

Detecting semantic shifts in biomedical literature through an intra-year and inter-year approach

This manuscript ([permalink](#)) was automatically generated from [greenelab/word_lapse_manuscript@9b53e46](#) on May 31, 2022.

Authors

- **David Nicholson**

 [0000-0003-0002-5761](#) ·  [danich1](#) ·  [dnicholson329](#)

Genomics and Computational Biology · Funded by Grant XXXXXXXX

- **Faisal Alquaddoomi**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [janeroe](#)

Center for Health Artificial Intelligence (CHAI)

- **Vincent Rubinetti**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [janeroe](#)

Center for Health Artificial Intelligence (CHAI)

- **Casey S. Greene**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [cgreene](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

Introduction

Language is constantly evolving, and the meaning that we ascribe to words changes over time. For example, the word “nice” was used to mean foolish or innocent back in the 15th-17th century; then, it underwent a positive shift to its current meaning of “pleasant or delightful”[1]. These shifts occur for many reasons. For example, writers may use new metaphors or substitute words for others with similar meanings in a process known as metonymy [1]. Studying these shifts can provide a nuanced understanding of how language adapts to describe our world.

Scientific fields of inquiry also change, sometimes rapidly, as researchers devise and test new hypotheses and applications. For example, the repurposing of the CRISPR-Cas9 system to a pervasive tool for genome editing has altered how we discuss molecular entities. Microbes use this as an immune system to defend against viruses. Scientists repurposed this system for genome editing [2], leading to changes in the use of the term. Science is a field with substantial written communication [3], both via published papers [4] and preprints [5,6]. Examining scientific manuscripts with computational linguistics can reveal longitudinal trends in scientific research.

Studying changes in the use of word meanings is called semantic shift detection. Approaches for semantic shift detection examine time series datasets that capture word usage patterns, both with respect to frequency and structure. Typically, these time series are generated for individual words by training a unique model on text binned by a selected time period [7,8,9]. Methods are then applied to identify “changepoints” where a word’s meaning has changed [11].

Semantic shifts have been examined in many sources. Analysis has included newspapers [12,13,14], books [7], reddit [8], and Twitter [15]. Researchers have examined topics in information retrieval [16], and in biomedicine COVID-19 has been examined multiple times [18]. The amount of open access biomedical literature has dramatically increased in the last two decades, laying the groundwork for the large-scale analysis of semantic shifts in biomedicine.

We examine these semantic shifts in this rapidly growing body of open access text. We include both published papers and preprints in our analysis. We found that novel strategies integrating multiple models for each year sidestepped the challenge of instability in the machine learning models and allowed us to estimate intra- and inter-year variability. We identify semantic changepoints for each token. We examine key cases and provide the full set of research products, including changepoints and machine learning models, as openly licensed tools for the community. We also created a webserver that allows users to analyze tokens of interest on the fly, examining both the most similar terms within a year and temporal trends.

Methods

Biomedical Corpora Examined

Pubtator Central

Pubtator Central is an open-access resource containing annotated abstracts and full-text annotated with entity recognition systems for biomedical concepts [19]. The methods used are TaggerOne [20] to tag diseases, chemicals, and cell line entities, GNormPlus [21] to tag genes, SR4GN [22] to tag

species, and tmVar [23] to tag genetic mutations. We initially downloaded this resource on December 07th, 2021, and processed over 30 million documents. This resource contains documents that date back to the pre-1800s to the year 2021; however, due to the low sample size in early years, we only used documents published from 2000 to 2021. The resource was subsequently updated with documents from 2021. We also downloaded a later version on March 09th, 2022, and merged both versions using each document's doc_id field to produce the corpus used in this analysis. We divided documents by publication year and then preprocessed each using spacy's en_core_web_sm model [24]. We replaced each tagged word or phrase with its corresponding entity type and entity id for every sentence that contained an annotation. Then, we used spacy to break sentences into individual tokens and normalized each token to its root form via lemmatization. After preprocessing, we used every sentence to train multiple natural language models designed to represent words based on their context.

Biomedical Preprints

BioRxiv [5] and MedRxiv [6] are repositories that contain preprints for the life science community. MedRxiv mainly focuses on preprints that mention patient research, while bioRxiv focuses on general biology. We downloaded a snapshot of both resources on March 4th, 2022, using their respective Amazon S3 bucket [25,26]. This snapshot contained 172,868 BioRxiv preprints and 37,517 MedRxiv preprints. These resources allow authors to post multiple versions of a single preprint. To prevent duplication bias, we filtered every preprint to its most recent version and sorted each preprint into its respective posted year. Unlike Pubtator Central, these filtered preprints do not contain any annotations. Therefore, we used TaggerOne [20] to tag every chemical and disease entity and GNormplus [21] to tag every gene and species entity for our preprint set. Once tagged, we used spacy to preprocess every preprint as described in our Pubtator Central section.

Constructing Word Embeddings for Semantic Change Detection

Word2vec [27] is a natural language processing model designed to model words based on their respective neighbors in the form of dense vectors. This suite of models comes in two forms, a skipgram model and a continuous bags of words (CBOW) model. The skipgram model generates these vectors by having a shallow neural network predict a word's neighbors given the word, while the CBOW model predicts the word given its neighbors. We used the CBOW model to construct word vectors for each year. Despite the power of these word2vec models, these models are known to differ both due to randomization within year and year-to-year variability across years [28,29,30,31]. To control for run-to-run variability, we examined both intra-year and inter-year relationships. Each year, we trained ten different CBOW models using the following parameters: vector size of 300, 10 epochs, minimum frequency cutoff of 5, and a window size of 16 for abstracts. Every model has its own unique vector space following training, making it difficult to compare two models without a correction step. We used orthogonal Procrustes [32] to align models. We aligned all trained CBOW models for the Pubtator Central dataset to the first model trained in 2021. Likewise, we aligned all CBOW models for the BioRxiv/MedRxiv dataset to the first model trained in 2021. We used UMAP [33] to visually examine the aligned models. We trained this model using the following parameters: cosine distance metric, random_state of 100, 25 for n_neighbors, a minimum distance of 0.99, and 50 n_epochs.

Detecting semantic changes across time

Once word2vec models are aligned, the next step is to detect semantic change. Semantic change events are often detected through time series analysis [34]. We constructed a time series sequence for every token by calculating its distance within a given year (intra-year) and across each year (inter-year). We used the model pairs constructed from the same year to calculate an intra-

year distance. Then, we calculated the cosine distance between each token and its corresponding counterpart for every generated pair. Cosine distance is a metric bounded between zero and two, where a score of zero means two vectors are the same, and a score of two means both vectors are different. For the inter-year distance, we used the Cartesian product of every model between two years and calculated the distance between tokens in the same way as the intra-year distance. Following both calculations, we combined both metrics by taking the ratio of the average inter-year distance over the average intra-year distance. Through this approach, tokens with high intra-year instability will be penalized and vice-versa for more stable tokens. Along with token distance calculations, it has been shown that including token frequency improves results compared to using distance alone [35]. We calculated token frequency as the ratio of token frequency in the more recent year over the frequency of the previous year. Then, we combined both the frequency and distance ratios to make the final metric.

Following time series construction, we performed change point detection, which is a process that uses statistical techniques to detect abnormalities within a given time series. We used the CUSUM algorithm [11] to detect these abnormalities. This algorithm uses a rolling sum of the differences between two timepoints and checks whether the sum is greater than a threshold. A changepoint is considered to have occurred if the sum is greater than a threshold. We used the 99th percentile on every generated timepoint as the threshold. Then, we ran the CUSUM algorithm using a drift of 0 and default settings for all other parameters.

Results

One potential results panel: comparing results with abstract and full text (since you say above full text is not so often used).

1. Figure 1 - Visual to show that alignment works along intra year variation (Procrustes Validation)
2. Umap panel for an individual year to show models are separated
I would start with the simplest possible way that you can show things, and only bring in more complexity like UMAP when necessary. I'm guessing that you would be able to show this with just PCA or similar, unless I'm not thinking correctly.
3. Umap panel to show the same year after models have been aligned
4. Umap panel to show across years (inter-year variation)
 1. might not have this as a figure given the aligned umap results, but we shall see
5. Figure 2 - CUSUM validation
I think you might want another figure here that shows how you can use the variability in distance to address the challenges of variable training set size / model uncertainty. Let's see how things shake out as you're writing.
6. Table of found timepoint changes using this algorithm
7. Highlight pandemic as positive control
8. Nearest Neighbors upset plot ^^
9. Might look into lung cancer or other form of cancer results
10. Figure 3 - Website walkthrough for the work done here
11. Website screenshots
The figures that make up the website might be nice to show/explain individually - not as screenshots but as actual figures (remember the PLOS situation where using a screenshot was an issue we had to work through).
12. Basically a walkthrough of the website and how a user can operate the web resource (similar to preprint similarity search)

Discussion and Conclusion

1. We provided a website resource to allow users to see token association changes
2. Word2vec is unstable and we implemented an approach to account for that variation
3. Constructs groundwork for future research into token changes
4. Will implement biorxiv and other preprint resources as a next step.

References

1. **Semantic Change**
Elizabeth Closs Traugott
Oxford Research Encyclopedia of Linguistics (2017-03-29) <https://doi.org/gp574c>
DOI: [10.1093/acrefore/9780199384655.013.323](https://doi.org/10.1093/acrefore/9780199384655.013.323)
2. **A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity**
Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, Emmanuelle Charpentier
Science (2012-08-17) <https://doi.org/f22dgd>
DOI: [10.1126/science.1225829](https://doi.org/10.1126/science.1225829) · PMID: [22745249](https://pubmed.ncbi.nlm.nih.gov/22745249/) · PMCID: [PMC6286148](https://pubmed.ncbi.nlm.nih.gov/PMC6286148/)
3. **Scientific communication pathways: an overview and introduction to a symposium**
David F Zaye, WV Metanowski
Journal of Chemical Information and Computer Sciences (1986-05-01) <https://doi.org/bwsxhg>
DOI: [10.1021/ci00050a001](https://doi.org/10.1021/ci00050a001)
4. **PubMed Central: The GenBank of the published literature**
Richard J Roberts
Proceedings of the National Academy of Sciences (2001-01-09) <https://doi.org/bbn9k8>
DOI: [10.1073/pnas.98.2.381](https://doi.org/10.1073/pnas.98.2.381) · PMID: [11209037](https://pubmed.ncbi.nlm.nih.gov/11209037/) · PMCID: [PMC33354](https://pubmed.ncbi.nlm.nih.gov/PMC33354/)
5. **bioRxiv: the preprint server for biology**
Richard Sever, Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos, John R Inglis
Cold Spring Harbor Laboratory (2019-11-06) <https://doi.org/ggc46z>
DOI: [10.1101/833400](https://doi.org/10.1101/833400)
6. **Medical preprint server debuts**
Jocelyn Kaiser
Science (2019-06-05) <https://doi.org/gpxkkf>
DOI: [10.1126/science.aay2933](https://doi.org/10.1126/science.aay2933)
7. **Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change**
William L Hamilton, Jure Leskovec, Dan Jurafsky
arXiv (2018-10-26) <https://arxiv.org/abs/1605.09096>
8. **Multi-Label Zero-Shot Learning via Concept Embedding**
Ubai Sandouk, Ke Chen
arXiv <https://arxiv.org/abs/1606.00282>
9. **Deep Neural Models of Semantic Shift**
Alex Rosenfeld, Katrin Erk
Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (2018)
<https://doi.org/gp574f>
DOI: [10.18653/v1/n18-1044](https://doi.org/10.18653/v1/n18-1044)
10. **Bayesian Online Changepoint Detection**
Ryan Prescott Adams, David JC MacKay
arXiv (2007-10-22) <https://arxiv.org/abs/0710.3742>
11. **Adaptive filtering and change detection**

Fredrik Gustafsson, Fredrik Gustafsson
CiteSeer (2000)

12. **Tracing armed conflicts with diachronic word embedding models**
Andrey Kutuzov, Erik Velldal, Lilja Øvrelid
Proceedings of the Events and Stories in the News Workshop (2017) <https://doi.org/ghx5gj>
DOI: [10.18653/v1/w17-2705](https://doi.org/10.18653/v1/w17-2705)
13. **Words are Malleable: Computing Semantic Shifts in Political and Media Discourse**
Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, Jaap Kamps
arXiv (2017-11-16) <https://arxiv.org/abs/1711.05603>
14. **Reading Between the Lines: Prediction of Political Violence Using Newspaper Text**
HANNES MUELLER, CHRISTOPHER RAUH
American Political Science Review (2017-12-14) <https://doi.org/gdj77d>
DOI: [10.1017/s0003055417000570](https://doi.org/10.1017/s0003055417000570)
15. **Statistically Significant Detection of Linguistic Change**
Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena
arXiv (2014-11-13) <https://arxiv.org/abs/1411.3315>
16. **Semantic word shifts in a scientific domain**
Baitong Chen, Ying Ding, Feicheng Ma
Scientometrics (2018-07-13) <https://doi.org/gd7bd7>
DOI: [10.1007/s11192-018-2843-2](https://doi.org/10.1007/s11192-018-2843-2)
17. **Semantic Changepoint Detection for Finding Potentially Novel Research Publications**
Bhavish Dinakar, Mayla R Boguslav, Carsten Görg, Deendayal Dinakarpanian
Biocomputing 2021 (2020-11) <https://doi.org/gp574d>
DOI: [10.1142/9789811232701_0011](https://doi.org/10.1142/9789811232701_0011)
18. **How COVID-19 Is Changing Our Language : Detecting Semantic Shift in Twitter Word Embeddings**
Yanzhu Guo, Christos Xypolopoulos, Michalis Vazirgiannis
arXiv (2021-02-17) <https://arxiv.org/abs/2102.07836>
19. **PubTator central: automated concept annotation for biomedical full text articles**
Chih-Hsuan Wei, Alexis Allot, Robert Leaman, Zhiyong Lu
Nucleic Acids Research (2019-05-22) <https://doi.org/ggzfsc>
DOI: [10.1093/nar/gkz389](https://doi.org/10.1093/nar/gkz389) · PMID: [31114887](https://pubmed.ncbi.nlm.nih.gov/31114887/) · PMCID: [PMC6602571](https://pubmed.ncbi.nlm.nih.gov/PMC6602571/)
20. **TaggerOne: joint named entity recognition and normalization with semi-Markov Models**
Robert Leaman, Zhiyong Lu
Bioinformatics (2016-06-09) <https://doi.org/f855dg>
DOI: [10.1093/bioinformatics/btw343](https://doi.org/10.1093/bioinformatics/btw343) · PMID: [27283952](https://pubmed.ncbi.nlm.nih.gov/27283952/) · PMCID: [PMC5018376](https://pubmed.ncbi.nlm.nih.gov/PMC5018376/)
21. **GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
BioMed Research International (2015) <https://doi.org/gb85jb>
DOI: [10.1155/2015/918710](https://doi.org/10.1155/2015/918710) · PMID: [26380306](https://pubmed.ncbi.nlm.nih.gov/26380306/) · PMCID: [PMC4561873](https://pubmed.ncbi.nlm.nih.gov/PMC4561873/)
22. **SR4GN: A Species Recognition Software Tool for Gene Normalization**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
PLoS ONE (2012-06-05) <https://doi.org/gpq498>

DOI: [10.1371/journal.pone.0038460](https://doi.org/10.1371/journal.pone.0038460) · PMID: [22679507](https://pubmed.ncbi.nlm.nih.gov/22679507/) · PMCID: [PMC3367953](https://pubmed.ncbi.nlm.nih.gov/PMC3367953/)

23. **tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine**
Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, Zhiyong Lu
Bioinformatics (2017-09-01) <https://doi.org/gbzsmc>
DOI: [10.1093/bioinformatics/btx541](https://doi.org/10.1093/bioinformatics/btx541) · PMID: [28968638](https://pubmed.ncbi.nlm.nih.gov/28968638/) · PMCID: [PMC5860583](https://pubmed.ncbi.nlm.nih.gov/PMC5860583/)
24. **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**
Matthew Honnibal, Ines Montani
(2017)
25. **Machine access and text/data mining resources | bioRxiv** <https://www.biorxiv.org/tdm>
26. **Machine access and text/data mining resources | medRxiv** <https://www.medrxiv.org/tdm>
27. **Efficient Estimation of Word Representations in Vector Space**
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-09-10) <https://arxiv.org/abs/1301.3781>
28. **Factors Influencing the Surprising Instability of Word Embeddings**
Laura Wendlandt, Jonathan K Kummerfeld, Rada Mihalcea
arXiv (2020-06-05) <https://arxiv.org/abs/1804.09692>
DOI: [10.18653/v1/n18-1190](https://doi.org/10.18653/v1/n18-1190)
29. **Stability of Word Embeddings Using Word2Vec**
Mansi Chugh, Peter A Whigham, Grant Dick
AI 2018: Advances in Artificial Intelligence (2018) <https://doi.org/gpxkkc>
DOI: [10.1007/978-3-030-03991-2_73](https://doi.org/10.1007/978-3-030-03991-2_73)
30. **Evaluating the Stability of Embedding-based Word Similarities**
Maria Antoniak, David Mimno
Transactions of the Association for Computational Linguistics (2018-12) <https://doi.org/gf39k8>
DOI: [10.1162/tacl_a_00008](https://doi.org/10.1162/tacl_a_00008)
31. **Predicting Word Embeddings Variability**
Benedicte Pierrehjean, Ludovic Tanguy
Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (2018)
<https://doi.org/gh6qpc>
DOI: [10.18653/v1/s18-2019](https://doi.org/10.18653/v1/s18-2019)
32. **A generalized solution of the orthogonal procrustes problem**
Peter H Schönemann
Psychometrika (1966-03) <https://doi.org/dx77sz>
DOI: [10.1007/bf02289451](https://doi.org/10.1007/bf02289451)
33. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
Leland McInnes, John Healy, James Melville
arXiv (2020-09-21) <https://arxiv.org/abs/1802.03426>
34. **Statistically Significant Detection of Linguistic Change**
Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena
Proceedings of the 24th International Conference on World Wide Web (2015-05-18)
<https://doi.org/ghcv6k>
DOI: [10.1145/2736277.2741627](https://doi.org/10.1145/2736277.2741627)

35. **Improving semantic change analysis by combining word embeddings and word frequencies**

Adrian Englhardt, Jens Willkomm, Martin Schäler, Klemens Böhm

International Journal on Digital Libraries (2019-05-20) <https://doi.org/gpxkkd>

DOI: [10.1007/s00799-019-00271-6](https://doi.org/10.1007/s00799-019-00271-6)