Digital Hermeneutics in the History of Concepts.

The Tool *Sense Clustering over Time* (SCoT): Application, Workflow, and Methodological Questions

Alexander Friedrich[1] [0000-0002-2807-7638], Saba Anwar[2] [0000-0002-8344-5165] and Chris Biemann[2] [0000-0002-8449-9624]

[1]Institute of Philosophy, Technical University of Darmstadt, Germany
[2]Department of Informatics, University of Hamburg, Germany

Abstract In recent years, the field of historical semantics has experienced notable advancements. Despite the emergence of promising methodologies, a standardized procedure has not yet been established in the research of the history of concepts. A significant methodological challenge is the seamless and transparent integration of analyzing large datasets (distant reading) with the traditional workflows of conceptual historical research (close reading). Additionally, the effective capture of various word senses (polysemy) and tracking their change over time using computational methods remains a complex task. This article introduces the tool Sense Clustering over Time (SCoT), specifically designed to address these challenges in conceptual historical research. SCoT employs the method of Word Sense Induction (WSI) to facilitate the semi-automatic detection and visual representation of the historical semantics of conceptual words into sense clusters. Such sense clusters enable the reconstruction of different word senses. Through an open-access web interface, users can: (1) analyze the diachronic development of sense clusters within extensive text corpora, (2) explore the linguistic contexts driving these changes, and (3) identify and compile relevant references in the text corpus for further study. Consequently, SCoT empowers researchers to perform text mining and conduct scalable analyses of large historical corpora, significantly enhancing research in the historical semantics and conceptual history.

Keywords: computational humanities, digital hermeneutics, history of concepts, language technology, sense induction, graph clustering, visual analysis

1 Introduction

The increasing availability of large historical text corpora has opened up new and far-reaching possibilities for the research into the history of concepts. As a field of study that is interested in concepts of cultural, political or scientific significance – e.g. *power*, *liberty*, *life, matter, time*, *energy*, *justice* etc. – the history of concepts focuses on the changes in their meaning and use over time in different, often interconnected, contexts and discourses. Analyzing the meaning of such concepts and their change helps to shed light on the constitution and development of larger historical or social contexts or enables an appropriate understanding of these contexts in the first place.

Pioneering for conceptual history as an interdisciplinary discipline and method of research in the humanities were, among others, the "Historical Dictionary of Philosophy" (1971–2007) edited by Joachim Ritter, Karlfried Gründer and Gottfried Gabriel, or the "Historical Dictionary of Political-Social Language in Germany" (1972–1997) edited by Otto Brunner, Werner Conze and Reinhart Koselleck. The internationalization and a more cultural studies-oriented reflection of approaches to the history of concepts led to an ongoing critical methodological discussion [35, 53]. Although Koselleck's *Begriffsgeschichte* explicitly distances itself "from the approaches of modern linguistics, especially structuralist linguistics", it has nonetheless "prepared the ground for a future approach" [41].

Advances in computational and corpus linguistics in recent decades have given rise to the possibility of combining the methods of conceptual history and linguistics in a new way. In this context, the increasing digital availability of possible sources not only offers new possibilities and opportunities, but also new methodological challenges that are associated with a digital history of concepts and their (inter)relation to a digital history of ideas [16, 24, 59]. One of these challenges relates to the question how statistically aggregated data and quantitative analyses can be meaningfully and insightfully connected to qualitative inquiries and the disciplinary conventions and methods of scholars in conceptual history. In the field of digital humanities, this issue is articulated as the question of the relationship between *close reading* and *distant reading* respectively as *scalable reading* [14, 40, 50, 56, 65]. The importance of this question for a digital history of concepts increases with the degree of the methods it employs, and the size of the corpora being analyzed. The more sophisticated methods and tools for the digital analysis of large text corpora become, the more challenging it will be to relate quantitative findings to specific passages or concrete historical speech acts, which a conceptual history study must still be based on.

In this article, we will explore the current state of research on approaches and methods in digital conceptual history, highlighting their respective advantages and disadvantages (Section 2). Based on this analysis, we propose a set of features that digital research tools for conceptual history should ideally possess (Section 3). We will then present the Word Sense Induction (WSI) based online tool "Sense Clustering over Time" (SCoT) and demonstrate its application and workflow using a recent example from research in the history of concepts (Section 4). Finally, we will discuss some of the technological, methodological and epistemological challenges associated with this tool, including its conceptualization through a general model for digital hermeneutics (Section 5).

2 State of Research: Methods for the Digital History of Concepts

To date, there is no standardized method for a digital history of concepts but rather a multitude of approaches. Seven fundamental approaches can be identified that have been used for digital studies of historical semantics, often in combination or as basic elements for each other: (a) *keyword query,* (b) *frequency analysis*, (c) *co-occurrence*

or *collocation analysis*, (d) *topic modelling,* (e) *semantic tagging,* (f) *word embeddings,* and (g) *word sense induction*. All these methods and their different utilizations have certain advantages and disadvantages for a history of concepts.

(a) *Keyword query* is the most basic method of conducting historical text research with digital means. Despite its simplicity, it remains a powerful and widely used method, e.g. for identifying the first occurrences of specific word usages in historical contexts. Tools featuring Keyword in Context (KWIC) functionality enable further exploration of the specific usage of particular word forms [23, 45]. The practice of keyword querying is a fundamental tool for digital historical research, especially for the inspection and interpretation of specific text passages. This method is particularly useful when researchers already know exactly what they are looking for. In this form, keyword query will remain an indispensable tool for digital hermeneutics. However, the effectiveness of sole keyword queries diminishes with the size of the text corpus. Especially, when conducting diachronic studies of large corpora, more advanced methodologies are necessary. Advanced tools for historical linguistic research facilitate more sophisticated search queries and comprehensive linguistic analyses of text corpora by enabling targeted searches for specific word forms or to narrow their focus to particular syntactic relations [23, 58]. These functionalities are based on additional methods of digital text analysis, such as frequency or collocation analysis.

(b) A *frequency analysis* of word occurrences in relevant text corpora, especially when tracked over time, can provide useful indications of certain conceptual or discourse conjunctures. An advantage of this method is its low-threshold application, as it does not require any major computer linguistic or statistical expertise and it enables relatively quick results. Tools like Google Ngram Viewer [47] or Digitales Wörterbuch der deutschen Sprache [27] have therefore become popular in the community. A great disadvantage is that they alone do not provide information about differences and changes in the meanings and contexts of a word. Thus, a query for 'bank' would sum up occurrences of financial institutes, furniture, riversides and much more. And the observation of a strong change in the frequency of a search term often leads to the conclusion that this change indicates an increase or loss of relevance of the associated concepts(s). While the latter is a flawed interpretative short-cut, the former is a technical limitation that in turn favors the latter. Silke Schwandt [60] for example, shows on the basis of texts by Augustine and John of Salisbury that the decreasing frequency of the word *deus* does not indicate a loss of meaning, but rather its undisputed significance. Jani Marjanen has pointed out that four different historical phenomena may contribute to historical frequency changes: "They are:

- Topicality: a word becomes very topical in a given moment
- Expansion: a word enters new domains in language
- Polysemy: a word is associated with new meaning
- Idiomatization: a word is associated with larger linguistic structures, such as idoms (which also carry their own meaning)" [46].

In addition, other circumstances, such as censorship in the context of the source language or an imbalanced composition in the source material, may contribute to changes

in word frequencies. Every solid frequency analysis must either uncover the main reasons for such changes or account for the corresponding uncertainty in its interpretation. Therefore, digital hermeneutics always requires an expanded form of source criticism, both in a qualitative and quantitative sense. A best practice here includes a systematic comparison of the relative word frequency with the absolute word frequency. Additionally, more comprehensive analytical methods are necessary to investigate the underlying causes of statistical findings, such as determining whether an increase in frequency is attributable to the emergence of a new meaning.

(c) *Co-occurrence* or *collocation analysis* makes it possible to determine different contexts of word occurrences, even independently of prior knowledge, and to map their changes diachronically. Depending on the design of the user interface and access options to usable corpora, this method is also relatively accessible for users with no or not much expertise in computational linguistics, as the representation of left and right neighbors of a word and the counting of their frequency can be related to non-computational way of reading texts, i.e. scholarly reading, without detailed expertise in statistics or data science, although elaborate statistical measures and data analysis functions may have been deployed in the backend. Advanced tools for the diachronic analysis of typical word combinations, such as DiaCollo [23, 36, 37], are therefore becoming increasingly popular in the community. A significant limitation of this method is its inability to distinguish differences and changes in the meanings of search terms. Accordingly, typical word combinations are accumulated, necessitating researchers to discern different contexts by relying on their prior knowledge. Based on these observations, researchers must infer or examine whether a change in contexts signifies the emergence of a new sense of a term. Without additional methods such an examination must rely on somewhat arbitrary sampling. A second disadvantage to consider relates to the potential drawbacks accompanying its advantages. The easy accessibility of diachronic co-occurrence visualizations, as provided by tools such as DiaCollo, offers researchers a user-friendly entry point to advanced statistical analysis tools. However, this accessibility may inadvertently lead to the neglect of the correct application and comparison of different measures, weightings, ratios, scores, and functions provided by these tools in favor of the most plausible-looking graphical representations. Just as with frequency curves, it is essential to critically assess the underlying data and measures in collocation analyses. Alongside heuristic and hermeneutic guidelines [46], supplementary tools can also prove beneficial in this context.

(d) *Topic models* have the advantage that they can identify typical wider contexts in which a word is used. There are various designs and applications of this method, but they are all based on the assumption that identifiable subject areas in text collections, i.e., the topics, can be derived from the distribution of thematically related words in each corpus. Based on this assumption, topic models generate interpretable word fields from statistical regularities and a calculation of corresponding probability distributions. While these models are primarily employed to identify topics within corpora, assign documents to these topics, and track their distribution over time, they can also offer valuable insights for researching historical semantics [63]. In early historical applications of this method, it was shown, for example, that topics “corresponding to sequenc-

ing and cloning, structural biology, and immunology" became very popular in the Proceedings of the National Academy of Sciences of the United States of America (PNAS) around 1991 [30]; or that the increase and decline of government-related topics in historical newspapers like the Pennsylvania Gazette align clearly with trends in political history [30]. To achieve meaningful and reproducible results, the complex parameters of the models must be well adjusted [8]. While there may be relatively few parameters to modify [54], a reliable application of the procedure requires specialist knowledge and analysis routines. Based on this expertise, several further interpretative steps are still required to transition from identified topics to the level of word semantics and conceptual meanings. However, researchers have relatively little control over the topics, neither in terms of their discoverability nor in terms of their interpretability [33, 55]. This constraint poses challenges in searching for specific conceptual vocabularies and providing evidence of their absence with the help of topic models. While a major advantage of this method lies in its ability to analyze large volumes of text without prior knowledge about the sources and concepts involved, this limitation is certainly the most significant drawback for its broader application in the history of concepts.

(e) *Semantic tagging* or *semantic annotation* comprises of a set of language technology that involves analyzing and classifying texts through the application of structured information. This methodology entails the formal description of content or data to identify topics, concepts, and entities. Essential resources for semantic tagging encompass ontologies, taxonomies developed by experts, domain-specific knowledge, existing metadata, and reference materials such as dictionaries and thesauri [1, 11, 61]. An example of an initiative in this domain is the Dataset of Historical Ontologies (DHO), which focuses on creating and maintaining ontologies that are tailored to various historical periods and themes [12]. A recent application of semantic tagging in the field of conceptual history is the annotation of the Hansard Corpus—a digitized compilation of British parliamentary debates—using the Historical-Thesaurus-based Semantic Tagger (HTST) [28]. This tool categorizes lexical units based on semantic criteria, facilitating targeted searches of historical references and specific speech acts. Such applications raise expectations that a multitude of historical expressions and speech acts can be formalized as articulations of a manageable number of specific concepts and thereby rendered diachronically searchable within extensive text corpora. However, achieving this level of formalization presents challenges and is currently only partially realized [28]. A fundamental methodological challenge remains the question of how the dynamics of conceptual changes can meaningfully be related to a fixed nomenclature of structured knowledge, or whether historical word meanings can be adequately captured by ontologies or dictionaries at all. Future advancements in machine learning and natural language processing may provide opportunities to automate or enhance this process, even within the specialized field of conceptual history.

(f) *Word embeddings* hold significant promise in the domain of conceptual history, especially since the availability of the word2vec model [48], and most recently the advent of Large Language Models (LLM). Word embeddings situate words in continuous vector spaces, where semantically similar words are positioned proximally. This approach enables the numerical representation of semantic relationships among words. Static word embeddings are generated using algorithms like word2vec, which analyze

the context of words within large text corpora based on word co-occurrences and represent words with a single vector each. Contextual word embeddings are generated by LLMs like BERT [17] and embed every single occurrence of a word as a different vector. For both variants, it holds that when trained on different corpora, vector-spaces reveal typical patterns of word usage and semantical relations without the use of predefined dictionaries or ontologies. Thereby, word embeddings can assist in identifying word meanings: Represented as vectors – for example with cosine as a measure for similarity – semantically related words are positioned closely together, whereas unrelated words are spaced apart within the vector space. Consequently, alterations in these spatial relationships over time can indicate semantic and thus possible conceptual changes. Therefore, word embeddings are of particular interest for the research in the history of concepts [25, 26]. While recent studies can demonstrate a number of promising results from their application [25, 33, 34, 40, 64], word embeddings remain associated with various methodological challenges. Models must be trained on a given corpus, based on arbitrary parameters. To assess them effectively, a "ground truth" reference is essential, which could take the form of a predefined dictionary again, or human annotators. However, historical research often lacks such "ground truth" references, and creating them for each study would quickly strain the available resources of time, human labor, and computing power [33]. While there are some feasible workarounds, word embeddings remain opaque, leaving room for the question to which extent cosine similarity is really about semantic similarity [62, 66], and semantic similarity as such is a fuzzy notion. Another major limitation of classical static word embeddings is that they provide only one representation in vector space for polysemous words, effectively eliminating polysemic structure [33]. While there are approaches to modelling polysemic words with several embeddings, either by splitting single static embeddings into several [3] or by clustering contextual embeddings into sense clusters, it is still a challenge to relate sense embeddings between different corpora or different time slices since the vector representations remain opaque and subject to fluctuations [42] – something that is a great disadvantage in the research of historical semantics.

(g) *Word sense induction* (WSI) is a natural language processing technique used to identify different meanings (senses) of a word by statistically analyzing the contexts in which the word appears in a large corpus of text [13, 18, 43]. These contexts can be represented in co-occurrence matrices, word embeddings, or syntactical relations involving the target word [4, 6]. By quantitatively analyzing the contextual information of a search term, WSI typically employs clustering algorithms to group similar contexts together, without relying on pre-defined lexical resources such as dictionaries or thesauri. Each resulting cluster is assumed to represent a distinct sense of the target word, reflecting its varied uses in different contexts. For example, the word 'bank' might form one cluster related to 'finance' and another related to 'river.' This makes this method particularly suitable for detecting polysemy. The aggregation of word sense clusters is based on calculating similarity scores between all words in each corpus. Unlike word embeddings, where similarity scores represent proximity in a vector space, WSI uses statistical association measures, often derived from syntactical relations extracted from a parsed corpus. Words that share many features, such as frequent subjects or adjectives, are considered more similar than those that share fewer elements. Consequently,

word sense clusters and their changes over time can be interpreted as a representation of the semantic structure of a search term and its historical development [5, 24, 31]. This approach presents substantial advantages for researching historical semantics for three primary reasons. First, it does not depend on predefined lexical or ontological resources like dictionaries or thesauri. Second, unlike other computational methods, it is particularly effective at identifying ambiguity and polysemy in words. Third, its foundation in statistical techniques guarantees that the analysis outcomes—including similarity scores and resulting clusters—remain consistently transparent. Such transparency is crucial for understanding the factors that contribute to the emergence or disappearance of word similarities and the subsequent interpretation of the possibly related conceptual changes [46]. Therefore, WSI supports to the principle of provenance and provides the potential for a scalable reading. However, its major disadvantage is certainly the requirement for vast amounts of data, substantial computing power, and a relatively high level of user expertise to effectively utilize this language technology. Thus, after an initial proposal for its use in the history of concepts [24], this method has only recently found wider application [23, 52].

To address and mitigate the limitations of previous techniques while enhancing their strengths, several advanced approaches to semantic change detection have been proposed. Notable approaches that innovatively combine, extend, and advance existing methods are “context volatility” [38], “distributional probability” (dpf) based on “lexical co-association” [15, 16], or “local neighbourhood measure of semantic change (LNM)” [34]. In general, it seems helpful and advisable to combine several methods in historic semantical research. A mixed-methods approach not only strengthens the quantitative basis of arguments but also facilitates a thorough assessment of the outcomes derived from individual analyses. However, different findings also require a careful interpretation and thus further questions with regard to digital hermeneutics.

The emergence of increasingly advanced language technology and computational linguistic methods, along with the growing number of demonstrations and publications of their capability and utility, is prompting a dual response in the field of conceptual history research. On the one hand, there is a rising interest in integrating these tools into individual research efforts and personal workflows. On the other hand, even relatively simple applications, such as frequency analysis, increase the demand for methodological knowledge in the underlying language technology for proper use to achieve reliable result. The visual representation of complex data often provides a low-threshold access point or interface to complex quantitative text analyses, even for “occasional users.” A well-known example is certainly the Google Ngrams Viewer. However, as text analysis methods become increasingly complex, the potential for misinterpreting visual representations can lead to inappropriate or arbitrary conclusions about data analysis results. This risk is present not only when the underlying language technology lacks inherent transparency, such as with word embeddings, but also in more basic applications like frequency analysis which can be misinterpreted in several ways [46]. While the latter risk can be mitigated or solved by acquiring skills in critical graph interpretation, the former represents a severe technical limitation.

Therefore, in addition to fostering interdisciplinary education and collaboration in the field of computational humanities, there is a crucial need for digital tools that enhance data interpretability. These tools should be designed with a focus on user expectations and workflow efficiency to better support the interpretation of computational analyses.

3 Requirements and Challenges in Tools Design for Digital Research in the History of Concepts

Based on the desiderata identified in the state of research and on our own experiences in this field, we propose the following features that digital research tools should possess to efficiently support digital research in the history of concepts:

(a) *Data-drivenness*: The tool should enable the (semi)automatic detection of lexical or conceptual features from text data in large corpora with minimal or no prior knowledge or assumptions about its content.
(b) *Versatility*: The tool should allow to employ different methods of text analysis.
(c) *Sensitivity*: It should enable the identification of different senses of a word, as in case of polysemic or ambiguity, which are key in historical semantics.
(d) *Diachroneity*: The tool must allow to track the change of word senses over time.
(e) *Visualization*: Instructive visual representations, like diagrams of graphs, should support data analysis, enabling distant reading approaches.
(f) *Transparency*: The application must ensure that the relationship between visual representations and their underlying data, alongside the method of their computation, remains both comprehensible and interpretable by users.
(g) *Traceability*: Users should be able to trace visual representations and the reasons for their diachronic change back to relevant documents or text passages, thus enabling a transition from distant to close reading—or scalable reading.
(h) *Accessibility*: The application should be as user-friendly as possible, ideally open-source, and immediately ready for use.
(i) *Adaptability*: Users should be able to modify settings and parameters of the tool to obtain, test, and compare various research results and it should provide annotation features.
(j) *Reproducibility*: The analysis results should be easy to archive, to reference and reproduce for further study.

In the following, the word sense induction (WSI) tool "Sense Clustering over Time" (SCoT)[1] is presented that has been developed specifically for applications in the history of concepts [31] and which is being used on a larger scale for the first time in the research project "The 20th Century in Basic Concepts. A Dictionary of Historical Semantics in Germany" [52]. As its technical foundation and earlier stages of its development have been described elsewhere [4, 5, 9, 31, 39], this article will focus on its application and workflows in research in the history of concepts.

[1] https://scot.ltdemos.informatik.uni-hamburg.de.

4 Researching historical semantics with the tool Sense Clustering over Time (SCoT)

4.1 A structural semantics approach

As a WSI tool, SCoT is essentially based on the distributional hypothesis. This fundamental hypothesis of structural semantics suggests that: *Words can be considered similar if they have a similar distribution of all contexts in which they occur* [32, 49]. In the case of SCoT, 'context' refers to the syntactic relations in which a given word occurs. This approach is an operationalization of the idea of structural semantics to view language as a system of syntagmatic and paradigmatic relationships. This approach can be illustrated using an example sentence (Fig. 1).

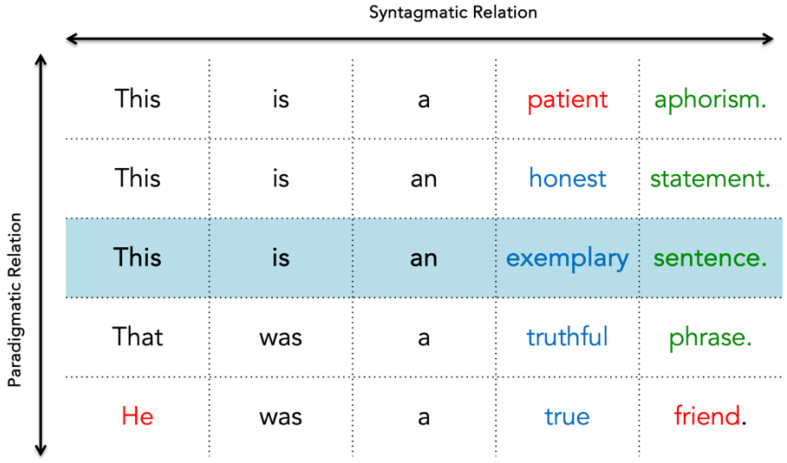

Fig. 1. The two dimensions of structural semantics: In the paradigmatic dimension, words can, to a certain degree, substitute each other within shared syntactic relationships.

In the paradigmatic dimension, the grammatical object of the highlighted exemplary sentence (*sentence*) could be replaced by a similar word (such as *statement*), without substantially altering the sentence's meaning. Therefore, we could accept the expressions: *This is an exemplary sentence,* and *This is an exemplary statement* as equivalent (although, of course, there are always contexts in which such little difference would change a lot). In general, however, we can say that due to such replaceability, the words *sentence,* and *statement* are more similar to each other than, for example, the words *sentence* and *friend*. In this sense, the green-colored words can be regarded as paradigmatic elements of each other. As such, they can easily share any of the blue adjectives with each other: *true* can be attributed to *friend* as well as to *sentence* without any semantic trouble. In contrast, the adjective *patient* that could connect with *friend* as well would induce a significant semantic irritation, perhaps a metaphorical effect, when combined with one of the green elements, e.g. in a sentence like: *This is a patient aphorism.* Therefore, *friend* would not be regarded as a paradigmatic element of *sentence*.

The open-source framework JoBimText [9], which is a natural language processing component of SCoT, can detect such paradigmatic relations of any word in a given text corpus based on part-of-speech recognition (Fig. 2). For each lexical unit (word), JoBimText calculates and rates the syntagmatic elements (contexts) that are most frequently associated with it. Key words need to be queried in connection with a specified part-of-speech (POS) tag. Thus, depending on POS tags, word inflection can also be considered in research tasks. Based on the user's query, common syntagmatic elements are counted for all pairs of lexical units, resulting in their similarity score [8, 9]. A value of 0 would mean no shared syntactic features and thus no similarity at all and the maximum value of 1000 would indicate the linguistic identity between two words. Words with such maximum value could replace each other anytime without any semantic difference. Of course, in natural language such perfect synonyms do not occur. Instead, high to very high similarity scores turned out to range between 200–500.

By comparing the associated paradigmatic elements of each word, JoBimText detects similarity clusters [4, 7]. In this way the words *decree*, *paragraph*, *decision,* for example, appear as elements of the same cluster to a shared vocabulary obviously related to the domain of jurisdiction. 'Jurisdiction', thus, can be interpreted as the domain of one sense of 'sentence' meaning 'the declaration of a punishment assigned to a defendant found guilty by a court.' Other words found to be similar to *sentence*, like *expression*, *verse*, *phrase*, *couplet*, *stanza* or *syllable,* appear as elements of a different cluster obviously consisting of linguistic entities, and so on.

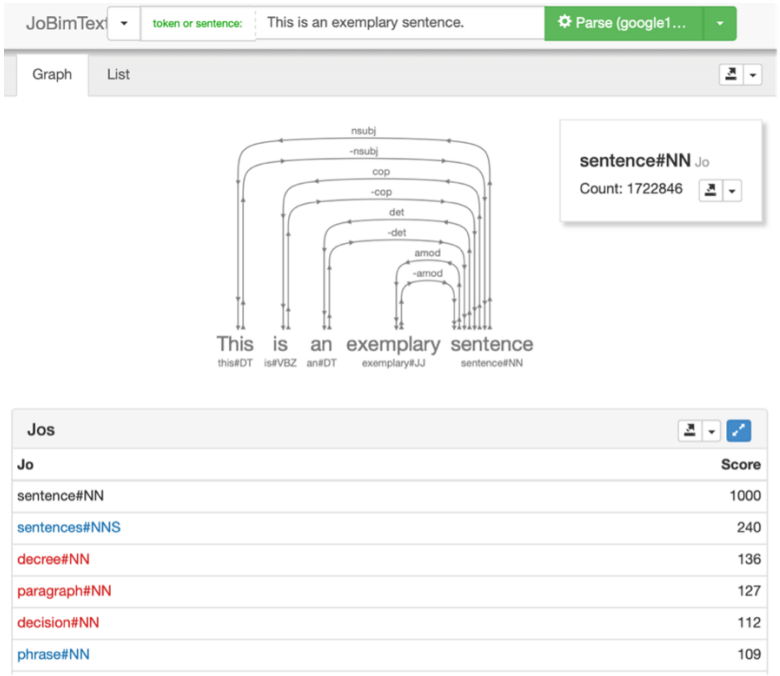

Fig. 2. The JoBimText framework. See http://ltmaggie.informatik.uni-hamburg.de/jobimviz/.

Such clusters apparently resemble topics models. The main difference, however, is that detected paradigmatic elements do not have to occur in the immediate neighborhood of the search word in order to be considered similar. A strong similarity may also indicate an onomasiological substitution, i.e. one word historically replaces another and takes over its previous function in a semantic network. In the German language, for example, the term *kreativ* increasingly replaced the older term *schöpferisch* in the second half of the 20th century, with both terms denote inventiveness and the productive use of imagination.

4.2 Analyzing the historical semantics of *kreativ* with SCoT. An exemplification

If we plot the diachronic development of the similarity value between these two terms in the German-language Google Books corpus with SCoT (Fig. 3), we find that a similarity between the two terms first appears in the second half of the 20th century, before peaking in the 1980s with a value of 480, which is enormously high, and then falling again significantly, before stabilizing at a level of around 300 – still indicating a very high similarity.

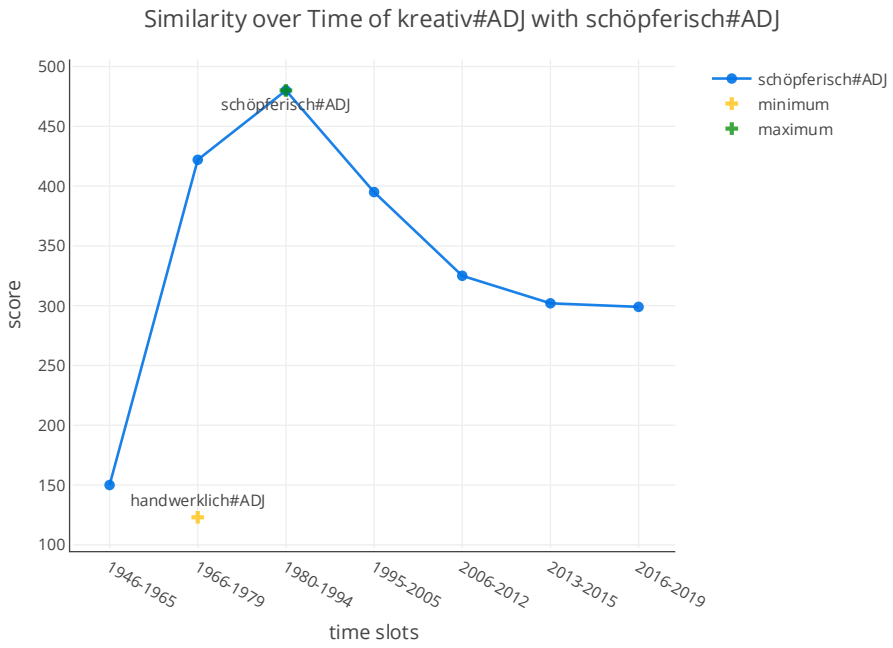

Fig. 3. A diachronic plot of the similarity score between the adjectives *schöpferisch* and *kreativ* in the German Google Books corpus

A comparison of the frequency development of both terms shows that *kreativ* becomes into regular use not before the 1960's (with a frequency above 1 ppm) – at a time when *schöpferisch* has its peak of use, which then drops again significantly and is

overtaken by the frequency of *kreativ* (Fig. 4). Taken together, the developments in the similarity score and word frequencies can be interpreted as an onomasiological substitution. However, a closer examination of the history of these terms in the 20th century shows that the (ex)change of the two words is also accompanied by a change in the concept they denote [10]. One major difference is that the older concept of *Schöpfertum* was still strongly associated with the concept of *Genie* (*genius*), whereas *Kreativität* in the post-war period is understood as something that can be scientifically analyzed and also enhanced by certain techniques – especially for economic and entrepreneurial interests in mobilizing human resources and productivity (which is also one reason why the term developed quite differently in the two opposing political systems of the German post-war period) [10].

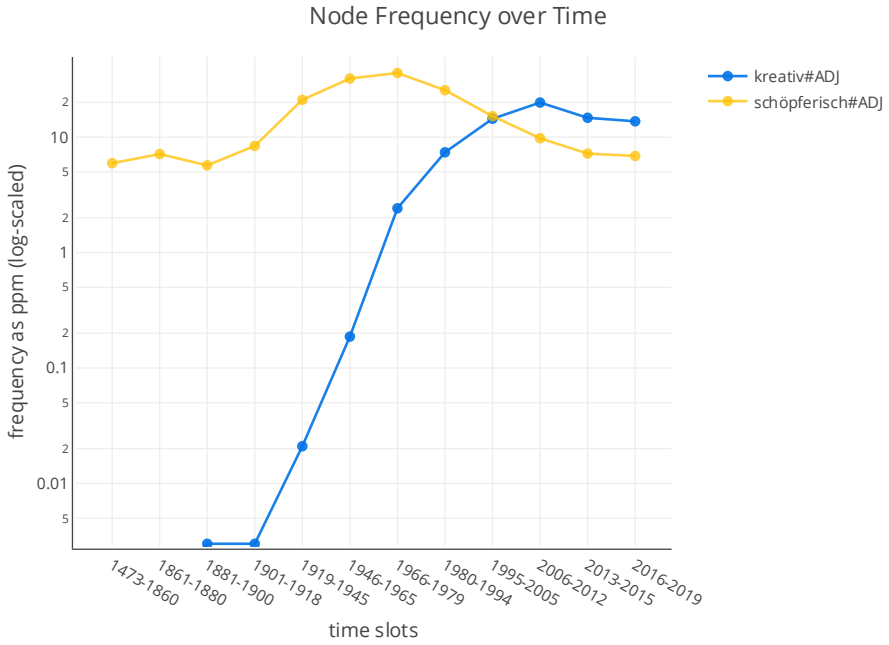

Fig. 4. A diachronic plot of the word frequency of the adjectives *schöpferisch* and *kreativ* in the German Google Books corpus

With the help of SCoT, the change in the historical semantics of both terms can be visualized. The following graph shows the semantic network of *kreativ* consisting of the 30 most similar words for the adjective in the German-language Google Books corpus for the entire period 1473–2019 (Fig. 5).[2] The degree of similarity is indicated by the size of the nodes: the bigger the node the more similar the word it represents. The word *schöpferisch,* for example, is one with the highest similarity value (480). The graph itself, its clusters and their coloring were generated automatically by SCoT based

[2] SCoT: *German Google Books*, period=1473–2019, query=kreativ#ADJ, N=30, D=15, graph type=NGoT fixed, https://scot.ltdemos.informatik.uni-hamburg.de/.

on three input values: the search word including part of speech tag (ADJ for adjective), the number of nodes (N) and the density of the Graph (D), controlling the number of possible links between the nodes. Such links (or edges) indicate a similarity score between each node. The density value adjusts the threshold of ranked similarity values of all existing edges in percent. In this case, the density value D has been set to 15, resulting in 131 of 870 max. possible edges. Based on these edges, clusters are calculated using the Chinese Whispers graph clustering algorithm [5]. In each cluster, elements exhibit a higher degree of similarity to each other than to those in other clusters. The degree of similarity depends on the number of shared contexts. In this way, clusters represent different senses of the search term (which itself, of course, is not included within the network). Consequently, the similarity graph gives a visualization of the semantic structure of a word within a corpus over a specified time span.

Let us examine this more closely:

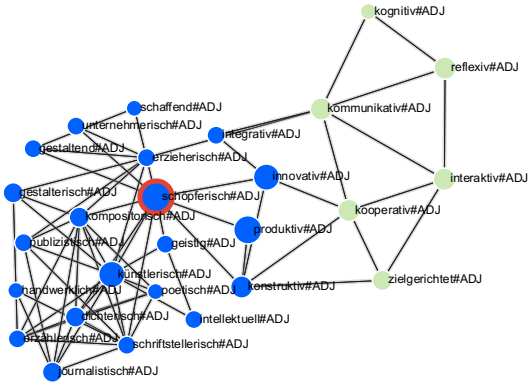

Fig. 5. Similarity graph of *kreativ* based on the thirty most similar words in the German Google Books corpus for the period 1473–2019

The graph in Fig. 5 consists of two clusters: the lefthanded blue one contains adjectives like *schaffend* (creative), *gestalterisch* (formativ), *künstlerisch* (artistic), *poetisch* (poetic) as well as *unternehmerisch* (entrepreneurial), *journalistisch* (journalistic), *handwerklich* (artisanal), and *produktiv* (productive); the righthanded green one, on the other hand, contains adjectives such as *kooperativ* (cooperative), *interaktiv* (interactive), *reflexiv* (reflexive), *kognitiv* (cognitive), and *konstruktiv* (constructive). In the blue cluster, we find the biggest nodes of the graph. This gives us a first indication that the blue sense is more dominant than the green one; we may interpret this finding as the "main sense" of the term within this time period. How does it relate to the green one? Reading the graph, it is already possible to guess the difference between both clusters: The blue cluster seems to contain elements related to the individual ability to make or craft things, while the green cluster seem to be primarily related to group activities and organizational processes. And there are also some connections shown between them.

Examining and understanding the exact reasons of this network structure more closely requires a detailed cluster analysis, made possible by the functionalities of the SCoT. Such an analysis reveals that the most significant contexts for the blue cluster are syntagmatic relations to nouns that refer to the human ability to create something or the productive exercise of imagination, with nouns for: *creation, imagination, design, talent, potency, production, practice, work, talent.*[3] In contrast, the most significant contexts of the green cluster include nouns that refer to professional qualification, teamwork and relationships such as *competence, cooperation, proposal, process, dialog, product, exchange,* and *collaboration*.[4] Thus, the context analysis confirms our hermeneutic anticipation: There seem to be two different meanings of *kreativ*, one of which has its place in the context of artistic or artisanal activities, in which the (highlighted) adjective *schöpferisch* is also located, and the other sense is more related to economic contexts.

However, both aspects have not been present throughout all times. The thirty nodes composing the clusters represent the most similar words for *kreativ* for the entire period of investigation. As a unique feature, SCoT makes it possible to visualize the distribution of paradigmatic elements over time in a diachronic representation.

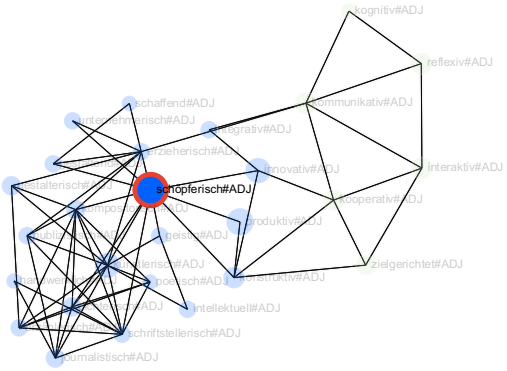

Fig. 6. Similarity graph of *kreativ* in German Google Books 1946–1965

The "Time-Slice" mode of SCoT reveals that none of the top thirty nodes occur among the most similar words to *kreativ* before 1946. As the frequency analysis already indicates (Fig. 4), this is because *kreativ* does not occur in German-language book publications before the middle of the 20th century on a relevant scale. It is not until the period 1946–1965 that a paradigmatic element for *kreativ* appears at all; the first and single one is: *schöpferisch* (Fig. 6). Then, in the following period, 1966–1979, something remarkable happens: almost all the other nodes occur at once (Fig. 7).

[3] The respective German nouns are *Schaffen, Phantasie, Gestaltung, Begabung, Potenz, Produktion, Praxis, Arbeit,* and *Talent*.

[4] The respective German nouns are *Kompetenz, Zusammenarbeit, Vorschlag, Prozess, Dialog, Produkt, Austausch,* and *Zusammenarbeit*.

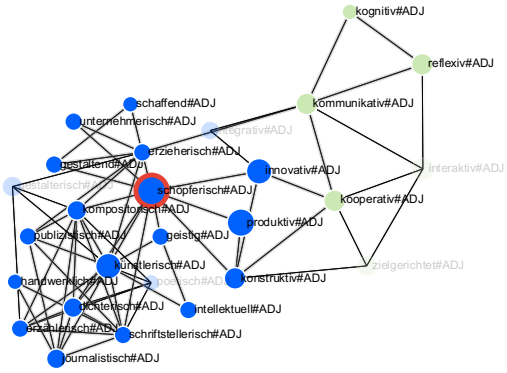

Fig. 7. Similarity graph of *kreativ* in German Google Books 1966–1979

As Morten Bierganns has shown, this development in the history of the concept results mainly from the adoption and Germanization of the English terms *creative* and *creativity* in German experimental psychology: "The noun *Kreativität* ('creative power,' 'creative ability') first appeared as a borrowing of the English *creativity* in the first half of the 20th century as a specialist psychological term and has seen a sharp increase in frequency of use since the mid-1960s" [10]. Based on this scientific development, the term and its related concepts was soon adopted in economic and labor science as well as educational policy discourses [10]. The latter development is reflected in the formation of the green cluster; the expansion of the blue cluster in turn indicates that the term assimilates parts of the semantics of *schöpferisch* in the course of this development.

A comparative look with SCoT at identical settings in the same corpus shows a significant overlap of the similarity graph of *schöpferisch* as it contains paradigmatic elements from the blue cluster of *kreativ* such as German adjectives for *poetic*, *intellectual*, *productive*, *artistic*, also *entrepreneurial* and *authorial*.[5] It is also interesting to note that the similarity graph of *kreativ* is reduced again towards the end of the 20th century, while the balance of power between the two clusters has shifted in favor of the green one (Fig. 8) – which speaks for the growing economic implication of the term within the German Google Books corpus. In this way, SCoT may help to detect changes in the meaning (sense) of conceptual terms.

[5] SCoT: *German Google Books*, period=1473–2019, query=schöpferisch#ADJ, N=30, D=15, graph type=NGoT fixed. The corresponding German nouns are *dichterisch, geistig, produktiv, künstlerisch*, *unternehmerisch* und *schriftstellerisch*.

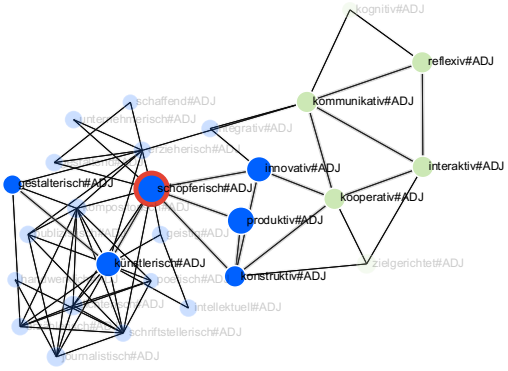

Fig. 8. Similarity graph of *kreativ* in German Google Books 2016–2019

4.3 Scalable reading with SCoT

The Sense Clustering over Time tool not only visualizes diachronic changes in similarity graphs but also offers a comprehensive suite of analytical tools. In addition to synoptic diagrams that chart frequency values and similarity scores over time (presented above), SCoT enables the validation of findings by referencing specific text passages – provided the entire corpus under study is available in full text. This is not always the case, for example with Google Books, which is a problematic text corpus also for other reasons, for example due to a lack of transparency of its composition. However, when querying other, fully available and transparent corpora, e.g. the German Reference Corpus provided by the IDS Leibniz Institute for the German Language with the same search term and POS-tag,[6] SCoT allows to examine corresponding example sentences for the term usage in question. For instance: "Although a 'violation of existence' […] stimulates creativity (*Schöpfertum*), creative potential (*kreatives Potential*) – Matussek comforts us – is waiting to be 'discovered, awakened and unfolded' in everyone, especially in mentally healthy people." (Fig. 9, left sidebar, my translation). This quote from a review of two books on creativity in the news magazine *Der Spiegel* [2] documents the popularization of the psychological concept in German language. Such example sentences can be traced via the syntagmatic relation *kreatives Potential* (creative potential) that is significant for *kreativ* in this corpus as revealed via context analysis (Fig. 9., right sidebar). The list of all example sentences can be filtered with further search terms. Thus, by enabling scholars to navigate from the graphical representation of the statistical data to the citation-based document evaluation, SCoT seamlessly supports scalable reading.

[6] SCoT: *IDS German Reference Corpus*, period=1945–2021, query=kreativ#ADJA, N=60, D=20, graph type=NGoT fixed.

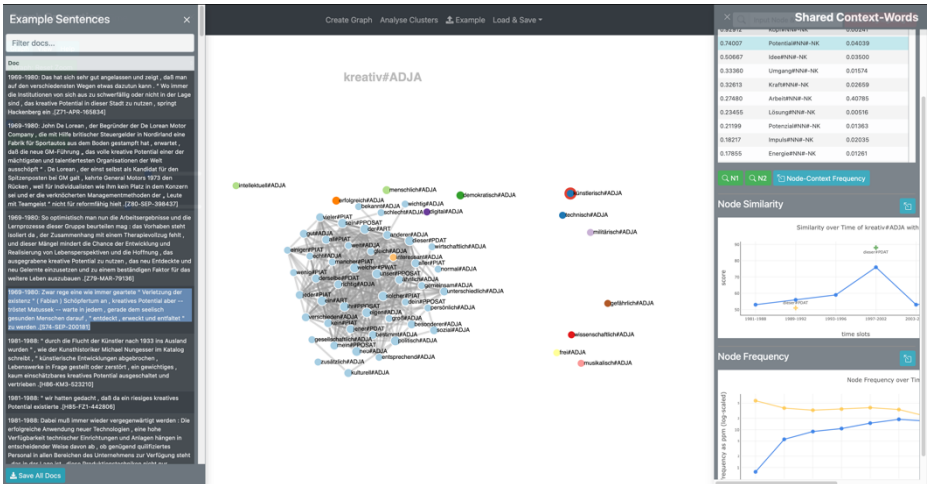

Fig. 9. Context analysis tools in SCoT: shared context words, node similarity, node frequency and example sentences for the similarity relation between *kreativ* (creative) and *künstlerisch* (artistic) in the German IDS corpus

5 Discussion

After presenting an example from the German conceptual history of creativity that demonstrates a user-centered workflow using SCoT, we will discuss the viability of this tool for research within the digital history of concepts and its implications for digital hermeneutics. In this discussion we will also consider the proposed features outlined in Section 4 and the IVIS reference model for digital hermeneutics.

5.1 Other application examples and use cases

Regarding its dedicated purpose, it is essential to consider and evaluate whether the application of SCoT can be effectively extended to other use cases. Several examples can be referenced in this context. In a first pilot study, it was shown with a corpus of German-language newspapers how a problem discourse about networks (as a structural feature of our interconnected world) emerging at the beginning of the 21st century affected and changed the semantic structure of the highly polysemic concept of *Netz/werk* (net/work) [24]. This work has later been approved and extended by a SCoT-based analysis on the historical semantic of *Netz, Netzwerk, Vernetzung* (web, network, interconnection) in large German language text corpora of the 20th century [22]. In the technical demonstration of the first published prototype of SCoT, the historical semantic of *crisis* in the English Google Books corpus has been used as a test case to show a shift of meaning of the term towards an increasing economic sense in the second half of the 20th century [31]. A further example is the history of the concept of *Energie* (energy): Using SCoT, it was possible to show that at the end of the 19th century, the German

concept of energy was primarily used to address human behaviors, but over the course of the 20th century, due to the quantitative increase in scientific texts, it was more frequently used in its physical-physiological sense and thereby developed a new, distinct sense – that could be related to an analogous development in English language [29]. The latter example comes from the first systematic application context of SCoT: the German lexicon project on "The 20th Century in Basic Concepts. A Dictionary of Historical Semantics in Germany" [52] already mentioned above. In this lexicon, a contribution on the history of the concept of *Aufklärung* (enlightenment, intelligence) has been published that we like to cite as our last example: In this article, SCoT was used to show how the historical semantics of the concept shifted in the course of the 20th century from a term used to describe or guide actions to a term denoting an epoch [51].

In the context of this lexicon, further works successfully utilizing SCoT have been published as part of the open access project running since 2020, developed and coordinated by the Berlin Leibniz Center for Literary and Cultural Research in cooperation with the Leibniz Institute for the German Language Mannheim and the Leibniz Centre for Contemporary History Potsdam [52]. However, it can also be reported from authors working in this context that attempts to apply SCoT have not yielded useful results in certain cases, partly due to restriction or unavailability of text corpora, partly to unsatisfying analysis results. In some cases, for example, certain syntactic features (such as frequent prepositions) dominate the similarity graph in a way that seems inappropriate und uninformative for a given research context. Similar findings have also been reported elsewhere [40]. While this kind of limitations are partly related to the accessibility and quality of the available text corpora (discussed in the next section) they also raise general questions of the viability of the method.

5.2 Reliability and viability of the method

SCoT is based on components whose reliability have been tested and proven elsewhere [4, 5, 7, 9]. A comprehensive system description has been given by Haase et al. [31]. Using the JoBimText framework, SCoT calculates semantic similarity utilizing distributional thesauri (DT). This approach has been selected due to its versatility, allowing it to accommodate a variety of context features, such as word n-grams, part-of-speech n-grams, and syntactic dependencies. For SCoT, JoBimText employs syntactic dependencies to process text, extracting syntactic features of single words based on POS-Tags and calculating their frequencies and association scores through statistical measures like Lexicographer's Mutual Information (LMI), Pointwise Mutual Information (PMI), and Log-Likelihood (LL) [20]. The default settings use LMI to rank and retain the top 1,000 features for each word based on shared syntactic contexts. The choice of JoBimText enhances the semantic modeling capabilities of the framework in an unsupervised manner, without relying on pre-existing lexical resources. When evaluated on tasks such as lexical substitution, this framework surpasses non-contextual models by effectively addressing challenges like ambiguity and synonymy [6]. Overall, it offers a robust, *data-driven*, *versatile*, and context-aware and thus *sensitive* methodology for modeling semantic relationships in natural language processing tasks. Con-

text-awareness also ensures that the framework maintains *transparency* and *traceability*, as the computed similarity scores can be directly associated and evaluated with the specific text-features of each word. The clustering and *visualization* functionalities of SCoT are based on the Chinese Whispers algorithm [5]; the front-end utilizes the Model-View-ViewModel (MVVM) framework Vue [31], and allows a diachronic analysis of the similarity graph as shown in Section 4, based on the DT of time-sliced corpora. SCoT also features a view configurator, supporting argument development in the sense of the Draft Reference Model (DRM@DH) presented in this volume.

The GUI offers further functionalities to display a list of syntagmatic contexts per selected word-nodes, including whole clusters, as ranked by LMI. The GUI therefore supports transparency and traceability. By supporting transparency and traceability, SCoT benefits historical research interests in accordance with the principle of provenance as it provides the possibility to explore specific contexts for given similarity scores or cluster representations. Users can track down relevant sentences from the text corpus for further investigation or CSV export. Thus, SCoT supports scalable reading – given that full access to the text corpus is provided. Which is, although desired, not always the case: Google Books, for example, does not come as a full text corpus but as n-gram dataset. Therefore, in this case, users are not able to read original documents. SCoT can process such data sets but only offers limited analysis options in these cases: The hermeneutic circle can then only operate at distant reading level – or must be connected to the level of close reading in other ways.

5.3 Provision and Processing of Data

The issue of corpus accessibility generally points to questions regarding the data basis and processing methods of SCoT. The transformation of text data into similarity values was explained in the previous section. Before these calculations can take place, suitable text corpora must first be acquired and prepared in an appropriate manner for the calculations. The acquisition of suitable corpora presents its own problem.

SCoT is only suitable for very large corpora, as small text collections cannot be meaningfully analyzed with symbolic WSI methods. The data-hungriness of SCoT is exacerbated by its transparent nature: While embedding-based methods can leverage word similarities for context similarities, JoBimText stays on the symbolic level. This approach preserves transparency but necessitates that contexts be exactly the same to be considered a signal of similarity. Large corpora, on the other hand, are not available for all languages and are not equally accessible. Full-text access is a prerequisite for computing the DTs. This access can be restricted or prevented by external obstacles, such as copyright or licensing conditions.

If access is possible, as for example in the case of the English-language Hansard corpus or the German-language Bundestag corpus, each corpus must be time-stamped, parsed and annotated with POS-tags, and finally split into time slices to enable meaningful diachronic analyses. Here, it is necessary to make trade-offs: on one hand, the chosen time segments must be large enough to provide meaningful statistical values; on the other hand, they must be small enough to allow for a desirable temporal resolution (in the history of concepts: ideally something between a year and a century). Once

a reasonable proportion has been determined, one must also decide *where* to place the temporal cuts. These can be oriented around familiar historical dates (e.g., 1945 or 1968) or other temporal units. Thus, raw data is already drawn into the hermeneutic circle during the pre-processing phase of texts.

Additionally, the corpora themselves can (and must) be subject to source criticism: Who compiled them under which criteria? How are they composed? Although criteria for a balanced composition play a decreasing role as the size of the corpus increases [44], this does not exempt us from the critical reflection on the composition of the sources and what possible biases might be associated with them. In the case of Google Books, for instance, it must be considered that the texts it contains are based on scans of unspecified collections of American university libraries. Thus, the selection of sources is pre-filtered not just by the collection criteria of the libraries and the selection process of Google, but also, for example, by the publication practices of certain academic disciplines: A large part of the scientific discourse of the second half of the 20th century will not be found there because it took place in journals, which are not included in the corpus. On the contrary, other corpora are very transparent: the German Parliament corpus, for example, contains all recorded communications of the Bundestag. However, the quality of the transcriptions and pre-processing processes remains a question to be considered.

Once a suitable corpus has been thoroughly processed for SCoT analyses, it can be made freely available online for interested users. As of the publication date, eight corpora in three different languages are accessible through the SCoT open-access platform, with plans for ongoing additions.

5.4 Evaluation and Interpretation of Research Results

Before a corpus is made accessible for SCoT analyses, it should be evaluated in terms of its preprocessing. Such an evaluation must be carried out through sampling or frequent applications by human experts: Do the queries produce plausible results? Answering this question is not yet proof of the validity of the calculated DTs. It is fundamentally possible that a Word Sense Induction produces counterintuitive results, which are not necessarily incorrect just because they do not align with our linguistic or historical expectations. However, they are always the basis of our data interpretation, and thus of the evaluation process. Alongside a hard-to-operationalize intuition regarding the plausibility of the results, critical questions aid in the assessment: Does a significant portion of the visualized paradigmatic elements of a similarity graph align with our knowledge of historical word semantics? Can clusters be assigned to recognizable word senses? Can their changes over time be related to meaningful contexts? Would or do many human experts agree on the answers to these questions?

The node and cluster analysis functions of SCoT play a crucial role in achieving answers, and every digital analysis of the historical semantics of a term should leverage these tools. By examining the top-n features of a node, one can gain valuable linguistic insights into its usage, leading to abductive conclusions about its conceptual role (as demonstrated in Section 4). Once a sufficient number of test queries and analyses have

been conducted, the corpus will be deemed ready for public release. However, the criteria for what constitutes 'sufficient' remain open to interpretation in this context: We can never rule out the possibility of issues arising later that were previously overlooked. These issues can only be addressed again through improved pre-processing, while the fundamental work and evaluation process remain unchanged. Consequently, in the long term, the evaluation process operates within a hermeneutic circle as well.

SCoT provides additional functions and visualization options that significantly enhance the evaluation process of analysis results. As highlighted in the state of research (Section 2), every method for a digital history of concepts presents its own set of advantages and limitations. Consequently, it is inadvisable to rely solely on a singular methodology. A comparative approach should always be a preferrable strategy. To facilitate this approach in terms of the desired feature of versatility (Section 3), SCoT offers a synoptic view, allowing for the comparison of the progression of average similarity and frequency values for all nodes within a cluster over time. The following figure (Fig. 10) presents the diachronic values for the two clusters of *Kreativität* previously examined in Section 4, based on the German-language Google Books corpus (cf. Fig. 5).

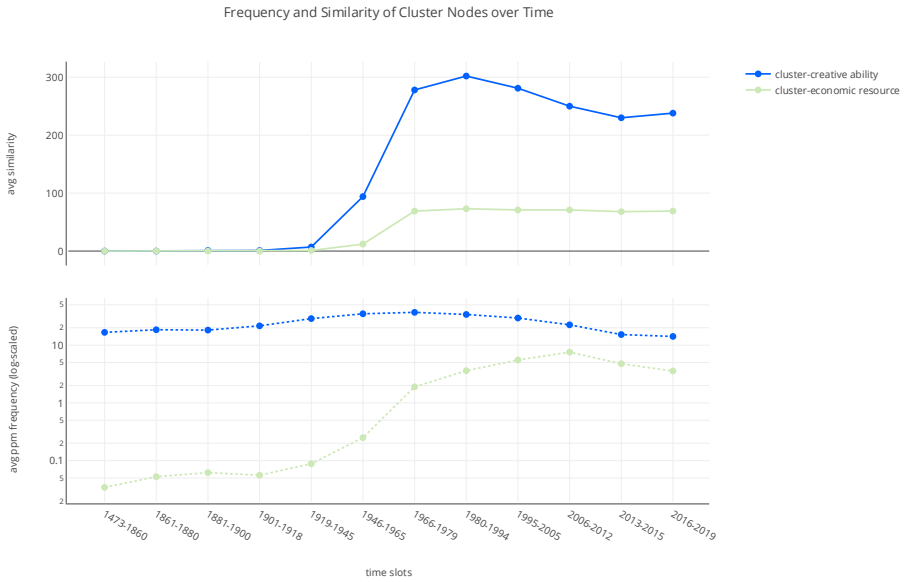

Fig. 10. Synoptic view of average similarity scores and frequency values of the nodes of different clusters over time – in this case: two clusters of *Kreativität* in German Google Books

In our exemplary analysis we found that the blue cluster represents creativity as an individual ability and the green one creativity as an economic resource. In addition to the diachronic cluster analysis, the synoptic diagram in Fig. 10 provides further evidence. It reveals that the blue 'ability'-cluster exhibits the highest similarity scores recorded during the observed period, reaching levels up to 300 (which can be regarded as

very high). Notably, these elevated values are attained only after end of the World War II, while the frequency of nodes within this cluster remains stable on average. Consequently, the rapid increase in similarity value cannot be attributed to significant fluctuations in frequency values. Conversely, the green economy cluster shows a different pattern. Although there is an increase in similarity values in the post-war period, it is not as pronounced as observed in the blue cluster. However, the average frequency of the cluster nodes experiences a substantial increase, rising a hundredfold overall. This significant increase could potentially impact the similarity value. Such correlations can now be taken into account in future studies of conceptual history. SCoT was recently updated with this synoptic analysis function, which is now available for public use. Together with the possibility of analyzing different corpora with different or identical parameters, SCoT supports the evaluation of analysis results and a comparative workflow.

5.5 Epistemic challenges

Beyond the presented advantages, the Sense Clustering over Time approach also poses a number of epistemic challenges. We will conclude with some methodological remarks concerning the digital hermeneutic process – which already starts with the process of semi-automatic graph clustering. This is not static, but non-deterministic. Which means it cannot be reproduced exactly. This in turn does not mean that the result of the clustering process is arbitrary. However, it always contains a certain amount of variability.

Let us take another look at the graph of *kreativ* in the Google Books corpus for demonstration purposes and run it through a series of repeated clustering cycles with identical parameter settings. It may happen, for instance, that some nodes, such as *konstruktiv* and *innovativ*, which previously belonged to the blue cluster (Fig. 5), become part of the green cluster (Fig. 11). This is because they maintain closer connections with both. By lowering the density value D, a much larger number of clusters can be created, i.e. even finer structural-semantic differences can be made visible which at some point may also lead to a fragmentation or partitioning of the graph (Fig. 12), or on the contrary, all differences can be blurred by drastically increasing the D value which will, at a certain point, lead to a ‘big blob’ telling us nothing interesting anymore. Which leads to the question of the right setting of the parameters.

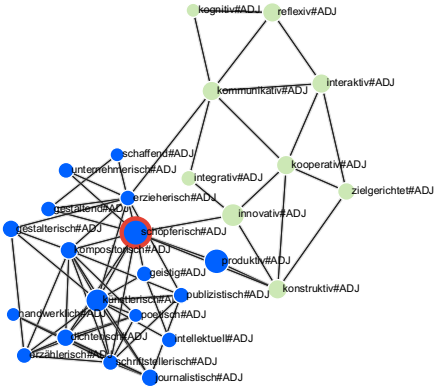

Fig. 11. graph with density D=15

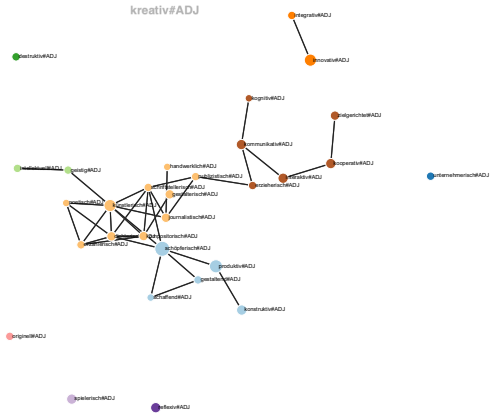

Fig. 12. same graph with density D=8

The question of the right setting is also a question of what 'right' can or should mean in this context. This question cannot be answered in a scientific manner, despite all the mathematics involved. It requires a critical judgement on the part of the researcher, who sets the parameters based on his prior understanding and epistemic interest – thereby further driving the hermeneutic circle. A fundamental heuristic can be applied, though, when searching for the 'right' parameters. This is, for instance, Rule No. 1: *Start with a high density value and lower it until at least two clusters appear for the first time. They should represent the most important difference in the semantic network.* Then follow Rule No. 2: *Repeat the re-clustering cycle several times and observe if the difference occurs on a regular basis. If so, then the difference is stable and thus significant. Now you can start with your interpretation of the senses.*

Once the analysis yields a satisfactory result, it can be saved. SCoT offers the capability to store visual representations as image files in SVG or JPG formats. Each graph can also be saved as a JSON file and reloaded later for future analysis or evaluation, ensuring reproducibility.

Regarding its epistemic challenges, the process of sense cluster analysis aligns in many ways with Rheinberger's concept of an experimental system [57], not only in terms of data and sources [21] but also, and even more so, regarding parameter manipulation. In this context, manipulation is not something to be avoided or debunked; rather, it is an integral part of the experimentation process. In this sense, SCoT functions as an experimental system, with the historical semantics of concepts as its epistemic object. The interaction of the researcher with the non-deterministic experimental system inevitably becomes part of the hermeneutic circle. It is precisely this circle between automatic induction, historical interpretation and parametric reconfiguration that characterizes the digital hermeneutic of SCoT.

In the next and last section, we will discuss the degree to which this parameter-related hermeneutic process is reflected by the DRM@DH model as presented in this volume.

5.6 SCoT in the context of the DRM@DH model

Digital hermeneutics can be understood as a digital or hybrid approach to analog sources [16]. However, when it comes to big data research tools, things become even more complex. By conceptualizing SCoT in terms of the DRM@DH model, it becomes evident that processes such as searching, filtering, selecting, sorting, mapping, and classifying, as well as parameterizing, testing, comparing, and evaluating 'data as capta' [19], initiate hermeneutic operations of different levels. These operations are integral to an experimental system where data visualizations oscillate between serving as technical (or: virtual) objects and functioning as epistemic things. During the analysis process, data visualizations serve as indicators, semiotic traces, or statistical evidence of the research object—in this instance: the historical semantics of a concept. However, at another stage of research and development, they transform into epistemic things that require examination, testing, evaluation, and validation. Digital experimental systems like SCoT make this transition especially evident during the research process itself and raise additional epistemological questions concerning digital hermeneutics.

The DRM@DH model seems particularly well-suited for hermeneutic research tools that are based on a qualified pre-selection and annotation of research data which are to be transformed from 'capta' into 'arguments.' Provided the model aims to describe workflows that also include big data analysis tools based on topic modeling, vector spaces, word sense induction or other advanced methods, it must adeptly capture the hermeneutic processes involved in computing, evaluating, and experimenting with data. This includes the aggregation and processing of data during the development and setup phase as well as the setting and adjustment of parameters for the calculation and visualization of data within the experimental system. 'Data transformation' in this context is therefore not only 'transcription' and 'description', but also *distributing*, *weighting* and *calculation*. Likewise, 'visual mappings' in this case do not amount to

'annotations', but to *render graphs* based on specific *parameter settings*. The graphs and their clusters can be annotated in a further step of data analysis and be compared with results achieved by different parameter settings. Accordingly, 'view transformation' and 'configuration' amount to *parameterization*, *experimentation* and *evaluation*. In order to take into account the transition of graphs from technical (virtual) objects to epistemic things – which can, and often must, go through several transitory cycles – dedicated phases of data *manipulation*, *testing* and *proofing* would have to be incorporated in relation to all transformation steps of the DRM@DH model.

To adequately represent these workflow steps, which are part of the hermeneutic circle in place, it is important to consider that the use of big data analysis tools typically does not commence with specific text documents to be collected and selected through conventional research and reading techniques conducted and overseen by human experts. Instead, it starts with ready-made, highly pre-processed, aggregated text data compilations to be queried with specific parameter settings that will result in visual representations. The discovery and identification of individual text documents for further reading are more likely to occur at the end of the work process with big data tools.

In turn, meaningful analyses at distant reading level require methodically controlled evaluation steps, both in the visual representations and the underlying data structures, e.g. the Distributional Thesaurus or the word embeddings. However, the criteria for evaluation utilized by users and developers of these tools are not identical, and both sets of criteria differ from those applied to the assessment of transcriptions or annotations of text documents, images, or multimedia files. Consequently, they all play different roles within the hermeneutic processes. This difference deserves careful consideration. For while in the case of annotation-based study of small corpora researchers are in most cases in control of the compilation and transformation of the sources, users of big data text analysis tools often depend on the support of or collaboration with language technology experts. Therefore, a continuous interdisciplinary exchange with the developers of such tools or a certain methodological education is essential in the case of advanced digital methods in the research of historical semantics [55]. This requirement should be adequately considered in a workflow-oriented development and therefore also in a comprehensive model for digital hermeneutic tools.

6 Conclusion

Up to date there are only few open-source and open-access online applications available that are dedicated for the research in historical semantics that are easily accessible to researchers working in the history of concepts. Prominent tools within the community, such as Google Ngrams Viewer and DiaCollo, have gained popularity due to their good accessibility and the methodologies they utilize. They allow for a relatively good integration of data analysis results in historical research methods even without advanced NLP expertise, thanks to the methodical transparency of the frequency and collocation analysis. Conversely, tools designed for more complex analyses, such as topic modeling or word embeddings, pose considerable challenges related to both technical and methodological accessibility and transparency. While some of these challenges can

be overcome with sufficient resources (knowledge, computing power, data, time, money), the problem of lacking transparency remains especially in the case of word embeddings which have become more popular for research in historical semantics recently.

SCoT is an open-access online tool specifically developed to support digital research in the history of concepts with a Word Sense Induction (WSI) approach. A key advantage of the WSI approach for historical research is its inherent transparency. It allows for *data-driven* detection and *visualization* of ranked word senses and their evolution over time by calculating context-based similarity values within time-stamped text corpora. This method enables researchers to identify and analyze word senses without the need for predefined dictionaries, thesauri, or training vector-based language models. While the latter also enable context-based representation of word meanings, they lose information about the corresponding contexts. In contrast, SCoT retains this context information, making it accessible for researchers and allowing them to trace similarity scores of word senses back to specific text passages. By maintaining *transparency* and *traceability*, SCoT thus enables a comprehensive scalable reading approach.

Another unique key feature of SCoT is the diachronic representation of different word senses. While other tools enable diachronic representations of word frequencies, contexts, similarities or other, more complex measures for the research in the history of concepts and ideas, SCoT can visualize different word senses (including polysemy) and their change over time, thus featuring *diachroneity* and *sensitivity*. By providing analytical functions to compare frequency values and similarity scores over time in a synoptic view, SCoT allows a detailed and *versatile evaluation* of analysis results. Options for modifying the queries and display parameters ensures *adaptability* and the save and load function for analysis results ensures *reproducibility*.

The most significant limitation of SCoT is that it demands vast amounts of data, substantial computing power, and a relatively high level of training in the field of language technology from its users. Without adequate training or interdisciplinary cooperation, methodological uncontrolled interpretation of visualizations of highly aggregated data can actually turn into an epistemological obstacle. These limitations make SCoT less accessible compared to other tools that have gained traction within the research community. However, to support *accessibility*, the design and documentation of the open-source online tool are structured in a workflow-oriented way to make the application and understanding of SCoT as inviting and convenient as possible. Further improvements are possible and planned. To compensate for restrictions on access in terms of data and computing power, all text corpora implemented for SCoT analyses to date have been made publicly available online for research purposes.

In conclusion, SCoT sufficiently fulfils the proposed requirements for digital research tools in the field of conceptual history, as outlined in Section 4. With its WSI approach, it offers valuable benefits for digital research in historical semantics by providing unique functionalities and thereby usefully complementing existing methods.

Challenges remain in the more precise description and reflection of the epistemological and hermeneutical implications of using SCoT as a digital experimental system. A further development of a general model for digital hermeneutics research tools could support this inherently interdisciplinary task.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Funding

This work was supported by the Leibniz Collaborative Excellence program as part of the project "The 20th Century in Basic Concepts. A Dictionary of Historical Semantics in Germany".

References

1. Andrews, P. et al.: A classification of semantic annotation systems. Semant Web. 3, 3, 223–248 (2012).
2. Anonymus: Ruhm des Hinkens. Gleich zwei neue Bücher sind zum Thema Kreativität erschienen. These: Schöpferisch kann jeder. In: Der Spiegel, 29.09.1974, Nr. 40, 175-177, https://www.spiegel.de/spiegel/print/index-1974-40.html (1974).
3. Bartunov, S. et al.: Breaking Sticks and Ambiguities with Adaptive Skip-gram. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. pp. 130–138 PMLR (2016).
4. Biemann, C. et al.: A framework for enriching lexical semantic resources with distributional semantics. Nat. Lang. Eng. 24, 2, 265–312 (2018). Doi:1017/S135132491700047X.
5. Biemann, C.: Chinese Whispers – an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing. pp. 73–80 Association for Computational Linguistics (2006).
6. Biemann, C.: Co-Occurrence Cluster Features for Lexical Substitutions in Context. In: Banea, C. et al. (eds.) Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing. pp. 55–59 Association for Computational Linguistics, Uppsala, Sweden (2010).
7. Biemann, C. et al.: JoBimText Visualizer: A Graph-based Approach to Contextualizing Distributional Similarity. In: Kozareva, Z. et al. (eds.) Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing. pp. 6–10 Association for Computational Linguistics, Seattle, Washington, USA (2013).
8. Biemann, C. et al.: Wissensrohstoff Text: eine Einführung in das Text Mining. Springer Vieweg, Wiesbaden, Heidelberg (2022). Doi:10.1007/978-3-658-35969-0.
9. Biemann, C., Riedl, M.: Text: now in 2D! A framework for lexical expansion with contextual similarity. J. Lang. Model. 1, 1, 55–95 (2013). https://doi.org/10.15398/jlm.v1i1.60.
10. Bierganns, M.: Kreativität. In: Müller, E. et al. (eds.) Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland. Schwabe Verlag, Basel, Berlin (2024). Doi: 10.31267/Grundbegriffe_65423837
11. Bjerva, J. et al.: Semantic Tagging with Deep Residual Networks. In: Matsumoto, Y. and Prasad, R. (eds.) Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 3531–3541 The COLING 2016 Organizing Committee, Osaka, Japan (2016).
12. Brennan, R. et al.: Building the Seshat Ontology for a Global History Databank. In: Sack, H. et al. (eds.) The Semantic Web. Latest Advances and New Domains. pp. 693–708 Springer International Publishing, Cham (2016). Doi:10.1007/978-3-319-34129-3_42.

13. Brody, S., Lapata, M.: Bayesian word sense induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. pp. 103–111 Association for Computational Linguistics, USA (2009).
14. Burckhardt, D. et al.: Distant Reading in der Zeitgeschichte. Möglichkeiten und Grenzen einer computergestützten Historischen Semantik am Beispiel der DDR-Presse. Zeithistorische Forschungen - Stud. Contemp. Hist. 16, 1, 177–196 (2019). Doi:10.14765/ZZF.DOK-1345.
15. De Bolla, P. et al.: Distributional Concept Analysis. Contrib. Hist. Concepts. 14, 1, 66–92 (2019). Doi:10.3167/choc.2019.140104.
16. De Bolla, P. ed: Explorations in the digital history of ideas: new methods and computational approaches. Cambridge University Press, Cambridge, New York (2024).
17. Devlin, J. et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J. et al. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). pp. 4171–4186 Association for Computational Linguistics, Minneapolis, Minnesota (2019). Doi:10.18653/v1/N19-1423.
18. Dorow, B., Widdows, D.: Discovering corpus-specific word senses. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics – Vol. 2. pp. 79–82 Association for Computational Linguistics, USA (2003). Doi:10.3115/1067737.1067753.
19. Drucker, J.: Humanities Approaches to Graphical Display. Digit. Humanit. Q. 5, 1, (2011).
20. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. Universität Stuttgart (2005).
21. Fickers, A.: Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik? Zeithistorische Forschungen – Stud. Contemp. Hist. 17, 1, 157–168 (2020). Doi:10.14765/zzf.dok-1765.
22. Friedrich, A.: Netz, Netzwerk, Vernetzung. In: Müller, E. et al. (eds.) Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland. Schwabe, Basel, Berlin (2024). Doi:10.31267/Grundbegriffe_13707165.
23. Friedrich, A. et al.: Tools und Korpora für 'Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland.' Digitale Begriffsgeschichte mit COSMAS II, DiaCollo und SCoT. In: Müller, E. et al. (eds.) Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland. Schwabe, Basel, Berlin (2024). (2024). Doi:10.31267/Grundbegriffe_66737084.
24. Friedrich, A., Biemann, C.: Digitale Begriffsgeschichte? Methodische Überlegungen und exemplarische Versuche am Beispiel moderner Netzsemantik. Forum Interdiszip. Begr. FIB. 5, 2, 78–96 (2016).
25. Gavin, M. et al.: Spaces of Meaning: Conceptual History, Vector Semantics, and Close Reading. In: Gold, M. and Klein, L. (eds.) Debates in the Digital Humanities 2019. pp. 243–267 University of Minnesota Press, Minneapolis, London (2019).
26. Gavin, M.: Vector Semantics, William Empson, and the Study of Ambiguity. Crit. Inq. 44, 4, 641–673 (2018). Doi:10.1086/698174.
27. Geyken, A. et al.: Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache" (DWDS). Z. Für Ger. Linguist. 45, 2, 327–344 (2017). Doi:10.1515/zgl-2017-0017.
28. Götzelmann, M. et al.: The Historical Semantics of Temporal Comparisons Through the Lens of Digital Humanities. In: Schwandt, S. (ed.) Digital methods in the humanities: challenges, ideas, perspectives. pp. 269–307 transcript, Bielefeld University Press, Bielefeld (2021).

29. Graf, R.: Energie. In: Müller, E. et al. (eds.) Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland. Schwabe, Basel, Berlin (2024). Doi: 10.31267/Grundbegriffe_81129735.
30. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. 101, suppl_1, 5228–5235 (2004). Doi:10.1073/pnas.0307752101.
31. Haase, C. et al.: SCoT: Sense Clustering over Time: a tool for the analysis of lexical change. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. pp. 198–204 Association for Computational Linguistics, Online (2021). Doi:10.18653/v1/2021.eacl-demos.23.
32. Harris, Z.S.: Distributional Structure. *WORD*. 10, 2–3, 146–162 (1954). Doi:10.1080/00437956.1954.11659520.
33. Hengchen, S. et al.: A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections. Digit. Scholarsh. Humanit. 36, Supplement_2, ii109–ii126 (2021). Doi:10.1093/llc/fqab032.
34. Heuser, R.: Computing Koselleck: Modelling Semantic Revolutions, 1720–1960. In: De Bolla, P. (ed.) Explorations in the digital history of ideas: new methods and computational approaches. pp. 256–301 Cambridge University Press, Cambridge, New York (2024).
35. Ifversen, J.: The Birth of International Conceptual History. Contrib. Hist. Concepts. 16, 1, 1–15 (2021). DOI:10.3167/choc.2021.160101.
36. Jurish, B.: Diachronic collocations, genre, and DiaCollo. In: Whitt, R.J. (ed.) Diachronic corpora, genre, and language change. pp. 41–64 John Benjamins, Amsterdam (2018).
37. Jurish, B., Nieländer, M.: Using DiaCollo for Historical Research. Presented at the Introduction July 3 (2020). https://doi.org/10.3384/ecp2020172005.
38. Kahmann, C. et al.: Detecting and Assessing Contextual Change in Diachronic Text Documents using Context Volatility: In: Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. pp. 135–143 SCITEPRESS - Science and Technology Publications, Funchal, Madeira, Portugal (2017). Doi:10.5220/0006574001350143.
39. Kempfert, I. et al.: Digital History of Concepts: Sense Clustering over Time. Presented at the 42. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS) , Universität Hamburg 6.03 (2020).
40. Kleymann, R. et al.: Conceptual Forays: A Corpus-based Study of "Theory" in Digital Humanities Journals. J. Cult. Anal. 7, 4, (2022). Doi:10.22148/001c.55507.
41. Koselleck, R.: Einleitung. In: Brunner, O. et al. (eds.) Geschichtliche Grundbegriffe: historisches Lexikon zur politisch-sozialen Sprache in Deutschland. p. XIII–XXVII Klett, Stuttgart (1972).
42. Kutuzov, A. et al.: Contextualized embeddings for semantic change detection: Lessons learned. North. Eur. J. Lang. Technol. 8, 1, (2022). Doi:10.3384/nejlt.2000-1533.2022.3478.
43. Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2. pp. 768–774 Association for Computational Linguistics, Montreal, Quebec, Canada (1998). Doi:10.3115/980691.980696.
44. Liu, V., Curran, J.R.: Web Text Corpus for Natural Language Processing. In: McCarthy, D. and Wintner, S. (eds.) 11th Conference of the European Chapter of the Association for Computational Linguistics. pp. 233–240 Association for Computational Linguistics, Trento, Italy (2006).
45. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Mass (2000).

46. Marjanen, J.: Quantitative Conceptual History: On Agency, Reception, and Interpretation. Contrib. Hist. Concepts. 18, 1, 46–67 (2023). Doi:10.3167/choc.2023.180103.
47. Michel, J.-B. et al.: Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 331, 6014, 176–182 (2011). Doi:10.1126/science.1199644.
48. Mikolov, T. et al.: Efficient Estimation of Word Representations in Vector Space, http://arxiv.org/abs/1301.3781, (2013). Doi:10.48550/arXiv.1301.3781.
49. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Lang. Cogn. Process. 6, 1, 1–28 (1991). Doi:10.1080/01690969108406936.
50. Moretti, F.: Distant Reading. Verso, London, New York (2013).
51. Müller, E.: Aufklärung/Gegenaufklärung. In: Müller, E. et al. (eds.) Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland. pp. 1–10 Schwabe, Basel, Berlin (2024). Doi:10.31267/Grundbegriffe_13772469
52. Müller, E. et al.: Einleitung. In: Müller, E. et al. (eds.) Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland. pp. 1–10 Schwabe Verlag, Basel, Berlin (2024). Doi:10.31267/Grundbegriffe_77012495.
53. Müller, E., Schmieder, F.: Begriffsgeschichte und historische Semantik: ein kritisches Kompendium. Suhrkamp, Berlin (2016).
54. Newman, D.J., Block, S.: Probabilistic topic decomposition of an eighteenth-century American newspaper. J. Am. Soc. Inf. Sci. Technol. 57, 6, 753–767 (2006). Doi:10.1002/asi.20342.
55. Oberbichler, S. et al.: Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians. J. Assoc. Inf. Sci. Technol. 73, 2, 225–239 (2022). Doi:10.1002/asi.24565.
56. Pääkkönen, J., Ylikoski, P.: Humanistic Interpretation and Machine Learning. Synthese. 199, 1, 1461–1497 (2021). Doi:10.1007/s11229-020-02806-w.
57. Rheinberger, H.-J.: Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube. Stanford University Press, Stanford, CA (1997).
58. Schreibman, S. et al. eds: A Companion to Digital Humanities. Blackwell Pub, Malden, MA (2004).
59. Schwandt, S. ed: Digital methods in the humanities: challenges, ideas, perspectives. transcript, Bielefeld University Press, Bielefeld (2021).
60. Schwandt, S.: Digitale Methoden für die Historische Semantik: Auf den Spuren von Begriffen in digitalen Korpora. Gesch. Ges. 44, 1, 107–134 (2018). Doi:10.13109/gege.2018.44.1.107.
61. Sevgili, Ö. et al.: Neural entity linking: A survey of models based on deep learning. Semantic Web. 13, 3, 527–570 (2022). Doi:10.3233/SW-222986.
62. 62. Steck, H. et al.: Is Cosine-Similarity of Embeddings Really About Similarity? In: Companion Proceedings of the ACM Web Conference 2024. pp. 887–890 ACM, Singapore (2024). Doi:10.1145/3589335.3651526.
63. Templeton, C.: Topic Modeling in the Humanities: An Overview, http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/, last accessed 2018/01/13.
64. Wevers, M., Koolen, M.: Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings. Hist. Methods J. Quant. Interdiscip. Hist. 53, 4, 226–243 (2020). Doi:10.1080/01615440.2020.1760157.
65. Willkomm, J. et al.: CH-Bench: a user-oriented benchmark for systems for efficient distant reading (design, performance, and insights). Int. J. Digit. Libr. 24, 4, 243–261 (2023). Doi:10.1007/s00799-023-00347-4.
66. Zhou, K. et al.: Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words. In: Muresan, S. et al. (eds.) Proceedings of the 60th Annual Meeting of the

Association for Computational Linguistics (Vol. 2: Short Papers). pp. 401–423 Association for Computational Linguistics, Dublin, Ireland (2022). Doi:10.18653/v1/2022.acl-short.45.