Visualization Methods for Diachronic Semantic Shift

Raef Kazi and **Alessandra Amato** and **Shenghui Wang** and **Doina Bucur**
Unversity of Twente
Drienerlolaan 5, 7522 NB Enschede, The Netherlands
{r.kazi,a.amato}@student.utwente.nl
{shenghui.wang,d.bucur}@utwente.nl

Abstract

The meaning and usage of a concept or a word changes over time. These diachronic semantic shifts reflect the change of societal and cultural consensus as well as the evolution of science. The availability of large-scale corpora and recent success in language models have enabled researchers to analyze semantic shifts in great detail. However, current research lacks intuitive ways of presenting diachronic semantic shifts and making them comprehensive. In this paper, we study the PubMed dataset and compute semantic shifts across six decades. We develop three visualization methods that can show, given a root word: the temporal change in its linguistic context, word re-occurrence, degree of similarity, time continuity, and separate trends per geographic location. We also propose a taxonomy that classifies visualization methods for diachronic semantic shifts with respect to different purposes.

1 Introduction

Diachronic semantic shift or concept drift studies how a language (the meaning and usage of words) evolve over time (Wang et al., 2011). Studying such semantic shifts is valuable for researchers who are interested in either the societal and cultural evolution, or the development of scientific research. In the latter case, innovations and groundbreaking discoveries often introduce new concepts, bring new meanings to existing ones, or shift existing meanings completely. Automatically identifying and understanding diachronic semantic shifts is thus desirable.

The availability of large-scale corpora (Hilpert and Gries, 2008) and recent success in language models (Tum, 2020) have enabled researchers to analyze semantic shifts in great detail (Jatowt and Duh, 2014; Hamilton et al., 2016; Azarbonyad et al., 2017; Gonen et al., 2020). Most of the research focuses on discovering general trends in semantic shifts, tracing the dynamics of the relationships between words, and elaborates on the methods used to detect such a shift (Kutuzov et al., 2018). Little research has explored the visual representation of such semantic shifts to help understand them intuitively. The visuals used across multiple studies include word graphs (Wijaya and Yeniterzi, 2011; Li et al., 2021), scatterplots (Kulkarni et al., 2015; Mahmood et al., 2016), and storylines (Mahmood et al., 2016). However, these methodologies currently fail in explicitly showing the temporal changes of a word and require the user to have some domain knowledge to fully comprehend the drifts. There is a strong need for further exploring the nature of this semantic shift by employing new visualization methods to make the semantic shift understandable, explainable, and explorable.

The goal of the this study, therefore, is to present a classification of visualization methods for a word's semantic shift based on the type of concept the user wishes to analyze, which leads us to the following objectives:

(a) Introduce intuitive methods for visualizing diachronic semantic shifts

(b) Propose a taxonomy that classifies visualization methods for diachronic shifts based on the type of concept one wishes to visualize

In this paper, we compute diachronic semantic shifts in PubMed across six decades, and propose three visualization methods utilising radial bars, spiral lines and word-cloud maps that can show, given a root word: the temporal change in its linguistic context, word re-occurrence, degree of similarity, time continuity, and separate trends per geographic location. We compare different visualization methods and propose a taxonomy that classifies methods for visualizing diachronic semantic shift.

Proceedings of the Third Workshop on Scholarly Document Processing, pages 89–94
October 12–17, 2022.

2 Data

Our study is based on PubMed (National Library of Medicine, 2022), a large, long-term corpus of citations and abstracts of biomedical literature. We include articles from 1970 onward (when abstracts are available). We randomly sample 106 out of the available 1114 `xml` data files, to keep the data relatively balanced over the 52 years, enough to create a word embedding for each decade. We then divide the corpus into six decades, from the 1970s to the 2020s. Table 1 shows the distribution of articles per decade.

Decade	No. articles	Decade	No. articles
1970s	222771	2000s	434448
1980s	258171	2010s	458802
1990s	247961	2020s	426790

Table 1: **PubMed**: the distribution of articles per decade

The biomedical abstracts have all their punctuation removed, and are tokenized into words which are then lowercased and lemmatized, with numerals and stop words also removed. On this preprocessed corpus, concept drift can be measured.

3 Method

3.1 Quantifying semantic drift

Inspired by Gonen et al. (2020), word embeddings are computed per decade (so the six decades are treated separately). We use the Continuous Bag of Words (CBOW) model of word2vec (Mikolov et al., 2013) with window size 10 to train each decade, and store the resulting embeddings. Then, the top $k = 50$ neighbours of a word in the embedded space make up the "linguistic context" of the word. Parameter k is tunable; Li et al. (2021) also showed that this model generally produces stable similarity scores across the corpus, when varying k. To measure concept drift, we look at the similarity of its contexts across time. The Jaccard index measures the context similarity of any word between any two decades. Let C_1 and C_2 be two different word contexts of the same word related to two different decades, the Jaccard index is defined as follows:

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} = \frac{|C_1 \cap C_2|}{|C_1| + |C_2| - |C_1 \cap C_2|}$$

3.2 Visualizing semantic drift

Using the word embeddings and contexts, we propose three **visualization methods**. The features of these visualization methods are summarized in Table 2, and compared to those of the related work. These features were chosen to help visualize at a glance the change of a word over time. The *similar words* and the *degree of similarity* are standard measures of closeness of a word or change in meaning. The *continuity* of a word is able to show precisely how its meaning changes over time. The *word re-occurrence* informs a user whether a linguistic shift has occurred or not, and *geography* adds an extra dimension to study the context of concept drift.

Table 2 also forms a taxonomy that classifies visualization methods for diachronic semantic shifts with respect to different purposes. While all related work is focused on showing the top similar words and the degree of similarity with the root word, our methods also capture a combination of word re-occurrence and continuity through time, so the temporal factor becomes clearer. We also add a geographic dimension in the word-cloud map.

Radial bar chart This visualization shows the concept drift of a *root word* over time periods at a glance. A circle is divided into time slices (here, per decade). In each slice, the k words in the context of that time period are arranged. Their order is informative, so the words are sorted twice: first by their *re-occurrence* (whether the word is present in the context of any other decade), and then by their *similarity* to the root word (their closeness in embedded space). The length of the bar shows this degree of similarity. The bars have one of two colours: orange if re-occurring and blue otherwise. The proportion of colours and the length of the bars provide a *global* picture of the concept drift over time. The user can then also read the individual words to understand the *local* context.

Spiral line chart This visualization focuses on showing the continuity of a context word across time. Per root word, each context word is represented as a line that runs through each decade. A spiral (rather than line) shape is chosen to compress multiple time periods and words in a small space. The spiral starts in summary: a bar chart with the root word's aggregate context over all time periods (taller bars mark context words which re-occur in the context of the root word). The spiral then continues through time periods. The continuity of a context word through time is seen in the continuity of its line through the spiral. A segment is present in a time period only if the respective word occurs in the word embedding for that particular decade.

	top similar words	word re-occurrence	degree of similarity	continuity	geography
Radial bar chart (this work)	✓	✓	✓		
Spiral line chart (this work)	✓	✓		✓	
Word-cloud map (this work)	✓		✓		✓
Word graph	✓		✓		
Scatterplot	✓		✓		
Storyline chart	✓			✓	

Table 2: **Features of visualizations compared with related work**: word graphs (Li et al., 2021; Wijaya and Yeniterzi, 2011), scatterplots (Kulkarni et al., 2015; Mahmood et al., 2016), storyline charts (Mahmood et al., 2016)

This visualization is similar to storyline visualization (Mahmood et al., 2016), but has the added benefit of determining, at a glance, the words that are a closest match to and retain the context of the root word.

Word-cloud map This visualization tracks changes geographically, as well as across time periods, in a word cloud of the top $k = 100$ neighbours (with k tunable). Kulkarni et al. (2015) performed studies on tracking geographic changes, but they use a scatterplot which doesn't make geographic differences explicit. The geographic data here is the home country of the journal publishing the work.

4 Results

First, the Jaccard index (the context similarity of any word between any two decades) shows a clear trend for each combination of decades, and various root words. Namely, many root words have a Jaccard similarity close to 0 (they exhibit a near-total change in their context across the time periods). However, these root words are not common scientific terms. The distribution of Jaccard index values has a long tail towards the maximum value 1, and the more interesting terms (such as those presented in the next examples) lie on this tail.

Figure 1 shows the radial bar charts for the terms "anxiety", "cigarette", "coronavirus", and "misinformation". We see that the context around "anxiety" and "cigarette" stay relatively stable, but the difference is that the context of "anxiety" is much stronger than that of "cigarette" as the length of the bars does not change much among the top 50 contextual words.

The lower left radial bar chart shows that the context of "coronavirus" (especially in the 2020s) has changed dramatically, suggesting how the research around coronavirus has shifted its focus due to the ongoing Covid-19 pandemic. For the relatively modern word "misinformation", interesting patterns are visualized. From being nearly completely missing in the 1970s, its context has largely changed, with quite different degrees of similarity. Words such as "journalist" and "trump" show up in the 2010s, owing to the rise of the "fake news" phenomena that was prevalent at the time, and words like "twitter", "antivaccination", "netizens", "celebrity", "socialmedia", "facebook", and "instagram" in the 2020s, signifying social media and the internet as primary modes through which information, or in this case misinformation, is being disseminated.

In Figure 2, we track the continuity of the word "anxiety" over time. We have manually annotated the category a word may belong to and assigned a colour to each of them. The continuity of the word can be seen if the line corresponding to the word is present in all the decades. Some words, such as "alienation" and "apprehension" are only present in a single decade, while "fear" is present in the initial few decades, moves out of the word context, and is present in the later decades again. Words such as "anger", "mood", and "selfesteem" are present across all decades and can be seen as uninterrupted lines across all the time periods.

Figure 3 shows the geography of the word "divorce" over three pairs of decades, respectively 1970s/1980s, 1900s/2000s and 2010s/2020s. The countries selected are USA, UK and The Netherlands and the choice was motivated by the large volume of articles that they presented in the dataset. As can be observed from the picture, an interesting pattern is identified; during the 1970s and 1980s divorce had context words related to social status and level of education. However, in more recent years, the word seems to be more associated to the negative consequences that divorce itself can cause, such as increased criminal behaviour, violence and self-harm. This phenomenon is more apparent in the USA and The Netherlands, but is not as evident in the UK, where divorce is still closely related to paternity and motherhood.

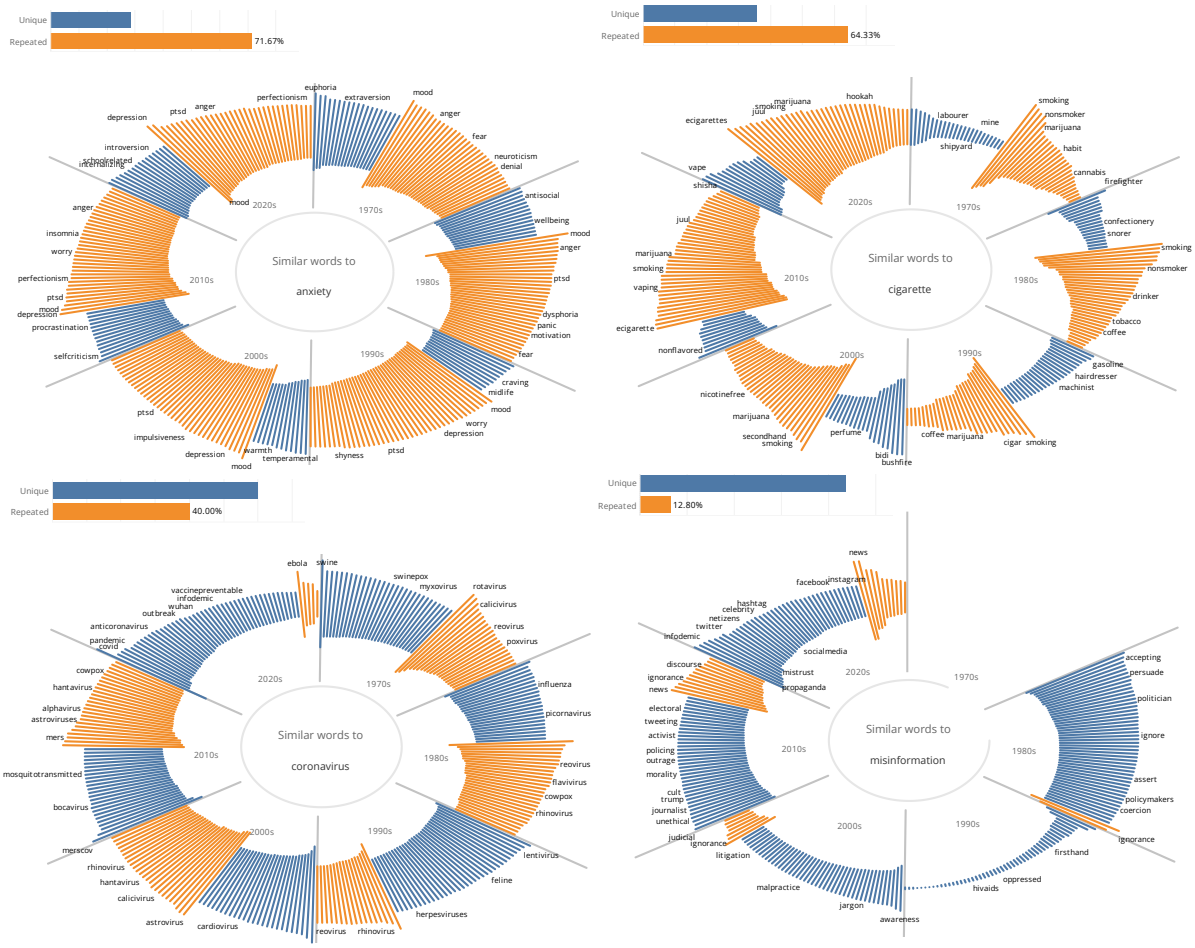

Figure 1: **Radial bar charts** ($k = 50$) for the terms "anxiety", "cigarette", "coronavirus", and "misinformation". Interactive versions online (Raef Kazi, 2022a).

5 User Study

To test the usability of these visualizations and measure whether they provide the desired benefit, we conducted a user study incorporating a self-assessment questionnaire. This study was conducted using the "VisEngage" questionnaire developed for interactive visualizations by (Hung and Parsons, 2017). Our questionnaire consisted of 11 questions which were grouped together based on type of characteristics they were meant to measure, which were, **aesthetics**, **ease of finding information**, **usability**, and **user engagement**. For each question, participants provide their response on a seven-point Likert scale, ranging from strongly disagree (1) to strongly agree (7).

Our study consisted of 5 visualizations, including previous visualization techniques from related work, the proposed methods in this paper, and the same information in tabular data for a comparison of usability. For each visualization, the study consisted of 2 task-related questions whose answers could be found within the visualization, the questionnaire for measuring the categories, and an open feedback form for the user to share their opinion on the visualization.

The results of this study from 8 participants have been summarized Figure 4. The heatmap shows that the average scores of the proposed Radial Bar Chart and Spiral Line Chart are *4.96* and *4.43* respectively, both above a *"neutral"* score (4), whereas, from the related work, only the Word Graph (4.7) scores above a *"neutral"* score, with the Scatterplot scoring a *3.74* and plain tabular data scoring *3.77*. The Radial Bar Chart performs best in categories of **aesthetics**, **usability**, and **user engagement**, and is second to the Word Graph in **ease of finding information**. We see that the Word Graph performs better overall compared to the Spiral Line Chart (except in the **aesthetics** category), and is a close second to the Radial Bar Chart overall, allowing it to also be a viable option for visualizing semantic shifts.

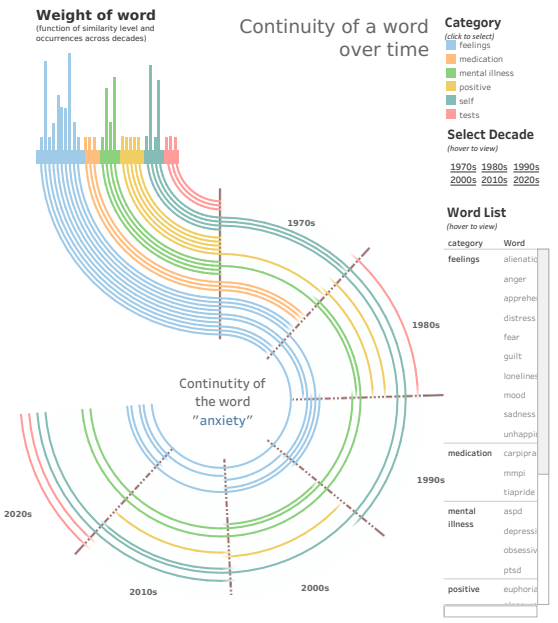

Figure 2: **Spiral line chart** for the word "anxiety". Interactive versions online ((Raef Kazi, 2022b)).

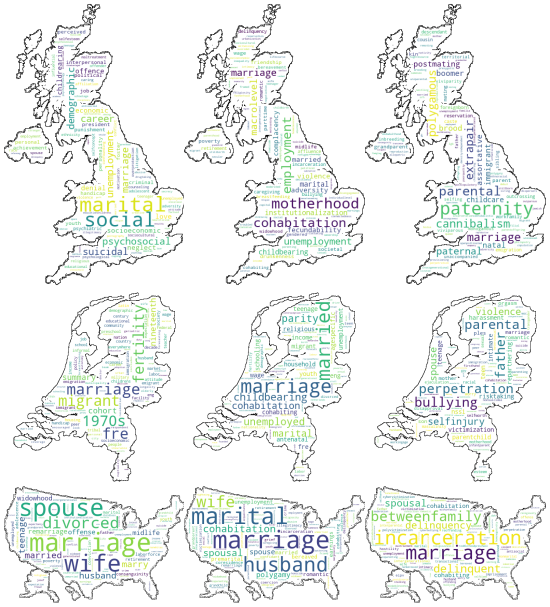

Figure 3: **Word-cloud maps** ($k = 100$) for the word "divorce" by the publisher location (UK, The Netherlands, and USA)

These scores from the study, taken in conjunction with the features visualized by each chart from Table 2, can aid in the selection of the right chart to use to best visualize a concept.

6 Conclusion

In this paper we have studied the diachronic semantic shift of words over time and have proposed methods to visualize these shifts. We perform a user study to test the usability of the proposed methods. Our interactive visualization tool helps users explore and understand these semantic shift. We also study previous visualization methods by other researchers and compare them to our proposed methods with a classification taxonomy.

In the future, we will study the complete PubMeb dataset and apply different methods to identify semantic shifts. More metrics need to be developed to further quantify the semantic shift so that highly dynamic words can be identified automatically. Furthermore, instead of manually annotating radial bar charts and selecting words for storylines, we will explore different possibilities of automatic annotation and selection.

This method does not capture the context of multiple words occurring together. For example, the meaning of the words "vaccine" and "news" may stay the same throughout time, but the context in which these words occur together can differ across time periods (these 2 words could have different neighbours if considered during the coronavirus pandemic). We will also develop additional ways of visualizing diachronic semantic shift.

	Question	Scatterplot	Word Graph	Radial Bar Chart	Tabular Data	Spiral Line Chart
Aesthetics	A	3.88	4.50	5.75	3.00	5.00
	B	3.75	4.00	5.63	3.38	4.63
	C	2.75	3.88	4.63	3.13	4.75
Ease of finding information	D	3.50	5.50	5.13	3.75	4.25
	E	3.50	5.38	4.63	3.75	4.25
Usability	F	4.38	5.13	5.38	4.63	4.38
	G	4.13	5.00	5.38	4.25	4.75
	H	3.00	4.63	4.13	3.38	4.00
User Engagement	I	4.25	4.00	4.88	4.75	4.38
	J	3.50	4.63	4.25	4.00	4.25
	K	4.50	5.13	4.88	3.50	4.13

Score Legend				
2.00	3.00	4.00	5.00	6.00

Figure 4: Heatmap of average scores for each question of each item across 5 visualizations. Values 1 and 7 have been removed from the score legend as they were never used by the participants. Sections have been superimposed for discussion purposes.

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Martin Hilpert and Stefan Th. Gries. 2008. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

Ya-Hsin Hung and Paul Parsons. 2017. Assessing user engagement in information visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1708–1717.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 229–238. IEEE.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Freshman or fresher? quantifying the geographic variation of internet language. *arXiv preprint arXiv:1510.06786*.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ruiyuan Li, Pin Tian, and Shenghui Wang. 2021. Study concept drift in 150-year english literature. In *CEUR workshop proceedings of the First Workshop on AI + Informetrics*, volume 2871, pages 153–163.

Salman Mahmood, Rami Al-Rfou, and Klaus Mueller. 2016. Visualizing linguistic shift. *arXiv preprint arXiv:1611.06478*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

National Library of Medicine. 2022. PubMed. https://pubmed.ncbi.nlm.nih.gov. Accessed 2022.

Raef Kazi. 2022a. Radial bar chart (concept drift). https://public.tableau.com/app/profile/raef6267/viz/RadialBarChartConceptDrift/RadialBarChartConceptDrift. Accessed 2022.

Raef Kazi. 2022b. Spiral line chart (concept drift). https://public.tableau.com/app/profile/raef6267/viz/SpiralLineChartConceptDrift/SpiralLineChart. Accessed 2022.

Phylypo Tum. 2020. A survey of the state-of-the-art language models up to early 2020. https://medium.com/@phylypo/a-survey-of-the-state-of-the-art-language-models-up-to-early-2020-aba824302c6.

Shenghui Wang, Stefan Schlobach, and Michel Klein. 2011. Concept drift and how to identify it. *Journal of Web Semantics*, 9(3):247–265. Semantic Web Dynamics Semantic Web Challenge, 2010.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40.