UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,
Mathematics & Computer Science

Navigating Semantic Shifts: A Visual Tool for Exploring Word Meaning Change

Raef Kazi

M.Sc. Thesis
September 2023

Committee:
dr. Shenghui Wang
dr. Doina Bucur
dr. Marcos Machado
Rob Koopman (OCLC B.V.)

Business Information Technology
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

Detecting and tracking semantic shifts in language is a complex task with significant implications for various fields. Computational methods like topic modeling have emerged as powerful tools for this purpose. However, challenges persist in evaluating, selecting, and interpreting suitable models. Existing metrics like coherence scores often fall short in capturing the nuances of word sense changes, leading to laborious manual inspection and parameter tuning. This research aims to bridge this gap by developing a user-friendly visualization tool that enhances the understanding of semantic shifts and facilitates model selection.

This study combines insights from semantic shift literature and data visualization techniques. It begins with a comprehensive review of semantic shift visualization methods to identify the requisite features to visualize for the creation of an innovative tool. This tool integrates various visualization elements, including intertopic maps, stacked bar charts, steamgraphs, and coherence scores. Using a topic modeling approach, the study detects word senses across seven decades, enabling users to visualize semantic shifts and assess the impact of different parameters. A user study, involving a diverse group of participants, evaluates the tool's effectiveness in exploring word senses and making model selections.

The research yields a comprehensive visualization tool that empowers users to explore diverse word senses and their evolution over time. Users leverage temporal features to navigate subtle changes in word senses. The tool aligns users' intuitive model choices with quantitative measures, bolstering their confidence in model selection. However, challenges related to user familiarity, visual complexity, and the subjectivity of semantic shift identification emerge.

This research contributes to the semantic shift analysis field by evaluating a visualization system on the basis of its usability and task support. It represents an interdisciplinary intersection of linguistics, data visualization, and user experience design. To further enhance the tool's utility, future work should integrate user feedback, broaden participant demographics, enable real-time computation, and offer multilingual support.

Keywords: semantic shift visualization, concept drift, word sense induction, computational linguistics, natural language processing.

Acknowledgements

This thesis marks the culmination of my two-year master's program journey. Yet, no journey is traversed alone, and there are individuals without whom this research would never have taken shape.

First and foremost, I extend my deepest gratitude to my supervisors for their boundless patience, wisdom, and invaluable guidance. Doina and Shenghui introduced me to the realm of semantic shifts more than a year ago, that piqued my interest in this field and helped set the stage for this research. Shenghui's infinite patience in helping me understand the intricacies of semantic shifts, coupled with her support while I learned and stumbeld, have been instrumental. Doina's guidance, reading recommendations, and conversations well before this thesis took its current form was a great source of inspiration to draw from. A huge thanks to Marcos for his expertise, profound feedback, and fresh perspectives that enriched this research. Special thanks go to Rob, whose support at OCLC ensured that my research was conducted as smoothly as possible, and whose conversations were a source of joy during the work days.

My heartfelt appreciation extends to my dear friends who made the Netherlands into a home away from home for me. Dea, Jesús, Tifani, Christian, Amy, Jeroen, Danniar, and Donika: without you, this journey would have been bereft of light.

No words of thanks can be enough for Evi's enduring patience and unwavering support during this all. Her reminders to celebrate the highs and her role as both cheerleader and therapist during the lows have been invaluable.

Lastly, but by no means the least, I am eternally grateful to my family – my parents, sister, and brother – for all your love, prayers, support, and encouragement that have extended well beyond the last two years. I felt it all despite the distance.

“Words do have power. Names have power. Words are events, they do things, change things. They transform both speaker and hearer; they feed energy back and forth and amplify it. They feed understanding or emotion back and forth and amplify it.”

- Ursula K. Le Guin

“Words are futile devices.”

- Sufjan Stevens

Contents

List of Figures

List of Tables

List of Acronyms

BERT	Bidirectional Encoder Representations from Transformers
c-TF-IDF	Class-Based Term Frequency-Inverse Document Frequency
DTM	Dynamic Topic Modeling
HCI	Human-Computer Interaction
HDBSCAN	Hierarchical Density-Based Spatial Clustering
ISO	International Organization for Standardization
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MARC	Machine-Readable Cataloging
NLP	Natural Language Processing
OCLC	Online Computer Library Center
PCA	Principal Components Analysis
t-SNE	t-distributed Stochastic Neighbor Embedding
TF-IDF	Term Frequency-Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection
UI	User Interface
UX	User Experience

Chapter 1

Introduction

Language constantly evolves over time. A word may take up a new meaning alongside the one it has, or may completely change its meaning to mean something new. A classic example of a semantic shift is of the word *"gay"*, which has changed in meaning over time. Originally, the word referred to feelings of being happy or cheerful, and, since the 20th century, is majorly used to refer to homosexuality. This exemplifies how words undergo shifts that reflect the evolving values, attitudes, and expressions of the societies in which they are used.

The study of semantic shifts holds relevance across various domains and disciplines. Computational linguists develop algorithms to automatically detect and track these shifts in large textual corpora. Linguists investigate the linguistic mechanisms and socio-cultural factors behind semantic change. Additionally, users from different fields, such as historians, sociologists, and data analysts, are interested in understanding how word meanings evolve over time.

With the emergence of computational detection and tracking of semantic shifts, visualization methods were proposed as an ideal way to interpret these shifts and make them accessible for researchers, practitioners, and users alike. It becomes necessary, then, to be able to select the right form of visualization for the type of shift one wishes to show, or the features considered most important to track.

While the computational detection and tracking of semantic shifts has been approached through various methods [3], we focus on the topic modeling approach, which works by grouping the different senses of a word from a corpus as a proxy for word senses and tracks the change of these senses to measure a shift in its semantics.

Although topic modeling has emerged as a powerful approach for detecting and analyzing semantic shifts [3], several challenges still persist. One major challenge lies in the lack of a universally agreed-upon metric for validating the performance and accuracy of topic models in capturing semantic shifts [3]. Evaluating and selecting the appropriate model often involves a process of parameter fine-tuning and manual inspection of topic labels. Metrics such as coherence scores [4] have been commonly used to measure the quality of topics. However, high values of these measures do not always guarantee better topics in the sense of being human interpretable [2], nor do they measure whether all usages of a word have been captured separately, in the context of semantic shift detection. This necessitates the need for manual inspection of topic labels with varying parameter values, and the trial-and-error process can be time-consuming and laborious for practitioners.

Another challenge is the absence of a ground truth for word sense detection and the semantic change of unknown words. This lack of a definitive benchmark also affects the model selection capabilities of other computational detection methods, and complicates the process of model selection and validation. It necessitates extensive back-and-forth iterations to determine a model most suitable based on the use-case and data available. Thus, fine-tuning parameters and visually inspecting the resulting topics becomes an essential part of this process, which can become laborious and time-intensive.

In response to these challenges, this study presents a specialized visualization tool designed to bridge the gap between the computational insights generated by topic modeling and the human understanding of semantic shifts in word senses. This tool offers a user-friendly interface for exploring and interpreting the results of different topic models, facilitating the selection of models for semantic shift analysis and enhancing the usability of such models for researchers and practitioners.

The visualization tool proposed in this study seeks to facilitate the exploration and understanding of semantic shifts by allowing users to visualize different word senses and their evolution over time, while comparing the results generated by different models for the same data.

Research Questions

Building upon the motivation and challenges outlined above, this research seeks to address the following research question (RQ):

RQ: **How effectively does the developed visualization tool facilitate the exploration and understanding of semantic shifts in word senses?**

To delve deeper into this main inquiry, the study is guided by the following sub-questions:

SRQ1: What are the different features to encode to aid in visualizing semantic shifts?

SRQ2: How does the visualization tool support users in exploring different senses of a word and their evolution over time?

SRQ3: How do users' intuitive choices of topic models align with quantitative measures such as coherence scores, and how does the tool influence their model selection process?

The main research question is answered with the help of a user study containing a blend of quantitative and qualitative analysis of task-based experiments, qualitative interviews, and post-experiment questionnaires to comprehensively assess the tool's impact on model selection, usability, and engagement. The first sub-question is answered with the help of a literature survey and formulation of a taxonomy for semantic shift visualization. The second and third sub-question is answered by analyzing the results of the user study to gauge the effectiveness of the proposed visualization tool.

The rest of the thesis is organized as follows: Chapter 2 elaborates on the related work in the domain of automatic semantic shift detection and the state of visualization methods in the field. Chapter 3 describes the data used for creating the models and performing the experiments. Chapter 4 introduces the methodology used for creating the detecting word senses and their evolution. Chapter 5 introduces the visualization tool, its design and implementation. Chapter 6 explains the design of the user study, its components, and the metrics to be measured. Chapter 7 discusses the results of the study, its implications on the research questions and the limitations. Chapter 8 concludes the thesis and suggests directions for future work.

Chapter 2

Related Work

In this section we survey the related work for literature related to the research question. The related work surveyed falls under two main categories: ways of automatically detecting semantic shifts, and ways of visualizing and/or interacting with the results of these shifts. We summarize the findings by listing down the different types of visualizations commonly used in literature, the features they measure, and their advantages and disadvantages.

2.1 Detecting Semantic Shifts

2.1.1 Defining a Semantic Shift

Before we move into the methods of detecting semantic shifts, we want to provide some background about them, specifically, considering the linguistic cases of polysemy and synonymy, by using Wang et al's [5] definition of a concept, where a concept C at some moment in time t is defined as a triple $(label_t(C), int_t(C), ext_t(C))$, where $(label_t(C)$ is a String, $int_t(C)$ refers to the intension of C and is a set of properties, and $ext_t(C)$ refers to the extension of C and is a set of things that the concept extends to. Either of the three elements in the triple changing can be cause for a concept drift to occur. Synonymy, then, occurs when the extension (or usage) of a concept is shared, but the labels differ, and polysemy occurs when the label of a concept is shared, but its extensions (usage) differ.

Considering the example of the word “bank”, if one were to detect its polysemy, then we need to look at the different concepts C being represented by the single label (bank), and would find its co-occurrence with different context words representing the concept of a financial institution, or a river bank. For detecting its synonymy, we consider one of the concepts represented by “bank” (this could be the major concept it represents, and in this case we will consider it to be a financial institution), and search for labels that describe this concept to a sufficient degree. Here we might find labels like “bank”, “safe” and “vault”, which all represent the concept of a financial safekeeping institution to varying degrees.

For the purpose of this research, we will consider the extension of a concept to be its usage in the corpus, and follow the distributional hypothesis of Firth [6], which implies that the semantics of a word are defined by its co-occurrence with other words. Alternatively, the extension of a word can be gathered by how it co-occurs with other words and in what context. In order to determine the co-occurrence and context, we explore the following methods that provide word contexts within a textual corpus.

There has been extensive work documented on the emergence of methods used to detect semantic shift, starting with frequency-based methods to word embeddings being the latest and most prominent method (refer to Kutuzov et al. [7], Tahmasebi et al. [3], Tang, X. [8] for a detailed discussion). For the context of this research, this section will briefly cover these methods, their benefits and limitations.

2.1.2 Mathematical Models for Semantic Shift Detection

Word embeddings are a type of Natural Language Processing (NLP) technique that represents words in a vector space by assigning a fixed vector to each word in the vocabulary [9]. They are learned by analyzing the co-occurrence patterns of words in a large corpus of text based on their meanings and relationships with other words. These resulting embeddings capture semantic and syntactic relationships between words based on their statistical distribution in the corpus.

The embeddings based method is the most commonly used method in the literature for automatic semantic shift detection, with the methods broadly falling into two types: static and dynamic. In parallel, topic-based models serve as another method for detecting semantic shifts. Comparatively, while embeddings-based methods rely on vector representations of words and their statistical patterns, topic models provide a different perspective by identifying latent topics within text corpora. This subsection briefly describes these methods.

Static Word Embeddings

In the literature of semantic shift detection, the pre-dominant approach is that of Hamilton et al. (2016) [10], which uses neural word embeddings to generate embeddings of a diachronic corpora. It works by training word embeddings on the two corpora, aligning the spaces, and then ranking the words by the cosine-distance between their representations in the two spaces, where large distance is expected to indicate significant change in meaning. The most popular way to create these word embeddings is to use neural networks such as Word2Vec [11] or GloVe [12].

Another method is based on the use of nearest neighbours of a word as a proxy for word meaning [13], [14], [15], where each word's meaning is represented as a set of its top-K neighbours and two words are said to be similar if their neighbours are sufficiently similar. Because this method is context-free, meaning that the word embeddings do not consider the context of the word while creating its embeddings, the embeddings need to be created for each time period one wants to observe.

Dynamic Word Embeddings

Dynamic or contextualized word embeddings differ from the static embeddings in that they capture not only the meaning of individual words but also their context within a sentence. Unlike traditional static word embeddings, which assign a fixed vector representation to each word regardless of its context, dynamic embeddings take into account the context in which a word appears. This means that they do not need to be trained specifically on different time periods. The popular approach is to use the transformer architecture [16] to generate these contextualized embeddings. During training, the transformer processes each word in a sentence in the context of the surrounding words, encoding the entire sentence as a sequence of contextualized embeddings. These embeddings capture not only the meaning of each word but also its relationships to the other words in the sentence.

These contextualized word embeddings help address the particular shortcoming of a static embedding in which each word is represented by a single vector for each time period, essentially having a single vector for all senses of a polysemous word. By retaining the context, the different senses of a word are able to be represented differently. They make use of pre-trained language models (most notably, Bidirectional Encoder Representations from Transformers (BERT) [17]) to generate these contextual embeddings.

Topic-Based Models

While embedding methods are useful for detecting whether a change has occurred, they do not tell us *what* changed since they do not recover the senses. Neural embedding based methods focus on word-level shifts in meaning, while shifts in concepts or senses are covered by topic-based models.

Topic modeling methods like Latent Dirichlet Allocation (LDA) [18] are most commonly used to generate topics from the corpus, which can be used as a proxy for the *concepts* present in the corpus. Tracking these concepts over time gives an indication of whether a shift in the concept has occurred by observing whether the words describing the topic have changed. Topic models work by analyzing the distribution of topics associated with specific words over time. For example, one approach is to train a topic model on a historical corpus of text and a contemporary corpus of text and compare the distribution of topics associated with specific words between the two corpora. If there is a significant difference in the distribution of topics associated with a word between the historical and contemporary corpora, it may indicate a semantic shift. The topic model can then be used to identify the specific topics associated with the word in each corpus and analyze how the topics have changed over time.

The usage of topic modeling to generate document level topics as a proxy for concept detection is not new and has been used before in previous works [19], [20], [21], [22]. The underlying assumption of this approach is that context shifts or shifts of semantic frames are closely related to topic shifts [23].

Recently, neural topic modelings, which also make use of pre-trained language models, have been widely used to generate topics as the results are considered to be more meaningful and coherent [24]. Of these, Top2Vec [25] and BERTopic [2] are popular choices and have found success in detecting concepts from a corpora and its change over time [21], [26], [19].

2.1.3 Model Selection

For detecting synonymy, the ideal way is to use a topic model by detecting similar concepts represented by different words.

In contrast to prior studies primarily utilizing the LDA method [18] for topic modeling, we have opted for the neural topic model BERTopic [2]. This choice is driven by BERTopic's demonstrated ability to enhance topic coherence through transformer-based pre-trained language models [24], [2], resulting in more contextually meaningful embeddings. Furthermore, we favor BERTopic due to its built-in feature for dynamic topic modeling, eliminating the need for pre-identifying stable topics before tracking them over time.

This is especially useful when the number or size of topics to be found from the data are not known beforehand. The input for this model is raw texts or summaries from each document along with their timestamps, and the expected output is a list of topics over time with a ranked lists of words describing each topic.

For polysemy detection, we want to observe each word individually, rather than a concept, and so contextualized BERT embeddings are useful for this scenario, since it encodes context and hence a single word can have multiple vectors defining it based on different contextual usage. The expected input is the same again, with raw texts or summaries from each document along with their timestamps, and the expected output is a vector describing the so-called co-ordinates of a word in vector space. The similarity between two words can then be compared by comparing the cosine distance between them.

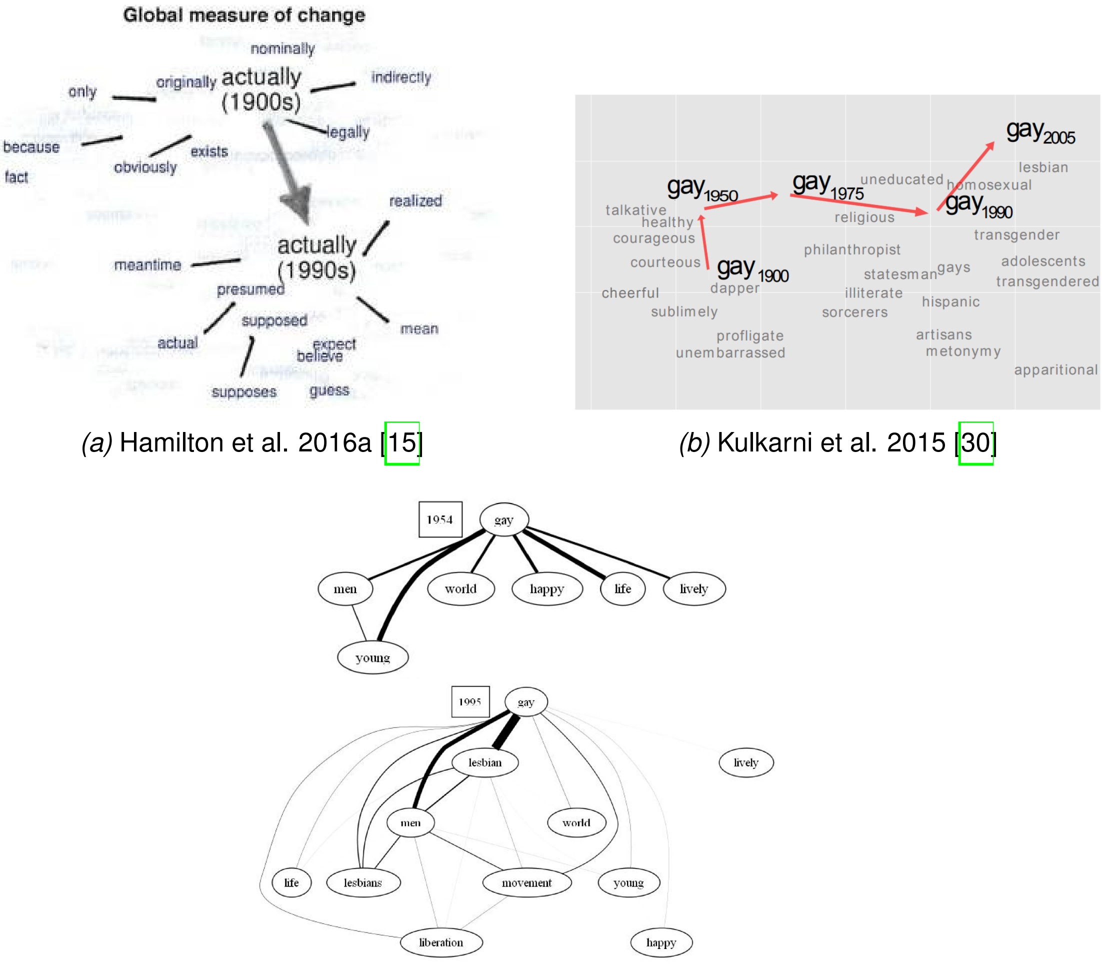
2.2 Visualizing Concept Drift

Visualization and analysis of diachronic conceptual change belong to an emerging and powerful research field of interactive visualization for computational linguistics [27]. Its purpose is to let users understand models of language and their abstract representations, and to visually uncover patterns in language [28]. Due to the nature of computation of semantic change data, which is large, complex, and multi-dimensional, it has become very common to find ways to visualize these shifts to understand and make sense of them.

The different visualizations covered in this survey, and explained in the subsequent sections, could be broadly divided into two types: static visualizations and interactive systems. The static visualizations consist of singular visualization methods, whether a novel method or an existing one, repurposed to visualize different kinds of concept or semantic shifts. The interactive systems, on the other hand, are designed to be used interactively by the user and usually consist of different individual visualizations showcasing a task-specific feature. Used in conjunction, such a system is meant to solve an end-to-end semantic shift query from different viewpoints.

2.2.1 Static Visualizations

One of the popular ways to visualize the semantic shift of words is the so-called "word graphs" that have been facilitated by dimensionality reduction techniques such as Principal Components Analysis (PCA), Latent Semantic Analysis (LSA) or the popular t-distributed Stochastic Neighbor Embedding (t-SNE) [29]. They are used to plot "trajectories" of word meaning over time in vector spaces using 2D plots. "By showing points that represent the meaning of the same words at different years or decades on the same 2D plot ([15], [30], [20]), and optionally connecting them with arrows, a single static view can show how the words changed their meaning over time, by simply following their "trajectories". Typically some background reference terms are added along these "trajectories" to ground and explain their meaning." [28] Figure 2.1 shows some examples of Word Graph visualizations displaying meaning change across time periods via change in meaning of its contextual words.



(a) Hamilton et al. 2016a [15]

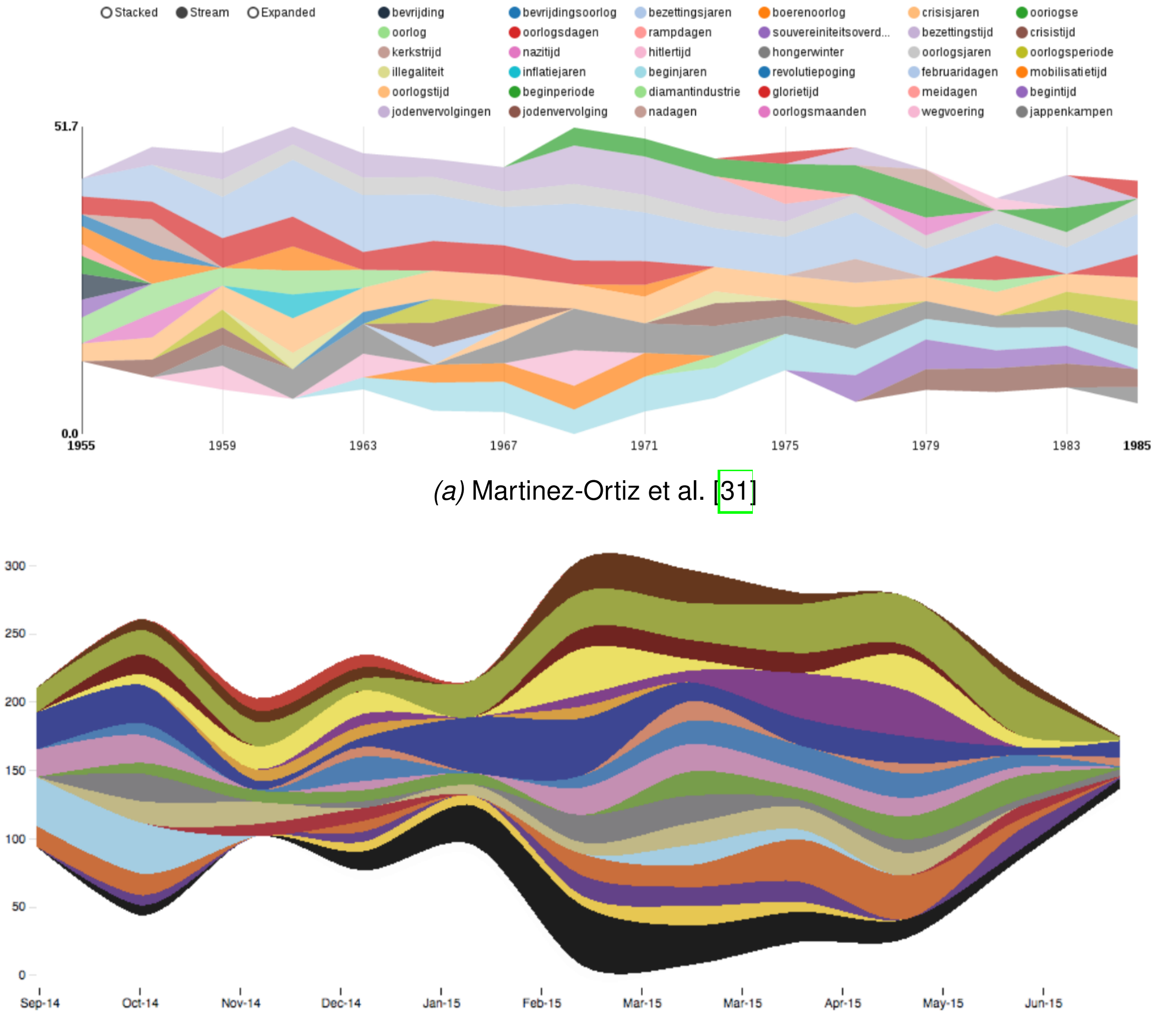
(b) Kulkarni et al. 2015 [30]

(c) Wijaya et al. 2011 [20] *(cropped to preserve space)*

Figure 2.1: Word Graph Visualizations showing the change of a meaning of a word across multiple time periods by change in meaning of their neighbouring words

Figure 2.2 shows a steamgraph, which is a common method for tracking changes in concepts and word sense change visually, as used by Martinez-Ortiz et al. [31] and Becher et al [32]. It shows color-coded streams for each term, where the stream sizes represent the relative importance of the term in a period. These are useful for showing the changes over time of multiple concepts a word belongs to.

Another way of showing the word sense change over time is through the use of percentage stacked bar charts (Figure 2.3), as used by Frermann & Lapata [33] and Montariol et al. [34]. These are different since they also show the percentage of the senses (and therefore its prevalence) of the word at each time interval, making it more convenient to observe semantic changes such as broadening, narrowing, semantic shift, pejoration and amelioration [8].

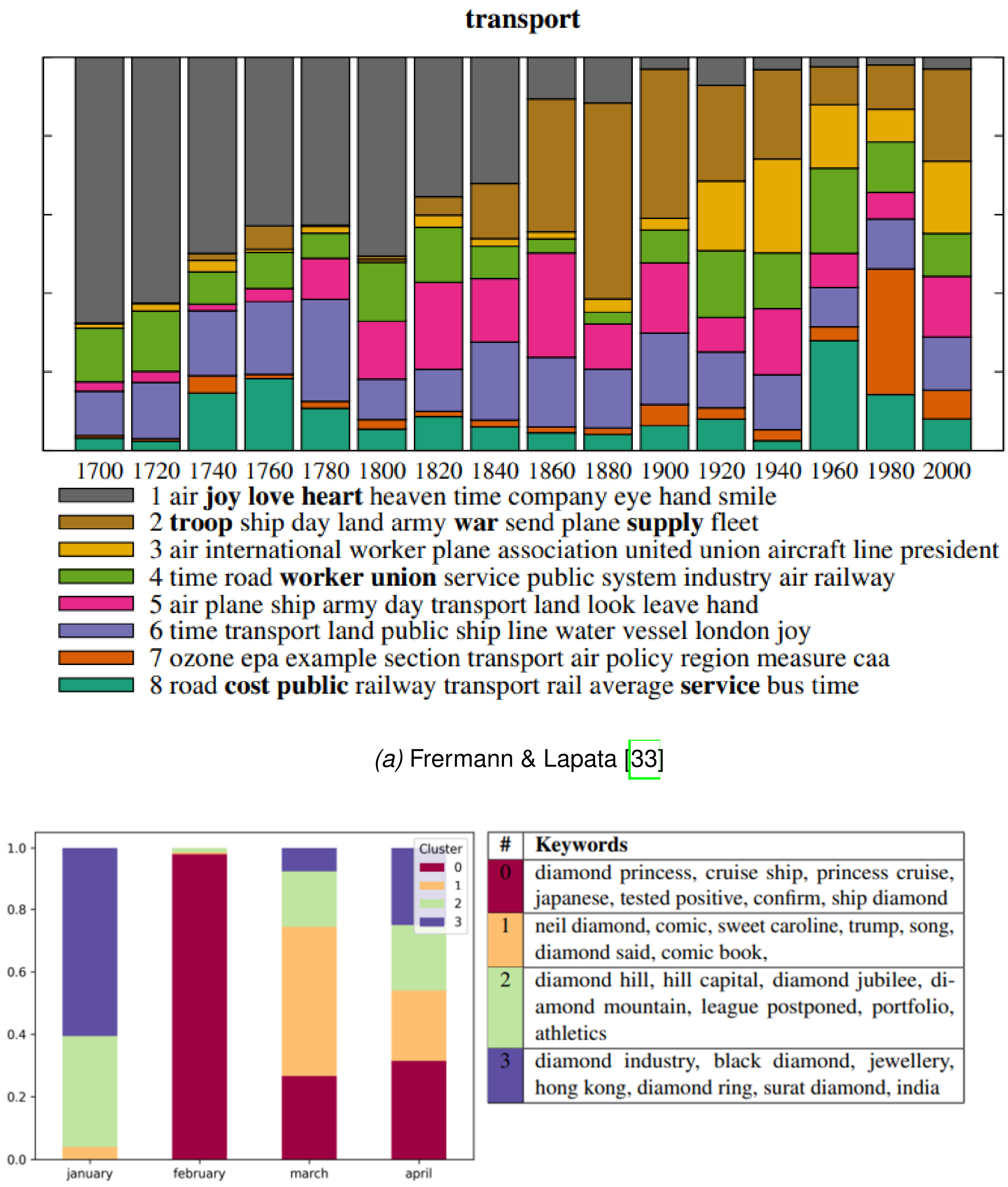


(a) Martinez-Ortiz et al. [31]

(b) Becher et al. [32]

Figure 2.2: Steamgraphs showing (a) relative importance of terms over time and (b) shifting concept representation over time

Both, the steamgraphs (Figure 2.2) and the percentage stacked bar charts (Figure 2.3), are able to encode *continuity*, i.e., the ability to track a particular concept as it is changing across time. The steamgraph is inherently able to do this due to the nature of its flow-like structure, and the percentage stacked bar charts can encode this if by keeping all stacks in the same position across all the bars.



#	Keywords
0	diamond princess, cruise ship, princess cruise, japanese, tested positive, confirm, ship diamond
1	neil diamond, comic, sweet caroline, trump, song, diamond said, comic book,
2	diamond hill, hill capital, diamond jubilee, diamond mountain, league postponed, portfolio, athletics
3	diamond industry, black diamond, jewellery, hong kong, diamond ring, surat diamond, india

(a) Frermann & Lapata [33]

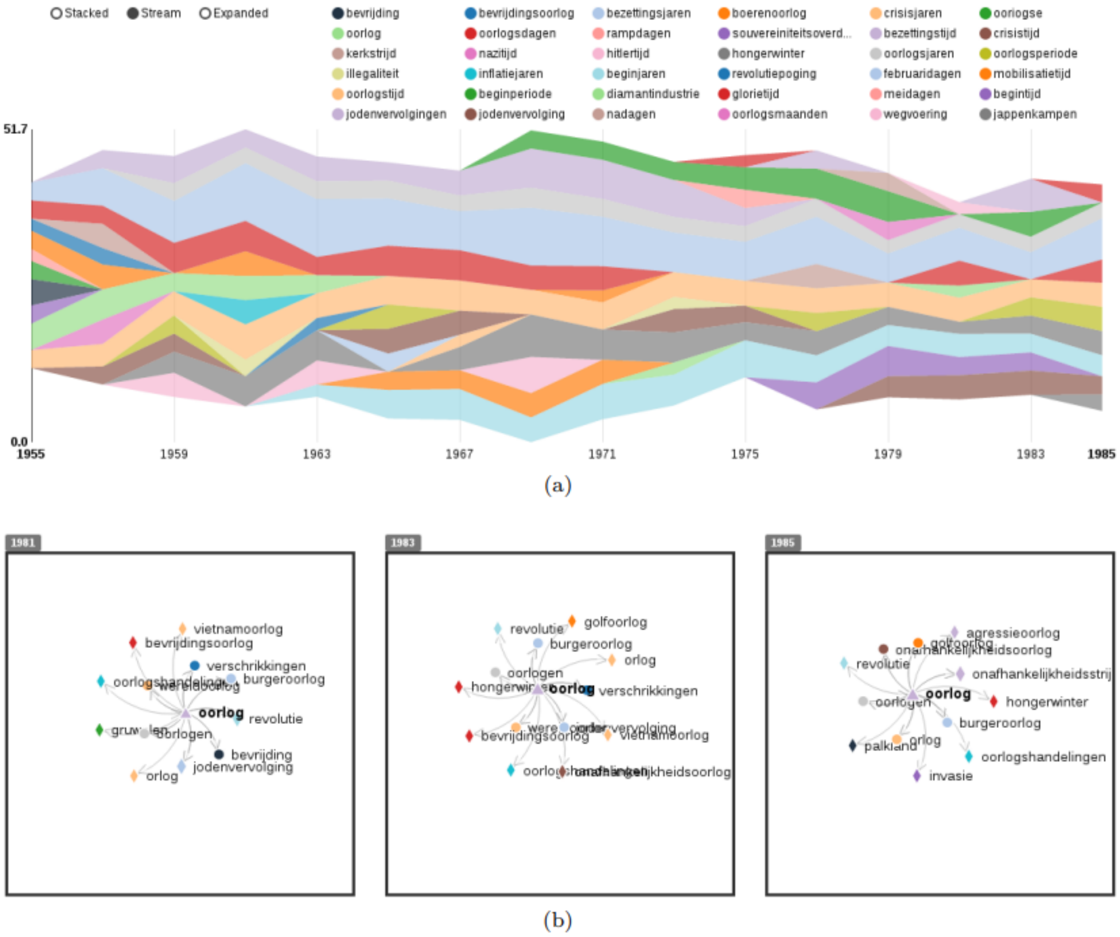
(b) Montariol et al. [34]

Figure 2.3: Percentage stacked bar charts showing proportion of concept usage and shift over time for the words a) transport and b) diamond

2.2.2 Interactive Systems

These visualization systems are usually built end-to-end, with a web-based front-end for the user to explore and with a back-end holding the semantic models. They consist of multiple visualization types used in conjunction, each one to aid in a different interpretation task for semantic shift understanding.

Martinez-Ortiz et al. (2016) [31], for example, designed *ShiCo*, a system for visualizing shifting concepts of Dutch words over time. Their system consists of two complementary graphs, a steam graph that shows a differently coloured stream for each term, and a network graph, that displays how different terms within a period are related (Figure 2.4).

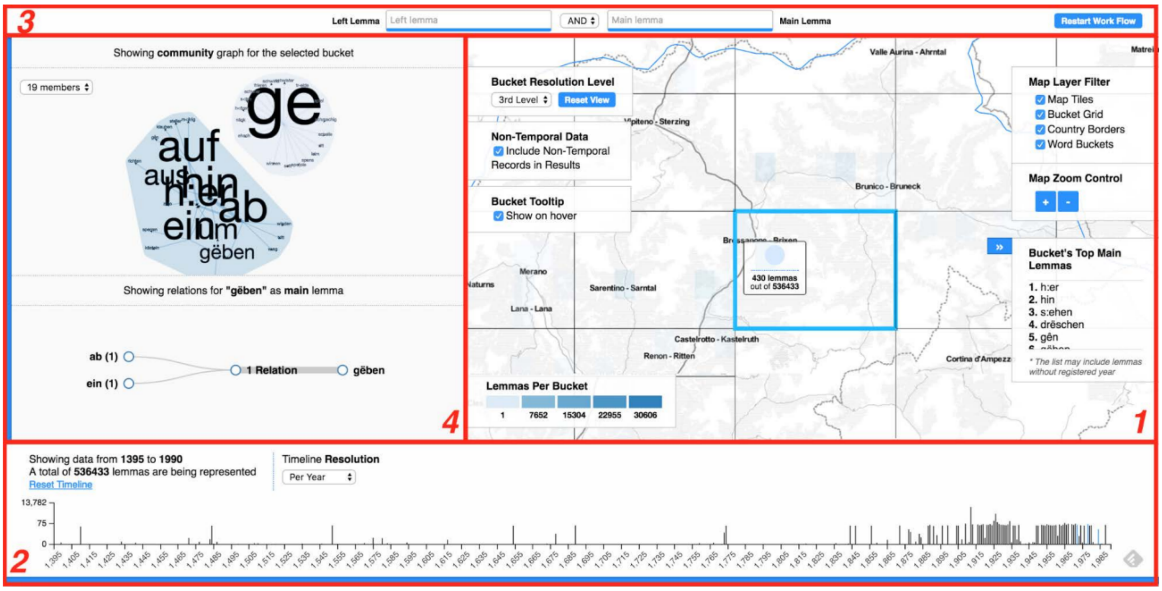


(a)

(b)

Figure 2.4: Screenshots of ShiCo interface, using a steamgraph (above) to denote temporal change, and a network graph (below) to showcase word similarity.

Benito et al. (2016) [1] (Figure 2.5) introduced an interactive spatio-temporal visual analysis tool that helps users to search for a lemma and get three types of



Showing **community** graph for the selected bucket

Showing relations for "**gëben**" as main lemma

Showing data from **1395** to **1990**
A total of **536433** lemmas are being represented

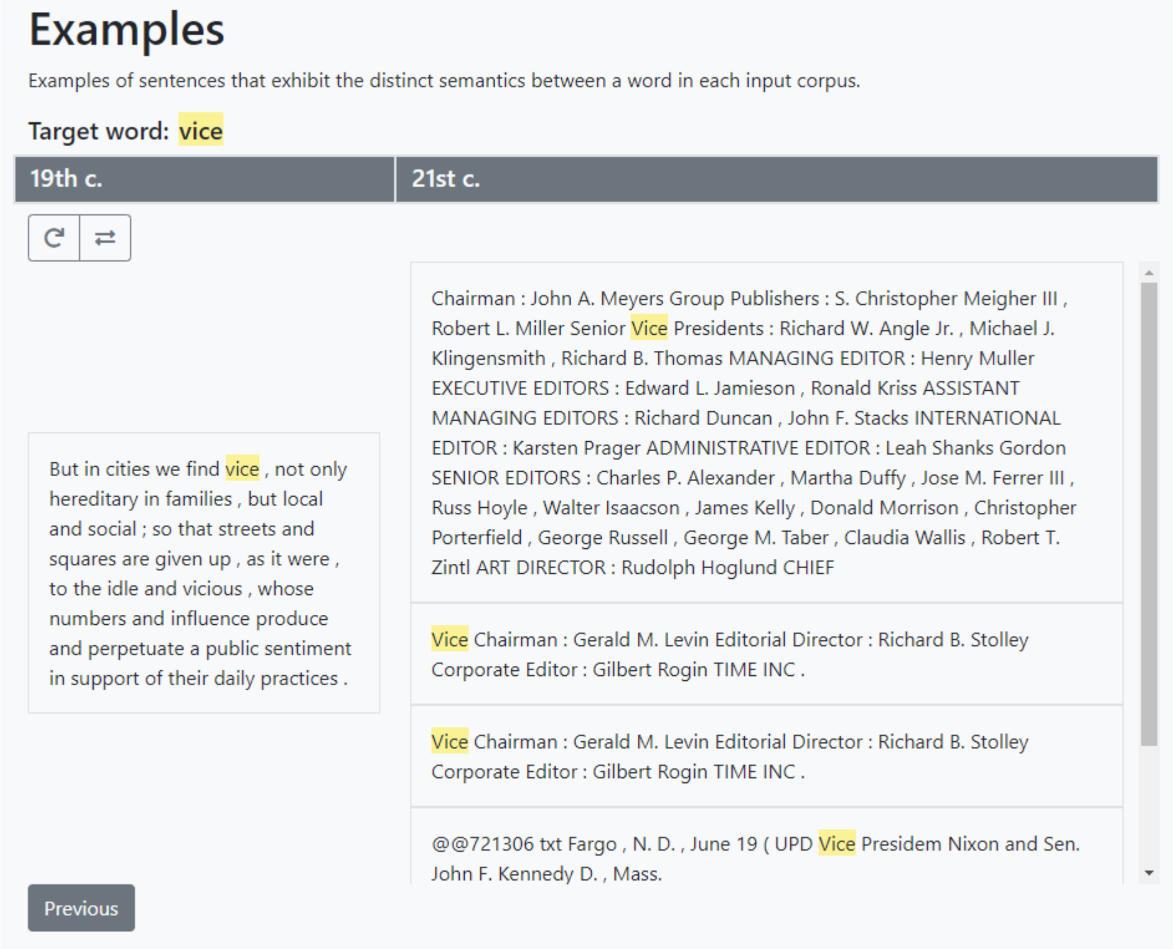
430 lemmas out of 536433

* The list may include lemmas without registered year

Figure 2.5: Interactive spatio-temporal interface. 1) Spatial projection/map. 2) Temporal projection or timeline. 3) Textual search bar. 4) Network analysis view (Benito et al. (2016)) [1]

information: the spatial distribution of corresponding lemmas over a map (View No. 1, Figure 2.5), network of relationships among the lemmas in question and other lemmas (View No. 4, Figure 2.5), and the temporal distribution of the lemmas along the timeline (View No. 2, Figure 2.5). The spatial view is shown with the help of a map and makes up the main view of the system. The timeline in View No. 2 makes up the temporal aspect of the system, and the network analysis view shows the words related to the lemma.

The SenSE Toolkit [35] is also an example of an interactive system that follows the same idea of having an interactive tool available to the user to assess results in real-time on a web interface. The SenSE toolkit is unique in the sense that it also offers example sentences of a word whose meaning has shifted in the different time periods of use (Figure 2.6), such that the user can themselves inspect a word used in two different ways, which, we feel, makes for a better understanding experience.



Examples

Examples of sentences that exhibit the distinct semantics between a word in each input corpus.

Target word: vice

19th c.	21st c.
But in cities we find vice , not only hereditary in families , but local and social ; so that streets and squares are given up , as it were , to the idle and vicious , whose numbers and influence produce and perpetuate a public sentiment in support of their daily practices .	Chairman : John A. Meyers Group Publishers : S. Christopher Meigher III , Robert L. Miller Senior Vice Presidents : Richard W. Angle Jr. , Michael J. Klingensmith , Richard B. Thomas MANAGING EDITOR : Henry Muller EXECUTIVE EDITORS : Edward L. Jamieson , Ronald Kriss ASSISTANT MANAGING EDITORS : Richard Duncan , John F. Stacks INTERNATIONAL EDITOR : Karsten Prager ADMINISTRATIVE EDITOR : Leah Shanks Gordon SENIOR EDITORS : Charles P. Alexander , Martha Duffy , Jose M. Ferrer III , Russ Hoyle , Walter Isaacson , James Kelly , Donald Morrison , Christopher Porterfield , George Russell , George M. Taber , Claudia Wallis , Robert T. Zintl ART DIRECTOR : Rudolph Hoglund CHIEF Vice Chairman : Gerald M. Levin Editorial Director : Richard B. Stolley Corporate Editor : Gilbert Rogin TIME INC . Vice Chairman : Gerald M. Levin Editorial Director : Richard B. Stolley Corporate Editor : Gilbert Rogin TIME INC . @@721306 txt Fargo , N. D. , June 19 (UPD Vice Presidem Nixon and Sen. John F. Kennedy D. , Mass.

Figure 2.6: SenSE Toolkit with example sentences for the word “vice”. Left hand side shows the meaning in the 19th century referring to the sin of vice. Right hand side shows increased frequency of usage of “vice” primarily in book credits as “Vice President”, “Vice Chairman” etc.

2.3 Features Measured

A goal of this research is also to find the appropriate features that must be measured in a visualization to most effectively convey the information needed. For diachronic semantic shifts, as the name suggests, time plays perhaps the most important role in understanding the shifts. Indeed, almost all visualizations incorporate time in their methods in one way or another. In this section we examine the different ways chosen to visualize the temporal dimension.

2.3.1 Visualizing Time as a Dimension

There are many ways that time can be visualized in general visualizations; the best use of which depends on the context it is used in, the type of data being visualized, and the information one wishes to convey. Since incorporating the temporal aspect is essential in visualizing any kind of diachronic semantic shift, we wanted to see the ways in which visualizations for semantic shift have made use of them.

From a user and interaction perspective, there are four main methods for temporal encoding of an interactive system (Figure 2.7): (linked) timelines, animation, superimposition, space-time cube. The linked timeline view is a prominent method among these, with Benito et al's timeline interface [1] (Figure 2.5) an example of it.

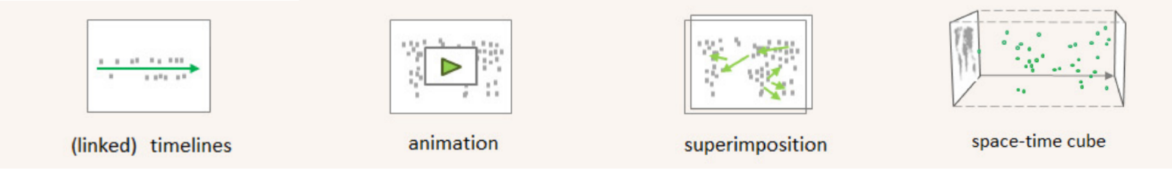
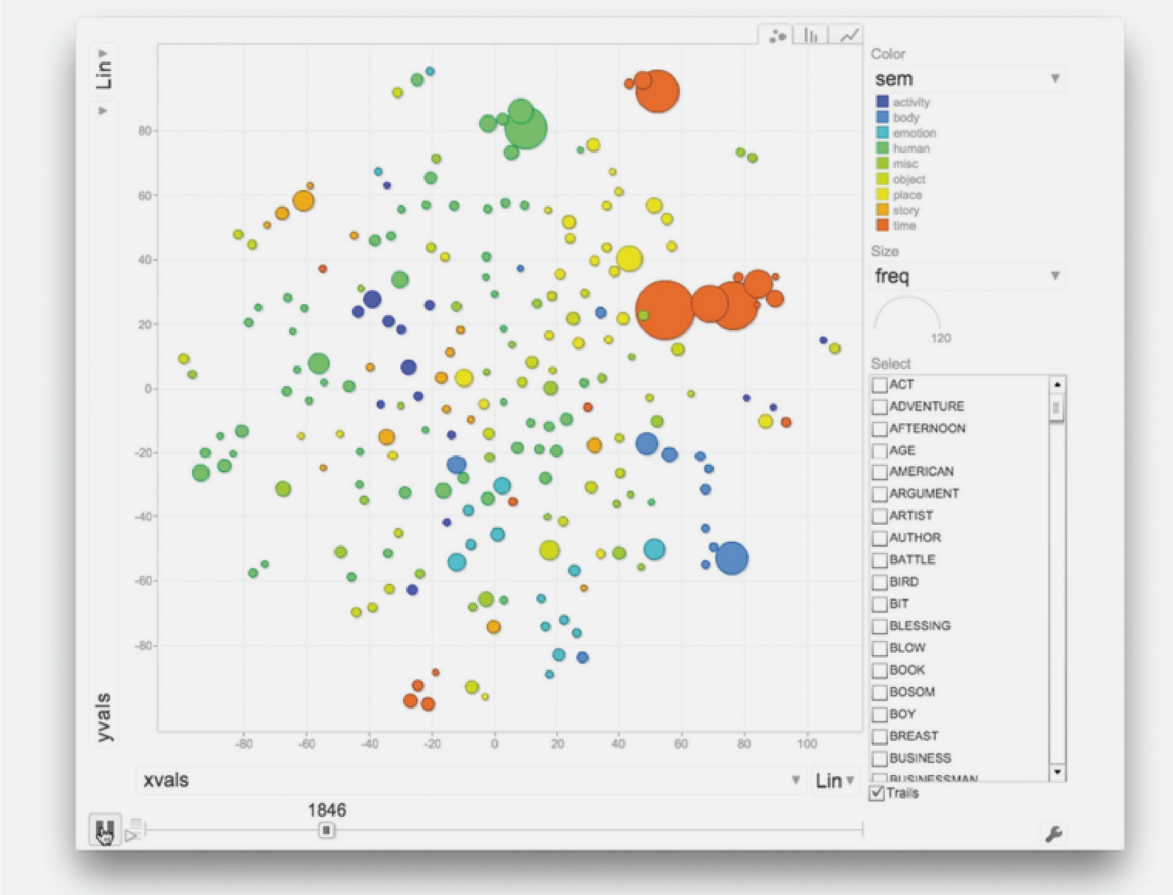


Figure 2.7: Methods of temporal encoding

Another option for encoding temporal aspects is to have an animated view, where the able is able to visually observe the changes as they happened temporally. As shown in Figure 2.8, Hilpert and Perek [36] made use of animations, which they called "motion charts", by animating scatterplots and showing how the changes unfolded over time. The "play" button at the bottom adds elements to the visualization based on when they made an appearance.

Superimposition is a technique that merges multiple temporal layers or snapshots into one visualization, with temporal data aspects often being distinguished by different colors. A simple example is from the SenSE toolkit [35] is shown in Figure 2.9, which visualizes the 19th century meaning of the word "staff" in a 21st century context and vice versa.

The space-time cube is a more advanced representation that can be fully appreciated only in an interactive dynamic visualization. It builds on 2D planes of encoded spatial data dimensions, and maps time to an additional spatial dimension, i.e. the orthogonal z-axis. We have not come across any works that make use of this visualization in a semantic shift context, but because of its usefulness in encoding the spatial and temporal aspect together interactively, it is still a useful way to view these changes, and its effectiveness needs to be experimented with.

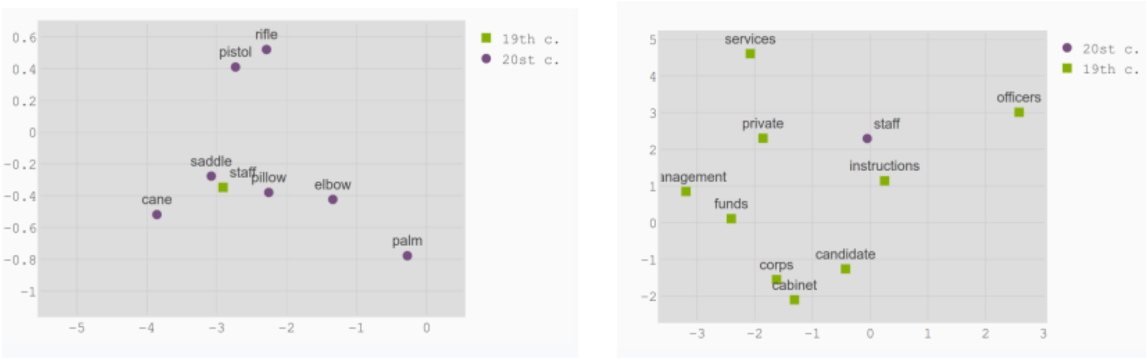


Color: sem

Size: freq

☑ Trails

Figure 2.8: Example of using animation to encode time.



(*a*) 19th century (*b*) 21st century

Figure 2.9: Example of using superimposition to encode time. Temporally different data distinguished by different colours

2.3.2 Visualizing Non-Temporal Measures

There are a number of other measures used to convey semantic shifts, each used for a different purpose. For example, showing similar words and the degree of their similarity is a way of placing the target word in the context of other words. Here, word graphs (Figure 2.1) use distance or background terms to show similar words and the thickness of the line joining the words indicate their degree of similarity.

Change of word sense is also used to show at a glance the different senses a word has taken over the years, with colours and width showing the proportion and frequency of the senses. They can be seen in the steamgraph visualization (Figure 2.2) and in the percentage stacked charts (Figure 2.3). It is also useful to show the type of shift occurring (whether narrowing or broadening) or a clear indicator of whether a word or concept's meaning has stayed stable or shifted through the years.

Showing the spatial component is important when trying to understand the variation in linguistic shift not (just) across time but also across location. Thus, maps remain the best way to show geographic information due to the ease of recall for the user, but we have not come across any works that use maps to show this. There have been other ways to show semantic shift across geography, albeit by choosing the locations of interest beforehand. For example, Kulkarni et al [37] showed the difference in linguistic drift between US and UK speakers by making use of a superimposed scatterplot with the colour indicating the location; although this can be considered more of a categorical encoding rather than a geospatial one. In our previous research [38], we made use of geographic word clouds to compare semantic shift across locations over time. This also, however, requires the prior selection of locations manually.

While many visualizations show that a change in meaning of a word has occurred, not all show the *continuity* of the word, i.e., tracking how the word has changed over time. Visualizations like the steamgraph are able to make use of their inherent temporal encoding to show this continuity. In previous work by the author [38], we also presented the spiral line chart that encoded the continuity of a word over different decades.

2.4 Evaluating Concept Drift Visualizations

Despite surveying plenty of visualization methods, we found no evaluations for the visualizations themselves. While some papers contained evaluations for the semantic shift, i.e., measuring the correctness of the obtained results, there were no evaluations from a visual analytics perspective to test the usability of the visualizations themselves.

Moreover, it was expected that interactive visualization systems should undergo usability testing compared to static visualizations, as direct user interaction is the ultimate goal. However, we were unable to find any such testing. Considering the importance of visualizations in detecting semantic shifts, it is crucial to establish

Table 2.1: Types of semantic shift visualizations and the features they measure

	word co-occurrence	degree of similarity	continuity	word sense change	concept change	word frequency	word context
Word Graph [15] [30] [20]	✓	✓					✓
Steamgraph [31] [32]			✓	✓	✓	✓	
Percentage Stacked Bar Chart [33] [34]			✓	✓		✓	
Network Graph [31]	✓	✓					✓
Radial Bar Chart [38]	✓	✓					✓
Spiral Line Chart [38]	✓		✓				

a framework for evaluating them from a usability perspective. Evaluating such a system is, therefore, one of the expected outcomes of this research.

2.5 Summary

In this section, we discussed the related work for each of the sub-questions. In Table 2.1 we list the different visualization methods mentioned and the features they measure, and in Table 2.2 we list down their advantages and disadvantages. What is missing in most of these visualizations is mainly the evaluation of their systems across the multiple criteria. There has been some evaluation to test how accurate they are in the information they visualize, but not much in terms of the function (the usability and effectiveness as a tool to help the user understand the information they are looking for), and their aesthetics (the quality and appeal to draw a new user in to use these tools). It is also important in the context of exploration of digital data to have these systems be interactive, mainly due to the sheer size of the data that is being attempted to visualize. While static charts are useful to show information for a particular word or concept, adding in interactivity to these charts, and making them work together as a system invites the user to explore the data from different views and draw novel conclusions from it. For the use case of seeing the history of a search term, it is useful to not only see how the word itself has changed over time, but also whether the concept being represented by the word has changed, and in this regards most systems focus on one or the other.

Table 2.2: Different visualization types along with the advantages and disadvantages they offer

Visualization Type	**Advantages**	**Disadvantages**
Word Graph	- Intuitive to understand - Easier to inspect context words to verify a shift	- Context words may need manual selection - Accurate context words need non-trivial filtering - Background knowledge may be needed to understand context words - Degree of shift of words is not shown
Steamgraph	- Includes a temporal feature with stream flow - Good for understanding concept level changes - Shift detection without depending on context word knowledge - Easy to see dominant concepts	- Hard to read with multiple concepts - Complex legend with multiple colors can cause information overload and accessibility issues - Can be hard to to understand if it is a static chart that doesn't allow hovering to view details
Percentage Stacked Bar Chart	- Shows lesser-used word senses that may be overlooked in context-based visualizations - Visualizes narrowing or broadening of senses and unstable senses - Shift detection without depending on context word knowledge	- Hard to compare word sense proportions if near senses are similar in size - Trade-off between ranking stacks by rank or temporality
Network Graph	- Provides a broader view of clusters of related words and their relationships - Good for model explainability of semantic similarity between words	- Unintuitive or noisy on a large scale - Needs filtering or prior knowledge for optimal use - Loss of information due to 2D scaling
Radial Bar Chart	- Shifts are easily visible at a glance - Does not depend on understanding context words to know whether a shift has occurred	- Requires manual selection of words to show most salient context words
Spiral Line Chart	- Tracks continuity of a word	- Requires manual encoding of categories

Chapter 3

Data

3.1 Data Collection

We use the Online Computer Library Center's (OCLC's) WorldCat[1] catalog as the source of our data. Online Computer Library Center (OCLC) is a global library co-operative that provides services to thousands of libraries around the world. OCLC collects and stores metadata about library resources, such as books, journals, and audiovisual materials, to help libraries manage their collections and provide access to information. It stores this data in the Machine-Readable Cataloging (MARC 21)[2] data encoding and transmission standard.

For the Machine-Readable Cataloging (MARC) 21 bibliographic data format, the possible bibliographic tags for a record take the form of a 3-digit code ranging from 001 to 999 (not all of these tags are present for every record and their inclusion depends on the type of resource being described). The detailed list of all the fields can be found on the Library of Congress page[3]; for the purpose of this research, however, the fields chosen are shown in Table 3.1, along with their description and feature type.

We use the `Summary` field as the main text input for creating word embeddings. To define time periods, we extract the `Year of Publication` from the `Control Field 008`, which is a 40-character field that provides bibliographic information. We also extract from it the `Place of Publication` to identify a geographical feature, and the `Language` to filter publications by the English language. The `Title` and `Author` fields are additional metadata to help identify a publication, while the `Topics` field contains subject entries that describe a publication, including names of events or objects. The `ISBN Number` and `Dewey Decimal Number` are unique identifiers used as classification metadata for unique publications. We also extract the `workID` field, which is

[1] https://worldcat.org

[2] https://www.loc.gov/marc/bibliographic/

[3] https://www.loc.gov/marc/bibliographic/

an OCLC identifier that has a unique number for each different work. We use this to keep only unique copies of a document that might have multiple variants, leading to duplicate values.

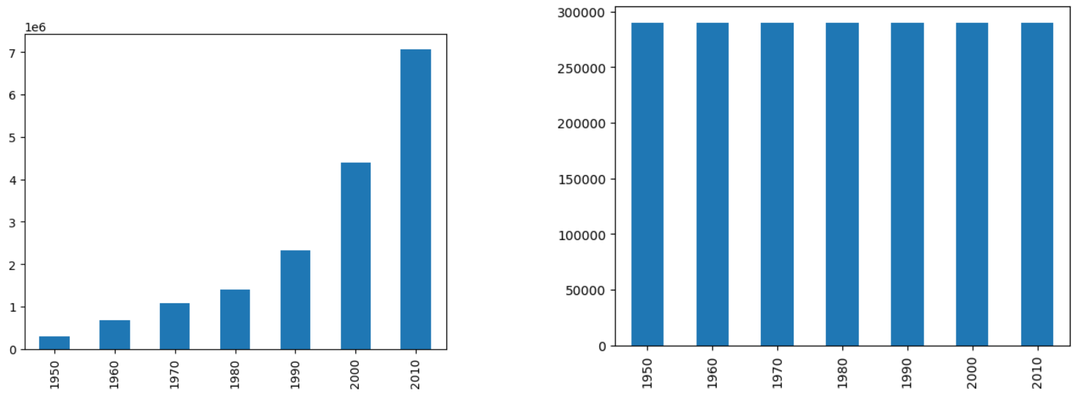
Type	Feature Name	MARC 21 Code and Description
Textual	Summary	**520**: A string containing a summary or abstract describing the material
Metadata	Title	**245**: A string containing the title and subtitle of a publication
	Year of Publication	**008(7-11)**: A 4-digit code in the control field giving the year of publication
	Place of Publication	**008(15-17)**: A 3-digit code giving the International Organization for Standardization (ISO) code of the place of publication
	Language	**008(35-38)**: A 4-digit code giving the language of publication
	Author	**700**: Author of the publication
	Topics	**650**: Manually added topical subjects for a publication
Classification	ISBN Number	**020**: Unique ISBN number for a publication
	Dewey Decimal Number	**082**: Dewey decimal classification number
	WorkID	**workID**: OCLC unique work identification number

Table 3.1: Selected features of the final dataset

3.2 Data Filtering

The data available consists of 400 zipped files of 4.05 GB each in the `xml` format. To select the records for analysis, we first filtered the dataset based on the availability of summaries, then filtered for whether the document language recorded was English, and finally selected all documents from 1950 to 2019. This led to a dataset of 27,796,256 documents. Due to the nature of bibliographic tagging, there are many instances of duplicate records, mainly due to multiple editions. Using the `WorkID` field provided by OCLC to detect unique works, we removed all the duplicate IDs and kept the earliest version of a document that contained a summary. This resulted in the final dataset of 17,247,530 final documents.

The decade-wise distribution for the final dataset is skewed (Figure 3.1(a)) due



(a) Original distribution (y-axis is millions of documents)

(b) Distribution after random under-sampling

Figure 3.1: Decade-wise document distribution

to much larger number of publications documented in the later decades than in the past. To make the decade distributions comparable so that there is equal representation of documents from all decades in the dataset, we randomly under-sampled data from all decades to keep 290,000 documents in each decade (since 290,000 was the lower limit of the smallest present decade data: 295,680 documents in 1950). The final dataset for use is shown in Figure [3.1](b). Table 3.2 shows an example of a random sample record. It provides an example record for each of the features shown in Table 3.1.

3.3 Corpus Preparation

The final corpus for analysis is prepared by selecting all documents from the filtered data relevant to the target word. For a target word *w*, all documents containing the target word are filtered to form the corpus used for topic generation. The search term is processed, escaping any special characters to avoid unintended regex patterns. A regular expression (regex) pattern is then constructed, matching the search term as a whole word in a case insensitive way and preventing partial matches from being considered. From the corpus, all documents with their summary field (MARC21 code 520) containing the search term are returned, with all the features as described in the previous section.

The document count is then calculated, and only if the filtering results in more than a 1000 documents is the search term passed on to the topic modeling step. This is to ensure an adequate corpus size for the procedure and that meaningful results are obtained in subsequent semantic shift analyses.

At the end of this step we extract two indexed lists, one containing all the docu-

Feature	**Value**
WorkID	4452553674
ISBN	['9780670016747', '0670016748']
Dewey Decimal	[E]
Title	Angelina Ballerina
Author	Craig, Helen,
Summary	Angelina is a little mouse who wants more than anything else to become a ballerina. Will she be able to make her dancing dreams come true? [Pg. 4 of cover]
Year of Publication	2013
Place of Publication	ny
Topics	['Angelina (Fictitious character : Holabird)', 'Ballet dancing', 'Mice', 'Dancers', 'Board books.', 'Board books.', 'Ballet dancing.', 'Dancers.', 'Mice.']
Language	eng

Table 3.2: Example of the retrieved data

ments, and the other containing all the timestamps, to pass on to the topic modeling section.

Chapter 4

Methodology

This chapter presents the methodology used to detect multiple word senses and the usage change of those senses across time. The steps in Figure 4.1 outline the structure for overall methodology. Step 1, or preprocessing, was covered in Chapter 3, while steps 7 and 8 (Visualization and Evaluation) will be covered in chapters 5 and 6 respectively. For this section we focus on steps 2-5, forming the crux of the topic modeling methodology for sense detection, and step 6, which outlines the steps to track the sense usage over time.

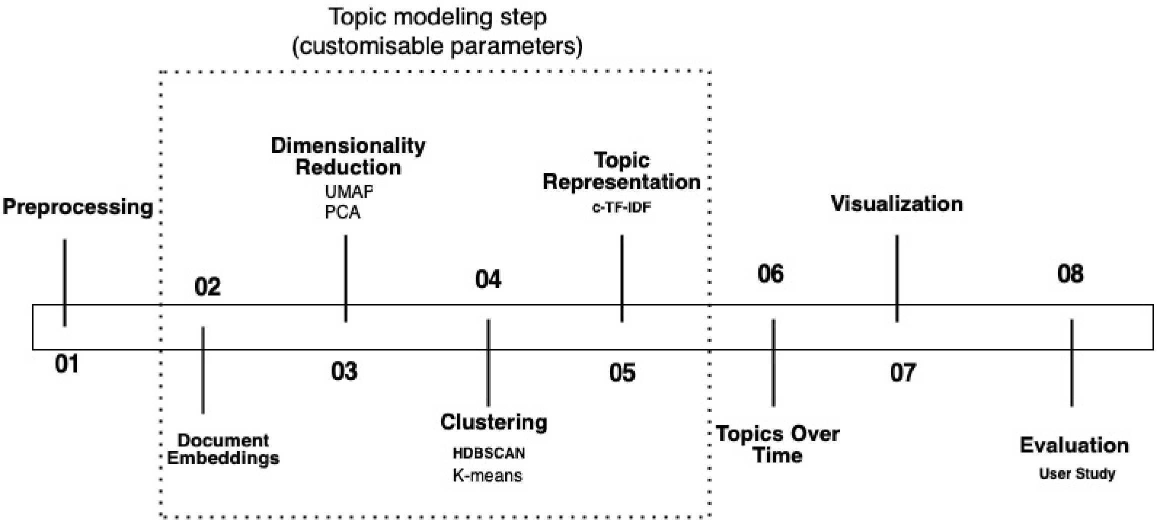


Figure 4.1: An overview of the methodology

4.1 Topic Modeling

To identify word usages, a topic modeling approach is adopted. We follow the assumption of the topic model approach for word sense modeling which assumes that each topic generated from the corpus is interpreted as a word sense.

While there are multiple approaches to identifying word senses via topic modeling (as discussed in Section 2.1.2), we choose the BERTopic [2] approach for several reasons. Since one of the research goals is to focus on the differences in outputs generated by models for various parameters, and since these parameters can vary in many different stages of the topic modeling life cycle, we wanted to choose a method that allows for high modularity and flexibility. BERTopic is designed as a series of four main stages, each of which can be performed using a different algorithm whose output can be compared separately. This modularity allows us to visually examine outputs for each stage of the process to better understand the contribution it makes. In addition to its modularity, experiments have shown that BERTopic has generally high coherence scores compared to other models, and remains competitive on diversity scores [2].

The topic modeling step consists of steps 2,3 and 4 from Figure 4.1. The prepared corpus is then used as input for the topic modeling algorithm. The following section provides an overview of the distinct phases within the BERTopic topic modeling process, and discusses the algorithms used for creating our model in each of these stages.

4.1.1 Document Embeddings

The first step is the document embeddings step. We use the default sentence-transformer `all-MiniLM-L6-v2`[1], which is an English language model that is optimized for the task of semantic similarity and for creating document- or sentence-embeddings. The model converts the text documents ("Summary" field from Table 3.2) to a high dimensional vector representation.

4.1.2 Dimensionality Reduction

Once the document embeddings are created, we apply dimensionality reduction methods to allow the ensuing clustering step to handle the high dimensional data. We use the Uniform Manifold Approximation and Projection (UMAP) [39] approach to reduce dimensionality.

Reducing high dimensional embeddings with UMAP has been shown to improve the performance of downstream clustering algorithms, such as k-Means and Hierarchical Density-Based Spatial Clustering (HDBSCAN), both in terms of clustering accuracy and time [40]. While traditional dimensionality reduction techniques like PCA and t-SNE are well-established, UMAP has demonstrated superior perfor-

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

mance in retaining both local and global features of high-dimensional data in lower-dimensional representations [39].

We opted for default parameters when configuring the UMAP dimensionality reduction method. This decision was motivated by the intention to maintain parameter uniformity across all target words, avoiding the risk of over- or under-optimization for any specific word. The number of neighbors was set to 15, with 5 components, and a minimum distance of 0.0 while employing cosine similarity as the similarity metric.

4.1.3 Cluster Documents

At this point the reduced embeddings are then clustered using the HDBSCAN [41] algorithm. It helps to find outliers and as a result does not force documents into clusters where they do not belong, leading to better topic representation.

One of the primary motivations for choosing HDBSCAN is its ability to automatically detect the number of clusters, a crucial factor in our topic modeling endeavor. In many clustering tasks, including topic modeling, manually specifying the number of clusters can be a formidable challenge. This necessitates a priori knowledge of the expected number of senses a word might possess, which may not be available or feasible for a broad range of search terms. Our objective, as part of exploratory analysis, is not to predetermine the number of unique senses a word might exhibit but rather to discover these senses organically. HDBSCAN aligns perfectly with this goal. In contrast, K-Means, while widely employed due to its simplicity and computational efficiency, is best suited when the value of ‘k’ (in our context, the number of word senses) is known in advance. For unsupervised classification tasks like automatically determining the number of word senses, HDBSCAN is better suited.

Additionally, while k-Means is the fastest and most scalable algorithm, it is essentially a partitioning algorithm and does not deal well with outliers [41], tending to explicitly assign a document to each cluster, whereas HDBSCAN separates outlier documents if they do not belong to a certain cluster.

4.1.4 Topic Representation

To retrieve topic representations from clustered documents, we use a modified version of the classic Term Frequency-Inverse Document Frequency (TF-IDF) [42] procedure known as the class-based TF-IDF procedure (c-TF-IDF), which was introduced by Grootendorst in the BERTopic paper [2]. This procedure represents topics based on the collection of all documents belonging to a single cluster with the following formula:

For a term $\mathtt{x}$ in class $\mathtt{c}$:

$$\mathrm{W}_{x,c} = \|\mathrm{tf}_{x,c}\| \cdot \log\left(1 + \frac{A}{\mathrm{f}_x}\right) \tag{4.1}$$

where

$\mathrm{tf}_{x,c}$= frequency of word x in class c
f_x= frequency of word x across all classes
A= average number of words per class

This class-based TF-IDF procedure models the importance of words in clusters instead of individual documents, allowing the model to generate topic-word distributions for each cluster of documents.

Regarding topic representations, BERTopic employs its proposed Class-Based Term Frequency-Inverse Document Frequency (c-TF-IDF) algorithm for the ranking of the terms of a topic. Although it offers two parameters, namely bm25_weighting and reduce_frequent_words, neither of these align with our specific use case.
bm25_weighting is designed for smaller datasets, which is not applicable to our scenario. reduce_frequent_words is intended to mitigate the appearance of stopwords in topic representations, a concern that did not affect our model creation. Therefore, default parameters were retained for this stage of the process.

4.1.5 Automatic Model Creation for Search Terms

As part of one of the usages of the tool, we wish to compare the results of multiple models generated for a single search term. A general problem for model creation of an unsupervised nature like topic modeling is the trial-and-error process of creating multiple models with different parameters, and to then inspect and compare the models to find the best one for the task at hand. To facilitate this comparison, we had to first create multiple models for each search term in order to compare them. While in theory, a practitioner can add all their model variants for visual comparison, in practice, we select 3 models per target word for the prototype tool to maintain consistency and manageability across all target words. This decision is motivated by cognitive limitations associated with working memory, which posits a capacity constraint typically for users, ranging from holding three to five separate items in one's working memory at a time [43]. By adhering to the lower end of this range with three models as a conservative estimate, we aimed to prevent overwhelming users during the study, particularly when one of the tasks involves the selection of an optimal model from the set.

Automating the number of word senses:

A common problem with the clustering stage of topic modeling is to manually assign the number of clusters to the model. In our case, this would require some pre-requisite knowledge of the number of word senses a word might have and create a model with the number of clusters set to that size. In our scenario, this would necessitate possessing prior knowledge regarding the probable number of senses associated with each target word and configuring each model variant with a corresponding number of clusters. However, for a multitude of search terms, this manual predetermination of the number of senses becomes impractical and unfeasible. Indeed, as an exploratory analysis, the task is usually to *find* how many such unique senses a word might have. For this purpose, we decided to use the HDBSCAN method for clustering as mentioned in the topic modeling clustering section before. This is because, while K-Means has been more commonly used for clustering tasks due to its simplicity and computational efficiency, it is best used when 'k', (in this case, the number of senses of a word) is known beforehand. For an unsupervised classification task like automatically determining the number of senses of a word, HDBSCAN is more suited to our goal. Moreover, HDBSCAN detects outliers that do not clearly belong to a specific topic, whereas K-Means is forced to assign it to a particular category, leading to possible noise in the output.

Setting the parameters for different models:

As explained in section 4.1, for the topic modeling stages we use the following parameters to create each model:

- Embeddings: `all-MiniLM-L6-v2`
- Dimensionality Reduction: `UMAP`
- Clustering: `HDBSCAN`
- Topic Representation: `c-TF-IDF`

There exist different possible combinations for creating a model, by substituting a different algorithm for each of the different components, along with setting different hyperparameter values for each of the algorithms. In order to automate the creation of different models, then, we decide to focus on the clustering step, and the HDBSCAN model in particular, while keeping the other parameters constant. This was decided since the clustering method directly affects the number of topics generated, and by keeping the sentence embeddings common to all three models, the embeddings do not have to be re-calculated for each model, saving processing time. Future work can experiment with allowing the user to select the parameter they intend to configure.

We vary the `min_topic_size` parameter in the BERTopic model, which is the equivalent of `min_cluster_size` in HDBSCAN. This parameter defines the minimum size criterion for a cluster to be considered a distinct sense. Specifically, it determines the minimum number of documents required to form a separate cluster, thereby influencing the granularity and fidelity of sense identification. For the purpose of creating distinct models, we configure the parameter of `min_topic_size` to 15 for model1, 25 for model2, and 50 for model3. While it's important to note that these specific values are selected without a predefined theoretical basis, they are chosen based on empirical observations of the resulting topics. These configurations have been determined to produce the most discernible and diverse clusters among the chosen words, allowing the user sufficient variance in the different models as a starting point to explore sense distinctions within the visualization tool.

An additional measure implemented during the model creation process is the imposition of an upper limit on the number of topics generated. Specifically, the number of topics is capped at 15, in line with prior research on automatic word sense induction using topic modeling that requires the number of senses to be set beforehand [44], [45]. This value of 15 is an arbitrary constraint to prevent the creation of too many fine-grained topics with a low overall topic size, and can be substituted or removed in future work depending on the task at hand.

Creating the list of words:

For the purpose of the prototype, which involves detecting semantic shifts, we wanted to be able to inspect words that have known to have exhibited a shift. A commonly used dataset for semantic shift analysis is the GEMS dataset [46], which consists of 100 English words labelled by five annotators according to the level of semantic change between the 1960s and 1990s, which broadly matches the time period our dataset encompasses. The words in the list were rated by 5 human annotators, who were asked to rank the words according to their intuitions about change in last 40 years on a 4-point scale (0: no change; 1: almost no change; 2: somewhat change; 3: changed significantly). In general, this dataset is used to validate the performance of a model by classifying whether or not a semantic shift has occurred. Since in this list, there exist some words for which no shift occurs, we filter the list to only select the ones where a shift does occur. For this, we take the average of the 5 annotation scores, and select those words greater than 2. The resulting list, however, contained only 8 words. We decided to then reduce the threshold to an average score greater than 1 to increase the number of words to pre-compute. This resulted in a list of 26 words occurring in at least 1000 documents in our dataset.

As an additional word selection method, we conducted a straightforward semantic shift analysis on our dataset based on the methodology proposed by [13]. To accomplish this, we initially extracted documents from the earliest decade (1950) and the most recent decade (2010) and organized them into two separate corpuses. To determine words that displayed semantic shifts, we compared the nearest neighbors of each word at both time periods. For this, we generated word embeddings from the tokenized dataset using a word2vec model [11] for each of the decades. The model's parameters and filtering criteria follow those specified in [13], with vector dimensions set at 300, a window size of 4, and a minimum word occurrence threshold of 20. The remaining hyperparameters remained at their default values. We identified neighbors of a word with a raw frequency exceeding 100 and selected 1000 such nearest neighbors (k = 1000) for further examination through intersection. Furthermore, to ensure good results are generated via topic modeling, we need to ensure a sufficient number of documents are passed to the model. For this, we include only those words that appear in at least 1000 documents across our whole dataset. This threshold of 1000 documents is arbitrary but aligns with general consensus for documents required for coherent topic modeling results. This method gave us a final list of 177 words, which were added to the initial set of 26 from the GEMS dataset for a total of 203 words that were pre-computed.

This selection of words for precomputation in our prototype represents our dataset and the specific selection criteria applied for the prototype, but it does not imply that only these words have experienced semantic shifts. The broader lexicon includes many more examples, and our selection serves the goals and scope of our research and prototype.

4.2 Topics Over Time

To track how the word senses or topics change over time, we make use of DTM, the architecture of which is shown in Figure 4.2. The process works by calculating the topic representation at each timestamp after obtaining the global topic representations for all the timestamps. For each topic and each timestamp, the c-TF-IDF representations are calculated again to find the representation of the topic at that particular timestamp.

For this step, we take the output from the topic modeling step, which consists of dividing the input data into its subsequent timestamps, and using the topics created, calculating new c-TF-IDF representations. The output of this step gives us the following features, the topic index, topic label, timestamp, frequency of the topic at the timestamp, and the top-5 words representing the topic at the timestamp. This is

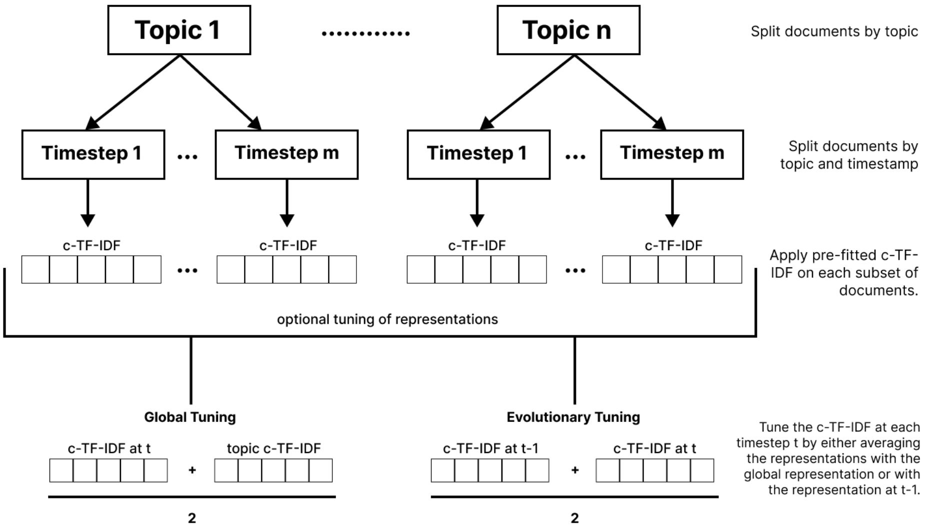

Topic 1
....
Topic n
Split documents by topic
Timestep 1
...
Timestep m
Timestep 1
...
Timestep m
Split documents by topic and timestamp
c-TF-IDF
c-TF-IDF
c-TF-IDF
c-TF-IDF
Apply pre-fitted c-TF-IDF on each subset of documents.
optional tuning of representations
Global Tuning
c-TF-IDF at t
topic c-TF-IDF
+
2
Evolutionary Tuning
c-TF-IDF at t-1
c-TF-IDF at t
+
2
Tune the c-TF-IDF at each timestep t by either averaging the representations with the global representation or with the representation at t-1.

Figure 4.2: Dynamic Topic Modeling (DTM) procedure from BERTopic [2] documentation

calculated for each model and search term.

We additionally fine-tune the dynamic topic model in two ways, setting the `global_tuning` and `evolution_tuning` parameters to `True`. These parameters influence the the topic representation or labels at each timestamp. Fine-tuning the topic representation globally averages its c-TF-IDF representation with that of the global representation. This allows each topic representation to move slightly towards the global representation whilst still keeping some of its specific words. Evolutionary fine-tuning averaging the c-TF-IDF representation with that of the c-TF-IDF representation at timestep 't-1'. This is done for each topic representation allowing for the representations to evolve over time. We set both to `True` to allow the topic at timestep 't' to retain its relevance to the global topic and also to the timestep at 't-1'.

4.3 Runtime Analysis

The model creation was performed on a high-performance computing cluster with 2 x Intel Xeon Silver 4109T CPUs (total 32 cores) CPU, 376 GB DDR4 RAM, and NVIDIA Tesla P100-PCIE (16 GB VRAM).

We will analyze the runtime for the model creation (topic modeling) and topics over time stages separately. For the 203 words computed, the total time taken for creating the models was 25096 minutes or 6.97 hours, and 17462 minutes or 4.85

hours for calculating the topics over time for each of the models. Below is the distribution of these times and the contributing factors towards it.

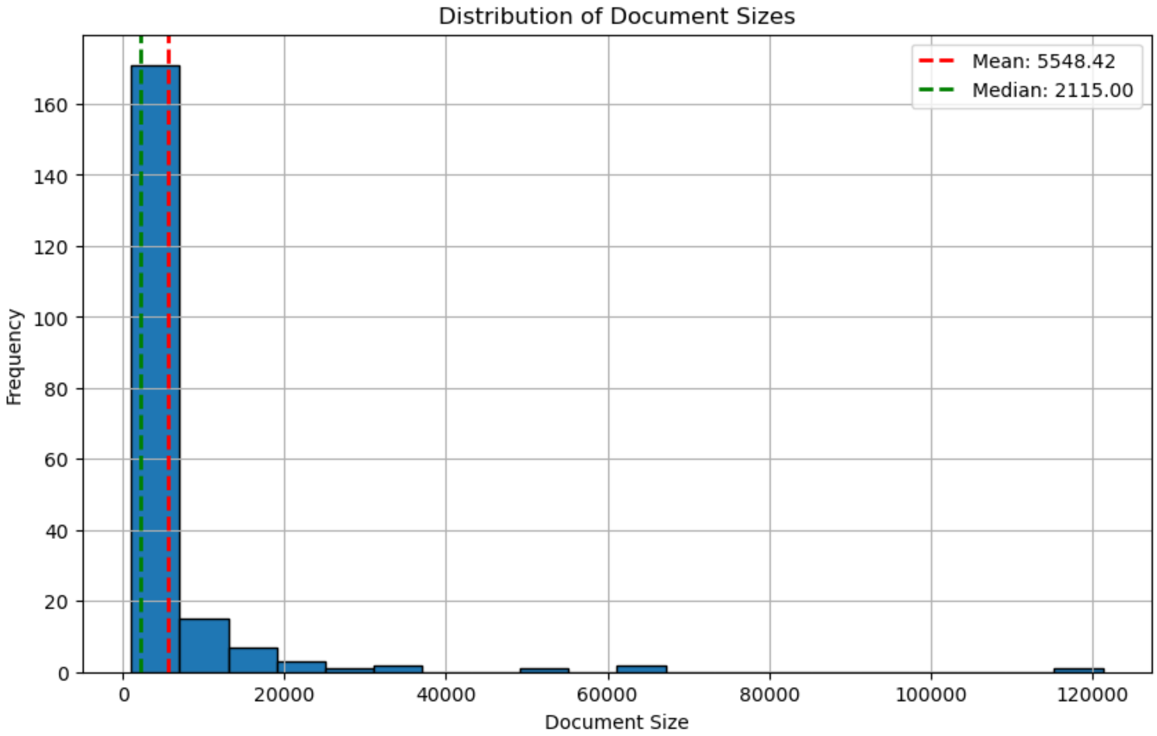
The time taken to generate the models for each word is directly correlated with the number of documents belonging to the target word (Pearson's correlation value of 0.974). Figure 4.3, (a) shows the distribution of the number of documents in the dataset for each word, considering the minimum threshold of 1000. The mean document size for the list of words is 5548.42, while the median size is 2115. The minimum number of documents for a word is 1003 for the word "notch", and the maximum is 121321 for the word "program". Figure 4.3, (b) shows the distribution of time taken in seconds for creating all three topic models for each word. The mean time taken is 123.63 seconds, while the median time is 69 seconds. The minimum amount of time taken is 22 seconds for the word "dupont", and the maximum time taken is 1857 seconds for the word "program". Figure 4.4 shows a linear correlation between required runtime for a word and the number of documents.

To gain an understanding into the steps contributing towards the model creation time, Figure 4.5 shows the distribution of time taken for each step within the topic modeling stage. The majority of time in the topic modeling step is spent in generating the sentence embeddings. For the dimensionality reduction step, we are able to keep the processing time low by rescaling the embeddings using PCA before reducing its dimensionality; while the clustering step is almost negligible in terms of time taken. The topics over time calculation (Figure 4.5) is another time-significant process, and is also linearly correlated with the document size (0.956 Pearson correlation score).

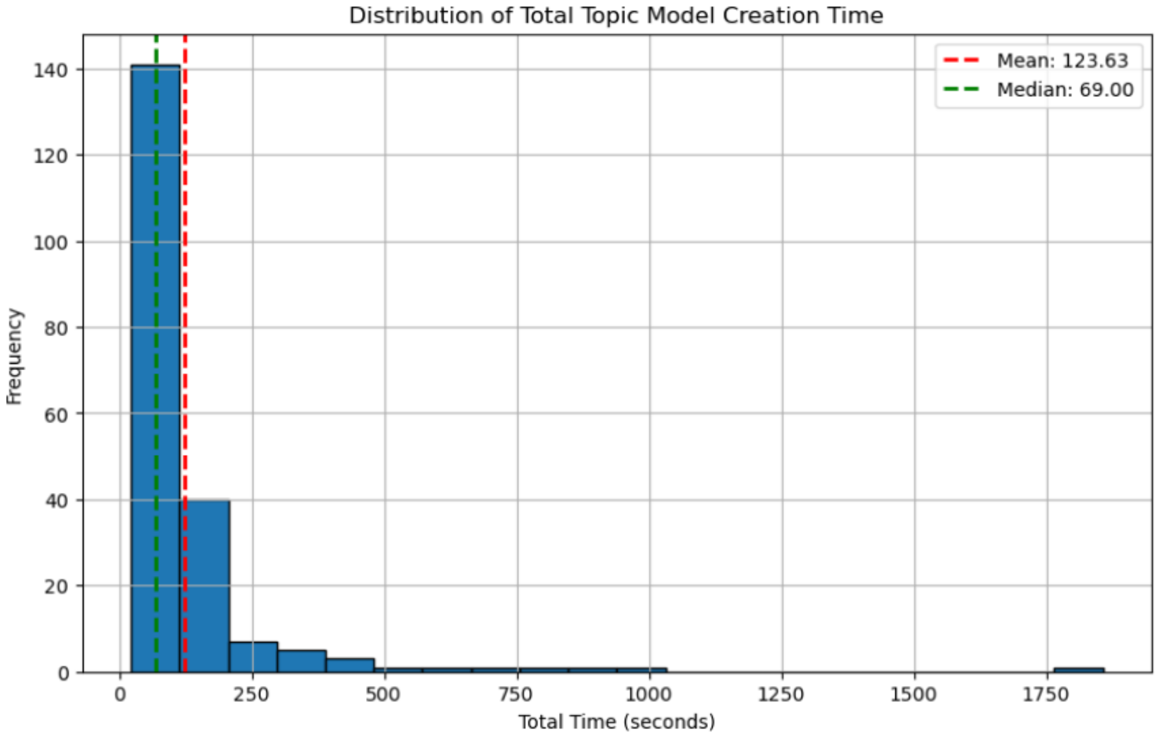
In conclusion, the runtime for a word is highly dependent on the number of underlying documents the word contains, with the runtime linearly increasing for that number. Within the topic modeling procedure, the embedding step is the most time consuming, and by keeping the embedding constant for the 3 models and only vary the clustering step, we are able to significantly reduce the runtime for the precomputation. The two main bottlenecks for the computation in terms of runtime is the embeddings generation step and the topics over time calculation. Future work can address this in two ways in order to facilitate real-time computation and analysis of words. Firstly, by experimenting with increasing the speed of the embedding and topics over time step, and secondly, by keeping document sizes under 20,000, the median processing time for each word can be kept low (as seen in Figure 4.5).

4.4 Summary

At the end of the topic modeling step, we have 3 different models created for each search term. These are stored as individual BERTopic objects. For each search term



(a) Distribution of number of documents for each word in the dataset.



(b) Distribution of time taken (in seconds) for model creation for each word in the dataset.

Figure 4.3: Distributions of number of documents and model creation times for three models

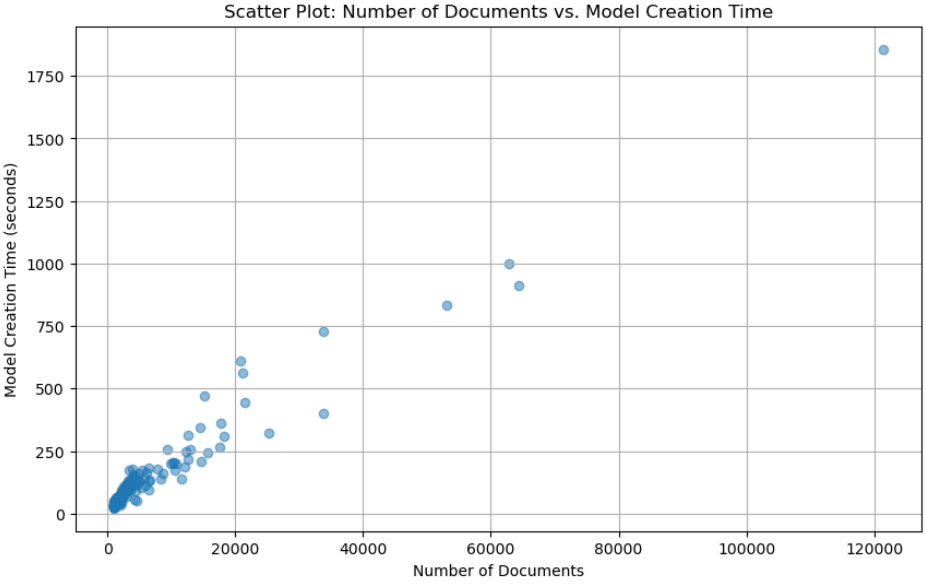


Figure 4.4: Correlation between number of documents and runtime

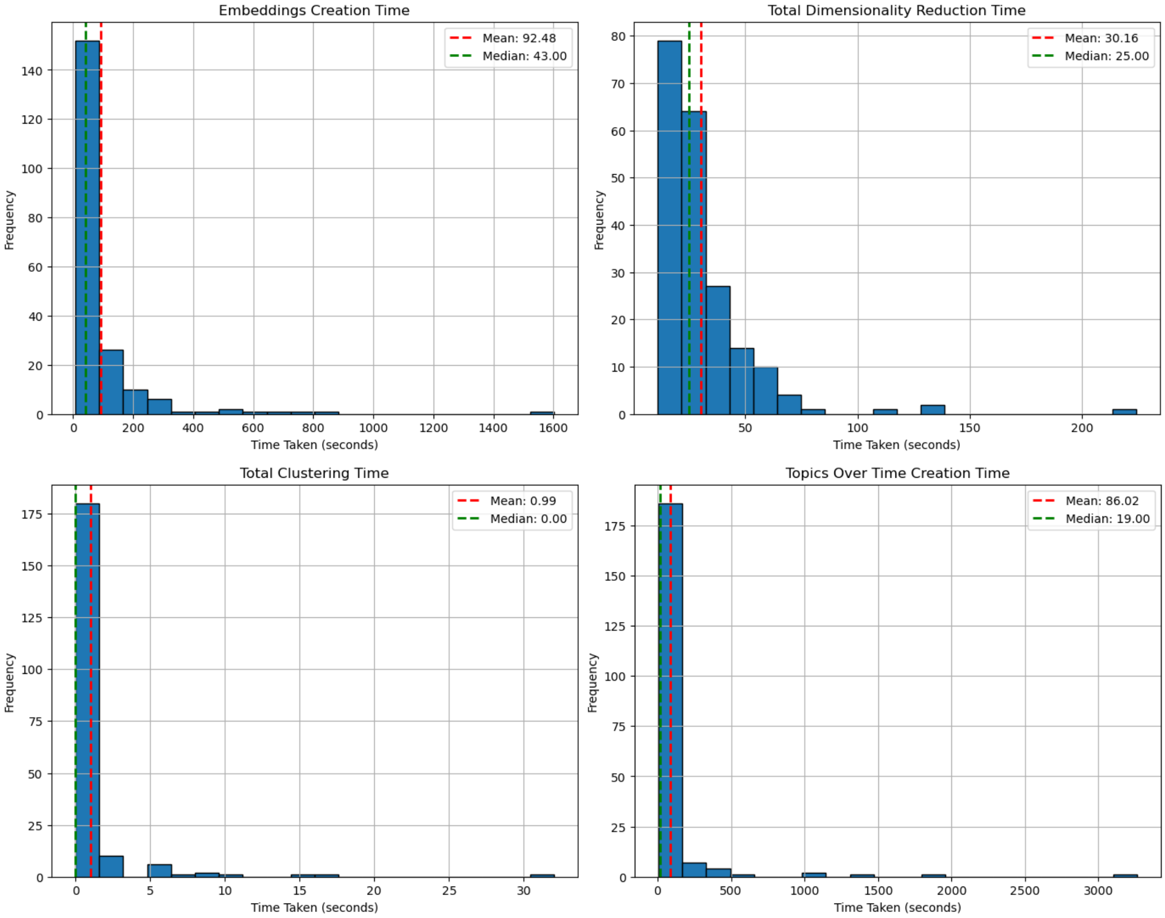


Figure 4.5: Histogram showing the distribution of runtime for different stages of processing

and corresponding model, we make use of the following features in the subsequent data processing: number of clusters, topic embedding, c-TF-IDF representations, cluster size, top-30 terms for each topic, clustered documents and timestamps.

The output of the Dynamic Topic Modeling step is similar to the topic modeling step, but with a different representation for each topic at each timestamp, and the number of documents at each timestamp based on its top terms at that timestamp.

This data generated from the topic modeling and dynamic topic modeling steps is then passed on to process separately for each visualization.

Chapter 5

Design and Implementation of the Visualization Tool

5.1 Introduction

In this chapter, we delve into the development and features of the visualization tool used in this study. We'll explore the libraries employed, the design decisions made, and the purpose behind these choices. The tool is designed to cater to two distinct user groups: firstly, those engaged in developing tools for detecting semantic shifts, seeking insights into the impact of various parameters on the outcomes; secondly, everyday users intrigued by the exploration of semantic shifts and the evolution of word meanings over time. This chapter is intended to provide a comprehensive understanding of how these visualizations contribute to a deeper comprehension of semantic shifts and word sense evolution.

The organization of this chapter mirrors the tool's structure, guiding us through each section as if we were actively engaging with the tool as users would. The first section is about the Introduction of the tool, and the second section lets the user select the word that they want to analyze. The third and fourth section are the ones that contain the visualizations. In these sections we will explain the following points for each chart: rationale behind the visualization selection, a screenshot of the diagram in the tool, the data preparation procedure, and the features that they measure.

To build interactive visualizations for our web-based tool, we opted for JavaScript libraries, as they are well-suited for web development. D3.js[1] [47] was our primary choice due to its flexibility and customization options, making it ideal for crafting tailored visualizations. We used version 4 of d3.js, and also made use of the 2.24.1

[1] https://d3js.org/

version of Plotly.js[2] and amCharts 5[3] libraries, for their out-of-the-box interaction functionalities. Our choice of visualization libraries, including d3.js, Plotly.js, and amCharts 5, is underpinned by their unique capabilities and functionalities. D3.js offers extensive flexibility, support with community-built libraries, and customization for crafting bespoke visualizations, while Plotly.js and AmCharts5 enrich the tool with out-of-the-box interaction functionalities, simplifying user engagement. While there are alternative technologies to create these visualizations, we opted for a web-based approach and selected only technologies that contain JavaScript functionalities in order to customize our system. For future customization, each individual chart can be created using another JavaScript library of choice provided it is customized to support the underlying data as explained in the following sections.

In all the visualizations, we dynamically assign a unique colour to differentiate topics, following the commonly used `d3.schemeCategory20` palette. This choice is grounded in principles of color selection, including considerations for colorblind accessibility and visual distinctiveness. The palette assigns a unique color to each subsequent topic per model per search term, organized by topic size, ensuring that the visualizations remain informative and user-friendly.

Design Guideline

For the design and implementation of the visualization tool, we adhere to the Information-Seeking Mantra [48], a framework that underscores three crucial aspects of information visualization: Overview First, Zoom and Filter, and Details on Demand. We use this as a guideline for all the visualizations proposed in the tool, in terms of presenting the information and guiding the user for further exploration.

Overview First: We begin by providing users with a high-level overview of the data, ensuring that the initial view offers a broad perspective of the data being visualized. For instance, the visualizations are designed to show an overview of the data, or a high-level trend of different word meanings and whether a shift has taken place. This serves as a starting point for users to familiarize themselves with the tool's scope.

Zoom and Filter: To empower users to explore specific aspects of the data, we incorporate interactive features like zooming, panning, and filtering controls. These functionalities allow users to focus on areas of interest. In our tool, users can select specific words to inspect, zoom into particular time periods, and filter data based on topics, among other interactions. These capabilities align with the ”Zoom and Filter” principle, enabling users to delve deeper into the information they seek.

[2] https://plotly.com/javascript/

[3] https://www.amcharts.com/

Details on Demand: To provide users with additional context or information when needed, we've integrated mechanisms such as tooltips, pop-ups, and detailed data views. For example, users can hover over data elements to access supplementary information in tooltips, or click on data points to reveal comprehensive details in a separate panel. This adheres to the ”Details on Demand” aspect of the mantra, ensuring that users can access more information as required.

Throughout the chapter, we emphasize how each visualization component aligns with the Information-Seeking Mantra framework. We illustrate the rationale behind visualization choices, provide screenshots of the tool's interface, explain the data preparation process, and clarify the features measured by each chart. Additionally, we highlight the interactivity options incorporated into the tool, allowing users to isolate, remove, zoom, select models, and more, thus enhancing their exploration and analysis capabilities.

5.2 Visualization Tool Interface

This section provides an in-depth exploration of the tool's user interface, explaining the different sections and their functionalities.

Introduction

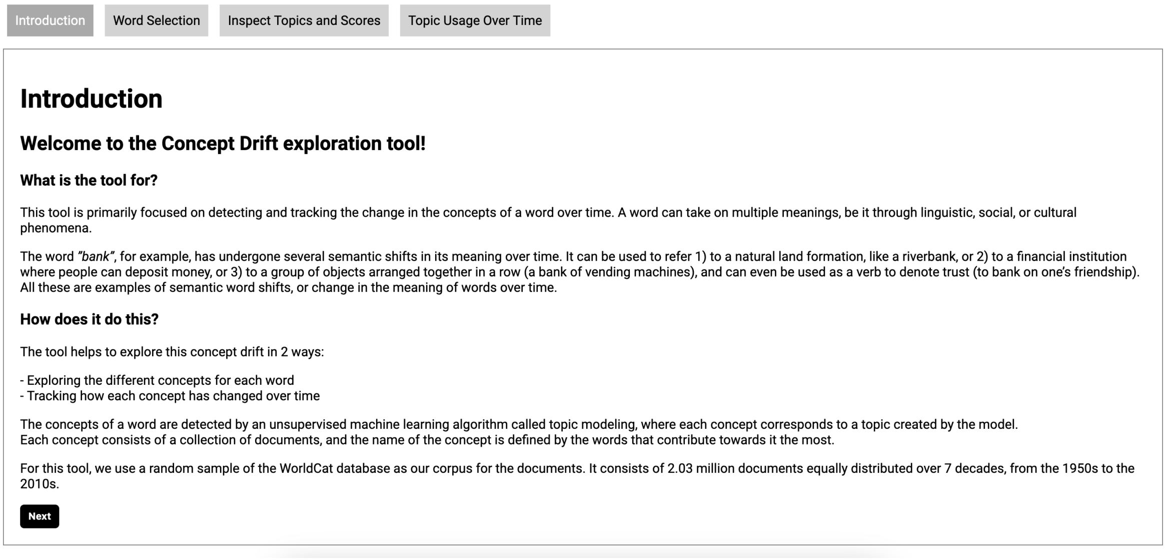
Upon opening the tool, the user is greeted with the Introduction screen explaining the tool. It provides a brief introduction of the purpose of the tool, its measures, and the underlying data. Clicking the “Next” button takes them to the next section for word selection.

Word Selection

Here, the user can select a word to analyse in the next sections. The words to select are from the list of 71 words introduced in section 4.1.5. Searching for the word brings the user to the next section.

5.2.1 Inspecting Word Senses

This section offers a comprehensive overview of the visualizations designed to inspect and understand word senses. It comprises of the following subsections:



Introduction

Welcome to the Concept Drift exploration tool!

What is the tool for?

This tool is primarily focused on detecting and tracking the change in the concepts of a word over time. A word can take on multiple meanings, be it through linguistic, social, or cultural phenomena.

The word *"bank"*, for example, has undergone several semantic shifts in its meaning over time. It can be used to refer 1) to a natural land formation, like a riverbank, or 2) to a financial institution where people can deposit money, or 3) to a group of objects arranged together in a row (a bank of vending machines), and can even be used as a verb to denote trust (to bank on one's friendship). All these are examples of semantic word shifts, or change in the meaning of words over time.

How does it do this?

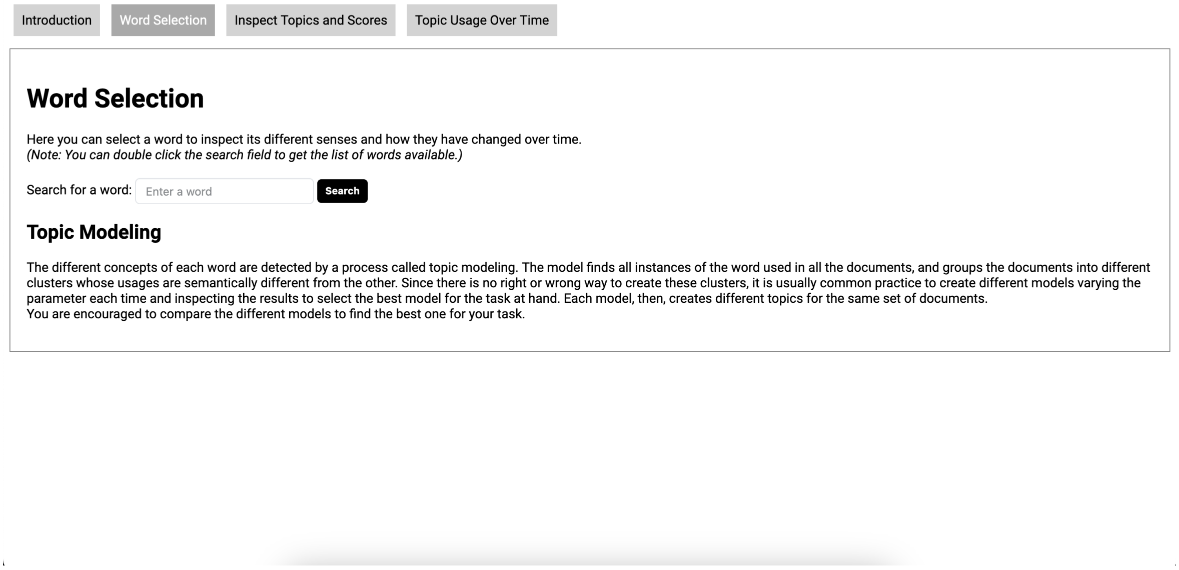
The tool helps to explore this concept drift in 2 ways:

- Exploring the different concepts for each word
- Tracking how each concept has changed over time

The concepts of a word are detected by an unsupervised machine learning algorithm called topic modeling, where each concept corresponds to a topic created by the model.
Each concept consists of a collection of documents, and the name of the concept is defined by the words that contribute towards it the most.

For this tool, we use a random sample of the WorldCat database as our corpus for the documents. It consists of 2.03 million documents equally distributed over 7 decades, from the 1950s to the 2010s.

Figure 5.1: Introduction Page of the interface



Word Selection

Here you can select a word to inspect its different senses and how they have changed over time.
(Note: You can double click the search field to get the list of words available.)

Search for a word: Enter a word

Topic Modeling

The different concepts of each word are detected by a process called topic modeling. The model finds all instances of the word used in all the documents, and groups the documents into different clusters whose usages are semantically different from the other. Since there is no right or wrong way to create these clusters, it is usually common practice to create different models varying the parameter each time and inspecting the results to select the best model for the task at hand. Each model, then, creates different topics for the same set of documents.
You are encouraged to compare the different models to find the best one for your task.

Figure 5.2: Page to select a word to inspect

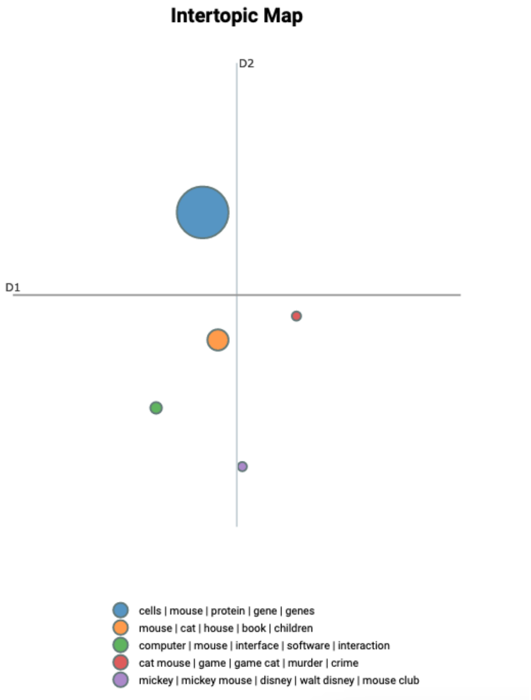


Figure 5.3: Intertopic Map: Topics represented as points with different colours, sized by document count, with proximity indicating similarity.

Intertopic Map

Reasoning: The first chart the user sees in this section is the intertopic map (Figure 5.3). This was chosen because the intertopic map provides a quick overview to inspect topics, their sizes, and their similarity to each other. The topic embeddings are reduced to two dimensions (D1 and D2) so that they can be represented visually. The chart is based on the global topic view of the LDAVis [49] visualization, which has become a common visualization method to inspect topics generated by topic models.

The general idea is to view the topic embedding space in two-dimensions in order to quickly inspect large and similar topics. We use UMAP to reduce the topic embeddings to two dimensions, positioning it in 2-D space, with `x` value taking the value of `D1` and the `y` value taking the value of `D2`. Each topic is represented by a circle whose size is proportional to the number of documents in the topic, with a unique colour assigned to each topic.

By viewing the placement of all the topics, this chart provides first the broad overview by showing number of topics, similarity, and topic size at a glance. Interactivity tools enable zooming and filtering for focusing on a single section. Hovering or clicking on a topic provides the details on demand.

Data Preparation: To plot the data 5 features are required:

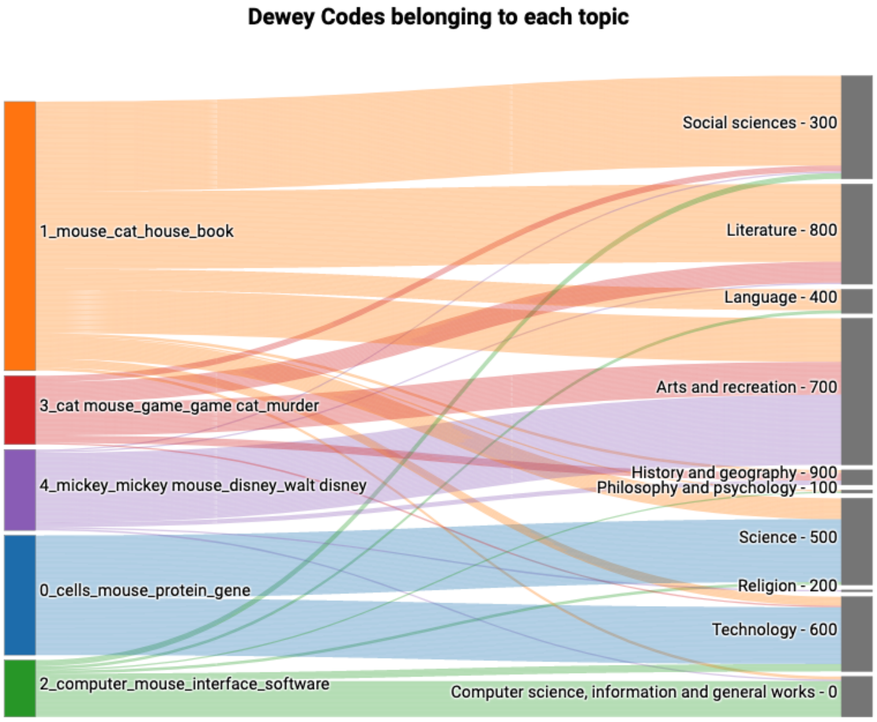


Figure 5.4: Dewey Code Mapping

`x, y, topic, words, size`

`x` and `y` are the co-ordinates defining the position of the topic, where `x` is the value of the first dimension, and `y` is the value of the second. `topic` is the index of the topic generated by the model, `words` is a string of the top-5 words defining the topic, which is shown as a label on hover, and `size` is the number of documents belonging to the topic, shown as the size of the circle denoting each topic. Here, we set the `sizemode` to “area”, telling the plot to scale the area as a whole based on the `size` parameter, with a scaling factor of 3.

Dewey Code Mapping

Reasoning: Since the underlying data is of a library classification system, we wanted to show the Dewey codes that the documents belong to as a form of topic categorization. To map word senses to Dewey codes visually, we make use of a Sankey plot (Figure 5.4). This is because Sankey plots are widely recognized to be an optimal way of visualizing many-to-many mappings between two domains.

By using Dewey codes, we aim to see the distribution of topics to human-annotated categories, which can also be useful for users to gauge the category a topic belongs to, if it was not clear through intuition. Additionally, it also provides a means to validate the coherence of a topic, as a topic belonging majorly to a specific category might provide more information that one equally distributed across multiple cate-

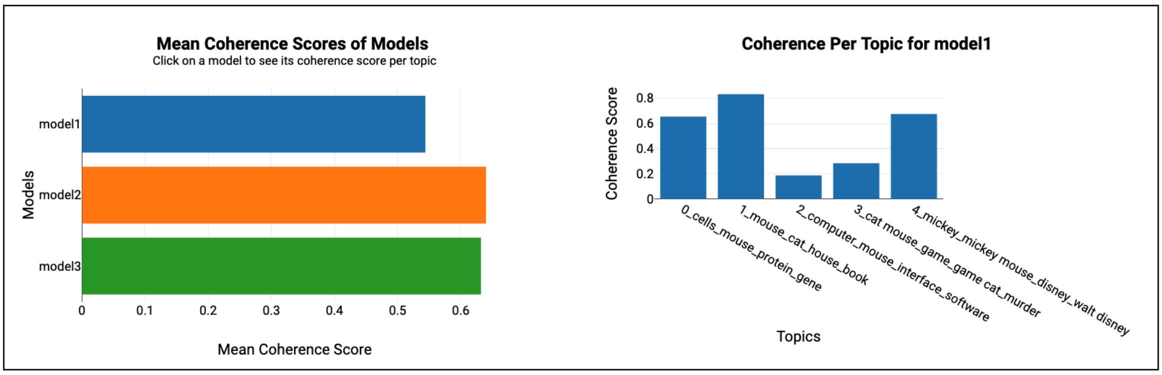


Figure 5.5: Left: Mean coherence scores for all models; Right: Coherence score per topic for each model which is accessed by clicking on the model

gories. It is important to note here that not all documents in the database have a Dewey code and only those documents containing a Dewey code are visualized here.

Data Preparation: We follow the Plotly[4] approach to creating a Sankey chart. The data is arranged in the form of a "source" (Topic) and "target" (Dewey Code) for each document, colour-coded by the topic the document belongs to, and "label" which assigns the document to the link that can be read on hover. We set the chart orientation to horizontal, and the auto-arrangement of the nodes is set to "snap", where the node arrangement is assisted by automatic snapping of elements to preserve space and minimize link overlap.

Coherence Scores

Reasoning: Since topic modeling forms the crux of the topic generation process, it is also important to have a common metric used for assessing models in addition to evaluation from solely a semantic lens. We, thus, calculate coherence scores for each of the model. Coherence scores help to assess the quality and interpretability of generated topics, and provide a quantitative measure of how well topics represent coherent and distinct concepts within a corpus.

Coherence Scores are a common metric used to measure the performance of a topic model and for selecting the best performing model. They measure the score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference.

Variants and Calculation: To calculate coherence scores, we employ four widely recognized variants: `u_mass`, `c_v`, `c_uci`, `c_npmi`. These variants are based on

[4] https://plotly.com/javascript/sankey-diagram/

the state-of-the-art work by Röder et al. [4]. Here's a brief overview of each variant:

- u_mass Coherence: This metric evaluates topic coherence by considering the co-occurrence of words within the same documents. It measures the degree of semantic similarity between high-scoring words in a topic based on their document-level co-occurrence.

- c_v Coherence: The c_v variant assesses topic coherence by comparing the co-occurrence statistics of top words within topics against a background corpus. It quantifies how well the words in a topic form meaningful and coherent concepts.

- c_uci Coherence: c_uci, or ”contextual unnormalized coherence,” calculates coherence by examining the co-occurrence patterns of top words within topics while considering their distribution across the entire corpus. It aims to capture the meaningful relationships between words in a topic.

- c_npmi Coherence: The c_npmi coherence score calculates the normalized pointwise mutual information between pairs of top words in a topic. It provides insights into how words within a topic are semantically related.

Insights from research: According to recent research by different variants have distinct behavior, sensitivity to intruder words, and varying levels of performance. For example:

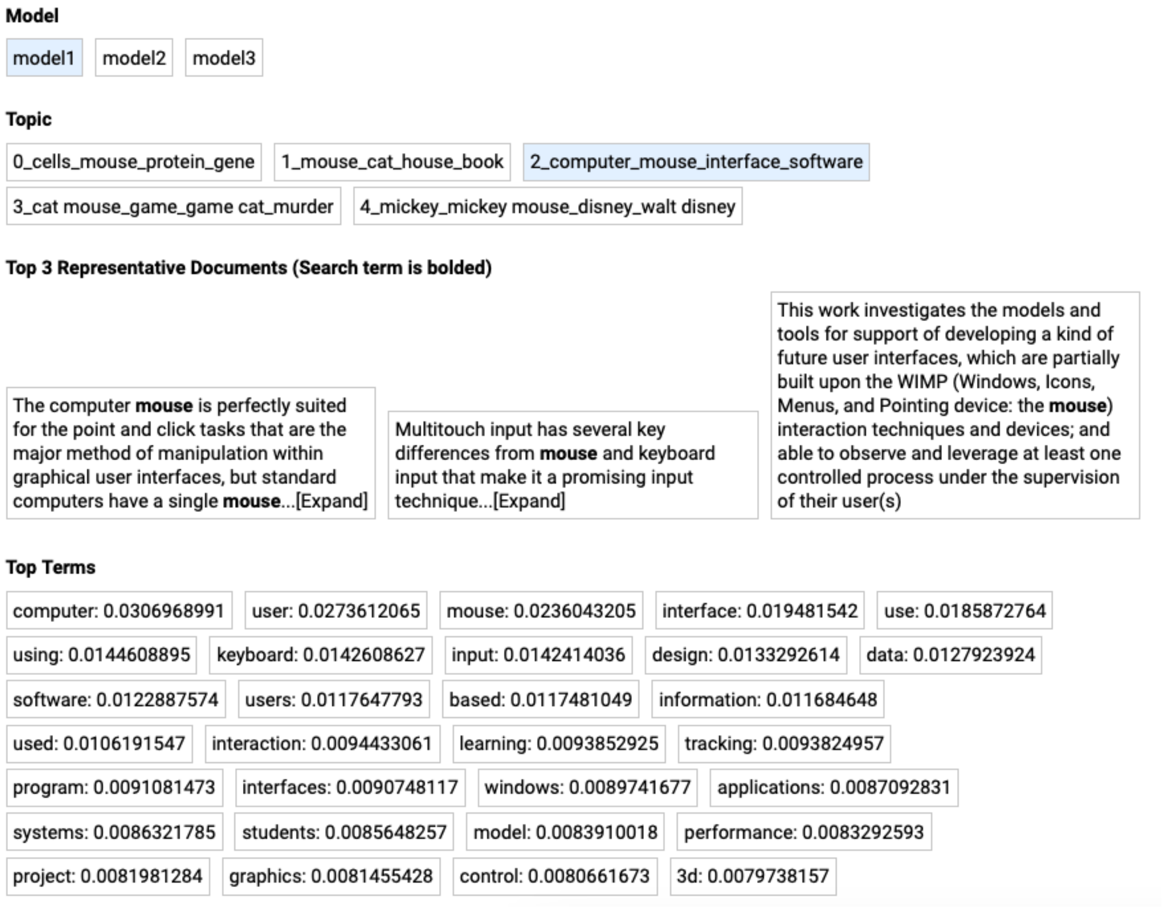
u_mass and c_uci were considered the best choices since they disregard small noises from affecting the scores. c_v is sensitive and affected by the number of intruder words, making it suitable for identifying noise in topics. c_npmi behave differently on 5-word and 10-word topics, behaving similarly to u_mass and c_uci for 5-word topics, and inverting its behaviour for 10-word topics, thereby offering choices based on the developer's preferences and objectives.

In order to display these measures, we make use of a bar chart (Figure 5.5) to show the computed scores per model and per topic.

For all the variants, the higher the score, the more coherent the topic is. In practice, however, coherence scores do not always represent the right topic or model for the task at hand. The interpretation of what constitutes a “good” coherence score is somewhat subjective, and there is no universally agreed-upon threshold for what qualifies as a coherent topic. Nevertheless, these scores provide a good baseline to evaluate the models against human intuition.

Data Preparation: We use the available variants of coherence scores from the `gensim`[5] implementation of topic coherence pipeline in [4], and calculate coherence

[5] https://radimrehurek.com/gensim/models/coherencemodel.html


Model
model1
model2
model3
Topic
0_cells_mouse_protein_gene
1_mouse_cat_house_book
2_computer_mouse_interface_software
3_cat mouse_game_game cat_murder
4_mickey_mickey mouse_disney_walt disney

Top 3 Representative Documents (Search term is bolded)

The computer **mouse** is perfectly suited for the point and click tasks that are the major method of manipulation within graphical user interfaces, but standard computers have a single **mouse**...[Expand]

Multitouch input has several key differences from **mouse** and keyboard input that make it a promising input technique...[Expand]

This work investigates the models and tools for support of developing a kind of future user interfaces, which are partially built upon the WIMP (Windows, Icons, Menus, and Pointing device: the **mouse**) interaction techniques and devices; and able to observe and leverage at least one controlled process under the supervision of their user(s)

Top Terms

computer: 0.0306968991
user: 0.0273612065
mouse: 0.0236043205
interface: 0.019481542
use: 0.0185872764
using: 0.0144608895
keyboard: 0.0142608627
input: 0.0142414036
design: 0.0133292614
data: 0.0127923924
software: 0.0122887574
users: 0.0117647793
based: 0.0117481049
information: 0.011684648
used: 0.0106191547
interaction: 0.0094433061
learning: 0.0093852925
tracking: 0.0093824957
program: 0.0091081473
interfaces: 0.0090748117
windows: 0.0089741677
applications: 0.0087092831
systems: 0.0086321785
students: 0.0085648257
model: 0.0083910018
performance: 0.0083292593
project: 0.0081981284
graphics: 0.0081455428
control: 0.0080661673
3d: 0.0079738157

Figure 5.6: Viewing the top 3 documents and top 30 terms per topic per model

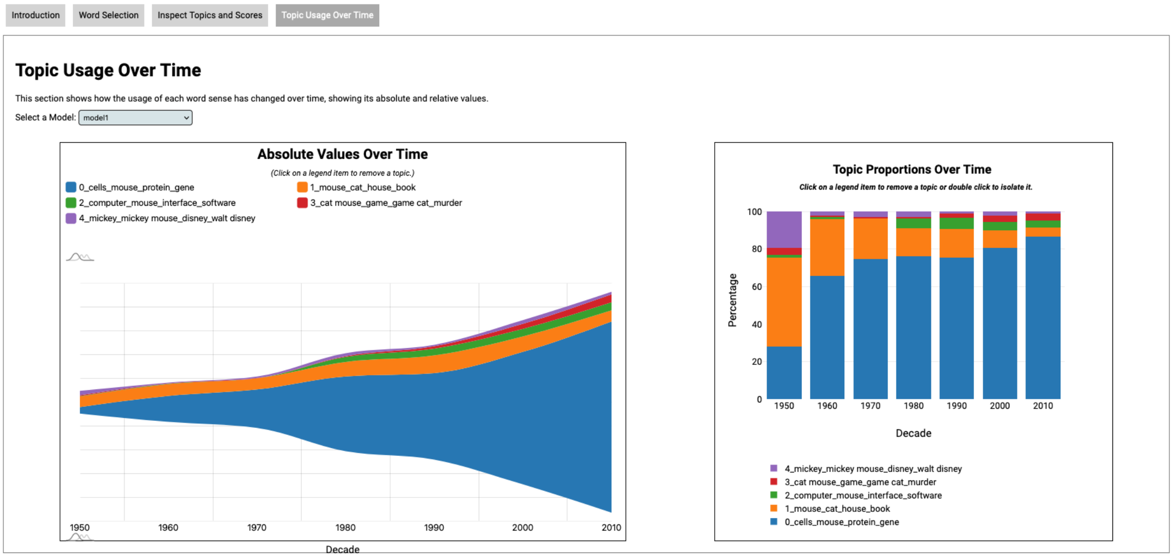
values for the following variants: `u_mass`, `c_v`, `c_uci`, `c_npmi`. These scores are calculated per topic for each model in each search term. A mean coherence score for each model is also calculated for a broader overall value.

The processed data is stored in two files: `mean_coherence_scores.csv` which stores the overall mean coherence scores of the model, and `coherence_per_topic.csv` which stores the individual coherence scores of each topic for each model.

Documents and Terms per Topic

Reasoning: This section is a non-visual way of showing information, designed as an option to allow users to read through the terms and documents that make up a topic. By choosing a model and a topic belonging to the model, the user is able to see the top documents and terms corresponding to that topic. Figure 5.6 shows a screenshot of this section.

Data Preparation: The top documents show the top 3 most representative documents belonging to the topic, which is calculated by comparing the c-TF-IDF representation of all the documents belonging to a topic, comparing it to the c-TF-IDF



Introduction
Word Selection
Inspect Topics and Scores
Topic Usage Over Time

Topic Usage Over Time

This section shows how the usage of each word sense has changed over time, showing its absolute and relative values.

Select a Model: model1

Absolute Values Over Time
(Click on a legend item to remove a topic.)
0_cells_mouse_protein_gene
1_mouse_cat_house_book
2_computer_mouse_interface_software
3_cat mouse_game_game cat_murder
4_mickey_mickey mouse_disney_walt disney
1950
1960
1970
1980
1990
2000
2010
Decade
Topic Proportions Over Time
Click on a legend item to remove a topic or double click to isolate it.
Percentage
100
80
60
40
20
0
1950
1960
1970
1980
1990
2000
2010
Decade
4_mickey_mickey mouse_disney_walt disney
3_cat mouse_game_game cat_murder
2_computer_mouse_interface_software
1_mouse_cat_house_book
0_cells_mouse_protein_gene

Figure 5.7: Section to visualize the topic change over time

representation of the topic, and selecting the 3 most similar documents from it. The top terms shows the top 30 terms that most contribute towards the topic, which is calculated via a topic-term matrix generated using the c-TF-IDF procedure.

5.2.2 Topic Usage Over Time

In this section, we visualize the evolution of word sense usage over time. As introduced in the earlier section on related work, to show the usage change of word senses over time, we employ two commonly used visualizations to aid in comprehending these changes: the steamgraph and the stacked bar chart, to encode continuity. A screenshot of this section is show in Figure 5.7.

Absolute Values Over Time

Reasoning: The steamgraph is a commonly used visualization in semantic shift visualizations as seen in the Related Word section. This prompted the usage due to the familiarity of its usage and the features it helps to visualize, namely: frequency, continuity, and separation of word senses.

The steamgraph in Figure 5.8 illustrates the absolute changes in word sense usage over time. By visualizing the rise and fall of various sense categories it offers insight into how the different senses have gained prominence or faded in significance over the timeline of analysis. A common issue with visualizing steamgraphs is the illegibility or difficulty in interpreting the chart when there are too many streams present in the diagram. The multiple colours or presence of too wide a stream can clash with or overshadow other senses. By enabling interactivity within the tool we

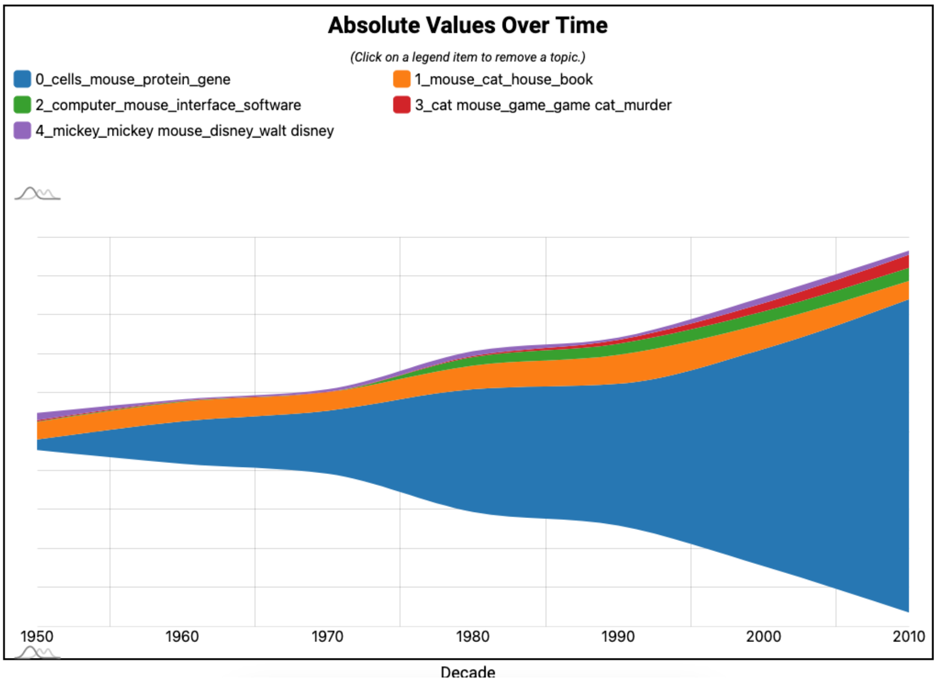


Figure 5.8: Steamgraph to show the absolute values over time

hope to address this difficulty by allowing users to isolate or hide certain streams, or zoom into to focus on a cluster of streams or a specific time period. Additionally, hovering over the streams reveals tooltip information (Figure 5.9. These interactivity controls can help address the problem of illegibility without compromising on cluttering the visual space.

Documents and Terms per Topic

Data Preparation: The steamgraph data is prepared by processing the output data generated by the topics over time procedure(Section 4.2). We use the following features from the topics over time data to create the steamgraph.csv file: “Topic” for the topic index, “Name” for the topic label, “Words” for the top 5 words belonging to the topic, “Frequency” for the width of the stream, “Decade” for the timestamp of the values.

Topic Proportions Over Time

Reasoning: The Stacked Bar Chart (Figure 5.10) is also a commonly used visualization, used to show the change in the most prominent senses over time. It encodes the same features as the steamgraph, but instead of showcasing the absolute value, it shows the value of each sense as a proportion out of hundred. This offers a different perspective by keeping the visual space the same and helps to observe more

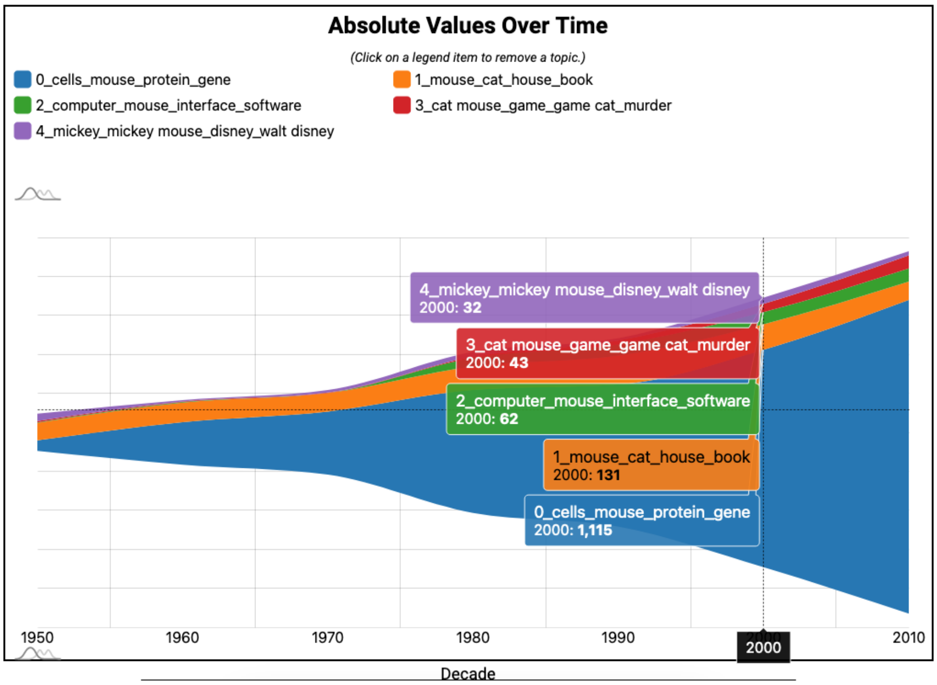


Figure 5.9: Tooltip information visibility when hovering over the steamgraph in the ”2000s” decade. The tooltip shows topic name, decade, and number of documents for all topics.

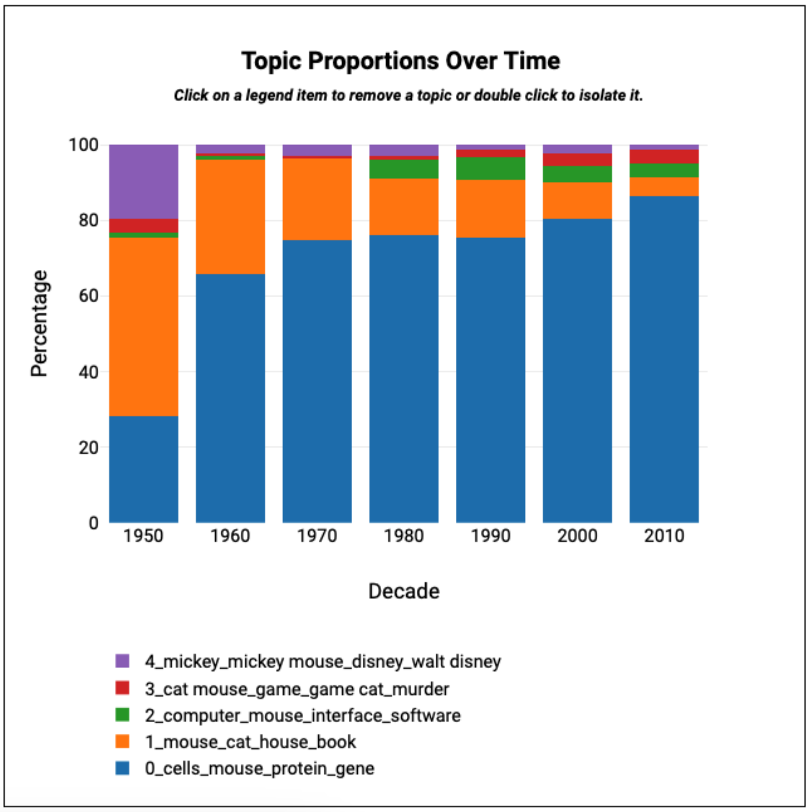


Figure 5.10: Stacked bar chart to show the proportion of topics over time

Table 5.1: The different types of charts used in the tool, the section they belong to, and the features they measure.

Section	Chart	Features Measured
Inspecting Word Senses	Intertopic Map	Number of topics, topic size, topic similarity
	Sankey Chart	Dewey Code mapping, categorisation of topics
	Bar Chart	Coherence Scores per topic per model
	Textual Information	Documents and Terms per topic per model
Topic Usage Over Time	Steamgraph	Absolute Values over time, word sense usage
	Stacked Bar Chart	Relative Values over time, word sense usage

closely the broadening or narrowing of a sense.

Data Preparation: It uses the same dataset as that of the steamgraph, namely steamgraph.csv, but encodes the value of frequency as a percentage of the total frequency for a particular sense of a word and model in a specific time period.

Table 5.1 shows an overview of the different types of charts used in the tool, the sections they belong to, and the features that they measure.

5.3 Visualization Interactivity

One of the major benefits of having an interactive visualization system as opposed to a static one is the ability for the user to interact with the visuals, and thus, the chance to add additional information or enhance usability through these interactions. Below are the different interactivity measures we incorporate into the tool. We use the built-in Plotly interactivity measures in the charts for most commonly used interactions.

- Isolate: Isolating specific word senses or topics of interest helps to inspect them individually. By double clicking on the legend entries corresponding to these senses, users can focus on visualizing the individual senses while temporarily hiding the others.

- Remove: Single clicking on a legend entry removes the sense's trace from the chart, allowing users to declutter the visualization when exploring specific word senses.

- Zoom: Zooming functionality enables users to inspect and scrutinize specific sections of the visual more closely. By clicking and dragging with the chart, users can select an area of interest to zoom in, revealing finer details of the data. Additionally, the scroll wheel of the mouse also helps to zoom in and out.

- Dropdown: The model dropdown menu allows users to render the visualizations for the respective model. Changing the model from the dropdown automatically updates the streamgraph to reflect the data associated with the newly selected model.

- Hover: Hovering over a segment of the streamgraph provides users with immediate information about the specific word sense represented by that segment. Tooltips display aadditional and all information for a visualization.
- Reset: To quickly return to the default view after any interactions, a 'Reset' button is available. Clicking this button resets the visualization to its initial state, displaying all senses. Double clicking anywhere on the chart resets the visualization.
- Download Plot as PNG: Users can download the plot as a PNG image as a static visual, aiding in sharing the visuals.

These interactive features collectively enhance the usability of our visualization tool, providing users with the flexibility to investigate word sense usage trends, conduct detailed analyses, and customize their visualizations to suit their specific research needs. Whether it's isolating individual senses, zooming in for a closer examination, or downloading visualizations for documentation, these interactions empower users to derive meaningful insights from the data.

5.4 Visually Identifying Known Shifts

To gain an understanding of the tool's ability to visually capture a semantic shift occurring, we provide two examples of identifying known shifts through the visualizations. For showing these examples, we select two words to inspect from the GEMS dataset (introduced in Chapter 4.1.5) that received a high average score of semantic change from the evaluators: "vector" and "net".

"vector": We use the steamgraph from the "Topic Usage Over Time" section to detect a new word sense. Figure 5.11 shows the different senses over time for the word "vector". By looking at the chart we can see three major topics increasing in width over time. Topic "0_gene_dna_vector_virus" has a sharp increase from the 1970 decade, and likely refers to the concept of viral vectors, which are tools for gene therapy and vaccines, and was introduced in 1972 [50]. Topic "0_classification_based_vector_svm" refers to the concept of support vector machines, which were introduced in the 60s [51] and gained popularity in the 90s [52]. Topic "0_shocks_monetary_models_economic" suggests a financial sense of the word. Further inspection of its top terms and documents ("var", "vector", and "autoregressive" comprise the 11th, 12th, and 15th most important terms for this topic respectively)

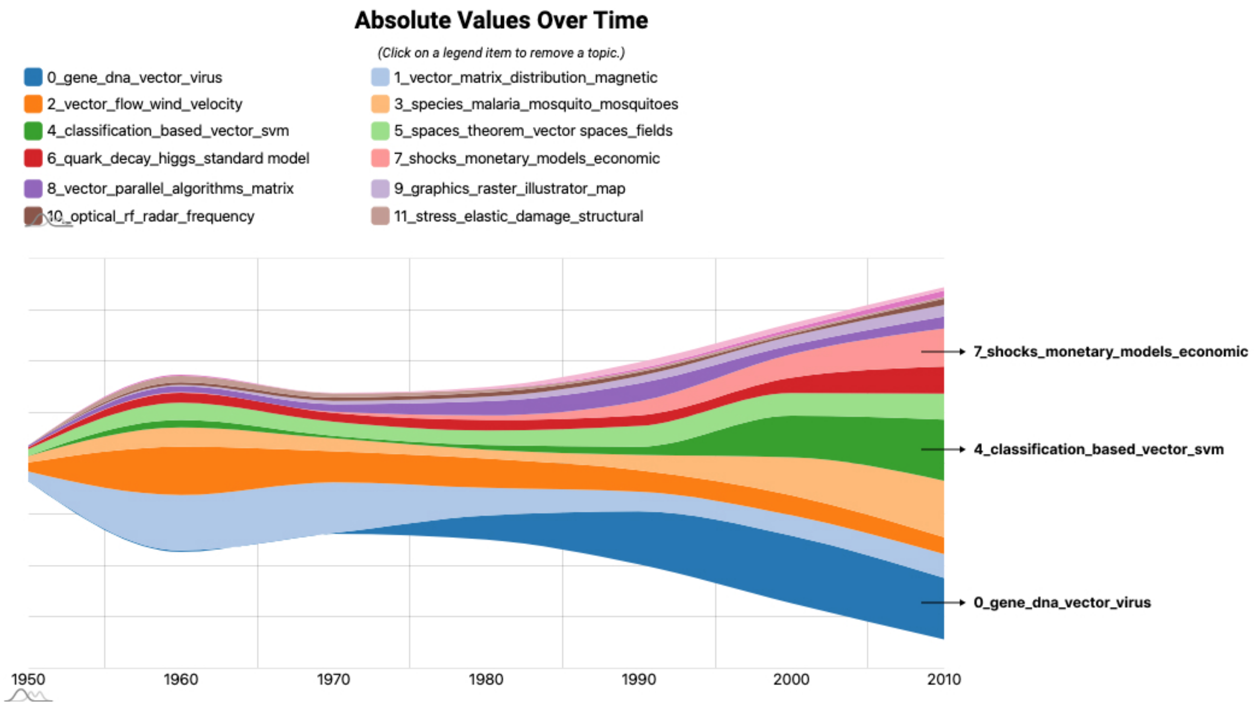


Figure 5.11: Absolute values over time for the word “vector”. Each stream represents a different topic.

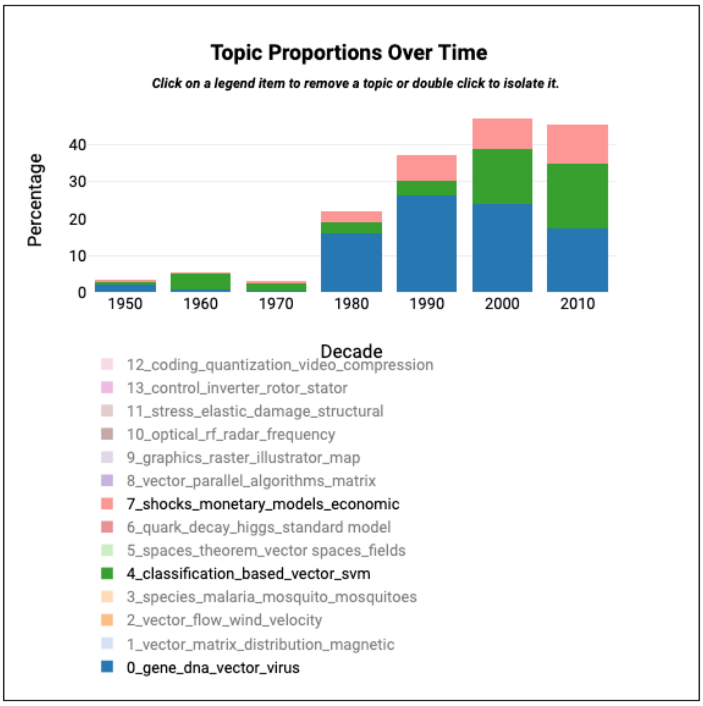


Figure 5.12: Usage proportion over time for the word “vector” for three topics. Remaining topics are hidden to increase readability.

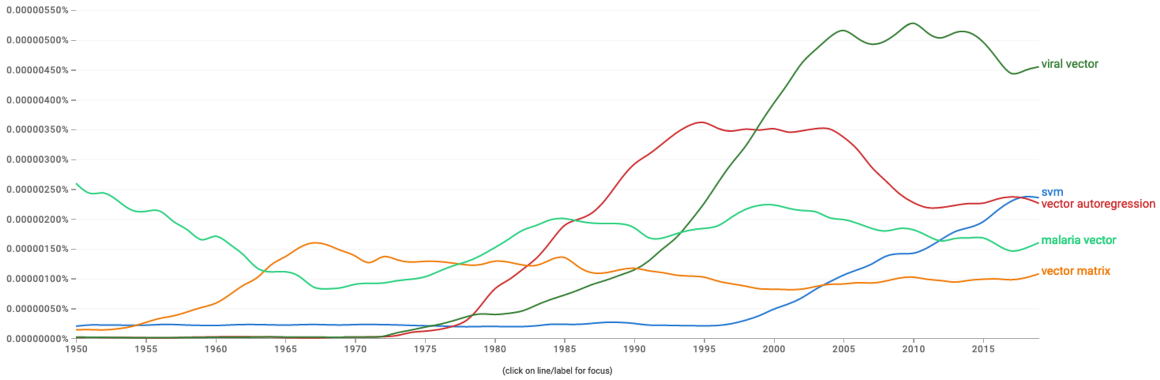


Figure 5.13: Counts of word usage across time from Google N-Gram Viewer for search terms relating to topics of the word "vector"

suggests that the concept is about the econometric framework of vector autoregression (VAR) models, which was introduced in the 1980s [53]. Figure 5.12 shows the stacked bar chart visualization, which shows the proportion increase of all the above topics over time.

While the primary purpose of this example is not to validate the "correctness" of the discovered senses, it is, nevertheless, useful to see how these detected new senses correlate to actual contextual usage of the word "vector". We cannot use the GEMS dataset for this purpose since it only provides information about whether a word has changed its meaning and to what extent, and not about what those different senses are and when they . For this, we can turn to the Google N-Gram Viewer [54] [6] since it is useful for observing counts of n-gram usage across time. It is commonly used for studies in the field of *culturonomics* [54] (the study of cultural and historical phenomena based on large textual data) rather than language change, but nevertheless provides a useful resource for this particular task. Because the N-Gram Viewer does not provide contexts of the word usage, we provide search terms in the form of n-grams for the different topics to observe. We use the terms "svm", "vector autoregression", and "viral vector" to describe the three topics we detected as having undergone shifts ("0_classification_based_vector_svm", "0_shocks_monetary_models_economic", and "0_gene_dna_vector_virus" respectively). We also add the terms "vector matrix" and "malaria vector" for the topics "1_vector_matrix_distribution_magnetic" and "3_species_malaria_mosquito_mosquitoes" respectively in order to include examples of "stable" topics.

Figure 5.13 shows the results of the word usage distribution for these terms. We see that the topics identified as new senses (represented by "svm", "vector autore-

[6] https://books.google.com/ngrams/

gression”, and “viral vector”) all emerge as novel search terms. Their periods of increased usage align closely with the patterns observed in the steamgraph. The “stable” senses (“vector matrix” and “malaria vector”) also show similar usage patterns as the steamgraph. This comparison, although highly specific, showcases the visualization tool in action and demonstrates its proficiency in identifying and correlating relevant information with word usage patterns.

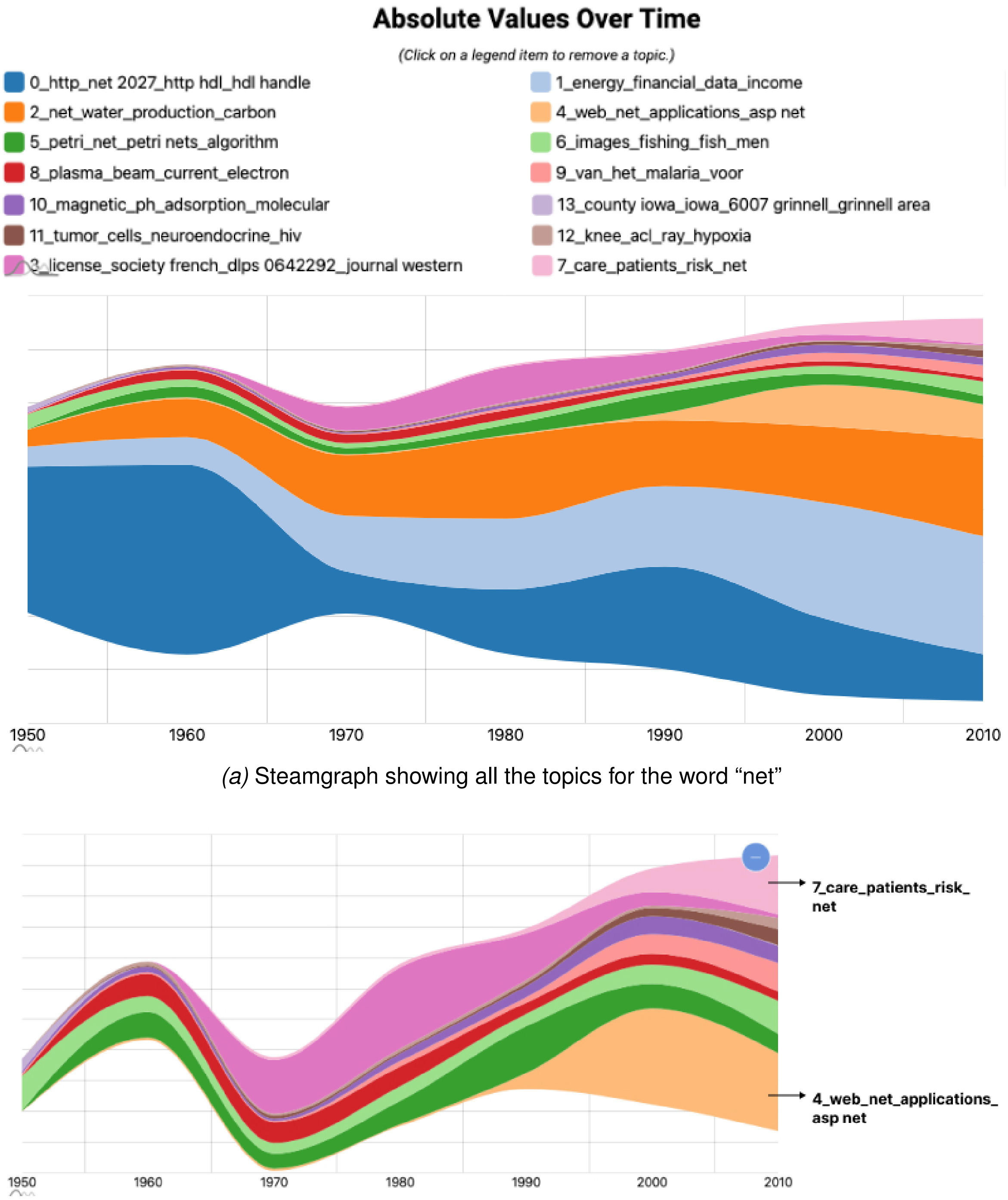
“net”:

Using another example, Figure 5.14 shows the absolute usage values of the word “net” over time. Topics “0_http_net 2027_http hdl_hdl handle”, “3_license_society french_dlps 0642292_journal western”, and “13_county iowa_iowa_6007 grinnell_grinnell area” are considered noise as they are separate clusters of distinct web addresses, while topic “9_van_het_malaria_voor” is a cluster of non-English documents. Two instances of novel word senses are detected from the remaining topics. Topic “4_web_net_applications_asp net” is about the .NET framework, which began developing the 90s. Topic “7_care_patients_risk_net” also emerges more prominently from the 90s onward, and seems to primarily deal with the concept of a healthcare safety net.

While viewing these examples, it is important to note that the detection of shifts is significantly influenced by the quality and breadth of the underlying data, as is the pinpointing of the exact moment when a word begins to take on a new sense. However, these examples illustrate that the tool effectively identifies the emergence of a new word sense and the specific time period in which it takes place.

5.5 Summary

To summarize, this chapter provides a detailed explanation of the methods and technology used to create the visualization tool. We examine the tool’s user interface, discuss its design principles and goals, and explain how its visual components work. Furthermore, we describe the interactive features added to enhance the user experience and analytical capabilities of the tool, and provide an example of using the tool to identify novel word senses for a word.



(a) Steamgraph showing all the topics for the word “net”

(b) Steamgraph for the word “net” with the 3 biggest topics hidden to make the other topics more visible

Figure 5.14: Absolute values over time for the word “net”. Each stream represents a different topic.

Chapter 6

User Study Design

In this chapter we outline the user study that was designed to gauge the tool's performance across various dimensions. The user study encompasses both qualitative and quantitative assessments, allowing us to gather insights from participants' experiences while also quantifying the tool's effectiveness in achieving specific tasks. This chapter serves as an explanation of the framework used to evaluate the tool as a whole and the visualizations in them on the basis of how they address the questions previously proposed. The results of this study will be discussed in the next chapter.

6.1 Goal of the Study

The objective of the user study is to evaluate the usability and functionality of the visualization tool designed for semantic shift analysis. The tool aims to help users visually analyze and understand the changes in word meanings and context over time. Specifically, the study aims to assess how the tool can aid in understanding context-based senses of words and their evolution over time. Due to constraints in time and resources, a convenience sampling strategy [55] was employed to select participants who were readily accessible and willing to participate in the study. While this method does not allow for generalization of the results to the wider population, it does provide valuable insights within the specific context of the study. The study aimed to engage a diverse user population, drawn from both OCLC and academic settings. This was to ensure that participants had a background in technology and language proficiency, which were essential for this study. These participants possessed technological backgrounds, held at least a university-level education, demonstrated proficiency in interpreting visualizations, and exhibited a good command of the English language. The participant pool was equally distributed across genders and spanned various age groups. Details of the participant demographics are discussed in Chapter 7.

The goal is to measure the tool's performance for the following tasks:

- **Exploratory Analysis:** Assess the tool's capability in aiding users to explore different senses of a word and their prominence over time.
- **Contextual Sense Understanding:** Determine if the tool enables users to comprehend word senses in the context of their usage.
- **Model Selection:** Ascertain whether the tool guides users in choosing appropriate models for downstream tasks. Additionally, understand if a user's intuition for model selection aligns with numerical measures such as coherence scores.
- **Usability and Engagement:** Evaluate the tool's user-friendliness, engagement factor, and visual appeal.

6.2 Introducing the study to the participants

The entire study is designed to have a duration of 45 minutes. The first 5 minutes are for introducing the study and the visualizations, followed by 30 minutes of task-based experiments, 5 minutes for a feedback interview, and 5 minutes for the post-study questionnaire. Our approach follows established guidelines for conducting user studies, drawing from the work of Bakalov et al. [56], as one of the references in the field of interactive tool evaluation. This framework divides the study into three key components: i) Task-based experiments, ii) feedback interviews, and iii) a post-experiment questionnaire.

For the introduction, we introduce the visualization tool with the help of a demo video outlining the goals of the tool, the different visualizations and how to read them, and general usability methods of the tool (how to filter, zoom, select etc.). The user can also hover over individual charts to receive an explanation of the visual.

6.3 Evaluating the Study

In order to evaluate the tool from the functional and aesthetic perspectives, we make use of 2 quantitative experiments: a task-based experiment to evaluate the functionality of the tool, and an online questionnaire post the study to evaluate the usability of the tool; along with a qualitative analysis in the form of a post-study interview asking for the user's specific feedback.

6.3.1 Task-Based Experiments

In general, task-based experiments are conducted to test how intuitive an interface is. To measure this, the participants are given specific tasks to solve using the interface. For our experiment, we identified 7 tasks representing common goals that the tool hopes to achieve: exploring word senses, exploring sense usages over time, and selection of models to create the word senses.

In the first 4 questions, the participant selects a word of their choice and the tasks are intended to measure how well the participant is able to use the right tool to complete that task. This is also intended as a way for the participant to familiarize themselves with the working of the tool. In these questions, the tasks' success will depend on whether the participant is able to find the right tool for the right task. The average time taken to complete the task across the range of participants will also be measured.

For the next 3 questions, the participants will be asked to select a particular word and complete tasks pertaining to that word. The goal for these tasks is to see whether participants are able to correctly answer the questions; and measuring the accuracy amongst all the participants. By having the participants choose the same word, we ensure that the results obtained are comparable across all participants. Moreover, by completing these tasks after the first 4, we hope that the possibility of an error occurring due to usability (e.g., not able to find the right tool) rather than functionality (e.g., not inferring the right insights from the data) is minimized.

Tasks:

Questions for word of choice: This is meant to familiarize participants with the interface, and check whether the right chart is being used to access the right information.

- **Task 1: Identifying Prominent Senses Over Time:** "When viewing the different word senses over time, is there any particular sense that has stayed the most prominent?"

- **Task2: Identifying Gaining Prominence Over Time:** "Can you identify any sense that has increased in prominence as time went by? If there is, can you also identify the biggest increase you see?"

- **Task 3: Identifying Similar Senses:** "For the different word senses, are there any that seem similar to each other? Think-out-loud during your process."

- **Task 4: Matching Senses to Dewey Codes:** "Which word sense has the closest alignment with its corresponding Dewey code category?"

Questions for word "mouse" selection: We ask participants to answer the next 3 questions for the same word: "mouse". This is done so that the participants conduct the following tasks on the same set of data, enabling their answers to be compared.

- **Task 5: Comparing Results from Different Models:** "What steps would you take to see how each model is performing?"
- **Task 6: Identifying New Sense Over Time:** "If you were a historian that had to report about whether this word had gained a new sense over time, what sense would you choose, and why?"
- **Task 7: Choosing Model based on Dewey categorization:** "As a librarian, if you had to categorize all documents containing the word "mouse" in terms of their different senses instead of their Dewey codes, what model would you choose? Why? Think out loud for your decision making process."

Tasks 1, 2, and 6 are questions about topic usage over time, whereas Tasks 3, 4, 5, and 7 are about inspection of the word senses.

6.3.2 Interviews

The interview is designed to gain feedback from the participants and perform a qualitative assessment based on 3 criteria:

- **Positive Characteristics:** "Starting with the positive, were there any aspects of the tool that you found particularly useful or engaging?"
- **Negative Characteristics:** "Were there any negative aspects of the tool that you found confusing or difficult to interpret?"
- **Suggested Improvements:** "Based on your experience with the visualization tool, do you have any recommendations for enhancing its usability, clarity, or effectiveness in conveying information?"

After performing the tasks, we ask the participant for their feedback on the above criteria, which are then evaluated qualitatively to observe any common phenomena, to find out what is working as intended, and to see where improvements can be made.

6.3.3 Questionnaire

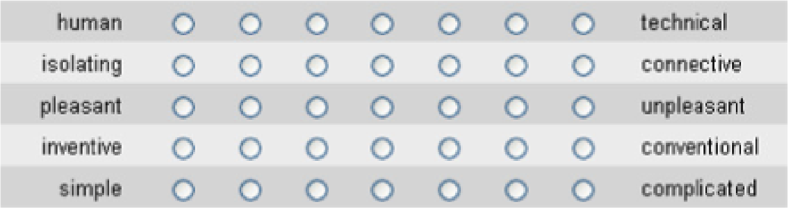
A common method to measure the usability of an interactive system is through standardized questionnaires, which are well-known, reliable, and inexpensive instruments to to evaluate User Experience (UX), composed of Likert scales and semantic differentials [57]. The two most commonly used questionnaires of this type are the AttrakDiff [58] questionnaire (60.24%), and the "User Experience Questionnaire" (UEQ) [59] (37.08%), according to [60]. Both these questionnaires measure usability as a function of the product's pragmatic and hedonic attributes. Both tools score similarly on their reliability and validity analyses [61]. However, we opt for the AttrakDiff questionnaire as our evaluation instrument based on its broader adoption and utilization amongst usability practitioners.

A key factor for selection of the AttrakDiff questionnaire is its accessibility; it is conveniently available as an online questionnaire on its dedicated website. This accessibility ensures that participants can easily complete it, and it also facilitates the generation of statistical analyses based on the responses. We use it in the "Single-Evaluation" mode, which is most suitable for our task of one-off evaluation for a product (instead of its other two modes: "Comparison A-B" and "Before-After"). The reliability is shown for the different parameters that it measures: hedonic quality - stimulation (Cronbach's alpha 0.79 - 0.90), hedonic quality - identity (Cronbach's alpha 0.73 - 0.83) and pragmatic quality (Cronbach's alpha 0.83 - 0.85), from studies in [58], [62]. These values demonstrate the questionnaire's reliability in measuring the constructs relevant to our research, ensuring the robustness of the analysis.

The AttrakDiff questionnaire addresses the subjective attractiveness of a product as a composite characteristic influenced by four qualities:

- Pragmatic Quality: The inherent usability of a product that indicates how successful the users can achieve their goals with the product.
- Hedonic quality-identity(HQ-I): The ability to develop the identity and help the user to establish personal connection with the product.
- Hedonic quality - stimulation (HQ-S): The ability to stimulate the need for further use.
- Attractiveness (ATT): The general outward appearance of a product.

The questionnaire consists of 28 word-pairs organized according to the four qualities. For each pair, the participant votes on a seven-value Likert scale (example in Figure 6.1).



human	○	○	○	○	○	○	○	technical
isolating	○	○	○	○	○	○	○	connective
pleasant	○	○	○	○	○	○	○	unpleasant
inventive	○	○	○	○	○	○	○	conventional
simple	○	○	○	○	○	○	○	complicated

Figure 6.1: Example of word-pairs in the questionnaire

6.4 Summary

This chapter serves as a comprehensive guide to the User Study Design, which aims to evaluate the visualization tool. It begins by articulating the study's core objectives, namely to assess usability and functionality across various tasks. It further outlines the study's structure, consisting of Task-Based Experiments, Feedback Interviews, and an Online Questionnaire.

In the next chapter, we delve into the results and findings from this user study, providing insights into the tool's performance, user experiences, and areas warranting enhancement.

Chapter 7

Results and Discussion

The study enlisted the participation of 10 individuals, exhibiting a diverse cross-section of characteristics. The demographic details below were self-entered by the participants as part of the AttrakDiff post-study questionnaire. Gender distribution among participants was fairly even, with 6 male and 4 female participants. All participants had achieved a university-level education. The participants' age groups were diversified, with 6 falling in the 20-40 years bracket, 2 in the 40-60 years category, and 2 aged 60 and above, ensuring representation across different generations. The range of occupations among participants, including students, software engineers, individuals from the psychology field, software developers, freelancers, and senior software engineers, provided a multifaceted view of the tool's usability and functionality. Two participants chose not to disclose their professions. 5 participants were associated with the company, OCLC, while the remaining half were university students.

The selection criteria for participants included having a background in technology, holding at least a university-level education, demonstrating proficiency in interpreting visualizations, and exhibiting a good command of the English language. These criteria were established to ensure that participants possessed the necessary qualifications to engage effectively with the semantic shift analysis tool under evaluation. As explained in Section 6, the participants were recruited via a convenience sampling strategy, which was necessitated by constraints in time and resources. Convenience sampling involves selecting participants who are readily accessible and willing to participate in the study, making it a pragmatic choice for this type of study, involving a detailed hands-on usage of the visualization tool while carefully noting down participant decisions and task results. While this convenience sampling strategy allowed for the inclusion of individuals who met the established criteria and were readily accessible, it does have limitations. It may introduce selection bias and restrict the generalizability of the study's findings beyond the specific cohort of participants. Consequently, the results and discussions in this section should be interpreted within the context of this sampling approach, recognizing the need for larger and more diverse participant samples to achieve broader generalizability.

Discussion on Participant Size

The question of how many participants are needed to reach the saturation point in usability testing has long been a subject of debate in the fields of Human-Computer Interaction (HCI) and usability testing. Early usability studies followed the "5-user assumption", which revealed that 5 users could reveal about 80% of all usability problems that exist in a product [63], [64], beyond which is a point of diminishing returns with more participants increasing the time and costs required. This widely held assumption has been challenged by empirical studies, finding that five users finding 80% of usability problems may not always hold true [65]. In some cases, sets of five participants revealed as low as 55% of the issues. While there is no "magic number" for the number of participants required for such a study, [66] suggests that a rule of 16±4 users gains validity in user testing.

Given the findings from these studies and recognizing the complexity of our own research, we chose to include 10 participants in our usability study. While 10 participants may still be considered a relatively small sample size to ascertain quantitative validity, it offers valuable qualitative data and user feedback, as well as identifying usability issues within the tool, which can serve as a foundation for further research and can provide guidance for designing similar tools for a broader audience. This sample size of 10 participants is then a practical compromise considering the resource constraints of our study (including time and budget).

7.1 Task-Based Results

This section describes the quantitative results obtained from the task-based experiments from two perspectives: time taken to complete the task, and selection of the chart for the task. Figure 7.1 shows a box plot of the time taken by a total of 10 participants for each of the tasks, while figure 7.2 shows the distribution of visualizations chosen to complete each of the tasks. Participants are allowed to use multiple visualizations to complete each task.

The results shown are based on the data collected during the task completion section of the study. This data consists of audio recordings of the users completing the tasks, who were encouraged to think out loud during the tasks, and explain their reasoning for taking actions. The system screen was also recorded during these tasks to keep track of the the user interactions. These recordings were manually processed for each participant. Timestamps were tagged to identify the initiation and

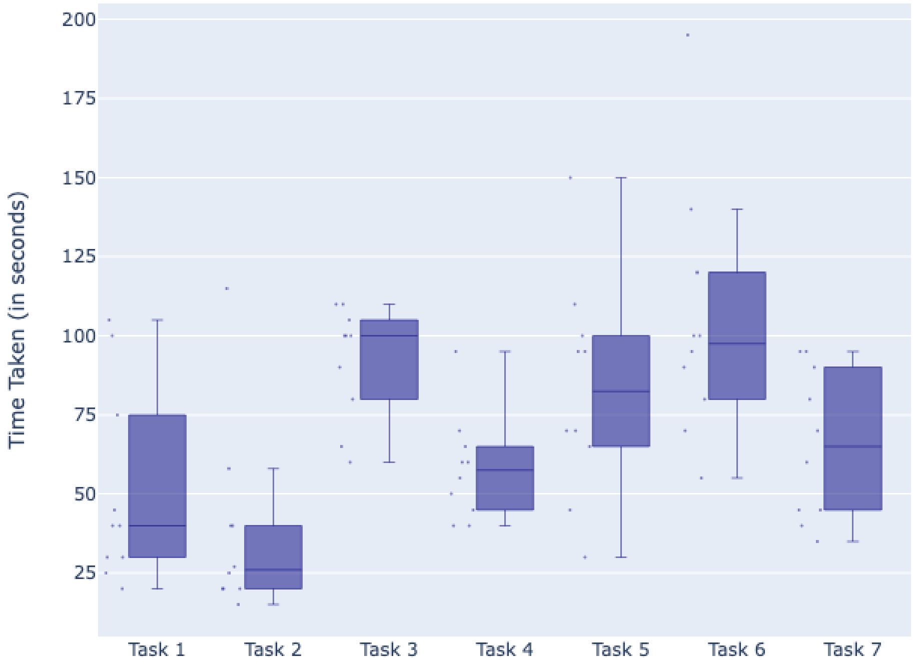


Figure 7.1: Distribution of time taken to complete each task (n=10 participants)

completion of specific tasks, providing insights into the time taken for task execution. User comments and spoken feedback, integral for understanding user preferences and thought processes, were transcribed verbatim to textual form for each task. The combination of timestamp annotation and transcription played a pivotal role in deriving quantitative and qualitative insights into user behavior, which are presented and discussed in the subsequent sections.

Task 1: Identifying Prominent Senses Over Time

"When viewing the different word senses over time, is there any particular sense that has stayed the most prominent?"

This task focuses on word sense usage over time, and each participant intuitively selected the "Topic Usage Over Time" section to solve it. The median time taken to complete this task was 40 seconds, with a minimum of 20 seconds by P2 and a maximum of 105 seconds by P5. P2 selected the word "creep" for their task, which had a total of 6 topics. The first topic ("0_creep_stress_strain_materials") for the default selection of `model1` had the highest proportion of documents and the participant immediately selected it as the most prominent sense. Participant P5 selected the word "virus" as their search term. This had 15 topics in `model1` with no distinct prominent sense. The participant isolated each topic individually and proceeded to also check the other models, talking out loud through their thought process and verbalizing the meaning of each topic as they inspected it.

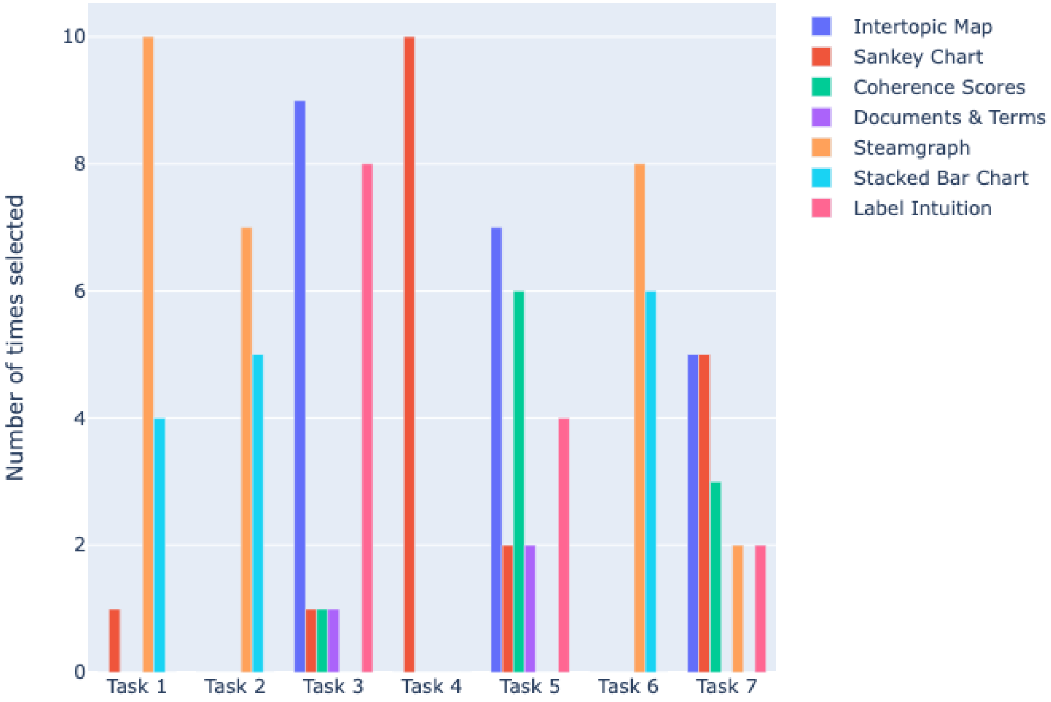


Figure 7.2: Distribution of charts selected by participants (n=10) to complete tasks

Here, both the steamgraph and the stacked bar chart can be used to select the sense that is most prominent. This can be done by choosing the word sense which has a high stream width, or the sense that makes up a high proportion in the stacked bar chart. For most target words, the first word sense contains the highest number of documents and this was the sense chosen by all the participants for their respective target words. As visible in the chart selection figure (Figure 7.2), all participants used the steamgraph. Four participants additionally consulted the stacked bar chart to see the proportion of the sense they selected.

Participants were able to correctly interpret the charts to search for the information needed. Their thought processes, shared during the task as they thought aloud, revolved around assessing the topic proportions over time and identifying the most prominent color in the charts. For example, Participant 4 mentioned, “I’m looking at the topic proportions over time and seeing which color is used the most,” while Participant 2 described their approach as selecting the topic that “at any point in time, makes up the most of the chart.” This common interpretation method was observed across all participants, affirming the correct understanding and utilization of the steamgraph and stacked bar chart for this task.

For the data point showing the Sankey chart being used (Task 1 in Figure 7.2, participant P1 initially tried searching for the most prominent sense using the Sankey chart visualization (Figure 5.4) and selected the most prominent sense from it. During the task, we then reiterated the question, asking to choose the most prominent sense over time; after which they moved to the “Topics Over Time” section and followed the procedure as expected.

Task 2: Identifying Gaining Prominence Over Time:

“Can you identify any sense that has increased in prominence as time went by? If there is, can you also identify the biggest increase you see?”

This task once again required the participant to refer to the temporal charts (steamgraph and stacked bar chart), and all the participants proceeded with the chart usage correctly. The median time taken to complete this task was 26 seconds, with a minimum of 15 seconds by P5, and a maximum of 115 seconds by P3. P3 was the only participant that went beyond the one-minute mark for this task. They chose the word “lecture”, which had its first two topics stable across the decades and taking up most of the visual space, which led the participant to spend time using the controls to isolate the remaining topics, while zooming in and out to select the topic. For the word, they commented that there seemed to be no apparent overall increase, but identified topics that “increased in a few decades”, while staying stable in the rest.

This task, notably, had the shortest median completion time among participants. This can be attributed to the use of the same chart as the previous task. Familiarity with interpreting word senses over time from the previous task enabled users to swiftly complete this task without the need to switch between different contexts or visualizations.

Task 3: Identifying Similar Senses:

“For the different word senses, are there any that seem similar to each other? Think-out-loud during your process.”

This task involved moving away from the “Topic Usage Over Time” section and to the “Inspect Topics and Scores” section. While six participants were able to make the change naturally, four participants continued trying to complete this task by looking at the topic labels in the steamgraph and stacked bar chart and trying to intuitively find semantically similar topics based on their labels, before moving on to the “Inspect Topic and Scores” section and using the visualizations there. This movement took a mean time of 45 seconds for the participant and involved them using their intuition to answer the question before exploring other charts. This behaviour could be explained by the fact that participants had to answer the first two questions using the same two charts in the same section and had become familiar with using the same visuals to solve the tasks. Future experiments could try experimenting with the order of the questions, asking them randomly for each participant to measure whether this plays a role in chart selection.

The median time taken to complete this task was 100 seconds, with a minimum of 60 seconds by P4, and a maximum of 110 seconds by P1 and P6. Participant P4 readily moved to the "Inspect Topics and Scores" section and used only the Intertopic Map distances to make their choice, explaining the reduced time taken. Of the remaining participants, 8 of them additionally used their intuition to validate the topics that were similar.

Nine out of the ten participants correctly identified the intertopic map and grouped similar topics together based on identifiable clusters formed within the map, while participant P3 used their intuition by checking the label names in the Coherence Scores section. The other charts (coherence scores, documents & terms information) were used to read topic labels and find similar topics based on prior knowledge of semantic similarity, with P1 inspecting the Documents & Terms to see whether semantically similar topics shared common top terms. Participant P7, when using the intertopic map to assess similarity, identified three closely positioned topics within a cluster as similar topics. These topics were associated with the word ”peaks” and were labeled as ”peaks_peak_data_spectra_temperature,” ”sea_peaks_records _carbonate_glacial,” and ”runoff_rainfall_drainage_flood peaks_catchments.” However, the participant also noted that these topics did not appear to be similar in semantic meaning. Despite their proximity on the intertopic map, the participant expressed a need for more information to be convinced of their actual similarity.

The usage of the Sankey Chart was not intended as a potential method to complete this task. Participant P2, however, used both the Intertopic Map and the Sankey chart to evaluate topic similarity. They speculated that topics with the same Dewey code might be indicative of similar word senses.

Overall, we found that almost all participants (9 in total) made use of a combination of the intertopic map and their prior intuition about word senses to find word similarity. Distance as a measure for the intertopic map was used as intended, with similar senses forming clusters and being identified correctly as a similarity measure by the participants. When using other charts for finding similarity, participants mostly utilized their prior knowledge to select senses that were similar to each other. This was noted down separately as "Label Intuition", and involved selecting topics whose labels had similar meanings, or, in the case of Documents & Terms, whether they shared any top terms. The term ”Label Intuition” in this study refers to the phenomenon where users rely on their intuitive faculties for pattern recognition and decision-making, in this case, relying on their intuitive faculties for deciding which labels are similar to each other. This concept is grounded in the theories of intuitive systems as pattern recognition and long-term (procedural) memory, as discussed in [67], drawing from the fundamental principles of dual systems theory, which posits that human cognition results from the interaction between an analytical reasoning system (comprising working memory and long-term declarative memory) and a swift, autonomous intuitive or implicit system characterized by implicit pattern recognition and procedural long-term memory.

The concept of “Label Intuition” is a term given specific to this study for the phenomenon when a user uses their intuitive systems for pattern matching and decision making. We draw this from the concepts of intuitive system as pattern recognition and long-term (procedural) memory in [67], which uses the core ideas of dual systems theory literature, which states that “human cognition derives from an interplay between an analytical reasoning system (composed of working memory and long-term declarative memory) and a rapid, autonomous intuitive or implicit system that entails implicit pattern recognition and procedural long-term memory”.

Task 4: Matching Senses to Dewey Codes:

“Which word sense has the closest alignment with its corresponding Dewey code category?”

The median time taken to complete this task was 57.5 seconds, with a minimum of 40 seconds by P7 and P9, and a maximum of 95 seconds by P1. P1 was an outlier in terms of time taken and spent the time thinking out loud and inspecting topics in the Sankey chart for all the models. This task seemed the most straightforward in terms of visualizations chosen to complete it. Mentioning the Dewey codes in the task instruction made it apparent for all the participants to use the Sankey chart (Dewey code mapping) to complete this task.

Additionally, all participants were able to interpret the chart correctly to answer this task question. Their reasoning, as expressed when asked to think out loud, was to discard topics that were diverging and flowing to multiple Dewey codes, while selecting the topic that flowed to only one or two codes.

The point of failure for this task seemed to lie with prior knowledge of Dewey codes. Four out of five participants from OCLC were able to interpret the meaning correctly (the other mentioned not using the Dewey codes in their work and hence being unfamiliar with the system), while most of the outside participants had to be briefed about what the codes represented, despite the presence of the ”information” circle. This observation highlights the potential limitations of instructional aids in bridging such gaps and can be overcome by surfacing the visualizations relevant to a user’s prior knowledge, as also suggested by Participant 6 (Appendix A).

Task 5: Comparing Results from Different Models:

"What steps would you take to see how each model is performing?"

In this task, participants were tasked with evaluating the performance of different models for the word "mouse". This task displayed notable subjectivity in terms of approach, evidenced by the considerable variation in completion times and the range of charts selected, making it the task with the most varied response in the study. The median time taken to complete this task was 82.5 seconds, with a minimum of 30 seconds by P3, and a maximum of 150 seconds by P6. The majority of participants finished this task within one to two minutes. The two outliers arose from a difference in approach: Participant 3 (P3) identified the best model based only on the highest coherence score and without using any other visualization, while Participant 6 (P6) took a more deliberative approach, thoroughly exploring the "Inspect Topics" section, examining the outputs of various visualizations, and vocally evaluating their alignment with their intuitive judgments. Participants mostly relied on coherence scores (6 participants) as a quantitative metric for assessing model performance, and the intertopic map (7 participants) to answer this question. Since the task did not directly ask the user to choose the best model, participants answered by thinking-out-loud about what visualizations they would choose to answer this question. All participants discarded model3 as a good model due to the lower number of topics it showed and its lack of granularity.

Task 6: Identifying New Sense Over Time:

"If you were a historian that had to report about whether this word had gained a new sense over time, what sense would you choose, and why?"

This task required going back to the "Topic Usage Over Time" section, which all participants did correctly, and then use the steamgraph and stacked bar chart. To keep the answers uniform, we asked the participants to select a sense from `model1` for target word "mouse". Figure 7.3 shows a snapshot of the initial plot of the steamgraph and stacked bar chart. The correct response to this query would be either the "2_computer_mouse_interface_software" topic or the "3_cat mouse_game_game cat_murder" topics (terms are separated by underscores; a space between words is an n-gram term). This involves identifying a new sense based on an increase in the width of the steamgraph from decades where the width of the stream is empty or negligible.

The median time taken to complete this task was 97.5 seconds, with a minimum of 55 seconds by P9, and a maximum of 195 seconds by P1. P1 is an outlier in terms of the time taken to complete the task, having spent most of the time inspecting each topic individually in both the steamgraph as well as the stacked bar chart.

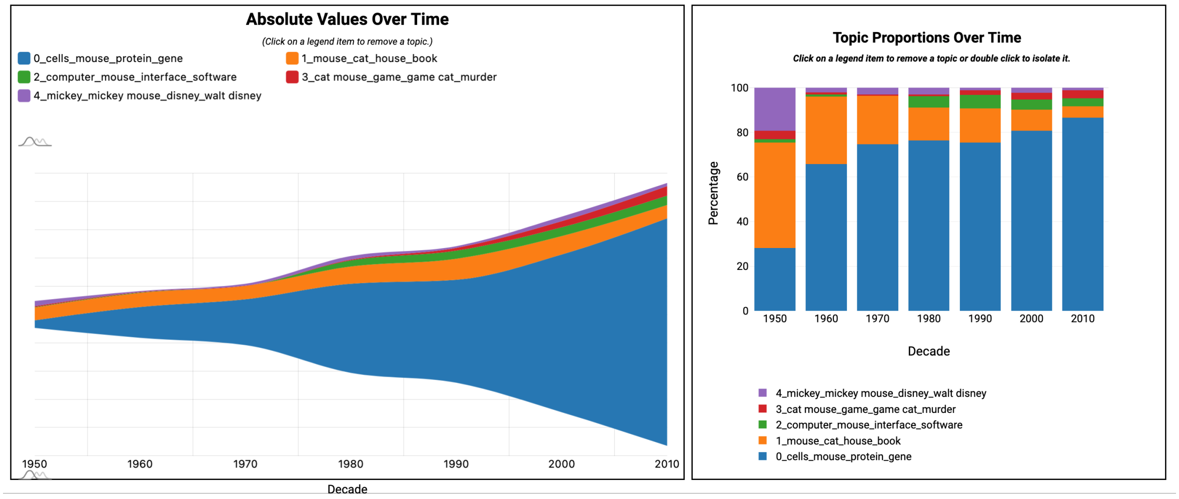


Figure 7.3: Overview of the Steamgraph and Stacked Bar Chart showing word sense evolution for the word “mouse”.

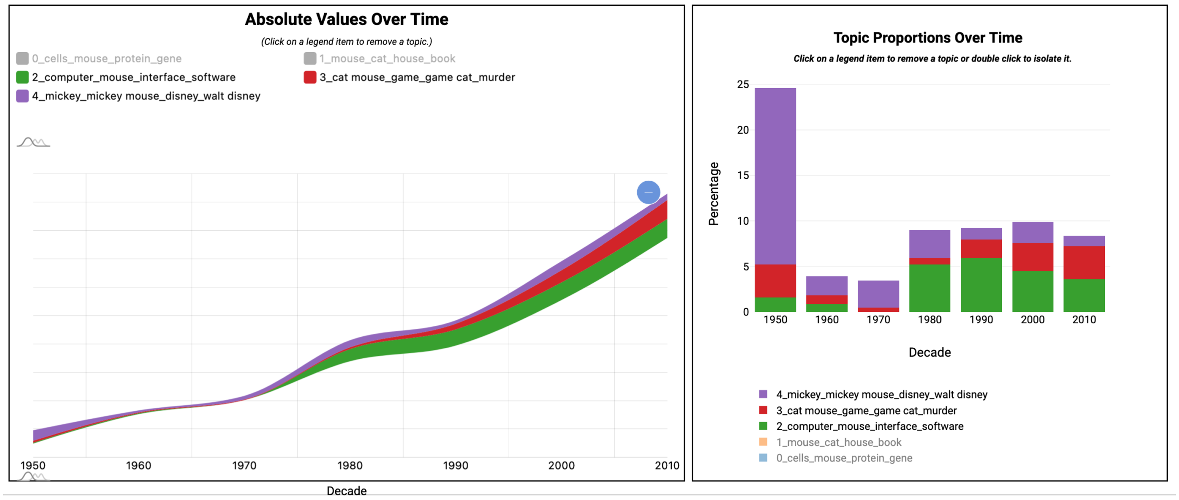


Figure 7.4: Overview of the Steamgraph and Stacked Bar Chart after hiding the prominent topics (1 and 2) from the plot to surface the less prominent topics.

Participant P9, who spent the shortest time on the task, already had an intuition about the "2_computer_mouse_interface_software" topic being a newly gained topic and utilized the steamgraph to confirm their intuition.

All participants made their decisions by "looking at senses which aren't there at the start but are present towards the end," for the steamgraph, or "checking to see if the proportions of a topic are increasing at a particular decade" for the stacked bar chart. They were able to identify the point at which the topic gained the new sense (during the 70s) and tried to intuitively make sense of the reason for doing so. Two participants found it hard to inspect the topics before they increased in prominence since the small area or proportion of the visualization made it difficult to clearly identify the smaller topics amidst the larger ones. They were prompted to use the interactivity features to zoom/isolate a topic to make it clearer in such cases. Figure 7.4 shows a snapshot of the charts when hiding the more prominent topics to enlarge the less prominent ones.

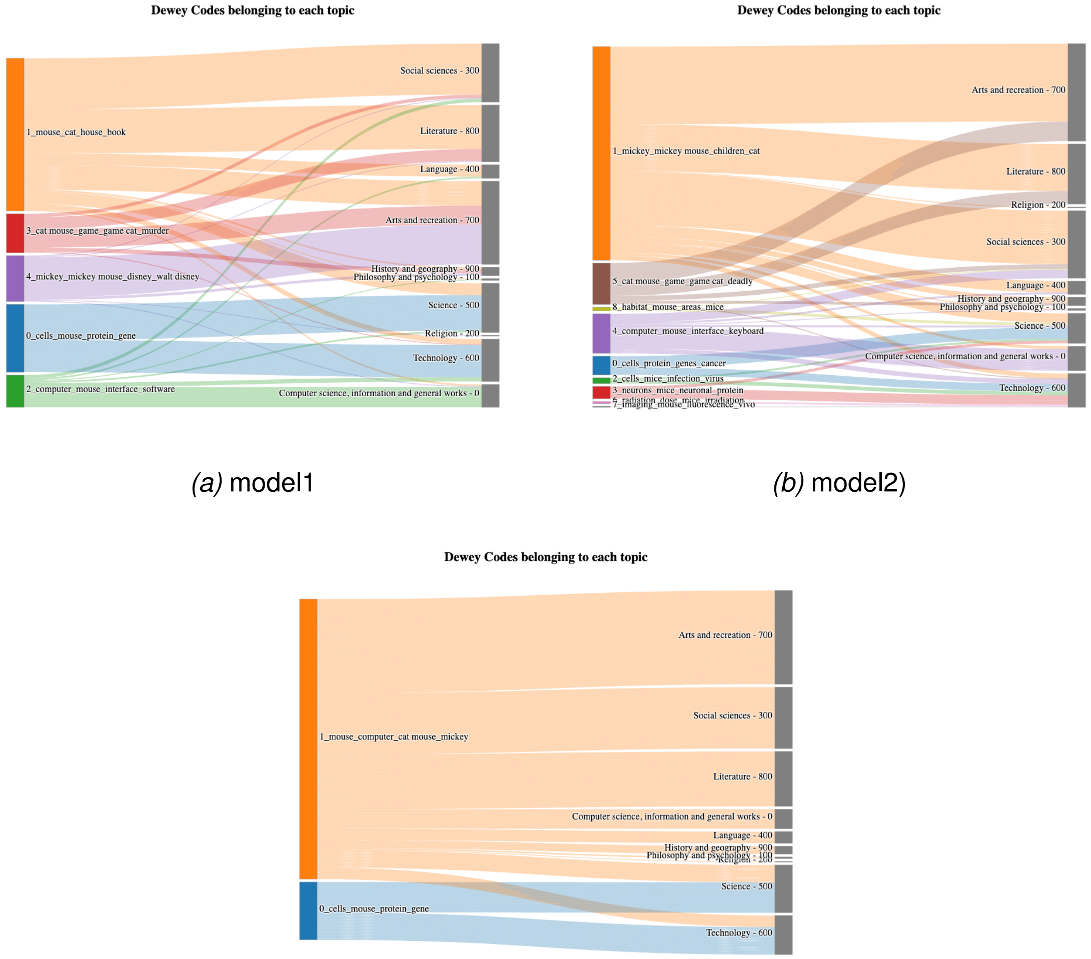
Four participants chose the "3_cat mouse_game_game cat_murder" sense of the word as the one that gained a new sense, while six participants chose "2_computer_mouse_interface_software". Two of these participants chose both the topics as their answer.

An unexpected topic selection was the "0_cells_mouse_protein_gene" topic, which was selected by two participants as having gained a new sense. Their reasoning for this was that the sense was increasing in prominence as the decades went by, attributing this increase to the gaining of a new sense for the word. This is a wrong interpretation of the chart and can be attributed to the more prominent topics taking up most of the plot space, rendering the smaller topics negligible at first glance. While this misinterpretation was observed in only one participant, it underscores the potential challenge when more dominant topics overshadow smaller ones in the visualization, making smaller topics seem negligible at first glance. To mitigate such misinterpretations in the future, interactive features like zooming or isolating specific topics for a closer examination can be made more prominent. Additionally, providing user guidance or tooltips within the tool itself, as well as offering user training that emphasizes the importance of inspecting all topics, can enhance comprehension and reduce the likelihood of similar errors.

Task 7: Choosing Model based on Dewey categorization:

"As a librarian, if you had to categorize all documents containing the word "mouse" in terms of their different senses instead of their Dewey codes, what model would you choose? Why? Think out loud for your decision making process."

The median time taken to complete this task was 65 seconds, with a minimum



(a) model1

(b) model2)

(c) model3

Figure 7.5: Mapping of the topics generated to the documents' corresponding Dewey Codes for each model

of 35 seconds by P9, and a maximum of 95 seconds by P8 and P10. Participant P9 quickly chose `model1` by utilizing the Intertopic Map, after having explored the different models during Task 5. They mentioned it `model1` being "the most selective model" due to "having the most distance between all of its topics". While all the participants looked up multiple charts (all from the "Inspect Topics and Scores" section) to inspect the topics, only half of them actually crossed checked their results with the Sankey charts and the Dewey codes.

Figure 7.5 shows the distribution of each of the topics to their corresponding Dewey codes for each of the models. All participants were quick to discard `model3` as the best model due to only having two topics, which they felt did not capture the range of meanings of the word, and it not being able to discriminate between the different senses that models 1 and 2 were able to.

Table 7.1: Different Coherence Scores measures for the different models of the word “mouse” (Rank of the model is in parentheses.)

Coherence Type	Model 1	Model 2	Model 3
u_mass	-0.387 (3)	-0.297 (2)	**-0.062 (1)**
c_v	0.544 (3)	**0.640 (1)**	0.632 (2)
c_uci	-1.872 (2)	**-1.090 (1)**	-2.153 (3)
c_npmi	0.040 (2)	**0.049 (1)**	-0.019 (3)

8 participants chose `model1` and 2 chose `model2`. For those who selected model1, the reasoning they provided was unanimous: it had the best categorized topics with the largest distance between topics, leading it to be sufficiently discriminative, whereas `model2` contained topics that were similar to each other and some very small topics, while `model3` only contained 2 topics and did not discriminate between senses to the same extent as the other two models. The two participants who selected `model2` explained their reasoning as follows: P3 chose the model due to it having the highest coherence score, while P5 mentioned `model2` having a better level of granularity compared to the other two.

Based on the results of Task 7, several conclusions can be drawn:

Preference for Model1: The majority of participants (8 out of 10) chose model1 as the best model for categorizing documents containing the word “mouse” based on different senses. Their reasoning centered on model1 having well-defined and distinct categories that effectively captured the various meanings of the word. These participants found it sufficiently discriminative and aligned with their intuitive understanding of the word’s senses.

Reasons for Discarding Other Models: Participants were quick to discard model3 due to its limited number of topics (only two). They felt that these two topics did not adequately capture the range of meanings associated with the word “mouse”. model2 was also less favored by most participants due to its topics being similar to each other, as well as the presence of very small topics.

Coherence Scores vs. Participant Choices: While coherence scores are often used as a quantitative measure to assess the quality of topic models, participants’ choices did not uniformly align with these scores. Only two participants (P3 and P5) mentioned coherence scores in their decision-making process. P3 chose model2 based on its higher coherence score, while P5 selected it for having a better level of granularity compared to the other two models. This indicates that coherence scores were not the primary factor influencing participants’ decisions.

Table 7.1 shows an overview of the different coherence scores for each of the models. While each of the coherence type calculates the values differently, they all attribute a higher score to a more coherent model. Interestingly, participants' preferences for `model1` did not align with the models' coherence scores. `model2` had the highest coherence score for 3 of the coherence types, and `model3` had the highest for the `u_mass` coherence_type. Despite this, the majority of participants (8 out of 10) favored `model1` due to its superior ability to represent distinct categories for word senses.

In summary, participants' selections in Task 7 were primarily guided by the interpretability and discriminative power of the topic models rather than coherence scores. This suggests that when choosing a model for semantic categorization, factors beyond quantitative measures like coherence scores are crucial, emphasizing the importance of visual interpretability and alignment with users' intuitive understanding of the data.

7.2 Usability Results

We assessed the usability of the tool with user feedback provided on the AttrakDiff questionnaire as the final step of the study. The usability was measured based on the tool's hedonic and pragmatic qualities, and its results are elaborated below.

Portfolio of Average Values

Figure 7.6 shows an overview of the received feedback with respect to the pragmatic (x-axis) and hedonic (y-axis) qualities. In this view, the values of the hedonic quality are represented on the vertical axis (bottom = low value), and the values of the pragmatic quality are represented on the horizontal axis represents (left = low value). As visible from the figure, the tool has a high value of hedonic quality and slightly above average value of pragmatic quality, leading it to be placed in the "self-oriented" character-region.

In terms of hedonic quality, the tool received high scores, indicating its success in aspects related to novelty, appeal, engagement, and stimulation—qualities associated with its user-centric objectives. However, in terms of pragmatic quality, the tool's performance, while above average, did not reach the highest level. This suggests that there is room for improvement in aspects related to its practical usability and effectiveness in achieving desired outcomes. The confidence rectangle, which reflects the reliability and consistency of participant feedback, predominantly falls within the "self-oriented" character region. This suggests that the feedback collected from participants was relatively consistent and reliable, without significant variations.

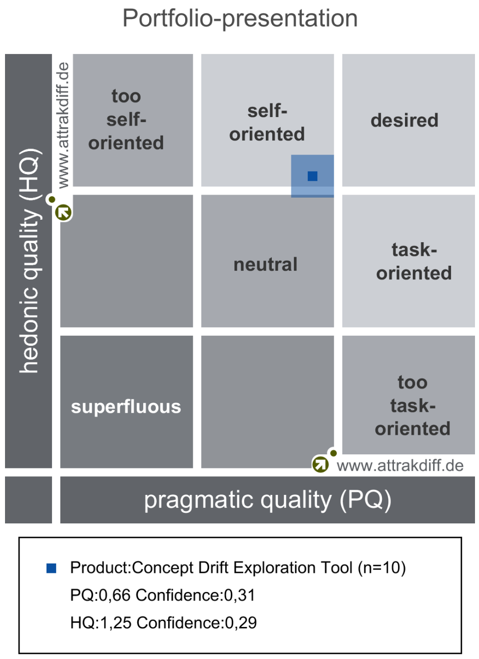


Figure 7.6: Portfolio view showing average values of the dimensions PQ (Pragmatic Quality) and HQ (Hedonic Quality). The blue square shows the tool's position on this scale with a confidence rectangle around it.

Diagram of Average Values

Figure 7.7 shows the mean value of each of the four measures influencing the overall attractiveness of the tool. To reiterate, the four measures are:

- **Pragmatic Quality (PQ):** Describes the usability of a product and indicates how successful users are in achieving their goals using the product.
- **Hedonic Quality - Identity (HQ-I):** Indicates the ability of the tool to communicate a unique identity.
- **Hedonic Quality - Stimulation (HQ-S):** Indicates the extent to which the tool facilitates novelty, innovation, and challenge.
- **Attractiveness (ATT):** Describes the overall attractiveness of the tool for the user, or the general outward appearance of a product.

These measures are evaluated with the help of a bipolar semantic differential scale [68] representing opposites (negative/ positive), with 7 adjectives describing each measure (the list of adjectives can be found in Figure 7.8). The middle value is 0, the leftmost value -3, and the rightmost value $+3$. The final values shown in Figure 7.7 are the mean of the 7 adjectives describing that measure.

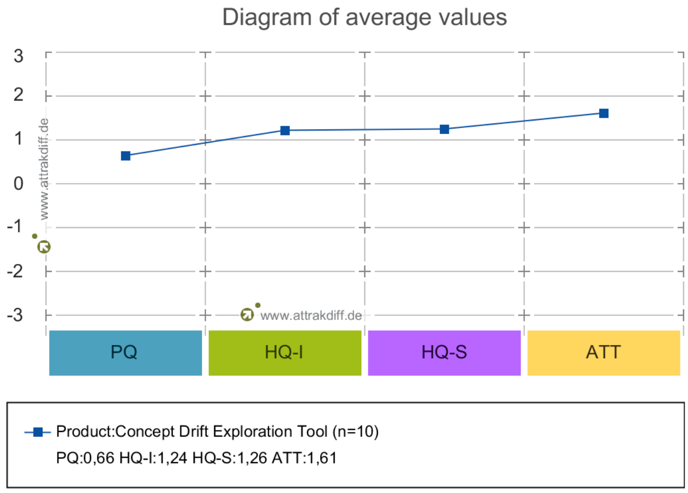


Figure 7.7: Mean values of the four dimensions measured

From Figure 7.7, we see that the hedonic qualities and attractiveness of the tool are above the average range for the tool, suggesting that participants were both, drawn towards the tool, and were engaged by usage of the tool. The pragmatic quality, or the ability of the tool to complete tasks using the tool was positive, but in the average range, which suggests that there is room for improvement in tool functionality for deriving insights.

Description of Word-Pairs

Figure 7.8 shows the mean rating of all the participants for the 28 word-pairs that contribute towards the four measures, and how the participants rated them. The values -3 to $+3$ represent the extent to which the participants rate the tool for any set of opposing adjectives. As seen from the previous results (portfolio and diagram of average values), the hedonic qualities and attractiveness are positive and above average, and that is reflected in the ratings for their corresponding word-pairs.

Two word-pairs, “unpredictable-predictable” and “undemanding-challenging” are closer to the average line (0) than all the others, showing that participants found it relatively challenging to definitively assess the tool’s performance along these dimensions. The “unpredictable-predictable” dimension may pertain to how well the tool presents data or information in a way that users can anticipate or comprehend. This could be interpreted in two ways: that the visualization isn’t showing information to the user in the way they expected it to, or that the nature of the word itself presented some meanings that the user was not expecting.

The “undemanding-challenging” dimension also does not reach a consensus. This dimension is a measure of the tool’s ease of use or the cognitive effort required to interact with it. This subjectivity underscores the potential for refinement

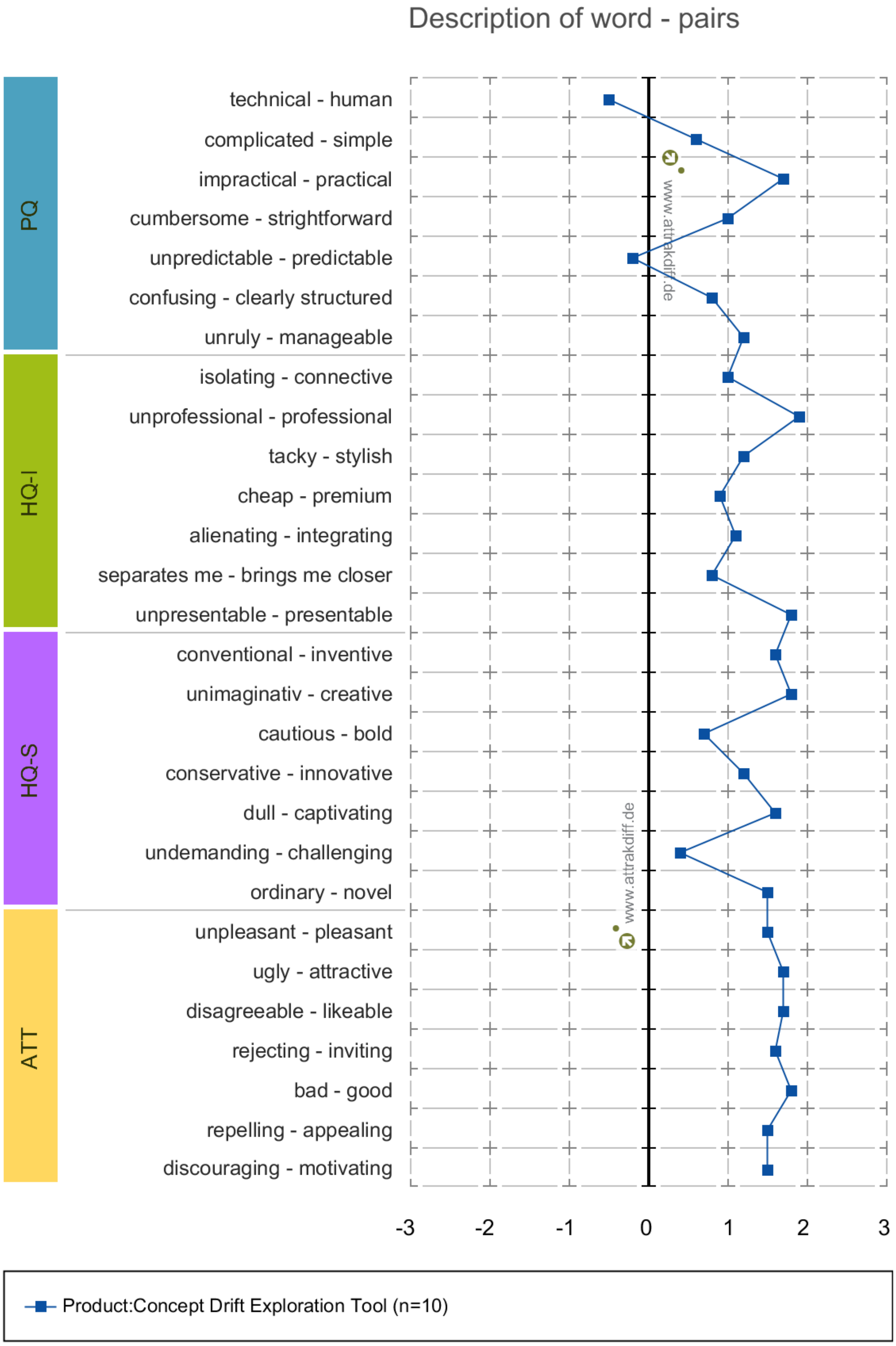


Figure 7.8: Mean values of the word pairs selected by the participants

and enhancement of the tool's usability. It is worth noting that the perception of the tool's ease of use is highly contingent on the specific objectives, particularly whether the tool is intended for exploratory or confirmatory analysis. In the context of exploratory analysis, a higher level of 'challenging' can be advantageous as it encourages users to delve deeper into the data, fostering the discovery of novel insights. On the other hand, when the tool is employed in confirmatory analysis, minimizing demands on cognitive resources may be more desirable to facilitate efficient hypothesis testing and validation [69]. Therefore, the optimal positioning of the tool on the 'undemanding-challenging' spectrum is contingent upon the intended purpose and research methodology, emphasizing the importance of tailoring the tool's usability to the specific needs and goals of the user.

7.3 Feedback Interviews

Like the task-based results, the user feedback interviews after the tasks were also audio recorded. This user feedback was then manually transcribed by the author for each of the participants, with a summary of the positive characteristics, negative characteristics, and suggested improvements shown below.The list of feedback interviews by the participants can be seen in Appendix A.

Positive Characteristics

Participants provided several positive comments regarding the visualization tool as part of the post-study feedback interview process:

- **Interactivity:** All users started off appreciative of the interactive nature of the tool, particularly the ability to select and compare different models. They found it intuitive and enjoyable to work with.
- **Interesting Insights:** "Interesting" was the most used word by the participants when describing the tool. They expressed fascination with the tool's ability to reveal how word meanings change over time, and saw potential applications not only for classification but also for linguistic studies. They appreciated that the tool encourages playing around with different models and visualizations, which helped them to uncover insights about a word in a serendipitous way.
- **Tool Information:** Some participants highlighted the importance of the information provided, both at the start of the study and through the chart information, which helped them understand the tool's context, especially if they were not familiar with the underlying method. They mentioned that this made it easy to get started quickly, and immediately knew where to look for the answer.

- **Preferred Visualizations**: Some users mentioned their preferences for specific visualizations without being prompted for it, with the Steamgraph being a clear favorite. The Intertopic Map also received praise for its utility in visualizing individual and similar topics. In general, the ability to view topics over time seemed to be the most appealing feature. P3 specifically mentioned the coherence scores as being useful since it provided them a statistical method of selecting the best model when they were unsure of their decision. P5 appreciated the ability to view representative documents to aid in their decision making.
- **Clarity:** Users found the tool to be clear and user-friendly, which greatly contributed to their overall positive experience. Participant P1, for instance, praised the interactive nature of the tool, highlighting that it allowed them to easily navigate through different models and customize their analysis. They described it as ”intuitive” and stressed that it provided a much clearer understanding compared to merely working with document scores. The ability to experiment with various aspects of the tool was seen as a significant advantage.

Negative Characteristics

- **Learning Curve:** While the tool was generally not considered unintuitive, some users felt that they needed more time to fully explore and discover all of its capabilities in order to provide more beneficial feedback.
- **UI Design:** Users mentioned issues with the User Interface (UI), including the dropdown not being prominent enough and a lack of a “back” button to switch between sections. Some felt that larger streams or colors dominated the visualizations, potentially overshadowing smaller, but still relevant, topics.
- **Interactions:** Participants noted that making the most of the tool’s interactions might depend on the audience’s familiarity with tools like Plotly. Beginners might find certain interactions, like zooming and filtering, challenging, if not used to double clicking for isolating a topic, or dragging to zoom.
- **Audience Knowledge:** Some users felt that the tool’s usefulness depended on the audience’s prior knowledge of topic modeling. Coherence scores, for example, were considered more useful for those familiar with the concept.
- **Longer-Term Use:** A few participants mentioned that more extended use would be needed to provide more in-depth general feedback.

Suggested Improvements

- **Audience Customization:** One of the suggestions was to separate the two sections to cater to a user's backgrounds and needs. They suggested asking the user a question at the start about their need or level of understanding, and to show the "Topics Over Time" section to a lay user, and the "Inspect Topics and Scores" section to a user more interested in building the models.
- **Dewey Code Chart Enhancements:** Two of the participants suggested further enhancements to the Sankey diagram for Dewey codes. They suggested a drill-down capability that allows the user to view the sub-categories belonging to a single Dewey code, allowing for more insight into the categorization of the documents. One of them remarked that the tooltips on hover for each document were not providing much information, and suggested clicking on a particular topic or Dewey Code to just get an overview of the types of documents present.
- **UI Improvements:** One of the suggestions was to make the section tabs floating for easier access to section selection while scrolling down in the tool. Another participant suggested adding a reset button for the charts to make it more user-friendly and obvious to beginners. Note that such a reset button already existed in the charts, but was not obvious to the user.

7.4 Discussion

7.4.1 General Discussion

Preference for Visual Inspection Over Coherence Scores: We observed that when it comes to selecting a model for word usage change detection, participants rely more on their intuition and visual examination of topics than on statistical coherence scores. This preference for visual inspection is particularly prominent during the initial phase when determining the ideal number of word senses. The reason behind this is that coherence values can be influenced by the number of senses a word has. Since a model's coherence is calculated by averaging the coherence scores of its topics, it can fluctuate significantly if one topic is perceived as less coherent compared to the rest. This phenomenon was evident in the case of `model3`, which had two topics with relatively high coherence scores, contributing to its overall score. However, despite its higher coherence score, participants unanimously rejected `model3` as a suitable model because the number of topics and their discriminative power played a more critical role in their selection. Therefore, the recommendation for employing a metric like coherence score is during the later stages of model selection, once the optimal number of word senses and their meanings have been identified through visual methods. Coherence scores can then be used to refine the model and enhance the interpretability of its topics.

Underutilization of ”Documents and Terms” Section: We noticed that the “Documents and Terms” section was not utilized often by the users to make their decision. One reason could be its placement at the bottom of the section; after already viewing the different topics and models through the intertopic map, Sankey chart, and coherence scores, the Documents and Terms section might not provide more value. Additionally, being a non-visual format might also be a reason why this section is not enticing enough for the users. Converting the data shown in this section to a visual format might prompt the users to inspect more the terms contributing to a topic or read the word being used in an example sentence.

Detecting Subtle Semantic Shifts: Participants also shared their thoughts on the tool’s performance in detecting subtle semantic shifts. Participant P6 commented, “The tool was great at capturing major shifts, but I felt it struggled with more nuanced changes. It missed some subtle shifts that I could identify.” Participant P9 mentioned that having the tool be able to differentiate between a “noun” meaning of the word or a “verb” meaning of the word would also be beneficial for better understanding of the different usages. This feedback highlights some additional scope for such a tool; while differentiating between word usages and contexts is beneficial for the user, additional abilities like parts-of-speech tagging, or classifying more subtle features of the change, like highlighting word stability, narrowing, or broadening could be valuable additions to such a tool.

Clarifying Tool Scope Based on User Intent: The usage of the charts, and the comments from a participant about dividing the tool sections based on intended purpose can be used to make the scope of the tool clearer for the user. If the user is a researcher or modeler keen on finding the best model for their task, they can be directed to the “Inspect Topics and Scores” section, where the ability to change parameters or models and inspect the outputs would be more beneficial. Lay users, or users more interested in the general change of word senses can be directed to the “Topic Usage Over Time” section, while also including some tools in the section that might aid in the comprehension of those senses.

Not using information hovers: Each chart contained an information button next to its title that surfaced information about the corresponding chart and how to read it. We found that this feature was not used by most participants, who preferred to experiment with the tool’s interactions and even asking the researcher for clarity at times, rather than refer to the information provided. While some participants did utilize the feature, the lack of utilization makes it clear that it is either not informative enough or enticing enough for the participant to use.

Dimension Labeling in the Intertopic Map: Many participants inquired about the meaning of the axis labels “D1” and “D2” on the intertopic map. As this map visually condenses a high-dimensional topic space into two dimensions, these labels simply denote the dimension numbers. It’s worth noting that in this representation, the distance between topics holds more significance than their specific placement on the axes. The quadrant-like layout of these axes, intersecting perpendicularly at the chart’s center, prompted some participants to wonder if this configuration implied a quadrant classification system and if topics’ placement within a particular quadrant had any specific meaning.

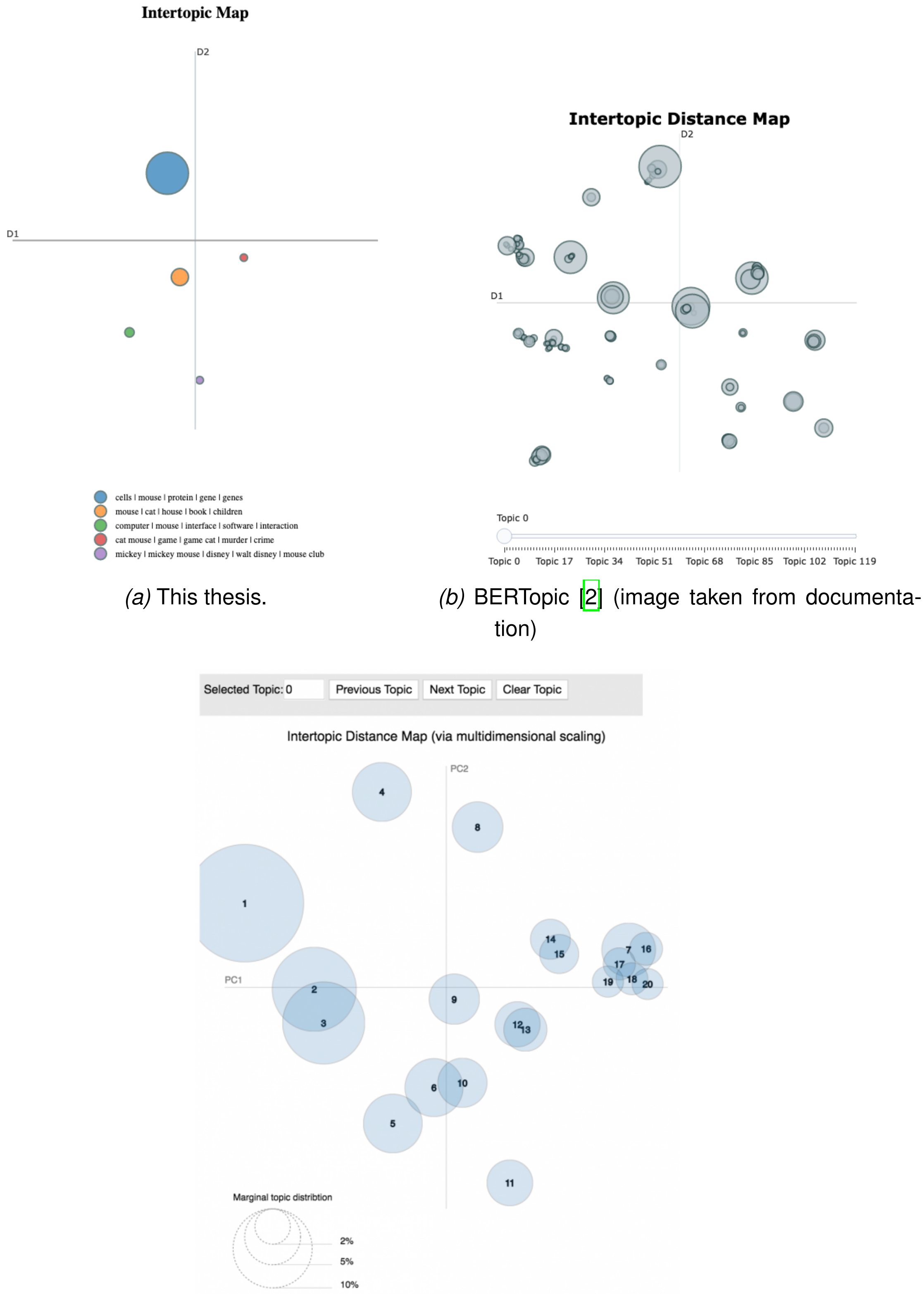
We note that this type of visual is also used both in the BERTopic and LDAvis intertopic distance map, with BERTopic using the “D1” and “D2” notation, while LDAvis uses “PC1” and “PC2”. Figure 7.9 shows an example of the different variants in use. This type of visualization might be informational to topic modeling practitioners as a common method to visualize reduced dimensions, but as observed, for an unfamiliar user it seems to be a point of confusion. An additional point to note is that the explanation about the axes labels was also provided in the chart informational hover but was not read by the participants, a problem that links back to the previous point.

Usage (or lack thereof) of Interactivity Features: The usage of the interactivity options in the charts was mixed. While some users were able to select, isolate, zoom or, in general, use the interactivity of the charts, the others viewed the charts as static and commented, especially in case of the steamgraph, about how the most prominent senses overshadow the smaller ones, making it difficult to track their stream over time. This was exacerbated for target words and models with multiple word senses. These users had to be prompted to use the interactivity features to help alleviate these problems, but the fact that its usage is not as intuitive for the user is a point to address going forward.

7.4.2 Further Tool Optimizations

The evaluation of the semantic shift analysis visualization tool has provided valuable insights into its usability and functionality. While the tool demonstrates promise, several areas for improvement and future development have been identified based on user feedback and observations. These potential enhancements aim to further enhance the user experience, increase the tool’s effectiveness, and address specific usability challenges encountered during the study. This section outlines the key areas and features that are planned for implementation in future iterations of the tool.

Enhanced Visual Representation of Top Terms: Currently, the tool ranks



(a) This thesis.

(b) BERTopic [2] (image taken from documentation)

(c) LDAvis [49] (image taken from documentation)

Figure 7.9: An overview of the variants of the intertopic distance map used.

terms and numerically lists their values, which may not facilitate user interaction effectively (Figure 5.6). To improve user engagement and understanding, an enhanced visual representation of the top terms will be explored. This visual approach aims to make the presentation of top terms more intuitive and actionable, allowing users to quickly grasp the significance of specific terms in the context of semantic shifts.

Additional Dewey Codes Functionality: The tool will incorporate additional drill-down functionality for Dewey codes, as suggested by Participant 9, allowing users to explore sub-Dewey codes associated with each topic. Additionally, a temporal dimension will be introduced to the Dewey codes visualization, enabling users to track how Dewey codes belonging to topics have evolved over time.

Improved Intertopic Maps Presentation: Efforts will be made to present intertopic maps in a more user-friendly manner that avoids the use of dimension names like 'D1' and 'D2,' which users found to be confusing. One potential approach involves removing the axes and labels, testing how users respond to this simplified representation without axes names, while still retaining the features on intertopic distance, number of topics, and topic sizes.

Task-Based Visualization Demarcation: User feedback highlighted the need for task-specific demarcation of visualizations within the tool. Participant 6 (Appendix A) suggested that users could be asked about their specific information needs before presenting visualizations. This would allow tailoring the tool's interface to match the user's goals, ensuring that relevant visualizations are prominently displayed, thus enhancing user experience and efficiency. Additionally, the usability questionnaire indicated that pragmatic values were not as high as hedonic values, suggesting room for improvement in assisting users in achieving results. Future iterations of the tool will focus on tailoring visualizations based on the user's knowledge level or task complexity to. For example, the tool may offer a simplified version for common users and a more advanced section for users familiar with topic modeling concepts.

Increased Interaction Prominence: To mitigate the chance of misinterpretation as seen in Task 6, especially when dominant topics overshadow smaller ones, interaction possibilities within the tool will be made more prominent. Features like zooming, isolating specific topics, and interactive guidance tooltips will be emphasized. These enhancements aim to facilitate a closer examination of visualizations, reducing the likelihood of misinterpretations.

Improved Understanding of Coherence Scores: The tool will explore better ways to convey coherence scores through visualizations, moving beyond displaying raw values. Enhancing the visual representation of coherence scores can assist users in understanding the quality and coherence of topics, providing them with more actionable insights.

User Interface (UI) Enhancements: User feedback will inform several UI changes, including the addition of floating tabs and a back button for seamless navigation between sections. Moreover, elements such as the visualization reset button, information hover icons, and other visualization interactions like zooming and filtering will be made more prominent and user-friendly.

Experimentation with Random Question Ordering: While not a direct change in the tool, it is a change in the user study that can lead to additional insights into user behaviour. In the user study, it was observed that participants tended to use the same charts in Task 3 as they did in Task 2, possibly due to the sequential ordering of questions. To mitigate this potential bias, future studies may experiment with randomizing the order of questions. This approach seeks to measure whether question ordering plays a role in chart selection and participant responses, providing insights into user preferences and decision-making processes.

7.4.3 Tool Applications

The real-world impact of this research extends to addressing a gap in the realm of semantic shift visualization tools. Prior to this study, there was a notable deficiency in the evaluation of usability within semantic shift visualization systems (Section 2.4). By delving into how users perceive changes, their preferences for visualization methods, and the crucial aspects to visualize for conveying desired information, this research starts to bridge that gap in the usability assessment of semantic shift visualization tools.

One immediate practical applications of this research lies in the domain of topic modeling explainability, particularly for comprehending unstructured, voluminous text data. Practitioners and researchers alike can significantly benefit from the interactivity provided by this tool. It empowers them to gain a more profound understanding of their data, thereby aiding in the selection of model parameters aligned with their specific tasks. Moreover, the study can serve as a useful exploratory tool for subsequent downstream tasks like word sense induction or disambiguation, rendering the process more transparent and interpretable through visualizations, which, based on the results, are as important, if not more, than traditional coherence scores. Additionally, enabling statistical validation measures like coherence scores, and allowing to see terms and documents contributing to the creation of the topic enhances the trust in the system for users who want to look beyond the senses generated and the shifts over time.

The tool helps to serve as an intermediary between practitioners who build the tool, and linguistic researchers, who may be interested in the study of semantic shifts but not necessarily well-versed in the technical intricacies of model creation. It empowers them to identify and interpret shifts with the help of a usable interface, thus facilitating more informed decision-making and analysis.

The modularity and focus on output visualization enables future work to use multiple methods of topic generation beyond traditional topic modeling. This addition enriches the evaluation discourse within the semantic shift evaluation field, where most model evaluation is performed accuracy scores of annotated shifted words [70]. This approach assists in the evaluation by allowing users a visual understanding of the different models and their outputs.

Chapter 8

Conclusion

This research journey aimed to develop a user-friendly semantic shift visualization tool while recognizing the growing importance of tracking word meaning changes over time. We began by identifying the significance of this endeavor in language understanding, information retrieval, and natural language processing. To meet this need, we created an innovative tool that combines topic modeling and interactive visuals.

Our study consisted of several key stages. We explored essential features for effective semantic shift visualization by reviewing existing literature, highlighting instruments like coherence scores, intertopic maps, stacked bar charts, and streamgraphs. We also delved into the interplay between quantitative measures and users' intuitive judgments during model selection.

Our primary goal was to assess the tool's effectiveness in facilitating semantic shift exploration and comprehension. We addressed this through research questions and sub-questions. Our secondary aim was to validate the visual exploration of semantic shifts through a user study, a critical step given the limited prior usability assessments of such systems.

The secondary objective was to conduct a user study to validate the visual way of exploring semantic shifts, since previous research of visualizations systems mostly delves into the different types of visualizations and does not necessarily validate the systems, especially from a usability context, measuring how effectively the system helps the user for its intended task.

In the context of semantic shift visualization tools, this study represents one of the first systematic efforts to evaluate them from a usability perspective. While prior research established the groundwork for visualizing semantic shifts, our study advances this by scrutinizing a dedicated tool's effectiveness from a functional and usability standpoint. Through user feedback, task data, and insights into participants' decision-making processes, we gained valuable insights into these tools' practical utility.

Table 8.1: Types of semantic shift visualizations and the features they measure

	word co-occurrence	degree of similarity	continuity	word sense change	concept change	word frequency	word context
Word Graph [15] [30] [20]	✓	✓					✓
Steamgraph [31] [32]			✓	✓	✓	✓	✓
Percentage Stacked Bar Chart [33] [34]			✓	✓		✓	
Network Graph [31]	✓	✓					✓
Radial Bar Chart [38]	✓	✓					✓
Spiral Line Chart [38]	✓		✓				
This thesis		✓	✓	✓	✓	✓	

In summary, this research contributes to the semantic shift analysis field by assessing a visualization system's usability and task support. The interdisciplinary intersection of linguistics, data visualization, and user experience design will continue to shape the future of semantic shift analysis tools. In the following sections, we present our findings, challenges, and prospects for future research, highlighting the broader implications for language analysis and natural language understanding.

8.1 Answer to Research Questions

RQ: How effectively does the developed visualization tool facilitate the exploration and understanding of semantic shifts in word senses?

To address this central research question, we delved into the following sub-questions:

SRQ1: *What are the different features to encode to aid in visualizing semantic shifts?*

To determine the key features necessary for effective semantic shift visualization, we conducted a comprehensive review of existing literature on semantic shift visualization methods. In Table 8.1, we revisit the commonly used chart types in semantic shift visualizations and the specific features they capture (as previously shown in Section 2.5, and included the features from that list that are encoded in this tool. We integrated the features shown in the table into our visualization tool and then conducted a user study to understand how users interacted with these features.

Our study resulted in the development of a comprehensive visualization tool that incorporates various features, including intertopic maps, stacked bar charts, steamgraphs, and coherence scores. Our findings highlight the importance of these visual elements in helping users effectively visualize and understand how word senses evolve over time.

SRQ2: *How does the visualization tool support users in exploring different senses of a word and their evolution over time?*

The results of our study demonstrate that the visualization tool effectively supports users in exploring different senses of a word and their evolution over time. Users overwhelmingly utilized the temporal features of the tool, such as the steamgraph and stacked bar chart, to delve into the nuanced changes in word senses across different time periods. The interactive nature of these visualizations allowed users to zoom in on specific time intervals, isolate particular word senses, and compare their prominence. This interactivity empowered users to gain a comprehensive understanding of how word senses evolved over time.

The steamgraph, in particular, emerged as a favored visualization method for exploring word senses. Users appreciated its ability to visually represent the dominance of different senses and how this dominance shifted over decades. They were able to easily identify and track the most prominent senses, which greatly facilitated their exploration of word sense evolution. The intertopic map, most commonly used for visualizing topic modeling results, was also a popular choice to inspect the senses of a word and find similar senses.

However, it's worth noting that some participants mentioned challenges related to the overshadowing of smaller senses by more prominent ones in the steamgraph, especially for target words with multiple senses. While this highlights the tool's effectiveness in showcasing dominant senses, it also points to the need for further refinements to make smaller senses more discernible, and to make the interactive nature of a system like this more apparent.

SRQ3: *How do users' intuitive choices of topic models align with quantitative measures such as coherence scores, and how does the tool influence their model selection process?*

Our study revealed that participants' intuitive choices of topic models often differed from quantitative measures such as coherence scores. While coherence scores played a role in the model selection process, participants primarily relied on visual inspection and their intuitive judgments to determine the most suitable model. This highlights the importance of a tool that integrates both quantitative and visual elements to aid in model selection.

The tool helped to enhance users' confidence in model selection. The visualizations provided visual evidence to support users' intuitive preferences for models. Participants reported feeling more assured in their selections, as the visualizations provided a clear data representation that reinforced their intuition. This alignment between intuition and visualization not only boosted confidence but also streamlined the decision-making process.

While users often relied on intuition, it's important to note that the tool also encouraged them to consider quantitative metrics, particularly coherence scores, when making model selections. Participants cited coherence scores as useful indicators of model performance, especially when they were uncertain about their intuitive choices. This dual approach, combining intuition and quantitative metrics, highlights the tool's capacity to provide a holistic framework for model selection.

8.2 Limitations

While our study yielded valuable insights, it is essential to acknowledge certain challenges and limitations:

User Familiarity: The usability and effectiveness of the tool might vary based on users' prior knowledge of topic modeling and related concepts. Some participants with prior experience in the field found it easier to navigate and interpret the tool, while beginners faced a steeper learning curve.

Visual Complexity: The tool's effectiveness in visualizing semantic shifts might be hindered by the visual complexity of certain representations, such as intertopic maps. Ensuring that users can easily interpret and interact with these visualizations is an ongoing challenge.

Moderator Bias: The choice of conducting tests in a controlled usability lab with moderators versus online tests without direct interaction can potentially lead to variations in results due to biases. Usability lab tests may be susceptible to biases like experimenter effects [71], social desirability bias [72], [73], and the Hawthorne effect [74]. These biases can influence participants' behaviors and responses, affecting the overall outcomes.

Subjectivity: The evaluation of model performance and the identification of semantic shifts involve subjective judgments. While we provided quantitative metrics like coherence scores, the final decision often relies on users' intuition and preferences.

Limited Interactivity: One challenge of the study was the inability to incorporate the option for the user to analyze a target word in real time, due to the large processing time a word with a large number of documents (greater than 20,000) might take. Participants in the study were constrained in their ability to experiment with their own word choices or a broader range of word choices. This limitation in interactivity could impact user engagement and the tool's applicability to a wider range of scenarios. A more interactive approach allowing users to input and explore their own word choices could offer a fundamentally different level of engagement and utility.

Participant Size and Homogeneity: While our study engaged a diverse group of participants, it is important to acknowledge that the sample size was relatively modest. The challenge lay in recruiting a large number of participants that were able to complete the entrire user study within the time and resource constraints. A larger and more varied participant pool could provide deeper insights into how different user backgrounds and expertise levels influence tool usability and decision-making processes.

Scale of Precomputed Data: It's important to note that the scale of the precomputed data used for visualization, particularly the number of words included, is a limitation. The tool's effectiveness may be influenced by the extent and diversity of the data. A larger and more diverse dataset could potentially provide more comprehensive insights into semantic shifts, and the tool's performance might vary accordingly.

8.3 Future Work

Building upon the findings and recognizing the challenges, there are several avenues for future research and tool enhancement:

Incorporate User Feedback into Tool Enhancements: Future work should consider integrating the feedback and suggestions provided by users during this study into the tool's design and functionality. This iterative approach to tool development can lead to refinements that better align with users' needs and preferences. Additionally, incorporating features that address the specific challenges and preferences voiced by participants, such as enhanced support for detecting subtle semantic shifts or finer-grained analysis of word usages, could enhance the tool's effectiveness and user satisfaction.

Wider Participant Demographics: Expanding the scope of participant demographics is crucial for gaining a more comprehensive understanding of how diverse user backgrounds and expertise levels influence the usability and applicability of the tool. Future studies should aim to include a larger and more diverse participant pool, encompassing individuals from various linguistic, cultural, and professional backgrounds. Conducting user studies with participants who have varying levels of expertise in topics related to natural language processing and semantic analysis can help identify specific user segments for which the tool is particularly well-suited and those that may require additional support or customization.

Real-Time Computation and Parameter Tuning: To enhance the tool's usability and efficiency, future work could focus on enabling real-time model computation and parameter tuning within the tool itself. This would empower users to interactively adjust model parameters and instantly observe the effects on topic modeling outcomes. Real-time computation and parameter tuning can streamline the process of exploring different models, making it more intuitive and user-friendly. Additionally, providing users with immediate feedback on how changes in parameters impact model coherence and topic interpretability can facilitate more informed decision-making during the model selection process.

Multilingual Support: Expanding the tool's capabilities to encompass multiple languages is a promising direction for future development. While the current version of the tool focuses on semantic shift analysis in a English, extending its functionality to include various languages can significantly broaden its utility. Multilingual support can cater to researchers, linguists, and professionals working with diverse linguistic datasets, enabling them to explore semantic shifts in different languages and cross-linguistic contexts.

User-Customized Experiences: Tailoring the tool's interface and functionalities to cater to users' varying levels of expertise could enhance its accessibility and effectiveness. This might involve providing introductory materials for beginners and advanced features for experts.

Integrating Linguistic Resources: Leveraging linguistic resources, such as part-of-speech tagging and word sense disambiguation, can provide more detailed insights into word sense changes and help users differentiate between subtle shifts.

Bibliography

[1] A. Benito, A. G. Losada, R. Therón, A. Dorn, M. Seltmann, and E. Wandl-Vogt, "A spatio-temporal visual analysis tool for historical dictionaries," in *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2016, pp. 985–990.

[2] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[3] N. Tahmasebi, L. Borin, and A. Jatowt, "Survey of computational approaches to lexical semantic change detection," *Computational approaches to semantic change*, vol. 6, p. 1, 2021.

[4] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[5] S. Wang, S. Schlobach, and M. Klein, "Concept drift and how to identify it," *Journal of Web Semantics*, vol. 9, no. 3, pp. 247–265, 2011.

[6] J. R. Firth, "Papers in linguistic analysis 1934–1951," 1957.

[7] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: a survey," *arXiv preprint arXiv:1806.03537*, 2018.

[8] X. Tang, "A state-of-the-art of semantic change computation," *Natural Language Engineering*, vol. 24, no. 5, pp. 649–676, 2018.

[9] F. Almeida and G. Xexéo, "Word embeddings: A survey," *arXiv preprint arXiv:1901.09069*, 2019.

[10] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," *arXiv preprint arXiv:1605.09096*, 2016.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[13] H. Gonen, G. Jawahar, D. Seddah, and Y. Goldberg, “Simple, interpretable and stable method for detecting words with usage change across corpora,” *arXiv preprint arXiv:2112.14330*, 2021.

[14] H. Azarbonyad, M. Dehghani, K. Beelen, A. Arkut, M. Marx, and J. Kamps, “Words are malleable: Computing semantic shifts in political and media discourse,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1509–1518.

[15] W. L. Hamilton, J. Leskovec, and D. Jurafsky, “Cultural shift or linguistic drift? comparing two computational measures of semantic change,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2016. NIH Public Access, 2016, p. 2116.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[19] O. Kellert and M. M. U. Zaman, “Using neural topic models to track context shifts of words: a case study of covid-related terms before and after the lockdown in april 2020,” in *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 2022, pp. 131–139.

[20] D. T. Wijaya and R. Yeniterzi, “Understanding semantic change of words over centuries,” in *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, 2011, pp. 35–40.

[21] B. Chen, Y. Ding, and F. Ma, “Semantic word shifts in a scientific domain,” *Scientometrics*, vol. 117, pp. 211–226, 2018.

[22] J. Gordon, L. Zhu, A. Galstyan, P. Natarajan, and G. Burns, “Modeling concept dependencies in a scientific corpus,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 866–875.

[23] E. Sagi, D. Diermeier, and S. Kaufmann, “Identifying issue frames in text,” *PLoS one*, vol. 8, no. 7, p. e69185, 2013.

[24] F. Bianchi, S. Terragni, and D. Hovy, “Pre-training is a hot topic: Contextualized document embeddings improve topic coherence,” *arXiv preprint arXiv:2004.03974*, 2020.

[25] D. Angelov, “Top2vec: Distributed representations of topics,” *arXiv preprint arXiv:2008.09470*, 2020.

[26] Q. Gao, X. Huang, K. Dong, Z. Liang, and J. Wu, “Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec,” *Scientometrics*, vol. 127, no. 3, pp. 1543–1563, 2022.

[27] C. Collins, G. Penn, and S. Carpendale, “Interactive visualization for computational linguistics,” in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Tutorial Abstracts*, 2008, pp. 6–6.

[28] A. Jatowta, N. Tahmasebib, and L. Borinb, “Computational approaches to lexical semantic change: Visualization systems and novel applications,” *Computational approaches to semantic change*, vol. 6, p. 311, 2021.

[29] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[30] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, “Statistically significant detection of linguistic change,” in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 625–635.

[31] C. Martinez-Ortiz, T. Kenter, M. Wevers, P. Huijnen, J. Verheul, J. Van Eijnatten, M. Düring, A. Jatowt, J. Preiser-Kappeller, A. v. Den Bosch *et al.*, “Design and implementation of shico: Visualising shifting concepts over time,” *HistoInformatics 2016*, vol. 1632, pp. 11–19, 2016.

[32] O. Becher, L. Hollink, and D. Elliott, “Exploring concept representations for concept drift detection.” in *SEMANTiCS Workshops*, 2017.

[33] L. Frermann and M. Lapata, “A bayesian model of diachronic meaning change,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 31–45, 2016.

[34] S. Montariol, M. Martinc, and L. Pivovarova, “Scalable and interpretable semantic change detection,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 4642–4652.

[35] M. Gruppi, S. Adalı, and P.-Y. Chen, “The sense toolkit: A system for visualization and explanation of semantic shift,” in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 283–287.

[36] M. Hilpert and F. Perek, “Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts,” *Linguistics Vanguard*, vol. 1, no. 1, pp. 339–350, 2015.

[37] V. Kulkarni, B. Perozzi, and S. Skiena, “Freshman or fresher? quantifying the geographic variation of internet language,” *arXiv preprint arXiv:1510.06786*, 2015.

[38] R. Kazi, A. Amato, S. Wang, and D. Bucur, “Visualisation methods for diachronic semantic shift,” in *Proceedings of the Third Workshop on Scholarly Document Processing*, 2022, pp. 89–94.

[39] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

[40] M. Allaoui, M. L. Kherfi, and A. Cheriet, “Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study,” in *International conference on image and signal processing*. Springer, 2020, pp. 317–325.

[41] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering.” *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.

[42] T. Joachims *et al.*, “A probabilistic analysis of the rocchio algorithm with tfidf for text categorization,” in *ICML*, vol. 97. Citeseer, 1997, pp. 143–151.

[43] N. Cowan, “The magical number 4 in short-term memory: A reconsideration of mental storage capacity,” *Behavioral and brain sciences*, vol. 24, no. 1, pp. 87–114, 2001.

[44] R. K. Amplayo, S.-w. Hwang, and M. Song, “Autosense model for word sense induction,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6212–6219.

[45] J. Knopp, J. Völker, and S. P. Ponzetto, “Topic modeling for word sense induction,” in *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, 2013. Proceedings*. Springer, 2013, pp. 97–103.

[46] K. Gulordava and M. Baroni, “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*. Edinburgh, UK: Association for Computational Linguistics, Jul. 2011, pp. 67–71. [Online]. Available: https://aclanthology.org/W11-2508

[47] M. Bostock, V. Ogievetsky, and J. Heer, “D^3 data-driven documents,” *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[48] B. Craft and P. Cairns, “Beyond guidelines: what can we learn from the visual information seeking mantra?” in *Ninth International Conference on Information Visualisation (IV’05)*. IEEE, 2005, pp. 110–118.

[49] C. Sievert and K. Shirley, “Ldavis: A method for visualizing and interpreting topics,” in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63–70.

[50] T. Ura, K. Okuda, and M. Shimada, “Developments in viral vector-based vaccines,” *Vaccines*, vol. 2, no. 3, pp. 624–641, 2014.

[51] V. N. Vapnik, “Pattern recognition using generalized portrait method,” *Automation and remote control*, vol. 24, no. 6, pp. 774–780, 1963.

[52] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[53] C. A. Sims, “Macroeconomics and reality,” *Econometrica: journal of the Econometric Society*, pp. 1–48, 1980.

[54] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig *et al.*, “Quantitative analysis of culture

using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[55] S. J. Stratton, "Population research: Convenience sampling strategies," *Prehospital and Disaster Medicine*, vol. 36, no. 4, p. 373–374, 2021.

[56] F. Bakalov, B. König-Ries, T. Hennig, and G. Schade, "Usability study of a semantic user model visualization for social networks," in *Proc. Workshop Visual Interfaces to the Social and Semantic Web (VISSW'11),*, 2011.

[57] C. E. Osgood, "The nature and measurement of meaning." *Psychological bulletin*, vol. 49, no. 3, p. 197, 1952.

[58] M. Hassenzahl, M. Burmester, and F. Koller, "Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität," *Mensch & Computer 2003: Interaktion in Bewegung*, pp. 187–196, 2003.

[59] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*. Springer, 2008, pp. 63–76.

[60] I. Díaz-Oreiro, G. López, L. Quesada, and L. A. Guerrero, "Ux evaluation with standardized questionnaires in ubiquitous computing and ambient intelligence: a systematic literature review," *Advances in Human-Computer Interaction*, vol. 2021, pp. 1–22, 2021.

[61] A. Hodrien and T. Fernando, "A review of post-study and post-task subjective questionnaires to guide assessment of system usability." *Journal of Usability Studies*, vol. 16, no. 3, 2021.

[62] M. Hassenzahl, "The interplay of beauty, goodness, and usability in interactive products," *Human–Computer Interaction*, vol. 19, no. 4, pp. 319–349, 2004.

[63] J. Nielsen, "Why you only need to test with 5 users," 2000.

[64] R. A. Virzi, "Refining the test phase of usability evaluation: How many subjects is enough?" *Human factors*, vol. 34, no. 4, pp. 457–468, 1992.

[65] L. Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," *Behavior Research Methods, Instruments, & Computers*, vol. 35, pp. 379–383, 2003.

[66] R. Alroobaea and P. J. Mayhew, “How many participants are really enough for usability studies?” in *2014 Science and Information Conference*. IEEE, 2014, pp. 48–56.

[67] R. E. Patterson, L. M. Blaha, G. G. Grinstein, K. K. Liggett, D. E. Kaveney, K. C. Sheldon, P. R. Havig, and J. A. Moore, “A human cognition framework for information visualization,” *Computers & Graphics*, vol. 42, pp. 42–58, 2014.

[68] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The measurement of meaning*. University of Illinois press, 1957, no. 47.

[69] J. C. Castro-Alonso, P. Ayres, and J. Sweller, “Instructional visualizations, cognitive load theory, and visuospatial processing,” *Visuospatial Processing for Education in Health and Natural Sciences*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201136862

[70] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi, “Semeval-2020 task 1: Unsupervised lexical semantic change detection,” *arXiv preprint arXiv:2007.11464*, 2020.

[71] R. Rosenthal, “Experimenter effects in behavioral research,” 1976.

[72] M. F. King and G. C. Bruner, “Social desirability bias: A neglected aspect of validity testing,” *Psychology & Marketing*, vol. 17, no. 2, pp. 79–103, 2000.

[73] T. L. White and D. H. McBurney, *Research methods*. Cengage Learning, 2012.

[74] J. Payne and G. Payne, “Key concepts in social research,” *Key concepts in social research*, pp. 1–248, 2004.

Appendix A

Participant Interview Feedback

A.1 Positive Characteristics

Question: “Starting with the positive, were there any aspects of the tool that you found particularly useful or engaging?”

Participant 1

- “The best thing is that it is interactive and you can click through all the models and prepare them as you like.”
- “It is quite intuitive to use.”
- “It is much more difficult to get a feeling for a model if you just have a document with some scores assigned to it.”
- “The ability to play around with it is a big plus.”

Participant 2

- “The debriefing at the start to understand the context of the tool was helpful because I was not very familiar with how the underlying method works.”
- “I was able to quickly get the hang of it. When you asked me a question I immediately knew where to look for the answer.”

Participant 3

- “I really liked the inclusion of the coherence scores because it helped me to objectively select the best model.”

- “I also liked the proportions over time (stacked bar chart) because it helps to see clearly which senses are being used more over time.”

Participant 4

- “It’s very interesting, as a visual way, to be able to see the topics changing over time.”

Participant 5

- “It’s really nice and interactive. The visualizations are nice and let you analyze all the topics.”
- “The representative documents section was nice and helped to understand each topic more.”

Participant 6

- “I can see this being very useful if I was a linguist or if I was working in the library field.”

Participant 7

- “The tool is really useful to see the context and meaning of a word over time.”
- “I can see it being used for linguistic research purposes.”
- “The steamgraph was my favourite visualization to use and the intertopic map would be on spot number 2.”

Participant 8

- “It’s quite an interesting tool seeing how it functions.”
- “All the overviews are clear and easy to use.”

Participant 9

- “It’s a very interesting and intelligent tool. It could replace laboriously created categorizations like Dewey codes in the future.”

Participant 10

- "It takes some getting used to but it is nice to work with."

A.2 Negative Characteristics

Question: "Were there any negative aspects of the tool that you found confusing or difficult to interpret?"

Participant 1

- "I don't know now. I would have to work with it a bit more and there might be some details which can improve the interface a little bit more."

Participant 2

- "Although I did receive debriefing it would help to have more information too because I was not very familiar with the underlying method of creating the topics."
- "It was not necessarily an un-intuitive chart, but I noticed I did not really look at the stacked bar chart to complete any of the tasks. It felt like I could get all the same information from the steamgraph."

Participant 3

- "The main visualizations were good, but some of the UI components were not intuitive. I expected a back button to go back to the previous section but I had to scroll up and click on the tabs."

Participant 4

- "This (Dewey Code visualization) had a large proportion (for the main topic) and mainly caught my eye. The other topics were quite small in comparison and I did notice them as much as they were supposed to."

Participant 5

- "Someone who is not very familiar with Plotly would have a harder time using it when it comes to the interactions like zooming and filtering."

Participant 6

- “It depends a lot on the audience’s prior understanding of topic modeling to make full use of the tool.”

Participant 7

- “More on the usability side. This (the Sankey chart nodes) is not really telling me much information. The ’incoming flow count’ and ’outgoing flow count’ (which can be seen when hovering over the nodes).”
- “The coherence scores section would probably be more useful for people who are more familiar with the concept.”

Participant 8

- “It would be good to have floating tabs so that they’re always visible on the top, otherwise I have to keep scrolling to go to another section.”
- “For me the colours were very close to each other on the spectrum so it would be nicer to have more contrast.”

Participant 9

- “It’s worth considering that it identified bank robbery and bank institutions as separate contexts even though they are both about the same concept, a bank as a financial institution. Then again, that depends on the viewpoint a user would want to see, considering its level of granularity.”

Participant 10

- “I would have to work with it professionally for a week to be able to give some better feedback, but can’t say anything negative about it for now.”

A.3 Suggested Improvements

Question: “Based on your experience with the visualization tool, do you have any recommendations for enhancing its usability, clarity, or effectiveness in conveying information?”

Participant 1

- “Some of the instructions could be clearer through the UI. If you play around with the tool a bit you understand which controls affect which visualizations, but things like adding a box around the control and a UI can make it clearer.”

Participant 2

- “Not sure whether there would be any.”

Participant 3

- “I would like a quick reset button that resets the chart to its original state after using interactive features like zooming and filtering.”

Participant 4

- “I was confused when different topics had common words between their labels. So it would be nice to have some explanation for that.”

Participant 5

- “It would be useful to search for some words of interest for yourself.”

Participant 6

- “There could be some kind of question at the start to test the audience’s knowledge of the subject. If the user is not too familiar, they can be shown only one or two visualizations instead of all of them.”
- “You could also ask the user beforehand what they want to know. If they only want to see usage change over time, they can be shown the Usage Over Time section, and if they are more interested in model selection, they can be shown the Inspect Topics section.”

Participant 7

- “The tool might be more useful if you take more documents from more time frames into account.”
- “Maybe having the option to click on each Dewey code and getting a summary of the documents that are part of each category would be more beneficial than showing each document on hover.”

Participant 8

- “I don’t really have anything else to add. Everything else was quite clear.”

Participant 9

- “It would be nice to have the option to drill-down into the sub-categories of a Dewey Code.”

Participant 10

- “I would need to use it more to provide some suggestions.”