model quality, whereas *reliability* between repeated experiments ensures that qualitative judgments can indeed be transferred between experiments. Based on the identification of critical conditions in the experimental set-up of previously employed protocols, we recommend improved training strategies for more adequate neural language models dealing with diachronic lexical change patterns. Our results concerning reliability also cast doubt on the reproducibility of experiments where semantic similarity between lexical items is taken as a computationally valid indicator for properly capturing lexical meaning (and, consequently, meaning shifts) under a diachronic perspective.

2 Related Work

Neural language models for tracking semantic changes over time typically distinguish between two different training protocols—*continuous training* of models (Kim et al., 2014) where the model for each time span is initialized with the embeddings of its predecessor, and, alternatively, *independent training* with a mapping between models for different points in time (Kulkarni et al., 2015). A comparison between these two protocols,

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 111–117,
Berlin, Germany, August 11, 2016.

such as the one proposed in this paper, has not been carried out before. Also, the application of such protocols to non-English corpora is lacking, with the exception of our own work relating to German data (Hellrich and Hahn, 2016b; Hellrich and Hahn, 2016a).

The `word2vec` algorithm is a heavily trimmed version of an artificial neural network used to generate low-dimensional vector space representations of a lexicon. We focus on its skip-gram variant, trained to predict plausible contexts for a given word that was shown to be superior over other settings for modeling semantic information (Mikolov et al., 2013a). There are several parameters to choose for training—learning rate, downsampling factor for frequent words, number of training epochs and choice between two strategies for managing the huge number of potential contexts. One strategy, *hierarchical softmax*, uses a binary tree to efficiently represent the vocabulary, while the other, *negative sampling*, works by updating only a limited number of word vectors during each training step.

Furthermore, artificial neural networks, in general, are known for a large number of local optima encountered during optimization. While these commonly lead to very similar performance (LeCun et al., 2015), they cause different representations in the course of repeated experiments.

Approaches to modelling changes of lexical semantics not using neural language models, e.g., Wijaya and Yeniterzi (2011), Gulordava and Baroni (2011), Mihalcea and Nastase (2012), Riedl et al. (2014) or Jatowt and Duh (2014) are, intentionally, out of the scope of this paper. In the same way, we here refrain from comparison with computational studies dealing with literary discussions related to the Romantic period (e.g., Aggarwal et al. (2014)).

3 Experimental Set-up

For comparability with earlier studies (Kim et al., 2014; Kulkarni et al., 2015), we use the fiction part of the GOOGLE BOOKS NGRAM corpus (Michel et al., 2011; Lin et al., 2012). This part of the corpus is also less affected by sampling irregularities than other parts (Pechenick et al., 2015). Due to the opaque nature of GOOGLE's corpus acquisition strategy, the influence of OCR errors on our results cannot be reasonably estimated, yet we assume that they will affect all experiments in an equal manner.

The wide range of experimental parameters described in Section 2 makes it virtually impossible to test all their possible combinations, especially as repeated experiments are necessary to probe a method's reliability. We thus concentrate on two experimental protocols—the one described by Kim et al. (2014) (referred to as *Kim protocol*) and the one from Kulkarni et al. (2015) (referred to as *Kulkarni protocol*), including close variations thereof. Kulkarni's protocol operates on all 5-grams occurring during five consecutive years (e.g., 1900–1904) and trains models independently of each other. Kim's protocol operates on uniformly sized samples of 10M 5-grams for each year from 1850 onwards in a continuous fashion (years before 1900 are used for initialization only). Its constant sampling sizes result in both oversampling and undersampling as is evident from Figure 1.

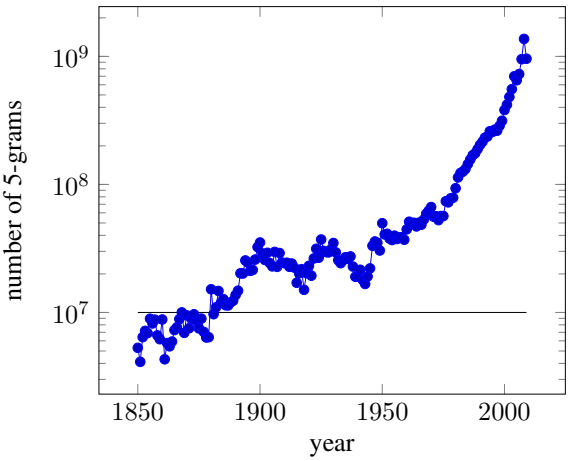

Figure 1: Number of 5-grams per year (on the logarithmic y-axis) contained in the English fiction part of the GOOGLE BOOKS NGRAM corpus. The horizontal line indicates a constant sampling size of 10M 5-grams according to the Kim protocol.

We use the PYTHON-based GENSIM[1] implementation of `word2vec` for our experiments; the relevant code is made available via GITHUB.[2] Due to the 5-gram nature of the corpus, a context window covering four neighboring words is used for all experiments. Only words with at least 10 occurrences in a sample are modeled. Training for each sample is repeated until convergence[3] is achieved or 10 epochs have passed. Following both protocols, we use word vectors with 200

[1] `https://radimrehurek.com/gensim/`

[2] `github.com/hellrich/latech2016`

[3] Defined as averaged cosine similarity of 0.9999 or higher between word representations before and after an epoch (see Kulkarni et al. (2015)).

Table 1: Accuracy and reliability among top n words for threefold application of different training protocols. Reliability is given as fraction of the maximum for n. Standard deviation for accuracy ± 0, if not noted otherwise; reliability is based on the evaluation of all lexical items, thus no standard deviation.

Description of training protocol			top-n Reliability					Accuracy
			1	2	3	4	5	
independent	negative	in all texts	0.40	0.41	0.41	0.40	0.40	0.38
		in 10M sample	0.45	0.48	0.50	0.51	0.52	0.25
		between 10M samples	0.09	0.10	0.10	0.10	0.10	0.26
	hierarchical	in all texts	0.33	0.34	0.34	0.34	0.34	0.28
		in 10M sample	0.38	0.40	0.42	0.42	0.43	0.22
		between 10M samples	0.09	0.09	0.10	0.10	0.10	0.22 ± 0.01
continuous	negative	in 10M sample	0.54	0.55	0.56	0.56	0.57	0.25
		between 10M samples	0.21	0.21	0.22	0.22	0.22	0.25
	hierarchical	in 10M sample	0.31	0.32	0.32	0.32	0.33	0.22
		between 10M samples	0.12	0.13	0.13	0.13	0.13	0.23

dimensions for all experiments, as well as an initial learning rate of 0.01 for experiments based on 10M samples, and one of 0.025 for systems trained on unsampled texts; the threshold for downsampling frequent words was 10^{-3} for sample-based experiments and 10^{-5} for unsampled ones. We tested both negative sampling and hierarchical softmax training strategies, the latter being canonical for Kulkarni's protocol, whereas Kim's protocol is underspecified in this regard.

We evaluate *accuracy* by using the test set developed by Mikolov et al. (2013a). This test set is based on present-day English language and world knowledge, yet we assume it to be a viable proxy for overall model quality. 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*.

We evaluate *reliability* by training three identically parametrized models for each experiment. We then compare the top n similar words (by cosine distance) for each word modeled by the experiments with a variant of the Jaccard coefficient (Manning et al., 2008, p.61). We limit our analysis to values of n between 1 and 5, in accordance with data on `word2vec` accuracy (Schnabel et al., 2015). The 3-dimensional array $W_{i,j,k}$ contains words ordered by similarity (i) for a word in question (j) according to an experiment (k). If a word in question is not modeled by an experiment, as can be the case for comparisons over different samples, $\emptyset$ is the corresponding entry. 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 any of these experiments and normalized by n, the maximally achievable score for this value of n:

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

4 Results

We focus our analysis on the representations generated for the initial period, i.e., 1900 for sample-based experiments and 1900–1904 for unsampled ones. This choice was made since researchers can be assumed to be aware of current word meanings, thus making correct judgments on initial word semantics more important. As a beneficial side effect, we get a marked reduction of computational demands, saving several CPU years compared to an evaluation based on the most recent period.

4.1 Training Protocols

Table 1 depicts the assessments for different training protocols. Four results seem relevant for future experiments. First, reliability at different top-n cut-offs is rather uniform, so that evaluations could be performed on top-1 reliability only without real losses. Second, both accuracy and reliability are often far higher for negative sampling than for hierarchical softmax under direct comparison of the evaluated conditions; under no condition hierarchical softmax outperforms negative sampling. Third, continuous training improves reliability, yet not accuracy, for systems trained on samples. Fourth, reliability for experiments between samples heavi-

ly degrades compared to reliability for repeated experiments on the same sample.

4.2 Detailed Investigation

As variations of Kulkarni's protocol yield more consistent results, we further explore its performance considering word frequency, word ambiguity and the number of training epochs. All experiments described in this section are based on the complete 1900–1904 corpus. Figure 2 shows the influence of word frequency, negative sampling being overall more reliable, especially for words with low or medium frequency. The 21 words reported to have undergone traceable semantic changes[4] are all frequent with percentiles between 89 and 99. For such high-frequency words hierarchical softmax performs similar or slightly better.

Entries in the lexical database WordNet (Fellbaum, 1998) can be employed to measure the effect of word ambiguity on reliability.[5] The number of WordNet synsets a word belongs to (i.e., the number of its senses) seems to have little effect on top-1 reliability for negative sampling, while hierarchical softmax underperforms for words with a low number of senses, as shown in Figure 3.

Model reliability and accuracy depend on the number of training epochs, as shown in Figure 4. There are diminishing returns for hierarchical softmax, reliability staying constant after 5 epochs, while negative sampling increases in reliability with each epoch. Yet, both methods achieve maximal accuracy after only 2 epochs; additional epochs lead to a small decrease from 0.4 down to 0.38 for negative sampling. This could indicate overfitting, but accuracy is based on a test set for modern-day language, and can thus not be considered a fully valid yardstick.

5 Discussion

Our investigation in the performance of two common protocols for training neural language models on historical text data led to several hitherto unknown results. We could show that negative sampling outperforms hierarchical softmax both in terms of accuracy and reliability, especially

[4]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*.

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

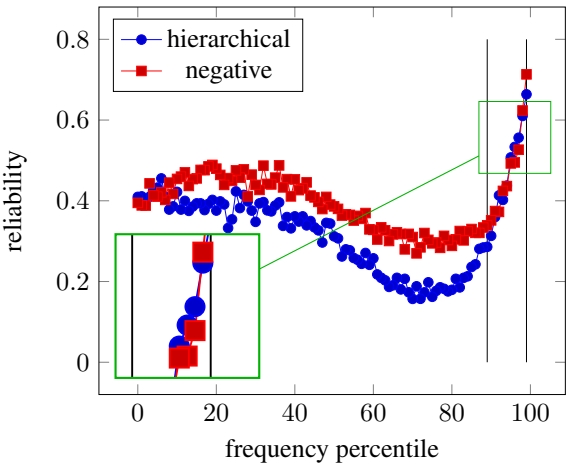

Figure 2: Influence of percentile frequency rank on reliability for models trained for 10 epochs on 1900–1904 data. Words reported to have changed during the 20th century fall into the rank range marked by vertical lines.

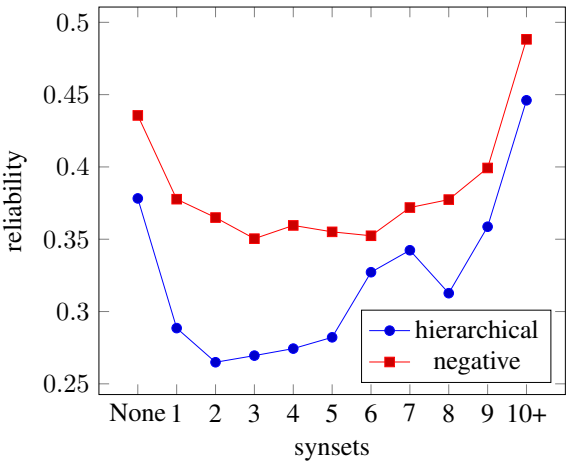

Figure 3: Influence of ambiguity (measured by the number of WordNet synsets) on top-1 reliability for models trained for 10 epochs on 1900–1904 data.

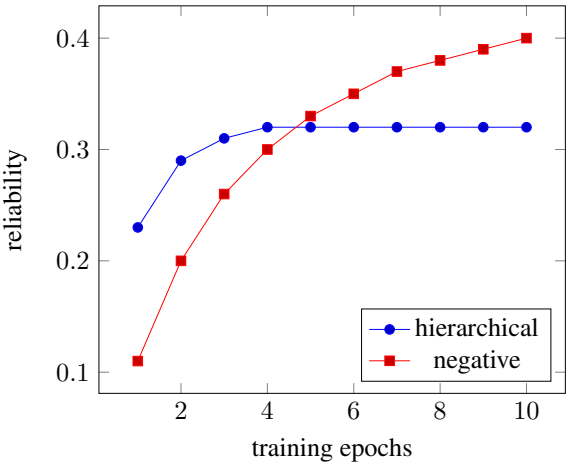

Figure 4: Top-1 reliability as influenced by the number of training epochs, for 1900–1904 data.

for infrequent and low-ambiguity words, if time for sufficient training epochs is available.[6] Our synchronic experiments provide evidence for the superiority of Kulkarni's over Kim's protocol, especially if modified to use negative sampling. Longer training time, due to unsampled corpora, can be mitigated by training models in parallel, which is impossible for Kim's protocol. We strongly suggest to train only on full corpora, and not on samples, due to very low reliability values for systems trained on different samples. If samples are necessary, continuous training can somewhat lower its negative effect on reliability between samples.

Even the most reliable system often identifies widely different words as most similar. This carries unwarranted potential for erroneous conclusions on a words' semantic evolution, e.g., *"romantic"* happens to be identified as most similar to *"lazzaroni"*[7], *"fanciful"* and *"melancholies"* by three systems trained with negative sampling on 1900–1904 texts. We are thus skeptical about using such similarity clouds to describe or visualize lexical semantics at a point in time.

In future work, we will explore the effects of continuous training based on complete corpora. The selection of a convergence criterion remains another open issue due to the threefold trade-off between training time, reliability and accuracy. It would also be interesting to replicate our experiments for other languages or points in time. Yet, the enormous corpus size for more recent years might require a reduced number of maximum epochs for these experiments. In order to improve the semantic modeling itself one could lemmatize the training material or utilize the part of speech annotations provided in the latest version of the GOOGLE corpus (Lin et al., 2012). Also, recently available neural language models with support for multiple word senses (Bartunov et al., 2016; Panchenko, 2016) could be helpful, since semantic changes can often be described as changes in the usage frequency of different word senses (Rissanen, 2008, pp.58–59). Finally, it is clearly important to test the effect of our proposed changes, based on synchronic experiments, on a system for tracking diachronic changes in word semantics.

[6]Using parallel 8 processes on an Intel Xeon E5649@2.53Ghz, completing a training epoch for 1900–1904 data takes about three hours, while 5 days are necessary for 2005–2009 data.

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

Acknowledgments

This research was conducted within the Research Training Group *"The Romantic Model. Variation – Scope – Relevance"* (`http://www.modellromantik.uni-jena.de/`) supported by grant GRK 2041/1 from the *Deutsche Forschungsgemeinschaft (DFG)*. The first author (J.H.) wants to thank the members of the GRK for their collaborative efforts.

References

Nitish Aggarwal, Justin Tonra, and Paul Buitelaar. 2014. Using distributional semantics to trace influence and imitation in Romantic Orientalist poetry. In Alan Akbik and Larysa Visengeriyeva, editors, *Proceedings of the AHA! Workshop on Information Discovery in Text @ COLING 2014. Dublin, Ireland, August 23, 2014*, pages 43–47. Association for Computational Linguistics (ACL).

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.

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

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).

Johannes Hellrich and Udo Hahn. 2016a. Measuring the dynamics of lexico-semantic change since the German Romantic period. In *Digital Humanities 2016 – Proceedings 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*.

Johannes Hellrich and Udo Hahn. 2016b. Romantik im Wandel der Zeit – eine quantitative Untersuchung. In *DHd 2016 – 3. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum. Modellierung-Venetzung-Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Leipzig, Germany, March 7-12, 2016*, pages 325–326.

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.

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 *WWW '15 — Proceedings of the 24th International Conference on World Wide Web. May 18-22, 2015, Florence, Italy*, pages 625–635, New York, N.Y. Association for Computing Machinery (ACM).

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

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444, May.

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, *Proceedings of the System Demonstrations @ 50th Annual Meeting of the Association for Computational Linguistics — ACL 2012. Jeju Island, Korea, 10 July 2012*, pages 169–174, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York/NY, USA.

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.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, July 8-14, 2012*, volume 2: Short Papers, pages 259–263, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. 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*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26 — NIPS 2013. Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA, December 5-10, 2013*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé, and Katrin Kirchhoff, editors, *NAACL-HLT 2013 — Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, GA, USA, 9-14 June 2013*, pages 746–751, Stroudsburg/PA. Association for Computational Linguistics (ACL).

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, Helene Mazo, Asuncion Moreno, Jan 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.

Martin Riedl, Richard Steuer, and Chris Biemann. 2014. Distributed distributional similarities of Google Books over the centuries. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 1401–1405, Paris. European Language Resources Association (ELRA).

Matti Rissanen. 2008. Corpus linguistics and historical linguistics. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, number 29/1 in Handbücher zur Sprach- und

Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK), chapter 4, pages 53–68. de Gruyter Mouton, Berlin, New York.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP 2015 — Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 17-21 September 2015*, pages 298–307. 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).