The University of Iowa

Report for Qualifying Exam

Understanding the flow of Temporal Context Shift in Data Voids

Author:
Manisha Keim
manisha-keim@uiowa.edu

Advisor:
Rishab Nithyanand
rishab-nithyanand@uiowa.edu

Department of Computer Science
July 30, 2024

Contents

1 Introduction

In the digital age, online information seekers frequently encounter *search results* laden with *misinformation*, *biased narratives*, and *conspiracy theories*. This exposure can potentially lead users to accept and propagate false information. However, this phenomenon is not universal across all search queries. Certain searches, known as *data voids* are particularly susceptible to these issues. *Data voids* occur when searching for terms that yield limited, non-existent, or highly problematic relevant information. Unlike common searches that produce abundant data, queries falling into data voids may return no results or present irrelevant and often inaccurate information. Data voids can be mainly categorized into two main types:

- newly coined terms with no established information base, and
- existing terms whose meanings have evolved over time.

This report focuses on the latter, examining data voids where context drift occurs. The following papers has been selected to explore this phenomenon.

- Statistically Significant Detection of Linguistic Change./ ACM WWW 2014 [8]
- Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change./ ACL 2016 [5]
- The Pandemic in Words: Tracking Fast Semantic Changes via a Large-Scale Word Association Task. [10]

1.1 What is a Data Void?

Data voids, as first introduced by Golebiewski and Boyd [2], are identified while searching terms for which relevant information is either limited, non-existent, or deeply problematic.
For example, unlike a typical web search for 'basketball' that yields a vast amount of data, data voids present a problematic scenario such as searches on these terms may return no results or irrelevant, often inaccurate information. This situation could prompt users to pursue information, click on misleading search outcomes, and be exposed to false or misleading content, ultimately shaping their perceptions and beliefs with misinformation.

Data Voids often originate in fringe communities. When vulnerable users search for information using this term, they may encounter results that are misrepresentations of the topic, as these fringe communities have already shaped the narrative. Mainstream media, particularly those with far-left or far-right biases, may pick up and amplify these narratives, contributing to the misinformation cycle.

1.2 Example of a data void?(Migrant Caravan)

Let's understand data void through an example of a term - *migrant caravan* The term 'migrant caravan' originally described a group of migrants traveling together for safety and mutual support. However, its usage, particularly in political discussions, has taken on a more negative tone. Today, it is commonly associated with large groups of migrants, often from Central America, journeying through Mexico to reach the United States.

1.3 Lifecycle of data void (from long tail to spike)

1.4 Types of Data Void

1.5 Technical Challenge

analyzing data void involves context drift over historical time. (How context drift happens with data void.)

2 Related Work

Tracking how word meanings change over time. The evolution of word meanings over time has been a subject of significant interest. Researchers employed a diverse array of methodologies, including *word embeddings* [8], *neural language models* [7], and *diachronic analysis* [5, 9], to track and examine semantic shifts across historical periods. Much of this research utilized the *Google Books Ngram corpus* [3–5, 7–9, 11], a vast collection spanning from 1900 to 2009, encompassing over 500 billion books in seven languages. This dataset provides n-grams with corresponding yearly occurrences and frequencies.

Words can change meaning over time through several linguistic processes. Words have the ability to undergo shifts in meaning over the course of time as a result of various linguistic mechanisms, including but not limited to:

- *semantic drift* [3, 7, 8]
  - Refers to the evolution of word *meanings* over time.
  - Example: The word 'gay' originally meant 'happy' or 'carefree', but now it predominantly means 'homosexual'.
- *syntactic alterations* [1, 4, 8]
  - Syntax focuses on the *structure* of language. This type of change pertains to modifications in the arrangement of words in sentences.
  - Example: The word 'apple', which transitioned from being used as a 'common noun' (e.g., a fruit) to a 'proper noun' (referring to the Apple company) after the company's rise in the 1980s.
- *broadening* [1, 4, 8]
  - A word's meaning becomes more general than its earlier meaning, also known as *generalization.*
  - Example: 'Holiday' originally meant a religious festival, but now it can refer to any day of celebration or time off.
- *narrowing* [1, 8]
  - A word's meaning becomes more specific than its earlier meaning, also known as *specialization.*
  - Example: 'Literally' used to mean 'figuratively' or 'symbolically'. Now, it is used to emphasize the truthfulness of a statement.
- *amelioration*
  - A word takes on a more *positive* meaning over time.
  - Example: The word 'thrifty' once meant 'cheap' but now suggests responsible use of resources.
- *pejoration* [3, 10]
  - A word takes on a more *negative* meaning over time.
  - Example: The word 'awful' meant 'full of awe' which transitioned to 'terrible' or 'appalling'.

Hamilton et al. (2016) highlighted that *cultural* or *linguistic* factors could have driven these transformations. Certain studies concentrated on semantic changes, some worked on syntactic changes, [4, 8, 10] and others explored the broader evolution that words underwent [3, 5, 7, 8]. Table 1 outlines various types of linguistic changes and the approaches employed to identify them.

Building on the extensive research on *semantic* change. Semantic change refers to any change in the meaning(s) of a word over time or acquiring a new sense [3, 10]. Kutuzov et al. (2018) conducted a survey on semantic shifts, consolidating the existing academic research in this domain. Their work provides a comprehensive overview of the methodologies and findings related to tracking semantic changes over time using computational techniques.

Linguistic	Approach	TCD	Example		
			Word	Former Usage	New Usage
Semantic	**FREQUENCY [3, 5, 8]**				
	Log Ratio	1960 & 1990	disk	-	-
	Word Frequency	1900–2005	bitch	female dog	slang
	DISTRIBUTIONAL [1, 3–5, 7, 8, 10, 11]				
	Local Mutual Information	1960 & 1990	sleep	deep sleep	sleep disorder
	Continuous Word Embeddings	1900–2009	cell	closet, dungeon	phone, cordless
	Change Point Algorithm	1900–2005	gay	cheerful, dapper	lesbian, homosexual
	Clustering (DBSCAN)	1900–2000	mouse	mice, rat	cursor, pointer
	Cultural and linguistic Drift	1800–2000	virus	infected with the virus	spreading computer virus
	Diachronic Word Embeddings	1800–1999	awful	full of awe	terrible or appalling
	Contextual Word Embeddings	1910 – 2009	tenure	short term leases, insecurity of tenure	tenure of office, employment and tenure
	Semantic Similarity Analysis	2014–2022	immunity	politics ('legislator', 'representative').	health-related ('prevention', 'AIDS')
	PART OF SPEECH [4, 8]				
Syntactic	POS Tags	1900–2000	windows	doors and windows of a house	Microsoft Windows
	Mixed-model Regressions	1800–2000	actually	originally, nominally	presumed, believe

Table 1. **Comparative Overview of Linguistic Change Detection Approaches.** This table provides a comprehensive summary of various approaches and contributions to the study of linguistic changes, categorizing the type of changes (semantic or syntactic), and detailing the methodologies used, including frequency and distributional approaches. It highlights each study's methods, such as word embeddings, statistical analysis, and clustering, along with the time periods of change detection and examples of changing words. The table provides an overview of the advancements and differences in analyzing language evolution, capturing the breadth of approaches from traditional frequency counts to modern embedding techniques.

- **Investigating Semantic Shifts Through Word Contexts.** Gulordava and Baroni (2011) detected semantic change by focusing on words used in the 1960s and 1990s. They compared the similarity of the surrounding words (words that co-occur with the target word) in these two time periods.
  - To assess similarity, the researchers calculated the *Local Mutual Information (LMI)* score between the central word and its surrounding words. A low LMI score between the target word and its surrounding words across the two time periods indicated a potential semantic shift.
  - The study demonstrated that their distributional similarity models were effective in capturing *cultural shifts* in word meaning. For example, they found that the word 'sleep' acquired more negative connotations related to sleep disorders when comparing its contexts in the 1960s to those in the 1990s.
- **Tracking Meaning Evolution Through Neural Nets.** Kim et al. (2014) focused on analyzing how word meanings evolved between 1900 and 2009.
  - They developed the first method that employed *prediction-based model* to trace semantic shifts. This involved training a model on data from a specific year y_i and then using the resulting word vectors as the starting point for training the model on the next year's data $y_i + 1$.
  - Their method analyzed *global shifts* in a word's vector semantics. Additionally, by plotting the time series of a word's distance to its neighboring words in the model's vector space, they visualized the period during which the semantic shift occurred.
  - They demonstrated this for the word 'cell' compared to its early neighbors, 'closet' and 'dungeon,' and the more recent neighbors, 'phone' and 'cordless.' For 'cell,' the identified period of change (1985-2009), which interestingly coincides with the introduction and widespread adoption of cell phones by the public.
- **Pinpointing Significant Shifts Statistically.** Kulkarni et al. (2015) proposes a novel computational approach to identify and quantify the semantic and usage changes in words across various media (new products, movies and books).
  - Building on the concept of distributed representations proposed by Hinton [6], they map words into a continuous vector space where words with similar meanings are positioned close together.
  - The approach hinges on constructing *property time series* for each word. They propose three methods for constructing these time series (§4).:
    - **Frequency:** This method analyzes changes in a word's overall frequency of use, assuming a sudden shift in frequency might indicate a semantic shift.
    - **Syntactic:** This method examines the distribution of a word's part-of-speech tags (e.g., noun, verb) across different time periods, aiming to capture changes in how the word functions grammatically.
    - **Distributional:** This approach leverages word embeddings which are created for each year, and then alignment is done to represent them in joint embedding space, and it's utilized to construct distributional time series for a word's displacement.
  - Finally, they employed statistically sound *change point detection algorithm* to identify significant moments in these time series, pinpointing the periods where word meaning or usage likely underwent a shift. Their results indicated that computational methods for the detection of semantic shifts can be robustly applied to time spans less than a decade.
- **Word Semantic Modelling of Polysemant.** Liao and Cheng (2016) explored the semantic changes of words. Their approach built on the understanding that word meaning is closely

tied to its context. When a word's meaning changes, the surrounding words used with it *(context words)* are likely to change as well. They focused on *polysemous* words (words with multiple meanings) and aimed to detect when new meanings emerged.
  - They used the *skip-gram architecture with negative sampling* [12] to obtain word embeddings. This technique helped in capturing the contextual meaning of words by representing them in a continuous vector space.
  - They employed *DBSCAN* to group word embeddings. DBSCAN helped in identifying clusters of similar word contexts and distinguishing them from noise, which could indicate semantic changes.
  - To find similar words, they used a nearest neighbor search method called *Random Project Forest*. This method helped in identifying words that are contextually similar to a given word.
  - Finally, they compared the stability of *similar words* with the stability of their *context words*.
- **Distinguishing Cultural Shifts from Linguistic Drift.** Hamilton et al. (2016) addressed the challenge of distinguishing between *cultural shifts* and *linguistic drift*, both of which can contribute to semantic change.
  - They proposed two distinct measures based on distributional semantics to distinguish between these two types of semantic change:
    * **Local Neighborhood Measure:** This measure focused on the closest neighbors (most similar words) in a word's embedding. A drastic shift in these nearest neighbors suggests a significant change in core meaning, potentially driven by a *cultural shift* (e.g., 'gay' changing from carefree to referring to homosexuality).
    * **Global Measure:** This measure considered the overall distribution of a word's surrounding words in a larger context window. Gradual changes in this broader distribution are more likely to reflect *linguistic drift*, the natural evolution of language due to regular processes (e.g., 'promise' expanding from a declaration to also suggesting a likelihood).
  - Prior research often treated semantic change as a single phenomenon. Hamilton et al. offered a novel approach by distinguishing between cultural shifts and linguistic drift using their two measures.
- **Diachronic Analysis on Historical Data:** The primary goal of Hamilton et al. (2016) is to track semantic changes and understand how the meanings of words shift in different historical contexts.
  - They created word embeddings for different time periods using both the PPMI matrix with SVD and the SGNS model. These embeddings were generated from historical text corpora to capture the contextual usage of words in each period.
  - After creating the embeddings, they aligned the embeddings. Once the embeddings are aligned, they calculate the *semantic displacement* of a word. This essentially measured how much a word's vector representation has moved in the embedding space between two time periods.
  - The study demonstrates that their method effectively identified semantic change in words and uncovered two statistical 'laws' of semantic change:
    * **Law of Conformity**, which suggests that the rate of semantic change is inversely proportional to a word's frequency. High-frequency words tend to change meaning more slowly.
    * **Law of Innovation**, on the other hand, proposes that words with multiple meanings are more likely to undergo semantic transformations over time.

  - They leveraged the rich information within word embeddings to quantify the degree of semantic change (semantic displacement) and identified potential patterns.
- **Semantic Shifts with Contextual Embeddings:** Giulianelli et al. (2020) explored the phenomenon of lexical semantic change using an unsupervised approach using *BERT*.
  - They utilized the *BERT* model to generate contextual word embeddings, which captured the meaning of a word based on its surrounding context.
  - The extracted word usage vectors are then clustered into different *usage types* using the *k-means clustering algorithm*. This helped to identify distinct senses or meanings of the words as they appear in various contexts.
  - They analyzed these clusters for a specific word across different time periods. The study proposed three metrics to quantify semantic change:
    * **Entropy Difference (ED)**: Measures the change in uncertainty (entropy) of a word's usage distribution over time.
    * **Jensen-Shannon Divergence (JSD)**: Compares the similarity of word usage distributions across time intervals.
    * **Average Pairwise Distance (APD)**: Computes the average distance between word usage vectors from different periods, indicating shifts in word meaning.
  - The qualitative analysis indicated that the approach could capture various linguistic phenomena, including both synchronic (current usage) and diachronic (historical changes) aspects.
- **Semantic Similarity:** The impact of significant events like the COVID-19 pandemic on language and semantic change has also been a subject of study. Laurino et al. (2023) explores tracking fast semantic changes through a large-scale word association task, aiming to understand how the collective mental lexicon evolves in response to such global events. Their research highlights the dynamic nature of language and how it incorporates new senses. For instance, words like 'quarantine', 'mask', and 'social distancing' took on new and prominent meanings in everyday conversation.

Linguistic change studies have shifted to include *syntax* alongside semantics. The field of linguistic change has traditionally focused on *semantics*. However, recent studies have begun to explore the role of *syntax* in language evolution as well. The *syntactic* functionality of a word can evolve by transitioning into a new part-of-speech (POS) category. Nouns, due to their inherent flexibility in meaning, exhibit a greater tendency to undergo these changes driven by cultural shifts. While verbs are more likely to participate in gradual semantic changes that follow established linguistic patterns.

- **Acquiring a new POS.** Kulkarni et al. assigned part-of-speech (POS) tags to a large collection of text. They calculated the likelihood of a word appearing in specific grammatical contexts over time.
  - To quantify temporal change, they compared the probability distributions of POS tags for a particular word across different time periods. This essentially measured the divergence between these distributions.
  - An example they cited was the word 'windows'. Its POS tag shifted from a common noun (referring to doors and windows of a house) to a proper noun ('Microsoft Windows'). This highlights how a word's grammatical function can change alongside its meaning.
- **Capturing Differences Between Nouns and Verbs.** Hamilton et al. (2016) highlight how *cultural changes*, often influenced by new technologies, are closely tied to transformations within *local neighborhoods*, particularly sensitive to shifts in nouns. On the other hand,

linguistic changes are more associated with *global measures* and are particularly responsive to variations in verbs.
  - To validate this hypothesis, the authors employed a statistical technique called a *linear mixed model* where word type (noun or verb) is a fixed effect, amount of change measured by each metric (local or global) treated as the dependent variable. By analyzing the model's results, they could assess whether there's a significant difference in the way nouns and verbs exhibit change.
  - The evolution of words like 'actually', 'must', and 'promise' demonstrate these changes. For instance, 'must' has transitioned from expressing obligation to indicating necessity, showcasing a common pattern seen in modal verbs.

Several studies from our literature review provide valuable insights for addressing data voids in our analysis.

- The first being the work by Kulkarni et al. which introduces us to time series construction. Their distributional methods focuses on finding subtle semantic shits to determine the context where a word in used. This concept aligns perfectly with our goal of uncovering shifts in word usage when encountering data voids.
- Second paper by Hamilton et al. introduces us to word emebeddings alignment. Since we want to compare word vectors from different time periods, vectors should be aligned in same coordinate axes. After aligning the embeddings for individual time periods, we can use the aligned word vectors to compute the semantic displacement that a word has undergone during a certain time-period.
- Third one by Laurino et al. investigates the impact of the COVID-19 pandemic on word meaning. One key finding from their work is that words directly related to the pandemic exhibited a greater difference in semantic similarity between pre-pandemic and pandemic time periods. This suggests that these words underwent a more significant and rapid semantic shift compared to control words not associated with the pandemic. They also employ semantic similarity analysis to quantify the shifts in meaning for pandemic-related words and provides evidence that the COVID-19 pandemic acted as a catalyst for rapid semantic change.

3 RP1: Statistically Significant Detection of Linguistic Change

3.1 Introduction

Kulkarni et al. (2015) explores the problem of detecting changes in word usage patterns over time. They introduce methods to identify words that have undergone significant shifts in *meaning*, usage *frequency*, or *context*. This is achieved by analyzing historical corpora, which contain vast amounts of text data spanning multiple years or even centuries. The study utilizes diverse datasets, including Twitter posts, Amazon product reviews, and Google Book Ngrams, to construct time series for individual words.

The research conducted by Kulkarni et al. exhibits a compelling alignment with our own research objectives, specifically focusing on the identification of data voids characterized by shifting semantic meanings over time. Their exploration into these linguistic shifts has revealed interesting insights, such as the evolution of the word 'gay,' which transitioned from being cheerful or dapper to signifying homosexual or lesbian aspect around the mid-20th century. To achieve understanding of this phenomenon, they intend to study the following research questions.

- **RQ1.** *What methods can quantify the statistical relevance of observed changes in a word's usage across different time periods?*
  Kulkarni et al.'s study uses three technical methods to understand how words change over time: *frequency* analysis, *syntactic* analysis, and *distributional* analysis. The frequency method tracks how often words are used, revealing spikes in usage when significant events occur, like how *breaking news* data voids can create a surge in specific keywords. The syntactic method looks at changes in the grammatical roles of words, while the distributional method examines word co-occurrence patterns to see how word meanings shift. Their approach not only uncovers the dynamic nature of language but also highlights "data voids," – areas where terms have *evolved*, become *outdated*, or *fragmented* in meaning.
- **RQ2.** *Given that a word's usage has changed, how can the precise moment or period of this shift be determined?*
  Understanding data voids involves not only recognizing that a change is occurring but also analyzing the specifics of how that change unfolds over time. It's crucial to pinpoint not just the fact that a shift has happened, but also the exact moment when it took place. Kulkarni et al. tackle this challenge by implementing a *change point detection algorithm* based on the *Mean Shift* model. This method determines whether a word has experienced a significant shift and, if so, identifies the precise point at which this change occurred.

Next two sections, §3.2 and §3.3 focuses on the two research questions that this paper is studying, the approach they took to study and their results.

3.2 What methods can quantify the statistical relevance of observed changes in a word's usage across different time periods?

Overview. The challenge is to identify how the meaning of words changes over time. To address this, they examine a corpus containing data from various domains like books, tweets, and reviews, known as the Temporal Corpus (C) which spans over a time period (S). This corpus is divided into smaller segments called snapshots C_t, each of length P. From these snapshots, a common vocabulary V is created having words which are common across all snapshots. To analyze changes in these words, a time series $T(w)$ is constructed for each word $w \in V$. Kulkarni et al. propose several methods to create this time series, aimed at understanding the evolution of words over time.

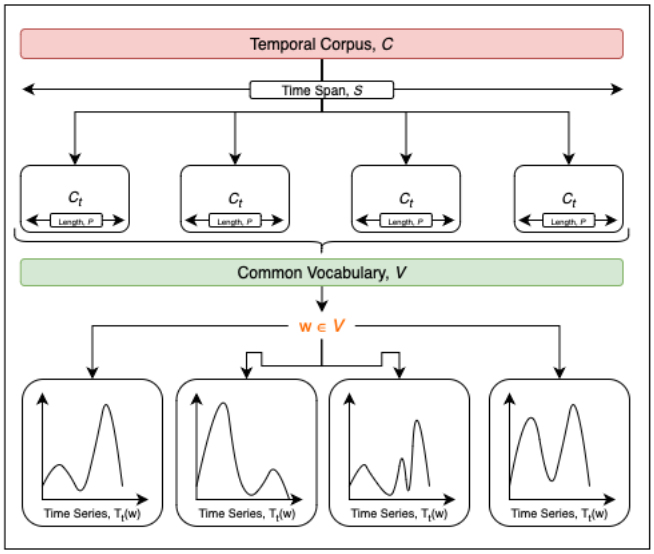

Fig. 1. My caption

Construction of Time Series. In this paper, three methods are explored for statistically modeling the evolution of words over time: frequency analysis, part-of-speech tagging, and word co-occurrence to create time series for each word.

(1) Frequency-based method: The simplest and most direct method for detecting sudden changes in word usage is by analyzing frequency trends.

This approach examines how often a word is used over a given period, providing insights into shifts in its popularity or relevance. Changes in frequency can indicate whether a word is gaining new meanings or losing old ones, reflecting broader cultural or societal trends.
For example, the word "gay" in Figure 2 shows a noticeable spike in usage during the 1980s, signaling a shift in its meaning or cultural relevance. Tools such as the Google Books Ngram Viewer and Google Trends are used for this purpose, as they provide large datasets and visual representations of word usage across different timeframes. They calculate the change in probability of a word appearing over time.

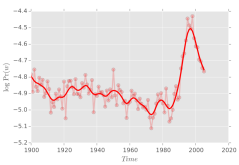

Fig. 2. My caption

$$T_t(w) = \log \frac{(w \in C_t)}{|C_t|}$$

(2) Syntactic method: Frequency-based metrics, while simple to compute, are susceptible to errors caused by imbalances in the corpus's domain and genre distribution. Fluctuations in word usage due to temporal events or the popularity of specific entities can obscure genuine shifts in word meaning. Moreover, a word's grammatical role can evolve over time, such as acquiring a different part of speech category. To study this, the corpus is annotated with part-of-speech (POS) tags, and the probability distribution of these tags is calculated for each word across different time snapshots. This approach allows researchers to observe how a word's syntactic role changes over time.

(3) Distributional method: For detecting subtle semantic changes, which are not changed either due to frequency or through change in part of speech, this method was developed to understand in which context a word is used in and based on that understand the semantic changes. Distributional methods focusses on creating a semantic space that maps words to continuos vector space, where each word is represented by a vector. Once they created a temporal word embeddings for each word in each time snapshot, then they track the changes of the represtantations across the embedding space.

- The researchers aimed to *learn word embeddings* by training neural language models on a corpus at different time snapshots. They initialized word vector representations randomly and optimized the model parameters using stochastic gradient descent. During training, the objective was to maximize the probability of context words appearing around a target word. After training, word embeddings were normalized by their L2 norm to ensure consistent representation.
- The *alignment* process involved aligning word embeddings from different time snapshots into a unified coordinate system to characterize changes between them. Assumptions were made to aid the alignment process, including the equivalence of spaces under linear transformation and the preservation of local structure of most words over time. When the alignment model failed to align a word correctly, it indicated a potential linguistic shift, highlighting the importance of accurate alignment for tracking semantic changes over time.
- By aligning embedding spaces across various time snapshots into a joint embedding space, the distributional method constructs a *distributional time series* that captures the semantic evolution of words over time.
- The paper shows the evolution of the word 'tape' over time. Initially, the word 'tape' referred to an 'adhesive tape' but underwent a semantic shift to also mean a 'cassette tape' during the early 1970s.

3.3 Given that a word's usage has changed, how can the precise moment or period of this shift be determined?

After constructing time series data for each word in the corpus—whether using frequency-based, syntactic, or distributional methods—the next step is to determine if the word has experienced a significant change. If a significant shift is detected, the algorithm then identifies and returns the estimated change point (ECP), which marks the time when this change occurred.

- Since language exhibits a stochastic drift. In this context, 'stochastic' refers to the randomness or probabilistic nature of the training process, where the models are trained on the same dataset but may produce different results each time due to this randomness. To resolve this, time series was normalized for each word by transforming the time series into a Z-score series.
- Then the algorithm utilizes a Mean Shift model to detect changes in the mean of the time series.
- Change points can be thus identified by detecting significant shifts in the mean.

3.4 Results

Time Series Analysis & Historical Analysis.

Frequency & Distributional - transmistted, bitch, sex, her The sharp increase of the word her in frequency around the 1960's could be attributed to the concurrent rise and popularity of the

Method	Examples
Frequency	transmistted, bitch, sex, her
Syntactic	apple, hug, sink, click, handle, windows, bush
Distributional	diet, tape, plastic

Dataset Source	Examples
Amazon Reviews	streaming, ray, combo, rays, twilight
Twitter Tweets	candy, myster, rally, sandy

feminist movement. Sudden temporary popularity of specific social and political events could lead the Frequency method to produce many false positives.

Syntactic - apple, hug, sink, click, handle, windows, bush The word apple was detected uniquely by the Syntactic method as its most frequent part of speech tag changed significantly from "Noun" to "Proper Noun" The Syntactic method is indicative of having low false positive rate, but suffers from a high false negative rate, given that only two words in the table were detected.

Distributional - diet, tape, plastic The popularity of books on dieting started with the best selling book Dr. Atkins' Diet Revolution by Robert C. Atkins in 1972. This changed the use of the word diet to mean a life-style of food consumption behavior and not only the food consumed by an individual or group.

Cross Domain Analysis. Amazon Reviews and Twitter (content that spans a much shorter time scale as compared to Google Books Ngram Corpus) What is the effect of applying distributional method on Amazon Review, Twitter dataset

Words - streaming, ray, combo, rays, twilight

Twitter - candy, myster, rally, sandy The word sandy acquired a new sense after the incident of Hurricane Sandy hitting the East Coast of USA, which was also mentioned in the data voids example by danah boyd.

These examples show that their method can detect the introduction of new products, movies, and books.

3.5 Takeaways

The paper by Kulkarni et al. (2015) contributes a novel and rigorous approach to detecting linguistic changes using statistical methods and word embeddings. Their contribution in using large historical corpora and word embeddings to capture the dynamic nature of language. A statistical framework that ensures the detected changes are significant, avoiding false positives due to noise or minor fluctuations.

Despite the fast pace of change of the web content, our method is able to detect the introduction of new products, movies and books. This could help semantically aware web applications to better understand user intentions and requests.

4 RP2. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

4.1 Introduction

Hamilton et al. (2016) explores the dynamic nature of language by analyzing how word meanings evolve over time. Using *diachronic* (historical) word embeddings, a technique that maps words into vector spaces based on historical corpora, the study uncovers patterns and laws governing semantic shifts. Their approach provides a quantitative framework to examine how words change meaning, influenced by cultural, and linguistic factors.

The pace at which words undergo semantic change differs, with some words changing meaning more frequently compared to others. For example, the word 'cat' has remained relatively stable in its meaning, while 'cast' has evolved to have multiple meanings [5].

There are several hypotheses about the patterns in semantic change, including the increasing subjectification of meaning or the grammaticalization. This paper, however, focuses on following two specific questions about semantic change:

- **RQ1.** *What role does word frequency play in the evolution of word meanings?*
  Frequency is a significant factor in linguistic changes, with high-frequency words often changing faster, while low-frequency words tend to be more resistant to change. The connection between word frequency and semantic change is a key unanswered question in the field of linguistics. The authors introduce the **law of conformity** to address this gap, demonstrating that frequent words change more slowly, and it clarifies the role of frequency in semantic change.
- **RQ2.** *How does polysemy relate to semantic change?*
  Another unresolved question in linguistics is the relationship between semantic change and polysemy. Polysemous words, which have multiple meanings, appear in a variety of contexts. It remains unclear whether the diverse contextual use of these words makes them more or less prone to undergoing semantic change. The authors propose the **law of innovation**, which demonstrates that polysemous words are more likely to experience faster semantic changes. If a word has multiple meanings, it is more likely to lead to semantic change, especially in the case of rare senses.

The paper aims to develop a robust methodology for quantifying semantic change using word embeddings and comparing different approaches to analyzing semantic change. This methodology is then applied in a large-scale cross-linguistic analysis spanning 200 years and four languages (English, German, French, and Chinese) to propose the above two statistical laws relating frequency and polysemy to semantic change.

4.2 Constructing Word Embeddings.

The authors employ three methods to construct word embeddings for different time periods. Initially, they create embeddings for each distinct period and then align these embeddings over time to maintain consistency. To quantify semantic change, they use various metrics. Specifically, they utilize Singular Value Decomposition (SVD), Positive Pointwise Mutual Information (PPMI), and Skip-Gram with Negative Sampling (SGNS). These distributional techniques represent each word by a vector that encapsulates information about the word's co-occurrence statistics.

(1) **PPMI:** PPMI is a statistical measure used to quantify the association between a word and its context within a corpus. In the paper, it is used to create word embeddings by capturing the strength of association between words and their co-occurring contexts over time.

- **Co-occurrence Matrix Construction.** First, a co-occurrence matrix is built, where each entry represents the frequency with which a word (target) and its context (typically a window of neighboring words) appear together in the corpus.
- **Joint Probability Calculation.** The joint probability of a target word w_i and a context word c_j is calculated by dividing the co-occurrence frequency of w_i and c_j by the total number of word pairs in the corpus.
- **Marginal Probability Calculation.** The marginal probabilities of the target word w_i and the context word c_j are calculated based on the total occurrences of each word in the corpus.
- **Calculating PPMI.** The Positive Pointwise Mutual Information (PPMI) between the target word w_i and the context word c_j is calculated using the formula:

$$\mathbf{M}_{i,j}^{\text{PPMI}} = \max\left\{\log\left(\frac{\hat{p}(w_i, c_j)}{\hat{p}(w_i)\hat{p}(c_j)}\right) - \alpha, 0\right\}, \tag{1}$$

  where:
  - $\hat{p}(w_i, c_j)$ is the estimated *joint probability* of the target word w_i and the context word c_j occurring together.
  - $\hat{p}(w_i)$ and $\hat{p}(c_j)$ are the estimated *marginal probabilities* of the target word w_i and the context word c_j, respectively.
  - $\alpha > 0$ is a positive constant used to smooth the values and prevent extreme positive values in the PPMI calculation.

  This formula refines the PPMI measure by incorporating α to adjust the log probability ratio. The inclusion of α helps manage the sparseness of the data and the impact of low-frequency events, making the measure more robust and stable.

(2) **SVD:** SVD is a mathematical technique that factorizes a matrix into three other matrices, revealing important features of the original matrix. In the context of word embeddings, SVD helps to reduce the dimensionality of the data while preserving the essential structure and patterns in word co-occurrences. SVD decomposes the co-occurrence matrix into three matrices in the following manner:

$$\mathbf{w}_i^{\text{SVD}} = (\mathbf{U}\Sigma^{\gamma})_i, \tag{2}$$

where:
- $\mathbf{w}_i^{\text{SVD}}$ is the vector representation of the word w_i in the SVD-based embedding space.
- $\mathbf{U}$ is the matrix of left singular vectors obtained from the decomposition of the co-occurrence matrix.
- Σ is the diagonal matrix of singular values.
- γ is a parameter that scales the singular values, allowing for tuning the influence of the different components.

In this formula, the word vector $\mathbf{w}_i^{\text{SVD}}$ is derived by scaling the singular values in Σ by γ and then multiplying by the corresponding vectors in $\mathbf{U}$. This approach provides a flexible way to adjust the contribution of the components to the final word embeddings.

(3) **SGNS:** SGNS is a technique used to learn high-quality word vectors by training on large corpora. The primary goal of SGNS is to learn word embeddings such that words that appear in similar contexts have similar vector representations.
For a given target word w and a context word c , the model tries to maximize the probability P(c|w)

$$\hat{p}(c_i \mid w_i) \propto \exp(\mathbf{w}_i^{\text{SGNS}} \cdot \mathbf{c}_j^{\text{SGNS}}), \tag{3}$$

where:

- $\hat{p}(c_i \mid w_i)$ is the estimated probability of a context word c_i given a target word w_i.
- $\mathbf{w}_i^{\text{SGNS}}$ is the vector representation of the target word w_i in the SGNS embedding space.
- $\mathbf{c}_j^{\text{SGNS}}$ is the vector representation of the context word c_j in the SGNS embedding space.

The probability $\hat{p}(c_i \mid w_i)$ that a word c_i is a context word of w_i is proportional to the exponential of the dot product of their corresponding vector representations. The dot product $\mathbf{w}_i^{\text{SGNS}} \cdot \mathbf{c}_j^{\text{SGNS}}$ measures the similarity between the target word and the context word in the vector space. A higher dot product indicates that the words are more likely to co-occur, leading to a higher estimated probability.

The model uses this probability estimation to distinguish true word-context pairs from randomly sampled "negative" pairs. The training objective is to maximize the probability of observed (target, context) pairs and minimize the probability of randomly sampled pairs, thereby learning embeddings that reflect the semantic relationships between words.

4.3 Aligning Embeddings

Since the embeddings are trained independently for each time period, their vector spaces may not be directly comparable. The paper addresses this by aligning the embeddings from different time periods. They aligned the emebddings across time periods using orthogonal Procrustes, which transforms the embeddings to minimize the distances between corresponding words in different periods. This step ensures that similar meanings are represented similarly across time.

4.4 Measuring Semantic Change

Once the embeddings are aligned, following methods are performed to quantify how much the meaning of word changes over time.

- **Pair-wise similarity time-series** This method measures changes in the similarity between pairs of words over different time periods using cosine similarity.

$$s^{(t)}(w_i, w_j) = \text{cos-sim}(w_i^{(t)}, w_j^{(t)}) \tag{4}$$

The Spearman correlation (ρ) is employed to measure the relationship between the similarity scores of word pairs and time. This non-parametric method assesses whether the similarity series shows a significant increase or decrease over time, which is crucial for understanding semantic shifts.

- **Measuring semantic displacement** This method quantifies how much a word's meaning has changed by calculating the displacement of its vector representation across different time points. They use the aligned word vectors to compute the *semantic displacement* that a word has undergone during a certain time-period. It is measured by calculating the distance between the vector representations of a word in different time periods. For a given word w, if $\mathbf{w}^{(t1)}$ and $\mathbf{w}^{(t2)}$ are its embeddings in two time periods $t1$ and $t2$, respectively, then the semantic displacement Δ is computed as:

$$\Delta(w) = \text{cos-dist}(\mathbf{w}^{(t1)}, \mathbf{w}^{(t2)}). \tag{5}$$

4.5 Evaluation of various approaches

They compared the different approaches discussed in **??** on *synchronic accuracy* (similarity within time-period) and *diachronic validity* (semantic changes over time).

- **Synchronic Accuracy** Synchronic accuracy refers to the accuracy with which word embeddings capture the semantic relationships and meanings of words at a *specific point in time*. It assesses how well the models represent the semantic relationships among words in a given time period. SVD performed best here.
- **Diachronic Validity** Diachronic validity evaluates how effectively the methods can detect and quantify changes in *meaning over time*. It involves testing the methods against historical data to see if they can accurately identify known shifts in word meanings. SGNS perfomed best on both of the tasks:
  - **Detecting Known Shifts.** The goal in this task is for the methods to correctly capture whether pairs of words moved closer or further apart in semantic space during a predetermined time-period.
  - **Discovering Shifts from data.** Tested whether the methods discover reasonable shifts by examining the top-10 words that changed the most from the 1900s to the 1990s according to the semantic displacement metric (**??**).

4.6 Statistical Laws of Semantic Change

The authors establish two statistical laws of semantic change by analyzing a large dataset across multiple languages and time periods. They demonstrate these laws by involving the comparison of different word embedding models (PPMI, SVD, SGNS) against known historical semantic shifts and novel benchmarks.

How diachronic embeddings can be used to reveal statistical laws that relate frequency and polysemy to semantic change.

- **Law of conformity:** *Frequently used words change at slower rates.*
  Words with higher frequencies tend to experience slower rates of semantic change. This is statistically supported by the observation that high-frequency words have smaller semantic displacements over time.
- **Law of innovation:** *Polysemous words change at faster rates.*
  Polysemous words tend to change more rapidly. The study finds that these words show larger semantic displacements, indicating more significant changes in meaning over time.

4.7 Takeaways

This paper explores how different types of word embedding models, such as PPMI, SVD, and SGNS, can be used to study diachronic shifts in word meanings. It demonstrates the use of metrics like Spearman correlation to quantify semantic drift and employs Procrustes alignment techniques to align word embeddings across different time periods. By analyzing these aligned embeddings, the authors identify two key statistical laws of semantic change. These findings provide a quantitative framework for understanding how and why languages evolve over time.

References

[1] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3960–3973. https://doi.org/10.18653/v1/2020.acl-main.365

[2] Michael Golebiewski and danah boyd. 2019. Data Voids: Where Missing Data Can Easily Be Exploited.

[3] Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus.. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Sebastian Pado and Yves Peirsman (Eds.). Association for Computational Linguistics, Edinburgh, UK, 67–71. https://aclanthology.org/W11-2508

[4] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 2116–2121. https://doi.org/10.18653/v1/D16-1229

[5] 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)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1489–1501. https://doi.org/10.18653/v1/P16-1141

[6] Geoffrey E. Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*. Amherst, MA, Amherst, MA, 1–12.

[7] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, Cristian Danescu-Niculescu-Mizil, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith (Eds.). Association for Computational Linguistics, Baltimore, MD, USA, 61–65. https://doi.org/10.3115/v1/W14-2517

[8] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically Significant Detection of Linguistic Change. arXiv:1411.3315 [cs.CL] https://arxiv.org/abs/1411.3315

[9] 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*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1384–1397. https://aclanthology.org/C18-1117

[10] Julieta Laurino, Simon De Deyne, Álvaro Cabana, and Laura Kaczer. 2023. The Pandemic in Words: Tracking Fast Semantic Changes via a Large-Scale Word Association Task. *Open Mind* 7 (06 2023), 221–239. https://doi.org/10.1162/opmi_a_00081 arXiv:https://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi_a_00081/2133848/opmi_a_00081.pdf

[11] Xuanyi Liao and Guang Cheng. 2016. Analysing the Semantic Change Based on Word Embedding. In *Natural Language Understanding and Intelligent Applications*, Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng (Eds.). Springer International Publishing, Cham, 213–223.

[12] 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* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.