Topics in Periodicals from the Swedish Diabetes Association 1949 – 1990: Extending the Topic Modelling Tool Topics2Themes with a Timeline Visualisation

Maria Skeppstedt[1], **Gijs Aangenendt**[1,2], **Vera Danilova**[2], **Ylva Söderfeldt**[2]
[1]Centre for Digital Humanities and Social Sciences Uppsala,
Department of ALM, Uppsala University, Sweden
maria.skeppstedt@abm.uu.se
[2]Department of History of Science and Ideas, Uppsala University, Sweden
{gijs.aangenendt,vera.danilova,ylva.soderfeldt}@idehist.uu.se

Abstract

Existing methods for visualising temporal topic models typically present the information in an aggregated form, and do not offer any possibility to track the specific texts responsible for the change in topic prevalence over time. We present a new type of topic modelling-based timeline visualisation. It still provides an overview with aggregated topic information suitable for distant reading, while also allowing the user to gradually zoom into the image for more detail. At the most detailed level, the individual texts can be reached, which makes it possible to switch to close reading. The timeline visualisation was implemented as an extension of the topic modelling tool Topics2Themes, but this visualisation technique can be adapted to other topic modelling tools and algorithms. We showcase the timeline visualisation on a corpus of periodicals from the Swedish Diabetes Association, which is one of the patient organisation corpora studied within the interdisciplinary project ActDisease. One timeline visualisation was generated for the entire corpus. Additionally, we generated a timeline focusing on the texts that contain the word “dietitian”. The two timelines, including the functionality to zoom into the graphs and reach the texts, were used to analyse the topics and how they vary. It could be concluded that some of the topics and topic timelines were predictable, while others revealed content that might be less expected. These results indicate validity of the method applied, and they also show that this visualisation technique could help us learn something new.

1 Introduction

A text collection covering a longer time period offers a wealth of possibilities for studying temporal change. Among many potential aspects of change, we focus on exploring prominent topics and the longitudinal variation of their prevalence. With the support of natural language processing (NLP) and text visualisation methods, it is possible to carry out such explorations on large text collections as well.

In this paper, we showcase how the NLP technique *topic modelling* can be used to automatically extract topics from a large text collection, and how the topics extracted can be visualised on a timeline. For this aim, we use Topics2Themes, which is a topic modelling tool maintained and further developed at the CLARIN node at the Language Council of Sweden (Skeppstedt et al., 2018). Topics2Themes has previously been applied to many different types of text collections, and the main contribution of the work presented here is an extension of the tool in the form of a timeline visualisation of the tool’s topic modelling output.

The main difference between Topics2Themes and the existing approaches for presenting topic modelling output is that the tool uses topic modelling as a means for selecting and topically sorting texts that might be relevant for a manual analysis. Our timeline visualisation takes the same approach. More specifically, building on the basic information visualisation principle of "Overview first – details on demand", the timeline first provides an overview of how the topics vary over time, and thereafter makes it possible to zoom into the image to gradually reach more detailed information. At the most detailed level, it is possible to access the actual texts upon which the timeline is built. (See Figure 1 for an illustration of this approach.) By providing a topic overview, as well as functioning as a point of departure for locating potentially interesting texts, we aim to create a visualisation that makes it possible to switch between a *distant* reading of temporal topic variation and a *close* reading of the actual texts.

Previous development iterations of this topic modelling-based timeline visualisation (Skeppstedt, 2022, 2023; Stede et al., 2023) had a technical focus and were carried out by groups consisting of only NLP researchers. In contrast, the work presented here was conducted in an interdisciplinary team, more specifically within the historical research project ActDisease that investigates the history of patient organisations in twentieth century Europe. We applied the timeline visualisation on one of the corpora investigated in the project – a collection of digitised periodicals issued by the Swedish Diabetes Association. This enabled us to test to what extent the visualisation was useful in exploring the content of the corpus. In addition to reporting on the results of this interdisciplinary work, we describe the design of the timeline visualisation and discuss the ideas underlying the design choices.[1]

2 Background

Topic modelling is a form of unsupervised technique for automatic text exploration and categorisation. That is, no manually annotated training data or pre-defined categories are fed to the topic modelling algorithm. Instead, patterns in the text collection itself, e.g. word co-occurrence patterns, are used to automatically extract frequently re-occurring topics. Each topic extracted by the algorithm is typically represented by (i) an ordered list of texts in which the topic occurs, and (ii) an ordered list of words prevalent in texts associated with the topic. The texts and words are ordered based on their closeness to the topic. There are a number of different topic modelling algorithms, from more classic ones such as LDA (D. Blei & Lafferty, 2006; D. M. Blei et al., 2003) and NMF (Greene & Cross, 2017; Lee & Seung, 2001) to transformer-based approaches, such as BERTopic (Grootendorst, 2022).

For the specific task of applying topic modelling on temporally extended text collections, a number of different strategies/topic modelling algorithms are available. Dynamic topic models is a frequently used method, which differs from standard topic models in that the timestamps of the texts are included as one of the parameters (D. Blei & Lafferty, 2006; Grootendorst, 2022; Wang & McCallum, 2006). It is also possible to first use standard topic models and then apply additional statistical methods for analysing their temporality (Meaney et al., 2022), or to apply standard topic models on many short time frames and combine those into topics spanning longer time periods (Greene & Cross, 2017). Another possibility is to combine dynamic and standard topic models (Hida et al., 2018), or to use a more straight-forward approach of applying standard topic modelling on a temporal text collection and simply visualise the temporal variation of the topics (Stede et al., 2023).

There are also many examples of approaches developed for the task of visualising topic models applied to temporal data. To the best of our knowledge, however, there is no standard or best practice yet. Examples of visualisation techniques used are line charts (D. Blei & Lafferty, 2006; Grootendorst, 2023), stacked bars (Sheehan et al., 2021), a Sankey diagram (Malik et al., 2013), a heatmap (Meaney et al., 2022) and horizontal trend lines (Gad et al., 2015). There is also the "Theme river" visualisation (Günnemann, 2013; Günnemann et al., 2013; Havre et al., 2000) where coloured "streams" (or "rivers") with changing widths represent topic variation, and (in a more developed format) these streams can split and merge as the topics they represent become more or less semantically close (Cui et al., 2011). Common to

[1] Topics2Themes is a language-independent, open source topic modelling tool, found at:
https://github.com/sprakradet/topics2themes.
The open source programming code for generating the timelines is provided at:
https://github.com/CDHUppsala/topic-timelines.

all these approaches – in their original form – is that they all aim to visualise aggregated topic prevalence over time, i.e. they provide a distant view of the corpus without any possibility to easily access the texts that contain the topics.

Topic modelling is one, among many, NLP methods that have been used in visualisations aimed to support *distant reading* (Jänicke et al., 2015). Distant reading emerged within the field of literary studies as a means of analysing literature at scale, often with the help of computational methods (Moretti, 2013; Underwood, 2017). Since its emergence, the approach has spread to other disciplines within the (digital) humanities and social sciences and is also applied on other forms of text besides literary data (Gelfgren & Drakman, 2022). Distant reading is often used in tandem with *close reading* to identify patterns in large textual datasets that merit further close-up investigation through the reading of individual texts.

Topic models are often criticised for being difficult to interpret due to the numerical and statistical nature of their outputs (Hagen, 2018). Assessing the quality of the extracted topics cannot solely be done based on the statistical output, but often requires domain knowledge and a close reading of the associated texts. Despite this, most topic modelling-based timeline visualisations only provide the user an aggregated view of the corpus. We believe that adding support for close reading could be one way to address this limitation. This hypothesis is supported by the usability evaluation of the Theme river approach, which showed that although the users appreciated the information provided by the aggregated distant view of the corpus, this information was not sufficient. Instead, the need to read the actual texts that contributed to a topic at a given timestamp was recognised (Havre et al., 2000).

The usability evaluation of Theme river also resulted in additional feature requests. These included the ability to see the total number of texts at any time period, and to be able to provide a user-defined ordering of the theme rivers. The original, static Theme river visualisation was later developed, and incorporated in an interactive text visualisation tool called D-Vita (Günnemann, 2013; Günnemann et al., 2013). Here, the timeline is divided into user-defined time periods, e.g. a year, and close reading is made possible by allowing the user to access the most typical texts for a topic for each such time period. The theme river visualisation is also incorporated in another interactive tool, called Tiara (Liu et al., 2009). In this tool, each stream is overlaid with word clouds, and a possibility to zoom into the word clouds using the fisheye view technique is provided. A form of close reading functionality is provided also for this tool, by making it possible for the user to read an automatically selected set of semantically diverse texts that represent the topics.

3 Design objectives and requirements

In addition to the overall design objective of creating a topic modelling-based timeline visualisation, we also had four main design requirements. These were based on studying previous research, as well as on reflections made in the preceding development iterations and on discussions within the team.

(1) The most important design requirement was to *not* create a visualisation limited to showing aggregated information, disconnected from the texts from which the topics are derived. This requirement is also what most clearly sets the visualisation created here apart from the previous examples of temporal topic visualisations described above. Instead, we aimed to employ the standard information visualisation workflow "overview first [...] then details on demand" (Shneiderman, 1996), to make it evident that the high-level/aggregated information has been derived by combining information from individual texts. That is, when the original-size version of the visualisation is presented to the user, aggregated information from many texts should be shown (overview first). It should then be possible to demand more details by zooming into interesting areas of the visualisation, letting representations of the individual texts – and topics that occur in these texts – be shown. Finally, in search for potentially interesting text content, it should be possible for the user to demand even more detailed information, by shifting from exploring the representations of the texts to reading the actual text content. (See Figure 1 for an illustration of this approach.) We do not aim to limit the texts that can be directly accessed to a subset of the texts from which the timeline is created, as has been the case for previous tools that include some support for close reading (Günnemann, 2013; Günnemann et al., 2013; Liu et al., 2009). Instead, all texts for which the occurrence of one or more topics is indicated in the graph should be possible to access.

(2) Another design objective was to avoid showing any concrete numerical values (except timestamps) in the visualisation, or using any visualisation technique that is closely associated with the visualisation of numerical values, such as bar charts or line graphs. The rationale for this objective is that when concrete numerical values are introduced, the importance of their exact meaning might easily be overestimated. For topic modelling, the output generated is more fuzzy than e.g. word count statistics. Not only because many topic modelling algorithms are randomised, but also because the output generated is very sensitive to what configuration parameters are used (Da, 2019, p 625). Topic modelling offers *a* window through which a text collection can be viewed, rather than offering *the* window for viewing it. Also, each numerical output value of the topic modelling algorithms is often not meaningful in itself, but only in relation to other output values. Yet a reason to communicate uncertainty or fuzziness has to do with the nature of the underlying datasets used in humanities research. These datasets often consist of incomplete historical sources and are created through subjective and implicit decisions made during the data collection process, e.g. when selecting what sources to include in the dataset (Panagiotidou et al., 2022). For these reasons, we aimed for a visualisation that would somehow convey a fuzziness, and that would encourage active exploration of the data as well as an interpretation of the topic modelling output in relation to other values in the graph.

(3) It was also important for the resulting visualisation to consist of one, static image. That is, the topic timeline should be displayed in a single static graph, without using any form of dynamic functionality to convey the information required to explore the topic modelling output. The only dynamic element of the exploration should consist of the user zooming in and out in the image. This restriction enforces a simplicity (and thereby hopefully also an increased usability) of the timeline design. In addition, we believe that if a visualisation can be included as a zoomable image in an article or as a large image on a printed poster, it lowers the threshold for using and re-using it. Although an interactive graphical user interface is needed for some types of text exploration tasks, we believe that one dimension is lost when the simplicity of a static image is traded for an interactive system. An example of this is the original Theme river visualisation (Havre et al., 2000) in relation to its interactive version (Günnemann, 2013; Günnemann et al., 2013). We, however, employed one important exception to this restriction: To let the user easily switch from exploring the graph to reading the texts upon which the graph is based, we allowed the image to be dynamic in the sense that it can be configured to contain links to web pages where the texts can be read.

(4) The final main design objective consisted of including information about the temporal text frequency in the visualisation. That is, the timeline should visualise the variation in the number of texts that stem from different time periods. This information is interesting in itself, e.g. to inform on variation in text publication frequency or on the temporal prevalence of certain keywords that have been used to extract the text collection. The text frequency also helps the user to interpret the variation in topic prevalence, since the number of texts associated with a topic during a time period is in part dependent on how many texts stem from that period (Da, 2019, p 627). This design objective is also in line with one of the feature requests from the user evaluation of the Theme river approach (Havre et al., 2000).

In addition to the four main design requirements, we also had a number of smaller requirements. It should, for instance, be possible to compare different timelines, e.g. timelines from different corpora or those resulting from different topic modelling configurations. To make visualisations resulting from texts in various languages understandable to an international audience, it must also be possible to translate the topic labels. Building on our own reflections when using the first versions of the timeline, as well as on results from previous studies (Baumer et al., 2017; Havre et al., 2000; Stede et al., 2023), we also saw the usefulness of making it possible to manually categorise and re-order the automatically extracted topics. Finally, previous research has shown the importance of the visualisations being fairly scalable, i.e allowing both few and many topics to be visualised within the same graph (Gad et al., 2015).

4 Implementing the design

The basic components of the design are described in the caption of Figure 1. This figure also includes a symbolic illustration of how the design makes it possible to gradually move from “overview first” to more “details on demand”. Our approach is to (i) let the original-size graph consist of *combinations* of small graphical components that represent the texts and their topic associations, and (ii) make it possible to see *each individual* graphical component by zooming into the graph (**1:H**). For instance, that a topic occurs in a text is indicated by a *vertical bar* (**1:D**) and by a *circle* (1:**G**), both with a size proportional to the strength of the text-topic association. In the original-size graph, the combination of partly overlapping circles provides an overview of topic prevalence, while zooming in makes it possible to see each vertical bar (**1:D**) and each circle (1:**G**) that represents the topic-strength for the text.

The most detailed level, i.e. the text itself, can not be reached by zooming into the static image. We therefore implemented the possibility to associate a unique HTML link to each of the texts represented by a bar/circle in the graph. When clicking on the circle (**1:I**), the web page associated with the text is opened, e.g. a web page that contains the original text with the original page layout (**1:J**). This overview first-details on demand approach is our solution for how to make it possible to switch between distant and close reading.

In addition to these two most important features, i.e. to be able to **gradually zoom in for more detail** and **click in the graph to reach the actual texts**, we also provided several other configuration options to meet the design requirements.

One configuration regards **texts that share the same timestamp**. That is, to be able to plot texts on a timeline, each text must be provided with a timestamp. If several texts share the same timestamp, they will be plotted on the exact same position in the graph, obscuring each other. We therefore implemented a configuration possibility that spreads texts with the same timestamp along the x-axis. The configuration lets the user specify the following three pieces of information: (i) A small time fraction indicating the distance by which to move the x-position of a text when another text is already positioned at this timestamp, (ii) the width of the vertical bar indicating the topic-strength for the text, and (iii) the transparency of the topic-strength indicators. By configuring these three parameters, it is possible to use overlapping transparent circles to indicate topic-strength for each individual text. This results in a pattern of partly overlapping, transparent circles, which will not only show topic-strength variations for texts with the exact same timestamp, but also for texts that are positioned very close to each other on the x-axis (**1:H**).

Another configuration option regards **how to scale the vertical bar** (and circle) that indicates topic-strength. The height is scaled to make sure the tallest bar fits within the horizontal lane that represents the topic (**1:L**). This scaling can either be configured to be performed on a graph-global level, using the overall maximum topic-strength as the factor with which to scale the bar. It can also be carried out on a topic-local level, using the maximum topic-strength for each specific topic in question as the scaling factor. The first option makes it possible to compare association strengths between topics, but it also makes it difficult to see temporal variations for topics with weaker text associations.

Finally, a configuration which made it possible to **manually order the topics**, and/or to create groups of topics, was implemented.[2] The original topic sorting is still shown by the number associated with the topic, but the topics are resorted based on the user input. The user-defined groups of similar topics are indicated by alternating colours (**1:M**).

5 Applying the timeline visualisation on a corpus

As mentioned in the introduction, the timeline visualisation development presented here was conducted within an interdisciplinary team, as part of the historical research project ActDisease. The project investigates the history of patient organisations in twentieth century Europe, and has digitised a number of patient organisation periodicals from four different countries (Aangenendt et al., 2024). To develop and apply the timeline within the context of a historical research project made it possible to (i) add functionality to the timeline visualisation that would be immediately useful to historical research, and (ii) apply the timeline visualisation on a corpus upon which research had already been carried out using traditional historiographical methods. We chose the journal *Diabetes*, published by the Swedish Diabetes

[2] Practically, it was implemented by providing a parameter to the timeline generation function, which takes the form of a nested list with numbers associated with the topics.

Association.[3]

This corpus contains 8 891 pages from 233 individual issues, covering the period 1949-1990 (Aangenendt et al., 2024). The raw files for the corpus derive from scans made by Gothenburg University Library, available in full through GUPEA[4].

In addition to applying the timeline visualisation on the entire corpus, we also applied it on a subset. This subset was selected to be more focused on one specific issue important to the organisation, and only included text data from those pages that contain the word "dietitian". The dietitian profession was introduced in Sweden in the 1960s through a process that the SDA was heavily involved in. From the early 1950s, when they first started using the term, until the 1980s when the role was a clearly defined healthcare profession, the SDA participated in defining and negotiating the boundaries and position of the dietitian. Since this process is one that we had studied extensively through close reading of the corpus and archival material, a dietitian sub-corpus allowed us to view the topic model visualisation from a standpoint of familiarity with the text it represents.

The Diabetes corpus is organised with pages as the text unit and the same unit was used for topic modelling. The texts were lemmatised using Efselab[5] (Östling, 2018), and thereafter the topic modelling tool Topics2Themes (Skeppstedt et al., 2018) was applied on the corpus. We used the Swedish stop word list included in NLTK (Bird, 2002), which we expanded by iteratively running the topic modelling tool on the corpus and inspecting the output. Topics2Themes provides an interactive graphical user interface, which can be used to inspect the output, e.g. regarding which words should be added to the stop word list. In the final iteration, we used the functionality in Topics2Themes where the topic modelling algorithm is run several times, and only stable topics are retained[6], an approach previously used by Baumer et al., 2017. We instructed the topic modelling algorithm to return 70 topics for the full corpus and 30 topics for the sub-corpus with texts containing the word "dietitian". This resulted in 51 stable topics being generated for the full corpus and 21 topics for the sub-corpus. We configured the tool to extract the 20 most typical words for each topic. For the texts, we used a cut-off of the 200 most closely associated texts in the visualisation, provided that the text contained at least one of the top 20 most typical words.

We automatically assigned the timestamp based on the issue in which the text appeared, assuming that the publication dates for the issues were evenly distributed over the year. All texts from one issue were thus assigned the same timestamp, but were configured to be slightly moved along the x-axis when visualised, in order not to collide. The texts within an issue were ordered according to page number, resulting in the first page being positioned on the timestamp for the issue and subsequent pages being slightly moved to the right. A local scaling of the topic strengths was applied, which makes it easier to see temporal variations for a topic, but more difficult to compare topic strengths between different topics. When clicking on the topic-strength symbols, the visualisation was configured to direct the user to a web page containing a pdf with the original page layout (see Figure **1:J**). English translations for the topic labels were obtained by applying Google translate on the Swedish labels and then manually correcting the automatic translations.

6 Exploring the timeline visualisations

As a first step in using the timeline visualisations created, the topics were manually combined into larger groups based on the topic labels. For the full corpus, five groups were created, and there were three topics not included in any of these groups. Five groups were also created for the dietitian-subset, and one topic could not be assigned to any of the five groups. The timelines using these groups are shown in Figure 2 (for the full corpus) and in Figure 3 (for the sub-corpus containing the word "dietitian"). The grouping of topics, as well as the exploration of the timeline described below, was carried out by the historian leading the project, who through close reading of the corpus and archival material had the knowledge required to

[3] *Svenska Diabetesförbundet* in Swedish (*Riksförbundet för sockersjuka* before 1956).

[4] https://gupea.ub.gu.se/handle/2077/64597

[5] https://github.com/robertostling/efselab

[6] More specifically, the topic modelling algorithm was run 50 times, and from these re-run outputs we kept the five most typical ones, and only retained topics that occurred in all five outputs. Similarity between topics was measured based on their associated words, and two topics with a 60% overlap of the top 20 most similar words were counted as the same topic.

assess the information visualised.

The difference between the timeline characteristics for different topics is immediately evident when looking at the original-size graphs. E.g. in Figure 2, it can be seen that some topics occur more or less the entire time period studied, e.g. the first two topics in the image, while others are limited to certain time periods – e.g. topic 49 – or to only having strong occurrences in certain time periods, e.g. topics 16 and 10. There are topics that occur with regular intervals, e.g. topics 3 and 6, while others do not show any obvious regularity, e.g topics 51 and 13. By zooming in, it is possible to see that some of the topics are represented with a single transparent circle per aggregated text-line, which shows that these topics occur only once per issue for a period of time, e.g. topics 2, 16 and 17. Zooming in further shows that topic 2 generally occurs towards the end of the publication and topic 17 in the beginning, while 16 at first occurs in the beginning and then later in time towards the end of the publications. Other topics, in contrast, have opaque segments created by many, partly overlapping circles positioned close to each other, which shows that the topic occurs several times in an issue, e.g. topics 6, 19 and 21. The two semantically similar topics 6 and 28 seem to be the same topic, since they have very similar timeline characteristics, and since topic 28 replaces topic 6, around 1979. As expected, the vertical lines that represent the texts and their timestamps occur with a regular frequency in Figure 2, except for a short time period in the beginning of the 1950s, when no SDA periodicals were published. In Figure 3, in contrast, the frequency of the text-lines vary, corresponding to a variation in how often there are texts published that contain the word “dietitian”.

Exploring the visualisation of topics in a familiar corpus makes it possible to evaluate the extent to which it captures known trends in the material. In Figure 3, topic 16 relates to the training of dietitians, an issue that the SDA pioneered and were instrumental in implementing. Their campaign to establish the dietitian as a new profession in Swedish healthcare began around 1960, and led to a trial course in 1969, which was followed by repeated efforts for a permanent solution. Finally, in 1978, a higher education program for dietitians started (Söderfeldt, 2024). Here, the visualisation closely aligns with the period that this issue was prominent in the SDA, and shows spikes in those phases when important steps took place.

Grouping the topics in categories makes it possible to see trends within thematic fields. For instance, the topics 2-51 in Figure 2 are related to diabetes treatment. It can be seen how certain manufacturers of insulin, glucose tests, and injection material come and go over time (e.g. 5, 16, 33) whereas other appear more stable (e.g. 42 and 51, regarding fundamental pathophysiology and treatment). Similarly, the group 4-50, which relates to food, has different brands of artificial sweeteners and sugar-free foods that appear for particular, limited time periods (e.g. 18, 30, 34), but also some more stable topics like 4, 10, and 39 that are more generally related to diets. By grouping topics, we are able to view a chronological map of a particular theme in the corpus, which provides an initial orientation in overall trends. An interesting find are also topics that were not related to particular themes, such as topic 9 in Figure 2, which represents texts with a certain type of language. It relates to texts dealing with personal experience, which show a remarkable increase in the 1980s.

In the case of dietitian education (topic 16 in Figure 3), the visualisation did not allow us to make new discoveries in the corpus. Rather, it served to confirm that the method depicts meaningful trends. Similarly, topic 46 in Figure 2, relating to nursing (and dietitian) education, is in zenith from the early 1970s to the early 1980s, during which time the SDA also engaged in nursing education by arranging annual courses on diabetes care. Topic 7 in Figure 2 regards the SDA owned and operated retreat centre Nordanede, which came to them through a donation in 1963 and was sold in 1984. Also topic 49 corresponds to known events, since insulin pumps, which were first developed in 1974, hardly appear at all until 1982 and then quickly become a frequent topic. Other topics, however, give indications of change that we had not previously detected and point towards interesting fields of further study. Using the feature of zooming in and accessing the texts, we can determine that the more transient treatment-related topics in Figure 2 (5, 16, 33) are found in advertisements that are repeated over a certain period of time. The more stable topics (42 and 51), however, are drawn from a mix of both editorial and advertising content. A close reading of the texts from these topics can therefore help in characterising the interplay between the way that the patient organisation discussed their illness, and how pharmaceutical companies marketed their products. Findings like these provide helpful directions for qualitative research in the corpus texts as well as supplementary sources like archival material. In the thematic group related to the association (topics 3-45 in Figure 2) we see a trend in that many of these administrative topics taper off in the 1980s. It is crucial to note that changes like these could be due to changes in layout and format, for instance title pages or other elements that are frequently repeated in the corpus. Therefore, the possibility to directly zoom in on the texts that the topics appear in offer a more transparent solution that allow the historian to not only view trends, but to investigate what they consist of. Sampling some of the text pages from this thematic field, we find that several topics are indeed drawn from elements like address lists (6, 28) or front matter (29). However, topics 3 and 19, both relating to annual meetings, reveal a drastic change in how these matters are presented in the periodical not just quantitatively, but also in style. From using the publication as an outlet for meeting reports, the significance of internal congresses not only decreases, but also becomes more journalistic than administrative. This gives indications to cultural changes in the organisation.

7 Final discussion and future directions

A method previously suggested for evaluating the validity of topic models is to investigate whether the model is able to extract known topics (Da, 2019, p. 628). If that is the case, also previously unknown topics and trends indicated by the model might potentially be interesting to explore. The timeline visualisation created helped us to not only detect such known topics, but also to investigate to what extent the prevalence of these topics corresponded to known historical events. The timeline and the possibility to read the actual texts giving rise to the topic-trends was equally useful in helping us to interpret unknown topics.

We are not aware of any previous approaches for the topic-timeline visualisation task that offer the same possibilities for text and topic exploration. This applies both to the functionality which allows the user to click on the timeline to access the actual texts, as well as to the level of expressiveness when showing aggregated topic prevalence in the original-size timeline visualisation. None of the approaches described in the background makes it possible to display this type of timeline characteristic for each topic, i.e. the characteristic that we achieve by letting partly overlapping transparent circles indicate the interplay between topic strength and topic prevalence. Our approach also scales well to displaying a timeline for many topics, and fulfils the requirement to visualise how the text frequency varies over the time period studied. The use of partly overlapping transparent circles as topic strength indicators does not only result in expressiveness, but we also believe – albeit rather subjectively – that they help us communicate the fuzziness we aim for.

Despite this practical usefulness of the timeline visualisation created, there are still many potential improvements for future development iterations. The approach used here to direct the user to a pdf that contains the original page layout – without providing any indications from the topic modelling output – has the drawback that the relevance of some texts to the topic is enigmatic to the human reader. The feature provided by, e.g. the graphical user interface of Topics2Themes, to markup words associated with the topic, might provide a help to understand the model better. This would be especially helpful for the type of corpus used here, where the text unit consists of a page, since a page sometimes includes one or more separate texts and it is unclear if the topic has been detected in one, both, or the texts together. Another problem associated with the pdf files that were linked to, was that they lacked the context of neighbouring pages, i.e. it would have been practical to be able to easily reach the previous or following page in the periodical. A problem, which is more related to the actual timeline visualisation, is that labels for the timeline and topics are no longer visible when the user has zoomed in. To facilitate navigation, it might therefore be useful with local indications of timestamps and topics that are provided with a font size small enough to only be visible when the user has zoomed in.

A number of additional features might also be added to the timeline visualisation. One such feature, which could be added alongside the manual categorisation and ordering of the topics, is the option to select an automatic ordering of the topics, by using a hierarchical topic model or a semantic similarity measure of texts or words associated with the topics. Another potentially useful feature would be to allow the researcher to build sub-corpora for close reading, based on texts that contain interesting topics.

Future work could also focus on making the timeline visualisation compatible with additional types of topic modelling algorithms and outputs. The implementation that we currently provide uses the json output format generated by Topics2Themes. However, the timeline design is by no means specific to Topics2Themes and could easily be adapted to another output format. This would also make it easier to compare the usefulness of different types of topic modelling algorithms for creating the topic timeline.

It can also be noted that the focus of the work presented here has been to visualise the *texts*, in particular how the prevalence of the topics occurring in the texts vary over time. We have used the words associated with the topics only as topic-labels in the visualisations developed. To also create a static timeline visualisation of the *words* associated with the topics is an equally relevant task. For instance, it might be interesting to create visualisations that convey the contrast of the fuzziness of the topic modelling output with more objective – but not necessarily fully objective in all aspects (Panagiotidou et al., 2022) – measures, such as the frequency count of the words associated with the topics.

Finally, we intend to perform a more structured user evaluation, which could include the visualisation technique presented here and some of the updates we have suggested. Such an evaluation might, for instance, be performed on the Diabetes corpus – possibly with developed methods for selecting and segmenting texts – or on some of the other corpora investigated within the ActDisease project.

Acknowledgements

We thank Andrew Burchell for his constructive comments that greatly helped to improve the manuscript.

The work described here was conducted within the ActDisease project. ActDisease is funded by the European Union (ERC ActDisease, ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The development of the topic-timeline programming code was also partly supported by InfraVis (the Swedish Research Council, grant 2021-00181) and Huminfra (the Swedish Research Council, grant 2021-00176). The Topics2Themes tool is maintained by Swe-CLARIN/The National Language Bank of Sweden (the Swedish Research Council, 2017-00626).

References

Aangenendt, G., Skeppstedt, M., & Söderfeldt, Y. (2024). Curating a historical source corpus of 20th century patient organization periodicals. *Proceedings of the Huminfra Conference (HiC 2024)*, 76–82. https://doi.org/10.3384/ecp205011

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, *68*(6), 1397–1410.

Bird, S. (2002). NLTK: The natural language toolkit. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Blei, D., & Lafferty, J. (2006). Dynamic topic models. *ACM International Conference Proceeding Series; Vol. 148: Proceedings of the 23rd international conference on Machine learning; 25-29 June 2006*, 113–120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., & Tong, X. (2011). Textflow: Towards better understanding of evolving topics in text. *IEEE Trans. Vis. Comput. Graph.*, *17*(12), 2412–2421. https://doi.org/10.1109/TVCG.2011.239

Da, N. (2019). The computational case against computational literary studies. *Critical Inquiry*, *45*, 601–639. https://doi.org/10.1086/702594

Gad, S., Javed, W., Ghani, S., Elmqvist, N., Ewing, T., Hampton, K. N., & Ramakrishnan, N. (2015). Themedelta: Dynamic segmentations over temporal topic models. *IEEE Transactions on Visualization and Computer Graphics*, *21*(5), 672–685. https://doi.org/10.1109/TVCG.2014.2388208

Gelfgren, S., & Drakman, A. (2022). How to combine close and distant reading within the history of science and ideas: Two examples from ongoing research. *Lychnos*, (1), 85–108.

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political analysis*, *25*(1), 77–94.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Grootendorst, M. (2023). Dynamic topic modeling, visualization.

Günnemann, N. (2013). D-vita: A visual interactive text analysis system using dynamic topic mining. *Datenbanksysteme für Business, Technologie und Web*. https://api.semanticscholar.org/CorpusID:15848321

Günnemann, N., Derntl, M., Klamma, R., & Jarke, M. (2013). An interactive system for visual analytics of dynamic topic models. *Datenbank-Spektrum*, *13*(3), 213–223. https://doi.org/10.1007/s13222-013-0134-x

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*, *54*(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006

Havre, S., Hetzler, B., & Nowell, L. (2000). Themeriver: Visualizing theme changes over time. *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, 115–123. https://doi.org/10.1109/INFVIS.2000.885098

Hida, R., Takeishi, N., Yairi, T., & Hori, K. (2018, July). Dynamic and static topic model for analyzing time-series document collections. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 516–520). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2082

Jänicke, S., Franzini, G., Scheuermann, G., & Cheema, M. (2015). On close and distant reading in digital humanities: A survey and future challenges. a state-of-the-art (star) report. *Eurographics Conference on Visualization (EuroVis)*.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 556–562.

Liu, S., Zhou, M. X., Pan, S., Qian, W., Cai, W., & Lian, X. (2009). Interactive, topic-based visual text summarization and analysis. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 543–552. https://doi.org/10.1145/1645953.1646023

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., & Shneiderman, B. (2013). Topicflow: Visualizing topic alignment of twitter data over time. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 720–726. https://doi.org/10.1145/2492517.2492639

Meaney, C., Escobar, M., Stukel, T. A., Austin, P. C., & Jaakkimainen, L. (2022). Comparison of methods for estimating temporal topic models from primary care clinical text data: Retrospective closed cohort study. *JMIR medical informatics*.

Moretti, F. (2013). *Distant reading*. Verso Books.

Östling, R. (2018). Part of speech tagging: Shallow or deep learning? *North. Eur. J. Lang. Technol.*

Panagiotidou, G., Lamqaddam, H., Poblome, J., Brosens, K., Verbert, K., & Moere, A. V. (2022). Communicating uncertainty in digital humanities visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 1–11. https://doi.org/10.1109/TVCG.2022.3209436

Sheehan, S., Luz, S., & Masoodian, M. (2021, April). TeMoTopic: Temporal mosaic visualisation of topic distribution, keywords, and context. In H. Toivonen & M. Boggia (Eds.), *Proceedings of the eacl hackashop on news media content analysis and automated report generation* (pp. 56–61). Association for Computational Linguistics. https://aclanthology.org/2021.hackashop-1.8

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. https://doi.org/10.1109/VL.1996.545307

Skeppstedt, M. (2022). The topic modelling tool Topics2Themes applied to different types of climate change-related texts. [The CLARIN Bazaar 2022. https://www.clarin.eu/content/clarin-bazaar-2022].

Skeppstedt, M. (2023). Topics in Swedish news on climate change: A timeline 2016 – 2023. *CLARIN Annual Conference Proceedings 2023*, 150–154.

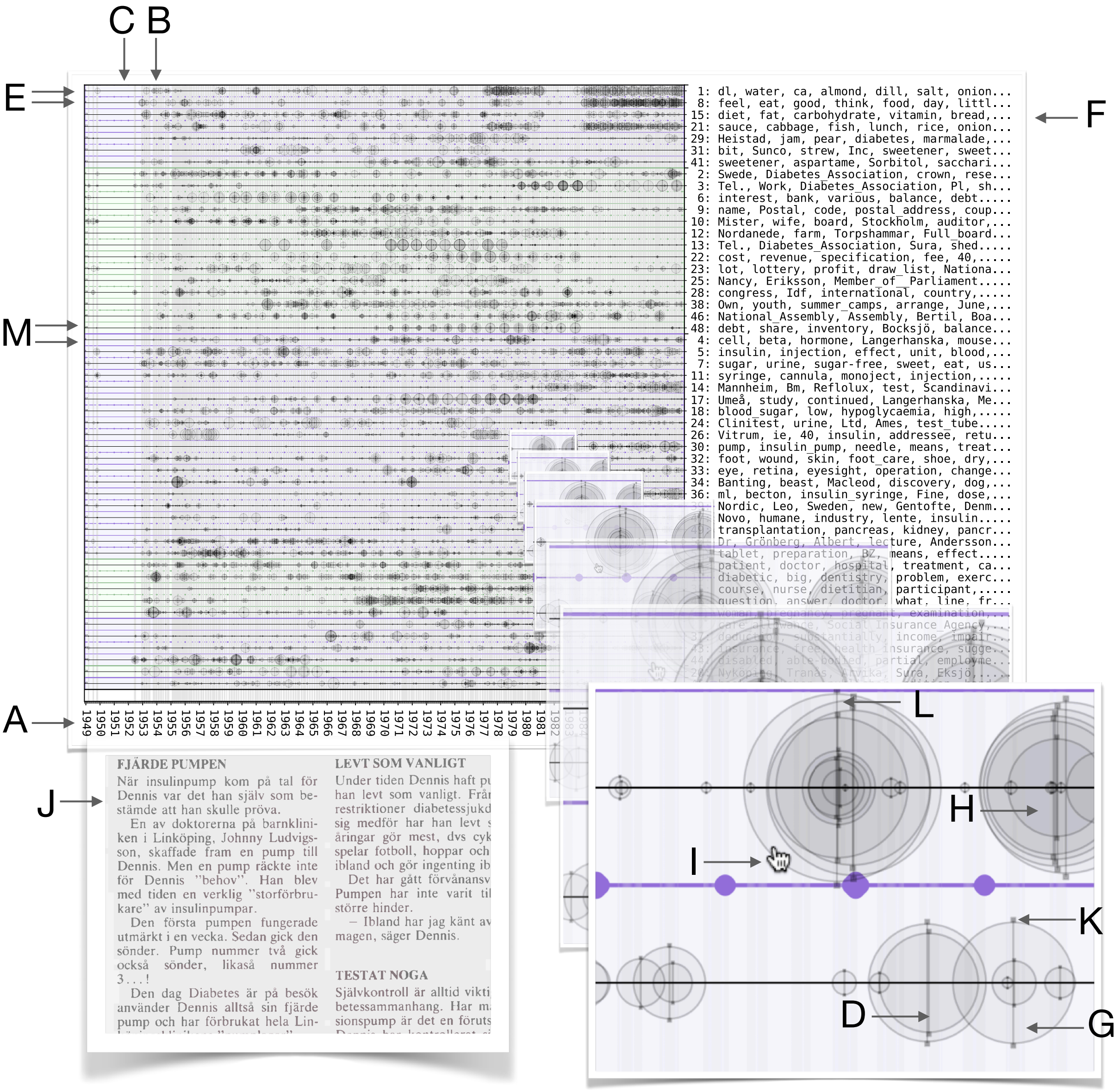
Skeppstedt, M., Kucher, K., Stede, M., & Kerren, A. (2018). Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics. *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, 9–16.

Söderfeldt, Y. (2024). Joint efforts in the Swedish model: The Swedish Diabetes Association under Nancy Eriksson (1956-1978) [Manuscript submitted for publication]. Department of History of Science; Ideas, Uppsala University.

Stede, M., Bracke, Y., Borec, L., Kinkel, N. C., & Skeppstedt, M. (2023). Framing climate change in Nature and Science editorials: applications of supervised and unsupervised text categorization. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-023-00199-7

Underwood, T. (2017). A genealogy of distant reading. *DHQ: Digital Humanities Quarterly*, *11*(2).

Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. *Conference on Knowledge Discovery in Data: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining; 20-23 Aug. 2006*, 424–433.


C
B
E
M
A
J
F
L
H
I
K
D
G
1: dl, water, ca, almond, dill, salt, onion...
8: feel, eat, good, think, food, day, littl...
15: diet, fat, carbohydrate, vitamin, bread,...
21: sauce, cabbage, fish, lunch, rice, onion...
29: Heistad, jam, pear, diabetes, marmalade,...
31: bit, Sunco, strew, Inc, sweetener, sweet...
41: sweetener, aspartame, Sorbitol, sacchari...
2: Swede, Diabetes_Association, crown, rese...
3: Tel., Work, Diabetes_Association, Pl, sh...
6: interest, bank, various, balance, debt.....
9: name, Postal, code, postal_address, coup...
10: Mister, wife, board, Stockholm, auditor,...
12: Nordanede, farm, Torpshammar, Full_board...
13: Tel., Diabetes_Association, Sura, shed.....
22: cost, revenue, specification, fee, 40,.....
23: lot, lottery, profit, draw_list, Nationa...
25: Nancy, Eriksson, Member_of__Parliament.....
28: congress, Idf, international, country,.....
38: Own, youth, summer_camps, arrange, June,...
46: National_Assembly, Assembly, Bertil, Boa...
48: debt, share, inventory, Bocksjö, balance...
4: cell, beta, hormone, Langerhanska, mouse...
5: insulin, injection, effect, unit, blood,...
7: sugar, urine, sugar-free, sweet, eat, us...
11: syringe, cannula, monoject, injection,.....
14: Mannheim, Bm, Reflolux, test, Scandinavi...
17: Umeå, study, continued, Langerhanska, Me...
18: blood_sugar, low, hypoglycaemia, high,.....
24: Clinitest, urine, Ltd, Ames, test_tube.....
26: Vitrum, ie, 40, insulin, addressee, retu...
30: pump, insulin_pump, needle, means, treat...
32: foot, wound, skin, foot_care, shoe, dry,...
33: eye, retina, eyesight, operation, change...
34: Banting, beast, Macleod, discovery, dog,...
36: ml, becton, insulin_syringe, Fine, dose,...
Nordic, Leo, Sweden, new, Gentofte, Denm...
Novo, humane, industry, lente, insulin.....
transplantation, pancreas, kidney, pancr...
Dr, Grönberg, Albert, lecture, Andersson...
tablet, preparation, BZ, means, effect.....
patient, doctor, hospital, treatment, ca...
diabetic, big, dentistry, problem, exerc...
course, nurse, dietitian, participant,.....
question, answer, doctor, what, line, fr...
woman, pregnancy, pregnant, examination,...
care_allowance, Social_Insurance_Agency,...
37: deduction, substantially, income, impair...
43: insurance, free, health_insurance, sugge...
44: disabled, able-bodied, partial, employme...
20: Nyköping, Tranås, Arvika, Sura, Eksjö,.....
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981

FJÄRDE PUMPEN

När insulinpump kom på tal för Dennis var det han själv som bestämde att han skulle pröva.

En av doktorerna på barnkliniken i Linköping, Johnny Ludvigsson, skaffade fram en pump till Dennis. Men en pump räckte inte för Dennis "behov". Han blev med tiden en verklig "storförbrukare" av insulinpumpar.

Den första pumpen fungerade utmärkt i en vecka. Sedan gick den sönder. Pump nummer två gick också sönder, likaså nummer 3...!

Den dag Diabetes är på besök använder Dennis alltså sin fjärde pump och har förbrukat hela Lin-

LEVT SOM VANLIGT

Under tiden Dennis haft pu
han levt som vanligt. Från
restriktioner diabetessjukd
sig medför har han levt s
åringar gör mest, dvs cyk
spelar fotboll, hoppar och
ibland och gör ingenting ib

Det har gått förvånansv
Pumpen har inte varit til
större hinder.

– Ibland har jag känt av
magen, säger Dennis.

TESTAT NOGA

Självkontroll är alltid vikti
betessammanhang. Har ma
sionspump är det en föruts

Figure 1: A visualisation of the approach to gradually zoom in for more detail and then finally click on the topic-strength indicator to reach the text.
Texts: Time is represented by position on the x-axis (**A**). Each text in the collection is visualised by a long vertical line positioned at the timestamp for the text (**B**). The visualisation of how the number of texts vary between different time periods is exemplified by the contrast between (**B**) and (**C**), i.e. many texts were published at (**B**) and none at all at (**C**).
Topics: The topics are represented by position on the y-axis, i.e. each topic is represented by a horizontal lane. (**E**) indicates the first and second of the horizontal topic lanes. The topic labels, which are created by the words most closely associated with the topic, are positioned to the right of the horizontal lanes (**F**). Ellipsis indicates that not all of the most closely associated words fit into the space available for labels. The centre of each horizontal topic lane is marked by a thin, black horizontal line, which we refer to as the *topic-line* (**E**).
Text-topic associations: If a text is one of the *n* texts most closely associated with a topic, this is indicated at the point where the *text-line* intersects the *topic-line*, by two indicators: by a vertical bar (**D**) – which runs along the text-line, and which has a height proportional to the text's topic-strength – and by a circle (**G**) – which has its centre at the intersection and which has a radius proportional to the topic-strength.
Reach the text: By clicking on the circle representing topic-strength (**I**), the web page associated with the text is opened (**J**). It is also possible to click on the small circles at the top and bottom of the topic-strength bars to reach the linked page (**K**). (**H**, **L** and **M** are described elsewhere in the text.)

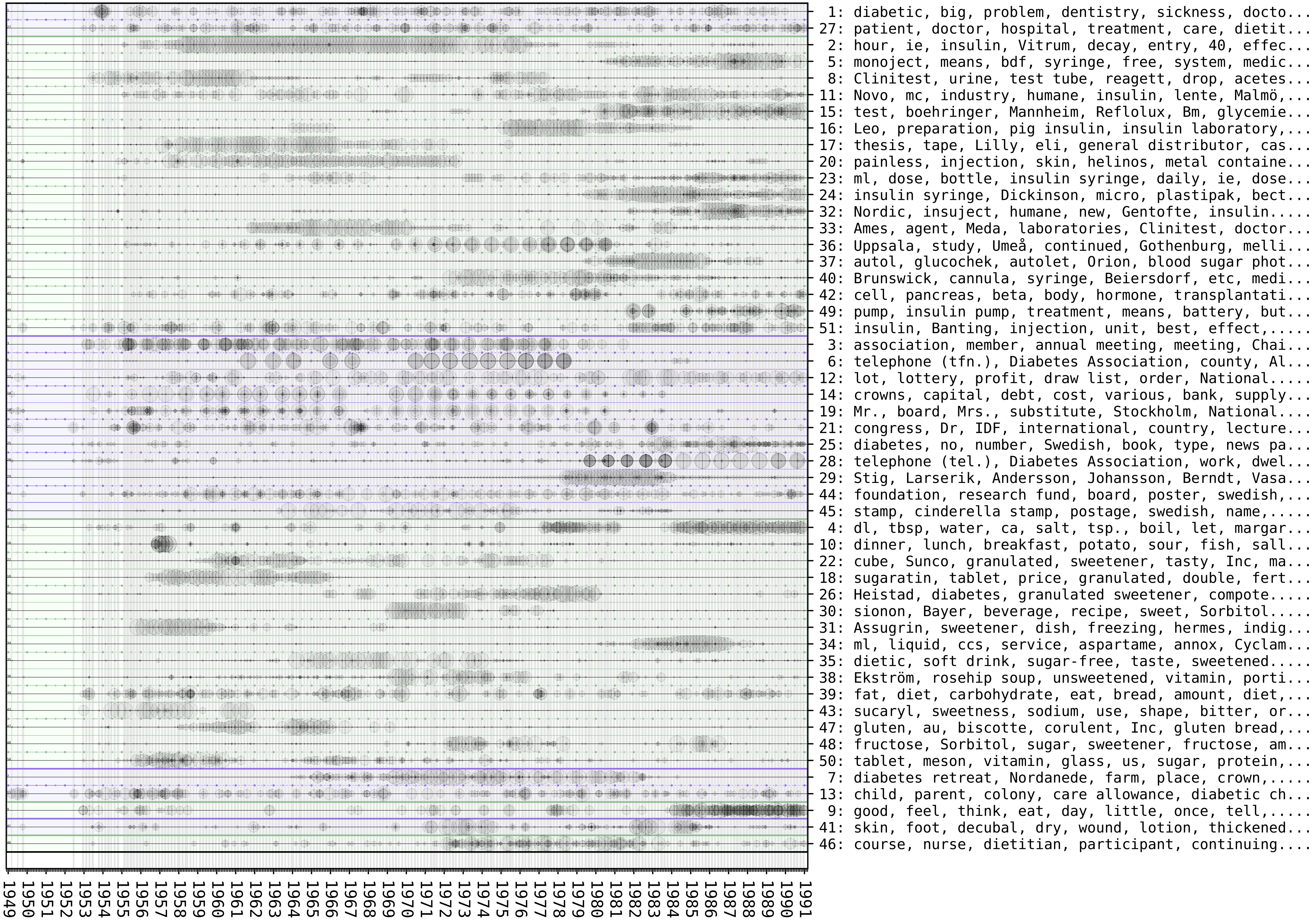


Figure 2: Timeline for the entire corpus from the Diabetes periodical. Each vertical line represents a text. The circles on the horizontal topic-lines indicate the occurrence of a topic in the text.

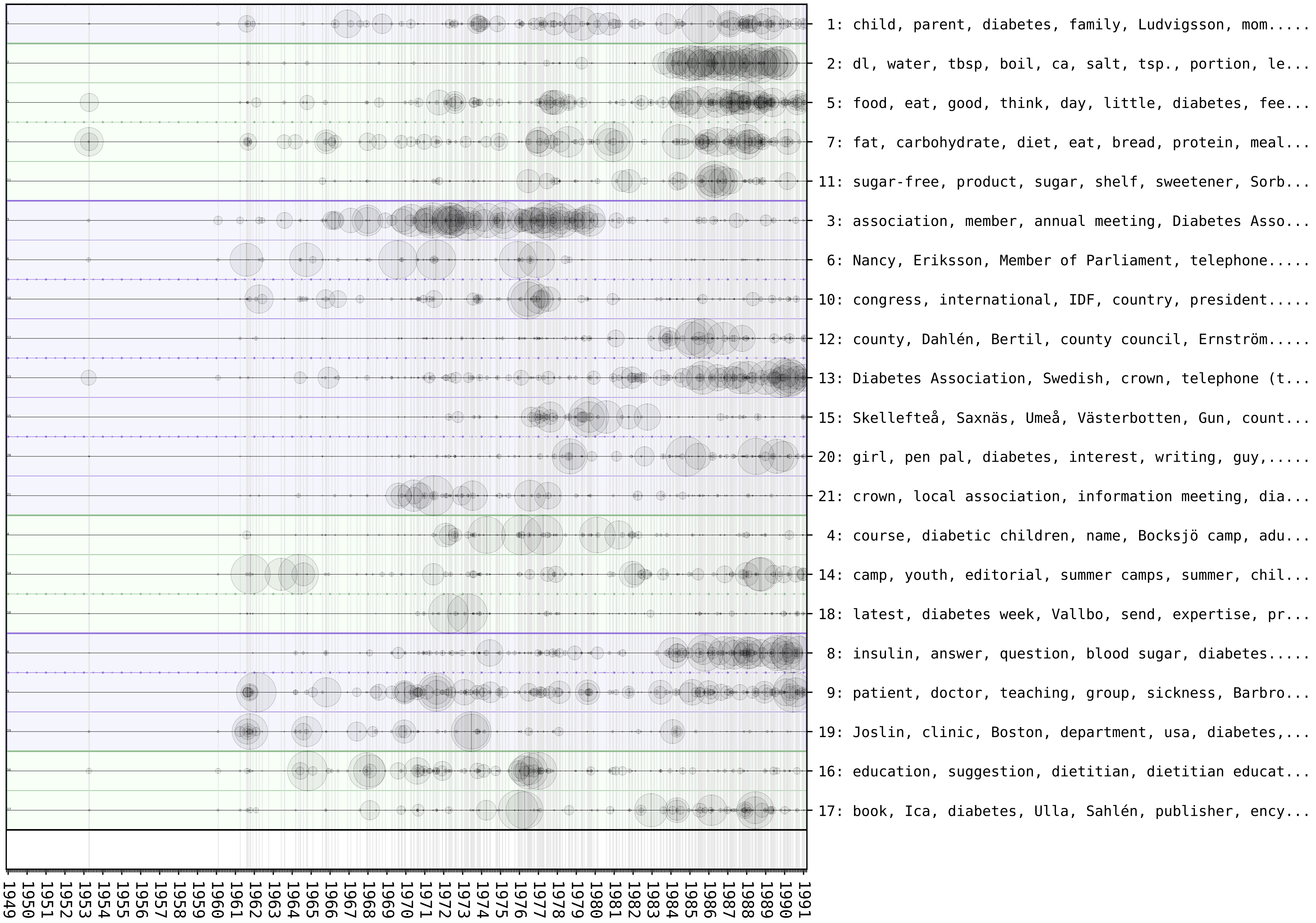


Figure 3: Timeline for a subset of the corpus containing the word “dietitian”. Each vertical line represents a text. The circles on the horizontal topic-lines indicate the occurrence of a topic in the text.