How to read the 52.000 pages of the British Journal of Psychiatry? A collaborative approach to source exploration

Eva Andersen, Maria Biryukov, Roman Kalyakin, Lars Wieneke

Journal of Data Mining and Digital Humanities

How to read the 52.000 pages of the British Journal of Psychiatry? A collaborative approach to source exploration

Eva Andersen[1], Maria Biryukov, Roman Kalyakin, Lars Wieneke

University of Luxembourg, Luxembourg

*Corresponding author: Lars Wieneke Lars.Wieneke@uni.lu

Abstract
Historians are confronted with an overabundance of sources that require new perspectives and tools to make use of large-scale corpora. Based on a use case from the history of psychiatry this paper describes the work of an interdisciplinary team to tackle these challenges by combining different NLP tools with new visual interfaces that foster the exploration of the corpus. The paper highlights several research challenges in the preparation and processing of the corpus and sketches new insights for historical research that were gathered due to the use of the tools.

keywords
topic modelling; visualisations; text mining; history of psychiatry; United Kingdom

INTRODUCTION

Contemporary historians face an overabundance of digitised and digital born sources. As Roy Rosenzweig pointed out more than 15 years ago "Surely, the injunction of traditional historians to look at 'everything' cannot survive in a digital era in which 'everything' has survived." (Rosenzweig 2003) Navigating, exploring and analysing these sources can form a major research obstacle for historians and humanists alike. In this paper we therefore want to discuss how to foster the process of corpus exploration through the application of Natural Language Processing (NLP) and interface design while closely supporting the research process. Our case study focuses on an ongoing PhD project concerned with the dissemination of psychiatric knowledge across Europe between 1843 and 1925 through five different psychiatric journals in different languages[2] with a total of about 250000 pages. The sheer quantity of this material formed a severe obstacle to perform a valuable and thorough analysis of the sources without computational support and to provide answers to the specific research questions. To mediate this, we set up a probing exercise to explore the feasibility of potential solutions to the problem, creating at the same time an interesting challenge for computer science due to the unstructured nature of the historical sources as data.

In the following we will briefly outline the specific challenges posed by historical psychiatric journals and describe the context of our case study. After this we will discuss how we performed a semi-automatic cleaning of the dataset and describe different approaches to topic modelling in order to enable the historian to find relevant material in the sources. We will then introduce *histograph* as an interface that enables corpus exploration. Here we will present a new type of visualisation that improves the practical usability of the topic modelling output for the sake of content exploration as well as the addition of further content lenses or

[1] The work of Eva Andersen has been supported by the Luxembourg National Research Fund (FNR) (10929115)

[2] The *British Journal of Psychiatry* (English); the *Bulletin de la Société de Médecine Mentale de Belgique (French)*; the *Annales Medico-Psychologique* (French); the *Psychiatrische Bladen (Dutch)*; the *Allgemeine Zeitschrift für Psychiatrie* (German).

perspectives that support the exploration tasks of the historian. We will conclude the paper with a discussion of the historical findings and in how far the use of digital methods enabled additional insights beyond what is feasible through purely paper based research.
Overall we would like to highlight that the problem at hand mandated a highly interdisciplinary approach, bringing together the work and professional perspectives of a historian, a computer scientist and a software engineer. The mutual understanding that this project has created amongst its participants on the structure and content of the (digital) corpus, as well as the different processes that were involved, ranging from data cleaning and NLP and topic modelling, to the validation of the results and its ultimate integration into an interface (*histograph*) was integral to obtain relevant results, and hence the collaboration between all the researchers involved cannot be underestimated.

Psychiatric journals and the case of the Asylum Journal

Psychiatric journals were at the time multipurpose publications. Aside from meeting reports, they included original scientific contributions and observations, along with literature reviews of domestic and foreign periodicals and books. In addition to these evident features, they also contained announcements for conferences and contests for prize essays, while also keeping track of the rotating job-positions and deaths of asylum physicians and university professors. This variety posed a serious challenge to the research endeavour, as the exploration of analogue sources with this extent and diversity through close reading only, is an almost impossible task. To tackle this issue we started a probing exercise where only one journal was taken as a case-study in order to reduce the complexity that different languages and source structures would create. For this we chose the *Asylum Journal*[3] which was created in 1853, and is currently known as the *British Journal of Psychiatry* (BJP) which was, and still is, published under the auspices of the *Royal College of Psychiatrists* (°1841)[4]. This specific journal covers a period of 72 years (1853-1925)[5], resulting in 52167 pages to explore.
As a first step, we collected the different issues of this journal in a digital form[6]. The material in question was downloaded in PDF format, along with plain text transcripts when available. If transcripts were unavailable, an optical character recognition (OCR) process was performed using ABBYY recognition server[7] to process all missing segments and to produce plain text output. Both the existing text transcripts as well as the post-OCRed text showed significant recognition errors. We did not perform any post-OCR corrections, and although this is an important aspect to take into account, we did not have the time nor the means to create an OCR gold standard for the corpus and apply thorough OCR cleaning. Currently the OCR errors do not seem to affect the creation of a useful topic modells. We do however plan to review this in more detail in the future.

[3] We chose this specific journal for pragmatic reasons: due to the different nationalities of the researchers an English corpus is more easily understood by everyone involved.
[4] In its long existence the British association changed names four times. In its earliest form they were called the *Association of Medical Officers of Asylums and Hospitals for the Insane*. A detailed history of the society can be found in (Bewley 2008)
[5] The year 1853 was excluded from the final analysis. The volume contained only 16 pages (i.e., documents), which resulted in a continuous repetition of the same topics during our early experiments. These results were not meaningful from a qualitative point of view. Furthermore, close evaluation by the historian revealed that this year did not contain crucial material for the particular research interests and was therefore omitted from the analysis.
[6] The BJP journal was accessed via the website of *The Royal Association of Psychiatrists*, who made *The British Journal of Psychiatry* and its predecessors available online, as well as via the Internet Archive. ("The British Journal of Psychiatry" n.d.; "The Journal of Mental Science" n.d.). Due to the copyright status of the material provided by *The Royal Association of Psychiatrists* we can not publish our full corpus.
[7] We used ABBYY Recognition Server 4.0 Extended Edition for this purpose

The complex interaction of different transformations (pdf -> OCR -> plain text) on the data in various degrees of quality posed a challenge for the historian. Particularly because, to work with such corpora the historian needs to be able to manipulate digitised "originals" in order to use them in a more consistent manner. Damerow and Wintergrün have put forward that historians always need full control of a corpus as even within "a digital framework, historical research relies on trust in its sources"(Damerow and Wintergrün 2019). This trust in sources is a precarious balancing act for the historian. How *can* we control a large and, in terms of data quality, inconsistent digitised corpus — such as the *Asylum Journal* — and give the researcher as much exploration possibilities as possible to do historical research? This was one of the prevailing issues as well as the driving force within our project.

I NLP/TOPIC MODELLING

In this part we provide a motivation for the entire text analysis pipeline, from cleaning the raw data to topic modeling.

1.1 Corpus preprocessing

The discovery of domain-specific ideas in the corpus would call for an analysis of all relevant scientific publications, such as articles, book reviews, honorary lectures and to a certain extent asylum reports. However, manual inspection of our input data revealed other types of heterogeneous content in the corpus such as financial accounts, obituaries, letters to the editor, etc. Importantly, these publications were almost as equally present in the journal volumes as the more relevant content types. Absence of clear separation marks between the publications, missing or incomplete table of contents, coupled with less than perfect OCR quality, made it impossible for us to automatically extract only relevant publication types from the corpus.

Following the practice of customisation of document length for the topic modelling (Schofield, Magnusson, and Mimno 2017) and the author style detection (Tschuggnall et al. 2017), we defined one page as the document unit, and splitted the entire scope of 72 volumes into 52396 pages. Next, we aimed to reduce pages that bear little content, such as membership lists as well as drawings, sketches and tables. To identify such pages in an unsupervised way, we applied a corpus-statistics methodology (Gries 2009), exploiting the observation that such pages will have far less stopwords[8] than a page with "regular" content. After the tokenization[9], we calculated the average number of stopwords per page in every given year, and removed all the pages that either had less stopwords than the threshold defined for the given year, or had less content-bearing words than stopwords while the number of content bearing words was below the threshold set to 30. In this way we reduced[10] the number of pages to 47085.

1.2 Topic modelling

The diversity of the corpus in terms of publication types prevented us from proceeding directly towards idea extraction. An intermediate step, which would help to structure the entire corpus into semantically distinct groups, was necessary. Doing so would allow to identify document clusters potentially pertinent to the main research questions and suitable for more in-depth investigation; at the same time, we expected to isolate irrelevant clusters which could be excluded from future analysis. A widely used unsupervised text clustering technique

[8] We use the stopwords list included in Mallet package http://mallet.cs.umass.edu and extend it with some additional words, such as addressing titles or words which appear in OCR production statements.

[9] We used Stanford Core NLP suite for the tokenisation and lemmatisation("Stanford NLP Tools" n.d.)

[10] The minimum of 30 content-bearing words per page has been selected by examining the data. In this way we also implicitly defined the minimal length of a valid document.

which meets our requirements is called "topic modelling" and aims at unsupervised identification of latent topics within a collection of text documents. It has been employed in digital humanities in a variety of ways, such as to derive and analyse topics in eighteenth century newspapers (D. J. Newman and Block 2006); reason about Classics as a field (Mimno 2012); or to analyze the development of themes in the field of computational linguistics (Hall, Jurafsky, and Manning 2008). As opposed to those use cases, we do not recur to topic modelling in order to reason about a field as a whole; rather we use it in order to get access to relevant material which would enable further investigation.

Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) and Non-Negative Matrix Factorization (NMF) (Lee and Seung 1999; Greene and Cross 2016; Luo et al. 2017) are popular algorithms underlying topic modeling. Both methods, in their original definition, use stochastic initialisation which leads to an instability of generated topics and make experiment results hardly reproducible. Several studies have been performed with the idea of finding ways to mitigate such instability issues. For LDA, approaches such as the selection of the most frequently assigned topics or the clustering of topics generated during repetitive runs (Riedl and Biemann 2012; Mäntylä, Claes, and Farooq 2018); freezing the topic labels or lists of top topic descriptors, generated during model updates (Yang et al. 2016), or optimisation via differential evolution (Agrawal, Fu, and Menzies 2018) have been proposed. For NMF, initialisation with Non-negative Double Singular Value Decomposition (NNDSVD), which does not contain a stochastic element, has been shown to effectively reduce instability of the generated topics (Belford, Namee, and Greene 2018).

1.3 Selection of the topic modelling algorithm

Being aware of these properties of LDA and NMF, we made preliminary runs with both and compared the performance in terms of the stability of the generated topics. We ran LDA[11] and NMF[12] on our data to generate sets from 2 to 10 topics, and repeated runs for 50 times using each of the methods. Then, we compared the similarity between the topic sets generated by LDA and NMF across all the runs, applying Average Jaccard Similarity measure (Greene, O'Callaghan, and Cunningham 2014), which accounts for both term overlap and their ranking. The results show that NMF stability outscores that of LDA on all topic sets, ranging from [0.36 - 0.45] for LDA, versus [0.65 - 0.90] for NMF. This concern became relevant for us due to the interdisciplinary nature of our project as we were aware that unstable topic sets make data exploration more difficult for the historian.

Besides topic stability, we considered two more factors when choosing between LDA and NMF. One of them is the number of hyper-parameters to be specified. Both algorithms require parameter *k* which indicates the number of desired topics. In addition, LDA requires two more parameters: α, which is responsible for topic distribution over the corpus, and β, which controls the word distribution over topics. Even though "off the shelf" values are sometimes applied, (e.g., MALLET package defaults which correspond to those suggested in (Steyvers, Griffiths, and Kintsch 2006)), it has been demonstrated that there are no reliable defaults and that parameter values should be learned from the given data set. This makes the use of LDA more demanding in terms of preparation.

The second aspect concerns topic properties. It has been noticed that LDA tends to produce rather generic topics with substantial overlap between the topic descriptors. NMF on the other

[11] We use the MALLET toolkit (http://mallet.cs.umass.edu) We tried various values for the hyper-parameters α and β aiming to maximise topic specificity. Stability scores increased slightly with smaller values of α and β (i.e., with more specific topics).

[12] For this particular task we use the NMF implementation by D. Greene et al., available at https://github.com/derekgreene/topic-stability.

ISSN 2416-5999, an open-access journal

hand yields more specific topics and may be more suitable for the analysis of narrow "non-mainstream" domains (O'Callaghan et al. 2015).
Furthermore the historical research questions are concerned with temporal changes in psychiatric knowledge acquisition. This perspective calls for a technique that would allow us to follow the dynamic evolution of the identified topics. Various approaches have been proposed to track topic evolution over time. Many of them extend the LDA technique to associate topics with time frames (X. Wang and McCallum 2006; Blei and Lafferty 2006; Cui et al. 2011; Beykikhoshk et al. 2018). However we opted for the NMF-based approach to maintain integrity of the workflow. We use a toolkit[13] which was designed by D. Greene and J. P. Cross, to model topics sequentially, moving from discrete time frames (called "window topics") towards their combined representation over time (called "dynamic topics")[14].
In the first phase, documents are organised into disjoint sets, with each set corresponding to one time window (i.e. the BJP of 1880). Topics generated from this input are called "window topics" and independently represent every time window. Each document in the window is scored with respect to each topic in the topic set.
For the second, dynamic phase, the original corpus is represented in an abstract manner by "topic documents", where each topic document is composed of the top-ranked terms from each topic in the window topic model. The underlying assumption is that thematically close topics coming from different windows will share similar topic documents[15]. Dynamic topics can be seen as a generalisation of the window topics[16], such that multiple individual window topics can be associated with a single dynamic topic. Each window topic is scored with respect to each dynamic topic which quantifies their relatedness.

1.4 Corpus preparation for the topic modelling with NMF

As mentioned earlier, one printed page is considered a document in our corpus. Documents which have been retained after the initial cleaning, are lemmatised and lower-cased. The advantage of working with the lemmatised text is that it helps to reduce the vocabulary size and ties morphologically distinct words with the same meaning into one lexical unit, thus making future topic descriptors more diverse. A document is considered valid if it contains at least 50 terms after the removal of stop words and terms with less than 3 characters. Additionally, terms which occur in less than 10 documents are removed. A term-document matrix is constructed for each time window. In our experiments window size is equal to 1 year, which corresponds to one yearly volume of the journal. Tf-idf term weighting and document length normalization are applied. Overall, our entire data set consists of 72 time windows, spanning the period from 1854 to 1925. These are composed of 47069 documents and 139422 terms. Matrix size ranges from 160 documents in 1896 to 1140 in 1881, with an average of 661.2 documents and 1921.08 terms per window.

1.5 Topic coherence evaluation and selection of the number of topics

We have mentioned earlier that topic modeling algorithms require the user to specify the number of topics to be generated. As it is often the case, such a number is not known in advance. The dynamic-nmf toolkit allows the user to adjust the number of topics via

[13] Available at https://github.com/derekgreene/dynamic-nmf.
[14] In what follows we use the original notation proposed by the authors of the toolkit, to designate the topic types. "Window topic" refers to a discrete time frame, which in our study is equal to one year. "Dynamic topic" refers to an aggregation of window topics, generated for a specific time span, into one topic.
[15] Detailed description of the entire procedure of window and dynamic topic generation implemented in dynamic-nmf toolkit is described in (Greene and Cross 2016).
[16] The span of window topics which serves the ground for the dynamic topic generation is defined by the user and may range from a subset of windows to the entire set.

generation and "on the fly" evaluation of multiple topic sets. Topics are evaluated from the point of view of their semantic interpretability, or "coherence". The intuition behind this is that a coherent topic consists of descriptors that tend to co-occur or belong to the same semantic space in the reference corpus[17]. Various metrics for topic coherence calculation have been proposed, including Pointwise Mutual Information (D. Newman et al. 2010) or log conditional probability (Mimno et al. 2011). The dynamic-nmf toolkit calculates topic coherence as follows: the coherence of an individual topic is calculated as the mean cosine similarity between vectors corresponding to the topic descriptors, where the vector space is constructed using the Word2Vec algorithm (Mikolov et al. 2013). Coherence of the entire model is represented as the mean coherence across all the topics. In practice, we can plot model coherence values calculated for each topic set in range of [k_{min}, k_{max}] and select k with the highest score. The number of topics used in the experiments described here range from 5 to 10. This choice was initially motivated by the limited amount of time we had for the experiment and the number of domain experts that could evaluate the topics.

1.6 Selection of the reference corpus

The Word2Vec algorithm responsible for the construction of the vector space can be applied to external data, which is not used for topic modeling (D. Newman et al. 2010), or to the same modelled text (Mimno et al. 2011; Greene and Cross 2016). To select an appropriate strategy we explored three different Word2Vec models, generated from potentially relevant corpora. The models we considered are the one generated from our data[18]; a model generated from full texts of the biomedical articles available via the PubMed Central (PMC) portal[19]; and GloVe(Pennington, Socher, and Manning 2014), trained on the Wikipedia, newswire and web crawled sources. To choose which of these models suits best to our task, we select "syphilis" and "paralysis" as keywords, which are in the research scope of the historian, and generate lists of their closest semantic neighbours.

0	syphilitic	0.7447	gonorrhea	0.7804	toxoplasmosis	0.7632	paralytic	0.8000	blindness	0.6630	paresis	0.7847
1	tabe	0.6727	tuberculosis	0.7051	chlamydia	0.7424	tabe	0.7155	convulsions	0.6536	paraplegia	0.7231
2	paralysis	0.6709	chlamydia	0.6984	STIs	0.7114	paresis	0.7139	spinal	0.6260	quadriparesis	0.7181
3	tertiary	0.6389	hepatitis	0.6536	trichomoniasis	0.7012	syphilis	0.6709	atrophy	0.6192	fasciculations	0.7053
4	lue	0.6225	venereal	0.6363	gonorrhea	0.6946	general	0.6435	neurological	0.6174	weakness	0.7052
5	metasyphilitic	0.6093	hiv	0.6345	STI	0.6894	tabetic	0.6416	debilitating	0.6052	quadriplegia	0.7013
6	disease	0.6071	typhus	0.6100	serology	0.6810	dorsalis	0.6390	dislocation	0.6022	spasticity	0.6974
7	syphi	0.5962	leprosy	0.6053	gonorrhoea	0.6807	paralysis	0.6293	deafness	0.6001	flaccid	0.6948
8	chancre	0.5943	herpes	0.6025	HIV	0.6800	locomotor	0.6272	cord	0.5768	convulsions	0.6944
9	parasyphilitic	0.5881	malaria	0.5999	Syphilis	0.6717	alysis	0.6177	respiratory	0.5735	tremors	0.6940
10	wassermann	0.5836	tb	0.5930	71/300	0.6608	disease	0.6169	numbness	0.5720	hyperreflexia	0.6886
11	luetic	0.5827	typhoid	0.5878	STDs	0.6563	ataxy	0.6159	hemorrhage	0.5718	incoordination	0.6879
12	parasyphilis	0.5777	infections	0.5846	co-infections	0.6466	metasyphilitic	0.6146	degeneration	0.5669	tetraparesis	0.6865
13	venereal	0.5726	infection	0.5785	sero-positive	0.6454	paralytica	0.6101	paralyzed	0.5619	hypopotassemic	0.6770
14	dorsalis	0.5716	dysentery	0.5760	neurosyphilis	0.6421	syphilitic	0.6100	palsy	0.5617	myoclonus	0.6705
15	tabetic	0.5684	leptospirosis	0.5731	chancroid	0.6389	tabo	0.6071	dystrophy	0.5570	paralytic	0.6666
16	gummata	0.5609	smallpox	0.5670	coinfection	0.6381	palsy	0.6028	fatigue	0.5569	areflexia	0.6665
17	aortitis	0.5597	prevalence	0.5652	toxoplasma	0.6376	aortitis	0.5934	complications	0.5567	spasms	0.6619
18	fournier	0.5594	transfusions	0.5608	serologic	0.6358	generalparalysis	0.5915	illness	0.5563	hypotonia	0.6588
19	paralytica	0.5584	measles	0.5570	transfusion-transmitted	0.6336	dementia	0.5889	epilepsy	0.5511	convulsion	0.6561

Figure 1. 20 closest semantic neighbours for the term "syphilis" (columns 1-3) and "paralysis" (columns 4-6), extracted from Word2Vec BJP, GloVe and Word2Vec PMC for each term. Vector dimensionality: 200 for the BJP and PMC models; 100 for GloVe.

Comparing the lists (see Figure 1) we notice that despite occasional commonalities between them, PMC-based and GloVe models demonstrate rather contemporary vocabulary, which does not exist in the same form within our corpus. On the other hand, non-words that are present in abundance in our corpus due to OCR errors, might not be found in the two other

[17] "Reference corpus" designate an external corpus outside the modeled data (D. Newman et al. 2010) or the same corpus which is used for topic modeling (Mimno et al. 2011; O'Callaghan et al. 2015).

[18] After having experimented with a variety of parameters, we finally used the skip-gram algorithm, apply context window ± 5 words and set vector dimensionality to 200.

19 PubMed Central free full-text archive of biomedical and life sciences literature. https://www.ncbi.nlm.nih.gov/pmc/. Modeling of the data is described in (Pyysalo et al. 2013) Models were downloaded from http://evexdb.org/pmresources/vec-space-models/

models. Both will contribute to the "out of vocabulary"[20] problem. Based on these observations we trained a Word2Vec model on our data[21], as we believed that it would be the most representative for evaluating the coherence of the topic models. As a side remark, we notice that semantic neighbors of "paralysis" and "syphilis" yielded by the BJP-based model, are almost the same. It is not the case for the corresponding lists generated by the PMC-based and GloVe models. It suggests that the BJP-based model reflects an appropriate nineteenth century view on the two diseases.[22] We could further assume that a combination of models generated from a variety of corpora might benefit some large-scale historical study of psychiatry, highlighting changes that happened over time.

1.7 Topic modelling experiments

With all the preliminaries being defined, we proceeded with the window and dynamic topic generation. Window topics are reviewed by the historian from the point of view of their interpretability and agreement with the automatic recommendation regarding the best number of topics. Evaluation at this stage has a two-fold role: assess the quality of the topics and, importantly, help to find the best setup for the generation of dynamic topics as they are built upon the per-window perspective.

Rank	1915_01	1915_02	1915_03	1915_04	1915_05	1915_06	1915_07	1915_08
1	reaction	blake	member	dream	fibre	nietzsche	asylum	drug
2	case	vision	meeting	freud	cell	anger	medical	habit
3	paralysis	william	association	idea	nerve	man	hospital	morphia
4	general	write	president	sexual	vessel	write	patient	sedative
5	syphilis	gilchrist	division	process	cord	friedrich	board	victim
6	spirochaete	poem	committee	thought	posterior	life	work	drink
7	fluid	mad	meet	emotion	cavity	time	psychiatry	opium
8	injection	time	secretary	psychic	artery	thing	act	state
9	positive	mental	council	interpretation	change	german	treatment	recover
10	wassermann	work	read	attitude	lesion	letter	lunacy	addiction

Figure 2. 8-topics partitioning of the BJP volume from 1915.

Among the window topics some have been judged generic while others appeared more focused and could be easily labelled. This turned out to be a recurrent phenomenon observed throughout the entire time span. Figure 2 shows an 8-topics set, generated for the window 1915. Note for example how topic # 5 (marked as 1915_5 in figure 2) is focused on biology or topic # 1 (1915_1) brings together experimental methods, disease diagnostics, and particular diseases (syphilis and paralysis) which involve the brain and nervous system. On the other hand, topic # 7 (1915_7) is generic in nature and most likely discusses the administrative practices within psychiatric institutions.

It is interesting to note that in most of the cases the number of topics suggested by the toolkit, based on the topic coherence evaluation, corresponds to the expert choice. In a few cases of disagreement, the expert mostly preferred more topics that offered a more fine-grained view on the domain and would include topics that otherwise would have stayed underscored. There was also an opposite example, when the expert suggested to collapse certain topics into one for the sake of logical generalisation. These observations guided our strategy in the next phase. Here, for every window which is supposed to participate in the dynamic topic construction, one topic model has to be specified out of several, generated with respect to the

[20] Out-of-vocabulary words are words which occur in the test corpus but do not exist (or are accounted for) in the training corpus.

[21] We use the entire dataset of 47069 documents for the training of the Word2Vec model.

[22] General paralysis and syphilis were throughout most of the 19th century seen as separate diseases. Via continuous research scientists over time began to understand that the symptoms of general paralysis were caused by syphilis. The characteristics of general paralysis are now known as belonging to neurosyphilis which is one of the forms that can manifest itself within the tertiary stage of syphilis.

requested range. All 72 windows were accounted for the dynamic topic stage, as well as all 10 topics. For each year, we picked up the model with the number of topics approved or suggested by the expert. In cases where explicit instruction was not available, we selected the model with the highest number of topics *k* out of three top-ranked topic partitionings. The reason why we preferred this solution to a mere selection of the partitioning with the highest number of topics (i.e., *k*=10 in our experiments) is that even though the expert inclined for a higher number of topics, there were cases when too many topics led to over-specialisation of the model. The specific use of the dynamic topics will be further discussed in the historic case study.

1.8 Remarks on the number of window topics and window topic coherence

Even though our working set was composed of maximum 10 topics per window, we did generate topic sets with larger k, thus creating sets of models with k in range [4:20]. By doing so we wanted to address two questions: a) if there were topics which remained hidden from the researcher's view in the smaller partitioning; b) how much topic coherence scores help to identify the most appropriate number of topics. We analyse partitioning into 10 and 20 topics from the researcher's and model coherence perspectives.

The researcher's analysis of topics, generated via various partitionings, revealed the following: a) certain topics remained stable and were visible in both - 10 and 20 topic partitioning. Specifically, these are topics related to biology (especially nerve, blood and brain cells and tissues, which is logical, given the thematic orientation of the entire corpus), criminal insanity, general paralysis or psychoanalysis. Another omnipresent topic had to do with various aspects of psychiatric institutions.

b) the constant interplay between the advantages and disadvantages of the growing number of topics. While higher topic partitioning (k>10) often offered a more detailed view of the field and rescued small yet important topics, there were also clear cases of over-specialisation and redundancy of certain topics. These observations lead to the second question of whether coherence measure provides reliable guidance for selection of the number of topics.

As suggested in (Röder, Both, and Hinneburg 2015; Greene and Cross 2016), we analyse median coherence scores, computed over the entire time span of the corpus. Figure 3 shows median coherence scores for the sets of topics generated with k ∈ [4,10] and k ∈ [4,20], calculated based on 10 and 20 top-ranked topic descriptors for every partitioning in BJP.

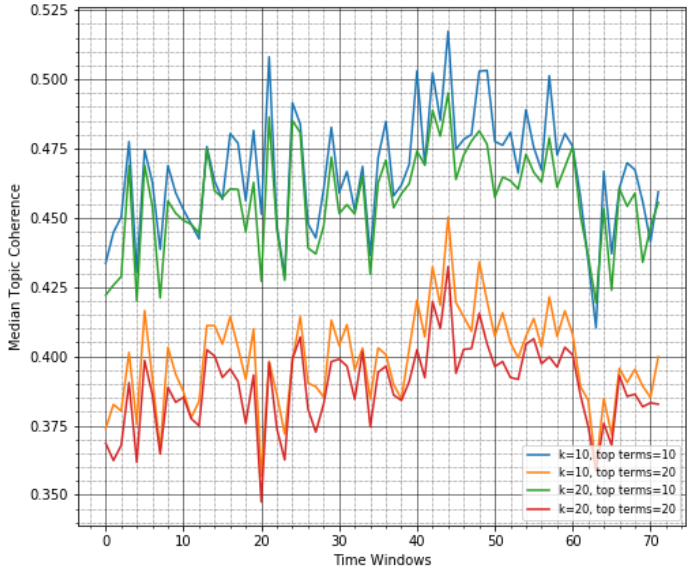


Figure 3. Median coherence scored calculated for calculated for the topic sets with k ∈ [4,10] and k ∈ [4,20] based on 10 and 20 top ranked topic descriptors for BJP in 1854-1925.

We observe that coherence scores depend mainly on the number of topic terms taken into consideration and that the fluctuation in the scores between the number of topics is only marginal.
We also calculated the average number of best suited topics for 10 and 20-topic partitioning, and the impact of the partitioning change onto the number of most appropriate topics suggested by the coherence score for every window over the whole time span. The change rate is calculated as the number of years in which the best number of suggested topics in the 20-topic partitioning differs with respect to the 10-topics partitioning, divided by the number of years in the time span. We observe the following: a) the average number of topics does not not exceed 5.9 even with the 20-topic partitioning (which is not in full agreement with the researcher's feedback described above); b) the average number of topics, as well as the change rate depend strongly on how many topic descriptors participate in the coherence calculation. The latter is 0.3 with top 10 descriptors, and goes down to 0.1 with top 20 descriptors.
It also turned out that the mean difference between the maximal and minimal coherence values computed over the entire time span was 0.01 for 10 and 0.03 for 20 topics partitioning with the corresponding standard deviation of 0.09 and 0.11. It means that the values as such do not have a discriminatory power. This observation corresponds to the researcher's remarques when several partitionings for a window could be accepted and seemed equally valid.
Taking into account the researcher's feedback on the quality of the topic partitioning and its statistical properties, we can conclude that coherence scores can only serve as guidance but cannot substitute the expert's assistance in selecting the best number of topics and even more for the number of topics in the partitioning.
Another observation has to do with the compatibility of the Word2Vec coherence measure with the very essence of the topic modelling. One of the strengths of the latter is its ability to capture polysemy[23]. On the contrary, Word2Vec constructs one vector per word, thus collapsing all possible meanings and contexts of a polysemic word. As a result, as long as the topic descriptors occur in contextual proximity, the topic will be considered coherent, irrespective of whether or not such polysemic words have been placed into one or multiple topics. It suggests that even though a Word2vec-based measure is suitable for topic coherence evaluation, it is less efficient for judging the best number of topics. It would be interesting to experiment with the contextual word embeddings, such as those proposed within Flair (Akbik, Blythe, and Vollgraf 2018) or Bert (Devlin et al. 2019) and see if they can help approximate topic modelling results and thus help to estimate the most appropriate number of topics.

II HISTOGRAPH

The massive digitisation of sources during the past decade by (national) libraries and archives has produced an abundance of interfaces to search collections. These search capabilities have also evolved over the years, now consisting of a mixture of simple keyword queries, n-gram frequencies, NER and topic modelling. However, not all platforms make all these options available and the most common feature is still keyword search[24]. A specific characteristic of these kind of repositories[25] is that sources and tools are often merged. A researcher is required

[23] Polysemy refers to the capacity of a word to have multiple meanings.
[24] (Ehrmann, Bunout, and Düring 2017) distinguish between "regular search" (sometimes supported by boolean operators and options to limit a date range or place), "fuzzy search" and "proximity search". The latter two are not often integrated in digital repository platforms (p12).
[25] There are many examples, but to name a few: Delpher (Dutch newspapers, books and journals) https://www.delpher.nl/; Impresso: media monitoring the past (French and German newspapers)

to use the digitised sources that these national and private institutions provide with the tools they offer. This aspect is quite revealing within the HathiTrust "data capsules"[26]: the data the researcher can use is encapsulated within a very specific framework and with very specific material. In essence, this does not have to be problematic if the researcher can find the material(s) he or she needs within one database. However, when this is not the case — journal X is available in digital library A and journal Y in digital library B — a consistent analysis becomes unattainable because digital libraries use different interfaces and a varied range of algorithms to structure and search their content. Furthermore, it is not always likely that the researcher can export data in a PDF, image or text format for analysis. Since the psychiatric journals in our possession came from multiple platforms it was key that the interface and algorithms could operate on different materials, and that the researcher was able to decide which sources needed to be implemented within the tool in order to be systematically analysed.

Based on our initial experiments with the historical dataset we decided to employ *histograph*, a tool initially built for graph based exploration of multimedia collections (Novak et al. 2014; Wieneke et al. 2014; Düring, Marten, Wieneke, and Croce 2015). *Histograph*[27] is a tool for the exploration and collective annotation of historical source material. It allows users to browse and search documents but also applies advanced visualisation tools for content exploration. As such it offers for example the ability to use the results of face recognition and named entity extraction processes to visualise the co-occurrence of persons within documents in a social graph of relationships. This makes the discovery of unexpected patterns possible.

The decision to use *histograph* as a platform for the exploration of the BJP corpus was based on two observations. First, while the raw topic modelling output was readable for the researcher it required a constant switch between the output and the source documents (PDFs) to review the content, thereby slowing down the exploration process. An integration into *histograph* promised to speed-up this process by enabling a direct link between the scores of the individual documents and the ability to view and access the relevant content both in the form of a transcript as well as a scan of the original page. Second, in discussions with the historian it became apparent that, even though topic modelling allowed us to structure the corpus, an efficient exploration strategy would demand multiple perspectives on the content which up to now were achieved through the use of different tools. In a spirit of building the historians macroscope (Graham, Milligan, and Weingart 2015) it became our goal to build an exploration facility that would provide multiple perspectives on the whole corpus in one view with the ability to identify and zoom into relevant sections that require a more detailed investigation. While the technologies and some of the visual approaches used show an overlap with projects such as Antconc (Anthony, Laurence 2019), Paper Machines (Guldi, Jo and Johnson-Robertson, Chris 2019) or even Googles NGram viewer (Michel et al. 2011) the main difference lies in the focus on exploration: data and visualisations are not used to make statements about the content of the corpus on their own, but provide a tool to find "the needle in the haystack" or simply the occurrence of very specific information. In so far, our approach is similar to other projects in the digital humanities such as the impresso project[28]. Clearly, this approach falls in the broader domain of Information Retrieval, defined as "[…] finding

---

https://impresso-project.ch/; HathiTrust Digital Library (newspapers, books, journals, etc) https://www.hathitrust.org/.

[26] For general information see HathiTrust Research Center Analytics (HTRC Analytics) https://analytics.hathitrust.org/. On HTRC Analytics data capsules see: https://analytics.hathitrust.org/staticcapsules#top. For the related topic of useable algorithms within HTRC Analytics see: https://analytics.hathitrust.org/statisticalalgorithms#top

[27] The source code as well as additional documentation is available here https://github.com/C2DH/histograph.

[28] see https://impresso-project.ch

material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).” (Manning, Raghavan, and Schütze 2008). While we also exploit conventional approaches such as full text search, we want to show in the following that we transcend this scope by incorporating a contextualised view that not only returns a list of potentially relevant documents but also lays out an integral map for exploration by providing additional indicators of relevance for the research question at hand.

As a point of departure, we imported the corpus as individual documents, each containing a single page, and linked in sequence with each other. With this approach we remained very close to the original concept of *histograph* thereby enabling us to make use of various existing filters in the interface. As it turned out, the time granularity of the corpus (one issue per year) led to a very sparsely populated timeline that made the selection of individual documents more complex than necessary. To mitigate this issue, we decided to evenly map each individual issue to the whole year in question, starting with page one on the first of January up to the last page on the thirty-first of December. While this process could be considered a-historical in the sense that the individual pages were obviously not published on the specific dates within the year, it allowed an easier integration into the tool and had the added effect of making the number of pages per year more visible through displaying the density of documents per year.

1.1 On the development of a new type of corpus visualisations

Based on the topic modelling results described in the previous chapter, our first task was to implement a visualisation that would enable users to cross reference the topics with the individual pages. We also wanted to experiment with different means to use and visualise the topic modelling scores as structural indicators – based on the assumption that individual sections formed by pages within a journal would show a coherence across topics – and to foster the exploration of individual as well as combined topics.

To this end we developed two visualisations to foster the corpus exploration: the first interface (topic view) gives quick access to the topic terms, and provides filtering capabilities for topic scores, keyword mentions in the documents/pages as well as ordering by time or relevance, a display of documents that match the criteria on a timeline as well as direct access to the different documents that match the selected criteria (see figure 4).

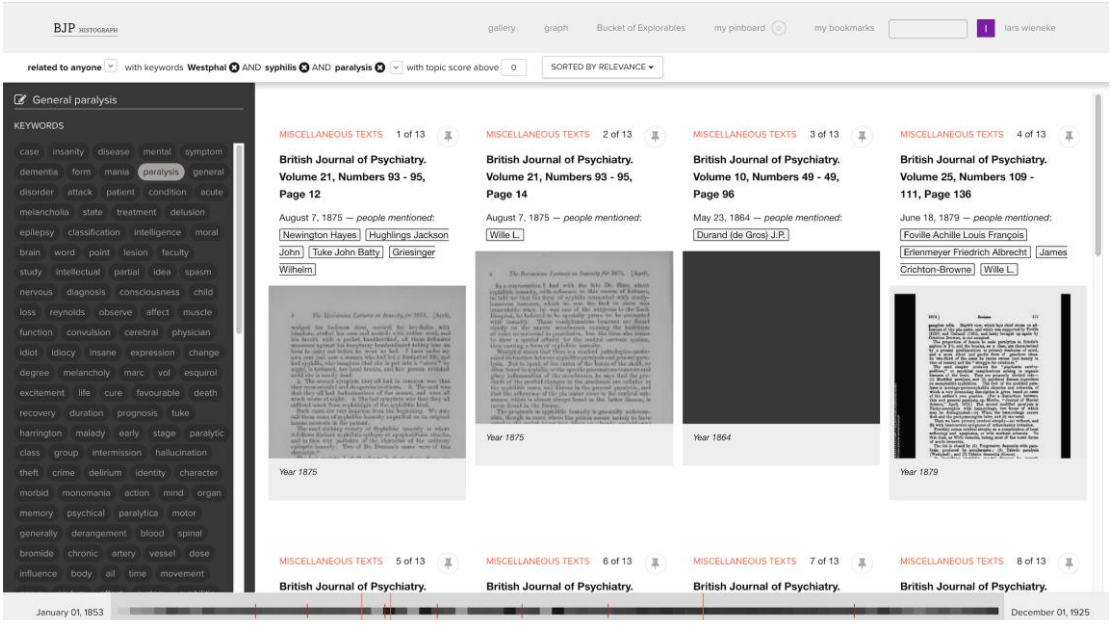


Figure 4. topic view with topic terms, filtering features, timeline view and access to individual documents.

While this interface builds on established mechanisms such as a filtering pane and direct display of the documents, surplus value emerges from access to the topic term list (which can become keyword filters when selected) as well as the visualisation of hits across the corpus.
The second visualisation we developed aims in providing an extensive perspective on the overall corpus with the ability to drill down into relevant sections by zooming with the added feature of adding multiple lenses on the corpus and to display matches in an integral view on the corpus. This view is currently titled *Bucket of Explorables* and is based on the idea of content "buckets" that aggregate a variable number of documents in a visual unit (see figure 5). On the largest zoom level, the full corpus is distributed into the available buckets and is based on two user controlled methods: an equal number of documents in a bucket and an aggregation by year. The former method is based on the fact that comparing groups of the same size is statistically correct while the latter method is closer to the working practices of historians for whom units of time have great significance.
Selecting an individual bucket opens a preview of all documents in the bucket below the visualisation. With the "Zoom into current bin" feature, users are able to open the specific year or an aggregated selection of documents to get a more fine grained view. Initial experiments with different kinds of visualisations for the topic modelling scores led to the conclusion that bubble charts could provide a relevant point of departure for our task as they allow to show two dimensions at the same time through colour and circle size which we used to represent the relative intensity of a topic in a group of documents.

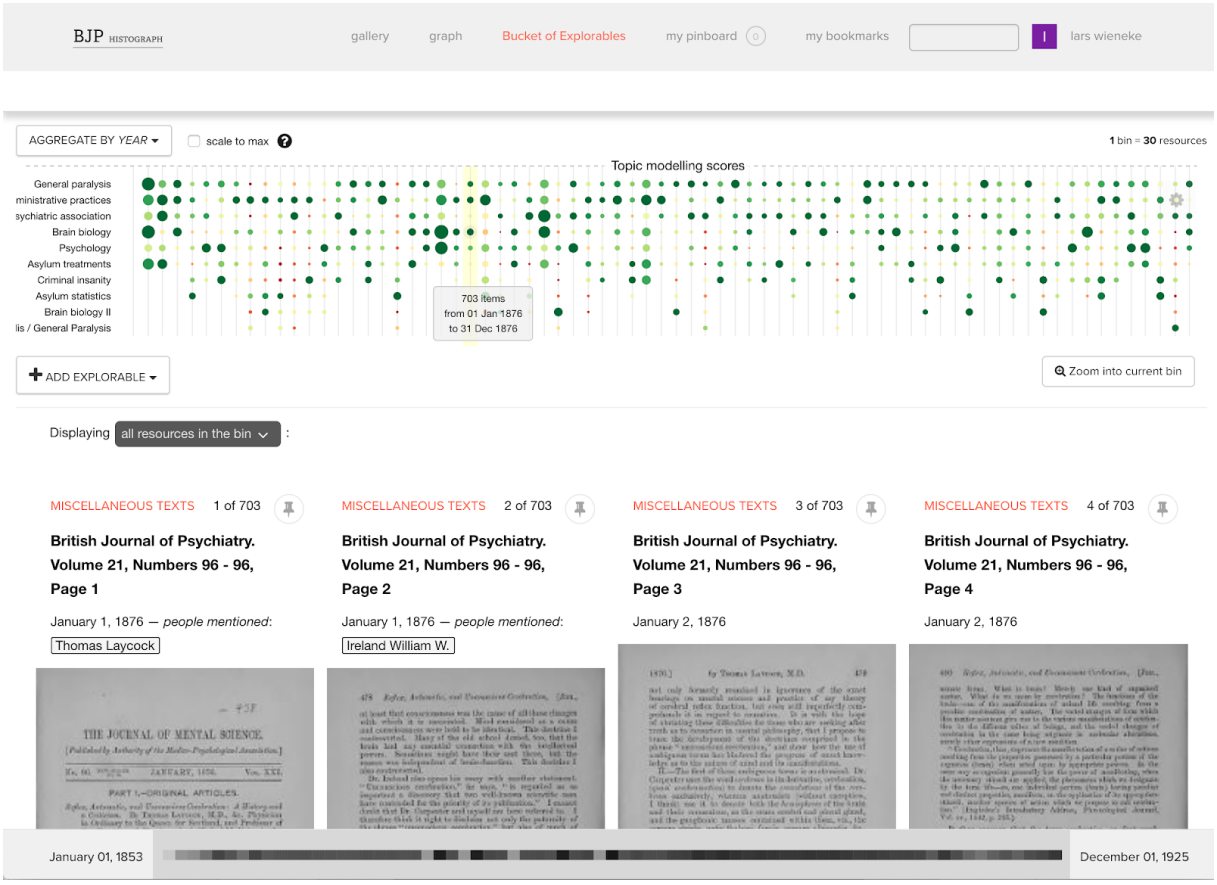


Figure 5. Aggregated topic model scores displayed in individual bins by year. The bucket for the year 1876 is highlighted and the associated documents for this bucket are displayed below the visualisation.

Based on this initial view we added several other features through an iterative process in close cooperation with the historian. First, we experimented with a number of standard ways to calculate topic modelling scores for a group of documents. In the end we decided to let the

user choose between two methods: mean value and maximum value. Grouping by the mean value of topic scores in a group of documents allows the researcher to see the general presence of the topic in the group. Whereas the maximum value allows the user to pinpoint the group that contains at least one document with a high score for a particular topic and "zoom in" into this group for further exploration.

Furthermore, we added an option to edit the names of the topics which were initially only numbered in sequence to make them more meaningful for the user and to foster the evaluation of the results.

The "Persons" view contains an aggregated number for each bucket with the number of individual persons mentioned. This is based on the results of a named entity recognition (NER) and named entity linking (NEL) task that was performed during the import of the documents into *histograph* For the NER task we used the flair framework and the 'ner-fast' model trained on the Conll-03 dataset[29]. Identified entities were either linked to the Google Knowledge Graph which yielded mainly Wikipedia references for our dataset[30] or, based on a string comparison, linked to a Nodegoat dataset procured from the 2TBI and TIC collaborative project[31]. The settings of this layer permit users to filter people with certain characteristics – depending on the available data – which is in turn displayed in the visualisation. As an experimental feature we integrated the ability to exclude people based on their nationality which allows, in the case of a British journal, to quickly identify contributions of, or references by actors outside of the British community. While this feature could become useful to identify relevant sources in the transnational migration of ideas, the current quality of the data poses significant limitations on its practical use: linking to Wikipedia entries relies on information available in the Wikipedia which limits results to people that actually have a Wikipedia entry and at the same time introduces noise by providing false positives through non-contemporary persons that have (vaguely) similar names. To mitigate this issue, we currently implement mass annotation features that will speed up manual annotation and cleaning but plan to further investigate the development of NEL tools that take historical context into account.

The "Keyword mentions" layer allows users to plot the occurrence of user defined keywords across the corpus and the selected bucket range. Users can combine keywords through logical operators (AND, OR) and can add multiple layers to visualize the occurrence over time across the corpus (see figure 6).

Both visualisations were used by the historian with a clear preference for the performance of the first view as illustrated in the following chapter. Nevertheless, we believe that the second visualisation (*Bucket of Explorables*) shows a significant potential for corpus exploration as it provides a unique and extendable view on a large number of pages over time. We foresee in particular further finetuning in the display of the documents within a bucket by including the ability to access document ranges with particular features. A first version of this additional filtering capability has been implemented but initial testing demonstrated the need for more granular access to the filter definitions.

We see further potential in the development of additional layers for the visualisation, allowing for example additional access to features of entities and more complex query mechanisms that we will implement and evaluate in the future.

[29] https://github.com/zalandoresearch/flair/blob/master/resources/docs/TUTORIAL_2_TAGGING.md#fast-english-models.

[30] see https://developers.google.com/knowledge-graph.

[31] see https://www.tic.ugent.be/content/2tbi-project.

ISSN 2416-5999, an open-access journal

Figure 6. View of the corpus with multiple layers, which enables e.g. the identification of parts of the corpus that score high in certain topics while containing a set of keywords.

III CASE STUDY GENERAL PARALYSIS

Developing an algorithm and interface to solve problems of an overabundance of historical documents with less than optimal scanning and OCR quality meant that the problem not only needed to be unpacked in all its facets — which we outlined above — but it also required a case study to validate our tool and to verify the usefulness of our approach and collaboration efforts.

3.1 Selecting general paralysis as a case study

Choosing a case study started with observing the different topics that had been generated by the topic modelling algorithm. The medical historian in our team evaluated the keyword output [case insanity disease mental symptom dementia form mania paralysis general disorder attack patient condition acute melancholia state treatment delusion epilepsy], as well as explored and close read multiple pages that were suggested by the system for this particular topic. This information correlated with a subject that we could identify and describe as "general paralysis", a common disease within psychiatric institutions during the nineteenth and early twentieth century. Furthermore, it was a stable and frequently present topic throughout the corpus (see figure 4).

However, to thoroughly vet our approach, we also required material against which to test our selected case study. Before illustrating the usefulness of the chosen topic modelling algorithm and visualisation facilities in *histograph* as an aid for researching the history of psychiatry, it is important to give the reader some background knowledge about general paralysis and the status of current historical research about this disease.

General paralysis was regularly witnessed within asylums in the nineteenth century, and had by some contemporaries been dubbed as "the disease of the nineteenth century".[32] Its physical

[32] The original quote in Dutch: "de ziekte der 19de eeuw" in: "Eenige beschouwingen over krankzinnigheid, hare oorzaken & hare behandeling door Dr. A. O. H. Tellegen, *Psychiatrische bladen*, 1884, volume 2, p13;

characteristics ranged from speech and writing impairment, over a diminishing of locomotion (i.e. difficulty with walking), ataxia and seizures, ultimately progressing into complete paralysis. Some of its mental symptoms were the presence of (grandiose) delusions as well as the onset of dementia, resulting in a total loss of intellectual capabilities.[33] Once a patient was diagnosed with the disease, it was a certain death sentence as no real cure was available.
Psychiatrists had difficulty with disambiguating and understanding general paralysis because of its resemblance to other diseases. Its symptoms, causes and treatment stayed for the larger part of the nineteenth and the beginning of the twentieth century a question mark for many physicians and psychiatrists, and was debated in extenso. Early on physicians laid a link with (an excess of) sexual "indulgences", but also many other factors were speculated to be the cause of general paralysis. Psychiatrists attributed a "fast life", prolonged mental efforts, excessive use of alcohol, hereditary, syphilis, and even sunstroke or a combination of them as a possible cause[34]. It was only from 1913 onwards that general paralysis would become known as neurosyphilis when its cause and the link with syphilis was acknowledged by the medical world. Before this revelation, general paralysis was for a considerable amount of time categorised as a separate disease entity with its own symptoms and disease pattern, while the relationship with syphilis drifted in and out of focus in the medical world across Europe.

3.2 Topic model keyword explorations and research validation

As mentioned earlier, the link between general paralysis and syphilis only slowly gained traction amongst physicians and psychiatrists, which is something we can also observe within the keyword lists of our topic models. We can observe that in the topic called "general paralysis", the words "syphilitic" and "syphilis" are listed as keywords. Furthermore, another topic that was identified by our algorithm consisted of the keywords [paralysis, general, syphilis, reaction, fluid, case, spinal, positive, syphilitic, blood, serum, cerebro, test, negative, paralytic, disease, wassermann, cent, result, organism]. This topic could be identified by the historian as either "Syphilis" or as "general paralysis". In a first instance we categorised it as the former since the Wasserman test used to detect syphilis (keywords: wasserman, serum, reaction, fluid, test) was a clear reference to this disease. In a later stage we renamed this topic into "Syphilis/general paralysis" because both became in the early twentieth century so closely connected and keywords related to general paralysis were also visible within the list of terms.
This primary exploration illustrates that the diagnostic and conceptual difficulties accompanying this disease shimmers through the keywords and topics proposed by the system. It also highlights that the annotation of topics with a keyword is not only a convenience on the level of the interface but also a critical step of historical interpretation of simple keyword lists that require reflection. When we investigate the aspect of time for the topic "general paralysis" and "syphilis/general paralysis" via keyword tracking, a couple of interesting aspects came to the surface (figure 7).[35] The "general paralysis" topic shows that words related to syphilis only show up sporadically from the 1890's onwards. In addition, the "syphilis/general paralysis" topic shows how keywords related to syphilis are more present

---

"Bijdrage tot de statistiek der dementia paralytica in Nederland door v. C.", *Psychiatrische bladen*, 1884, volume 2, p.55-56.

[33] A full overview on the range of symptoms can be found in:(Davis 2008, 87–96).

[34] "The Pathology of General Paresis. By W. H. O. Sankey,M.D. Lond., Medical Superintendent, Female Department, Middlesex County Asylum, Hanwell", *The Journal of Mental Science*, 1864, volume 9, number 48, p. 467-493; "Pseudo-General Paralysis. By THEO. B. HYSLOP, M.D., Assistant Physician, Bethlem Royal Hospital", *The Journal of Mental Science*, 1896, volume 42, number 177, p. 314.

[35] This exploration was done using text files to track keyword changes throughout the years. These files were made by the computer scientist to explore and assess the correctness of our pipeline (the tracking of keyword changes is not yet available within the HG environment due to its internal structure).

from the 1890's onwards. Also note that, for our second topic, the system mostly shows years from the second half of the nineteenth century as relevant for this specific topic. In essence it is a continuation of the previously mentioned general paralysis topic, while at the same time illustrating that a change occurred in how psychiatrists talked about general paralysis and syphilis as (a standalone) medical subject(s).

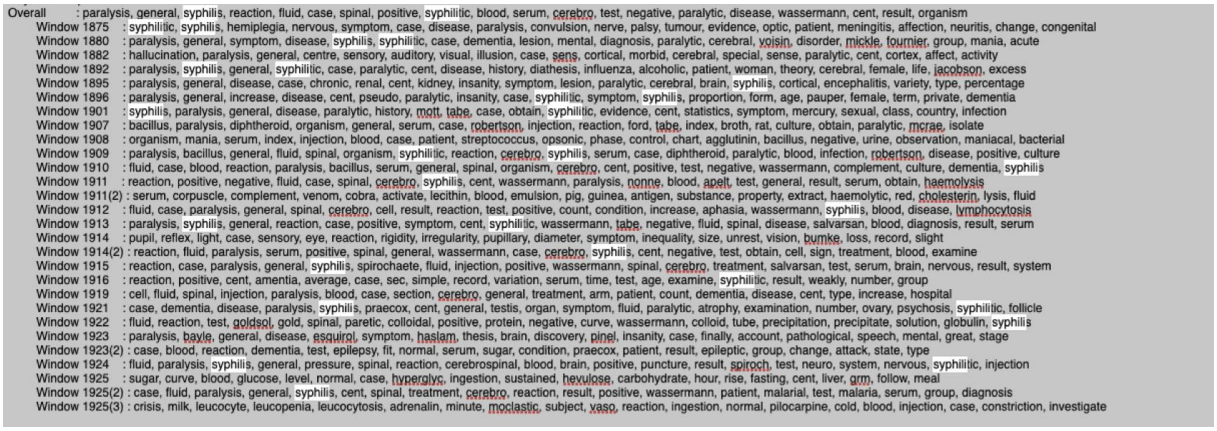


Figure 7. list of keywords per year for the topic "syphilis/general paralysis". The tracking of keyword changes through time was done using plain text files. These files were made by the NLP expert to explore and assess the correctness of part of our topic modeling algorithm. The tracking of keyword changes is not yet available within the histograph environment due to its specific internal structure.

While we could interpret these findings as an indication that the system is capable to capture the delicacy of historical subjects, it should be noted, that topic modelling remains a rather coarse tool. Due to its statistical nature only very strong "signals" will create an output and their interpretation is often not as straightforward as in our case.

3.3 Validating existing knowledge and beyond

In order to validate our system from a content point of view, we compared it with current historical research about general paralysis. We especially contrasted our outcomes with the findings of Juliet Hurn (Hurn 1998). Her thesis deals partially with the same corpus which she had to manually investigate via the table of contents or indexes of these journals.[36] In addition, she asked similar questions about general paralysis and syphilis. The comparison of existing research with our explorations (keyword tracking, reading through the highest scoring pages, etc.) proved to be consistent with the broad lines depicted in existing historiography. We were in particular able to confirm that the pages used by Hurn in her thesis were also picked up by our algorithm and scored accordingly for the specific topic.

While confirmation of existing knowledge is a necessary precondition to demonstrate that new digital tools are capable to reproduce this knowledge, our goal was not limited to a reproduction of the status quo but also to identify how tools can enable researchers to contribute to existing research by identifying new and relevant content in large amounts of source material.

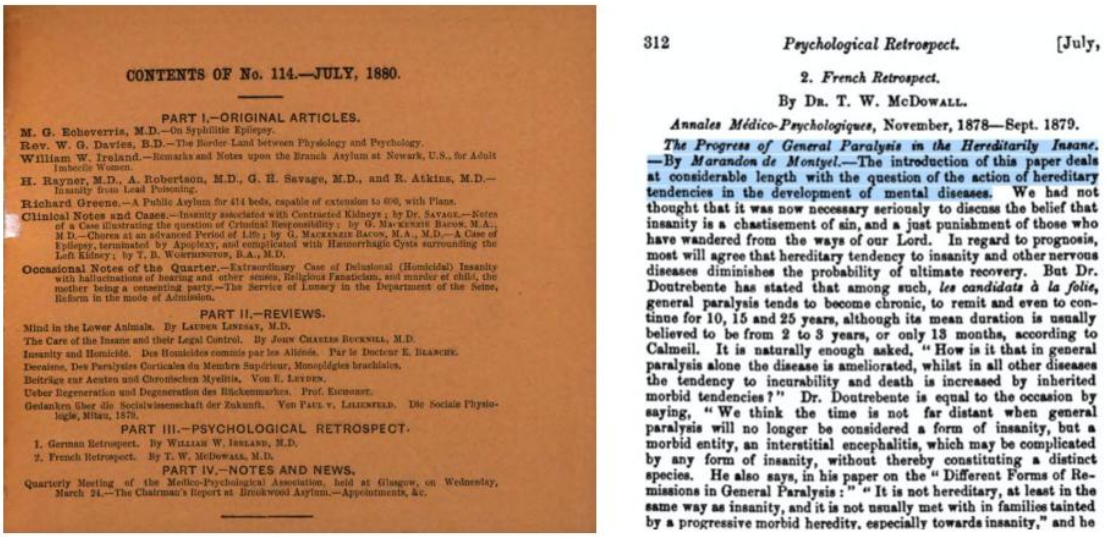
The combination of topic modelling and the exploration of the corpus through *histograph* highlighted two particular domains where an application of the toolchain provided new inputs for historical research.

[36] Keep in mind that this is something we only presume since the author does not reveal how she found or used the sources incorporated in her thesis. However, the use of the physical table of contents and indexes seems the most likely course of action for the period in which this thesis was written.

1. In the identification of "hidden" pages that are not captured through classical approaches such as table of content analysis
2. Through limiting the search space by narrowing down the amount of relevant pages in combination with topic modelling and keyword search

3.4 Exploring "Hidden" pages

One of the added values of our topic modelling and page ranking algorithm, is that it can supply us with more specific and relevant content. The basis of historical research are the sources that are available to the historian, which is furthermore guided by the particular *ways* through which we have access to them (i.e. physical or digital; in full or only partially). To the best of our knowledge, Judith Hurn only had the opportunity to manually investigate the BJP journal to acquire relevant content. This is not to say that researchers should not use the table of contents at all, far from it, as these finding aids give the researcher an idea of the original structure of the information. Nonetheless, human-made indexes and table of contents can contain mistakes (keywords that are missing, wrong page number attributions, etc). In addition, the historian could, while sifting through these indexes and table of contents, accidently skip across useful material. Furthermore, article titles do not capture the full breadth of an article and it is difficult to judge from the title alone if it could contain useful information. Instead, our system allows the researcher to not only rely on human made indexes or table of contents, but to trace the presence of themes and ideas in the full corpus.
In particular we were also able to identify pages in other sections of the journal which were not visible or accessible through analogue table of contents and index analysis.
For example, the index of the year 1880[37], only contains a limited number of articles that would be considered relevant in a TOC analysis (figure 8), such as for example the article "On Syphilitic Epilepsy" shown in the *original article* section of the journal.



CONTENTS OF No. 114.—JULY, 1880.

PART I.—ORIGINAL ARTICLES.

M. G. Echeverria, M.D.—On Syphilitic Epilepsy.
Rev. W. G. Davies, B.D.—The Border Land between Physiology and Psychology.
William W. Ireland.—Remarks and Notes upon the Branch Asylum at Newark, U.S., for Adult Imbecile Women.
H. Rayner, M.D., A. Robertson, M.D., G. H. Savage, M.D., and R. Atkins, M.D.—Insanity from Lead Poisoning.
Richard Greene.—A Public Asylum for 414 beds, capable of extension to 600, with Plans.
Clinical Notes and Cases.—Insanity associated with Contracted Kidneys; by Dr. Savage.—Notes of a Case illustrating the question of Criminal Responsibility; by G. Mackenzie Bacon, M.A., M.D.—Chorea at an advanced Period of Life; by G. Mackenzie Bacon, M.A., M.D.—A Case of Epilepsy, terminated by Apoplexy, and complicated with Hæmorrhagic Cysts surrounding the Left Kidney; by T. B. Worthington, B.A., M.D.
Occasional Notes of the Quarter.—Extraordinary Case of Delusional (Homicidal) Insanity with hallucinations of hearing and other senses, Religious Fanaticism, and murder of child, the mother being a consenting party.—The Service of Lunacy in the Department of the Seine, Reform in the mode of Admission.

PART II.—REVIEWS.

Mind in the Lower Animals. By Lauder Lindsay, M.D.
The Care of the Insane and their Legal Control. By John Charles Bucknill, M.D.
Insanity and Homicide. Des Homicides commis par les Aliénés. Par le Docteur E. Blanche.
Decaisne, Des Paralysies Corticales du Membre Supérieur, Monoplégies brachiales.
Beitrage zur Acuten und Chronischen Myelitis. Von E. Leyden.
Ueber Regeneration und Degeneration des Rückenmarkes. Prof. Eichhorst.
Gedanken über die Socialwissenschaft der Zukunft. Von Paul v. Lilienfeld. Die Sociale Physiologie, Mitau, 1879.

PART III.—PSYCHOLOGICAL RETROSPECT.

1. German Retrospect. By William W. Ireland, M.D.
2. French Retrospect. By T. W. McDowall, M.D.

PART IV.—NOTES AND NEWS.

Quarterly Meeting of the Medico-Psychological Association, held at Glasgow, on Wednesday, March 24.—The Chairman's Report at Brookwood Asylum.—Appointments, &c.

312 *Psychological Retrospect.* [July,

2. *French Retrospect.*

By Dr. T. W. McDowall.

Annales Médico-Psychologiques, November, 1878—Sept. 1879.

The Progress of General Paralysis in the Hereditarily Insane.—By *Marandon de Montyel.*—The introduction of this paper deals at considerable length with the question of the action of hereditary tendencies in the development of mental diseases. We had not thought that it was now necessary seriously to discuss the belief that insanity is a chastisement of sin, and a just punishment of those who have wandered from the ways of our Lord. In regard to prognosis, most will agree that hereditary tendency to insanity and other nervous diseases diminishes the probability of ultimate recovery. But Dr. Doutrebente has stated that among such, *les candidats à la folie*, general paralysis tends to become chronic, to remit and even to continue for 10, 15 and 25 years, although its mean duration is usually believed to be from 2 to 3 years, or only 13 months, according to Calmeil. It is naturally enough asked, "How is it that in general paralysis alone the disease is ameliorated, whilst in all other diseases the tendency to incurability and death is increased by inherited morbid tendencies?" Dr. Doutrebente is equal to the occasion by saying, "We think the time is not far distant when general paralysis will no longer be considered a form of insanity, but a morbid entity, an interstitial encephalitis, which may be complicated by any form of insanity, without thereby constituting a distinct species. He also says, in his paper on the "Different Forms of Remissions in General Paralysis:" "It is not hereditary, at least in the same way as insanity, and it is not usually met with in families tainted by a progressive morbid heredity, especially towards insanity," and he

Figure 8. Left: table of contents from the BJP of 1880 (26/113-115) with only one or two articles relating to general paralysis or syphilis. Right: within the section of "psychological retrospect" we find information extracted from a French journal about general paralysis. This could not have been discovered via the table of contents

When we explore the same year of the BJP with the topic modelling scores, we can identify pages that carry titles or headings that do not explicitly refer to general paralysis or syphilis,

[37] "Contents of no. 114 - July 1880", The Journal of Mental Science, 1880, volume 26, number 113-115.

but instead are called for example "psychological retrospect".[38] A strict table of contents analysis therefore misses this particular section of information about general paralysis.
Such weak and "hidden" signals are just as important to form an idea about the history of general paralysis and syphilis as are the original or translated articles we can find in the BJP. Especially since it gives us a more nuanced view of where the information that was distributed amongst British physicians originated from, as well as what foreign physicians found important to inform British psychiatrists about. These features enable the historian to trace the finer and sometimes implicit connections and references that are made to general paralysis, expanding and fine tuning our understanding of the topic at hand. This means that the information that Juliet Hurn derived from the journals, and the information we extracted is somewhat different. Hurn for example placed an emphasis on French influences (Hurn 1998, 24), but through the sources compiled by our algorithm this should be re-assessed to a certain extent. The pages selected by our system show that information did not only arrive in Britain via France, but also in some instances from Germany. Not only did the BJP include reviews of German books, but also translations of German articles, which indicates the importance British psychiatrists placed upon this German research. All of this indicates that German influences should be studied more. The possible importance of this German influence, will also resurface in the following section about "limiting the search space".
The possibility of finding "hidden pages" is a very important feature of histograph, however it is difficult to absolutely quantify how many pages are identified (and how many are still missing) because the creation of a gold standard that we could use to measure the amount of hidden pages would require a stable set of characteristics. This kind of definition is not possible because what qualifies as a hidden page, and how many there are, largely depends on the topic under study and the specific facets the historian is interested in, in a given moment. Hence we would be only able to give a number of hidden vs visible pages for a very specific research subject and under a significant time investment of the historian, as they would have to perform close reading and annotation of the corpus to create this gold standard. While this time investment was not feasible within the limits of the project, we will explore and review this issue in future research.

3.5 Limiting the search space

To get a more precise idea of the development of the syphilis-general paralysis connection it is useful to specifically look at those pages within the corpus that mention both words together. The "general paralysis" topic for example contains 46.847 pages that include information on this topic. As therefore almost all pages of the corpus have a score for the topic – even though this score diminishes tremendously towards the end of the set – further indicators become necessary to identify relevant pages within the corpus. The occurrence of specific keywords became a very practical indicator to limit the amount of pages that need to be reviewed. Histograph fosters an iterative approach where a user selects and tests different keywords directly in the interface and sees the results plotted across the full corpus in time. By combining ["paralysis + syphilis"] or ["paralytic + syphilitic"] within the general paralysis topic, only pages are shown where both words occur. This not only leads to even more relevant pages to explore but also to *less* information that needs to be consulted. The "paralysis + syphilis" keywords resulted into 1103 pages, the "paralytic + syphilitic" keywords result in 1455 pages to explore (see figure 9).

[38] These 'retrospects' were often a mix of foreign journals that were consulted by the editors of the BJP or correspondence written by designated foreign psychiatrists. Both made it possible to give a brief recap of the interesting and useful information that could be gleaned from other journals or countries.

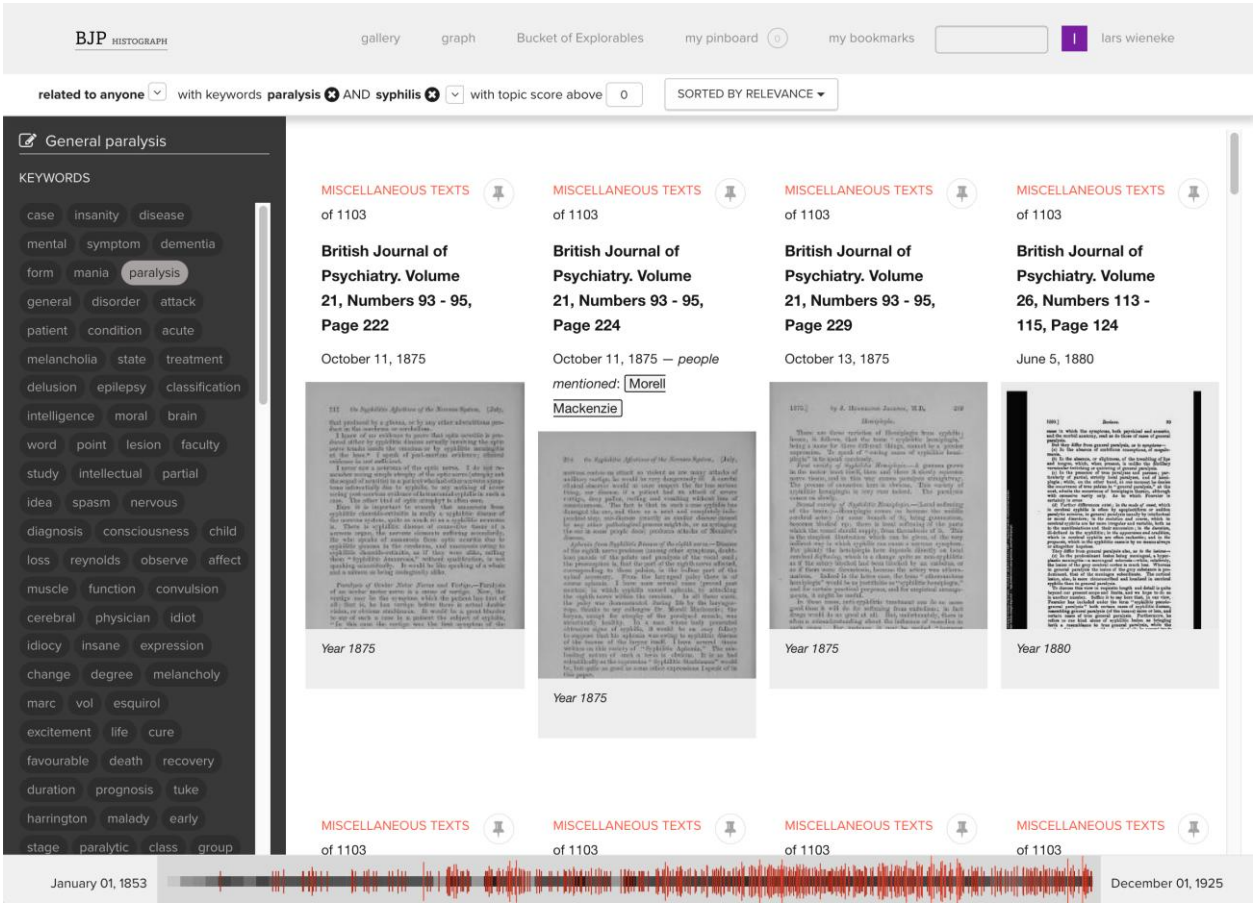


Figure 9: keyword exploration of co-occurrences of "syphilis" and "paralysis" in the topic view printed over time

What became overwhelming while reading current historical research about general paralysis and syphilis, is the repetitiveness of the story that has been told. A constant rehashing of material. More specifically, it tells a whiggish, triumphalist history wherein only a few important physicians are highlighted (Hurn 1998; Kragh 2010; Pearce 2012; Tuong-Vi, Kompanje, and van Praag 2013; Stewart et al. 2017; Swain 2018). People like Esmarch and Jessen, Fournier, Wasserman, or Noguchi and Moore have taken up a central position within the history of general paralysis because of their groundbreaking contributions to this disease. However, they were not the only ones interested in solving the medical mysteries this subject entailed.

When analysing the high ranked pages in *histograph*, we see that some of these articles/pages were also investigated in former historical research. While the information on these particular pages are in line with current historiography, its content can in addition be used as a basis for further (re)examination through a more focused use of the "keyword mentions" feature in *histograph*. One specific article for example mentioned more than 85 psychiatrist who were concerned with syphilis and general paralysis.[39] No historian has taken an interest in these different psychiatrists nor did they take (or have) the opportunity to investigate the relevance of these European psychiatrists throughout a substantial corpus. Using the "keyword mentions" in *histograph* allows us to limit the search space to target very specific pages where these persons are mentioned in the relevant context. In prospect, a better quality of NER/NEL

[39] "General Paralysis and syphilis: a Critical Digest. By W.H.B. Stoddart, M.D. P.R.C.P.", The Journal of Mental Science, 1901, volume 47, number 198, p. 445.

that correctly annotates relevant persons could further streamline the process of drilling down to relevant pages.

To further experiment with this, we selected one name (Karl Friedrich Otto Westphal, an important German psychiatrist) from the earlier mentioned list. Through logical reasoning we know that Westphal's name will most likely be mentioned in relation to a very diverse range of topics. To counteract this, we opened the "general paralysis" topic and specified that the system should only search for documents where "Wespthal", "syphilis" and "paralysis" co-occur on the same page. This resulted in an output of 13 pages. While opening the first page from within this specific selection we read the following: "[...] Several German writers, Meyer, Westphal, Oedmansson, and Griesinger, have specially studied this kind of case [general paralysis]: and, while they do not altogether agree as to their conclusions, the weight of evidence is in favour of the view that those are cases of general paralysis, who, having previously had syphilis, have the character of their symptoms influenced by it to some extent, a casual re-lation only existing between the two diseases".[40] This statement was made in 1873 in a lecture by the British physician David Skae. This once more confirms that German physicians were actively involved in the research of general paralysis and that, although we have forgotten many of them now, psychiatrists at the time were well aware of whom across Europe participated in solving the medical questions surrounding this disease.

CONCLUSION AND FUTURE WORK

While historians and humanists are grateful for the availability of sources through online repositories and archives, this can, nonetheless, create various research obstacles. One of these is that researchers can easily lose focus and oversight while exploring large corpora, such as *The British Journal of Psychiatry* containing more than 50000 pages. A potential remedy to these challenges lies in the development of new tools that are tightly aligned with the research questions at hand and support the researcher in the exploration task. As discussed in this paper, we experimented with different forms of natural language processing – in particular topic modelling – and data visualisation to provide new modes for interrogating the corpus. Throughout this probing exercise, communication between the four team members was crucial for the completion of this project.

In terms of text pre-processing of the source documents our current work did not include specific measures to clean up the text itself. We can observe that overall the topics make sense but face occasional occurrences of non-words. Cleaning of the OCR errors would improve the text consistency and allow us to generate more informative topics.

On the part of developing new interfaces for historical research, our approaches to distant reading (*Bucket of Explorables*, *keyword time tracking*) can provide the historian a first "lay of the land", a map to navigate a corpus through dominant topics, especially in cases wherein the researcher is not familiar with the corpus. Quick overviews like this provide a distant reading or abstraction that cannot be acquired through either analogue or digital close reading. Furthermore, these mechanisms can help to observe changes across multiple journals within the same country or across borders. This can aid researchers for example to identify the inception of technical terms and, later on via close reading, the particular definitions and points of view of psychiatrists and physicians that lay behind these terms. The acceptance of the general paralysis-syphilis connection for example, is often seen as globally acknowledged between the late 1890's and early 1910's. Our system could help historians to be more precise about these developments in other journals or countries by making it easier to identify and

[40] "The Morrisonian Lectures on Insanity for 1873. By the late David Skae, M.D., F.R.C.S.E., Physician-Superintendent of the Royal Edinburgh Asylum, &c., &c. Edited by T. S. Clouston, M.P., F.R.S.E., Lecture V", The Journal of Mental Science, 1875, volume 21, number 93, p. 4.

ISSN 2416-5999, an open-access journal

review relevant sections of the corpus without manually having to skim read through large amounts of sources.

Observing similar page suggestions from the topic modelling in comparison to existing historical knowledge provided a baseline from which we were able to further investigate how the topic modelling algorithm and the *histograph* interface can support the exploration of sources and advance historical research. The main contribution of our approach to historical research lies within its ability to re-assess existing historical knowledge by allowing a more fine-grained view of the historical developments that took place, contributing to a (slightly) different historical reading of the development, conceptualisation and contextualisation of general paralysis throughout the decades. However, the interpretation of the page content remains the historian's task and can not be replaced by the machine. Algorithms and visualisations in *histograph* remain only a tool for facilitating research by pointing out potentially interesting instances to the historian.

Aside from these results, our project will continue to follow up on some of the new research challenges we discovered in the course of the project and we actively aim in providing the tool to a broad range of users[41]. Through constant feedback within our team, we are continuously evaluating our approach and the results, as well as planning new features to be implemented and tested. Three challenges will be reviewed in particular:

1. Automatic selection of the most suitable number of topics. Currently, this number is chosen based on an agreement between automatic evaluation of the topic coherence, and expert opinion. As an alternative to this semi-supervised approach, we plan to apply algorithms such as Hierarchical Dirichlet Process (C. Wang, Paisley, and Blei 2011) or divisive clustering, for example Bi-secting k-means (Steinbach, Karypis, and Kumar 2000) or Rank-2 non-negative matrix factorization (Kuang and Park 2013), which may help to set the most appropriate number of topics automatically. Introducing hierarchical processing will also have the advantage of assembling the topics into groups of different granularity, allowing an historian to perceive semantic structure of the data.
2. The integration of "autocomplete topics". During discussions with the medical historian in our team, it was pointed out that, although most of the generated topics were useful for further historical research, the historian knew through her expertise in the domain of medical and psychiatric history that the topic modelling did not always capture every single subject that nineteenth century psychiatrists were involved with. To support the exploration of the corpus we therefore want to integrate a custom autocomplete feature based on word embeddings. Through this approach we will be able to identify relevant co-located terms within the corpus with higher granularity than provided by the topic modeling (because the latter already gives us collocated terms). In a next step we want to use this custom query to retrieve documents with a high similarity to the newly build query. At the same time, the terms can be also used to reformulate keyword queries for the fulltext search.
3. Creating a generic tool. The problems we encountered during this extended probing exercise are issues that many historians and humanists face within their research, irrespective of the sources they use or their specific research domain. Due to this potential we aim to build a more generic version of the tool. This will imply the application of the tool to a wider variety of historical research projects to better understand the limits and the potential of a generic approach.

[41] The current version of the tool as well as installation and use instructions are available here: https://github.com/C2DH/histograph as of the time of writing, the current version does not yet contain an automatic pipeline for topic extraction but allows to import topic scores as annotations.

4. In depth testing of a topic model for configurations with a larger number of topics and for different language corpora. We recently started experimenting with the creation from 4 up to 20 different topics to better understand the impact of variable topic numbers for the exploration of corpora. Allowing for the option of different numbers of topics makes the exploration of certain themes even more fine-grained. Besides this experiment, we have also started to explore topic modelling for corpora in other languages (French). These experiments are still in an early stage. However, it became apparent rather quickly that certain topics (i.e. "the criminally insane" or "general paralysis") can be clearly observed in the British as well as the Belgium corpus, signifying a certain unity in psychiatrists' interests across different countries.

References

Agrawal, Amritanshu, Wei Fu, and Tim Menzies. 2018. "What Is Wrong with Topic Modeling? (And How to Fix It Using Search-Based Software Engineering)." *Information and Software Technology* 98 (June): 74–88. https://doi.org/10.1016/j.infsof.2018.02.005.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. 2018. "Contextual String Embeddings for Sequence Labeling." In *COLING*, 1638–49. https://aclanthology.coli.uni-saarland.de/papers/C18-1139/c18-1139.

Anthony, Laurence. 2019. *AntConc (Version 3.5.8)*. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/antconc/.

Belford, Mark, Brian Mac Namee, and Derek Greene. 2018. "Stability of Topic Modeling via Matrix Factorization." *Expert Syst. Appl.* 91: 159–169. https://doi.org/10.1016/j.eswa.2017.08.047.

Bewley, Thomas. 2008. *Madness to Mental Illness: A History of the Royal College of Psychiatrists*. London: RCPsych Publications.

Beykikhoshk, Adham, Ognjen Arandjelovic, Dinh Q. Phung, and Svetha Venkatesh. 2018. "Discovering Topic Structures of a Temporally Evolving Document Corpus." *Knowl. Inf. Syst.* 55 (3): 599–632. https://doi.org/10.1007/s10115-017-1095-4.

Blei, David M., and John D. Lafferty. 2006. "Dynamic Topic Models." In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, 113–120. https://doi.org/10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *J. Mach. Learn. Res.* 3: 993–1022.

Cui, Weiwei, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. 2011. "TextFlow: Towards Better Understanding of Evolving Topics in Text." *IEEE Trans. Vis. Comput. Graph.* 17 (12): 2412–21.

Damerow, Julia, and Dirk Wintergrün. 2019. "The Hitchhiker's Guide to Data in the History of Science." *Isis* 110 (3): 513–21. https://doi.org/10.1086/705497.

Davis, Gayle. 2008. *"The Cruel Madness of Love": Sex, Syphilis and Psychiatry in Scotland, 1880-1930*. The Wellcome Series in the History of Medecine. Amsterdam; New York: Editions Rodopi.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/n19-1423.

Düring, Marten, Lars Wieneke, and Vincenzo Croce. 2015. "Interactive Networks for Digital Cultural Heritage Collections - Scoping the Future of HistoGraph." In *Engineering the Web in the Big Data Era*, edited by Philipp Cimiano, Flavius Frasincar, Geert-Jan Houben, and Daniel Schwabe. Vol. 9114. Lecture Notes in Computer Science. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-19890-3.

Ehrmann, Maud, Estelle Bunout, and Marten Düring. 2017. "Historical Newspaper User Interfaces: A Review." In . http://library.ifla.org/2578/.

Graham, Shawn, Ian Milligan, and Scott Weingart. 2015. *Exploring Big Historical Data*. IMPERIAL COLLEGE PRESS. https://doi.org/10.1142/p981.

Greene, Derek, and James P. Cross. 2016. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *CoRR* abs/1607.03055. http://arxiv.org/abs/1607.03055.

Greene, Derek, Derek O'Callaghan, and Padraig Cunningham. 2014. "How Many Topics? Stability Analysis for Topic Models." In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, 498–513. https://doi.org/10.1007/978-3-662-44848-9_32.

Gries, Stefan Th. 2009. "What Is Corpus Linguistics?" *Language and Linguistics Compass* 3 (5): 1225–1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x.

Guldi, Jo, and Johnson-Robertson, Chris. 2019. *Paper Machines*. http://papermachines.org.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. "Studying the History of Ideas Using Topic Models." In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the*

Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A Meeting of SIGDAT, a Special Interest Group of the ACL, 363–371. https://www.aclweb.org/anthology/D08-1038/.
Hurn, Juliet D. 1998. *The History of General Paralysis of the Insane in Britain, 1830 to 1950*. University of London.
Kragh, Jesper Vaczy. 2010. "Malaria Fever Therapy for General Paralysis of the Insane in Denmark." *History of Psychiatry* 21 (4): 471–86. https://doi.org/10.1177/0957154X09338085.
Kuang, Da, and Haesun Park. 2013. "Fast Rank-2 Nonnegative Matrix Factorization for Hierarchical Document Clustering." In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, 739–747. https://doi.org/10.1145/2487575.2487606.
Lee, Daniel D., and H. Sebastian Seung. 1999. "Learning the Parts of Objects by Nonnegative Matrix Factorization." *Nature* 401: 788–791.
Luo, Minnan, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, and Qinghua Zheng. 2017. "Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling." In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2308–2314. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14469.
Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. USA: Cambridge University Press.
Mäntylä, Mika V., Maëlick Claes, and Umar Farooq. 2018. "Measuring LDA Topic Stability from Clusters of Replicated Runs." In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2018, Oulu, Finland, October 11-12, 2018*, 49:1–49:4. https://doi.org/10.1145/3239235.3267435.
Michel, Jean-Baptiste, Yuan Shen, Aviva Aiden, Adrian Veres, Matthew Gray, Joseph Pickett, Dale Hoiberg, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science (New York, N.Y.)* 331: 176–82. https://doi.org/10.1126/science.1199644.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781.
Mimno, David M. 2012. "Computational Historiography: Data Mining in a Century of Classics Journals." *JOCCH* 5 (1): 3:1–3:19. https://doi.org/10.1145/2160165.2160168.
Mimno, David M., Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A Meeting of SIGDAT, a Special Interest Group of the ACL*, 262–272. https://www.aclweb.org/anthology/D11-1024/.
Newman, David J., and Sharon Block. 2006. "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." *JASIST* 57 (6): 753–767. https://doi.org/10.1002/asi.20342.
Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. "Automatic Evaluation of Topic Coherence." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. HLT '10. USA: Association for Computational Linguistics.
Novak, Jasminko, Lars Wieneke, Marten Düring, Isabel Micheel, Mark Melenhorst, Javier Garcia Morón, Chiara Pasini, Marco Tagliasacchi, and Piero Fraternali. 2014. "HistoGraph: A Visualization Tool for Collaborative Analysis of Historical Social Networks from Multimedia Collections." *IV2014 - DHKV: Cultural Heritage Knowledge Visualisation*, 2014.
O'Callaghan, Derek, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. "An Analysis of the Coherence of Descriptors in Topic Modeling." *Expert Syst. Appl.* 42 (13): 5645–5657. https://doi.org/10.1016/j.eswa.2015.02.055.
Pearce, J. M. S. 2012. "Brain Disease Leading to Mental Illness: A Concept Initiated by the Discovery of General Paralysis of the Insane." *European Neurology* 67 (5): 272–78. https://doi.org/10.1159/000336538.
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "Glove: Global Vectors for Word Representation." In *In EMNLP*.
Pyysalo, S., F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. 2013. "Distributional Semantics Resources for Biomedical Text Processing." In *Proceedings of LBM 2013*, 39–44.
Riedl, Martin, and Chris Biemann. 2012. "Text Segmentation Wit h Topic Models." *JLCL* 27 (1): 47–69.
Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, 399–408. https://doi.org/10.1145/2684822.2685324.
Rosenzweig, Roy. 2003. "Scarcity or Abundance? Preserving the Past in a Digital Era." *The American Historical Review* 108 (3): 735–762.
Schofield, Alexandra, M\aans Magnusson, and David M. Mimno. 2017. "Pulling Out the Stops: Rethinking Stopword Removal for Topic Models." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, 432–436. https://doi.org/10.18653/v1/e17-2069.
"Stanford NLP Tools." n.d. Accessed April 20, 2015. http://nlp.stanford.edu/software/.
Steinbach, M., G. Karypis, and V. Kumar. 2000. "A Comparison of Document Clustering Techniques." In *KDD Workshop on Text Mining*. http://citeseer.ist.psu.edu/steinbach00comparison.html.
Stewart, Mary, Joseph Debattista, Lisa Fitzgerald, and Owain Williams. 2017. "Syphilis, General Paralysis of the Insane, and Queensland Asylums." *Health and History* 19 (1): 60–79. https://doi.org/10.5401/healthhist.19.1.0060.
Steyvers, Mark, Tom Griffiths, and W. Kintsch. 2006. "Probabilistic Topic Models." In *Latent Semantic Analysis: A Road to Meaning*, edited by T. Landauer, D. McNamara, and S. Dennis. Laurence Erlbaum.

http://books.google.de/books?hl=de&lr=&id=JbzCzPvzpmQC&oi=fnd&pg=PA427&dq=Probabilistic+Topic+Models&ots=aMG3M0R_IK&sig=DpWCn5nSwZdnO-NYvZts27g9wu0#v=onepage&q=Probabilistic%20Topic%20Models&f=false.

Swain, Kelley. 2018. "'Extraordinarily Arduous and Fraught with Danger': Syphilis, Salvarsan, and General Paresis of the Insane." *The Lancet Psychiatry* 5 (9): 702–3. https://doi.org/10.1016/S2215-0366(18)30221-9.

"The British Journal of Psychiatry." n.d. Cambridge Core. Accessed January 26, 2020. https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/all-issues.

"The Journal of Mental Science." n.d. Archive.Org. Accessed January 26, 2020. https://archive.org/search.php?query=creator%3A%22Royal+Medico-psychological+Association%22.

Tschuggnall, Michael, Efstathios Stamatatos, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2017. "Overview of the Author Identification Task at PAN-2017: Style Breach Detection and Author Clustering." In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*. http://ceur-ws.org/Vol-1866/invited_paper_3.pdf.

Tuong-Vi, Nguyen, Erwin J.O. Kompanje, and Marinus C.G. van Praag. 2013. "Enkele Mijlpalen Uit de Geschiedenis van Syfilis." *Nederlandsch Tijdschirft Voor Geneeskunde* 157 (A6024). https://doi.org/10.1086/429626.

Wang, Chong, John Paisley, and David Blei. 2011. "Online Variational Inference for the Hierarchical Dirichlet Process." In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, edited by Geoffrey Gordon, David Dunson, and Miroslav Dudík, 15:752–760. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR. http://proceedings.mlr.press/v15/wang11a.html.

Wang, Xuerui, and Andrew McCallum. 2006. "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends." In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, 424–433. https://doi.org/10.1145/1150402.1150450.

Wieneke, Lars, Marten Düring, Ghislain Silaume, Carine Lallemand, Vincenzo Croce, Marilena Lazzarro, Francesco Nucci, et al. 2014. "Building the Social Graph of the History of European Integration." In *Social Informatics*, edited by Akiyo Nadamoto, Adam Jatowt, Adam Wierzbicki, and JochenL. Leidner, 8359:86–99. Lecture Notes in Computer Science. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-55285-4_7.

Yang, Yi, Shimei Pan, Yangqiu Song, Jie Lu, and Mercan Topkara. 2016. "Improving Topic Model Stability for Effective Document Exploration." In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 4223–4227. http://www.ijcai.org/Abstract/16/635.