Journal of Open Humanities Data

A Collection of Swedish Diachronic Word Embedding Models Trained on Historical Newspaper Data

DATA PAPER

SIMON HENGCHEN
NINA TAHMASEBI
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This paper describes the creation of several word embedding models based on a large collection of diachronic Swedish newspaper material available through Språkbanken Text, the Swedish language bank. This data was produced in the context of Språkbanken Text's continued mission to collaborate with humanities and natural language processing (NLP) researchers and to provide freely available language resources, for the development of state-of-the-art NLP methods and tools.

CORRESPONDING AUTHOR:
Simon Hengchen
Språkbanken Text, Department of Swedish, University of Gothenburg, SE
simon.hengchen@gu.se

KEYWORDS:
word embeddings; semantic change; newspapers; diachronic word embeddings

TO CITE THIS ARTICLE:
Hengchen, S., & Tahmasebi, N. (2021). A Collection of Swedish Diachronic Word Embedding Models Trained on Historical Newspaper Data. *Journal of Open Humanities Data*, 7: 2, pp. 1–7. DOI: *https://doi.org/10.5334/johd.22*

DOI: 10.5334/johd.22

1 OVERVIEW

We release diachronic word2vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017) models in their skip-gram with negative sampling (SGNS) architecture. The models are trained on 20-year time bins, with two temporal alignment strategies: independently-trained models for post-hoc alignment (as introduced by Kulkarni et al. 2015), and incremental training (Kim et al., 2014). In the incremental scenario, a model for t_1 is trained and saved, then updated with the data from t_2. The resulting model is in turn saved, then updated with data from t_3, etc. Given that space alignment is shown to be noisy (Dubossarsky et al., 2017, 2019), we release the independently trained models without alignment and leave the choice of alignment algorithm to the end user.[1]

To make this data release useful out-of-the-box, as well as to foster reuse by researchers from various fields, we release models, code to run the whole pipeline, code examples to load and use the models, and documentation.

We would like to underline that language change is driven by humans. Humans learn from their mistakes and what was once considered acceptable thankfully is not anymore. Machine learning models trained on data from the past inevitably learn biases of their time and as a result, the models shared with this paper contain characteristics of the past. These characteristics include sexism, racism, antisemitism, homophobia, and other types of unacceptable characteristics of their time (see e.g. Tripodi et al., 2019). The authors do not endorse them, and neither does the University of Gothenburg. Nonetheless, whilst this is not the aim of this paper, we hope it can also help shed light on these representations, as ignoring them would mean they have never existed.

1.1 REPOSITORY LOCATION

The data set is available on Zenodo at *https://zenodo.org/record/4301658*.

1.2 CONTEXT

This data was produced in the context of Språkbanken Text's continued mission to collaborate with humanities and natural language processing (NLP) researchers and to provide freely available language resources for the development of state-of-the-art NLP methods and tools.

2 METHOD

We retrieved all Sparv-pipeline (Borin et al., 2016) processed XML files in the Kubhist 2 newspaper archive: Språkbanken Text makes several corpora – including Kubhist 2 – available through web interface Korp (Borin et al., 2012). The data in Korp has been processed (dependency parsing, semantic annotation, lemmatisation, etc.) with the Sparv pipeline. The original dataset and the specific steps are described below.

2.1 ORIGINAL DATA

The entirety of the Kungliga bibliotekets historiska tidningar ('The Royal Library's historical newspapers,' Kubhist 2) corpus (Språkbanken, 2019) was used. For a detailed description of the corpus, we refer to Adesam et al. (2019) and to a blog post by Dana Dannélls.[2] Kubhist 2 contains over 5.5 billion tokens, and it is made up of newspapers from all over Sweden.

2.2 STEPS

- Extracted all words from the XML
- Given the relative quality of the optical character recognition (OCR) output and to reduce the amount of OCR errors in the data set, we cleaned the resulting text with the following procedure:[3]

1 We refer to the LSCDetection repository (Schlechtweg et al., 2019; Schlechtweg & Schulte im Walde, 2020) for a selection of alignment methods: *https://github.com/Garrafao/LSCDetection/tree/master/alignment*.

2 *https://spraakbanken.gu.se/blogg/index.php/2019/09/15/the-kubhist-corpus-of-swedish-newspapers/*, last accessed 2020/11/27.

3 For a discussion on how OCR can affect quantitative text analyses, we refer the reader to Hill & Hengchen (2019) and van Strien et al. (2020).

DOI: 10.5334/johd.22

  - lowercasing;
  - removing digits;
  - removing all characters not belonging to the (lowercased) Swedish alphabet, which consists of the 26 letters in the Latin alphabet and å, ä, ö. This includes the removal of punctuation marks;
  - removing tokens the length of which is two characters or smaller

- Joined files belonging to the same double decade, starting with our earliest time bin of 1740 and ending in 1899 (i.e. 1740–1750; 1760–1770; ⋯; 1880–1890, where e.g. 1740 covers 1740; 1741; ⋯; 1749)
- For each time bin, trained two type-embedding language models with two "alignment" strategies:
  - word2vec, independently-trained and incrementally trained
  - fastText, independently-trained and incrementally trained

For both language models, we use the default hyperparameters in gensim[4] (Řehůřek & Sojka, 2010) aside from: vector dimensionality of 100, frequency threshold of 50, seed of 1830. The choice for default parameters is explained in Subsection 2.3.

2.3 QUALITY CONTROL

Several sanity checks were made during the preprocessing of the original data, including:

- Manual matching of records between the original XML and resulting corpus files as well as against the Korp version of Kubhist 2;
- Manual matching of metadata between the original XML and resulting corpus files as well as against the Korp version of Kubhist 2.

It is notoriously difficult to evaluate the quality of word embeddings for historical data, as annotated test sets are either lacking or extremely costly (Schlechtweg et al., 2020; Hengchen et al., 2021b). While synthetic evaluation procedures exist (Cook & Stevenson, 2010; Kulkarni et al., 2015; Rosenfeld & Erk, 2018; Dubossarsky et al., 2019; Shoemark et al., 2019), they are tailored for the specific task of semantic change (usually, the task of determining if there is a change of a word's meaning over time) and are not suited for general-purpose diachronic word embeddings as they might lead to privileging a (set of) hyperparameter(s) that might be detrimental to other tasks. As a result, we use default parameters and carry out a small-scale quality control by a) verifying that the code written does what it is expected to do; and b) making sure that models output semantic similarity, as expected.

The code to train word embeddings was read (but not run) by two computational linguists who have extensive experience with diachronic word embeddings and are not authors of this paper, and no errors were found. Once the word embeddings were trained, we selected several target words and systematically extracted the most similar terms for every model trained. Similar terms were then evaluated by a native speaker of Swedish who confirmed that such terms were indeed, to the best of their knowledge, semantically similar. In many cases and especially so for the fastText models that harvest subword information, the most similar words consist of OCR errors and spelling variations, an interesting avenue to pursue in future research. A (non-native speaker of Swedish) reviewer, whom we thank, also performed checks on the local neighbourhoods of selected terms as well as vector arithmetics, and confirmed the models behaved as expected.

We would like to note that the first time bins are very scarce in data, and warn researchers that previous work indicates this has a large influence on the stability of nearest-neighbour distances (Antoniak & Mimno, 2018). We would also like to acknowledge that different temporal alignment strategies might benefit from different hyperparameters for specific tasks (see e.g.

4 I.e.: alpha = 0.025, window = 5, sample = 0.001, min_alpha = 0.0001, negative = 5.

DOI: 10.5334/johd.22

Kaiser et al. 2020 for vector dimensionality in lexical semantic change (LSC)), however, for the reasons stated above we do not perform any tuning on any specific task either.

3 DATASET DESCRIPTION

3.1 OBJECT NAME

The dataset is named `HENGCHEN-TAHMASEBI_-_2020_-_Kubhist2_diachronic_embeddings.zip`.

3.2 FORMAT NAMES AND VERSIONS

The data is shared as a ZIP file containing gensim binary files (`.ft` for fastText models, `.w2v` for word2vec models) and Python (`.py`) scripts. For the larger models, matrices and vectors are stored separately as `NumPy` arrays (Oliphant, 2006, `.npy`). Given the relatively large size of the archive, we recommend that Windows users decompress the file (right-click → 'Extract all') instead of double-clicking it. The directory structure is as follows:

```
ROOT/
    README.md
    code/
        *.py files
        requirements.txt
    fasttext/
        incremental/
           *.ft files
           *.npy files
        indep/
           *.ft files
           *.npy files
    word2vec/
        incremental/
           *.w2v files
           *.npy files
        indep/
           *.w2v files
           *.npy files
```

The `README.md` file contains basic information about this release, while the `code/requirements.txt` file contains a list of required Python packages to run the provided code.

3.3 CREATION DATES

The models were trained on 2020-09-15.

3.4 DATASET CREATORS

The original data was scanned and OCRed by the National Library of Sweden. It consists of Swedish newspapers from all parts of Sweden. It has since been run through the Sparv annotation pipeline by Martin Hammarstedt at Språkbanken Text. As described in Subsection 2.2 the authors of this paper have extracted the text from the original XML, processed it, and trained the models.

3.5 LANGUAGE

The diachronic word embedding models have been trained on Swedish data. The variable names in the accompanying Python code and documentation are in English.

3.6 LICENSE

3.7 REPOSITORY NAME

The data is released on Zenodo, and named 'A collection of Swedish diachronic word embedding models trained on historical newspaper data.' A link to the Zenodo repository as well as a description of the dataset are also available on the Språkbanken Text website, along with other resources.[5]

3.8 PUBLICATION DATE

The data was released on Zenodo on 2020/12/2.

4 REUSE POTENTIAL

We believe that this data release can be re-used by a relatively large community of researchers from different fields. This fact is reinforced by the release of documented code – bypassing the need for advanced technical skills, which is one of the key challenges in interdisciplinary collaborations (McGillivray et al., 2020).

Since the models span several decades, they present an interesting view of words over time, useful for researchers interested in diachronic studies such as culturomics (Michel et al., 2011), semantic change (see Tahmasebi et al. (2018); Kutuzov et al. (2018), for overviews), historical research (van Eijnatten & Ros, 2019; Hengchen et al., 2021a; Marjanen et al., 2020), etc. They also can be further fed as input to more complex neural networks tackling downstream tasks aimed at historical data such as OCR post-correction (Hämäläinen & Hengchen, 2019; Duong et al., 2020) or more linguistics-oriented problems (Budts, 2020). Since we release the whole models and not solely the learned vectors, these models can be further trained and specialised, or used by NLP researchers to compare different space alignment procedures.

ACKNOWLEDGEMENTS

The authors would like to thank the researchers and developers at Språkbanken Text for creating and releasing the processed XML for Kubhist 2 used as a basis for training the embeddings. Additional thanks go to Haim Dubossarsky and Dominik Schlechtweg for their insightful comments on the robustness of the pipeline that trains word embeddings. Finally, our sincere thanks go to both reviewers for their useful remarks.

FUNDING STATEMENT

This work has been funded in part by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184), and *Nationella Språkbanken* (the Swedish National Language Bank) – jointly funded by the Swedish Research Council (2018–2024; dnr 2017-00626) and its 10 partner institutions, to Nina Tahmasebi. The authors wish to thank the Department of Swedish at the University of Gothenburg for providing financial support for the Open Access fee.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Simon Hengchen *orcid.org/0000-0002-8453-7221*
Språkbanken Text, Department of Swedish, University of Gothenburg, SE
Nina Tahmasebi *orcid.org/0000-0003-1688-1845*
Språkbanken Text, Department of Swedish, University of Gothenburg, SE

5 *https://spraakbanken.gu.se/en/resources*.

REFERENCES

Adesam, Y., Dannélls, D., & **Tahmasebi, N.** (2019). Exploring the quality of the digital historical newspaper archive KubHist. In *Proceedings of the 2019 DHN conference* (pp. 9–17).

Antoniak, M., & **Mimno, D.** (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119. DOI: *https://doi.org/10.1162/tacl_a_00008*

Bojanowski, P., Grave, E., Joulin, A., & **Mikolov, T.** (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. DOI: *https://doi.org/10.1162/tacl_a_00051*

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & **Schumacher, A.** (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University* (pp. 17–18).

Borin, L., Forsberg, M., & **Roxendal, J.** (2012). Korp — the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 474–478). Istanbul, Turkey: European Language Resources Association (ELRA). URL: *http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf*

Budts, S. (2020). A connectionist approach to analogy. On the modal meaning of periphrastic do in Early Modern English. *Corpus Linguistics and Linguistic Theory*, *1*(ahead-of-print). DOI: *https://doi.org/10.1515/cllt-2019-0080*

Cook, P., & **Stevenson, S.** (2010). Automatically identifying changes in the semantic orientation of words. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Dubossarsky, H., Hengchen, S., Tahmasebi, N., & **Schlechtweg, D.** (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* (pp. 457–470). Florence, Italy: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/P19-1044*

Dubossarsky, H., Weinshall, D., & **Grossman, E.** (2017). Outta control: Laws of semantic change and inherent biases inword representation models. In *Proceedings of the 2017 Conference onEmpirical Methods in Natural Language Processing* (pp. 1136–1145). Copenhagen, Denmark: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/D17-1118*

Duong, Q., Hämäläinen, M., & **Hengchen, S.** (2020). An unsupervised method for OCR post-correction and spelling normalisation for Finnish. *arXiv preprint arXiv:2011.03502*.

Hengchen, S., Ros, R., Marjanen, J., & **Tolonen, M.** (2021a). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*.

Hengchen, S., Tahmasebi, N., Schlechtweg, D., & **Dubossarsky, H.** (2021b). Challenges for computational lexical semantic change. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, & S. Hengchen (Eds.), *Computational Approaches to Semantic Change, Language Variation, chap. 11*. Berlin: Language Science Press.

Hill, M. J., & **Hengchen, S.** (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, *34*(4), 825–843. DOI: *https://doi.org/10.1093/llc/fqz024*

Hämäläinen, M., & **Hengchen, S.** (2019). From the paft to the fiiture: a fully automatic NMT and word embeddings method for OCR post-correction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 431–436). DOI: *https://doi.org/10.26615/978-954-452-056-4_051*

Kaiser, J., Schlechtweg, D., Papay, S., & **Schulte im Walde, S.** (2020). IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Barcelona, Spain: Association for Computational Linguistics.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & **Petrov, S.** (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (pp. 61–65). DOI: *https://doi.org/10.3115/v1/W14-2517*

Kulkarni, V., Al-Rfou, R., Perozzi, B., & **Skiena, S.** (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 625–635). International World Wide Web Conferences Steering Committee. DOI: *https://doi.org/10.1145/2736277.2741627*

Kutuzov, A., Øvrelid, L., Szymanski, T., & **Velldal, E.** (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397).

Marjanen, J., Kurunmäki, J., Pivovarova, L., & **Zosa, E.** (2020). The expansion of isms, 1820–1917: Data-driven analysis of political language in digitized newspaper collections. *Journal of Data Mining and Digital Humanities*.

McGillivray, B., Alex, B., Ames, S., Armstrong, G., Beavan, D., Ciula, A., Colavizza, G., Cummings, J., De Roure, D., Farquhar, A., et al. (2020). The challenges and prospects of the intersection of humanities and data science: A white paper from The Alan Turing Institute.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. DOI: *https://doi.org/10.1126/science.1199644*

Mikolov, T., Chen, K., Corrado, G., & **Dean, J.** (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Oliphant, T. E. (2006). *A guide to NumPy*, vol. 1. USA: Trelgol Publishing.

Řehůřek, R., & **Sojka, P.** (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. *http://is.muni.cz/publication/884893/en*

Rosenfeld, A., & **Erk, K.** (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 474–484). New Orleans, Louisiana. DOI: *https://doi.org/10.18653/v1/N18-1044*

Schlechtweg, D., Hätty, A., del Tredici, M., & **Schulte im Walde, S.** (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* (pp. 732–746). Florence, Italy: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/P19-1072*

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & **Tahmasebi, N.** (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.

Schlechtweg, D., & **Schulte im Walde, S.** (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, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*. DOI: *https://doi.org/10.17617/2.3190925*

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & **McGillivray, B.** (2019). Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 66–76). Hong Kong, China: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/D19-1007*

Språkbanken. (2019). *The Kubhist Corpus, v2*. Department of Swedish, University of Gothenburg. URL: *https://spraakbanken.gu.se/korp/?mode=kubhist*, version downloaded in 2019.

Tahmasebi, N., Borin, L., & **Jatowt, A.** (2018). Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.

Tripodi, R., Warglien, M., Levis Sullam, S., & **Paci, D.** (2019). Tracing antisemitic language through diachronic embedding projections: France 1789–1914. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (pp. 115–125). Florence, Italy: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/W19-4715*

van Eijnatten, J., & **Ros, R.** (2019). The eurocentric fallacy. A digital approach to the rise of modernity, civilization and Europe. *International Journal for History, Culture and Modernity*, *7*. DOI: *https://doi.org/10.18352/hcm.580*

van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., & **Colavizza, G.** (2020). Assessing the impact of OCR quality on downstream NLP tasks. In *ICAART, 1*, 484–496. DOI: *https://doi.org/10.5220/0009169004840496*

TO CITE THIS ARTICLE:
Hengchen, S., & Tahmasebi, N. (2021). A Collection of Swedish Diachronic Word Embedding Models Trained on Historical Newspaper Data. *Journal of Open Humanities Data*, 7: 2, pp. 1–7. DOI: *https://doi.org/10.5334/johd.22*

Published: 27 January 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.