# Changing word meanings in biomedical literature reveal pandemics and new technologies

*This manuscript ([permalink](#)) was automatically generated from [greenelab/word_lapse_manuscript@79360d0](#) on April 5, 2023.*

## Authors

- **David N. Nicholson**
  0000-0003-0002-5761 · danich1 · dnicholson329
  Genomics and Computational Biology Program, University of Pennsylvania, Philadelpia, PA, USA · Funded by The Gordon and Betty Moore Foundation, GBMF4552; The National Human Genome Research Institute, R01 HG010067

- **Faisal Alquaddoomi**
  0000-0003-4297-8747 · falquaddoomi
  Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA; Center for Health Artificial Intelligence (CHAI), University of Colorado School of Medicine, Aurora, CO, USA · Funded by The Gordon and Betty Moore Foundation, GBMF4552; The National Human Genome Research Institute, R01 HG010067

- **Vincent Rubinetti**
  · vincerubinetti
  Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA; Center for Health Artificial Intelligence (CHAI), University of Colorado School of Medicine, Aurora, CO, USA · Funded by The Gordon and Betty Moore Foundation, GBMF4552; The National Human Genome Research Institute, R01 HG010067

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · greenescientist
  Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA; Center for Health Artificial Intelligence (CHAI), University of Colorado School of Medicine, Aurora, CO, USA · Funded by The Gordon and Betty Moore Foundation, GBMF4552; The National Human Genome Research Institute, R01 HG010067

To whom correspondence should be addressed:

```
CSG: casey.s.greene@cuanschutz.edu
```

## Abstract

While we often think of words as having a fixed meaning that we use to describe a changing world, words are also dynamic and changing. Scientific research can also be remarkably fast-moving, with new concepts or approaches rapidly gaining mind share. We examined scientific writing, both preprint and pre-publication peer-reviewed text, to identify terms that have changed and examine their use. One particular challenge that we faced was that the shift from closed to open access publishing meant that the size of available corpora changed by over an order of magnitude in the last two decades. We developed an approach to evaluate semantic shift by accounting for both intra- and inter-year variability using multiple integrated models. Using this strategy and examining year-by-year changes revealed thousands of change points in both corpora. We found change points for tokens including 'cas9', 'pandemic', and 'sars' among many others. The consistent change-points between pre-publication peer-reviewed and preprinted text were largely related to the COVID-19 pandemic. We developed a web app for exploration (https://greenelab.github.io/word-lapse/) that enables users to investigate individual terms. To our knowledge, this analysis is the first to examine semantic shift in biomedical preprints and pre-publication peer-reviewed text, and it lays the foundation for future work to examine how terms acquire new meaning and the extent to which that process is encouraged or discouraged by peer review.

**Keywords**: Linguistic shift, pandemic, software, novelty

# Introduction

Language constantly evolves, and the meaning we ascribe to words changes over time. 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, repurposing 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.

The task of examining if words change their meanings is called semantic shift detection. This process involves capturing word usage patterns such as frequency and structure over a set period of time [7]. Once these patterns have been captured, the last step is generating a time series showing possible shift events, commonly called changepoints [9]. Scientists have used this approach to discover many changepoints within publically available English corpora [10,11,12,13,14]. These discoveries included semantic changes like the meaning of awful shifting from majestic to horrible [15]. Along with individual discoveries, scientists have identified global patterns that semantic shifts can follow [15,16]. One pattern discovered is that words with similar meanings, i.e., synonyms, tend to change over time and undergo similar changes [16]. The other patterns are that words undergo semantic shifts inversely proportional to their frequency, and words with multiple meanings will have higher rates of change [15]. The majority of these discoveries have been within regular English text. There has been an effort to investigate if any of these patterns are upheld within biomedical literature [17]. The researchers only found strong evidence of words that change meaning inversely scales to their usage

frequency [18]. Despite the conflicting evidence for these change patterns, biomedical words and concepts change over time.

Most recent work to detect semantic shifts has been outside of biomedical corpora. Common mediums include newspapers [19,20,21], books [15], reddit [22], and Twitter [23]. Researchers have examined topics in information retrieval [24], and in biomedicine COVID-19 has been examined multiple times [25,26,27]. Also, researchers have examined how drugs and disease pairs changed in their usage over time [18]. Despite these discoveries, 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. There remains a gap in uncovering semantic change on a global scale within published and preprint works and this work takes a deeper dive.

To fill this gap, we sought to uncover instances of semantic shifts within the rapidly growing body of open-access texts, published papers, and preprints. We took a novel approach that integrates multiple models to account for the challenge of instability in the machine learning models trained across various years. Using this approach, we identified semantic changepoints for each token and examine key changepoint cases. We provided the full set of research products, including changepoints and machine learning models, as openly licensed tools for the community. Lastly, we created a web server that allows users to analyze tokens of interest on the fly with respect to 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 [28]. The methods used are TaggerOne [29] to tag diseases, chemicals, and cell line entities, GNormPlus [30] to tag genes, SR4GN [31] to tag species, and tmVar [32] to tag genetic mutations. We initially downloaded this resource on December 07th, 2021, and processed over 30 million documents. This resource contains documents from the pre-1800s to 2021; however, due to the low sample size in the 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 [33]. 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.
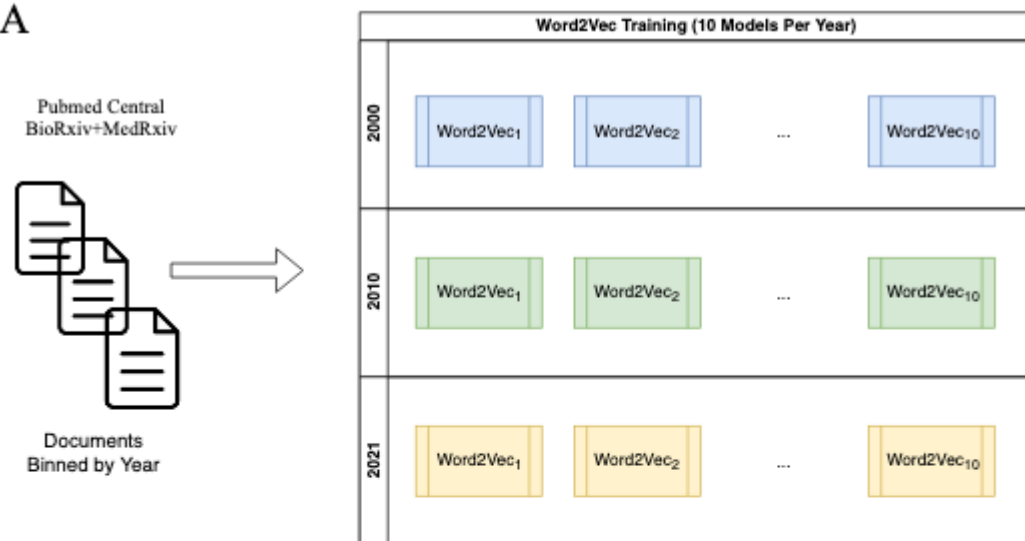
### Biomedical Preprints

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 [34,35]. This snapshot contained 172,868 BioRxiv preprints and 37,517 MedRxiv preprints. These resources allow authors to post multiple versions of a single preprint. We filtered every preprint to its most recent version to prevent duplication bias and sorted each preprint into its

respective posted year. Unlike Pubtator Central, these filtered preprints do not contain any annotations. Therefore, we used TaggerOne [29] to tag every chemical and disease entity and GNormplus [30] 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.
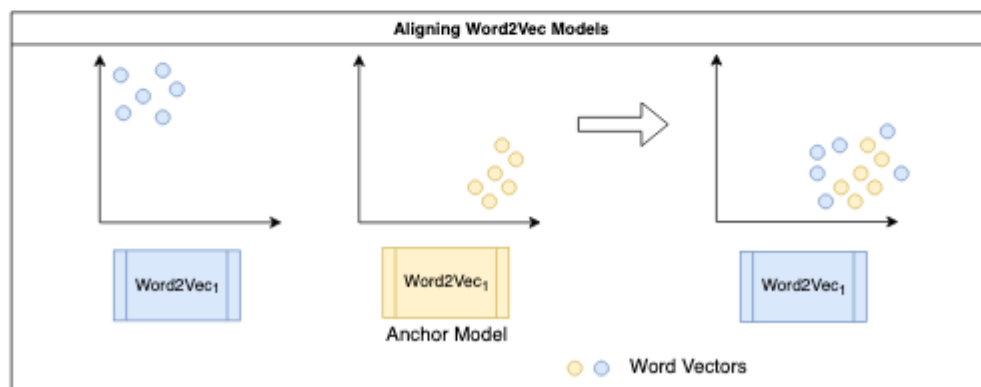
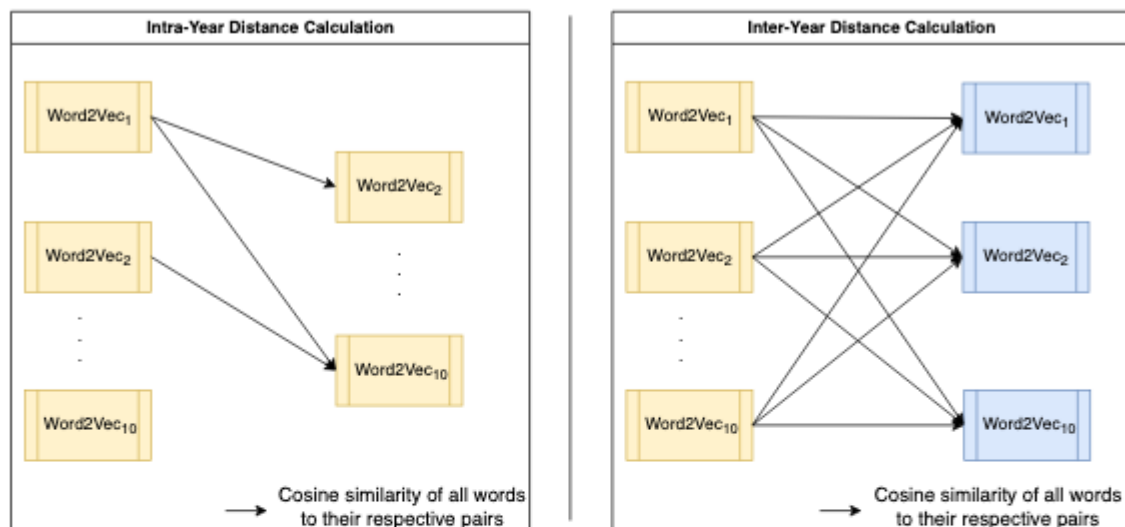## Constructing Word Embeddings for Semantic Change Detection

**A** Word2Vec Training (10 Models Per Year)

Pubmed Central
BioRxiv+MedRxiv

Documents
Binned by Year

**B** Aligning Word2Vec Models

Word2Vec₁

Word2Vec₁
Anchor Model

Word2Vec₁

○ ○ Word Vectors

**C**

Intra-Year Distance Calculation

Cosine similarity of all words
to their respective pairs

Inter-Year Distance Calculation

Cosine similarity of all words
to their respective pairs

**D**

Changepoint?

Combined Single Metric
for each year pair

$$\frac{interyear_{x,y}}{intrayear_x + intrayear_y} + \frac{freq_x}{freq_y}$$

Time Points for Each Year

**Figure 1:** A. The first step of our data pipeline is where PMCOA papers and BioRxiv/MedRxiv preprints are binned by their respective posting year. Following the binning process, we train ten word2vec models for each year's manuscripts. B. Upon training each individual word2vec model, we align every model onto an anchor model. C. We capture token differences using an intra-year and inter-year approach. Each arrow indicates comparing all tokens from one model with their respective selves in a different model. D. The last step combines the above calculations into a single metric to allow for a time series to be constructed. Once constructed, we use a statistical technique to autodetect the presence of a changepoint.

Word2vec [36] 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 due to randomization within a year and year-to-year variability across years [37,38,39,40]. 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 (Figure 1A). 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 [41] to align models (Figure 1B). 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 [42] 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 [10]. 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) (Figure 1C). 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. This approach will penalize tokens with high intra-year instability 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 [43]. 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 the frequency with distance ratios to make the final metric (Figure 1D).

Following time series construction, we performed changepoint detection, which is a process that uses statistical techniques to detect abnormalities within a given time series (Figure 1D). We used the CUSUM algorithm [9] 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 exceeds 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

## Models can be aligned and compared within and between years

We examined how the usage of tokens in biomedical text changes over time. Our evaluation was derived from machine learning models designed to predict the actual token given a portion of its surrounding tokens. Each token was represented as a vector in a coordinate space constructed by these models. However, training these models is stochastic, which results in arbitrary coordinate spaces. Model alignment is essential in allowing word2vec models to be compared [44,45]. Before alignment, each model has its own unique coordinate space (Figure 2A), and each word is represented within that space (Figure 2B). Alignment projects every model onto a shared coordinate space (Figure 2C), enabling direct token comparison. We randomly selected 100 tokens to confirm that alignment worked as expected. In aligned models, tokens in the global space were more similar to themselves within the year than between years, while identical tokens in unaligned models were completely distinct (Figure 2D). Local distances were unaffected by alignment (Figure 2D), as token-neighbor distances were unaffected by the alignment procedure.
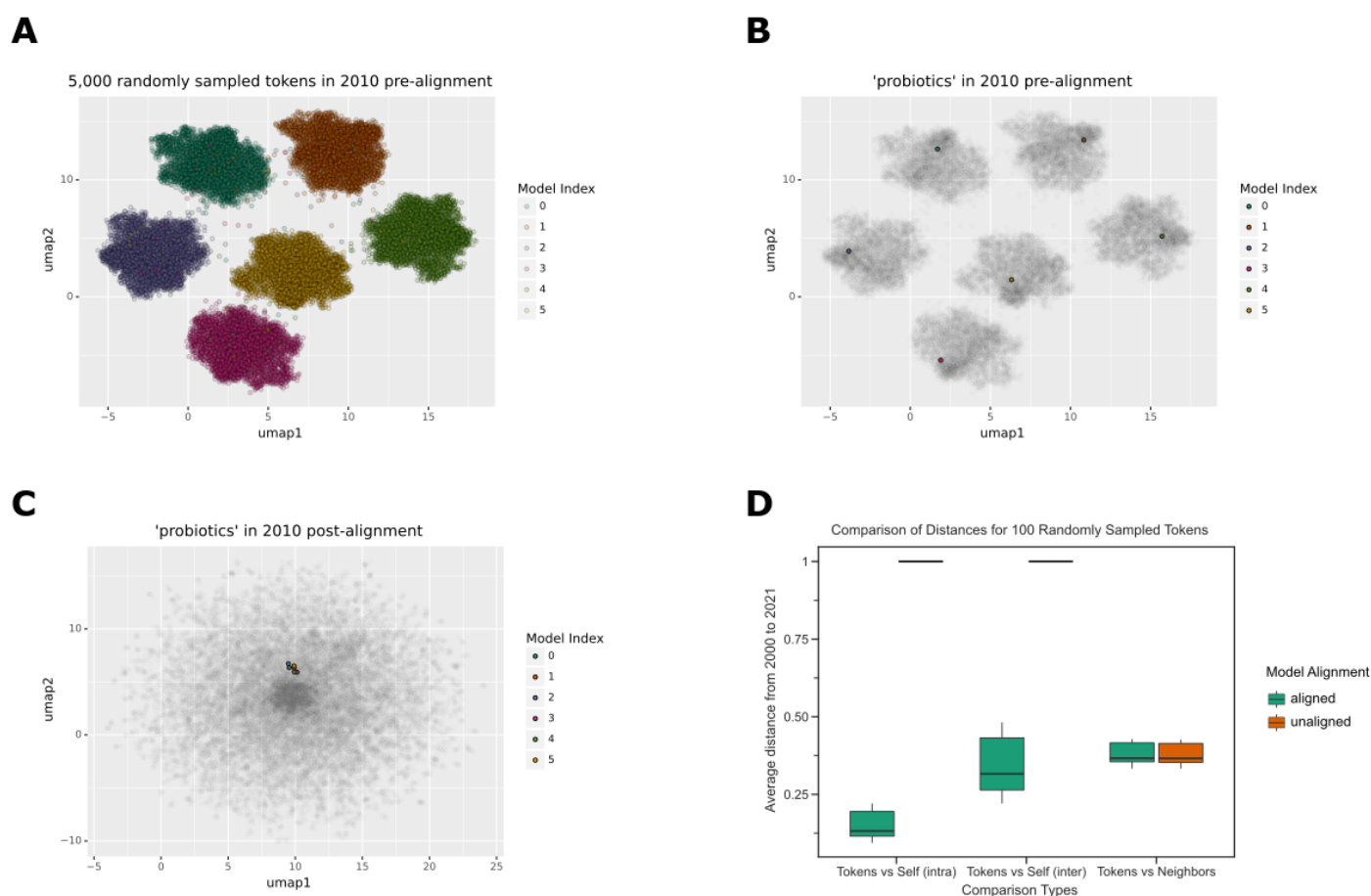


**Figure 2:** A. Without alignment, each word2vec model has its own coordinate space. This is a UMAP visualization of 5000 randomly sampled tokens from 5 distinct Word2Vec models trained on the text published in 2010. Each data point represents a token, and the color represents the respective Word2Vec model. B. We greyed out all tokens except for the token 'probiotics' to highlight that each token appears in its own respective cluster without alignment. C. After the alignment step, the token 'probiotic' is closer in vector space signifying that tokens can be easily compared. D. In the global coordinate space, token distances appear to be vastly different when alignment is not applied. After alignment, token distances become closer; tokens maintain similar distances with their neighbors regardless of alignment. This

boxplot shows the average distance of 100 randomly sampled tokens shared in every year from 2000 to 2021. The x-axis shows the various groups being compared (tokens against themselves via intra-year and inter-year distances and tokens against their corresponding neighbors. The y-axis shows the average distance for every year.

The landscape of biomedical publishing has changed rapidly during the period of our dataset. The texts for our analysis were open-access manuscripts available through PubMed Central. The growth in the amount of available text and the uneven adoption of open-access publishing during the interval studied was expected to induce changes in the underlying machine learning models, making comparisons more difficult. We found that the number of tokens available for model building, i.e., those in PMC OA, increased dramatically during this time (Figure 3A). This was expected to create a pattern where models trained in earlier years were more variable than those from later years simply due to the limited sample size in early years. We aimed to correct for this change in the underlying models by developing a statistic that, instead of using pairwise comparisons of token distances between individual models, integrated multiple models for each year by comparing tokens' intra- and inter-year variabilities. We defined the statistic as the ratio of the average distance between two years over the sum of the average distance within each year respectively.

**A**



**B**



**C**
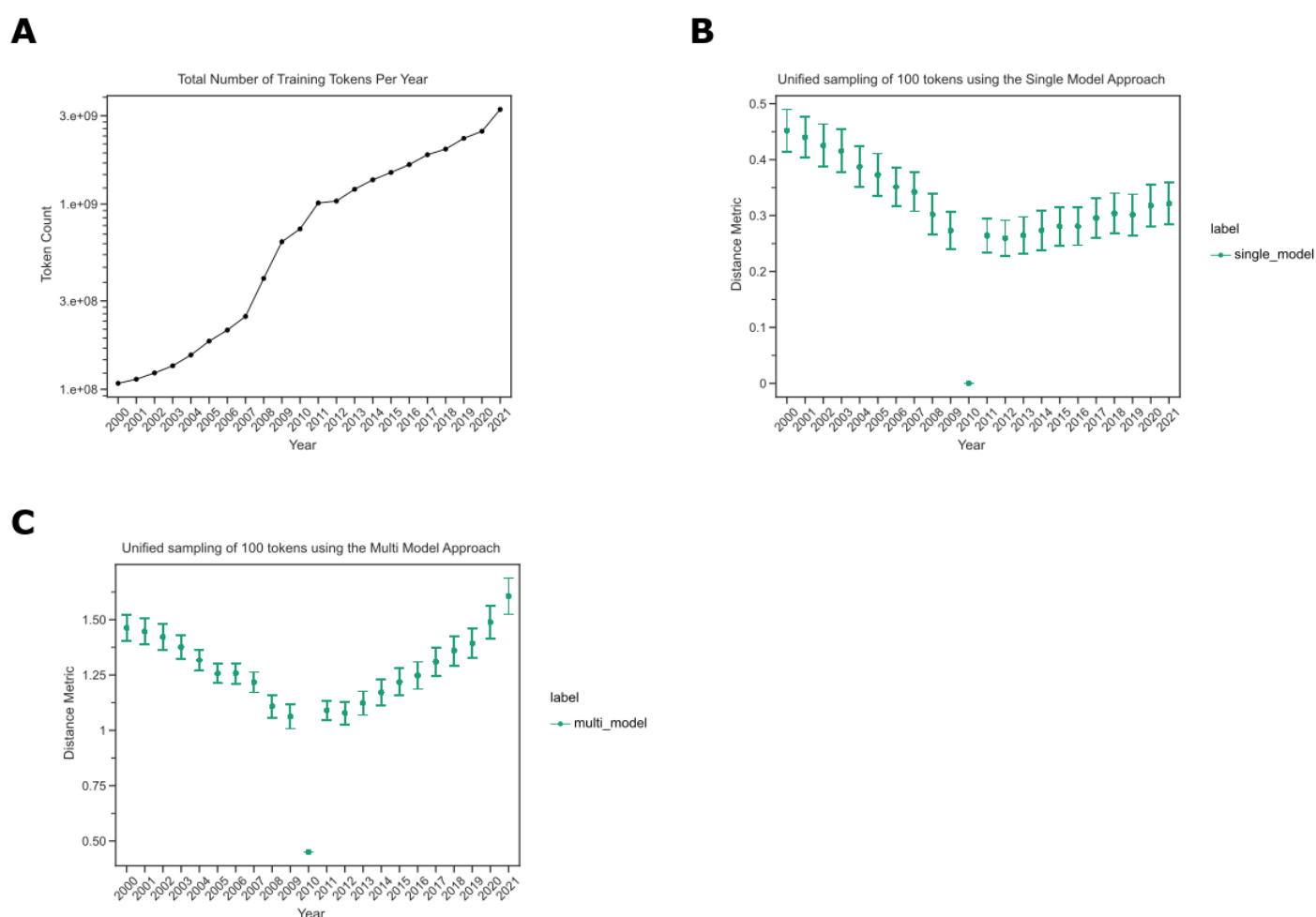


**Figure 3:** A. The number of tokens our models have trained on increases over time. This line plot shows the number of unique tokens our various machine-learning models see. The x-axis depicts the year, and the y-axis shows the token count. B. Earlier years compared to 2010 have greater distances than later years. This confidence interval plot shows the collective distances obtained by sampling 100 tokens present from every year using a single model approach. The x-axis shows a given year, and the y-axis shows the distance metric. C. Later years have a lower intra-distance variability compared to the earlier years. This confidence interval plot shows the collective distances obtained by sampling 100 tokens present from every year using our multi-model approach. The x-axis shows a given year, and the y-axis shows the distance metric.

We expected most tokens to undergo minor changes from year to year, while substantial changes likely suggested model drift instead of true linguistic change. We measured the extent to which tokens differed from themselves using the standard single-model approach and our integrated statistic. We

filtered the token list to only contain tokens present in every year and compared their distance to the midpoint year, 2010, using the single-model and integrated-models strategies. We found that distances tended to be markedly larger in the earliest years, where we expected models to be least stable, using the traditional approach (Figure 3B). The integrated model approach did not display the same pattern in the earliest years (Figure 3C). Both trends reinforce that training on smaller corpora will lead to high variation and that an integrated model strategy is needed [39]. Based on these results, we used the integrated-model strategy to calculate inter-year token distances for the remainder of this work.

## Terms exhibit detectable changes in usage
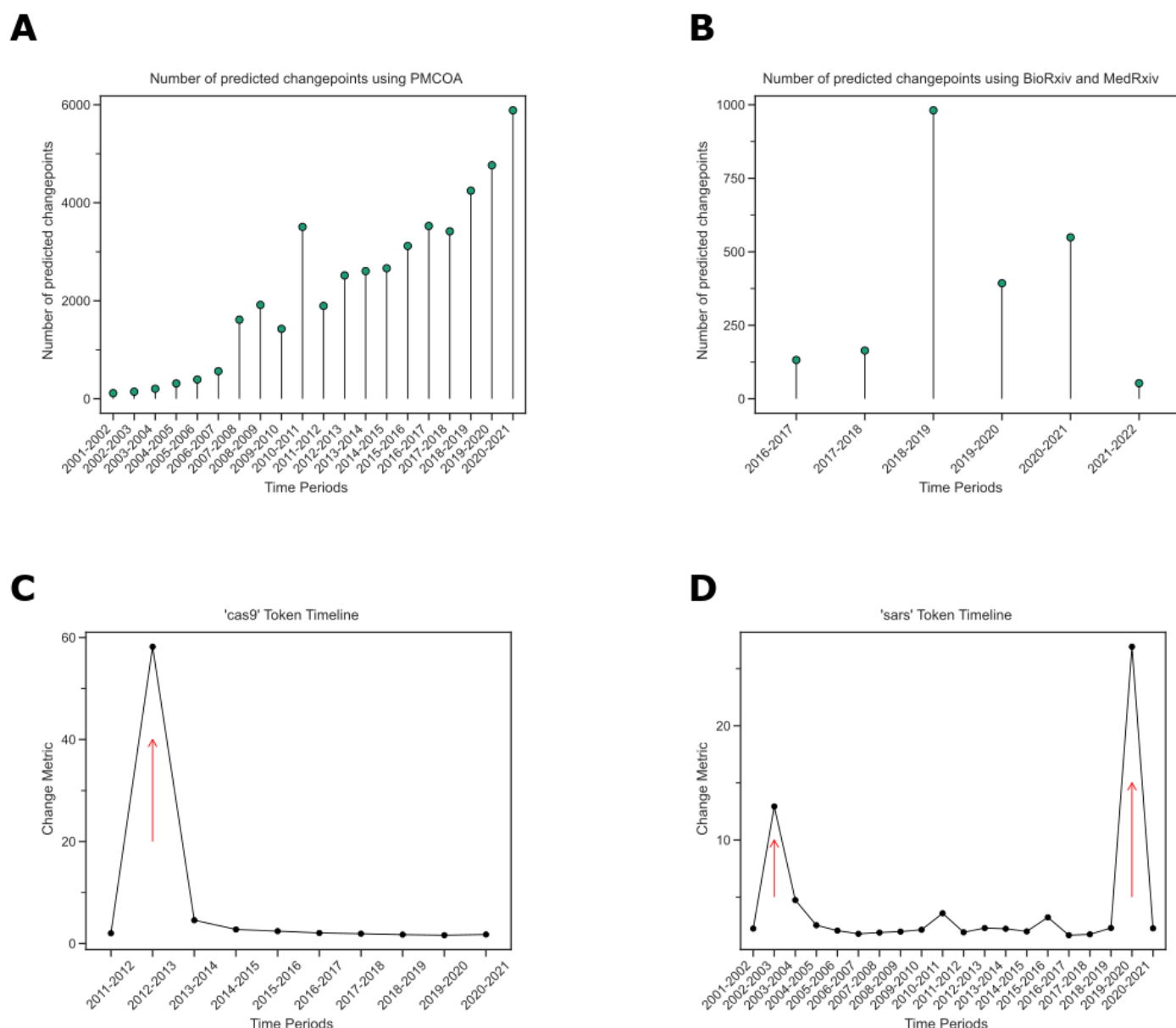
**Figure 4:** A. The number of change points increases over time in PMCOA. The x-axis shows the various time periods, while the y-axis depicts the number of detected changepoints. B. Regarding preprints, the greatest number of change points was during 2018-2019. The x-axis shows the various time periods, while the y-axis depicts the number of detected change points. C. The token 'cas9' was detected to have a changepoint between 2012 and 2013. The x-axis shows the time period since the first appearance of the token, and the y-axis shows the change metric. D. 'sars' has two detected changepoints within the PMCOA corpus. The x-axis shows the time period since the first appearance of the token, and the y-axis shows the change metric.

We next sought to identify tokens that changed during the 2000-2021 interval for the text from PubMed Central's Open Access Corpus (PMCOA) and the 2015-2022 interval for our preprint corpus.

We performed change point detection using the CUSUM algorithm with distances calculated with the integrated-model approach to correct for systematic differences in the underlying corpora. We found 41281 terms with a detected change point from PMCOA and 2266 terms from preprints (Figures 4A and 4B), and the vast majority (38019 for PMCOA and 2260 for preprints) had just a single change-point.

We explored individual change points. We detected one in PMCOA for 'cas9' from 2012 to 2013 (Figure 4C). Before the change point, its closest neighbors were related genetic elements (e.g., 'cas'1-3). After the change point, its closest neighbors became terms related to targeting, sgRNA, and gRNA, as well as other genome editing strategies, 'talen' and 'zfns' (Table 1). For some terms, we detected multiple change points within the studied interval. We detected change points for 'SARS' from 2002 to 2003 and 2019 to 2020 (Figure 4D), consistent with the emergences of SARS-CoV [46] and SARS-CoV-2 [47,48] as observed human pathogens. We found miscellaneous neighbors before each change point, with use consistent with the acronym for Severe Acute Respiratory Syndrome after each (Tables 2 and 3).

Out of all change points, we observed 200 tokens with at least one change point in each corpus. Only 25 of the 200 terms were detected to have simultaneous changes between the preprint and PMCOA corpora. We examined the overlap of detected change points between preprints and published articles. Many of these 25 were related to the COVID-19 pandemic (Supplementary Table S1). The complete set of detected change points is available for further analysis (see Data Availability and Software).

**Table 1:** The fifteen most similar neighbors to the token 'cas9' for the years 2012 and 2013.

| 2012 | 2013 |
| --- | --- |
| cas2 | sgrna |
| crispr1 | talen |
| cas3 | spcas9 |
| cas1 | zfns |
| cas10 | grna |
| crispr3 | zfn |
| tracrrna | dcas9 |
| crispr | nickase |
| csn1 | pcocas9 |
| crispr4 | crispr |
| cas7 | sgrnas |
| cas6e | meganuclease |
| cas4 | tracrrna |
| cse1 | crispri |
| cas6 | crrna |

**Table 2:** The fifteen most similar neighbors to the token 'sars' for the years 2002 and 2003.

| 2002 | 2003 |
| --- | --- |
| qsar | species_227859 |
| herbicidal | mesh_c000657245 |

| 2002 | 2003 |
|---|---|
| antiplasmodial | severe acute respiratory syndrome-related coronavirus (species_694009) |
| arylpiperazine | unidentified human coronavirus (species_694448) |
| a]pyridine | SARS1 (gene_6301) |
| leishmanicidal | ebola virus sp. (species_205488) |
| naphthyridine | pandemic |
| indolo[2,1 | coronavirus infections (mesh_d018352) |
| b]quinazoline-6,12 | coronavirus |
| nematocidal | ebola virus (species_1570291) |
| f]isoxazolo[2,3 | severe acute respiratory syndrome (mesh_d045169) |
| 5-(4 | paramyxovirus |
| cholinephosphotransferase | viruse |
| oxovanadium(iv | drosten |
| catecholase | virologist |

**Table 3:** The fifteen most similar neighbors to the token 'sars' for the years 2019 and 2020.

| 2019 | 2020 |
|---|---|
| g.o. | sar |
| nsp13 | mers |
| 40/367 | cov |
| lissodendoryx | sars-1 |
| lutken | severe acute respiratory syndrome-related coronavirus (species_694009) |
| sarr | coronaviruse |
| sar | middle east respiratory syndrome-related coronavirus (species_1335626) |
| ophiura ophiura (species_72673) | cov. |
| verrill | coronavirus infections (mesh_d018352) |
| hirondelle | mers- |
| kobelt | covs |
| azorean | severe acute respiratory syndrome coronavirus 2 (species_2697049) |
| rusby | severe acute respiratory syndrome (mesh_d045169) |
| d'orbigny | sarscov |
| psychropotes longicauda (species_55639) | sarscov-2 |

# The word-lapse application is an online resource for the manual examination of biomedical tokens
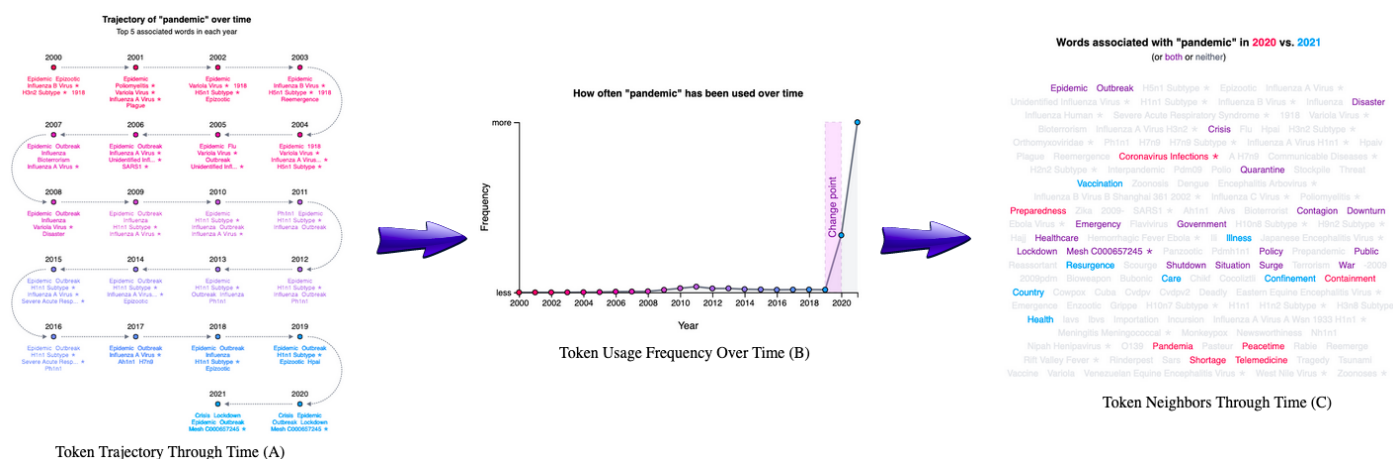
**Figure 5:** A. The trajectory visualization of the token 'pandemic' through time. It starts at the first mention of the token and progresses through each subsequent year. Every data point shows the top five neighbors for the respective token. B. The usage frequency of the token 'pandemic' through time. The x-axis shows the year, and the y-axis shows the frequency for each token. C. A word cloud visualization for the top 25 neighbors for the token 'pandemic' each year. This visualization highlights each neighbor from a particular year and allows for the comparison between two years. Tokens in purple are shared within both years, while tokens in red or blue are unique to their respective year.

We constructed an online application that allows users to examine how tokens change through time. The application supports token input as text strings or as MeSH IDs, Entrez Gene IDs, and Taxonomy IDs. Users might elect to explore the term 'pandemic', for which we detected a change point between 2019 and 2020. Users can examine the token's nearest neighbors through time (Figure 5A). Using the token 'pandemic' as an example, users can observe that 'epidemic' remains similar through time, but taxid:114727 (the H1N1 subtype of influenza) only entered the nearest neighbors with the swine flu pandemic in 2009 and that MeSH:C000657245 (COVID-19) appeared in 2020. The application also shows a frequency chart depicting how often the particular token is used each year (Figure 5B), which can be displayed as a raw count or adjusted by the total size of the corpus. When change points are detected, they are indicated on this panel (Figure 5B). The final visualization shows the union of the nearest 25 neighbors from each year ordered by the number of years that neighbor was present (Figure 5C). This visualization has a comparison function allowing users to examine years' differences. All functionalities are fully supported across the PMCOA and preprint corpora, and users can toggle between the two.

# Discussion and Conclusion

Language is rapidly evolving, and the usage of words changes over time with words assimilating new meanings or associations [1]. Some efforts have been made to study semantic change using biomedical text [25,26,27]; however, no such work has examined the changes evident in both pre-publication peer-reviewed and preprinted biomedical text. We examined semantic change in both open-access biomedical corpora from Pubmed (PMCOA) and bioRxiv/MedRxiv for the 2000-2021 interval. Studies like this one have only become feasible recently with the rapid uptake of open access publishing. To be able to span two decades with very different data availability, we developed a novel statistic that incorporated multiple models using both inter- and intra-year distances. Without the correction, comparing between stable and unstable models is challenging as previously reported [39,49]. Our analysis revealed more than 41,000 different change points, including tokens such as 'cas9', 'pandemic', and 'sars'. Many change points overlapping between PMCOA and preprints were related to COVID-19, indicating that the COVID-19 pandemic has been strong and immediate enough to induce rapid semantic change across both publishing paradigms.

As the amount of preprinted text grows, future work may be able to determine the consistency and time-lag of semantic change between preprint and pre-publication peer-reviewed text - potentially predicting future change in pre-publication peer-reviewed text. We developed a web application to

enable users to investigate individual tokens - automatic approaches that use orthogonal metrics to estimate change point validity would make analysis at scale more straightforward. Furthermore, including corpora such as the arXiv [50] or psyArXiv [51] repositories may reveal consistencies across a broader swath of fields or within-field analyses may reveal the earliest starting points of semantic changes that ultimately sweep through biomedicine.

# Availability of Data and Materials

An online version of this manuscript is available under a Creative Commons Attribution License at https://greenelab.github.io/word_lapse_manuscript/. The source for the research portions of this project is licensed under the BSD-2-Clause Plus Patent at https://github.com/greenelab/biovectors. Our Word Lapse website can be found at https://greenelab.github.io/word-lapse, and the code for the website is available under a BSD-3 Clause at https://github.com/greenelab/word-lapse. Full-text access for the bioRxiv repository is available at https://www.biorxiv.org/tdm. Full-text access for the medRxiv repository is available at https://www.medrxiv.org/tdm. Access to Pubtator Central's Open Access subset is available on NCBI's FTP server at https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/.

# Funding Declaration

# 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

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 · PMID: 22745249 · PMCID: PMC6286148

3.  **Scientific communication pathways: an overview and introduction to a symposium**
    David F Zaye, WV Metanomski
    *Journal of Chemical Information and Computer Sciences* (1986-05-01) https://doi.org/bwsxhg
    DOI: 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 · PMID: 11209037 · PMCID: 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

6.  **Medical preprint server debuts**
    Jocelyn Kaiser
    *Science* (2019-06-05) https://doi.org/gpxkkf
    DOI: 10.1126/science.aay2933

7.  **Diachronic word embeddings and semantic shifts: a survey**
    Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, Erik Velldal
    *arXiv* (2018-06-14) https://arxiv.org/abs/1806.03537

8.  **Bayesian Online Changepoint Detection**
    Ryan Prescott Adams, David JC MacKay
    *arXiv* (2007-10-22) https://arxiv.org/abs/0710.3742

9.  **Adaptive filtering and change detection**
    Fredrik Gustafsson, Fredrik Gustafsson
    *Citeseer* (2000)

10. **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

11. **A framework for analyzing semantic change of words across time**
    Adam Jatowt, Kevin Duh

*IEEE/ACM Joint Conference on Digital Libraries* (2014-09) https://doi.org/gp8zpm
DOI: 10.1109/jcdl.2014.6970173

12. **Understanding semantic change of words over centuries**
Derry Tanti Wijaya, Reyyan Yeniterzi
*Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web* (2011-10-24) https://doi.org/cmxz2v
DOI: 10.1145/2064448.2064475

13. **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

14. **A state-of-the-art of semantic change computation**
XURI TANG
*Natural Language Engineering* (2018-06-18) https://doi.org/gkkswt
DOI: 10.1017/s1351324918000220

15. **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

16. **A computational evaluation of two laws of semantic change.**
Yang Xu, Charles Kemp
*CogSci* (2015)

17. **Tracking word semantic change in biomedical literature**
Erjia Yan, Yongjun Zhu
*International Journal of Medical Informatics* (2018-01) https://doi.org/grwsdh
DOI: 10.1016/j.ijmedinf.2017.11.006 · PMID: 29195709

18. **Exploring Diachronic Changes of Biomedical Knowledge using Distributed Concept Representations**
Gaurav Vashisth, Jan-Niklas Voigt-Antons, Michael Mikhailov, Roland Roller
*Proceedings of the 18th BioNLP Workshop and Shared Task* (2019) https://doi.org/grwsdj
DOI: 10.18653/v1/w19-5037

19. **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

20. **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

21. **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

22. **Detection of emerging drugs involved in overdose via diachronic word embeddings of substances discussed on social media**
Austin P Wright, Christopher M Jones, Duen Horng Chau, R Matthew Gladden, Steven A Sumner
*Journal of Biomedical Informatics* (2021-07) https://doi.org/gp8zph
DOI: 10.1016/j.jbi.2021.103824 · PMID: 34048933

23. **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

24. **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

25. **Semantic Changepoint Detection for Finding Potentially Novel Research Publications**
Bhavish Dinakar, Mayla R Boguslav, Carsten Görg, Deendayal Dinakarpandian
*Biocomputing 2021* (2020-11) https://doi.org/gp574d
DOI: 10.1142/9789811232701_0011

26. **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

27. **Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study**
Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, Satrajit S Ghosh
*Journal of Medical Internet Research* (2020-10-12) https://doi.org/ghm9v2
DOI: 10.2196/22635 · PMID: 32936777 · PMCID: PMC7575341

28. **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 · PMID: 31114887 · PMCID: PMC6602571

29. **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 · PMID: 27283952 · PMCID: PMC5018376

30. **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 · PMID: 26380306 · PMCID: PMC4561873

31. **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 · PMID: 22679507 · PMCID: PMC3367953

32. **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://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860583)

33. **<span class="nocase">spaCy 2</span>: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**
Matthew Honnibal, Ines Montani
(2017)

34. **Machine access and text/data mining resources | bioRxiv** [https://www.biorxiv.org/tdm](https://www.biorxiv.org/tdm)

35. **Machine access and text/data mining resources | medRxiv** [https://www.medrxiv.org/tdm](https://www.medrxiv.org/tdm)

36. **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](https://arxiv.org/abs/1301.3781)

37. **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](https://arxiv.org/abs/1804.09692)
DOI: [10.18653/v1/n18-1190](https://doi.org/10.18653/v1/n18-1190)

38. **Stability of Word Embeddings Using Word2Vec**
Mansi Chugh, Peter A Whigham, Grant Dick
*AI 2018: Advances in Artificial Intelligence* (2018) [https://doi.org/gpxkkc](https://doi.org/gpxkkc)
DOI: [10.1007/978-3-030-03991-2_73](https://doi.org/10.1007/978-3-030-03991-2_73)

39. **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](https://doi.org/gf39k8)
DOI: [10.1162/tacl_a_00008](https://doi.org/10.1162/tacl_a_00008)

40. **Predicting Word Embeddings Variability**
Benedicte Pierrejean, Ludovic Tanguy
*Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (2018)
[https://doi.org/gh6qpc](https://doi.org/gh6qpc)
DOI: [10.18653/v1/s18-2019](https://doi.org/10.18653/v1/s18-2019)

41. **A generalized solution of the orthogonal procrustes problem**
Peter H Schönemann
*Psychometrika* (1966-03) [https://doi.org/dx77sz](https://doi.org/dx77sz)
DOI: [10.1007/bf02289451](https://doi.org/10.1007/bf02289451)

42. **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](https://arxiv.org/abs/1802.03426)

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

44. **Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change**
William L Hamilton, Jure Leskovec, Dan Jurafsky
*arXiv* (2016) [https://doi.org/gp8zpp](https://doi.org/gp8zpp)
DOI: [10.48550/arxiv.1605.09096](https://doi.org/10.48550/arxiv.1605.09096)

45. **DUKweb, diachronic word representations from the UK Web Archive corpus**
Adam Tsakalidis, Pierpaolo Basile, Marya Bazzi, Mihai Cucuringu, Barbara McGillivray
*Scientific Data* (2021-10-15) https://doi.org/gqbkx4
DOI: 10.1038/s41597-021-01047-x · PMID: 34654827 · PMCID: PMC8520005

46. **SARS: clinical virology and pathogenesis**
John NICHOLLS, Xiao-Ping DONG, Gu JIANG, Malik PEIRIS
*Respirology* (2003-11) https://doi.org/cxjwrc
DOI: 10.1046/j.1440-1843.2003.00517.x · PMID: 15018126 · PMCID: PMC7169081

47. **Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges**
Chih-Cheng Lai, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, Po-Ren Hsueh
*International Journal of Antimicrobial Agents* (2020-03) https://doi.org/ggpj9d
DOI: 10.1016/j.ijantimicag.2020.105924 · PMID: 32081636 · PMCID: PMC7127800

48. **The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2**
, Alexander E Gorbalenya, Susan C Baker, Ralph S Baric, Raoul J de Groot, Christian Drosten, Anastasia A Gulyaeva, Bart L Haagmans, Chris Lauber, Andrey M Leontovich, … John Ziebuhr
*Nature Microbiology* (2020-03-02) https://doi.org/ggqj7m
DOI: 10.1038/s41564-020-0695-z · PMID: 32123347 · PMCID: PMC7095448

49. **Factors Influencing the Surprising Instability of Word Embeddings**
Laura Wendlandt, Jonathan K Kummerfeld, Rada Mihalcea
*arXiv* (2018) https://doi.org/gqcn9m
DOI: 10.48550/arxiv.1804.09692

50. **arXiv.org: the Los Alamos National Laboratory e-print server**
Gerry McKiernan
*International Journal on Grey Literature* (2000-09-01) https://doi.org/fg8pw7
DOI: 10.1108/14666180010345564

51. **PsyArXiv**
Margie Ruppel
*The Charleston Advisor* (2021-10-01) https://doi.org/gqbk4k
DOI: 10.5260/chara.23.2.38

# Supplemental Tables

**Table S1:** The intersection of changepoints found between published papers and preprints.

| Token | Changepoint |
|---|---|
| lockdown | 2019-2020 |
| 2021 | 2020-2021 |
| distancing | 2019-2020 |
| 2019 | 2018-2019 |
| ace2 | 2019-2020 |
| pandemic | 2019-2020 |
| 2020 | 2019-2020 |
| coronavirus | 2019-2020 |

| Token | Changepoint |
|-------|-------------|
| bcl2a1 | 2018-2019 |
| peak3 | 2020-2021 |
| 3.6.2 | 2019-2020 |
| quarantine | 2019-2020 |
| cobl | 2020-2021 |
| injectrode | 2020-2021 |
| nrc3 | 2020-2021 |
| 4.0.5 | 2020-2021 |
| TMPRSS2 (gene_7113) | 2019-2020 |
| n262 | 2019-2020 |
| bin1 | 2017-2018 |
| n3c | 2020-2021 |
| tip1 | 2020-2021 |
| omicron | 2020-2021 |
| pangolin | 2019-2020 |
| adrn | 2020-2021 |
| seir | 2019-2020 |