Towards Formalizing Concept Drift and Its Variants: A Case Study Using Past COSIT Proceedings

Meilin Shi[1]
Department of Geography and Regional Research, University of Vienna, Austria

Krzysztof Janowicz
Department of Geography and Regional Research, University of Vienna, Austria

Zilong Liu
Department of Geography and Regional Research, University of Vienna, Austria

Kitty Currier
Department of Geography, University of California, Santa Barbara, CA, USA

Abstract

In the classic *Philosophical Investigations*, Ludwig Wittgenstein suggests that the meaning of words is rooted in their use in ordinary language, challenging the idea of fixed rules determining the meaning of words. Likewise, we believe that the meaning of keywords and concepts in academic papers is shaped by their usage within the articles and evolves as research progresses. For example, the terms *natural hazards* and *natural disasters* were once used interchangeably, but this is rarely the case today. When searching for archived documents, such as those related to disaster relief, choosing the appropriate keyword is crucial and requires a deeper understanding of the historical context. To improve interoperability and promote reusability from a Research Data Management (RDM) perspective, we examine the dynamic nature of concepts, providing formal definitions of *concept drift* and its variants. By employing a case study of past COSIT (Conference on Spatial Information Theory) proceedings to support these definitions, we argue that a quantitative formalization can help systematically detect subsequent changes and enhance the overall interpretation of concepts.

2012 ACM Subject Classification Information systems → Digital libraries and archives; Information systems → Similarity measures; Computing methodologies → Information extraction

Keywords and phrases Concept Drift, Semantic Aging, Research Data Management

Digital Object Identifier 10.4230/LIPIcs.COSIT.2024.23

Category Short Paper

Supplementary Material *Software (Source Code and Data)*: `https://github.com/meilinshi/Concept-Drift-Formalization-COSIT-Case-Study`

1 Introduction

With the exponential growth of data from various domains and sources, the need for FAIR (Findable, Accessible, Interoperable, and Reusable) data [19] has become increasingly important for Research Data Management (RDM). For instance, the reuse of data is one of the pillars of data science. By embracing the FAIR principles, data become easier to find and access, and they become more interoperable across domains, thereby promoting reproducibility, reusability, and transparency in academic research. Creating semantically rich digital asset management systems that store data in a human- and machine-readable manner is among the most promising directions for implementing these FAIR principles.

[1] Corresponding author

However, the keywords used to semantically describe the data or research topics evolve, which is also known as the challenge of *semantic aging* for digital records preservation [12]. This can be due to the dynamic nature of language, cultural or political changes, or advances in science and technology. For example, the term *climate change* has become more frequently used than *global warming* over time because of the growing awareness that the concern is not just an increase in Earth's temperature but a broader range of changes in the climate system. Moreover, a spatial component can be involved, as the adoption of terms is likely to occur at different times across regions. From an RDM perspective, limited knowledge of semantics in the past will hinder information retrieval and understanding of archival data. Schlieder argued that a time span of 100 years constitutes a proper temporal frame within which to address semantic aging [12]. However, the vast and continuously expanding volume of data, together with new data formats, multimedia types, and tools available today, may accelerate the rate of semantic aging.

Previous work has approached semantic aging in historical records by identifying temporal counterparts as a way to establish connections and bridge different time periods, e.g., `Walkman` is considered to be a counterpart of `iPod` over time [2, 21]. However, if we approach this issue by tracing the evolution of `Walkman` to `iPod`, there is no need to search for its counterpart. This leads to the study of *concept drift*, a phenomenon in which the meaning or interpretation of concepts changes over time, possibly also across space. By examining concept drift, we can achieve a more comprehensive understanding of the concepts than by searching for counterparts one at a time within a given time frame. Hence, research data management systems would benefit from measures to alert their users to changes in conventional terminology. Failure to address such changes may lead to data reuse and semantic interoperability issues [7].

Prior investigation into concept drift [17] identified that a concept can undergo a split, resulting in two or more distinct concepts over time. However, other forms of change, such as merging, have been overlooked. Furthermore, various temporal characterizations of concept drift, including sudden, gradual, incremental, and reoccurring changes [9], have not been addressed in previous formalization attempts [17, 4, 3]. This work aims to fill these gaps by proposing a formal definition, incorporating temporal scales, of concept drift and its related variants, using a case study for illustration.

The remainder of this paper is structured as follows. Section 2 provides related work on the various notions of concept drift used in different domains, as well as different approaches to detect it. Next, in Section 3, we conduct a case study using keywords from past COSIT (Conference on Spatial Information Theory) proceedings. Section 4 discusses and categorizes our findings to formulate formal definitions. Finally, we conclude our work in Section 5, and provide directions for future work.

2 Related Work

Concept drift has become a research focus in many domains, including linguistics, history, machine learning, and the Semantic Web community. There are many notions describing similar phenomena, such as *semantic drift, concept change, semantic change*, etc. [4, 13]. In linguistics, *semantic drift* or *semantic change* refers to changes in the meaning of words or phrases over time [6]. However, even *concept drift* itself holds different meanings across research fields. In machine learning, it refers to the problem of models becoming less accurate on prediction tasks in unforeseen ways over time [18].

In this work, concept drift, aligned with research in the Semantic Web community, is defined as a change in the meaning of a concept over time, possibly also across locations or cultures [17]. According to Wang et al., [17], a concept is formed by its label, intension, and extension. Here, following the *classical theory of concepts* [14], the *intension* can be seen as the TBox for class properties, while the *extension* serves as the ABox for class instances, i.e., the set of cases successfully categorized under the given class. Concept drift can be understood as the evolution of an ontology (i.e., the set of statements that defines the terminology used) at the TBox level. To give a concrete example of concept drift, Wang et al., [17] used the concept of `European Union (EU)`, which has evolved from an economic cooperation initiative to a political union over the years, with multiple name changes such as the `European Economic Community` and `European Community`. The number of EU member states has also fluctuated due to countries joining and leaving. These can be seen as changes in its intension, label, and extension, respectively.

To detect concept drift, approaches are not limited to examining ontology versioning or relying on the hierarchies (e.g., subclasses and superclasses [3]) within an ontology. Early work by Raubal [11] explored the representation of concepts over time through the movement of conceptual spaces along space-time paths in a semantic space. Other efforts include the AdvoCate system [5], which models the evolution of concepts and categories, demonstrated through a use case capturing changes in land cover classification taxonomies. Tietz et al. [15] investigated the evolution of concepts and presented challenges from a Natural Language Processing (NLP) perspective, conducting a preliminary analysis of recipe data extracted from newspapers. The meaning of concepts in natural language can be captured via latent word representations [10]. Another approach involves constructing a time series using word embeddings to capture the evolution of words over time [8]. Considering the overarching goal of addressing semantic aging in RDM, this work adopts a natural language-based approach to identify concept drift, as no existing ontology covers all concepts of interest.

3 Case Study

Before providing formal definitions, we conduct a case study to identify concept drift and its related variants by examining COSIT proceedings[2] from four time points, 1992, 2001, 2011, and 2022. Keywords are extracted from these proceedings, sourced either from authors (for 2011 and 2022) or generated by machine (for 1992 and 2001), as provided by the publisher Springer Link[3]. Table 1 provides a statistical summary of the dataset.

Table 1 Number of papers and keywords in COSIT proceedings from 1992, 2001, 2011, and 2022.

Year	Number of Papers	Number of Keywords
1992	25	122
2001	33	171
2011	23	107
2022	29	147
Total	110	547

To compute embeddings for keyword similarity measurement, we use SciBERT [1], a pre-trained language model tailored for scientific text. Drawing inspiration from Wittgenstein's *meaning as use* theory [20], we use contextual information to assess similarity in usage.

[2] `https://geosensor.net/cositseries/past-proceedings`
[3] According to Springer Link, this process is still experimental, and the keywords may be updated as the learning algorithm improves.

We compute two types of keyword similarity: (1) textual similarity, represented by the cosine similarity between the embeddings of the two keywords, and (2) contextual similarity, measured as the cosine similarity between two contextual embeddings generated from the associated paper titles and abstracts. All keywords are converted to lowercase. For identical keywords within the same year, only one contextual embedding is generated from their aggregated titles and abstracts. We assign weights α as 0.2 and β as 0.8 to emphasize contextual similarity when aggregating textual and contextual similarities. The overall similarity between Concept C_i and C_j at time t can be expressed as:

$$Sim(C_i, C_j) = \alpha \cdot text\_sim(C_i, C_j) + \beta \cdot context\_sim(C_i, C_j) \tag{1}$$

Figure 1 shows the distribution of top keywords that appeared more than five times in total and their frequency by year. Among these, `Spatial Reasoning` and `Spatial Cognition` are the two most frequently occurring keywords, appearing 11 and 8 times, respectively. Pairwise similarities of keywords are computed between `Spatial Cognition` in 1992 and all other keywords in 2001, 2011, and 2022. For simplicity in illustration, only the most similar keyword and/or the keyword itself are documented in Figure 2a. Figure 2b documents keyword similarity between `Spatial Reasoning` in 1992 and other keywords in 2001, as well as pairwise keyword similarity for `Qualitative Spatial Reasoning` from 2001 onwards.

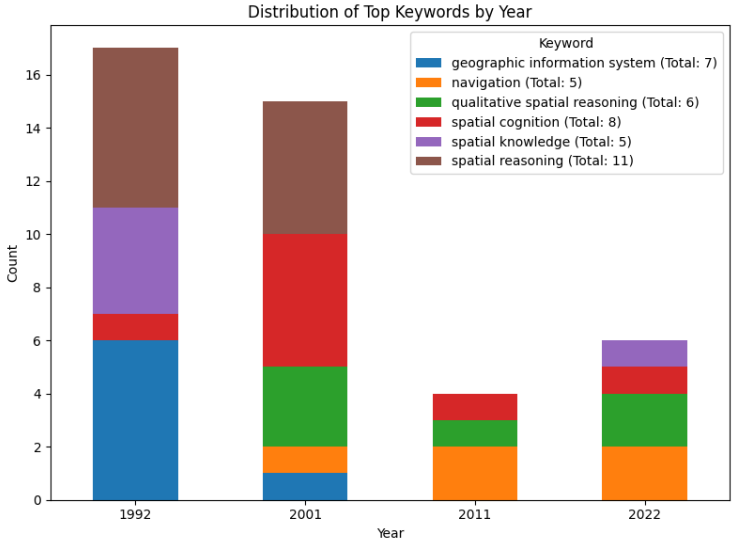
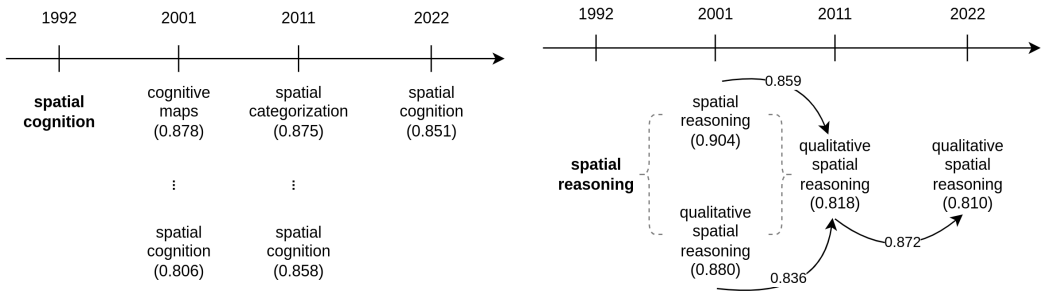


Figure 1 Distribution and frequency of top keywords by year.

4 Towards Formal Definitions

Interestingly, while concept change has been studied in linguistics, ontology engineering, and machine learning, most have done so by providing ad-hoc definitions [9, 16], making comparisons between papers difficult. Here, we will introduce more formal notions of these changes, aiming to serve as the basis for future studies and increase the reproducibility of research on semantic aging in general. Finally, it is worth mentioning that the definitions presented below are meant as *indicators* of change rather than measures. For instance, even if we identify a likely drift (or other change) according to our formal definition, we cannot



(a) Keyword Similarity for `Spatial Cognition`. **(b)** Keyword Similarity for `Spatial Reasoning`.

Figure 2 Evolution of keyword similarity for `Spatial Cognition` and `Spatial Reasoning`. Similarities in parentheses are computed between the 1992 keyword and the keywords from other years. Dashed curly brackets indicate potential split and merge relations.

readily imply that this is not caused by stochastic fluctuations or simply the absence of submissions on certain keywords in the dataset we used. However, without formal definitions, we would not be able to state what we want to statistically test in the first place.

4.1 Concept Drift

▶ **Definition 1.** *A concept C_i, existing at both time t and $t+n$ ($n \geq 1$), undergoes a concept drift from time t to $t+n$, if and only if the most similar concept to C_i at time t among all concepts C_j at time $t+n$ is not C_i itself*[4]*. Formally:*

$$concept\_drift_{t,t+n}(C_i) \iff \underset{j}{argmax}\ sim(C_i^t, C_j^{t+n}) \neq i,\ n = 1, 2, ... \tag{2}$$

In Figure 2a, we can see that the keyword `Spatial Cognition` exists in all four time points. According to our definition, it undergoes a concept drift from 1992 to 2011, while by 2022, it reverts to being most self-similar. When concept drift only exists between t and $t+1$, it can be seen as a sudden drift. However, if this drift persists between t and $t+2$ and onwards, and subsequent similarities such as $sim(C_i^t, C_j^{t+2}) < sim(C_i^t, C_j^{t+1})$ occur, it can be seen as an incremental drift. If we extend the time axis for a longer time frame, we can explore whether the pattern of concept similarity oscillates, indicating a reoccurring drift.

4.2 Concept Shift

▶ **Definition 2.** *A concept C_i undergoes a concept shift to concept C_k from time t to $t+1$, if and only if concept C_i does not exist at time $t+1$, and the most similar concept to C_i at time t among all concepts C_j at time $t+1$ is C_k*[5]*. Formally:*

$$concept\_shift_{t,t+1}(C_i) \iff \nexists C_i^{t+1} \text{ and } \underset{j}{argmax}\ sim(C_i^t, C_j^{t+1}) = k \tag{3}$$

In Figure 2b, we observe that `Spatial Reasoning` only exists in 1992 and 2001. However, from 2001 to 2011, the keyword disappears and becomes most similar to `Qualitative Spatial Reasoning`. Following our definition, this indicates that `Spatial Reasoning` undergoes a concept shift from 2001 to 2011. Upon further investigation, we discover that

[4] The intuition here is that assuming C_i is not drifting, then at $t+1$, it should be more self-similar than similar to another concept C_j.
[5] The intuition here is that if C_i no longer exists, it shifts to its most similar concept C_j at $t+1$.

`Qualitative Spatial Reasoning` also exists in 2001 and 2022. Hence, we hypothesize that a concept split and subsequent merge may have occurred between 1992 and 2011, as denoted by the dashed curly brackets.

4.3 Concept Split and Merge

▶ **Definition 3.** *A concept C_i undergoes a concept split to two (or more) concepts $\{C_i, C_j, ...\}$ from time t to $t+1$, if and only if the similarity of $\{C_i^{t+1}, C_j^{t+1}, ...\}$ to C_i^t satisfies a threshold θ_1 for overall similarity and a threshold θ_2 for textual similarity*[6]*. Formally:*

$$concept\_split_{t,t+1}(C_i) \iff Sim(C_i^t, C_{i,j,...}^{t+1}) > \theta_1 \text{ and } text\_sim(C_i^t, C_{i,j,...}^{t+1}) > \theta_2 \quad (4)$$

▶ **Definition 4.** *Two (or more) concepts $\{C_i, C_j, ...\}$ undergo a concept merge into concept C_i from time $t-1$ to t, if the similarity of $\{C_i^{t-1}, C_j^{t-1}, ...\}$ to C_i^t satisfies a threshold θ_1 for overall similarity and a threshold θ_2 for textual similarity. Formally:*

$$concept\_merge_{t-1,t}(\{C_i, C_j, ...\}) \iff Sim(C_{i,j,...}^{t-1}, C_i^t) > \theta_1 \text{ and } text\_sim(C_{i,j,...}^{t-1}, C_i^t) > \theta_2 \quad (5)$$

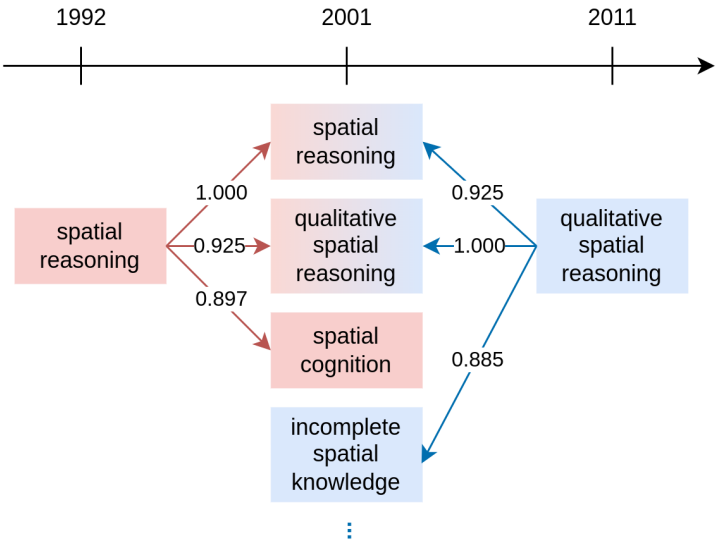


Figure 3 Workflow of identifying Concept Split and Merge. Keywords in 2001 with higher than threshold overall similarities with `Spatial Reasoning` in 1992 are highlighted in red, and those associated with `Qualitative Spatial Reasoning` in 2011 are highlighted in blue. The ones highlighted with a color gradient are associated with both. For simplicity in illustration, not all keywords that meet the threshold are shown. The values indicated by arrows represent the textual similarities between keywords.

Figure 3 shows the result when the thresholds θ_1 and θ_2 are set to be 95% and 90% of their respective maximum similarity values. Keywords highlighted in red are associated with an overall similarity exceeding θ_1 set at 0.859 (95% of the maximum similarity) to

[6] The intuition behind concept split and merge is under the assumption that these processes only occur among keywords with maximal similarity and should exhibit maximal similarity in labeling.

`Spatial Reasoning` in 1992, and those in blue are associated with `Qualitative Spatial Reasoning` in 2011. Textual similarities are indicated by arrows, with a maximum similarity value of 1.000. Therefore, θ_2 is set at 0.900 (90% of the maximum similarity). Based on our definitions, only `Spatial Reasoning` and `Qualitative Spatial Reasoning` (highlighted with a color gradient) meet the criteria for concept split and merge, among all 2011 keywords.

Therefore, we conclude that `Spatial Reasoning` undergoes a concept split to `Spatial Reasoning` and `Qualitative Spatial Reasoning` from 1992 to 2001, followed by a merge into `Qualitative Spatial Reasoning` from 2001 to 2011.

5 Conclusions

Due to semantic aging, keywords that are used to describe research topics evolve, hindering their potential reusability and interoperability. Identifying concept drift over time has the potential to mitigate this issue and facilitate searching within longitudinal archival documents. In this work, we investigate concept drift and its variants by examining the dynamic nature of keywords used in academic literature. We provide formal definitions to indicate potential *concept drift*, *concept shift*, *concept split*, and *concept merge*. A case study using past COSIT proceedings is adopted to illustrate our definitions in practical scenarios. This work highlights the importance of understanding semantic aging, and by detecting concept drift, we hope to enhance reusability and interoperability in Research Data Management. Furthermore, semantic aging may occur at varying rates across regions or cultures, suggesting that there is also a spatial dimension in concept drift. While our formalization provides a preliminary framework to indicate changes, further investigation is needed to quantify them into numerical measures. Moving forward, additional related variants, such as *concept radiation*, and more temporal patterns, such as reoccurring changes, can be explored using larger datasets with longer time spans.

References

1 Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*, 2019. `doi:10.18653/v1/D19-1371`.

2 Klaus Berberich, Srikanta J. Bedathur, Mauro Sozio, and Gerhard Weikum. Bridging the terminology gap in web archive search. In *International Workshop on the Web and Databases*, 2009. URL: `https://api.semanticscholar.org/CorpusID:6709650`.

3 Giuseppe Capobianco, Danilo Cavaliere, and Sabrina Senatore. Ontodrift: a semantic drift gauge for ontology evolution monitoring. In *CEUR Workshop Proceedings*, 2020. URL: `https://api.semanticscholar.org/CorpusID:233432249`.

4 Antske Fokkens, Serge ter Braake, Isa Maks, and Davide Ceolin. On the semantics of concept drift: Towards formal definitions of concept drift and semantic change. In *Drift-a-LOD@EKAW*, 2016. URL: `https://ceur-ws.org/Vol-1799/Drift-a-LOD2016_paper_2.pdf`.

5 Prashant Gupta and Mark Gahegan. Categories are in flux, but their computational representations are fixed: That's a problem. *Transactions in GIS*, 24:291–314, 2020. `doi:10.1111/tgis.12602`.

6 William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *ArXiv*, abs/1605.09096, 2016. URL: `https://api.semanticscholar.org/CorpusID:5480561`.

7 Francis Harvey, Werner Kuhn, Hardy Pundt, Yaser Bishr, and Catharina Riedemann. Semantic interoperability: A central issue for sharing geographic information. *The annals of regional science*, 33:213–232, 1999. `doi:10.1007/s001680050102`.

8 Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web*, pages 625–635, 2015. doi:10.1145/2736277.2741627.

9 Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. doi:10.1109/TKDE.2018.2876857.

10 Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013. URL: https://api.semanticscholar.org/CorpusID:5959482.

11 Martin Raubal. Representing concepts in time. In *Spatial Cognition VI. Learning, Reasoning, and Talking about Space: International Conference Spatial Cognition 2008, Freiburg, Germany, September 15-19, 2008. Proceedings 6*, pages 328–343. Springer, 2008. doi:10.1007/978-3-540-87601-4_24.

12 Christoph Schlieder. Digital heritage: Semantic challenges of long-term preservation. *Semantic Web*, 1(1-2):143–147, 2010. doi:10.3233/SW-2010-0013.

13 Thanos G Stavropoulos, Stelios Andreadis, Efstratios Kontopoulos, and Ioannis Kompatsiaris. Semadrift: A hybrid method and visual tools to measure semantic drift in ontologies. *Journal of Web Semantics*, 54:87–106, 2019. doi:10.1016/j.websem.2018.05.001.

14 Wolfgang G Stock. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, 61(10):1951–1969, 2010. doi:10.1002/asi.21382.

15 Tabea Tietz, Mehwish Alam, Harald Sack, and Marieke van Erp. Challenges of knowledge graph evolution from an nlp perspective. In *WHiSe@ ESWC*, pages 71–76, 2020. URL: https://ceur-ws.org/Vol-2695/paper8.pdf.

16 Stella Verkijk, Ritten Roothaert, Romana Pernisch, and Stefan Schlobach. Do you catch my drift? on the usage of embedding methods to measure concept shift in knowledge graphs. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 70–74, 2023. doi:10.1145/3587259.3627555.

17 Shenghui Wang, Stefan Schlobach, and Michel Klein. Concept drift and how to identify it. *Journal of Web Semantics*, 9(3):247–265, 2011. doi:10.1016/j.websem.2011.05.003.

18 Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996. doi:10.1007/BF00116900.

19 Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016. doi:10.1038/sdata.2016.18.

20 Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953.

21 Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, 2016. doi:10.1109/TKDE.2016.2591008.