

MANREET KAUL

HISTORY IN THE DIGITAL AGE

VOL 2

History in the Digital Age: Vol 2

History in the Digital Age: Vol 2

Manreet Kaul



Published by The InfoLibrary,
4/21B, First Floor, E-Block,
Model Town-II,
New Delhi-110009, India

© 2022 The InfoLibrary

History in the Digital Age: Vol 2
Manreet Kaul
ISBN: 978-93-5590-117-0

This book contains information obtained from authentic and highly regarded sources. All chapters are published with permission under the Creative Commons Attribution Share Alike License or equivalent. A wide variety of references are listed. Permissions and sources are indicated; for detailed attributions, please refer to the permissions page. Reasonable efforts have been made to publish reliable data and information, but the authors, editors and publisher cannot assume any responsibility for the validity of all materials or the consequences of their use.

Trademark Notice: All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy. Furthermore, the publisher ensures that the text paper and cover boards used have met acceptable environmental accreditation standards.

Table of Contents

Part 3	Close and Distant Reading Visualizations and Debates	
Chapter 1	Use of MiKARELIA Database in Study of Karelia Migration and Evacuees	2
Chapter 2	Digital Collections and History of Women's Suffrage Movements	21
Chapter 3	Audio Visual Heritage and the Digital Universe	36
Chapter 4	Digitised Parliamentary Talk and Nordic Allemansrätten	52
Chapter 5	Internationalism in the UK Parliament: Computer-assisted Analysis and Digital Methods	69
Chapter 6	Metadata of an Image Database: Use of Digital Methods to Map East German Opposition Movement	90
Chapter 7	Transit of Venues: Discourses of Astronomy	105
Chapter 8	Topic Modelling German Humanism with DTM and MALLET	126
Chapter 9	Computational Authorship Attribution in Manuscripts and Other Texts	145
Chapter 10	Towards the Sun of Socialism: Turning our Macroscopes	168
Part 4	Epilogue	

Chapter 11 Concluding Remarks: The Universal Borderlands of the
Digital Humanities

190

Part 3
Close and Distant Reading
Visualizations and Debates

Use of MiKARELIA Database in Study of Karelia Migration and Evacuees

Mirkka Danielsbacka, Lauri Aho, Robert Lynch,
Jenni Pettay, Virpi Lummaa and John Loehr

Introduction

The consequences of forced migrations are felt globally and are faced by millions of people each year. A critical question is how these refugees adjust to their new environments and eventually integrate into the host population. A number of factors can influence the ultimate assimilation of migrant populations and these are frequently related to the characteristics of the migrants (for example, demographic variables and socio-economic background), flight (for example, cause of flight), host country or region (for example, natural resources) and the resettlement policies of host populations. One way to measure the successful settlement and assimilation of displaced populations is to look at the number of times an individual relocates after their initial arrival in a host country and to analyse which factors affect these moves. In general, the more individuals move, the less likely they are to integrate.¹ In this

chapter, we investigate the evacuation of Karelia—a unique forced migration event that took place in Finland during the Second World War—and the socio-demographic and environmental factors associated with the relocation and settlement of Karelian evacuees during and after the war.

Our methodological approach builds on the socio-economic tradition of conducting digital history in which we use quantitative data and methods to analyse and interpret a question of historical importance (the assimilation of Karelian evacuees). Therefore, it might be more accurate to define the methodological approach used here as quantitative history, rather than ‘new digital history’, even though it can certainly be argued that all quantitative history is essentially digital history. In line with many of the new digital history projects, our data has been digitised and compiled with the help of new digital tools (see ‘Material and Methods’, below) and we have an interdisciplinary research team of biologists, computer scientists, sociologists and historians with considerable experience working on these and similar datasets. However, analysing these newly extracted data has been executed with rather common statistical methods (for example, regression models) that are in line with the quantitative tradition of social and economic history.

During the Second World War, an estimated 40 million Europeans fled their homes in what is widely considered to be the worst refugee crisis in modern history.² Finland faced this problem after it ceded Karelia to the Soviet Union in the aftermath of the Winter War (1939–1940) and once again after the Continuation War (1941–1944). Almost all Karelians were evacuated to the remaining parts of Finland. It has been said that Karelians have had the ‘sad privilege of being the only refugee group in the world to have been displaced three times within a short period of four years—1940–1944.’³ Two of these displacements were forced and resulted from the Soviet occupation of Karelia in the Winter War and at the end of the Continuation War, but one, during the Continuation War, was a voluntary migration back to recaptured Karelia.

Previous historical and cultural studies of Karelians have concentrated on describing the Karelian evacuees and their assimilation in Finnish society,⁴ Karelians’ memories of the evacuations and the land they lost in Karelia,⁵ and the resettlement policy of the Finnish government.⁶ In addition, previous sociological and epidemiological studies of Karelian evacuees have mainly focused on the long-term effects of forced migrations on mortality,⁷ income⁸ or socio-economic status⁹ by comparing displaced Karelians with the rest of the Finnish population. These studies have frequently been conducted with the same 10% sample data ($n = 411,629$) from the 1950 population census, which was the first full census implemented in Finland. Karelians can be extracted from these data because there is information on the place of residence from the year 1939 which is prior to the initial evacuation. A constraint of this dataset, however, is that it is limited to variables available in the years 1939, 1950 and in follow-up datasets from 1970s onwards. Studies conducted with these data have found that after the Second World War Karelian men had higher socio-economic standing

and higher income than their non-displaced counterparts.¹⁰ Sarvimäki and colleagues suggest that one reason for this is that Karelians were more likely to move from their initial placement areas to other regions in Finland in search of better employment opportunities than non-displaced Finns. This suggests at least one explanation for the finding that younger male Karelians reached a higher socio-economic position when compared to non-displaced Finnish males. In addition, displaced people transitioned faster from agrarian to modern occupations than non-displaced Finns, which could have also affected their improved socio-economic standing.¹¹ On the other hand, a study by Haukka and colleagues found that displaced Karelians had higher overall mortality and ischemic heart disease mortality than the rest of the Finnish population. It is interesting to note, however, that when compared to international research on the long-term effects of forced migration, Karelians had lower suicide rates.¹²

Previous studies have not used micro-level data to explore whether Karelian evacuees differed in their migration histories as a function of their background characteristics. Background characteristics of Karelians as well as environmental factors of a host country may associate with likelihood of migration; furthermore, migration history could reflect the assimilation of Karelians. More specifically we explore: Who moved back to Karelia when they had the opportunity? And who remained in western Finland? Which environmental factors affected the likelihood of return? How much, on average, did Karelians move after the second evacuation (that is, how easily did they settle after the Second World War)? What were the characteristics of the evacuees who moved most frequently and what factors predict faster assimilation?

Karelian Evacuees

Two separate wars were fought with the Soviet Union on the eastern border of Finland. The Winter War started on 30 November 1939, when the Soviet Union attacked Finland, and lasted until 13 March 1940. During this first war, Finland lost 11% of its land territory, including its second biggest city, Vyborg. The Soviet occupation of Karelia forced approximately 407,000 people to flee their homes and to be placed elsewhere in western Finland.

Before 1950, Finland was predominantly an agrarian country, and agricultural occupations were even more common in Karelia than in other parts of the country. Approximately 230,000 evacuees (57%) earned their living from agriculture. Not all of them were farmers, however, and some were agricultural workers who did not own the farm they worked on. These farms were, on average, smaller than farms in other parts of Finland.¹³

The initial placement of the evacuees was poorly planned and organised due to the sudden start to the Winter War and rapid advance of the Soviet troops. Migrants were initially housed in public buildings that were used as shelters and only later were transferred to private residences. In the summer of 1940, an

Emergency Settlement Act and compensation law were passed. With the settlement law, farmers could obtain new land to farm and, with the compensation law, the state would pay for the lost property.¹⁴

Land for evacuees was acquired from the state, the Church and municipalities, but it was also frequently seized from private owners. Although Finnish authorities attempted to carry out land acquisitions with voluntary purchases, many farmers were forced to sell their land. The purpose of the Emergency Settlement Act was not to ensure that Karelian farmers would be fully compensated for the land they had lost, but rather was to make sure that those Karelians who made their living from agriculture could continue to do so.¹⁵

Between the Winter and Continuation Wars, evacuees who made their living from agriculture, especially those who had their own farms in Karelia, had the hardest time adjusting because the Emergency Settlement Act forced them to wait before they received land. This may have caused additional friction between evacuees and the host population because of the hard labour shortage with which evacuees were expected to help.¹⁶ But it was the placement of Karelian evacuees among Swedish-speaking Finns that aroused the most criticism. This was because placing Finnish-speaking Karelians in bilingual municipalities could have endangered the delicate relationship between Swedish- and Finnish-speaking populations. The language question came to the fore once again after the Continuation War, when Karelians had to be settled permanently in the remaining parts of Finland.¹⁷

Because carrying out the Emergency Settlement Act was slow, only about 13,000 new small farms were actually founded and only 6,000 of these contracts were finalised by the summer of 1941, even though many more applications were received. With the onset of the Continuation War in the summer of 1941, Finnish troops reconquered the Russian occupied regions, which gave Karelians the opportunity to return to Karelia. Evacuees who had received emergency settlement farms were then allowed to cancel their contracts; more than half of them did so by March 1943 and returned to Karelia. Nevertheless, a few hundred households kept their emergency settlement farms and gave up their claims on their land in Karelia.¹⁸ Approximately 70% of the original evacuees (280,000) who had initially settled elsewhere in Finland voluntarily moved back to their previous home in Karelia, while the remaining 30% decided to remain in their new location. The number of evacuees who returned was higher in some locations of origin (for example, over 80% for Sortavala) and lower in others (for example, 40% to 60% for Viipuri). Farmers were more likely to return (~75%), and although returning to locations near the front line was not allowed, some disobeyed and returned anyway. A long period of trench warfare kept the front line quite stable from January 1942 until summer 1944 when the final Soviet offensive began.¹⁹

The Continuation War ended in the autumn of 1944 and the border was redrawn back to where it had been in 1940 and everyone who had returned to Karelia between the wars was evacuated once again.²⁰ This time, the evacuation

and placement of evacuees were much more systematic than they had been after the Winter War. In May 1945, the Parliament approved the Land Acquisition Act (*Maanhankintalaki*), which guided the settlement policy.²¹ According to the Act, groups that were entitled to receive land were evacuees who had made their living from agriculture, disabled soldiers, war widows, war orphans, soldiers who had served on the frontline or had a family and several other smaller groups. Evacuees submitted almost 48,500 applications for land; 92% of these were accepted and evacuees were placed in certain initial placement areas. In the summer of 1945, they started to move to other places, partly because they were ordered to and partly due to their own initiative. The official placement plan only applied to the agricultural population, which meant that townspeople and industrial workers were free to choose where they wanted to settle. Resettlement of the agricultural population was based on the idea that people from the same villages would be able to stay in the same areas and that their placements would correspond to the climatic, economic and religious circumstances of the area from which they were evacuated. The official placement plan was only applied in its strictest form to farmers, and among them those who were entitled to farm. This constituted about 35% of all evacuees. Although they were in the minority, the final resettlement plan resulted in most of the farmers having to move again. As a result, in the years immediately following the war, movements may have been more prevalent among farmers than other evacuees.²²

Material and Methods

Here we use the recently digitised Migration Karelia (MiKARELIA) database, which contains over 160,000 adult Karelians and a wide range of data on births, marriages, occupations and movements of these forced migrants. The original source material for the database comes from a register compiled in the book series *Siirtokarjalaisten tie* (Anon. 1970; the title directly translates to: *Karelian migrants' road*), which systematically recorded the experiences of evacuees.

Interviews took place between 1968 and 1970 and were performed by approximately 300 trained interviewers. Each entry lists the full name (maiden name if applicable), profession, birth date, birth place and all movements (towns or cities of residence) from birth until the date of the interview, as well as their spouse's names, professions, birth dates, birth places and years of marriage for those who married. Children's names, birth years and birth places are also listed. These basic demographic data are presented in a standardised format for each entry. There is a variety of other data as well, including, for example, whether men had served in the army during the war and whether women had participated in the Lotta Svärd organisation (an all-female paramilitary organisation).

The resulting registers contain a vast amount of data on the Karelian migrants, but in book format, they are poorly suited to quantitative analysis. Therefore,

a project was initiated to digitise these data, which ultimately resulted in the generation of the MiKARELIA database. Data entries were scanned at 300 dpi using a Canon c5250i copier and saved in pdf format. ABBYY Fine Reader 12 (ABBYY production LLC 2013) was used to scan pdf documents for optical character recognition (OCR) and the output saved in html format. An open source software program²³ was written to convert Fine Reader produced html files to a simpler xml format containing the data entries. The program reads and extracts the source text to produce a JSON file containing all extracted data. These data can then be used to populate a structured database.²⁴

Obviously, the MiKARELIA database represents only those Karelians who were alive in 1968 to 1970 when the interviews were conducted. However, Loehr and colleagues have estimated that these data include records on approximately 75% of the Karelian migrants who were alive at this time. Therefore, MiKARELIA can be considered to be a population-based database and not just a statistical sample of Karelians.²⁵ The MiKARELIA database is being further improved and replenished by combining it with other datasets (for example, the Karelia database ‘Karjala-tietokanta’, which contains digital demographic information from about 70 parish registers of the ceded Karelia from the end of the 17th century until the start of Second World War).

One key advantage of the MiKARELIA database, for example, as compared to the Statistics Finland 10% sample data from the 1950 population census, is that while in the sample data individual level variables (for example, migration of Karelians) are only available for the year 1950 and from 1970 onwards,²⁶ in MiKARELIA there are individual level data on evacuees during the Second World War (for example, whether they served during the war and whether they returned back to Karelia during the Continuation War). Therefore, the MiKARELIA database offers excellent opportunities to explore with considerable detail the migrations of Karelians during and after the Second World War in addition to a variety of socio-demographic and environmental factors that were associated with their decisions to migrate.

To determine whether Karelian evacuees differed in their migration histories as a function of their background characteristics, the current study involved analysing the already existing MiKARELIA database and combining it with a database on the location of all the cities and towns involved in the evacuations, and their population sizes. Populations of towns located in Finland and Karelia were obtained from the Statistical Yearbooks of Finland 1939.²⁷ In addition, for each place, we obtained coordinates to locate them on the map and calculate the effect of several geographical dimensions on the probability of returning home during the war (1941–1944). To do this, information was gathered from multiple sources on the internet and maps. Most of the coordinates could be found directly from the history books of the *Suomen sukututkimusseura* (the Finnish Genealogy Society), while the rest were searched from Google Maps—a map utility served by Google Incorporation (Google Maps, Finnish Genealogy Society).

Sample selection

Although interviewees provided some information on other members of their family (for example, spouses and children), in our analyses we focused solely on individuals who were interviewed and on which we had the most complete and systematically recorded information. Thus, the statistical unit for this research is the family, rather than each family member separately, given that families were presumed to have moved together. In addition, children (individuals who were born after 1925) were excluded from these analyses. These individuals would have been 15 years old or younger in 1941 when the first opportunity to return to Karelia was possible. The birth location, rather than the location in Karelia at the moment of the evacuation, was used because the location of the evacuees immediately prior to the evacuation was only recorded for a small subset of individuals, whereas birth place was available for more than 90% of the total sample. Finally, only those who were born in Karelia were chosen. These selection criteria left us with a sample of 59,477 Karelian evacuees. Each population size parameter (birth population, population of first destination in Finland and population of return destination in Karelia, which was used in the maps) was log transformed for reasons of statistical inference (that is, the effects of population size are not expected to be merely additive) and to aid fitting the models.

Variables

As a dependent variable, we use a binary variable: whether an individual returned to Karelia or not (0 = no, 1 = yes). In our sample, 52% returned, which is a somewhat lower number than the overall proportion of evacuees who returned, which was reported to be approximately 70%. This may be because we are both missing the oldest Karelians who might have been more likely to return than the younger ones (and had died by the time of the interviews in 1968 to 1970), and also missing those who were children (less than 15 years old) during the war. Our other dependent variable is the total number of moves after 1945 and up until 1970. In our sample, Karelians had on average 1.02 (SE 0.005) moves, which varied between 0 and 19. The majority, however, moved at least once (54%).

As independent variables we use: sex, age, occupation in 1970 (farmer or non-farmer; we are expecting that farmer was the most 'stable' occupation, that is, one can assume that they were already farmers in Karelia), whether he or she had children in 1940, longitude and latitude of birth location, longitude and latitude of first destination in Finland and population size in birth location and first destination in Finland (see Table 8.1 for descriptive statistics).

We used generalised linear logistic and Poisson regression models to analyse these data. In the case of returning to Karelia during the Continuation

Table 8.1: Descriptive statistics for those who returned Karelia and those who did not (%/mean, (SE)).

	Returned Karelia 1941–1944	
	Yes	No
Sex (%)		
Women	53.1	46.9
Men	52.2	47.8
Age (mean)	30.7 (0.06)	28.9 (0.06)
Farmer (%)		
Yes	73.5	26.5
No	42.9	57.1
Have children (%)		
Yes	60.6	39.4
No	48.5	51.5
Birth destination longitude (mean)	29.8 (0.01)	29.7 (0.01)
Birth destination latitude (mean)	61.0 (0.004)	60.9 (0.005)
Destination longitude (mean)	24.8 (0.02)	25.3 (0.02)
Destination latitude (mean)	61.9 (0.01)	61.6 (0.01)
Destination population size (mean)	18280.1 (351.6)	43327.9 (743.5)
Birth population size (mean)	12559.33 (116.9)	17915.5 (212.3)

Note: Demographic variables $n = 49,780$; environmental variables $n = 29,622$.
Source: Authors.

War, we explain our binary dependent variable (returned to Karelia = 1, did not return = 0) using a logistic regression and the coefficients of the predictors are interpreted as odds ratios. An odds ratio above 1 indicates a greater likelihood of the event compared to the reference category, and an odds ratio below 1 indicated a smaller likelihood when all other covariates entered into the model are held constant. To model the number of moves after the Continuation War, we used both a logistic regression (no moves = 0, at least one move = 1) and a Poisson regression which fits these count data (namely, the number of moves) better than a normal distribution. Poisson regression coefficients can be interpreted in a similar manner as linear regression coefficients such that negative coefficients indicate a negative relationship and positive coefficients indicate a positive association with the outcome variable when all other covariates are held constant.

Results

The evacuees in the MiKARELIA database we used here were from areas west and north of Lake Ladoga, which are the regions Finland lost to the Soviet Union. The distribution of evacuees' homes at the time of evacuation are illustrated in Map 8.1. Geographically, the top destination as well as the rest of the distribution are generally widely spread across southern Finland (see Map 8.2). The distribution of returning evacuees to Karelia during the Continuation War is similar to the baseline in 1939, with a few exceptions in southern Karelia. The number of people who returned is, of course, fewer than the number who left from those same places (see Maps 8.1 and 8.3).

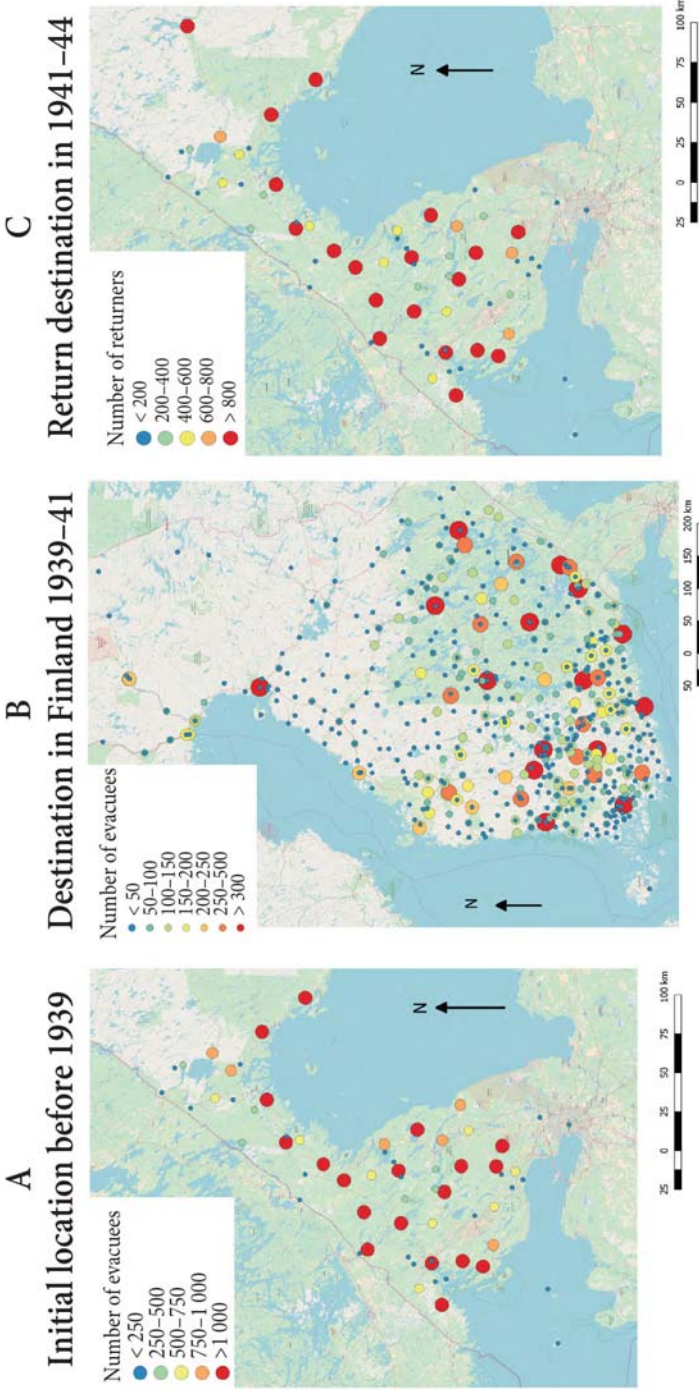
Return to Karelia during the Continuation War

Evacuees were spread across an area spanning 60 to 70 degrees latitude, with most people concentrated in the south, especially in areas below 64 degrees latitude (see Map 8.2). Map 8.2 shows the percentage who returned from each evacuation destination in western Finland. Here, it is evident that the degree to which the return rate depends on location is complex. However, return rates below 50% are more common at lower latitudes and a higher percentage of people returned to northern parts of Karelia. Evacuees were spread quite evenly across an area of Finland spanning 19 to 31 degrees longitude. At more western longitudes, the proportion of those who returned to Karelia is greater.

Overall, Karelians placed in northern and western Finland were more likely to return. Evacuees were spread across towns and cities of varying populations, but those evacuated to the areas in the largest category (population size greater than 20,000), fewer than 60% returned. No other relationships between population size and return rate were obviously evident (Map 8.2).

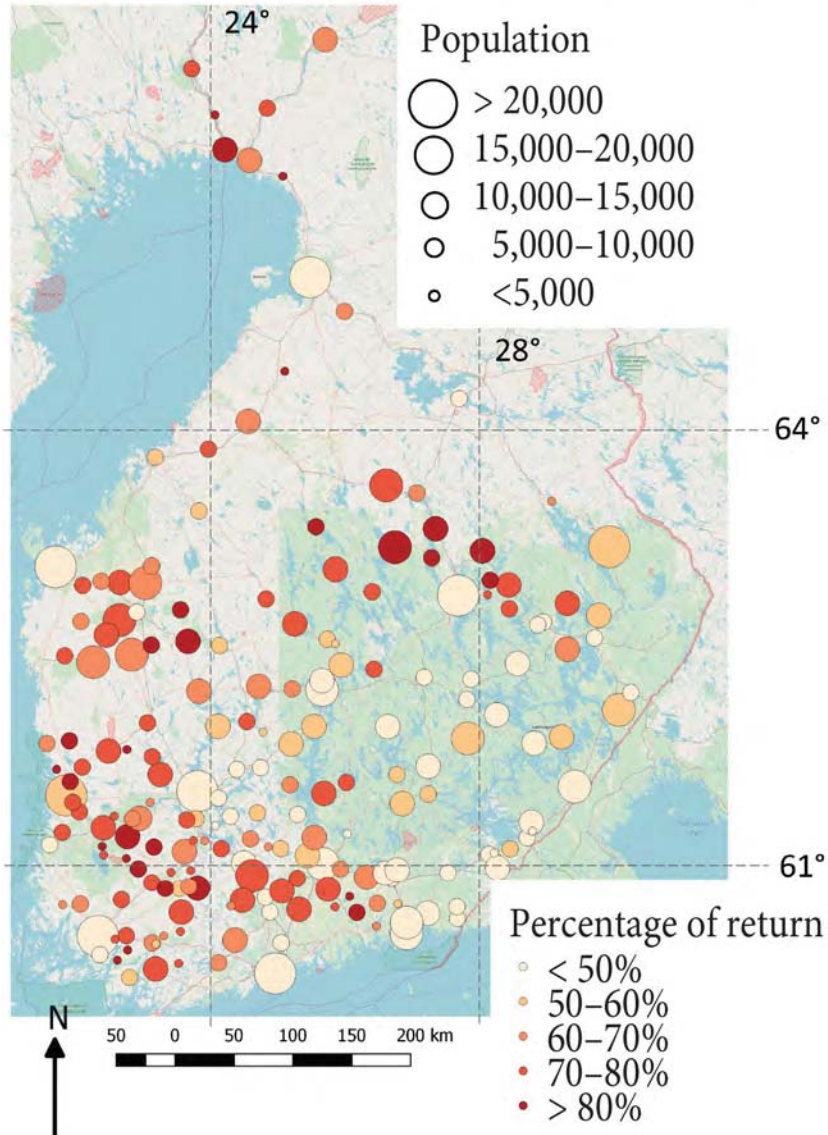
Results shown in Table 8.2 are from a two-stage stepwise logistic regression model in which the dependent variable is whether or not a person returned to Karelia, and the independent variables are added to the model in two stages: first, all socio-demographic variables and, second, all environmental variables. Results from Model 1 show that men were less likely to return than women. In subsequent sensitivity analyses, this difference disappeared, however, once the fact that many men were serving in the army was taken into account (results of sensitivity analyses are not shown in Table 8.2). In addition, results suggest that age was not a significant predictor of returning to Karelia. Being a farmer was, however, and the predicted probabilities (calculated from odds ratios) of returning for farmers was 73% as compared to non-farmers, which was 43%. (Note: In Model 2, which also takes into account environmental factors, the probabilities were 76% and 53% for farmers and non-farmers respectively.) Therefore, the adjusted probabilities did not differ much from the unadjusted distribution (see Table 8.1). Having children was also positively associated with returning to Karelia.

Percentage of returned evacuees



Map 8.1: Dispersal patterns of Karelian evacuees 1939–1944. First (A) the locations where Karelians were evacuated from. Secondly (B) the distribution of displaced evacuees in Western Finland. Finally (C) the locations of the returning Karelians to homeland during Continuation War. Source: Authors.

Percentage of returned evacuees



Map 8.2: The proportion of evacuees who returned to Karelia from towns in western Finland during the Continuation War. Source: Authors.

Table 8.2: Association between socio-demographic and environmental factors with the likelihood of returning to Karelia during Continuation War.

	Model 1			Model 2						
	OR	SE	p	95% CI		OR	SE	p	95% CI	
				lower	upper				lower	upper
Sex										
Women (ref.)										
Men	0.83	0.02	0.000	0.80	0.86	0.84	0.02	0.000	0.80	0.88
Age	1.00	0.001	0.091	1.00	1.00	1.01	0.002	0.000	1.01	1.01
Farmer										
No (ref.)										
Yes	3.64	0.08	0.000	3.49	3.79	2.78	0.08	0.000	2.61	2.95
Have children (%)										
No (ref.)										
Yes	1.38	0.03	0.000	1.32	1.44	1.19	0.04	0.000	1.12	1.27
Birth destination longitude						0.78	0.02	0.000	0.75	0.81
Birth destination latitude						2.31	0.08	0.000	2.15	2.48
Destination in Finland longitude						0.89	0.01	0.000	0.88	0.90
Destination in Finland latitude						1.04	0.01	0.002	1.01	1.06
Destination population size						0.74	0.01	0.000	0.73	0.76
Birth population size						0.98	0.02	0.126	0.95	1.01
n	49,780					29,622				
McFadden's Adj r ²	0.067					0.102				

Note: Results from two-stage stepwise logistic regression. Source: Authors.

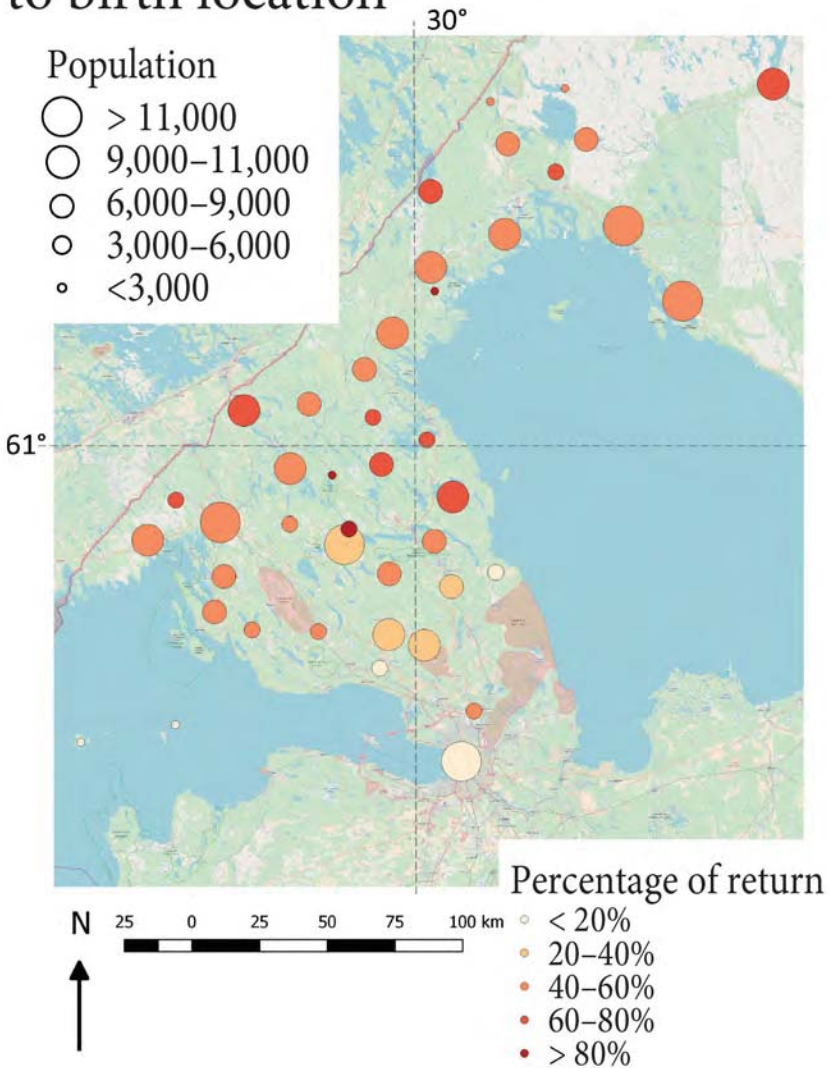
Results of Model 2 (Table 8.2) take into account several environmental variables, in addition to socio-demographic variables. Results from this regression model support the conclusions drawn from the maps shown above: people from more western and northern birthplaces were more likely to return and evacuees who went to more westerly and northerly destinations in Finland were more likely to return. In addition, the population size of the destination town or city in Finland was significantly and negatively associated with the likelihood of returning to Karelia during the Continuation War. In other words, people placed in less populated areas were less likely to remain and more likely to return to Karelia. Taking these environmental factors into account did not alter the effects of socio-demographic factors, although age was significantly and positively associated with returning, meaning that older people were more likely to return.

Map 8.3 indicates that evacuees from the larger populations have relatively fewer people returning, although the association was not statistically significant in the regression model. In addition, places located nearer to the front line, especially in the Karelian Isthmus, had relatively fewer people returning.

Migration after the Continuation War

Our second dependent variable considered the number of moves after the Continuation War. First, we investigated those Karelians who had moved at least once after their first placement (Table 8.3). Here, men were more likely than

Percentage of returned evacuees to birth location



Map 8.3: Proportion of evacuees who returned to their natal locations in Karelia. Source: Authors.

women to move at least once. Also, the younger these individuals were, the more likely they were to move multiple times. Being a farmer was also positively associated with the likelihood of moving at least once. In addition, those who had returned to Karelia during the Continuation War were more likely to move at least once after the war than those who did not return.

Table 8.3: Socio-demographic factors and whether or not individuals returned to Karelia associated with the likelihood of Karelians to move at least once after Continuation War

	OR	SE	p	95% CI	
				lower	upper
Sex					
Women (ref.)					
Men	1.26	0.02	0.000	1.22	1.31
Age	0.99	0.001	0.000	0.98	0.99
Farmer					
No (ref.)					
Yes	1.11	0.02	0.000	1.06	1.16
Returned Karelia					
No (ref.)					
Yes	2.22	0.04	0.000	2.14	2.31
n	49,241				
McFadden's Adj R2	0.034				

Note: Odds ratios from logistic regression model. Source: Authors.

Second, we examined whether the same socio-demographic factors were associated with the frequency of moves among Karelians after the Continuation War (Table 8.4). As was the case with any moves, men, younger Karelians and those who returned to Karelia during the Continuation War were all more likely to move more after the Continuation War than women, older Karelians and those who did not return to Karelia when they had a chance. However, farmers were less likely than non-farmers to move more after the Continuation War. This was the only factor that was differently associated with moves when compared to the previous model (Model 1).

Discussion and Conclusions

Our primary aim in this chapter was to study how the migration histories of Karelian evacuees during and after the Second World War were influenced by a variety of social, environmental and demographic characteristics. Which evacuees were more likely to move back to Karelia when they had the opportunity? Which environmental factors influenced an individual's decision to return or remain? How many times, on average, did the Karelians move after the second evacuation and who moved the most and who settled the fastest?

Table 8.4: Socio-demographic factors and whether or not one returned to Karelia after the Winter War is associated with the frequency of moves among Karelians after the Continuation War.

	coeff.	SE	p	95% CI	
				lower	upper
Sex					
Women (ref.)					
Men	0.10	0.01	0.000	0.09	0.12
Age	-0.02	0.0005	0.000	-0.02	-0.02
Farmer					
No (ref.)					
Yes	-0.16	0.01	0.000	-0.18	-0.14
Returned Karelia					
No (ref.)					
Yes	0.40	0.01	0.000	0.38	0.41
n	49,241				
McFadden's Adj. R2	0.022				

Note: Coefficients from Poisson regression model. Source: Authors.

We used the new MiKARELIA database which has unique individual level information on moves of Karelians during and after the Second World War. Our results are mainly in line with previous studies,²⁸ although very few of these have concentrated on factors related to returning to Karelia during the war. We found that both socio-demographic and environmental factors were associated with returning to Karelia during the Continuation War.

In detail, we found no sex differences in the likelihood of returning once we took into account the fact that many men were serving in the army during the Continuation War. Previous studies²⁹ have shown and this study confirms that farmers were more likely to return than non-farmers. In addition, once environmental factors are taken into account, the models show that older individuals were more likely to return than younger ones. This suggests that those Karelians who were in a more stable phase of life and who were probably more attached to their home districts (for example, had family and land and were older) were more likely to return to Karelia. Environmental factors also made a difference. People placed in more westerly and northerly destinations in Finland were more likely to return,³⁰ while at the same time Karelians who were from more western and northern birthplaces were also more likely to return. In addition, evacuees who were placed in smaller towns were also more likely to return. Importantly, these environmental factors, which had been documented

previously in other studies, were still significant predictors of returning to Karelia even when socio-demographic variables were controlled for. A major advantage of this study is that we can take several characteristics into account at once and draw the conclusion that, for instance, occupation (namely, being a farmer) did not alone explain the variation in returning to Karelia.

While analysing the number of moves after the Continuation War, we discovered that nearly half of the Karelians actually settled permanently in their first location (46%). Those who had moved at least once were also more likely to be farmers. This was probably the result of the resettlement policies and the Land Acquisition Act,³¹ which required farmers to wait to acquire their own land. However, after this initial displacement, the farmers were less likely than others to move, which suggests that they probably settled the fastest. Finally, younger people, men and those who did return to Karelia during the war were more likely to move at least once after the war ended and these evacuees also moved more overall. The positive association between evacuees who returned to Karelia and subsequent movements after the war ended is particularly interesting because it seems to contradict our findings on the characteristics of those who returned and those who moved more after the war. In other words, although farmers were more likely to return to Karelia when an opportunity came and individuals who returned were more likely to move after the war ended, those who moved more after the war were also less likely to be farmers. This indicates that there may be yet-to-be-determined factors influencing the relationship between returning to Karelia and geographic mobility after the war and suggests that these relationships need to be investigated further.

The main strengths of this chapter are that we were able to utilise individual level data on a large number of Karelians to study their migration during and after the Second World War. A key advantage of this kind of database and the methods used in this chapter are that we were able to simultaneously take into account several factors that are associated with the frequency of migration.

The main limitations of this study are data related. For example, we do not have data on the oldest Karelians and currently we only have occupations for people in these data from when they were interviewed in 1970. These issues are related to the fact that the original data were collected in 1968 to 1970. However, a crucial advantage of having digitised these data is that we can in the future continuously update and supplement these data with other source material and merge them with other large quantitative databases available for the Finnish population.

Future studies could investigate more closely the migration profiles of different sub-groups of Karelians. For instance, what happened to those farmers who did not return to Karelia during the Continuation War or to those evacuees who settled in their first location after the Continuation War? By examining more closely the movements of different groups of Karelians, we may also explore how the early settlement of evacuees is linked to the long-term outcomes associated with forced migration.³²

Notes

- ¹ Kuhlman 1991.
- ² Lowe 2012.
- ³ Kliot 2007: 57–78.
- ⁴ See, e.g., Waris et al. 1952; Paukkunen 1989; Kananen 2010.
- ⁵ See, e.g., Raninen-Siiskonen 1999; Armstrong 2004; Fingerroos 2010; Savolainen 2015.
- ⁶ See, e.g., Waris et al. 1952; Hietanen 1982; Laitinen 1995a.
- ⁷ Haukka et al. 2017.
- ⁸ Sarvimäki et al. 2009; Sarvimäki et al. 2018.
- ⁹ Sjöblom 2016.
- ¹⁰ Sarvimäki et al. 2009; Sjöblom 2016.
- ¹¹ Sarvimäki et al. 2009.
- ¹² Haukka et al. 2017.
- ¹³ Hietanen 1982: 67, 105.
- ¹⁴ Waris et al. 1952: 52, 327.
- ¹⁵ Ibid.: 104–111; Hietanen 1982.
- ¹⁶ Hietanen 1982: 114–115.
- ¹⁷ Kulha 1969: 80–82.
- ¹⁸ Waris et al. 1952: 110; Hietanen 1982: 226–230; Laitinen 1995b: 52–138.
- ¹⁹ Waris et al. 1952: 53–56, 328.
- ²⁰ Simonen 1965.
- ²¹ Waris et al. 1952; Sarvimäki et al. 2009.
- ²² Laitinen 1995b: 52–138; Waris et al. 1952: 65–74, 112–127.
- ²³ Salmi & Loehr 2017.
- ²⁴ Loehr et al. 2017. See also Lynch, Lummaa & Loehr 2019a; Lynch et al. 2019b who have used the same data.
- ²⁵ Loehr et al. 2017.
- ²⁶ See more on Statistics Finland data: Sarvimäki et al. 2009; Sjöblom 2016; Haukka et al. 2017; Sarvimäki et al. 2018.
- ²⁷ Statistics Finland 1940.
- ²⁸ See, e.g., Hietanen 1982; Lynch et al. 2019; Waris et al. 1952.
- ²⁹ See, e.g., Hietanen 1982: 114–115; Lynch et al. 2019.
- ³⁰ Waris et al. 1952.
- ³¹ Ibid.: 65–74, 112–127; Laitinen 1995b: 52–138.
- ³² Cf. Sarvimäki et al. 2009; Haukka et al. 2017.

References

- Armstrong, K.** (2004). *Remembering Karelia: a family's story of displacement during and after the Finnish wars*. New York, NY: Berghahn Books.
- Fingerroos, O.** (2010). *Karjala utopiana*. Jyväskylä: Jyväskylän yliopisto.

- Haukka, J., Suvisaari, J., Sarvimäki, M., & Martikainen, P.** (2017). The impact of forced migration on mortality: a cohort study of 242,075 Finns from 1939–2010. *Epidemiology*, 28(4), 587–593.
- Hietanen, S.** (1982). *Siirtoväen pika-asutuslaki 1940: Asutuspoliittinen tausta ja sisältö sekä toimeenpano*. Historiallisia tutkimuksia 117, Helsinki: Suomen historiallinen seura.
- Kananen, H.** (2010). *Kontrolloitu sopeutuminen: Ortodoksinen siirtoväki sotien jälkeisessä Ylä-Savossa (1946–1959)*. Jyväskylä: Jyväskylän yliopisto.
- Kliot, N.** (2007). Resettlement of refugees in Finland and Cyprus: a comparative analysis and possible lessons for Israel, In A. M. Kacowicz & P. Lutomski (Eds.), *Population resettlement in international conflicts: a comparative study* (pp. 57–78). Plymouth: Lexington Books.
- Kuhlman, T.** (1991). The economic integration of refugees in developing countries: a research model. *Journal of Refugee Studies*, 4(1), 1–20.
- Kulha, K.** (1969). *Karjalaiset Kanta-Suomeen: Karjalaisen siirtoväen asuttamisesta käyty julkinen keskustelu vuosina 1944–1948*. Jyväskylä: Gummerus.
- Laitinen, E.** (Ed.) (1995a). *Rintamalta raiviolle: Sodan jälkeinen asutustoiminta 50 vuotta*. Jyväskylä: Atena Kustannus Oy.
- Laitinen, E.** (1995b). Vuoden 1945 maanhankintalain synty, sisältö ja toteutus. In E. Laitinen (Ed.), *Rintamalta raiviolle: Sodan jälkeinen asutustoiminta 50 vuotta* (pp. 52–138). Jyväskylä: Atena Kustannus Oy.
- Loehr, J., Lynch, R., Mappes, J., Salmi, T., Pettay, J., & Lummaa, V.** (2017). Newly digitized database reveals the lives and families of forced migrants from Finnish Karelia. *Finnish Yearbook of Population Research*, 52, 56–69.
- Lowe, K.** (2012). *Savage continent: Europe in the aftermath of World War II*. London: Viking.
- Lynch, R., Lummaa, V., & Loehr, J.** (2019). Self sacrifice and kin psychology in war: threats to family predict decisions to volunteer for a women's paramilitary organization. *Evolution and Human Behavior*, 40(6), 543–550 DOI: <https://doi.org/10.1016/j.evolhumbehav.2019.06.001>
- Lynch, R., Lummaa, V., Panchanathan, K., Middleton, K. M., Rotkirch, A., Danielsbacka, M., O'Brien, D., & Loehr, J.** (2019). Integration involves a trade-off between fertility and status for World War II evacuees. *Nature Human Behaviour*, 3, 337–345. DOI: <https://doi.org/10.1038/s41562-019-0542-5>
- Paukkunen, L.** (1989). *Siirtokarjalaiset nyky-Suomessa: Jyväskylän yliopiston yhteiskuntapolitiikan laitoksen tutkimuksia*. Jyväskylä: Jyväskylän yliopisto.
- Raninen-Siiskonen, T.** (1999). *Vieraana omalla maalla: Tutkimus karjalaisen siirtoväen muistelukerronnasta*. PhD thesis, University of Joensuu.
- Salmi, T., & Loehr, J.** (2017). Kaira-core [computer software]. Lammi Biological Station. Retrieved from <https://github.com/Learning-from-our-past/Kaira-core>
- Sarvimäki, M., Uusitalo, R., & Jäntti, M.** (2009). *Long-term effects of forced migration*. IZA Discussion Paper No. 4003. Retrieved from <https://ssrn.com/abstract=1351167>

- Sarvimäki, M., Uusitalo, R., & Jäntti, M.** (2018). *Habit formation and the misallocation of labor: evidence from forced migrations*. Department of Economics, Aalto University, Working Paper. Retrieved from <http://hse-econ.fi/sarvimaki/>
- Savolainen, U.** (2015). *Muisteltu ja kirjoitettu evakkomatka: Tutkimus evakkolapsuuden muistelukerronnan poetiikasta*. Joensuu: Suomen Kansantietouden Tutkijain Seura.
- Sjöblom, E.** (2016). *Karjalan evakoiden elämänurat: Sopeutuminen elämään pakkomuuton jälkeen*. Master's thesis of Sociology. University of Helsinki.
- Simonen, S.** (1965). *Paluu Karjalaan: Palautetun alueen historiaa 1941–1944*. Helsinki: Otava.
- Statistics Finland.** (1940). *Statistical yearbook of Finland 1939*. Helsinki: Tilastollinen päätoimisto.
- Waris, H., Jyrkilä, V., Raitasuo, K., & Siipi, J.** (1952). *Siirtoväen sopeutuminen. Tutkimus karjalaisen siirtoväen sosiaalisesta sopeutumisesta*. Helsinki: Otava.

Digital Collections and History of Women's Suffrage Movements

Heidi Kurvinen

Introduction

During the past decade, the amount of digitised material has exploded,¹ but they are not available to all researchers in an equal manner. The entrance to the field of digital humanities requires cultural and technological capital which excludes or marginalises researchers who do not have the skills to conduct digital analyses by themselves or do not have access to the organisational support. According to Matthew K. Gold, it is research-intensive universities containing sufficient financial and human resources that have been able to embrace the digital turn.² Again, this ability to focus on digital research and hire personnel to carry out the analysis has formed a 'circle of good' stabilising their status within the field. Other universities, not to mention individual researchers, have been less fortunate, but simultaneously digital analysis tools increase the expectations that we as scholars are expected to accomplish.³

Gender is one of the factors that seems to affect the ability of researchers to take part in the digital turn. Farida Umrani and Rehana Ghadially discuss the

aspect of empowerment that is connected to the access of computers for women by using the division between 'information rich' and 'information poor' people.⁴ Even though their approach is connected to third-world countries and the use of computers in general, I find the division useful also in the context of digital humanities. Similarly, researchers within humanities and social sciences are nowadays divided into those who have skills and access to digital research methods and those who don't. Based on stereotypical gender role expectations (such as viewing technology as masculine coded) with which most generations of current historians have been raised, men have often better opportunities to explore the field even if the starting point is the same with their female colleagues. The reason for this, as Miriam Posner has pointed out, is that middle-class white men are more likely to have been encouraged to explore with computers at a young age than women or other marginalised groups.⁵ At the same time, most present-day researchers are regardless of gender already to some extent participating in the digital turn as the mixing of traditional and new (digital) research practices have become a self-evident part of our present-day work as scholars, in the form of digital voice recorders, digital cameras and the use and analyses of digital texts and images.⁶

In this chapter, I will discuss what is needed when a historical scholar with limited digital skills wants to take a step towards learning how to conduct digital analyses, towards becoming a digital historian. As a feminist historian, I will combine this approach with a discussion of the relation of feminist research and digital humanities. In line with practice in feminist research, I will be using a self-reflexive approach and asking how the increase in the understanding of digital methods influences our research questions in feminist history. Do digital humanities tools transform our work as feminist historians? How can digital analyses develop the field of gender history in general and the history of feminism in particular? Can a scholar who has limited technological skills engage with an informed and critical discussion with digitised materials?

Even though the main points of my chapter apply to all historical research, a focus on gender analysis is worth making as gender seems to have remained a rather limited category of analysis among digital historians. And although not all gender historians identify their work as feminist, there is a strong connection between the two,⁷ which makes the discussion of the relation between feminism and digital analysis a valid starting point for this chapter. The intention is not, however, to claim that there is a clear difference between feminist history and historical research in general, but to participate in the discussion of the meanings of feminist approaches to digital humanities and ponder why in particular feminist historians should be part of these discussions.

Feminism and Digital Humanities

At first, the combination of feminist research and digital analysis may seem as strange bedfellows, but for the past decade feminist digital humanities research

has been conducted in various fields in the Anglo-American world, in particular, and scholars have engaged in critical discussion of the relation between feminism and digital humanities. Some scholars have problematised this relation by deconstructing the gender-neutrality of digital analysis tools in order to find ways to overcome the divide between male producers and female users of computational tools.⁸ Others have asked how the use of a large amount of data fits with feminist research that relies on gender-sensitive reading.⁹ Scholars have also found similarities between feminist research and digital humanities approaches, such as collaborative research. According to Janine Solberg, feminist research and digital humanities may even form a fruitful pair since the idea of an ethical (feminist) researcher encourages the scholar to be open for multiple viewpoints and to position oneself as a researcher and conduct the research in a transparent manner (features that are also valued in digital humanities).¹⁰

Discussion on a feminist approach to digital humanities has mainly focused on pondering how gender, race and other marginalising factors can be taken into account when compiling datasets and digital archives. In addition, the responsibility of feminist scholars to unsettle digital humanities' 'retro-humanist' practices that maintain a canonical understanding of what is relevant to digitise have been pointed out.¹¹ However, less has been written about the actual methodological practice of conducting a feminist digital humanities project. Nevertheless, there are some exceptions. In her insightful article on the US suffragette Frances Maule, Solberg, for example, points out how the new technology made it possible for her to find information about Maule, whose life was relatively unknown when she discovered her. At the same time, new information that she was able to find thanks to digitised material changed the interpretations of Maule's texts used by Solberg in her work.¹² Thus, digitised data can help us to find sporadic information of our research subject and combine these pieces of information more easily than previously. Mass digitisation may also widen our opportunities to find traces of people who have been marginalised in the past or purely forgotten, which is consistent with the core ideas of feminist research.

In spite of the existing literature on feminism and digital humanities, feminist digital history seems to be an under-discussed area of study. For instance, scholars of feminist historiography of rhetoric, Jessica Enoch and Jean Bessette, argue that feminists have used digitally born materials to study women's lives, but historians have rarely pondered how digital methods could widen their scope of study.¹³ A few exceptions have appeared in feminist literary history, in particular, but the field is still narrow. This seems problematic because the digital turn has already started a revolution in history which will potentially profoundly change our scholarship and require us to learn new tools, as Alexis Lothian and Amanda Phillips have formulated.¹⁴ Also due to this, gender-sensitive historians should start to pay attention to the challenges and opportunities that the digital turn will cause in our field.

The argument follows Solberg as well as Enoch and Bessette, who have demanded more discipline-specific discussions on the role of digital analysis tools and digital research materials.¹⁵ According to my understanding, a focus

on gender history is particularly important because, as an already marginalised sub-field of a traditionally masculine discipline, it may not otherwise be able to answer the demands of financiers that have started to highlight the importance of digital methods. In addition, gender sensitivity is needed to guarantee that digitisation of archival sources and other material such as print media or books do not focus solely on canonical pieces or notable people of history. Even though gender has been written into history increasingly since the 1960s, digitised collections often maintain gender bias when men's history tends to be viewed as more important. Furthermore, a move towards more digitally aware research may require that gender historians, alongside gender scholars in other fields, start to discuss how women's and other minority groups' abilities to conduct digital humanities research could be supported. However, the intention of this chapter is not to strengthen the essentialist notion of women as less capable than men of conducting digital analysis, but by using my own research field as a starting point to instead problematise what is needed to support scholars such as myself who do not have the basic digital skills, making it difficult to start on their own.

Taking the First Steps in the World of Digital Humanities

In Finland, the first computer started to operate in 1958 and a relatively rapid computerisation has taken place in the country since the 1960s. This has also had its effect on research. Historian Viljo Rasila was already writing about computer-assisted research in 1967 and used these kinds of methods in his work.¹⁶ However, when I began my studies at university as a fresh undergraduate student in 1999, I did not own a computer, and neither did many other students at this time. Computer-assisted methods did not belong to the curriculum and in spite of the accelerating computerisation of the Western world, for many they remained primarily a tool for writing and for using publishing or photo programs. This applied also to my relationship with computers, which explains why I never learned to understand properly how computers work as operational systems. For me, computers remained tools that I used to write and I knew only as much of them as was needed to complete that task.

As a scholar who began her postgraduate studies in the mid-2000s, I was even able to conduct my PhD studies without ever hearing the words 'digital humanities'. I first became familiar with the field as late as in 2015 when editing an article on that matter for a Finnish peer-reviewed journal as part of my duties as a sub-editor. I became instantly intrigued; but for a person with limited IT skills; it felt overwhelming to even try to figure out how I could use the approach in research. As for many, I assume, the first push towards this took place after a year while I was writing a research proposal for a major financing body and tried to figure out how I could elevate the state of art so that my project would be successful in receiving funding. I started to read digital humanities literature and tried to understand what all this could mean for my project.

By reading the texts, I started to realise that digital humanities projects were often collaborative initiatives (that is, not everyone needs to know how to code). However, at that time, I was working as a visiting scholar at a Swedish university with a Finnish research grant, which meant I found it difficult to start looking for collaborative partners. How could I even start to look for them, I was asking myself. In short, I lacked the institutional support as well as technical skills that would have helped me to explore the field on my own.

For a year, every now and then, I read articles on digital humanities, some of which made more sense to me than others. I was, for instance, exhilarated to find out that distant reading could be combined with close reading, the latter of which is the method I am most familiar with. However, the basic question remained the same: How could I start to understand the process of the analysis? Simultaneously, I changed universities and my new colleagues helped me to find pages that offered guidance for different analysis tools, such as *The Programming Historian*, but it still felt difficult to start learning on my own. Luckily, in the spring of 2018, I was selected for the practical course on digital humanities organised by the project 'From Roadmap to a Road Show' led by Mats Fridlund. It turned out to be the first step towards understanding what digital humanities could actually mean for my historical research. As my project, I selected a small-scale case study of the use of the word '*naisasiaiike*' (women's suffrage movement) in the biggest national newspaper *Päivälehti* (1889–1904) / *Helsingin Sanomat* (1904–) at the turn of the 19th and 20th centuries.

Learning by One's Mistakes

Janine Solberg has argued that digital environments can be used as safe spaces to test our research ideas.¹⁷ In her work, Solberg did not rely on big data, but used digitised material to trace pieces of information of her research subject. However, her argument also seems suitable for a scholar who combines a big data approach with close reading of a relatively small pool of data, as is the case in this chapter.

Previously, I had used the National Library of Finland's digital newspaper archive from time to time to look for information. However, I had only used the search option, without trying to familiarise myself with the platform. Due to this, I had the habit of writing down the texts that interested me and it was only when preparing the data for the course that I realised that an OCR view of the text would make it easier to gather the data. However, from the literature, I had learned that not all letters would necessarily appear the same in the OCR text as they were in the original.¹⁸ This was painfully clear concerning the material from *Päivälehti*, in which the articles were published by using *fraktur*, an old German font type. For instance, the machine was unable to recognise the letter 'w' as such, but it was often written with 'm' or with a combination of two letters such as 'n' and 'i'. Also similar kinds of errors took place with other letters such as 's', which had different kinds of typographical variations in the

original text. Furthermore, wider spacing of words which was used to highlight certain words in the newspaper texts, such as surnames, caused trouble for the machine. More importantly, the machine was unable to recognise the newspaper columns which ran in an uneven manner at that time. This meant that the OCR view did not include texts in the order they were mentioned, and I had to copy the paragraphs manually. This included removing paragraphs that were not part of the article I was interested in and ascertaining that all the paragraphs of my article had been copied to the file. Due to these problems, the OCR view made it only slightly easier to prepare the material than my original manner of writing everything manually. While the newest articles might have had only a couple of errors, older texts were often impossible to understand based on the OCR. This meant that I had already read through some of the texts while correcting the OCR, making me familiar with the material. The same kinds of problems have also been noticed by other scholars who have problematised the idea of digital analysis as a rapid way to conduct research with newspaper material.¹⁹

My first experience with any kind of a data analysis software took place in May 2018, when I participated in the earlier-mentioned digital history course. Based on my project abstract, in which I had suggested that I would use a statistical natural language processing tool called MALLET for carrying out topic modelling analysis of the data, I had been assigned to a MALLET group led by the digital historian Mila Oiva. In addition, Juho Savela was providing technical support. My first challenge on the course was to learn to understand what can be made when using the command prompt of my computer. After that, I learned some basic commands for MALLET which helped me to start playing around with the material. Thus, the course gave me a basic understanding of how the command prompt functions worked and what I as a researcher could do with the material by using MALLET. However, the process of writing this chapter has been a test in which I have used MALLET and my own computer as a safe space for learning more by using the 'learn by your mistakes' method. Gradually, this has deepened my understanding of the process, even though there are still many things I do not understand.

One of the major revelations during the process has been that combining digitised material with technologically assisted analysis needs suitable research questions. According to Solberg, digital tools change our ways of discovering, accessing and making sense of the past. To be more specific, digital environments can reorientate us 'both physically and conceptually' if we choose to be active technology users instead of remaining as passive users of them.²⁰ Similarly, Jacqueline Wernimont has defined the division of male creators and female users of digital tools as one of the critical questions that feminist digital humanities needs to address.²¹ Furthermore, other feminist scholars have engaged in critical discussion of what is enough to make the field more diverse and whether the ability to code is a necessity for all digital humanities scholars.²² In my case, the move towards a more active user of digital tools meant

that my original research questions changed during the process. I realised that MALLET would not be the best option to trace the transnational influences of feminist ideas as I had originally thought. Instead, it offered a window towards the variety of topics connected to which the word *naisasialiike* was used. In the following, I will outline the process of carrying out the analysis as well as ponder whether digital analysis of a relatively small pool of newspaper articles can bring new information concerning the early feminist movement in Finland.

The Importance of Search Words

In the project plan, I outlined the research period to cover the years between 1889 and 1929. I used *Päivälehti*'s first publishing year as a starting point for the search period because the first women's organisation *Finsk Kvinnoförening* —*Suomen Naisyhdistys* had been established five years earlier in 1884.²³ The period ends in 1929, when the New Marriage Law was approved in Finland, forming one kind of an end point for the early feminist movement.²⁴ When outlining the period like this, I assumed that it would consist of a reasonable amount of data that I could use in my analysis. Surprisingly, the number of texts using the word *naisasialiike* proved to be relatively low. The search brought 51 results, one of which was a list of literature for Christmas presents. Because the list contained only one item that was relevant for the theme, I did not take it into account. Other hits were mentions as part of larger articles dealing with various women's organisations and their gatherings or books that dealt with women's issues. Additionally, the pool of data consisted of few notifications for meetings organised by the Women's Feminist Union (*Naisasialiitto Unioni*), among others.

The search also brought to light other astonishing revelations. At first, the word *naisasialiike* seems not to have belonged to the newspaper's vocabulary at all, since the first hit was from the year 1896 (12 years after the first women's organisation had been established). Before 1900, the word *naisasialiike* had been used only five times and continued to be used quite rarely until the 1920s: between 1900 and 1920, it appeared 12 times. Thus, it seems that the word made its breakthrough in the 1920s, even though it was still used only occasionally. This is slightly surprising because the 1920s was a relatively quiet period in the Finnish women's movement compared to earlier decades.

One explanation for the concentration of the use of *naisasialiike* may be that, during the 1920s, it was used as a retrospective term to look back on the history of the women's movement. However, throughout the studied period, it appears also as an umbrella term that was used to refer to women's emancipatory demands of its own time. In other words, *naisasialiike* seems to have become a label that was used both by the women's movement activists and their opponents, and it was accepted by the newspaper's editorial office. Thus, the question remains: Why did the most active years of early feminism as

a movement in Finland not cause wider coverage in the pages of *Päivälehti / Helsingin Sanomat*? The question is particularly interesting because the word *naisialiike* is the one that is commonly used in the research to refer to the early feminist movement in Finland. Due to this, it could have been assumed that the word had appeared in public discussion at the turn of the century.

When looking for answers to the above-mentioned question, it is worth taking into account that the results would have been different if I had used different search words. For instance, Hieke Huistra and Bram Mellink have reminded us that a digital humanities scholar needs to choose the right search words to receive reliable results.²⁵ However, the problem is that a topic can be described with a variety of words that appear at different times, the meaning of which may change in different contexts and throughout the studied period. In *Päivälehti / Helsingin Sanomat*, the words ‘*feminismi*’ (feminism) and ‘*feministi*’ (a feminist) received 11 hits respectively between 1889 and 1929, but the usage of them took place mostly before the 1920s. The word ‘*naisliike*’ (women’s movement), which refers to all kinds of women’s organising (both feminist and non-feminist), received 255 hits, which suggests that issues relating to women’s status in society were not yet directly connected to feminist organising at the turn of the 19th and 20th centuries. The most comprehensive word was ‘*naisasia*’, which received 897 hits. It was used for the first time in 1890, after which it appeared continuously throughout the research period, suggesting that women’s issue as a theme was part of the public discussion of its time, but it was not connected to a specific movement per se. An analysis of the usage of the word *naisasia* would, therefore, give us a more comprehensive understanding of the early feminism in Finland, but the cleaning of the material would also require a considerable amount of work, which was not possible in the scope of this chapter. Furthermore, such big data would have made it difficult to use this process as an opportunity to reflect on the relation between feminist research and digital humanities. Comparison of the usage of various words dealing with women’s issues nevertheless reveals the development of terminologies which has been shown to be one of the strongest sides of big data analysis. However, as Alex Mold and Virginia Berridge remind us, these kinds of results also need to be contextualised and triangulated with other sources/traditional research methods in order to receive a more accurate understanding of the results provided by the digital analysis.²⁶

As pointed out earlier, a close reading of research material is one of the corner stones of feminist research, and digitised computer reading of big data seems to be in contradiction to this. One solution to overcome this conflict is to combine computer-assisted analysis with close reading of the material or parts of it, as Johan Jarlbrink, Pelle Snickars and Christian Colliander have suggested, among others.²⁷ Based on my small-scale project, a combination of distant and close reading is not only necessary to validate the results, but the combination also gives new perspectives on the material—one example of which is the connection between national and transnational discussion in Finnish feminism.

Previous scholars who have worked with archival material or used media texts in a more traditional manner have extensively shown how the so-called first-wave feminism was committed to national issues in Finland even though the early feminists, at the same time, had wide transnational networks. For instance, Alexandra Gripenberg saw women's emancipation as necessary for human progress and therefore it had to be strived for universally. Simultaneously, her work towards women's emancipation was tied with nationalism.²⁸ Both sides of early feminism also appear in my data, which I recognised while cleaning the material for the analysis. However, the digital analysis also revealed different nuances in the dynamics between national and transnational aspects of early Finnish feminism, as will be shown in the last section of this chapter.

How the Analysis Was Made

The project was started by preparing the dataset of 50 texts for the analysis, after which I conducted topic modelling with 10, 15 and 20 topics. At first, the topics produced by MALLETT seemed like a foreign language to me and the fact that every round of analysis could bring different kinds of word lists was puzzling. Even though I was mechanically able to make the right commands, the ability to start the analysis required a new way of interpreting the lists produced by the computer. This I could not have done without the guidance of Mila Oiva, who patiently used her own research as an example to walk me through the process of shifting my way of thinking. Learning a new way of interpreting the word lists was not the only challenge: I also had some problems with stop-words. Some of them kept popping up in the topics even though I had added them to the list. However, the final round of topic modelling offered satisfactory clean topics even if they still contained some of the listed stop-words, such as the foreign words '*del*', '*und*' and '*des*', as well as abbreviations such as '*klo*'. Because the pool of data was small, I chose to make the analysis based on 10 topics, which brought the clearest image of the data (see Table 9.1).

Three out of 10 topics pointed out to transnational exchange of ideas with words that referred to foreign countries in general or by name and to nationalities or countries in plural. However, there were differences between these three topics. While the first one clearly referred to international connections in media in forms of news reports, the second one attached internationalism to the past of the women's movement and the third topic connected international connections and women's movement congresses, pointing to the transnational nature of the women's movement. Other topics were clearly national by nature, but nationalism became a marker for only one of them which included the word '*isänmaan*' (nation's), for instance.

Otherwise, the topics emphasised meetings of various women's organisations and particular individuals such as Maikki Friberg. Four of the topics include words referring to men. Two of them seem to point to the negotiation

Table 9.1: Topic modelling with 10 topics.

Topic	Keywords	Themes
0	kongressin ihmiset del ulkomailla suomalaist teki sisältää ranskan lehdet pitäen prete otettiin tavoin kansainliiton merkitsee helsingissä tahi monessa aihetta italialaista	kansainvälisyys, kongressit, uutisointi <i>internationality, congresses, news coverage</i>
1	suomen osaston naisliiton lucina tohtori puhui helsingin naisten ohjelma maikki friberg liiton olga lausui opettaja esitti hagmanin kuulla alkoi unionin	naisasiajärjestöt, kokoukset, raportointi <i>women's organisations, meetings, reporting</i>
2	puhuja tehdä elämän naisen lapset suurta mies elämä piti yleinen miehet tehtävä määrässä äiti olosuhteet elää nuori sai alustaja tehty	kokoukset, nainen, roolit, mies, lapsi <i>meetings, a woman, roles, a man, a child</i>
3	naisten naiset saa saada työtä naisia miesten kodin osaa yhteiskunnan osa naisille pois lopuksi maan saanut ulkopuolella muutamia olemassa nähden	nainen, roolit, julkinen elämä <i>a woman, roles, public life</i>
4	owat nainen naisen naisasialiikkeen maissa naisasialiike oliivat syntynyt toiminnan joukko vuotta naisyhdistyksen asema eiwät työn maassa osasto omasta toimintaa kehityksen	kansainvälisyys, ylijärjaisuus, historia <i>internationality, transnational, history</i>
5	naisten suomen kotitalouden liiton hyväksi klo alalla esitelmä seurasi suomessa nykyään kokous saksan liitto ohjelmassa kaikissa saapunut kansallisliiton suomi esitelmän	kokoukset, kansainvälisyys, naisasiajärjestöt <i>meetings, internationality, women's organisations</i>
6	mies professori miehen nim nainen naista naisen voinut esittää perheessä mielestä arvoa ensinkään prof suhteessa tunnettu pitää olevien määrin käy	mies, arvio, naisen rooli <i>a man, review, a woman's role</i>
7	ibsenin tuli ibsen lapsia väkijuoimien tyttöjen naimisiin perintönä isä ammatin runouden paloviinan lapsen jokaisen ominaisuudet vanha valtiopäivillä saivat selville vieläpä	Ibsen, raittius, mies, naisen rooli, äitiys <i>Ibsen, temperance, a man, a woman's role, motherhood</i>
8	suomen kansan lasten laki maamme oikeus eduskunnassa rouva äänioikeus pitäisi asioissa itselleen lain yleisesti miehen kansamme toimintaan isänmaan j.n.e tietä	naisten asema, äitiys, kansallisuusaaite <i>women's status, motherhood, nationalism</i>
9	warten dagmar von die hywin sai wiime saawat warsin rahaston walittiin prior naisasia hyväksi woi des anne tiedekunnassa und erityinen	naisasia, keskeiset henkilöt <i>women's issue, main persons</i>

Source: Author.

between women's and men's roles, while others bring men's point of view to the women's question to the centre. This can be explained with the variety of writers and their relation to the women's movement. Texts that appeared in the regular column of the Women's Feminist Union were most probably written by women's movement activists themselves, as were some other texts published in the paper. However, other texts presented opinions of prominent men. Motherhood is part of three topics, whereas temperance appears to be part of only one topic. The emphasis on women's maternal roles seems accurate because, during the early 20th century, bourgeois women argued on behalf of a social motherhood locating motherhood as women's most important task in the society. According to Sulkunen and others, this bipartisan citizenship was supported by a majority of Finnish women by the 1920s.²⁹ On the contrary, the marginal role of temperance within the topics is slightly surprising.

What do the above-mentioned outlines tell us about the data? First, as an unexperienced user of MALLETT, as pointed out earlier, I began the process by staring at the list of keywords offered by the software without really knowing how I could interpret them. Due to this, I used the scattered ideas I had gained of the texts while carrying out the above-presented classification, even though I had not read all the texts with a similar intensity. Thus, a combination of digital analysis and a close reading of the material helped me to pinpoint the topics that might have otherwise remained unnoticed.³⁰ Second, the strength of the digital analysis is in its manner of presenting the data in a different form, which highlights certain patterns. In this case, it is evident that the discussion on the women's movement was conducted as part of the Finnish public sphere and nationally topical issues. Simultaneously, foreign countries served as a standard reference point and the women's movement appeared as a transnational phenomenon even though it was connected to national discussions. Third, the public discussion offered room also for men to define their stance towards the women's issue. Fourth, rather surprisingly, certain milestones in the development of women's status were not connected to the women's movement in the public debate. For instance, themes such as women's suffrage (1906) and the New Marriage Law (1929) did not raise discussion in which the word *naisasiike* had been used.

Thus, it is evident that digitised material has the potential to show us surprising results, features that we don't expect to find from the material, as Mold and Berridge have pointed out.³¹ However, to be able to understand these results more profoundly, they need to be contextualised both thematically and journalistically. That is to say that computer-assisted analysis also needs a human to contextualise the results (an example of which are media texts that should not be seen as a number of separate articles, but instead as part of the publication context of their time).³² For instance, the length of articles had great variation at the turn of the 19th and 20th centuries. *Päivälehti / Helsingin Sanomat* published short news and reports as well as extensively long congress reports, which were often several pages in length. Potentially, this affects the results as I assume has been the case with the discussion on temperance. Based on the

topic modelling, temperance had a relatively minor role in the public discussion, but based on close reading of the data it was a recurring theme within the material. However, it was not discussed as part of the longer articles, but only mentioned briefly in other texts. Elsewhere, I have also argued the need to bear editorial practices in mind while historians use media texts to make interpretations of past phenomena. This is particularly important while using digitised materials that easily shadow the journalistic processes behind the texts by taking them out of the context.³³ In my opinion, contextualising may also form the bridge that brings digital humanities and feminist history closer together, moving them from being strange bedfellows to being a functional pair.

Conclusion

In this self-reflexive chapter, I have discussed my own road to digital humanities, a journey which has actually only just begun. I believe that my reflections correlate with those of many of my fellow historians and other humanists who have started their scholarly work before the increasing digitisation of the society and are now trying to figure out how the digital methods can be used. What did I learn when conducting my small-scale case study?

Based on my experience, I agree with Solberg, who has argued that digital environments create 'new ways of interacting with' the material.³⁴ I would like to add that, at least for scholars with limited digital skills, they offer an opportunity to conduct the research in a more self-aware manner, when every step of the process needs more thought than a traditional research day working with paper archives, for instance. For a feminist scholar, digital humanities may also serve as a channel for emancipation if the scholar chooses to actively participate in the process of analysis instead of relying on the results produced by IT support. However, to be able to do this, we need the support from our universities to focus on this kind of a large-scale project that also requires time for learning new skills.

My experience clearly demonstrates that conducting a basic digital analysis is possible even for a beginner if she receives sufficient support to carry it out. Additionally, the practice is the best way to increase one's understanding of digital analysis. When the understanding increases, the research questions become more accurate at the same time. Within the limits of this small-scale project, the results were not mind-blowing, but they merely strengthened the results of other scholars focusing on the intertwined relation between national and transnational in the history of early feminism. However, the data also reveals new and previously unresearched questions, such as the use and development of vocabulary relating to women's issues in Finland. Furthermore, results of big data analysis expose new ways of perceiving the material which may revolutionise gender history by revealing gender in places that previous research has been unable to grasp. By also challenging feminist scholars to take a step back

and examine the material from a distance, digital humanities has the potential to change our understanding of gendered patterns in the past.

These questions require more sophisticated digital analysis than has been possible in this chapter, but it is an inspiring direction towards which I hope to be able to move, alongside other feminist historians in the future. One way to do this is to develop grassroots digitisation projects in which gender, race and other marginalising factors could be taken into account when selecting the objects of digitisation. These kinds of projects have the possibility of developing the field by producing more localised and situated data collections that challenge the history we are writing and offer a broader participation in digital history work also for those with basic technological skills.³⁵

Notes

- ¹ Huistra & Mellink 2016: 221.
- ² Gold 2012. See also Earhart & Taylor 2016.
- ³ Solberg 2012: 67.
- ⁴ Umrani & Ghadially 2003: 359–360.
- ⁵ Posner 2012.
- ⁶ Solberg 2012: 69.
- ⁷ See, e.g., Kurvinen & Timosaari 2016.
- ⁸ Wernimont 2013.
- ⁹ See, e.g., Rhody 2016.
- ¹⁰ Solberg 2012, 66. See also, e.g., McKee & Porter 2010: 155–156.
- ¹¹ See, e.g., Risam 2015.
- ¹² Solberg 2012: 59–60, 70.
- ¹³ Enoch & Bessette 2013.
- ¹⁴ Lothian & Phillips 2013.
- ¹⁵ Solberg 2012; Enoch & Bessette 2013.
- ¹⁶ Rasila 1967: 140–146; Paju 2008: 85–87; Paju 2016: 5.
- ¹⁷ Solberg 2012: 67.
- ¹⁸ See, e.g., Jarlbrink, Snickars & Colliander 2016: 29–30.
- ¹⁹ See, e.g., Mold & Berridge 2018: 7.
- ²⁰ Solberg 2012: 54–55.
- ²¹ Wernimont 2013.
- ²² See, e.g., Bailey 2011; Posner 2012.
- ²³ See, e.g., Jallinoja 1983.
- ²⁴ Saarimäki 2018: 75.
- ²⁵ Huistra & Mellink 2016: 222–223.
- ²⁶ Mold & Berridge 2018: 7–11. See also Bergenmar & Leppänen 2017: 235.
- ²⁷ Jarlbrink, Snickars & Colliander 2016.
- ²⁸ Kinnunen 2016: 653.
- ²⁹ Sulkunen 1989; see also Saarimäki 2018: 74.

- ³⁰ See also Jarlbrink, Snickars & Colliander 2016.
³¹ Mold & Berridge 2018.
³² Kurvinen 2018; Huistra & Mellink 2016: 221.
³³ Kurvinen 2018.
³⁴ Solberg 2012: 68.
³⁵ Earhart & Taylor 2016. See also Bergenmar & Leppänen 2017: 240, 242.

References

- Bailey, M. Z.** (2011). All the digital humanists are white, all the nerds are men, but some of us are brave. *Journal of Digital Humanities*, 1(1). Retrieved from <http://journalofdigitalhumanities.org/1-1/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-brave-by-moya-z-bailey/>
- Bergenmar, J., & Leppänen, K.** (2017). Gender and vernaculars in digital humanities and world literature. *NORA—Nordic Journal of Feminist and Gender Research*, 25(4), 232–246.
- Earhart, A. E., & Taylor, T. L.** (2016). Pedagogies of race: digital humanities in the age of Ferguson. In M. K. Gold & L. F. Klein (Eds.), *Debates in the digital humanities*. Minneapolis, MN: University of Minnesota Press. Retrieved from <https://dhdebates.gc.cuny.edu/read/untitled/section/58ca5d2e-da4b-41cf-abd2-d8f2a68d2914>
- Enoch, J., & Bessette, J.** (2013). Meaningful engagements: feminist historiography and the digital humanities. *College Composition and Communication*, 64(4), 634–660.
- Gold, M. K.** (2012). *Whose revolution? Towards a more equitable digital humanities*. The Lapland Chronicles. Retrieved from <http://blog.mkgold.net/2012/01/10/whose-revolution-toward-a-more-equitable-digital-humanities/>
- Huistra, H., & Mellink, B.** (2016). Phrasing history: selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 49(4), 220–229. DOI: <https://doi.org/10.1080/01615440.2016.1205964>
- Jallinoja, R.** (1983). *Suomalaisen naisasialiikkeen taistelukaudet: Naisasialiike naisten elämäntilanteen muutoksen ja yhteiskunnallis-aatteellisen murroksen heijastajana*. Helsinki: WSOY.
- Jarlbrink, J., Snickars, P., & Colliander, C.** (2016). Maskinläsning: Om massdigitalisering, digitala metoder och svensk dagspress. *Nordicom Information*, 38(3), 27–40.
- Kinnunen, T.** (2016). The national and international in making a feminist: the case of Alexandra Gripenberg. *Women's History Review*, 25(4), 652–670.
- Kurvinen, H.** (2018). Toimittajat ja toimitukselliset prosessit mediatekstien takana. *Historiallinen Aikakauskirja*, 116(3), 310–322.
- Kurvinen, H., & Timosaari, N.** (2016). Feministitutkijat historiatieteen haastajina. In J. Pulkkinen & K. Väyrynen (Eds.), *Historian teoria: Lingvistiksestä käänneestä mahdollisuuksien historiaan* (pp. 209–236). Tampere: Vastapaino.

- Lothian, A., & Phillips, A.** (2013). Can digital humanities mean transformative critique? *E-media studies*, 3(1). DOI: <https://doi.org/10.1349/PS1.1938-6060.A.425>
- McKee, H. A., & Porter, J. E.** (2010). Rhetorica online: feminist research practices in cyberspace. In E. E. Schell & K. J. Rawson (Eds.), *Rhetorica in motion: feminist rhetorical methods and methodologies* (pp. 152–170). Pittsburgh, PA: University of Pittsburgh Press.
- Mold, A., & Berridge, V.** (2018). Using digitised medical journals in a cross European project on addiction history. *Media History*, pre-published. DOI: <https://doi.org/10.1080/13688804.2018.1506698>
- Paju, P.** (2008). National projects and international users: Finland and early European computerization. *IEEE Annals of the History of Computing*, 30(4), 77–91.
- Paju, P.** (2016). *Raportti digitaalisen historiantutkimuksen kyselyvastauksista*. Retrieved from <https://digihistfinlandroadmapblog.files.wordpress.com/2016/09/digitaalinen-historiantutkimus-kyselytuloksia5.pdf>
- Posner, M. K.** (2012). *Some things to think about before you exhort everyone to code*. Miriam Posner's Blog. Retrieved from <http://miriamposner.com/blog/some-things-to-think-about-before-you-exhort-everyone-to-code/>
- Rasila, V.** (1967). Tietokone historiantutkimuksessa. *Historiallinen Aikakauskirja*, 65(2), 140–146.
- Rhody, L. M.** (2016). Why I dig: feminist approaches to text analysis. In M. K. Gold & L. F. Klein (Eds.), *Debates in the digital humanities*. Minneapolis, MN: University of Minnesota Press. Retrieved from <http://dhdebates.gc.cuny.edu/debates/text/97>
- Risam, R.** (2015). Beyond the margins: intersectionality and the digital humanities. *DHQ: Digital Humanities Quarterly*, 9(2). Retrieved from <http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html>
- Saarimäki, P.** (2018). Bourgeois women and the question of divorce in Finland in the late 19th and early 20th centuries. *Scandinavian Journal of History*, 43(1), 64–90. DOI: <https://doi.org/10.1080/03468755.2017.1353192>
- Solberg, J.** (2012). Googling the archive: digital tools and the practice of history. *Advances in the History of Rhetoric*, 15(1), 53–76. DOI: <https://doi.org/10.1080/15362426.2012.657052>
- Sulkunen, I.** (1989). Naisten järjestäytyminen ja kaksijakoinen kansalaisuus. In R. Alapuro, I. Liikanen, K. Smeds & H. Stenius (Eds.), *Kansa liikkeessä* (pp. 157–175). Helsinki: Kirjayhtymä Oy.
- Umrani, F., & Ghadially, R.** (2003). Empowering women through ICT education: facilitating computer adoption. *Gender, Technology and Development*, 7(3), 359–377.
- Wernimont, J.** (2013). Whence feminism? Assessing feminist interventions in digital literary archives. *Digital Humanities Quarterly*, 7(1). Retrieved from <http://www.digitalhumanities.org/dhq/vol/7/1/000156/000156.html>

Audio Visual Heritage and the Digital Universe

Maiju Kannisto and Pekka Kauppinen

Introduction

Part of a nation's cultural heritage is produced and preserved through audio-visual archives. Whatever has made it into the archival collection has passed severe processes of selection, which secure for certain cultural products a place in the cultural memory of a society. This process is called canonisation.² In this chapter, we explore how the canon is built up in the national audio-visual archive in the digital age. We study this by asking which people and what events and periods are 'remembered' not by the members of a nation, but by a collective national memory resource, the online audio-visual archive of a national public service broadcaster. Like particular natural and historic sites considered as 'heritage', and thus referring to a cultural and historical resource for all generations, past radio and television is now protected through copyright and continued cultural recirculation and exploited as both private and public property.³ Our study focuses on the Finnish public service broadcasting company

Yle (former Yleisradio, founded in 1926) and its online archive. The dataset used in our research consists of Yle's archival metadata, which we analyse as a historical source material using the method of Named Entity Recognition (NER) as it is implemented in the digital tool the Finnish rule-based named-entity recogniser (FiNER).

Yle's online archive, The Living Archive (*Elävä arkisto*), presents part of Finland's audio-visual history. The archive is an illuminating case study as it is a large editorial historical audio-visual service with open access metadata, as well as a high profile among Yle's media output. It was first launched in 2006 on the 80th birthday of the Yle Company and the Living Archive enabled Yle to celebrate its history, while at the same time providing the public with new ways of watching television programmes online. The core idea behind the Living Archive is that audio-visual clips are historical source materials—documents from and about the national past representing Finland's national audio-visual heritage. In this way, the archive contributes the nation-building process by formulating collective national identity, as suggested by Benedict Anderson in his theory of nations as imagined communities.⁴ As media researcher Derek Kompare suggests, the audio-visual heritage serves as a base of legitimacy for audio-visual media and memory; namely, as something worthy of attention, preservation and tribute. It can be used as a cultural touchstone, instantly signifying particular times.⁵

The Living Archive continually publishes new material from Yle's archives for viewing and listening via the online service. What is published is based on topicality, new copyright licences, Yle's current programming strategies, audience wishes and new archival discoveries.⁶ Archival material has accumulated from the first audio recordings at the beginning of the 20th century to the present: Yle selects and presents the audio-visual archive material for current users in constantly new ways by adding and framing the material in 'background articles', written by journalists and archivists. Media history researcher Mari Pajala has analysed the ways in which the Living Archive attempts to make the material 'alive' and meaningful in the present through its journalistic front page, background articles and the possibility for interaction. In this way, by tying in moments of archive television with current events and television programmes, the archive continually connects the present with the past.⁷

The aim of this chapter is to produce new knowledge on the canon built in the Living Archive. However, at the same time, metadata as a source material and name recognition as a tool in a historical study must be analysed because the canon is connected with the limits and possibilities offered by the digital material and tool. In this chapter, we first introduce the research material and our method, the metadata used and the NER-based analysis. This analysis, using the FiNER tool, enables the identification of particular historical individuals, events and years from the metadata material. Finally, we discuss the limits and possibilities of our research process and results.

Finding Voices and Images of the Finnish Past: Metadata and FiNER Analysis

So far, since 2006, tens of thousands of audio-visual clips have been published in the Living Archive, so a computer-assisted method is needed to make larger sense of such a vast source material. The historical researcher's role in delimiting and selecting relevant material for analysis is crucial to the success of the computer-assisted analysis results. In this chapter, the research data does not consist of the media clips of the media files stored in the Living Archive as such, but of the metadata of the media information of these media files. Yle made this metadata material available on 3 January 2018.⁸ Of this metadata, we selected the columns describing the title of the media, the promotion title and the description of the content (in Finnish). These columns describe the content of media clips, but not in general the author information. We did not want to include the latter in this study as we preferred to focus on the constructed national heritage and the individuals presented therein—not the authors of the various media documents. This could, however, be a relevant topic for future historical research.

The use of metadata as data is problematic in some respects because the material is inconsistent and contains some double data and gaps, as well as, at times, false information. An essential part of the historical source criticism is thus to find out the classification and guidelines that have been used to create the metadata. In the Living Archive metadata, the media descriptions have been produced in three main ways. (1) Descriptions have been made by the authors of the original programmes in connection with their original production to serve as programme information for other media. (2) The reporter or administrator of the Living Archive has written a short description of the clip's content in its subject field. Since this description field has not been displayed to the end user, this field is not always completed, and thus about one-tenth of the fields are empty. (3) Due to technical failure in the migration of the archival clips in 2011 (the transfer from one technical platform to another), the media information on the clips' still images were incorrectly entered in the media clips' subject field.⁹ Therefore, in place of the description of the audio-visual content, this field contains the content of the still image, together with the photographer's name. Only the content of some video clips published before 2011 may have been edited and supplemented after that year. In the analysis, we have been aware of the above-mentioned issues and taken them into account as much as possible (for example, we excluded the photographers from the analysis).

NER is a task in Information Extraction consisting in identifying and classifying some types of information elements, called Named Entities (NE). It is stated that NER analysis usually responds to the five typical questions in the journalism domain: what, where, when, who and why.¹⁰ In our analysis, we utilised FiNER,¹¹ a rule-based NER tool loosely based on the Swedish NER

HFST-SweNER.¹² At the time of our study, FiNER had not yet been used extensively as a research tool and therefore this chapter is also partly a methodological experimental study.¹³ FiNER was created for the FIN-CLARIN consortium at the University of Helsinki, which is the Finnish part of the European CLARIN (Common Language Resources and Technology Infrastructure) collaboration for developing research infrastructure for language-related resources in humanities and social sciences.

FiNER utilises the Helsinki Finite-State Toolkit and its implementation of the *pmatch* (partial string matching) function,¹⁴ which allows the compilation and implementation of pattern-matching rules as computationally efficient finite-state transducers (FSTs). FiNER's pattern-matching rules employ a number of strategies in finding and disambiguation names, including hints in string structure (such as uppercase letters, affixation, etc.), collocations, runtime adaptation and gazetteers (lists of names). The FiNER tool accepts any plain-text input, but works best on running text that adheres to Standard Finnish spelling and typographic conventions. We used a set of UNIX text-processing utilities to extract relevant segments from the tabular metadata for tagging, as well as to filter out any superfluous data that might slow down or interfere with the NER process.

In order to calculate total frequencies for names in the data, each matched word segment in the output first had to be lemmatised (that is, reverted to its uninflected form). Thus, we had FiNER output lemma forms and morphological analyses for each word and created a Python script that extracted matched sequences and used morphological information to lemmatise each of them. Once all names had been printed out in their lemmatised forms, their frequencies in the data could be calculated with relative ease.

As digital history researchers Kimmo Elo and Olli Kleemola have pointed out, it is essential for the researcher to understand how the computer-assisted analysis produces the results.¹⁵ Thus, we looked at the frequency lists of different categories of FiNER analysis from the perspective of the possibilities and limits set by the technical characteristics of the tool, as well as possible errors. After having removed the names that occurred as a result of technical error, we put together TOP10 and TOP20 lists from the names that have received most mentions in different categories. Finally, in order to understand the results, it was important to check the background articles in the Living Archive and to contextualise them tentatively with Finnish cultural and political perspectives.

Great Men, Journalists and Musicians as Remembered Persons

People's names are generally better recognised than other name entity types.¹⁶ FiNER recognised nearly 12,000 people in our research material; in the analysis, we focused on those who had received most mentions. When analysing the people, we found three main groups dominating the historical personage of the Living Archive: great men, journalists and musicians (see Table 10.1). Among

Table 10.1: Top 20 of the most frequently mentioned people in the metadata of the Living Archive listed by FiNER.

Number of mentions	Top 20 people	Their professions
353	Urho Kekkonen (includes 'Urho Kekkonen' and 'Kekkonen')	President of Finland
223	Arto Nyberg	journalist
125	Kirka	musician
111	Marjukka Havumäki	journalist
105	Vesa-Matti Loiri	musician, actor
90	Mauno Koivisto	President of Finland
89	Mikko Alatalo	musician, journalist, politician
89	Matti	
86	Bettina Sâgbom	journalist
78	Matti Rönkä	journalist
78	Danny	musician
76	Päivi Istala	journalist
74	Göran Palm	journalist
73	Mannerheim	President of Finland
71	Paula Koivuniemi	musician
71	Jaakko Selin	journalist
69	Lasse Mårtenson	composer, musician
69	Antti	
68	Susanna Ström-Wilkinson	journalist
64	Juha Laaksonen	journalist

Note: Photographers are excluded. Source: Author.

the Top 20 of the most represented people, President Urho Kekkonen is the most frequently mentioned person. Kekkonen (1900–1986) was the longest-serving President of Finland (1956–1982) and a politician who achieved an almost unchallengeable political position during his long presidency. President Kekkonen's prevalent role in the archival clips is also emphasised by the fact that his trusted photographer Kalle Kultala received the second most mentions in the material, although this is due to an incorrect overwrite in the media information. Kekkonen is the protagonist in many articles, which include a number of video and audio clips from election campaigns, presidential visits and public speeches, as well as clips of him carrying out his hobbies. During his 25 years of presidency, Kekkonen gave 25 presidential New Year's speeches and hosted 22 Independence Day receptions, both of which are essentially national

audio-visual broadcasts, repeated annually and among the most watched programmes on Finnish television.

President Kekkonen was connected to the Finnish Broadcasting Company Yleisradio in many ways. First, he had close personal relations with people working there. Second, his political career was concurrent with the development of broadcasting in the 1930s and 1940s. Finally, he often appealed to the people in his radio speeches during the growth in radio licences. It can be argued that his long reign was comparable to the monopoly of Yleisradio.¹⁷ Media researchers Lotta Lounasmeri and Johanna Sumiala have argued that President Kekkonen used the new mass media skillfully. He managed to combine different roles by acting as a sovereign political leader in the mass media while, on the other hand, at other less formal occasions, performing as a man of the people by skiing, fishing and meeting people in the countryside.¹⁸ Iconic images of the photogenic president were spread across the mass media, making these audio-visual presidential representations part of the nation's public memory. In this way, the Living Archive played an important but previously mainly neglected role in preserving and presenting a political part of the national audio-visual memory. Furthermore, President Kekkonen not only appears in the archival clips, but also as a character in documentaries, drama series and sketch shows, and as a reference point to which other politicians were reacting. This is not just a past but also a very active part of the national memory, as the cult of President Kekkonen still strongly survives as a nostalgia and longing for strong leadership.¹⁹

Among the Top 20 of the most mentioned people, there was also another President of Finland, Carl Gustaf Emil Mannerheim (1867–1951), who could be characterised as a great man in national history. Mannerheim, Marshal of Finland, was frequently mentioned, even though he appears in few contemporary audio-visual sources. However, he is retrospectively presented in the archive through later historical photographs, film clips, and radio speeches and documents. Furthermore, Mannerheim's character became a part of Finnish popular culture in the form of a controversial doll-animated short film in 2008 and through other films. According to historian Tuomas Tepora, the contradicting views have been an essential part of Mannerheim's mythical role in Finnish cultural memory.²⁰ His legacy also appears in the material when talking of The Mannerheim League for Child Welfare and the Knights of the Mannerheim Cross, awarded to Finnish soldiers.

Along with the great men, men in general dominate the media clips: 72 of the first 100 people are men. Of the 10 women mentioned, the top eight are Yle journalists; only singer Paula Koivuniemi and President Tarja Halonen are also cited. As is well known, in the Western historiographic tradition, women have rarely been treated as active actors until recent years and therefore they have not been given much space in national historiography. Although the tradition has been challenged for almost 50 years now, women are still under-represented in most canonised historical narratives.²¹ This is also the case in the

Living Archive, in spite of equality work being in operation for a long time now. This is also related to the history of the feminisation of Finnish professions. President Tarja Halonen was the first female Finnish President (2000–2012). This can also be observed among the journalism profession, another dominant group in the archive. At Yleisradio in 1972, only 35% of the journalists were women,²² and the number of female journalists only increased significantly from the 1970s onwards.

In general, the journalists of Yle to a great extent appear in the material. The Top 20 listing includes eight Yle journalists (see Table 10.1). In the metadata, journalists are mentioned in many respects. News anchors are mentioned in connection with the archival news clips, while journalists working on the Living Archive are mentioned in connection with the background articles. Journalist Arto Nyberg is the second most frequently cited person in the archive. He has since 2004 hosted a popular talk show named after him, where many Finnish public figures and celebrities have visited and many clips therefrom are published in the Living Archive. In this audio-visual history, journalists are the key figures and agents behind the production of the media clips and thus appear frequently in the material more in this capacity as facilitators and producers, rather than as objects of newsworthy events. This important factor of the media production system needs to be taken into account.

In addition to journalists and presidents, many musicians are also mentioned in the audio-visual archive: six are included in the Top 20 listing. What unites these musicians is great popularity, which often comprises long careers stretching to several decades of hit songs. When compared to the most frequently mentioned bands listed by FiNER, the musicians are more focused on older pop singers and evergreen hits, while the bands category is more versatile and focuses on newer bands. The most frequently mentioned artists are those who have had visibility in Yle's music programmes, such as Eurovision Song Contests and festival recordings. For example, Mikko Alatalo's prominence in the archive emanates from his versatile TV work in Yle's music programmes in the 1970s and 1980s.

Eurovision Song Contests, Sports and Wars Unite the Nation

Music also plays an important role in FiNER's list of the most frequent events. In part, this is due to Yle's major role in recording music festivals and preserving them for the audio-visual heritage in the Living Archive. However, the largest national audio-visual event is not a national but a European event in the form of the Eurovision Song Contest. As an annual and long-lasting television event, the song contest has been an important event of popular culture. Historian Mari Pajala has argued that the Eurovision Song Contests have become a prominent part of Finnish national memory and history. The regular annual contests have particularly participated in Finnish nation-making by

creating media discussions surrounding the meaning of nationality in the collective cultural memory. The Living Archive reviews both the successes and the failures of Eurovision. Finland has historically been very unsuccessful in the contests, but this finally changed in 2006 with the hard rock band Lordi's victory. The long tradition of negative experiences and the decades of disappointment were then forgotten as public festivals reupted in market squares all over Finland.²³ In addition to the international contest, the national Eurovision qualifiers are presented extensively in the archive, which gives many artists a place in the national audio-visual history.

Another popular category of events is major sports competitions, such as the European and World Championships in ice hockey and track and field. The broadcasting of these major sport events has been strategically significant to the Finnish Broadcasting Company since they attract large audiences and are therefore legitimising the company's role as a public service broadcaster.

The third category relates to war. Wars and political conflicts identified by FiNER are listed as events, together with music, sport, art and entertainment, which can, however, be analytically difficult for comparisons. Among the wars, the Finnish Winter and Continuation Wars are particularly well represented in the media material. These wars are significant episodes in Finnish history and form part of the national mythology formulated through national publicity, as well as, for example, through their appearance in schoolbooks.²⁴ Therefore, these two wars also form an important part of Finnish cultural memory.²⁵ The Living Archive preserves the remembrance of wars in connection with different anniversaries.

According to Pajala, many of the traditional moments of the television year are explicitly related to nationality: Eurovision Song Contests and sport events unite the national public in excitement at the presence of their own representatives, while the state's sovereignty is celebrated at the President's Independence Day Reception, and in the national war films.²⁶ These national audio-visual events are accumulated and added to the Living Archive annually.

Nationally Significant Periods as Recognised by Yle

Different periods from the 1960s to the 2010s are remembered relatively evenly in the Living Archive, even though most of the programmes have been preserved only in the last two decades (see Figure 10.1). However, it is important to note that in FiNER's analysis, the recognition of the decade numbers is a challenge: FiNER may not recognise these without using clues from the text contexts, and may instead interpret them as ordinary words rather than as markers for historical periods. This produces gaps in the time series. There are three peak years in the FiNER frequency list: 1976, 1995 and 2008. When tracing these years from the media material, we noticed that during these periods there were an extraordinary number of significant political and cultural events.

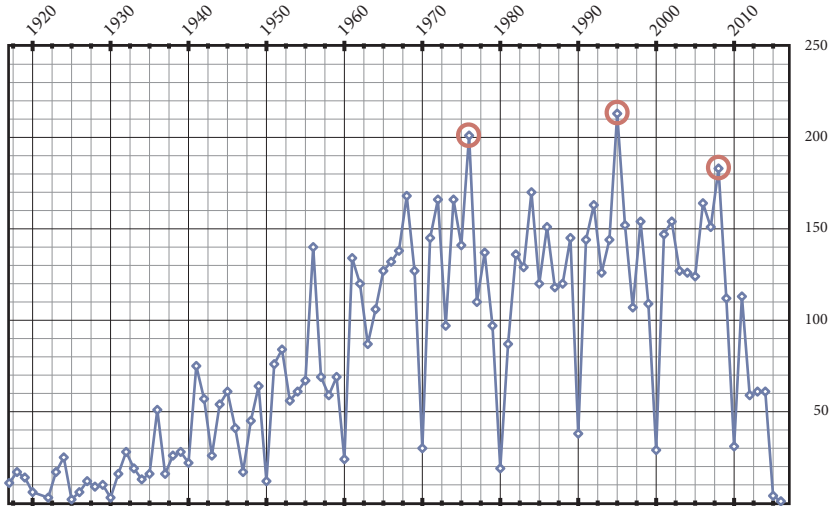


Figure 10.1: The occurrence frequencies of the years 1917–2017 in the time series identified by FiNER from Yle's metadata material. Note: The peak years 1976, 1995 and 2008 have been circled. Since the summer of 2015, all the programmes of the Living Archive have been published via the Areena platform, which causes a harsh decrease in the metadata material.²⁸ Source: Author.

We suggest that these events have become key experiences determining the past and are thus significant parts of the national memory.²⁷ These historical events are presented in the background articles and collectively remembered in the Living Archive.

Among the 1976 media clips, the long-distance runner Lasse Virén's wins in the 5,000- and 10,000-metre runs at the Montreal Olympics and the drama series *Myrskyluodon Maija* (*Maija of Myrskyluoto*), shown on Yle's main channel, are both highlighted as key collective experiences. The latter television series was set in the Finnish archipelago during the 19th century and pictured the life of the protagonist Maija and her family in six episodes. Lasse Mårtenson, who was among the Top 10 in FiNER's list of the most represented people, composed the music for the series. The series has been characterised as unforgettable by many, in large part due to the music. In actual fact, the piano transcription of the theme music is the best-selling Finnish music publication of all time.²⁹ The Living Archive contains several background articles about both the music and new versions of the songs, such as the highlights of 15 versions of the theme music performed by various artists.³⁰ As evidenced by its many occurrences in our source material, the melancholic composition apparently succeeded in touching the collective Finnish psyche and therefore in becoming a prominent part of the national audio-visual heritage.

In 1995, two main events dominate the media material. The first was that Finland joined the European Union. This was the topic of a large number of media clips, such as news and current affairs programmes, as well as sketches, and there are many background articles belonging to the EU thematic. The second event was winning the Ice Hockey World Championship, which was and has continued to be a key national experience. Ice hockey is Finland's largest sport in terms of media visibility and every spring the Ice Hockey World Championships attract a large national audience. The televised winning final in 1995 was watched by 46% of all Finns. Cultural historian Hannu Salmi has connected the boom of ice hockey to changes within the Finnish media landscape. In the 1980s, when the national team began to enjoy success, ice hockey became more visible and television rights became the subject of a struggle between public service and commercial media companies.³¹ The victory in 1995 was the first World Championship win for any Finnish sports team and the nation was united in celebrating the victory together. In public, the victory was made into a question of national self-esteem.³² Therefore, remembering the great victory also has an important role in the Living Archive.

During the third peak year of 2008, the key experiences are connected to the political elections, which received significant media attention. In Finland, the Finns Party (*perussuomalaiset*, formerly the True Finns Party) was the biggest winner in the municipal election and brought forward a strong criticism of the existing immigration policy in public discussion. Another notable political event that contributed to the 2008 peak was an event occurring outside Finland, when, in the United States, Barack Obama was elected President. However, most of the accumulated media clips during the year concerned the Kauhajoki school shooting, when a student shot and killed 11 people, including himself. After an earlier Finnish school shooting in Jokela the previous year, it was not possible to treat the Kauhajoki shooting as an individual incident, and journalists thus tried to find political and social explanations for the tragedy (for example, related to gun legislation).³³ The Living Archive documented this debate.

In addition to these events, many other 2008 media clips focused on the 30th anniversary of the Finnish rock festival *Provinssirock*. Even if many key experiences are connected with Finnish or world history, the audio-visual nature of the archive places a particular emphasis on audio-visual events. These events emphasise the particular role of Yle—for example, regarding television series, the Eurovision Song Contest, sports and footage of music festivals. Therefore, the national audio-visual heritage is Yle's particular heritage and its particular contribution to Finnish history; Yle owns a large proportion of the Finnish audio-visual heritage. Commercial media companies, like MTV founded in 1957, have no similar archives and there are no programmes left from the early decades of the independent Finland. During the early years of television, programmes reached large audiences, allowing programmes to create iconic images and become part of the national public memory.³⁴ However, in the

Living Archive, commercial television and radio are left aside, rendering it an important resource for the public service company Yle in the struggle among the various media to legitimise its cultural position.

Finally: The Limits and Possibilities of Interpretation

Digital archives are organic entities that grow and change their shape as new materials are added.³⁵ The strategy of Yle's Living Archive is to grow constantly, which produces a cyclical nature for the production of this national audio-visual heritage. Annually recurring events related to the audio-visual culture have generated the most archival material over the years. In this way, Yle has created a national audio-visual annual calendar to commemorate significant moments, like Independence Day. Another strategy of the Living Archive, tackling contemporary topical issues, in turn, binds the past to this present moment. The archive actively follows current topics and events. This strategy equals Derek Kompare's notion that the legitimation of television, as in the case of archive, was based not on the existing canon, as it had been with film, but on the idea of heritage. While the canon tends to be separated from everyday life and located in a refined, timeless sanctuary, a heritage is part of the lived, historical experience of a culture.³⁶ Both of Yle's strategies have succeeded in producing the most popular content in the archive, such as responses to topical issues and death notifications for public figures.³⁷

It is important to note that the choices of which media clips are published in the Living Archive and their output are political choices that remain invisible to archive users.³⁸ As Pajala has pointed out, the archive has limited possibilities to publish anything not shown by public service television. As a result, former legislation and cultural norms continue to restrict today's debate on the subject of audio-visual archives.³⁹ In addition, the copyright agreements significantly shape the publication choices, as not all of Yle's archive material can be published. During the first years, the only materials published were Yle's own journalistic programmes and films redeemed from old film companies. Another restriction is that all released music has to be reported to copyright organisations, but since 2015, a separate agreement has made it possible to publish many old music programmes, such as the Finnish qualification competitions for the Eurovision Song Contest and the music programme *Hittimittari* from the 1980s. A drama agreement in 2016 has also enabled Yle to release drama series and films, with long publishing rights. However, this is not only about a question of choice, since not all of the old material is actually available. Only since 1984 has all self-produced TV material been archived. In addition, the old programmes may have very poor metadata, making them more difficult to locate.⁴⁰

In addition to the choice of data, we also have to take into account FiNER's specific internal limitations, which are similar to those of other rule-based

NER systems: any shortcomings in rule formulation or gazetteers result in false positives and misclassification of matched names, which if no rule applies, can make a name go altogether unnoticed. Single-word names of organisations and events are particularly difficult to identify without context clues or structural hints. This is particularly true for short texts such as the metadata entries, where the reader is often assumed to be familiar with names and their referents. In the case of rule-based NER systems, any misses and tagging errors are human in nature as they arguably reflect the system developer's ability or inability to formulate exhaustive rules, as well as their oversights when building the gazetteers.

Metadata as a research material also raises wider ontological questions of historical knowledge: What can we really know about the audio-visual cultural heritage on the grounds of metadata and FiNER analysis? Digital history researcher Michael J. Kramer discusses the relationship between historians and digital archives and describes history as a meta-metadata.⁴¹ There are so many layers of interpretation: metadata is not merely descriptive, but is also already an interpretation of an archivist. Another issue stems from the fact that the results of the FiNER analysis are the lists of frequencies of names. These frequencies are not directly related to media material: FiNER as a tool has its own limitations and rules on how it interprets and categorises the metadata that has been pre-processed by the researcher, something which also brings its own specific limitations. Finally, the historical researcher interprets the results of FiNER. We suggest that it is actually already a meta-meta-metadata of the target of study. In order to understand all these layers of interpretation, we need collaboration between archivists and historians to make visible the guidelines and ways of writing metadata. The most fruitful approach would be for them to cooperate in negotiating how metadata best serve both parties.

A historian also needs to be aware of the functional logic of the digital tool for being able to recognise the bumps on the road of interpretation.⁴² Our NER-based analysis revealed an interesting emphasis on people, events and years in the audio-visual heritage constructed by the Living Archive. However, the FiNER analysis did not shed light on why and how these certain topics are represented in the archive. Instead, what the analysis did show is why audio-visual heritage is constructed in a certain way.

Metadata, like Yle's metadata, is often messy and requires a significant amount of selection and pre-processing. In addition, the metadata material, as well as the digital tool, has limitations, as described above.⁴³ The large volume of data can compensate for some of the technical limitations. However, it is important to acknowledge that, in the humanities, the shift from small smart data to big data is not just technological; in fact, it seems to be even more of a methodological shift. Methodologically, it means the shift from close reading to distant reading. In this paradigm, instead of reading a few selected texts, we can analyse an entire collection of relevant textual data.⁴⁴ However, the cultural contextualising and close reading of the themes pointed out by the results of NER-based analysis still play an important role in the analytical process. Only

after the cultural contextualising and close reading do the word lists come alive, so they can animate the role of the audio-visual heritage in the construction of the Finnish imagined national community broadcast on radio and television.

Acknowledgements

We would like to thank the research project From Roadmap to Roadshow, funded by Kone Foundation, for organising the Digital History Finland workshop. Assistance provided by Yle's Living Archive was greatly appreciated. Thanks also to editors and anonymous reviewers for their suggestions and criticisms, all of which made this a better text.

Notes

- ¹ An earlier, Finnish language version of this research entitled *Kekkonen, Euroviisut ja Helsinki—kansallinen audiovisuaalinen perintö NER-analyysin tunnistamana* was published in the journal *Ennen ja nyt* (2/2019). This edited version is published with the permission of the journal *Ennen ja nyt*.
- ² Assmann 2008: 100.
- ³ Kompare 2005: 103.
- ⁴ Anderson 2007/1983.
- ⁵ Kompare 2005: 102–103.
- ⁶ Yli-Ojanperä 2018.
- ⁷ Pajala 2010: 134, 142.
- ⁸ Material is in the form of a CSV file, which is 18 megabytes in size and contains 34,816 entities.
- ⁹ Elina Yli-Ojanperä's interview 23 May 2018.
- ¹⁰ Marrero et al. 2013: 484.
- ¹¹ In this chapter, we use a work version between two publications, which, however, is very similar to the version that has been published at the end of the year 2018. See <http://urn.fi/urn:nbn:fi:lb-2018091301>.
- ¹² Kokkinakis et al. 2014.
- ¹³ However, see Kettunen et al. 2017.
- ¹⁴ Karttunen 2011.
- ¹⁵ Elo & Kleemola 2016: 154.
- ¹⁶ van Hooland et al. 2015: 13.
- ¹⁷ Puro 2016: 23–24.
- ¹⁸ Lounasmeri & Sumiala 2016: 3–4.
- ¹⁹ On the cult of President Kekkonen and his popular cultural image, see Kallioniemi, Kärki & Mähkä 2016.
- ²⁰ Tepora 2015b: 8–11.
- ²¹ Halldórsdóttir, Kinnunen & Leskelä-Kärki 2016: 19.
- ²² Kurvinen 2013: 68.
- ²³ Pajala 2006: 351–368.

- ²⁴ Torsti 2012: 135–155; Ahonen 2017: 10–11, 92–95.
- ²⁵ Tepora 2015a.
- ²⁶ Pajala 2006: 24; Pajala 2012.
- ²⁷ Cf. Sumiala-Seppänen 2007: 283; Torsti 2012: 63–73. On collective memory, see Benton 2010: 1, 5.
- ²⁸ Elina Yli-Ojanperäs interview 20 June 2019.
- ²⁹ Helsingius 2010.
- ³⁰ Yle Elävä arkisto (Living Archive) 2015.
- ³¹ Salmi 2015: 13, 32–46.
- ³² Kannisto 2015: 95–96.
- ³³ Raittila et al. 2009: 112.
- ³⁴ Pajala 2010: 135.
- ³⁵ Jarlbrink & Snickars 2017: 1240.
- ³⁶ Kompare 2005: 105.
- ³⁷ Yli-Ojanperä 2018.
- ³⁸ Cf. Jarlbrink & Snickars 2017: 1229.
- ³⁹ Pajala 2010: 140.
- ⁴⁰ Yli-Ojanperä 2018.
- ⁴¹ Kramer 2014.
- ⁴² See also Elo & Kleemola 2016: 154.
- ⁴³ See also Kannisto 2016.
- ⁴⁴ Schöch 2013.

References

- Ahonen, S.** (2017). *Suomalaisuuden monet myytit: Kansallinen katse historian kirjoissa*. Helsinki: Gaudeamus.
- Anderson, B.** (2007/1983). *Kuvitellut yhteisöt: Nationalismin alkuperän ja leviämisen tarkastelua*. Tampere: Vastapaino.
- Assmann, A.** (2008) Canon and archive. In A. Erll & A. Nünning (Eds.), *Cultural memory studies: an international and interdisciplinary handbook* (pp. 97–107). Berlin and New York, NY: Walter de Gruyter.
- Benton, T.** (2010). Introduction. In T. Benton (Ed.), *Understanding heritage and memory* (pp. 1–5). Manchester: Manchester University Press.
- Elo, K., & Kleemola, O.** (2016). SA-kuva-arkistoa louhimassa: Digitaaliset tutkimusmenetelmät valokuvatutkimuksen tukena. In K. Elo (Ed.), *Digitaalinen humanismi ja historiatieteet* (pp. 151–190). Historia mirabilis 12. Turku: Turun Historiallinen Yhdistys.
- Halldórsdóttir, E. H., Kinnunen, T., & Leskelä-Kärki, M.** (2016). Doing biography. In E. H. Halldórsdóttir, T. Kinnunen, M. Leskelä-Kärki & B. Possing (Eds.), *Biography, gender and history: Nordic perspectives* (pp. 7–34). Turku: k&h, kulttuurihistoria, Turun yliopisto.
- Helsingius, B.** (2010). *Mårtenson, Lasse (1934–2016)*. National Biography of Finland. *Studia Biographica* 4. Helsinki: Suomalaisen Kirjallisuuden Seura. Retrieved from <https://kansallisbiografia.fi/kansallisbiografia/henkilo/1095>

- Jarbrink, J., & Snickars, P.** (2017). Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation*, 73(6), 1228–1243. DOI: <https://doi.org/10.1108/JD-09-2016-0106>
- Kallioniemi, K., Kärki, K., & Mähkä, R.** (2016). Kekkonen ja rock'n'roll. In M. Kaartinen, H. Salmi & M. Tuominen (Eds.), *Maamme—itsenäisen Suomen kulttuurihistoria* (pp. 395–414). Helsinki: Suomalaisen Kirjallisuuden Seura.
- Kannisto, M.** (2015). Sankarit kultajuhlissa. In B. Heiskanen & H. Salmi (Eds.), *Kiekkokansa* (pp. 72–101). Helsinki: Teos.
- Kannisto, M.** (2016). Uusi, ehempi, paras? Digitaaliset tekstilouhinnan työkalut televisiotutkimuksessa. *Lähikuva*, 29(1), 62–70. DOI: <https://doi.org/10.23994/lk.56578>
- Karttunen, L.** (2011). Beyond morphology: pattern matching with FST. In C. Mahlow & M. Piotrowski (Eds.), *Systems and frameworks for computational morphology, volume 100 of communications in computer and information science* (pp. 1–13). Berlin and Heidelberg: Springer.
- Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., & Löfberg, L.** (2017). Old content and modern tools: searching named entities in a Finnish OCRed historical newspaper collection 1771–1910. *Digital Humanities Quarterly*, 11(3). Retrieved from <http://www.digitalhumanities.org/dhq/vol/11/3/000333/000333.html>
- Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., & Borin, L.** (2014). HFST-SweNER—a new NER resource for Swedish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association.
- Kompare, D.** (2005). *Rerun nation: how repeats invented American television*. London and New York, NY: Routledge.
- Kramer, M. J.** (2014). Going meta on metadata. *Journal of Digital Humanities*, 3(2). Retrieved from <http://journalofdigitalhumanities.org/3-2/going-meta-on-metadata/>
- Kurvinen, H.** (2013). 'En mä oo mies enkä nainen. Mä oon toimittaja': Sukupuoli ja suomalainen toimittajakunta 1960- ja 1970-luvulla. *Acta Universitatis Ouluensis B Humaniora* 113. Oulu: Oulun yliopisto.
- Lounasmeri, L., & Sumiala, J.** (2016). Promootiokulttuurin kuningas Kekkonen. *Lähikuva*, 29(2), 3–5. Retrieved from <https://journal.fi/lahikuva/issue/view/3995>
- Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbis, J. M.** (2013). Named Entity Recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces* 35(5), 482–489. DOI: <https://doi.org/10.1016/j.csi.2012.09.004>
- Pajala, M.** (2006). *Erot järjestykseen! Eurovision laulukilpailu, kansallisuus ja televisiohistoria*. Nykykulttuurin tutkimuskeskuksen tutkimuksia 88. Jyväskylä: Jyväskylän yliopisto.

- Pajala, M.** (2010). Cultural memory and its limits on the Finnish public service broadcaster's online archive. *Critical Studies in Television*, 5(2), 133–145.
- Pajala, M.** (2012). Televisio kulttuurisen muistin mediana: Miten itsenäisyyspäivä alkoi merkitä sotamuistelua? In E. Railo & P. Oinonen (Eds.), *Media historiassa*, *Historia mirabilis* 9 (pp. 127–150). Turku: Turun historiallinen yhdistys ry.
- Puro, J.-P.** (2016). Puhumalla hallitseminen: Vahvan johtajuuden ilmentyminen Kekkonen radiopuheissa vuosina 1937–1967. *Lähikuva*, 29(2), 22–42. Retrieved from <https://journal.fi/lahikuva/issue/view/3995>
- Raittila, P., Haara, P., Kangasluoma, L., Koljonen, K., Kumpu, V., & Väliiveronen, J.** (2009). *Kauhajoen koulusurmat mediassa*. Tampereen yliopisto, tiedotusopin laitos: Journalismin tutkimusyksikkö. Retrieved from <http://tampub.uta.fi/bitstream/handle/10024/65360/978-951-44-7654-9.pdf>
- Salmi, H.** (2015). Kiekkokansa ryhmäkuvassa. In B. Heiskanen & H. Salmi (Eds.), *Kiekkokansa* (pp. 13–46). Helsinki: Teos.
- Schöch, C.** (2013). Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital Humanities*, 2(3). Retrieved from <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>
- Sumiala-Seppänen, J.** (2007). 'Joku raja ja loppu se on saatava tällaisille keskusteluille!' Jatkoaika ja kansallisten kertomusten konflikti 60-luvun suomalaisessa televisiossa. In J. Wiio (Ed.), *Television viisi vuosikymmentä: Suomalainen televisio ja sen ohjelmat 1950-luvulta digiaikaan* (pp. 280–291). Helsinki: Suomalaisen Kirjallisuuden Seura.
- Tepora, T.** (2015a). Mikä tekee 'sota-ajasta' muistettavan? Sota ja kollektiivinen muistaminen. *Ennen ja nyt*, 2015(2). Retrieved from <http://www.ennenjanyt.net/2015/08/mika-tekee-sota-ajasta-muistettavan-sota-ja-kollektiivinen-muistaminen/>
- Tepora, T.** (2015b). Sota, Tuntematon ja Mannerheim. *Futura*, 34(3), 5–11.
- Torsti, P.** (2012). *Suomalaiset ja historia*. Helsinki: Gaudeamus.
- van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., & Van de Walle, R.** (2015). Exploring entity recognition and disambiguation for cultural heritage collections. *Literary and Linguistic Computing*, 30(2), 262–279. DOI: <https://doi.org/10.1093/lc/fqt067>
- Yle Elävä arkisto** (2015). *Myrskyluodon Maija monella tavalla*. Retrieved from <https://yle.fi/aihe/artikkeli/2015/07/30/myrskyluodon-maija-monella-tavalla>
- Yli-Ojanperä, E.** (2018, helmikuu). *Yle Elävä arkisto*. Avoin digitaalinen kulttuuriperintö—course, slides. Yle.
- Yli-Ojanperä, E.** Producer-in-chief, Yle Living Archive. Interviewed by Maiju Kannisto by email 23 May 2018 and 20 June 2019.

Data systems

FiNER: <http://urn.fi/urn:nbn:fi:lb-2018091301>

Tools for processing FiNER output: <https://github.com/pkauppin/finer-utilities>

Digitised Parliamentary Talk and Nordic Allemansrätten

Matti La Mela

Introduction

Allemansrätten, a right of public access to nature, is an integral part of the identity and lifestyle of the people in the Nordic countries. This principle, which is commonly seen as an age-old tradition, allows everyone to access and use resources in the wild even without the landowner's consent. Despite its major role in contemporary Nordic societies, the roots and the development of this institution are little researched and not well known. This chapter contributes to the historical revision of *Allemansrätten* by studying public uses of the concept in Finland in the 20th century. Such a broad study is possible using the recently digitised documents of the Finnish Parliament, which offer a unique view on how central societal concepts have been defined and used in public discussion. The chapter asks how and when *Allemansrätten* actually emerged as a term, and to which discursive environments the concept was tied in the public debates of the 20th century.

The aim of the chapter is to challenge the common view of *allemansträtten* as an age-old and stable tradition, and to demonstrate how the concept is historically constructed and has been flexibly used as part of different political discourses. In particular, the chapter focuses on the principle of universality inherent to the modern *allemansträtten*. Even though it is acknowledged today that *allemansträtten* took its form only with the modern processes of urbanisation, growth of free time, and development of new ways of recreation in nature after the 1930s, many authors build continuity to pre-modern Nordic legal culture or access practices.¹ On the other hand, some critics have proposed that *allemansträtten* was actually an ideological move to socialise private land to the use of everybody.² Most recently, however, a mid-way has been sought, where the past and modern cultures of access to nature are discussed separately, to emphasise differences in the social contexts and to demonstrate the parliamentary political support given to *allemansträtten* in the 20th century.³

The chapter uses methods of text mining, and uses the Finnish parliamentary documents as its source. This digitised data has been created only very recently, and is discussed therefore rather thoroughly in the second section of the chapter. It is notable, however, that other digitised sources, for instance newspapers, have already been used for studying parliamentary debates.⁴ The digitised newspapers are currently available only until 1929. The new digitised parliamentary documents, therefore, are not only an important dataset for studying policy and law-making, but also offer an important perspective over the broader public debate after 1929, which no other complete digital collection in Finland currently represents. The digitised parliamentary sources of other countries have been used in historical and social scientific research. These include straightforward debate analysis through keyword searches, but also the use of more complicated methods such as sentiment analysis.⁵

This chapter studies the history of the uses of *allemansträtten* in the parliamentary data in two steps. First, in the section entitled ‘*Allemansträtten* Emerges’, below, keyword searches and frequency analysis is used for discovering the general trajectory of the term in the complete dataset from 1907 until 2000. The parliamentary debates are understood as reflecting common language use of the time, and thus reflecting topics which were central for the contemporary public discussion. Second, in the section entitled ‘*Allemansträtten* since the 1970s’, below, the aim is to study differences in the discursive environments where *allemansträtten* was used. Key parliamentary debates are identified, and the parliamentary data is text-mined into two detailed debate corpora from the 1970s and 1990s. These two corpora are analysed with collocation analysis and topic modelling, and the results are contrasted with each other. The hypothesis of the chapter is that the term appears only after the Second World War, which has been preliminarily confirmed from sporadic sources.⁶ However, in which discussions did this take place, and how was the modern and commonly acknowledged concept appropriated in the public discussion?

Digitised Parliamentary Talk of the 20th Century: Quantity and Quality

Parliamentary documents are a classic source in history. They are used to study the past national legislative work, but also offer a broader view on the political culture and political language, as well as major societal issues of the time.⁷ Internationally, parliamentary sources have been digitised only in the past 10 years and are now available for research use.⁸ In the Finnish case, the parliamentary documents of the latest decades have been available digitally for several years; however, the digitisation of the documents of the unicameral parliament from 1907 until 2000 were finished only recently: the online digital collection was inaugurated in September 2018.⁹

The digital collection includes the documents of the Finnish *Eduskunta*, unicameral Finnish Parliament, which convened for the first time in 1907, when the country still formed an autonomous Grand Duchy as part of the Russian Empire. The parliamentary reform leading to *Eduskunta* was radical at its time, as it expanded the suffrage to the whole of the male and female population and, first in the world, enabled women to stand for election. From a more concrete point of view, however, there was important institutional continuity, as the national and local representative rights had already been exercised since the 1860s. The national Diet or Assembly of Estates, which gathered the representatives of the four estates for legislative work, had convened in 1809 and after that regularly since 1863. As Pekonen has shown, foreign parliamentary practices were carefully studied in the 19th century and the basis for the parliamentary procedure was established, for example, regarding minute keeping.¹⁰

The digitised collection comprises the printed volumes that have been compiled during the parliamentary season (ranging usually from February to January).¹¹ The parliamentary sources contain both static documents, which were the basis of the parliamentary work or produced in the legislative process, and dynamic minute keeping, which recorded the speeches and the procedure of the sessions. It is important to note that even though the minutes were recorded in detail and directly, they have gone through minor editing in the transcription process at the Records Office (for example, regarding the use of dialects).¹²

The parliamentary documents follow until 1975 the publication practices established already in the 19th century, according to which the documents were grouped together per legislative case (minutes proceed chronologically). The materials of the season were published both in Finnish and in Swedish, of which the Finnish collection forms the complete collection, and Swedish texts include translations of the main documents and a summary of the minutes. The annual publications consist of two to four volumes of the Minutes of the parliamentary sessions (*Pöytäkirjat*), one volume of the Swedish summary (*Protokoll i sammandrag*), three to five volumes of the Documents in each language (*Asiakirjat, Handlingar*) and the Annexes (*Liitteet*, only in Finnish).¹³ For

the season 1975/II (which started in September), the Documents and Annexes began to be regrouped in Documents series (A–F in Finnish, A–D in Swedish) according to their type. Besides this, a separate Index (*Hakemisto, Register*) was published, which had been included prior to 1975/II in the last volume of the Minutes.¹⁴ The index of the Minutes has been extracted during the digitisation into a separate file for the seasons 1948 to 1975/I.

The digitised material has been published online by the Finnish Parliament.¹⁵ In the digitisation process, the separate volumes have been scanned, optical-character recognised (OCR) and stored as PDF/A files. The online interface allows keyword searching of the text content of the pdf files, which, however, has considerable challenges due to errors and typographical features included in the OCR output.¹⁶ The separate pdf files (the different printed volumes) can be downloaded and used separately. This complete dataset consists of 92.4 GB of optically recognised pdfs, of which 61.6 GB are in Finnish and 30.8 GB are in Swedish. The pdf files have been named according to their publication types (described above) and publication year, but the dataset does not include any text or metadata files.

For this chapter, the text content of the pdfs was extracted with the pdftotxt tool included in the open source Xpdf.¹⁷ The pdf format does not offer a good structure for the text, and the text output was cleaned for the analysis conducted here. In ‘*Allemansrätten Emerges*’ below, the raw documents could be used while they suited keyword searches and close reading of the search results. For the topic modelling carried out in ‘*Allemansrätten since the 1970s*’, below, two debate corpora were refined manually: only selected law cases were picked, and from these, only speeches by representatives were extracted and corrected into simple text.¹⁸

As the digitisation of the parliamentary documents has been carried out with printed material, the OCR quality of the material is generally very good. However, it varies a lot and needs cleaning to be used for detailed textual analysis. As there is currently no previous research or evaluation of the quality of the material, I conducted a very rough analysis of the word recognition rates concerning the complete material per decade. I used the LAS-tool,¹⁹ which has a functionality for word recognition rate.²⁰ The recognition was conducted per file, and an average was calculated for document type per year.

As shown in Figure 11.1, the recognition rates are between 60% and around 95%, mainly being over 80%. In comparison to the digitised Finnish newspapers, the quality is very good, as a word accuracy of around 70% to 75% has been reported for the historical newspapers.²¹ It is notable, however, that the quality varies to some degree in the parliamentary material, and is lowest for the documents published in the 1920s and 1930s. Furthermore, the LAS-tool uses only one language when detecting recognition rate in the complete file. The Minutes in particular have a lower detection rate for the earlier years, as Swedish was used more commonly by the MPs. Thus, the real recognition rate is slightly higher than shown in the graph, but the general trend surrounding the quality of the digitisation is clearly visible.

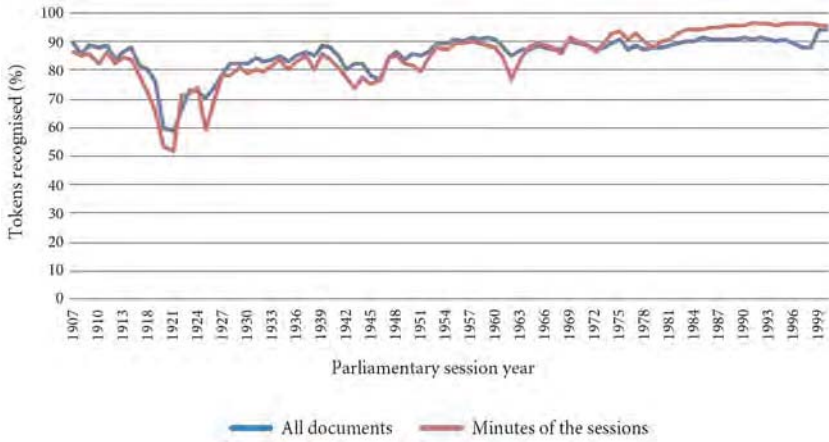


Figure 11.1: Tokens recognised on average per year in the Finnish-language, digitised parliamentary documents. Source: Author.

Allemansrätten Emerges: From Every Man to Fishing and to Modern Access Rights

In this section, I will explore the emergence of *allemansträtten* as a commonly known and shared concept in the Finnish public discussion of the 20th century. This history is not known, and offers an important perspective on the temporal tension inherent in *allemansträtten*. On the one hand, *allemansträtten* is commonly narrated as an age-old tradition or as deriving from the specific Nordic culture. For instance, in autumn 2018, the Finnish outdoor association *Suomen Latu* launched a public campaign broadly visible in the national media, which aims to add *allemansträtten* to UNESCO's list of intangible cultural heritage.²² On the other hand, the scholarship has acknowledged, mainly in the Swedish case, that *allemansträtten* saw the light of day only after the 1930s, with the development of urbanisation and modern mass outdoor recreation.²³ As Kardell writes, since its first appearances in the 1930s, *allemansträtten* 'grew to become part of the Swedish nation's soul (*folksjälen*)'.²⁴

How is the case with Finland, then? On a general level, it is known that also in Finland, the term *allemansträtten*, or *jokamiehenoikeus* in Finnish, becomes commonly used only after the Second World War.²⁵ The parliamentary data allows us to study the trajectory of the term during the whole century. It also fills an important gap, as the digitised national newspapers are available only until 1929. I will first trace the appearance of the term in the parliamentary data, and then focus on particular instances by studying, on a sentence level, in which ways *allemansträtten* could be used in public talk. In this mapping, I will use the complete dataset, and in the following section of the chapter, I will focus on more limited corpora to study the discursive environments in more detail.

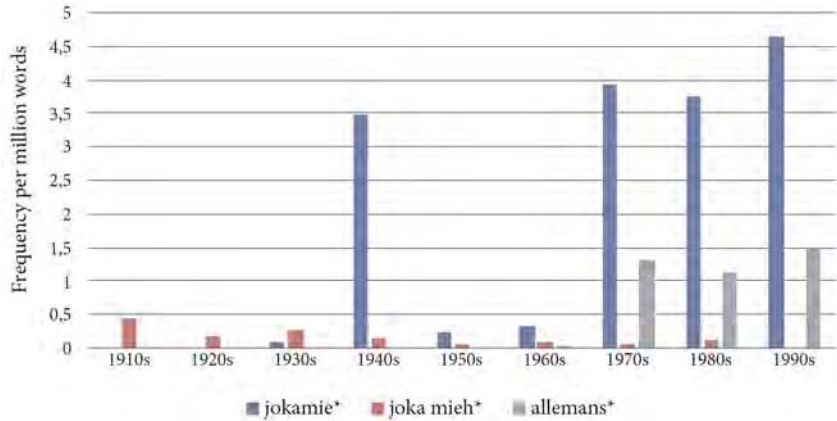


Figure 11.2: Frequencies of ‘everyman*’ (*jokamie**, *joka mieh** in Finnish, and *allemans** in Swedish) per million words in the parliamentary documents per decade. Source: Author.

I used multiple keywords in the search in order to shed light on the differences between the uses in the two national languages, Finnish and Swedish, and to capture a broader variety of terms used.²⁶ As we can see from Figure 11.2, there are differences with regard to how much the terms were used. Before the 1940s, the Swedish term *allemansrätten* is not used at all, and the Finnish *joka miehen* (every man’s) appears to some extent. In the 1940s, there is a curious peak in the use of the term in Finnish. Finally, only in the 1970s, the use of the written forms *jokamiehen* (everyman’s) and *allemans** really boom and become common. The other word forms are used less frequently and disappear.

When we look at the actual uses of the terms in the parliamentary data, we can shape three periods. In the pre-Second World War era, the term ‘everyman’ referred to a common, ordinary person. During the Second World War, ‘everyman’ was used in the context of nature and outdoor activities with the introduction of the wartime ‘everyman’s right of fish’. Finally, only gradually since the 1960s, ‘everyman’ became employed in its main contemporary meaning as describing *allemansrätten*.

The Finnish *joka mies*, everyman or every man, is obviously a term that has already appeared in the language for a longer amount of time. In this meaning, it refers to an ordinary person or generally everybody. It is important to note that this Finnish term translates into the Swedish *var man* (everyman), and does not bear similarity to the Swedish-language word root *alleman* pointing at *allmän* (common, general) and *allmänningar* (the commons). In general, then, the origins of the Finnish vocabulary are less bound to land ownership and social conflict.²⁷ Moreover, in the early century, the term was not yet used in the context of outdoor activities or access to nature. We find manuals, guidebooks and even magazines for the ‘everyman’ (for example, the ‘Everyman’s—and

every woman's—weekly' first published in 1907, later called the 'Everyman's weekly'.²⁸ In the parliamentary data of the early century, this is how '*joka mies*' (every man) was used. In the Minutes of 1935, national defence was supported by stating that the 'defence question is a question for every man'.²⁹

Interestingly, the first major uses of 'everyman' as explicitly concerning public access to nature took place in the early 1940s in relation to fishing rights. During the wartimes, temporary fishing rights were enacted in Finland in 1941 to alleviate the shortage of food. At first, the rights concerned fishing in one's local waters, but these were extended in 1943 to allow for some years fishing for everyone in all parts of the country. The right concerned non-commercial fishing by the household, but allowed special permissions for professional fishers, who were immigrants away from their home lakes. The right became labelled as 'everyman's right to fish' or 'everyman's fishing' (*jokamiehen kalastusoikeus*, *jokamieskalastus* or, in Swedish, *var mans fiskerätt*).³⁰ In Parliament, there was also criticism raised against this right; however, its temporary and exceptional nature was acknowledged. In 1946, in the aftermath of the war, it was noted how among the 'everyman fishers' (*jokamieskalastaja*)³¹ there were also many immigrants and locals who were in dire need of fish. It was noted by the social democrat MP Tuomas Bryggari how this right should not only be made an exception, but 'a general law, so that every citizen would have the right to fish'.³²

The discussion about 'everyman's fishing' continued in the 1950s, but it was only in the 1960s that the 'everyman' was used to refer to the modern *alle-mansrätten*. However, we still find discussion about fishing and 'everyman' as everybody or a common person as in 'everyman's sports'. In the parliamentary data, the first uses of *alle-mansrätten* is in Finnish in 1964. In his question to the government, the left-wing MP Kalevi Kilpi (and others) asked about how the future outdoor legislation would react to the question of no trespassing signs. Kilpi added how according to 'custom there existed a so-called everyman's right to roam on another's land without the permission of the owner'.³³ This is the period when the term *alle-mansrätten* became very common in Sweden. The word appeared as part of land planning and urban nature use in the late 1930s, and in the early 1950s, the public in Sweden could read in the newspapers how *alle-mansrätten* did not really appear in Swedish law, but gave everyone the right to move freely in the woods since the early times. In Finland, sporadic appearances are found in these years, but the country seemed to follow its western neighbour only in the following decade.

In the 1970s, then, the term *alle-mansrätten* became very widely used in the parliamentary data. This peak is explained also by the particular moment, while the new outdoor legislation was discussed in the parliament in the early 1970s. However, this expansion in the use of *alle-mansrätten* is also due to the term becoming common in the Finnish language, perhaps even a rhetorical motif describing public access rights in general. The peak in the uses of *alle-mansrätten* does not decrease, but stays at the same level and even increases in the following decades. Moreover, the term *alle-mansrätten* becomes mainly associated

with the area of public access rights to nature. Already at the time, the term was commonly used in public and scholarly discussion in Sweden. This is an indication that the interpretation and demarcation of the modern concept of *Allemansrätten* in Finland was influenced by the Swedish discussions, which is also confirmed by the legal scholarly discussion on access rights to public and private spaces in Finland.³⁴

Even though the digitised newspaper collections are incomplete, it is important to confirm the results with the newspaper material and use it as a parallel dataset to control possible quality problems related to the parliamentary data. I conducted similar keyword searches in the Finnish National Library newspaper dataset until 1929, and used the Sanoma Digital Archives, which host a handful of major Finnish newspapers with national coverage, such as *Helsingin Sanomat* and *Iltasanomat*. The results are very similar. From 1900 until 1929, there are no appearances of *Allemansrätt** (or *alle mans rätt**), and I found uses of ‘*var man*’ (everyman) similar to the Finnish-language use of ‘everybody’ and ‘common person’. The Sanoma Archives’ sources also demonstrate the peak in the 1940s with references to the ‘everyman’s right to fish’. The first uses of the modern term ‘*jokamiehen oikeus*’ (*Allemansrätten*) are found in the 1960s, and the term becomes common in the 1970s material. The first reference from 1962 is in a letter from a reader to the newspaper *Helsingin Sanomat* about the recent private road legislation and whether walking on private roads was permitted.³⁵

It seems clear, then, that the term *Allemansrätten* became commonly used only after the 1960s. How stable have the modern uses been? In the following section, I will explore more carefully the discursive environments where the modern term has been used by contrasting the parliamentary debates of the 1970s with those of the early 1990s—a moment of Finland’s economic and political opening, for example, concerning the EC/EU membership.

Allemansrätten since the 1970s: Mapping the Shifts in Discursive Environments

Today, *Allemansrätten* is a well-known concept which has extended outside its core meaning of public access rights to nature. *Allemansrätten* has been used in other areas than nature to designate the importance of public access rights—for example, how ‘public libraries are an everyman’s right’. Moreover, the concept has been branded as something uniting the Nordic countries, but also representing several key values of these societies, such as Nordic freedom, clean nature and equality.³⁶ What can we learn about the uses of the concept and its expansion beyond a mere ‘right to roam’ by looking at the parliamentary debates?

In this section, we investigate and contrast the uses of *Allemansrätten* at two moments in time: in the early 1970s, when *Allemansrätten* was becoming a commonly used concept, and in the early 1990s, when the term had become an irreplaceable part of discussion about public access rights to nature. This

is achieved by studying the co-occurrence of *allemansträtten* with other terms in the two moments in time with collocation analysis and topic modelling. As the focus is on public discussion, the minutes of the parliament are used in this section. Moreover, the uses of the Finnish term *jokamiehen/oikeus* only are followed, as the majority of the parliamentary speech was in Finnish.

When looking at the debates of the 1970s, concordance searches³⁷ for the term reveal important concentrations of *allemansträtten* in 1973, 1974, 1976 and 1978. They help to identify the key legislative debates in which *allemansträtten* was used and became defined. In the early 1970s, the new Outdoor Recreation Act was discussed and enacted in Parliament. In 1974 and 1976, the debates regarded the expropriation law and, in 1978, the public use of nature was a topical part of the legislative work on the chemical treatment of forest vegetation.

In the 1990s, several concentrations of the use of the term are found with concordance searches. As expected, the clusters in the 1990s are more numerous and regard a broader variety of themes. In the 1970s, 170 concordance hits are found, which consist of over 90 hits about the Outdoor Recreation Act only and of two smaller concentrations of 10 to 20 hits. In the minutes of the 1990s, on the contrary, 401 hits are found, which are divided rather equally in smaller clusters appearing every year. Most annual hits (56) are found in the 1996 minutes. In the 1990s, *allemansträtten* appeared as part of debates on nature use and the natural environment, such as hunting and fishing laws and natural protection legislation. *Allemansträtten* was, however, taken into political debate also in relation to questions about Finland's international relations: the EC/EU membership and legislation concerning foreign ownership in Finland. *Allemansträtten*, therefore, was used in the 1990s more broadly than merely in the context of access rights in the natural environment.

If we move closer to the level of text, this broadening in the uses of the term becomes more visible. For the comparison, I formed two equal-size debate corpora (about 250,000 characters), which were about similar legislative topics. Similarity of the corpora was sought to minimise the effects created by the mere variance in the legislative topics discussed at the two moments in time. The 1970s debates were used as a starting point, and were contrasted with the legislative projects of the 1990s, which regarded environmental protection, fishing and recreational use of nature.³⁸ Furthermore, the corpus was cleaned for obtaining more accurate results for the analysis of co-occurrences: first, the words in the two corpora files were lemmatised with the LAS tool presented in 'Digitised Parliamentary Talk of the 20th Century', above. Second, the texts were trimmed by removing other characters than alphabet letters and deleting the names of the MPs and commonly repeated phrases, such as the greetings addressed to the speaker at the beginning of talks.

When looking at the collocates of '*jokamie**' (everyman*) (statistically the most commonly appearing terms with '*jokamie**'³⁹), we find that *allemansträtten* retains a common core, but has been used to deliver rather diverse messages. The most frequent collocates shared by both corpora include 'right', 'citizen',

'nature', 'ice / lure fishing' and 'Finland'; but also 'Nordic'. When we look at differences in collocation (collocates that are not found in the other corpus), the picture becomes nuanced. There are collocates which related mainly to the differences in law projects discussed (even though similar debates were selected for the comparison), such as 'recreational use' or 'chemical treatment product' in the 1970s.

There seem to be differences in relation to how internal or external the uses of *allemansträtten* have been. The collocates in the 1970s point to the discovery, stabilisation and internal debate about the concept. The most frequent collocates include, for instance, the terms 'age-old', 'property right' / 'property' and 'landowner', and among the statistically most common one finds 'historical', 'heritage' and 'socialise'. In the 1990s, the horizon seems to be broader, and *allemansträtten* appears rather as something that is being challenged by or being related to the outside world. In the 1990s, the statistically most common collocates include 'unique', 'outsider', 'trample' and 'spoil', and also terms related to space, such as 'international', 'Europe', 'European' and 'integration'. Even though the 1990s corpus does not include direct debates about EC membership, the discussions on *allemansträtten* seem to generate questions related to the opening of the borders. Similar fears about the overuse or the weakening of the access rights to nature due to EC membership were also raised in neighbouring Sweden in the early 1990s.

Finally, the differences in co-occurrences between the two corpora are studied by using topic modelling. Topic modelling is a method where separate topics ('patterns of tightly co-occurring terms') are detected in the text corpora through probability analysis.⁴⁰ Topic modelling has proved to be a powerful tool, especially when organising and classifying a large quantity of text documents.⁴¹ In this section, we do not examine the different topics that are found in the corpus, but we study in detail the topics in which *allemansträtten* appears. The topic modelling was carried out using the MALLET tool.⁴² MALLET includes automatic removal of stop words, and after testing several shares of topics, MALLET was run to find 80 topics in both corpora.⁴³ Besides building the topics, MALLET produces an output file, where the words in the corpus are annotated by their topic number. The results of 'jokamie*' (everyman*) appearing in different topics is presented in Figure 11.3.

As illustrated very clearly in the graph above, MALLET clustered *allemansträtten* mainly to two different topics. To a large extent, this results due to the functionalities of the method, as the data had been lemmatised and the written word forms used in the 1970s and 1990s were different: topic 49 includes the term 'jokamies' (everyman), whereas topic 76 includes the term 'jokamiehe-noikeus' (*allemansträtten*), which, as we have seen, had become the common written form in the 1990s

In addition to this, we can discern again a different discursive environment for the two topics and the uses of *allemansträtten* in the 1970s and 1990s. Topic

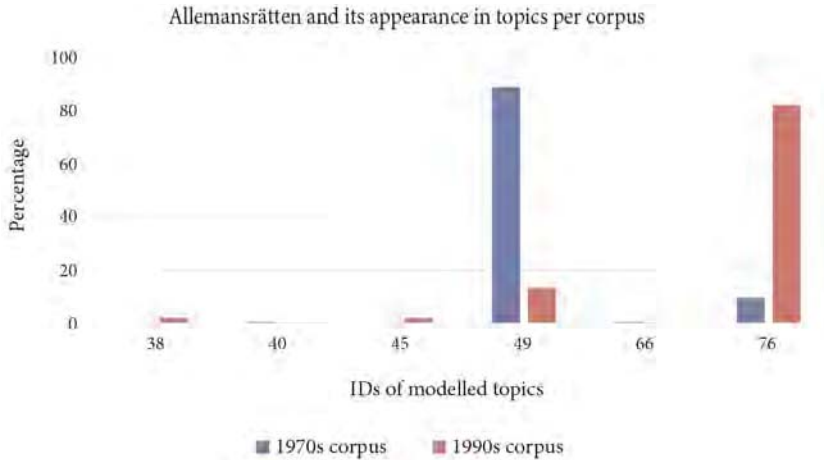


Figure 11.3: Topics in which the term ‘*jokamie**’ (everyman*) was grouped in the topic modelling (per corpus). Source: Author.

49 is mainly about *allemansträtten*: the main keyword given in topic 49 is ‘*jokamies*’ (everyman). This topic includes terms about political debate (political parties) on nature, nature use and access rights. Topic 76, on the other hand, does not have ‘*jokamiehen oikeus*’ (*allemansträtten*) as the main keyword, but the first words of the topic are ‘talk’, ‘damage’, ‘needs to’ and ‘pay’, and then ‘*jokamiehen oikeus*’. Further down the list of topic terms we find ‘damage’, ‘problem’, ‘forest fire’, ‘Finnish’, ‘companies’ and ‘integration’. No names of political groups are listed in topic 49, and its terms are more related to necessity and change, rather than political negotiation or argumentation.

It seems, then, that *allemansträtten* expands beyond its ‘traditional’ range of reference in the period. In the 1970s, *allemansträtten* appears as part of political debate in which the concept was contested, but the contestation limited itself to the question of access rights to nature. The discussion focused on *allemansträtten* itself and its roots. Based on a close reading of the corpus, we can highlight as an example the comment put by the agrarian party MP Mikko Kaarna in 1973: ‘whatever is meant by this everyman’s right, it seems that everybody aims to interpret it in his own way, some of us in a very broad sense.’⁴⁴ After this, Kaarna referred to the laws which already regulated outdoor access. In the 1990s, however, *allemansträtten* conveyed meanings in the political debate that were matters external to access to nature, and rather related to international relations and questions of national identity. For instance, in the debate on the law on natural protection, agrarian party MP Markku Koski emphasised ‘how the right, that was very broad for the Finns, should be valued, namely *allemansträtten*, which does not exist in other European countries to the same extent as in Finland.’⁴⁵

Conclusion

This digital history of *allemansträtten* in the recently digitised documents of the Finnish Parliament has allowed light to be shed on the trajectory of the Finnish *allemansträtten*. The methods used in this chapter (keywords search and collocation analysis) have been useful for identifying key debates and studying long-term changes in those debates related to *allemansträtten*. The chapter has shown how the vocabulary of ‘everyman’ was already being used extensively before the term *allemansträtten* came into use. The introduction of fishing rights in the early 1940s presents an important turning point, and depicts key tensions in access rights to nature. Universality in the fishing rights, which is at the core of today’s *allemansträtten*, was criticised with references to the dishonest ‘vagabond fishermen’ and damages done to the honest locals.⁴⁶ In fact, even though voices were heard for making fishing permanently open for everyone, ‘everyman’s fishing’ was soon limited to allow fishing in the areas where the ‘everymen’ resided. Importantly, economic reasons, including economic distress, were behind the decision to open the fishing rights. In a similar way, the discussions on wild berry picking in the late 19th century regarded the possibilities for the poor of the rural areas.⁴⁷ Only when strangers outside the local community arrived as ‘everymen’ to the woods and lakes did it become crucial to define the limits of access rights: the broadening (and creation) of the right was done for economic reasons, not based on a cultural tradition.

The Finnish vocabulary of ‘*allemansträtten*’ slowly began to point to universal access rights to nature after the 1950s. In fact, at some point in the 1950s, two conceptual traditions became united. The modern institution of *allemansträtten* in Finland was modelled after the Swedish example, yet, the Swedish term was not translated into Finnish, but the Finnish expression of ‘common man’s’ and ‘normal’ access to nature used for wartime fishing rights was taken into use. In the 1970s, then, *allemansträtten* was an established and commonly used concept in Finland. It appeared in the political debates of the decade, but mainly *internally*, as something at the centre of attention itself: a ‘national heritage’ that was discussed and defined by the different national political groups. In the 1990s, the uses of *allemansträtten* had expanded to various legislative debates. With the normalisation of the concept, *allemansträtten* could be used *externally*, to convey meanings related to national values outside the sphere of access rights to nature. In the context of European integration in the early 1990s, it was used to defend a particular Nordic way of life in contrast to European practices of ownership. It is this kind of national symbol, how it is understood (or felt) by the public, which is presented to foreign visitors and used in the branding of the Nordic countries today.⁴⁸

In general, the digitised parliamentary data opens up a new research horizon on the public matters of the 20th century. The data complements the digitised national newspapers, which are available comprehensibly for the first three decades of the century only. Yet, the Finnish case presents shortcomings that can

render digital historical research cumbersome and uninviting: the parliamentary data is not easily accessible through the current web user interface of the Finnish Parliament.⁴⁹ By their nature, parliamentary documents are structured according to time, theme and speaker, and basic search tools and filters that enable the use of these features would satisfy the needs of most historians.⁵⁰ In this chapter, I did not use the web interface, but applied textual analysis methods on the parliamentary data. These methods are applicable by historians with basic computational skills; however, a significant amount of data work is necessary when using such non-structured and partly weak quality data. It seems to me, therefore, that digital historians should pay extra attention to the workload and the trade-off related to the new digital sources: how much can be done ‘easily’ with the existing resources by the historian, and what data and development work can and should be left for broader cooperative projects? Moreover, this also implies that historians should actively partake in the digitisation processes of the key historical sources and the development of the related user interfaces.

Notes

- ¹ Sandell 1997; Husa, Nuotio & Pihlajamäki 2007: 25–26. For everyman’s rights in Finland generally, see Matilainen 2019: 34–35.
- ² Wiktorsson 1996.
- ³ La Mela 2014; La Mela 2016; Sténs & Sandström 2016.
- ⁴ Pekonen 2014; La Mela 2016.
- ⁵ Quinn et al. 2010; King 2016; Abercrombie & Batista-Navarro 2018.
- ⁶ La Mela 2016.
- ⁷ Ihalainen & Palonen 2009.
- ⁸ Ihalainen, Ilie & Palonen 2016. For a list of the available parliamentary corpora by CLARIN, see <https://www.clarin.eu/resource-families/parliamentary-corpora>.
- ⁹ Apilo 2018.
- ¹⁰ Pekonen 2014.
- ¹¹ Valtiopäivien 2006.
- ¹² Pekonen 2014: 181–225; Voutilainen 2016.
- ¹³ The Documents are the main collection of the parliamentary documents (government proposals and committee statements, for instance), and the Annexes contain the motions of the MPs.
- ¹⁴ See, e.g., Riksdagshandlingarna 2012.
- ¹⁵ Digitoidut valtiopäiväasiakirjat 1907–2000, <http://avoindata.eduskunta.fi/digitoidut/>.
- ¹⁶ For example, the speakers’ surnames are spelled in the Minutes with spaced text, which the search engine is unable to interpret. MP Koivisto’s speaking turn appears in the text as ‘Ed. K o I v I s t o.’
- ¹⁷ XpdfReader, <https://www.xpdfreader.com/about.html>.

- ¹⁸ For example, page numbers were removed and text column recognition errors corrected.
- ¹⁹ Linguistic Analysis Command-Line Tool, version 1.5.15, <https://github.com/jiemakel/las>.
- ²⁰ Mäkelä 2016.
- ²¹ Kettunen & Pääkkönen 2016.
- ²² Finnish NGO 2018.
- ²³ Sandell 1997.
- ²⁴ Kardell 2004: 234.
- ²⁵ La Mela 2016.
- ²⁶ The keywords used in the search were *jokamie** and *joka mieh**, which capture both written forms of the term *allemansträtten* in Finnish: *jokamiehen oikeus* and *joka miehen oikeus*.
- ²⁷ La Mela 2014: 271–272; La Mela 2016: 218–219.
- ²⁸ Jokamiehen —ja joka naisen —viikkolehti. No. 3 (19 October 1907).
- ²⁹ ‘Minä tahtoisin sanoa: puolustuskysymys on meillä joka miehen kysymys. Jokaikaisen terveen Suomen miehen on otettava puolustuskysymys omakseen ja omalta osaltaan vastattava siitä.’ Minutes of the Parliamentary Session of 1935 (Minutes 1935 II). Helsinki: Valtioneuvoston kirjapaino, 1936, p. 1623.
- ³⁰ Brofeldt 1943.
- ³¹ ‘... elintarviketilanne ei vielä ole korjautunut sellaiseksi, että voisimme kieltää jokamieskalastajilta mahdollisuuden omalla työllään ansaita itselleen vähäisen särpimen lisän.’ Minutes 1945 II, p. 2187.
- ³² ‘Minun nähdäkseni kalastusoikeus pitäisi saada vakiinnutetuksi, ei ainoastaan poikkeukseksi, vaan yleiseksi laiksi, että kaikilla kansalaisilla olisi kalastusoikeus.’ Minutes 1945 II, p. 2189.
- ³³ ‘Tosin on vanhastaan katsottu, että maantavan mukaan on olemassa ns. jokamiehen oikeus kulkea toisen maalla ilman omistajan lupaa.’ Documents of the Parliamentary Session of 1964 (Documents 1964 V), Question no. 67, p. 2.
- ³⁴ Legal scholar V. K. Noponen was already using variants of *allemansträtten* in Finnish in his work on public and private roads in the 1940 and 1950s. Noponen also refers to concepts developed by the legal scholar S. Ljungman, who was among the first to present this ‘newly-found catchphrase with legal value’, *allemansträtten*. La Mela 2016: 216.
- ³⁵ Laki yksityisistä teistä. Helsingin Sanomat, no. 264 (30 September 1962), p. 31.
- ³⁶ See, e.g., Sandell & Svenning 2011.
- ³⁷ The AntConc corpus analysis tool is used for concordance searches and file views. The concordance search was used to detect keywords in the data and study their uses on sentence level. See <http://www.laurenceanthony.net/software/antconc/>.
- ³⁸ The selected law projects from the 1990s regard: state outdoor recreation area in Teijo, natural protection act, remuneration of environmental damage and the fishing law.

- ³⁹ A range of 10 words before and after the term was used to study collocation.
- ⁴⁰ Blei 2012.
- ⁴¹ See, e.g., Weingart & Meeks 2012; Wehrheim 2018.
- ⁴² MALLET toolkit, <http://mallet.cs.umass.edu/>.
- ⁴³ A different number of topics were tested manually. The number of topics did not affect the clear distinction between topics presented in Figure 11.3. A larger number of topics was preferred for producing more nuanced topics.
- ⁴⁴ ‘Mitä tällä jokamiehen oikeudella tarkoitettaneenkaan, niin näyttää siltä, että jokainen pyrkii sitä tulkitsemaan omalla tavallaan, monet hyvinkin laajasti.’ Minutes 1973 II, p. 1454.
- ⁴⁵ ‘Mielestäni Suomessa nykyisin on pidettävä arvossa sitä oikeutta, mikä suomalaisilla on hyvin laajasti, eli jokamiehenoikeutta, jota ei muissa eurooppalaisissa valtioissa siinä mittakaavassa ole kuin Suomessa.’ Minutes 1996 II, p. 1700.
- ⁴⁶ Minister of Agriculture Eemil Luukka on the law proposal on temporary fishing rights: ‘... jokamieskalastajat tästä lähtien saavat harjoittaa pyyntiään vain vakinaisen tai tilapäisen asuinpaikkansa lähivesistössä. Tämän rajoituksen kautta on tahdottu estää sellainen kulkurikalastajien toiminta, joka juuri on osoittautunut kaikkein haitallisimmaksi paikallisten asukkaiden kalanpyynnille, olkootpa he sitten kalastusoikeutta omaavia tai sitä vailla olevia.’ Minutes 1945 II, p. 2187.
- ⁴⁷ La Mela 2016.
- ⁴⁸ Mission for Finland 2010; Tuulentie & Rantala 2013.
- ⁴⁹ Happily, the situation is improving. The research consortium “Semantic Parliament”, which aims to produce a linked open data and research infrastructure on Finnish parliamentary data, started its work in January 2020. See <https://seco.cs.aalto.fi/projects/sem parl/en/>.
- ⁵⁰ The digitised Canadian parliamentary debates webpage provides a very balanced and user-friendly interface for searching and browsing the parliamentary debates. See <http://www.lipad.ca/>. See also Beelen et al. 2017.

References

- Abercrombie, G., & Batista-Navarro, R.** (2018, May). *‘Aye’ or ‘no’? Speech-level sentiment analysis of Hansard UK parliamentary debate transcripts*. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan (pp. 4173-4180). European Language Resources Association (ELRA). Retrieved from <https://www.aclweb.org/anthology/L18-1659.pdf>
- Apilo, A.** (2018, May). *Digitized documents in Finnish Parliament*. Presentation at Aalto DH Pizza seminar, Otaniemi, Finland.
- Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., Polyanovsky, R., & Whyte, T.** (2017). Digitization of the Canadian parliamentary debates.

- Canadian Journal of Political Science*, 50(3), 849–864. DOI: <https://doi.org/10.1017/S0008423916001165>
- Blei, D. M.** (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1).
- Brofeldt, P.** (1943). *Jokamiehen kalastusoikeus ja muut poikkeukselliset kalastusmääräykset*. Helsinki: Otava.
- Documents of the Parliamentary Session of 1964.**
- Finnish NGO** (2018, 1 October). *Finnish NGO nominates ‘everyman’s right’ for UNESCO heritage list*. YLE News. Retrieved from https://yle.fi/uutiset/osasto/news/finnish_ngo_nominates_everymans_right_for_unesco_heritage_list/10432231
- Husa, J., Nuotio, K., & Pihlajamäki, H.** (Eds.) (2007). *Nordic law: between tradition and dynamism*. Antwerp: Intersentia.
- Ihalainen, P., Ilie, C., & Palonen, K.** (2016). Parliament as a conceptual nexus. In P. Ihalainen, C. Ilie & K. Palonen (Eds.), *Parliament and parliamentarism: a comparative history of a European concept* (pp. 1–16). Oxford and New York, NY: Berghahn Books.
- Ihalainen, P., & Palonen, K.** (2009). Parliamentary sources in the comparative study of conceptual history: methodological aspects and illustrations of a research proposal. *Parliaments, Estates and Representation*, 29(1), 17–34.
- Kardell, L.** (2004). *Svenskarna och skogen*. Vol. 2: *Från baggböleri till naturvård*. Jönköping: Skogsstyrelsens förlag.
- Kettunen, K., & Pääkkönen, T.** (2016, May). *Measuring lexical quality of a historical Finnish newspaper collection—analysis of garbled OCR data with basic language technology tools and means*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portoroz, Slovenia (pp. 956–961). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf
- King, L.** (2016). Future citizens: cultural and political conceptions of children in Britain, 1930s–1950s. *Twentieth Century British History*, 27(3), 389–411. DOI: <https://doi.org/10.1093/tcbh/hww025>
- La Mela, M.** (2014). Property rights in conflict: wild berry-picking and the Nordic tradition of allemansrätt. *Scandinavian Economic History Review*, 62(3), 266–289. DOI: <https://doi.org/10.1080/03585522.2013.876928>
- La Mela, M.** (2016). *The politics of property in a European periphery: the ownership of books, berries, and patents in the Grand Duchy of Finland 1850–1910* (doctoral dissertation, European University Institute, Florence). DOI: <https://doi.org/10.2870/604750>
- Matilainen, A.** (2019). *Feelings of psychological ownership towards private forests* (doctoral dissertation, University of Helsinki, Faculty of Agriculture and Forestry). Retrieved from <http://hdl.handle.net/10138/300433>
- Minutes of the Parliamentary Sessions of 1935, 1945, 1973, 1996.**

- Mission for Finland.** (2010). Mission for Finland: Country brand report 25.11.2010. Final report of the Country Brand Delegation. Retrieved from http://www.maakuva.fi/wp-content/uploads/2011/06/TS_Report_EN.pdf
- Mäkelä, E.** (2016). LAS: an integrated language analysis tool for multiple languages. *Journal of Open Source Software*, 1(6), 35. DOI: <https://doi.org/10.21105/joss.00035>
- Pekonen, O.** (2014). *Debating 'the ABCs of parliamentary life': the learning of parliamentary rules and practices in the late nineteenth-century Finnish Diet and the early Eduskunta*. Jyväskylä: University of Jyväskylä.
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R.** (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209–228.
- Riksdagshandlingarna.** (2012). *Riksdagshandlingarna det finländska parlamentstrycket: handledning* [A guide to the Finnish parliamentary documents]. Library of the Parliament of Finland. Retrieved from <http://www.eduskunta.fi/SV/tietoaeduskunnasta/kirjasto/aineistot/eduskunta/valtiopaivaasiakirjat-tietopaketti/Documents/Handledning-riksdagshandlingarna.pdf>
- Sandell, K.** (1997). Naturkontakt och allemansrätt: om friluftslivets naturmöte och friluftslivets tillgänglighet i Sverige 1880–2000. *Svensk geografisk årsbok*, 73, 31–65.
- Sandell, K., & Svenning, M.** (2011). *Allemansrätten och dess framtid: utredning om allemansrätten*. Stockholm: Naturvårdsverket.
- Sténs, A., & Sandström C.** (2014). Allemansrätten in Sweden: a resistant custom. *Landscapes*, 15(2), 106–118.
- Tuulentie, S., & Rantala O.** (2013). Will free entry into the forest remain? In D. K. Müller, L. Lundmark & H. Raynald (Eds.), *New issues in Polar tourism* (pp. 177–188). Dordrecht: Springer.
- Valtiopäivien.** (2016). *Valtiopäivien alkamis- ja päättymisajat* [The beginning and closing dates of the parliamentary seasons]. Retrieved from <http://www.eduskunta.fi/thwfakta/yht/ohje/valtiopaivat.htm>
- Voutilainen, E.** (2016). Tekstilajitietoista kielenhuolto: puheen esittäminen kirjoitettuna eduskunnan täysistuntopöytäkirjoissa. In L. Tiittula & P. Nuolijärvi (Eds.), *Puheen kirjallisen esittämisen alueita, keinoja ja rajoja* (pp. 162–191). Helsinki: Suomalaisen Kirjallisuuden Seura.
- Wehrheim, L.** (2018). Economic history goes digital: topic modeling the *Journal of Economic History*. *Cliometrica*. DOI: <https://doi.org/10.1007/s11698-018-0171-7>
- Weingart, S. B., & Meeks, E.** (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1).
- Wiktorsson, G.** (1996). *Den grundlagsskyddade myten: om allemansrättens lansering i Sverige*. Stockholm: City University Press.

Internationalism in the UK Parliament: Computer-assisted Analysis and Digital Methods

Pasi Ihalainen and Aleksi Sahala

Introduction¹

While the history of the diplomatic events and institutions of 20th-century international politics has been comprehensively explored,² macro-historical and long-term computer-assisted analyses of conceptualisations of the ‘international’ have not yet been attempted. With the increasing availability of digitised parliamentary debates, such an analysis of the everyday language of politics has become possible. In the conceptual history of internationalism, focus on Parliament is particularly pertinent in the British case, as the country has been one of the most active agents in the field of international cooperation while regarding Parliament as *the* forum of ideological debate. Parliament has had a say in foreign policy, too, regarding membership in international organisations.

After having previously analysed parliamentary debates with more conventional close-reading methods of the history of political discourse,³ we turn here to text analysis programmes to explore their benefits for conceptual history.

Our goal is to reconstruct meanings assigned to international issues in the UK Parliament in the long 20th century: How has the ‘international’ been experienced, understood, conceptualised, constructed, debated and redefined? How and why has the ‘international’ been given meaning and implicitly defined through its use in a variety of (ideologically motivated) political arguments, particularly in connection with membership in international organisations? To what extent and when, how, why and with what consequences has this attribute turned into an ‘ism’?

We argue that the distant reading of extensive series of digitised UK parliamentary debates by the means of a collocation analysis helps to extend and deepen conceptual analysis so that previously unnoticed ways to discuss international cooperation can be discovered and the close reading of sources is more effectively focused. We supplement the analysis of parliamentary discourse at macro level with collocation analyses of particular debates and contextualised conceptual analyses in concrete speaking situations. The latter correspond with the criteria of historical research for understanding meaning created in specific contexts and provide checks to premature conclusions drawn on the basis of computer-assisted distant reading. In this exploration, conclusions of the distant reading remain suggestive so that problems rising from decontextualised interpretations can be pointed out. Thus, our investigation provides an example of interaction between text analysis programmes and an analytical mind familiar with the genre and discourses of the primary sources.

While doubts about the application of collocation analyses to intellectual history have by no means been overcome,⁴ in corpus linguistics they have been used productively.⁵ For sociolinguists, collocation is ‘an accepted, linguistically meaningful measurement’⁶ referring to ‘the co-occurrence of two words within a pre-specified span, when the frequency of the co-occurrence is above chance, taking into account the frequencies of the “node” (the word in focus), its collocates, and the collocation itself.’⁷ Applications thus far include a diachronic analysis of UK parliamentary speaking on Ireland,⁸ an analysis of a parliamentary debate on the climate change⁹ and an analysis of adjective collocates qualifying capitalism. Foxlee has aimed at combining more semantically and more pragmatically oriented versions of conceptual history with computer-assisted text analysis.¹⁰ Guldi has demonstrated how word counts and text mining produce indices of historical change.¹¹ Lähdesmäki and Wagenaar have used collocation analysis to explore discourses of diversity within the Council of Europe by grouping the key terms in semantic fields (collocation networks) and measuring the frequencies of those fields in order to reveal how concepts were produced as policy.¹² The current authors share such an understanding of politics as primarily discursive and of the need to focus on conceptual innovations and active uses of language aimed at affecting policies.¹³

The Hansard Corpus (<https://www.english-corpora.org/hansard/>) contains nearly every speech given in the UK Parliament between 1803 and 2005,

1.6 billion words in total, and allows researchers to search on parliamentary debates, including collocation searches. While the scanned records have been proofread and optical character recognition (OCR) is not an issue, something like 5% of the debates have not been included in the Historic Hansard database (<https://api.parliament.uk/historic-hansard/index.html>) from which the data originates, which causes some uncertainty with search results. We started our distant reading with a collocation analysis of the noun collocates of ‘internationalism’ based on the tool of the Hansard Corpus, quoting selected examples from Historic Hansard to give more concrete content to the discernible trends. We then proceeded to collocation analyses of the entire vocabulary of ‘international’ in a selection of Commons (HC) and Lords (HL) debates concerning British membership in international organisations.

For the collocation analysis, we used a measure called PMI^2 ,¹⁴ which is a less low-frequency sensitive improvement of the Church and Hanks’ word association measure built around the idea of Pointwise Mutual Information (PMI).¹⁵ The core idea of the PMI-based measures is to divide the corpus into forward-looking or bi-directional windows of fixed size, which define the maximum distance between the keywords and their possible collocates. The keywords are paired with each word that can be found within the defined window size, and the actual joint probability of each pair is compared to the expected probability of those words co-occurring independently. *The maximum score of 0* indicates that the words are only found together, and the minimum of $-\infty$ that the words never co-occur within the given window size. For calculating the scores, we used a Python script called *pmizer*, which is an open source script for calculating different PMI-based association measures from tokenised text.¹⁶ We did not lemmatise our data, as we wanted to preserve singular and plural forms separately. To avoid our data being overcrowded with conjunctions, prepositions and pronouns, we filtered most of these out by using a simple stop-word list.

For the analysis of membership debates, there was no need to limit search terms as single debates varying from one to a few days were in question, and hence all references to the ‘international’ could be considered. A broad collocation window of 10 words both ways was used to discover every politically significant noun associated with the ‘international’. The scores calculated with PMI^2 are reported below in the form ‘(number of collocates within the span of ten words both ways/score/distance)’.

Politically interesting collocates picked from result lists that ranked the closeness of the collocations on the basis of their score were grouped into collocation networks and their relative importance in both Houses discussed. While collocations are usually considered statistically significant when they appear in the corpus at least twice, individual combinations of words also deserve attention as politically potentially meaningful innovative speech acts. The next analysis is primarily based on distant reading, though some general context is introduced to support interpretation. A close reading of some findings will follow.

‘Internationalism’ in the UK Parliament, 1803–2005

As the Hansard Corpus is so extensive and as the use of ‘international’ is often technical rather than ideological (referring to aviation, for instance), our distant reading focused on collocations of ‘internationalism’. The total number of co-occurrences of ‘internationalism’ in the Hansard Corpus with a nine-word collocation window is 1,542, with an emphasis on the 20th century and especially in the interwar period and the 1970s and 1960s. This leads to a manageable amount of results, even if ones that focus on the ‘extreme’ forms of international thinking that are expressed as an ‘ism’ word.

The noun collocates of ‘internationalism’ that were considered politically meaningful were divided into 13 loose semantic fields (groups of related terms, topical sub-categories or collocation networks), namely nationalism, party, socialism/labour, spirit, peace, democracy, imperialism, cosmopolitanism, globalism, collaboration, institutions, supra-nationalism and capitalism. The diachronic frequencies of the six most important of these semantic fields are visualised in Figure 12.1. The grouping of the terms was intuitive, building on previous empirical analyses of discourses on internationalism. Bringing in the historian’s subjective mind in this way helped in discerning relevant topics among diversified discourses.

The diachronic frequencies point at the centrality of discourses on patriotism and nationalism for conceptualisations of internationalism. As Clavin and

Table 12.1: The number of collocations of ‘internationalism’ in UK parliamentary debates according to the Hansard Corpus.

1870s	4
1880s	8
1890s	0
1900s	8
1910s	38
1920s	132
1930s	188
1940s	202
1950s	129
1960s	99
1970s	317
1980s	204
1990s	130
2000–2005	83

Source: Author.

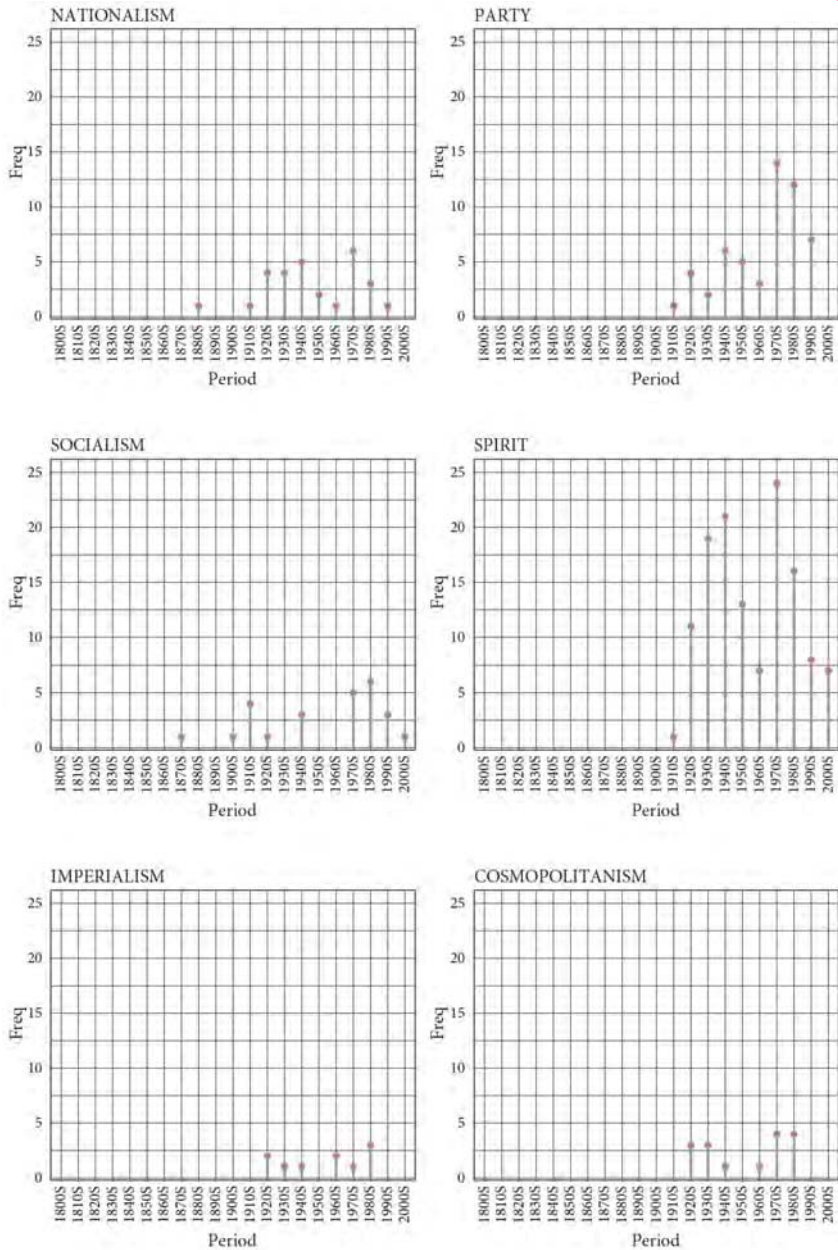


Figure 12.1: Politically significant collocate groups of ‘internationalism’ in the Hansard Corpus: nationalism, party, socialism (including the vocabulary of labour), spirit, imperialism, and cosmopolitanism. The visualisations were made by Kimmo Elo. Source: Author.

Sluga have argued, for much of the 20th century, internationalism was unthinkable without nationalism; nationalism was the basic premise, not a mere counter-concept of internationalism.¹⁷ The vocabulary of patriotism contrasted with socialist internationalism in the British parliament already in 1882 as P. J. Smyth (Home Rule Party) criticised Irish agitation of ‘the substitution of a vague, sickly, and godless internationalism for the manly patriotism’ (*HC Deb. 9 March 1882 vol. 267 c524*; the references from now on are to volume and column numbers (c) in accordance with the conventions of British parliamentary debates). Patriotism or nationalism and internationalism (59 collocations) became increasingly associated after the First World War as the promoters of the League of Nations tried to reconcile these ways of thinking. Yet, J. D. Rees (Conservatives), a former colonial administrator, pointed at tensions between patriotism or nationalism and (socialist) internationalism, concluding that ‘[i]nternationalism means the negation of patriotism and the abnegation of everything of which we should be proud. Instead of extending internationalism I long myself to see it abolished completely off the face of the earth’ (*HC Deb. 1 November 1920 vol. 134 c106*). Internationalism could only be based on nationalism, as Goronwy Owen (Liberals) put it: ‘I have no sympathy at all with the people who preach internationalism as such ... The basis of a proper internationalism is a good nationalism ...’ (*HC Deb. 27 April 1928 vol. 216 c1285*). Morgan Jones (Labour), a pacifist, agreed in the early 1930s as every European state was developing ‘not towards a growing internationalism but towards essential nationalism’ (*HC Deb. 10 May 1932 vol. 265 c1837*). Doubts about internationalism continued after the Second World War, especially among non-socialists: Ralph Rayner (Conservatives) was ironical when pointing out that ‘Russia is still internationalist in so far as internationalism will serve her nationalism’ (*HC Deb. 20 February 1946 vol. 419 c1192*).

Only from the mid-1950s can we find Conservatives conceptualising internationalism in more positive terms. Peter Smithers, a British delegate for the Council of Europe, believed that ‘[a] nationalist war today is a physical impossibility; and economically and socially the temptations of the benefits of internationalism are so great that nationalism in itself is no longer a very attractive proposition’ (*HC Deb. 27 July 1955 vol. 544 c1288*). More radical challenging of nationalism dated from the time of the EEC membership as the young David Owen (Labour) declared: ‘When I talk about European unity, I am talking in part about our concept of nationalism and internationalism. I find that one of the most dangerous facets of modern life and, indeed, of our history over the last 50 years is the scar of nationalism. I believe in internationalism as an article of faith’ (*HC Deb. 26 October 1971 vol. 823 c1634*). By 1981, even the former Conservative MEP Hugh Dykes argued that ‘Britain has always been internationalist in its nature. We changed the orientation of our internationalism by entering the Community in 1973. I wish that new, modern internationalism and Europeanism to continue, for the benefit of future generations’ (*HC Deb. 8 April 1981 vol. 2 c1000*). Such positive associations between integration and

internationalism, though interesting with hindsight, were not mainstream in the 1970s and 1980s either.

Conservative and liberal suspicions about internationalism had traditionally risen from its associations with socialism, as is revealed by 52 collocations of party vocabulary and internationalism, especially in the 1910s and 1920s, but still in the 1970s and 1980s as well. Party perspectives include associations between internationalism and socialism/socialist(s) (20), workers (10), labour (7), revolution (1), Marxism (1), bolshevism (1) and communism/communist(s) (2). Socialist or labour internationalism had emerged in the mid-19th century with the First International,¹⁸ came up in the British parliament in the 1870s and was welcomed by 1911 as Josiah Wedgwood (Liberals) observed how '[i]nternationalism is spreading rapidly, not only the internationalism of capital but the internationalism of labour' (*HC Deb. 14 December 1911 vol. 32 c2625*). The founding of the International Labour Organisation (ILO) in 1919 activated this discourse, although typically the political rivals attacked the internationalist background of the Labour Party. By 1950, Joseph Kenworthy, a Labour peer, nevertheless declared that 'the future of mankind lies in internationalism: That word "internationalism" is a word we do not hear nearly often enough to-day: Real internationalism properly applied would have avoided the terrible catastrophes of the world wars of this century and have raised the standard of life of the whole of humanity' (*HL Deb. 28 June 1950 vol. 167 c1186*). Philip Russell Rea, a Liberal peer, also came up with ideas about 'more internationalism, some relinquishment of national sovereignty' as 'necessary in the modern world' (*HL Deb. 2 November 1960 vol. 226 c52*). Ronald Leighton, too, believed in internationalism, but added in line with the interwar prioritisation of nation states: 'The word "inter" means between: Instead of supra-nationalism, I want to see a group of independent, self-governing countries co-operating together' (*HC Deb. 21 May 1984 vol. 60 cc730-731*).

The spirit of internationalism (the fourth most common collocate of internationalism) had been discussed before the First World War as Philip Snowden (Labour), an anti-capitalist trade unionist, declared that 'we believe in the spread of a spirit of internationalism' and urged Britain to lead 'a great international league of peace' (*HC Deb. 15 March 1910 vol. 15 c308*). The vocabulary of this discourse included collocations with spirit (24), principle(s) (22), idea(s) (17), sense (10), ideal(ism) (10), word(s) (13), belief/believer (12), concept (11) and values (7), as well as thought, thinking, theories, enthusiasm, vocabulary and term. There would seem to have been a slight rise in the 'idea' of internationalism between the 1920s and 1940s and again in the 1970s and 1980s, as in the general intensity of internationalism discourse. The 'spirit' of internationalism peaked from the 1930s to the 1950s, 'principles' peaked in the 1950s. Critique against the 'idealism' of internationalism appeared between the 1920s and the 1950s and again in the 1970s, in the same period as 'values' were discussed. 'Sense' and internationalism were co-textualised a few times from the 1920s to 1940s and again since the 1970s, but with diminishing frequencies.

Explicit references to the spirit of internationalism were not so many as could be expected on the basis of Britain's role as a herald of liberal internationalism. In the aftermath of Hitler's accession to power, James Henderson Stewart (National Liberals), a supporter of Anglo-American cooperation, nevertheless assured that the British government had done more than any other 'to establish the spirit of internationalism' (*HC Deb. 27 April 1933 vol. 277 c366*). After the Second World War, the British spirit of internationalism was emphasised every now and then. In 1956, Lord Rea demanded that 'international questions must be handled with international, and not with national, mentality' and that 'we must recognise this spirit of internationalism much more than we have done in the past' by joining European organisations (*HL Deb. 15 March 1956 vol. 196, cc461–462*). Reginald Prentice (Labour) welcomed development aid as a way of 'building up a spirit of internationalism which can be a factor towards world peace' (*HC Deb. 25 April 1961 vol. 639, c303*).

Internationalism has customarily been associated with peace. This discourse emerged with the League of Nations and peaked in the 1930s. Dennistoun Burney (Conservatives), an aviation expert, summarised the logic of internationalism from the point of view of national, European and imperial security: '... if you are to have peace you must have internationalism, and if you are to have internationalism, you can only have it by abrogating to some extent the sovereign rights of each nation and at the same time restricting the freedom of the elective assembly of each national Government' (*HC Deb. 7 March 1929 vol. 226, c670*). By the 1930s, the problem was, according to Seymour Cocks (Labour), that Germany 'removed pacifism and internationalism from her vocabulary' (*HC Deb. 13 November 1933 vol. 281, c665*) so that, for Ralph Rayner (Conservatives), it already appeared as 'extremely dangerous to teach pacifism, internationalism, and the brotherhood of man' (*HC Deb. 14 June 1937 vol. 325, cc112–113*). Discourse on the spirit of internationalism nevertheless emerged in the interwar era, peaked with the creation of the United Nations and became again rarer in the 1950s and 1960s. After EEC membership, the UK Parliament appeared quantitatively at its most 'internationalist'; thereafter, collocations between spirit and internationalism have declined. Democracy was associated with internationalism in the parliamentary context mainly in the 1930s and 1940s, although Francis Pym (Conservatives), a former foreign secretary who opposed Thatcherism, argued boldly in 1987 that '[t]he world is interdependent, and internationalism must be nurtured in every democracy' (*HC Deb. 7 April 1987 vol. 114, c196*).

International issues could be conceptualised in further alternative ways. The British parliamentary elite had seen the League of Nations as supportive of the interests of the Empire.¹⁹ They often discussed imperialism/imperialist(s) (commonwealth, empire, colonialism) and internationalism in the same context, with a break in the 1950s when decolonisation had started, reflected on the topic occasionally from the 1960s to the 1980s, and then dropped it from their vocabulary. Cosmopolitan ideals of internationalism—consisting of a variety

of notions ranging from brotherhood, altruism, solidarity, humanitarian issues, aid and assistance to friendship, neighbourliness, fellowship and reciprocity—were also defended between the 1920s and 1940s. After the Second World War, this discourse became marginal, only to peak in the 1970s and 1980s in a rising internationalist atmosphere supportive of development and humanitarian aid, and losing popularity from the 1990s onwards. This period saw the emergence of a normative discourse on cosmopolitan democracy in political science, but such theories did not find their way to Parliament. The 1990s did see the emergence of the alternative discourse on globalisation, but only in three collocations with internationalism.

A discourse on collaboration (1) or cooperation (1) as internationalism surfaced by the end of the century, but as the low frequencies show, remained surprisingly marginal. Conventions on human rights, including that of the Council of Europe, appeared to Oliver McGregor, an economic historian, as ‘the most remarkable features of recent history, a triumph for reason, co-operation and internationalism’ over nationalism (*HL Deb. 16 December 1987 vol. 491, c729*). EEC membership supported discourse on systems, organisations, institutions and associations on the one hand and internationalism on the other. Yet, such debate withered away by the early 2000s, which may be reflective of the lack of commitment to the institutions of the community. The membership gave rise to entirely new debates on the relationship between integration (3), union (6), community/ies (11), Europeanism (3), supra-nationalism (4) and internationalism as well, but not to any great extent. In the meantime, the membership does not seem to have made such a great difference in associations between internationalism and markets (including trade(s)/trading, capital/capitalism/capitalist(s), economy, finance, growth and competition), a discourse that had existed before the First World War and been on a higher level in the interwar era. The global free trade visions of British politicians do not seem to have changed much with the post-Second World War economic integration: the EEC, too, was mainly conceptualised as a question of markets.

Debates on Membership in International Organisations

Next, we shall complement the above distant reading of trends in discourse on internationalism with analyses of collocates of the ‘international’ in entire parliamentary debates that concerned the British membership in international organisations—as key moments of discourse on the ‘international’. The collocation analyses enabled a type of ‘topic modelling’ of the contents of the debates so that the results were not determined by previously selected search terms such as ‘internationalism’ only.

The selected membership debates concerned the League of Nations (LoN, in the Commons on 21 July 1919, in the Lords on 24 July 1919), the United Nations (UN, in both Houses on 22 August 1945), the Council of Europe (CoE,

debated only in the Commons on 13 November 1950), the European Economic Community (EEC, several days in February 1972 in the Commons and in July 1972 in the Lords) and the European Union (EU, a couple of plenaries in September 2017 in the Commons and in January 2018 in the Lords). A further possibility might have been membership in NATO, but the defence alliance differed in its character from the other more general forms of international cooperation. It is also debatable whether these are the most representative occasions and whether the EU should be seen as a mere 'international organisation' or rather as a project of transnational integration. Debates on the EEC/EU in particular had several stages, and in principle all of these could have been analysed, but for the sake of consistency only *the second readings* of the related bills were considered. The second reading is typically the stage of deliberative decision-making when most extensive ideological contributions to the debate are made and competing arguments presented, reflecting much of what had come up in other parliamentary discussions and the public debate.

The Commons debates on the LoN membership on 21 July 1919 was a key moment in the history of British internationalism.²⁰ It took place in the aftermath of the signing of the treaty of Versailles that not only concluded the First World War with tough peace terms on Germany, but also introduced the League Covenant. The collocation analysis suggests that internationalism surrounding the League was conceptualised by the MPs to a great extent through the general concept of 'international law', as could be expected on the basis of the British role in drafting the Covenant and the inclusion of the International Court in it. The 'international' in the context of the League was about 'court' (4/-1184/2.75), 'legislation' (2/-1238/2.0), 'regulation' (1/-1238/1.0), 'justice' (4/-1271/2.75) and 'law' (1/-1571/1.0). Yet, for a trading nation, the 'international' also stood for finances, as reflected by nine close collocations of 'international' and 'finance' (9/-1029/1.0) and two more with 'financiers' (1/-1338/1.0) and 'financial' (1/-1655/1.0). 'Labour' had numerous close collocates with 'international' (10/-1220/2.2) due to the connected founding of the ILO, aimed at appeasing revisionist Western socialists under the alternative of the Communist International.²¹ As an entirely new international organisation was being constructed, its institutions were discussed with terms such as 'bureau' (1/-1238/2.0) and 'machinery' (3/-1312/1.0). Discourses on 'experiment' (2/-1355/5.0), 'opportunity' (3/-1367/5.3) or 'cooperation' (1/-1397/3.0) and 'international' surfaced, but only rarely. Out of these findings, 'opportunity' will be analysed in more detail below.

The Lords paid plenty of attention to the moral aspects of the League, associating 'morality' (6/-800/1.0) tightly with 'international' and connecting 'morals' (1/-1059/3.0) and 'sanctity' (1/-1059/4.0) with it as well. The League was about 'treaties' (2/-1059/6.0), 'jurisprudence' (1/-1159/1.0), 'court' (2/-1178/1.0), 'sanction' (1/-1217/1.0), 'justice' (2/-1259/1.0), 'code' (1/-1259/1.0), 'rules' (1/-1259/6.0), 'law' (1/-1391/1.0) and 'treaty' (1/-1611/5.0), the total number of legal collocates rising to 13. Reflective of the more value- than interest-directed

discourse (different from the down-to-earth approach of the Commons) is talk about ‘international spirit’ (2/-1117/1.5), which was reinforced by synonymous collocates such as ‘friendship’ (1/-1059/7.0) and ‘reconciliation’ (1/-1159/4.0). A major difference was the lack of debate on ‘finances’ and ‘labour’, which shows how economic and social issues were left for the lower house to deal with, appearing as less relevant in the social context of the peers.

The League was generally regarded as a drastic failure after the Second World War, and the British government then agreed with the United States and the Soviet Union on the founding of a new international organisation.²² The Commons debated the UN membership in August 1945, in the aftermath of a victory over Nazi Germany and in the shadow of the first military use of the atomic bomb in Japan. Despite dissatisfaction with the League, British understandings of the ‘international’ had not changed much since 1919: the UN was likewise conceptualised through law, labour and institution. Next-door collocations of ‘international’ and ‘court’ (5/-1076/1.0), ‘justice’ (6/-1097/2.7), ‘law’ (4/-1228/1.0) and ‘lawyers’ (1/-1382/1.0) dominated, and collocations with ‘conventions’ (2/-1182/2.0) can be added to this discourse on international law. Associations with ‘labour’ (9/-1080/1.7) and ‘workers’ (1/-1582/7.0) continued to feature, which shows not only that the interests of the working class were central to the current Labour government, but also highlights a reaction to the strengthened international status of the Soviet Union and the connected need to appease the working classes of the West. Discourse on the institution focused in 1945 distinctly on ‘control’ (10/-1126/2.6), ‘security’ (5/-1476/3.2), ‘machinery’ (2/-1490/1.0) and ‘peace’ (4/-1544/1.0). ‘Economic’ (3/-1500/5.0) had rather loose connections with ‘international’ in comparison with the post-First World War situation. Collocates deserving further exploration include ‘collective’ and ‘security’.

The Lords viewed the UN much like the Commons, emphasising law and justice on the one hand and the functioning of the institution aimed at collective security on the other. The peers associated ‘international’ and ‘justice’ (3/-1156/1.0), ‘treaties’ (1/-1292/5.0) or ‘law’ (1/-1509/1.0) and produced collocations of ‘international’ with ‘supervision’ (2/-1192/-1292/1.0), ‘guards’ (1/-1192/3.0), ‘operation’ (2/-1292/3.0), ‘security’ (5/-1341/3.6), ‘peace’ (4/-1386/1.0) and ‘machinery’ (1/-1573/1.0). Noteworthy are close associations between ‘international’ and ‘collaboration’ (2/-1151/1.0). Comments on ‘patriotism’ (1/-1292/9.0) and the ‘commonwealth’ (1/-1473/3.0) were also made.

Also at the formation of the CoE in 1950, the Commons drew predominantly from conceptualisations of the international law. ‘Tribunal’ (1/-1170/1.0) and ‘court’ (4/-1202/1.0) had several close collocations with ‘international’, and collocations with ‘justice’ (3/-1085/2.3) and ‘jurists’ (1/-1170/9.0) appeared. Associations between ‘labour’ (4/-1228/3.25) or ‘workers’ (1/-1370/2.0) and ‘international’ remained part of the discourse. Loose associations between ‘international’ and ‘continental’ (1/-1540/4.0), ‘Brussels’ (1/-1540/7.0) and ‘federal’ (1/-1634/4.0) were emerging, which may be indicative of a tendency to

locate the ‘international’ out there in Europe. Economy or trade (1/-1578/9.0) had a marginal role in the CoE debates, nor do democracy or human rights feature,²³ which is surprising given the later role of the organisation, and associations between ‘peace’ (1/-1609/6.0) and ‘international’ were weak as well. The CoE appeared as a further body applying international law. Yet, it might have powerful tools of ‘international pressure’ (2/-1070/1.0, the strongest discovered association) in its possession, to which we shall return below.

The EEC, by contrast, was conceptualised in 1972 much less through law than economy. For the Commons, the EEC was about markets, trade, business and companies—*not* about law—which suggests a lack of dedication to common legislation as an aspect of the European integration. Associations between ‘international’ and ‘companies’ were exceptionally strong (7/-1022/1.0), but the ‘international’ was also associated with ‘monetary’ issues (6/-1171/2.0), ‘trade’ (5/-1345/2.6), ‘fund’ (3/-1362/2.0), ‘trading’ (2/-1442/1.0), ‘firm’ (1/-1515/9.0), ‘business’ (1/-1553/1.0) and ‘growth’ (1/-1659/5.0). An association between ‘international’ and ‘continent’ (1/-1529/4.0) can be found, but interesting is the remaining considerable semantic distance between ‘international’ and ‘Europe’ (2/-1781/6.5) or ‘European’ (1/-1971/5.0), ‘community’ (2/-1903/5.0) or even ‘British’ (1/-1933/3.0). The integration was not that much about partnership, with ‘partners’ only passingly associated with ‘international’ (1/-1383/9.0). Some associations with ‘scientific’ (2/-1215/6.0) aspects of integration appeared, while the defence aspect of the EEC was mentioned only in passing (1/-1622/2.0). At first sight, some concern on being located in an ‘international periphery’ (1/-1283/1.0) would seem to have been expressed, but close reading will lead to opposite conclusions.

The Lords talked much less about law than in connection with previous memberships. While ‘international law’ (1/-1732/1.0) was mentioned, ‘rights’ (1/-1505/4.0) and ‘rules’ (1/-1543/8.0) had few and relatively weak associations with ‘international’. The economic aspect was less distinct than in the Commons—associations with ‘companies’ (1/-1432/7.0), ‘capital’ (1/-1443/2.0), ‘monetary’ (1/-1512/1.0) and ‘trade’ (1/-1704/1.0) appearing, but those with ‘economic’ (1/-1727/5.0) and ‘market’ (1/-1799/9.0) being much weaker. Associations with ‘mobile’ (1/-1073/1.0), ‘cohesion’ (1/-1173/3.0) and ‘standards’ (1/-1454/1.0) reflect an understanding of the ideas of the economic community, although a need for ‘protecting’ (1/-1073/2.0) would suggest the opposite view. ‘International partnership’ (1/-1419/1.0) as an expression made an appearance. The semantic distance between both ‘British’ (1/-1779/6.0) or ‘community’ (1/-1926/2.0) and ‘international’ is observable, just as in the Commons. Worth close reading is an occasional association between ‘European’ and ‘internationalist’ (1/-1638/3.0), for instance.

Debates since the Brexit referendum of June 2016 have been complex and are far from completed at the time of writing (August 2018/2019), which means that the following remarks necessarily remain provisional. The early Commons debates reflected a considerable concern about ‘development’ (4/-990/2.25) in

the field of international questions and addressed international ‘obligations’ (4/-1042/1.5) to an exceptional degree. The latter way of speaking can be seen as an aspect of the traditional legal discourse on ‘treaties’ (3/-1209/1.0) and law (3/-1657/1.0), which seems stronger at the time of a prospective exit from the EU than during entrance negotiations. Yet, the Commons continued to understand the EU overwhelmingly through ‘trade’ (9/-1144/1.7, though this result is overemphasised by the title ‘International Trade Secretary’), and the prospective withdrawal gave rise to questions about ‘tax’ (1/-1368/1.0), ‘taxation’ (1/-1442/4.0) and ‘customs’ (1/-1690/8.0). ‘United Kingdom’ (2/-1518/5.5), ‘nation’ (1/-1534/5.0) and ‘UK’ (2/-1712/4.5) were viewed as only slightly more ‘international’ than in the 1970s, and ‘union’ (2/-1750/8.0) and ‘European’ (2/-1988/5.5) continued to be dissociated from ‘international’. The British MP refrained from conceptualising the EU which they were about to leave as ‘international’.

The Lords, known for their more pro-integration stands, were likewise dedicated to ‘international obligations’ (3/-1243/1.0) and ‘international treaty/treaties’ (4/-1371/-1482/1.0), but also to the ‘international reputation’ of Britain (2/-1371/1.0). Like the Commons, they felt some concern about international ‘development’ (2/-1440/3.0). Discourses on international addressed economic issues with ‘monetary’ (1/-1301/1.0), ‘trade’ (6/-1432/2.3), ‘fund’ (1/-1460/2.0), ‘manufacturing’ (1/-1571/1.0), ‘bank’ (1/-1591/9.0) and ‘financial’ (1/-1794/1.0), but this was by no means the dominant discourse. Legal discourse played a more diversified role than at the time of joining the EEC, seen in association with ‘divorces’ (2/-1101/1.0), ‘crime(s)’ (2/-1360/-1518/1.0-4.0), ‘law’ (7/-1516/3.6), ‘rules’ (2/-1582/6.5), ‘court’ (2/-1644/1.0), ‘standards’ (1/-1718/1.0), ‘regulatory’ (1/-1760/1.0), ‘justice’ (1/-1763/3.0) and ‘rights’ (1/-2060/2.0). Noteworthy is the use of the metaphor ‘divorce’ to describe Brexit, with an emotional connotation side by side with concrete legal discourse. An exceptional intervention addressing ‘internationalist heritage’ (1/-833/1.0) has been chosen for closer reflection below. ‘National’ (2/-1597/2.0), ‘nation’ (1/-1763/6.0) or ‘UK’ (2/-1877/8.0) had not become any more ‘international’ than in the membership debates, and there was really not anything ‘European’ that would appear as ‘international’ either (2/-1916/5.5). In the context of Russian interventions in Western elections in general and the British referendum in particular, ‘international’ found an association in ‘cyberattacks’ (1/-1201/3.0) as well. Otherwise, the British debates on Brexit show considerable trajectories in the prioritisation of economy and the rise of legal discourse as a consequence of the membership.

Individual Speech Acts Surrounding the ‘International’

The third and final step of our analysis proceeded as close and contextualising reading of some discovered collocations. Potentially interesting collocates were

pointed at in the above debate analysis. They were now located in their textual context in the House of Commons Parliamentary Papers database (Hansard 1803–2005 and, for the EU case, Hansard Online of the UK Parliament). This phase allowed some checking of the functioning of the collocation analysis programme and our preliminary conclusions. The quotations were analysed as individual speeches in which politicians defined the ‘international’ by the active use of language in political action in particular contexts. These could only be reconstructed on an exemplary basis in the confines of this report. While general trends of thought on internationalism (such as the centrality of law and economy) become obvious on the basis the macro-level collocation analysis and need no extensive discussion here, some peculiar points deserve attention as they demonstrate the importance of context in determining what exactly was done politically in Parliament, also revealing shortcomings in mere distant reading.

As the Commons debated the League Covenant, two Labour MPs came up with ‘opportunities’ opened up by it: for international relations in the spirit of the optimistic expectations of British internationalists, and for social reform central for the Labour Party. J. R. Clynes, a leading trade unionist and the deputy chairman of the Labour Party, welcomed the League ‘[a]s an instrument for providing, through the medium of International Courts and international action, an opportunity for considering differences as they arise’ (*HC Deb. 21 July 1919 vol. 118, c961*). George Barnes, a former Labour leader who represented in 1919 the pro-coalition National Democratic and Labour Party and had been one of the British negotiators in Paris, encouraged social reform in the spirit of labour internationalism by pointing that ‘for the first time Governments have put a chapter of Labour into an international Treaty [ILO] and made labour conditions a matter of international agreement’, which constituted an ‘opportunity’ for workers worldwide (*HC Deb. 21 July 1919 vol. 118, c976*). These quotes exemplify the high leftist expectations for international cooperation during the post-First World War reconstruction.

Liberal and Conservative internationalists in the Lords were, in the name of the government, also predominantly optimistic. James Bryce, a respected constitutional lawyer, member of the International Court at The Hague and Liberal politician, viewed the League as based on ‘the feeling that the world has now become one one in a new sense never dreamed of before’ surrounding ‘the belief that the community of the world requires that a new spirit should prevail in international relations—a spirit which seeks to substitute friendship for enmity’ (*HL Deb. 24 July 1919 vol. 118, c1019*). George Curzon (Conservatives), a major imperialist (as former Viceroy of India) and acting Foreign Secretary, echoed this belief in rising internationalism, stating, ‘the international spirit, the kind of idea that the future unit is not to be the race, the community, the small group, but is to be the great world of mankind, and that in that area you try and induce a common feeling, you try and produce co-operation which will be a better solvent of international difficulties ...’ (*HL Deb. 24 July 1919 vol. 35,*

c1029). The League was basically welcomed by all political groups in the British parliament, even if questioning its effectivity was also widespread.

Doubts about weaknesses in organisation were equally present as the UN was formed in 1945. Prime Minister Clement Attlee, the leader of a Labour majority government, then forcefully advocated the concept of 'collective security' as the foundation of the UN Security Council 'where the policies of the States ... could be discussed and reconsidered ... especially when they showed signs of such divergences as to threaten the harmony of international relations. Collective security ... is active co-operation to prevent emergencies occurring' (*HC Deb. 22 August 1945 vol. 413, c665*). Captain David Gammans (Conservatives), a former diplomat, was one of several MPs to question the definition of the UN as a provider of 'collective security', suggesting that such a concept should not be used at all (*HC Deb. 22 August 1945 vol. 413, c734*). The required unanimity of the permanent members, after all, would constitute a major limitation to the functionality of the world organisation. In the Lords, Robert Cecil, the key planner of the League, a leader of the League of Nations Union and a Nobel Peace Prize winner (1937), consistently assured that 'every attempt ought to be made, and must be made, to secure peace by international collaboration' (*HL Deb. 22 August 1945 vol. 413, c133*), bridging two major projects of British internationalism.

The debate on the CoE in 1950 provides a good reminder of the need for close reading and contextualisation. While a swift reading of the collocation results might suggest that 'international pressure' by the CoE was welcome, a closer analysis shows that the contrary was the case. Major Harry Legge-Bourke (Conservatives) was opposing restrictions to national or parliamentary sovereignty when arguing: 'I am in favour of the Council of Europe, but I am in favour of it only on one set of terms, and that is that it remains as a council and does not become an international pressure group. There seems to be very real danger of it becoming an international pressure group.' What particularly worried Legge-Bourke was a 'desire for the institution of a European Political Authority' (*HC Deb. 13 November 1950 vol. 480, c1479*).

Caution with far-reaching conclusions based on collocations is needed also in the case of the EEC membership. The MP who referred to 'international periphery' did not imply that Britain would become somehow peripheral outside the Community, but was concerned about the potential loss of sovereignty. Ronald King Murray (Labour), a leading Scottish lawyer, reacted to a suggestion that membership in the UN and NATO already implied a loss of national sovereignty and relativised the radicality of an EEC membership by pointing out that:

... it was sovereignty in a peripheral sphere, the international periphery of our being which did not involve the heart of our domestic constitutional being as the Bill unquestionably does. We are surrendering a portion of the inner core of our sovereignty because we are dealing

with two aspects of the constitution, first with an economic aspect and secondly with one which is more properly constitutional. (*HC Deb.* 15 February 1972 vol. 831, c363)

While general international or defence cooperation did not challenge national sovereignty, the economic and political aspects of the EEC did—a conclusion that has dominated much of the British press discourse ever since. Arguments in favour of membership won in 1972, although hardly in the extreme form presented by Frank Beswick (Labour), a former voluntary in the Spanish Civil War and a current party whip: ‘My own approach to this Bill is that of an internationalist. I have always been ready to surrender sovereignty in those areas where individual and national dignity and wellbeing are not impaired’ (*HL Deb.* 26 July 1972 vol. 333, c1368). Confrontations were even tougher in the Brexit debates. The supporters of ‘leave’, emphasising British national identity as distinct from the Continent, considered the vocabulary of the ‘international’ useless. ‘International’ was an attribute of the ‘remain’ side and often of backbenchers with limited political influence. An exceptional association between ‘international’ and ‘European’ was made by Liz Saville Roberts (Plaid Cymru), who emphasised the significance of the EU for British foreign relations as a whole: ‘Beyond the single market and customs union, there are upward of 40 pan-European agencies that form the basis of our international relations across a range of policy areas’ (*HC Deb.* 7 September 2017, c422). A suggestion that the government was acting against national values was heard in Helen Hayes’ (Labour) declaration that Brexit impacted negatively ‘our British values of tolerance, diversity and internationalism’ (*HC Deb.* 11 September 2017, c574). Roger Liddle, a former researcher, adviser of Prime Minister Tony Blair and of the President of the European Commission and a think-tank chairperson, contributed to an intra-Labour dispute on Brexit. He appealed both to the trajectory of British internationalism and to the tradition of labour internationalism to persuade his party fellows to oppose Brexit: ‘Europe is in a category of its own in terms of its impact on future generations ... I want our party to lead, to seize this opportunity to demonstrate that, in contrast to this wretched Government, we can live up to our national responsibilities and our internationalist heritage’ (*HL Deb.* 31 January 2018, c1534). Not only the Brexiteers but also their opponents were on the move, fighting on definitions of internationalism also with history-political arguments.

Results and the Added Value of Digital Methods for Conceptual History

This computer-assisted analysis combining the collocation tool of the Hansard Corpus, the collocation analysis of membership debates and a contextual analysis of instances of political speaking has provided us with an overview

of the evolving discourse on internationalism in the UK Parliament, while also revealing innovative speech acts of potential political significance. Both expected and more surprising general trends were demonstrated, specificities of associations of the 'international' in the context of decisions on membership in international organisations pointed out and some peculiar arguments by individual MPs reconstructed.

The concrete findings include the dominance of discourses on international law in the founding of international organisations other than the EU. Parliamentary discourse, rather than contributing much to the creation of its first major international institution, turned more internationalist as a consequence of the founding of the League of Nations. The first, rather weak, wave of British internationalism lasted from 1919 to the founding of the UN in 1945. Internationalism remained relatively weak during the Cold War, with the exception of a few internationally oriented politicians, advocates of the EEC membership and expressers of global solidarity in the 1970s, when the second wave of British internationalism peaked. Once the EEC membership had become a reality, the parliamentary elite lost its enthusiasm about internationalism, especially with reference to European cooperation, and anti-European rhetoric rose during Thatcher's governments. Economic debate dominated and the legal discourse was set aside as Britain joined the EEC, only to be restored with Brexit when especially the trajectory of discourse on national sovereignty versus internationalism resurfaced. Early 20th-century discourses on labour internationalism have mostly withered away with the rise of non-socialist internationalism, with some revival in the 1970s and 1980s and during the Brexit crisis. Several factors indicate that internationalism was in decline well before Brexit, starting in the 1980s, and that trends in public discourse had implications for followed policies. Our conclusions correspond with Glenda Sluga's suggestion that the 'global seventies' of new international society and international public sphere were followed by the 'post-international' 1990s,²⁴ but the British turn to post-international discourse clearly deserves more attention.

Computer-assisted collocation analyses can contribute to conceptual history in at least two ways:²⁵ First, the analysis produces quantitative data on associations between political concepts that enable us to estimate trends in political attitudes the reconstruction of which with traditional methods would not be possible. Second, distant reading reveals original political points that would have gone unnoticed in close reading or in full-text keyword searches. Such arguments must be subjected to close reading and contextual analysis so that premature conclusions based on distant reading can be corrected. Revealed peculiarities in argumentation frequently turn out to originate from leading politicians attempting to influence the course of policy, which warns against low frequency thresholds.²⁶ All in all, the collocation analysis of discourse on internationalism in Parliament works well, enabling a more efficient locating of meaningful speech acts, although their meanings can only be properly understood with close reading and appropriate contextualisation.

The collocation analysis could be extended on many levels: In distant reading, it would be important to proceed beyond the ‘extreme’ concept of ‘internationalism’, with its inherited associations of socialism and pacifism, and to include concepts such as the world, humanity, universal, global, Empire, Commonwealth, cosmopolitan, supranational, multinational, transnational, Europeanism and trans-Atlantic, for instance. The global (imperial) dimension has remained central for Britons also in the days of European integration, and the joint natural language with the United States has supported discourses of isolationism against internationalism. At the level of membership debates, synonyms and counter-concepts of the ‘international’ (foreign, abroad, domestic, national, etc.) could be considered. At the level of individual speech acts, the dynamics between political parties in arguing about internationalism would deserve more attention. A more extensive reconstruction of the temporal contexts of the arguments on the basis of digitised newspapers and other forums of political debate would be helpful then. Furthermore, now that the sub-themes of the discourse on internationalism in the UK Parliament have been identified, the analysis of the dynamics between nationalisms and internationalisms in other parliaments and transnational interconnections between these debates could be explored.

Notes

- ¹ Pasi Ihalainen was alone responsible for the planning of the research setting, analysis and the written report. Digital analysis method specialist Aleksu Sahala created the program script that enabled the distant reading of particular membership debates. Kimmo Elo produced the visualisations of the results of digital distant reading.
- ² Yearwood 2009; Laqua 2011; Mazower 2012; McCarthy 2012; Sluga 2013; Clavin & Sluga 2017.
- ³ Ihalainen & Palonen 2009; Ihalainen & Matikainen 2016; Ihalainen 2017; Holmila & Ihalainen 2018; Ihalainen 2018.
- ⁴ Edelstein 2016.
- ⁵ Baker et al. 2008.
- ⁶ Baker, Brezina & McEnery 2017: 105.
- ⁷ Gabrielatos & Baker 2008: 11.
- ⁸ Baker, Brezina & McEnery 2017.
- ⁹ Willis 2017.
- ¹⁰ Foxlee 2018: 77, 80.
- ¹¹ Guldi 2019.
- ¹² Lähdesmäki & Wagenaar 2015: 16.
- ¹³ Ihalainen 2006; Halonen, Ihalainen & Saarinen 2015; Ihalainen & Saarinen 2015; Steinmetz & Freedman 2017; Ihalainen & Saarinen 2019.
- ¹⁴ Daille 1994.

- ¹⁵ Church & Hanks 1990.
- ¹⁶ The technical description of the analysis program has been written by Aleksii Sahala. See <https://github.com/asahala/Collocations>.
- ¹⁷ Clavin & Sluga 2017: 5–6; Sluga 2013, 3, 5.
- ¹⁸ Sluga 2013: 4.
- ¹⁹ Holmila & Ihalainen 2018.
- ²⁰ Ibid.
- ²¹ Ibid.
- ²² Ibid.
- ²³ Cf. Häkkinen 2018.
- ²⁴ Sluga 2013: 6–7, 9.
- ²⁵ Cf. Steinmetz & Freedman 2017: 32, who are uncertain as to how to interpret semantic data rising from digital humanities.
- ²⁶ See also Kim 2014: 233.

References

- Baker, H., Brezina V., & McEnery T.** (2017). Ireland in British parliamentary debates: plotting changes in discourse in a large volume of time-series corpus data. In T. Säily, A. Nurmi, M. Palander-Collin & A. Auer (Eds.), *Exploring future paths for historical sociolinguistics* (pp. 83–107). Amsterdam: John Benjamins.
- Baker, P., Gabrielatos, C., Khosravini, M., Krzyzanowski M., McEnery, T., & Wodak, R.** (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
- Church, K. W., & Hanks, P.** (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.
- Clavin, P., & Sluga, G.** (2017). *Internationalisms: a twentieth century history*. Cambridge: Cambridge University Press.
- Daille, B.** (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques* (PhD dissertation, Université Paris 7).
- Edelstein, D.** (2016). Intellectual history and digital humanities. *Modern Intellectual History*, 13(1), 237–246.
- Foxlee, N.** (2018). Pivots and levers: political rhetoric around capitalism in Britain from the 1970s to the present. *Contributions to the History of Concepts*, 13(1), 75–99.
- Gabrielatos, C., & Baker, P.** (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005. *Journal of English Linguistics*, 36(1), 5–38.
- Guldi, J.** (2019). The measures of modernity: word counts, text mining, and the promise and limits of present tools as indices of historical change.

- International Journal of History, Culture and Modernity*, 7. DOI: <https://doi.org/10.18352/hcm.589>
- Halonon, M., Ihalainen, P., & Saarinen, T.** (2015). Diverse discourses in time and space: historical, discourse analytical and ethnographic approaches to multi-sited language policy discourse. In M. Halonon, P. Ihalainen & T. Saarinen (Eds.), *Language policies in Finland and Sweden: interdisciplinary and multi-sited comparisons* (pp. 3–26). Bristol: Multilingual Matters.
- Holmila, A., & Ihalainen, P.** (2018). Nationalism and internationalism reconciled: British concepts for a new world order during and after the World Wars. *Contributions to Conceptual History*, 13(2), 25–53.
- Häkkinen, T.** (2018). British parliamentary attitudes towards a supranational parliament and the Consultative Assembly of the Council of Europe, 1948–1949. *Parliaments, Estates & Representation*, 38(1), 63–75.
- Ihalainen, P.** (2006). Between historical semantics and pragmatics: reconstructing past political thought through conceptual history. *Journal of Historical Pragmatics*, 7(1), 115–143.
- Ihalainen, P.** (2017). *The springs of democracy: national and transnational debates on constitutional reform in the British, German, Swedish and Finnish parliaments, 1917–1919*. Helsinki: Finnish Literature Society. DOI: <http://dx.doi.org/10.21435/sfh.24>
- Ihalainen, P.** (2018). Internationalization and democratization interconnected: the Swedish and Finnish parliaments debating membership in the League of Nations. *Parliaments, Estates and Representation*, 39(1), 11–31. DOI: <https://doi.org/10.1080/02606755.2018.1483998>
- Ihalainen, P., & Matikainen, S.** (2016). The British parliament and foreign policy in the twentieth century: towards increasing parliamentarisation? *Parliamentary History*, 35(1), 1–14.
- Ihalainen, P., & Palonen, K.** (2009). Parliamentary sources in the comparative study of conceptual history: methodological aspects and illustrations of a research proposal. *Parliaments, Estates & Representation*, 29(1), 17–34.
- Ihalainen, P., & Saarinen, T.** (2015). Constructing ‘language’ in language policy discourse: Finnish and Swedish legislative processes in the 2000s. In M. Halonon, P. Ihalainen & T. Saarinen (Eds.), *Language policies in Finland and Sweden: interdisciplinary and multi-sited comparisons* (pp. 29–56). Bristol: Multilingual Matters.
- Ihalainen, P., & Saarinen, T.** (2019). Integrating a nexus: the history of political discourse and language policy research. *Rethinking History: The Journal of Theory and Practice*, 23(4), 500–519. DOI: <https://doi.org/10.1080/13642529.2019.1638587>
- Kim, K. H.** (2014). Examining US news media discourses about North Korea: a corpus-based critical discourse analysis. *Discourse & Society*, 25(2), 221–244. DOI: <https://doi.org/10.1177/0957926513516043>
- Laqua, D.** (Ed.) (2011). *Internationalism reconfigured: transnational ideas and movements between the World Wars*. London: I. B. Tauris.

- Lähdesmäki, T., & Wagener, A.** (2015). Discourses on governing diversity in Europe: critical analysis of the White Paper on Intercultural Dialogue. *International Journal of Intercultural Relations*, 44(January), 13–28. DOI: <https://doi.org/10.1016/j.ijintrel.2014.11.002>
- Mazower, M.** (2012). *Governing the world: the history of an idea, 1815 to the present*. New York, NY: Penguin.
- McCarthy, H.** (2012). *The British people and the League of Nations: democracy, citizenship and internationalism, c.1918–45*. Manchester: Manchester University Press.
- Sluga, G.** (2013). *Internationalism in the age of nationalism*. Philadelphia, PA: University of Pennsylvania Press.
- Steinmetz, W., & Freedon, M.** (2017). Introduction: conceptual history: challenges, conundrums, complexities. In W. Steinmetz, M. Freedon & J. Fernández-Sebastián (Eds.), *Conceptual history in the European space* (pp. 1–46). Oxford and New York, NY: Berghahn.
- Willis, R.** (2017). Taming the climate? Corpus analysis of politicians' speech on climate change. *Environmental Politics*, 26(2), 212–231.
- Yearwood, P. J.** (2009). *Guarantee of peace: the League of Nations in British policy 1914–1925*. Oxford: Oxford University Press.

Metadata of an Image Database: Use of Digital Methods to Map East German Opposition Movement

Melanie Conroy and Kimmo Elo

Introduction: Networks of the East German Opposition

This chapter shows how network graphs and analysis can be used to shed light on the structure and dynamics of the geospatial social networks of segments of the East German opposition movement between 1975 and 1990. What new knowledge can we uncover about a well-studied historical phenomenon if we combine the use of non-traditional source material, in this case metadata from an image database catalogue, with a non-traditional historical methodology, namely, social network analysis? This chapter studies the network of East German dissidents as reflected in the photographic database on the East German Opposition, which archives photos from the 1970s until the fall of the Berlin Wall.

In this chapter, we examine graphs of East German dissident networks, as well as sub-networks filtered by date and by place. We then discuss the general

principles of social network analysis and historical network analysis. Finally, we consider how network analysis could be further improved for historical purposes. The primary tool we use is Palladio, a suite of tools for the visualisation of historical networks developed in the Humanities + Design Lab at Stanford University. Unlike many tools for creating network graphs, Palladio was designed for humanists to visualise data without the need for a designer.¹ Palladio can be used to visualise historical datasets and discover patterns in the data that researchers may choose to analyse using network analysis or other means, including examining the original historical sources.

An Unconventional Source Material: Metadata of an Image Database

The photographic database maintained by the Robert Havemann Society in Berlin as part of its archive on the East German Opposition consists of approximately 60,000 digitised photos with a relatively rich metadata providing information, including the date the photo was taken, the photographer, a descriptive title, keywords, regional/geographical tags and information about the persons to whom the photo is related.² The sample used in this chapter consists of photos featuring selected prominent figures of the East German dissident scene with a connection to the city of Jena. These individuals included academics, artists and intellectuals of diverse socio-economic backgrounds, who were part of a range of movements, from youth movements to peace activism to environmental movements, during the period from 1975 to 1990. In most cases, the photos were taken to document opposition action and activities and used as illustrations in underground magazines, bulletins and leaflets. We should, however, keep in mind that since the German Democratic Republic (GDR) was a dictatorship, photographing these kinds of illegal actions was closely bound with the risk of becoming subject to counter-measures by the security authorities. From this perspective, the photos also document the courage of the people involved in oppositional activities.

In the history of the East German opposition, Jena and Berlin were the two most important regions when it came to the structure, means, motives and dynamics of the opposition groups in the GDR. In Jena, the discrepancy between democracy and dictatorship often led to open conflicts, making this city the primary region of political opposition in the GDR. Jena was also called the secret capital of the GDR opposition, reflecting the complex domestic conflict between the state apparatus, church and opposition in the GDR.³ The temporal focus of this chapter is the period between 1975 and 1990, a period heavily shaping the range of political action for the opposition and resistance groups. One key event was the Conference on Security and Cooperation in Europe (CSCE), held in Helsinki in August 1975, which caused the East German political leadership to become increasingly concerned with

the destabilising impact of the CSCE.⁴ As a consequence, the *Sozialistische Einheitspartei Deutschlands* (SED), the communist monopoly party of the GDR, took numerous repressive actions against dissidents and opposition groups, seeking to scatter the resistance and opposition by eliminating their leading personalities.⁵ To what extent can we see the repressive actions of the SED in the networked relations of Jena-linked figures of the opposition?

One of the main components of the Jena-based opposition movement was the Jena Peace Community (*Jenaer Friedensgemeinschaft*), established in March 1983 as one of the first major opposition communities outside the protective walls of the Evangelical church. The local Peace Community was a dissident platform of short duration, but of long-lasting impact. Its founders were disillusioned with the reluctant resistance of the Evangelical church against state repression and, hence, sought to establish a new, independent platform under the umbrella of the European Peace Movement. The community itself was short-lived, because by the spring of 1983, the security authorities had already decided to destroy the Jena Peace Community once and for all. But, because the security police did not achieve its main goal (a complete destruction of Jena's opposition), the community had a long-lasting impact, causing Jena to remain an unsettled city and one of the most important places for political opposition between 1983 and 1989.⁶ Members included Uwe Behr, Manfred Hildebrandt, Mario Dietsch, Edgar Hillmann, Michael Rost and Frank Rub, and non-church-members Roland Jahn and Petra Falkenberg. As we shall see, many of the members and allies of the Jena Peace Community remained active in the opposition movement in Jena for years after the crackdown.

The data culled from the database catalogue combine many of the elements of historical research which can be profitably analysed and give us an opportunity to study a network that is relatively circumscribed in both time and place. Further, the connections between individuals are consistent since they all represent co-occurrences in photos. Knowing the boundaries of the network and having an understanding of what the underlying data represent are fundamental to creating a data model and visualisations which contribute to a research problem rather than merely illustrating an archive. The individuals who appear most frequently in the data are known historical figures, many of whom were instrumental in the creation of the archive: Matthias Domaschk, Jürgen Fuchs, Roland Jahn, Robert Havemann, Katja Havemann, Bettina Wegner, Carlo Jordan, Gerd Poppe, Bärbel Bohley and Tom Sello.⁷

Many of these photos are of protests and actions by the opposition; other photos are casual portraits and group shots not obviously related to any political action. The photographers are recorded (where they are known) by the archivists; the photos were, for the most part, taken by members of the group and their acquaintances. We see one example of a group photo (Figure 13.1) taken at Ulrike and Gerd Poppe's home in Woltersdorf. In this photo, we see Robert Havemann, Ulrike Poppe and others who are gathered for a reading of



Figure 13.1: Reading at Ulrike and Gerd Poppe's property, photo by Gerd Poppe, courtesy of Robert-Havemann-Gesellschaft, 27 June 1981. Source: Robert-Havemann-Gesellschaft/Gerd Poppe/RHG_Fo_HAB_09781. All rights reserved.

the works of Gert Neumann on 27 June 1981, after the release of Havemann from police custody.

In this photo, we see the conviviality of the opponents of the regime who are drinking wine and enjoying one another's company despite the serious circumstances. This is a rather typical example from the photo collection documenting activities of East German opposition. At the same time, since photos like this successfully document social actions taken by individuals, they offer a reliable source to reconstruct historical social networks.

An Unconventional Method: Why Network Analysis?

Social network analysis has been used since at least the 1940s in the social sciences.⁸ Network analysis has only more recently been adopted within historical disciplines.⁹ Networks are a powerful analytical tool for understanding the structure of groups, especially at scale or when there are complex interrelations between large numbers of individuals. While the network is a 20th-century concept, social relations that could be described as networks have existed in all historical periods and in all societies.¹⁰ Whether we can profitably examine one particular social group through the lenses of network analysis is a reflection of how much we know about the internal structure of that group, the research question being asked and the completeness of the historical documentation

that could be used to reconstruct the network.¹¹ Historical network analysis (HNA) is deeply rooted in social network analysis (SNA), using the same basic structures and metrics. But HNA must engage with questions of historiography and the use of sources in a way that is different from sociological methods. In this chapter, we show how network graphs and analysis can be used to shed light on the structure and dynamics of the geospatial social networks of segments of the East German opposition movement between 1975 and 1990. This chapter's exploration of dissident networks serves as an example of how scholars can discover connections between individuals and sub-groups that might track the spread of dissident thought between people, as well as between geographical regions.

SNA allows us to study the meaning and importance of relationships in great detail and, thus, offers a promising tool to examine past communities in the aggregate:

With SNA, we are only interested in individuals as part of a much bigger whole. In fact, one advantage to the technique is that SNA helps us view an entire community and figure out which individuals we should be truly interested in and which ones were perhaps less significant. When we study past relationships systematically as SNA allows, the method will prevent us from misunderstanding the function of an individual's relationships or exaggerating the distinctiveness of those relations.¹²

In our eyes, the true power of HNA lies in its capability to untangle complex social interaction patterns by way of graphical visualisations, thus making those patterns easier to perceive and analyse. The limit of network analysis is in researchers reducing social relations to those patterns and potentially losing track of the ways in which the network abstracts from the source material. For this reason, HNA should be practised with a keen awareness of the source material and the cultural and historical context of the networks at hand.

Using Network Graphs to See the Big Picture

One of the difficulties in creating network graphs is to create legible diagrams which show the structure of the entire network. Many of the most famous network diagrams are both complex and vast. Such diagrams are often referred to derisively as 'hairball' graphs due to their illegibility.¹³ As we shall see, using a combination of complex graphs of the whole network and smaller, more precise graphs of network segments is an easy way to overcome this problem. Figure 13.2 shows the network as a whole. The cleaned-up and corrected database contains 841 records (photographs) with 171 unique person references. There are a total of 1,843 co-occurrences of these individuals in photographs documented in the database.

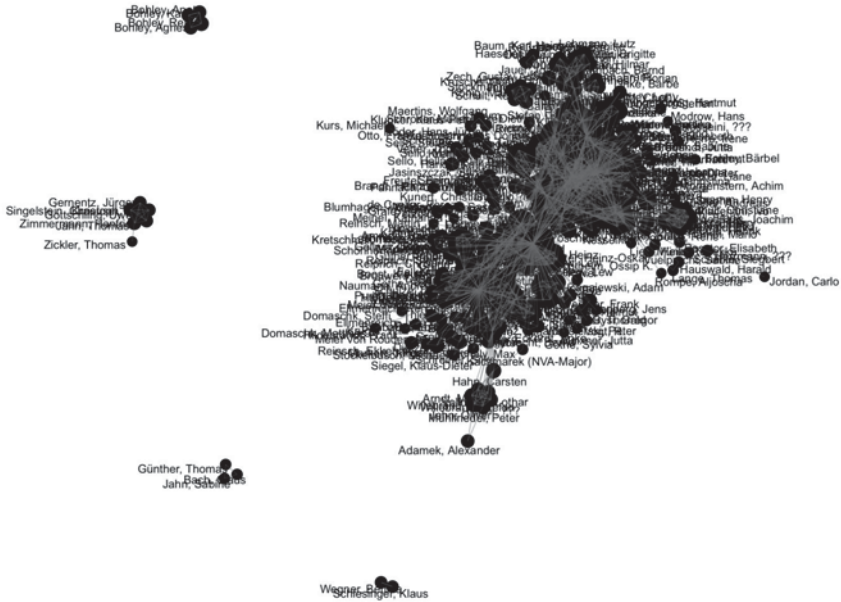


Figure 13.2: Network graph of GDR dissidents linked to Jena in photographs, 1975–1990. Source: Authors.

This network graph was created in one step in Palladio and was not modified, so it is relatively difficult to read, but it gives us a ‘quick and rough’ overall view of the network. There are many tools available for creating network diagrams which can make use of colour, refine the design and change the spacing and layout.¹⁴ A network is a set of nodes (in this case, people) linked by edges (in this case, photo co-occurrences); in this graph, the nodes are labelled with the name of the person and sized based on the number of photos that person appears in.

It is apparent in Figure 13.2 that the majority of people who appear in the photos appear together. Figure 13.2 evidences the existence of a core social network of densely connected people, in network terms, the giant component. Outside this core component there are four distinct yet very small social groups. These sub-networks emerge from photographs taken in apartments or in the outdoors, apparently just documenting daily events and activities and people involved in these. In fact, such photographs form a large part of the giant component as well, despite the repressive political circumstances in the late 1970s and the 1980s, which meant that taking photos documenting members of underground and dissident groups was a rather risky business.

One way to make network graphs more legible to humanities scholars is to combine graphs with other types of diagrams with which they may be more familiar.¹⁵ Palladio can be used to make maps, galleries and tables which can complement network graphs.¹⁶ In this case, we can use the same dataset to produce a map which shows the approximate number of photos taken at each

location. In order to obtain geographical data, we processed the keyword entries and collected all recognised geographical names and sub-regional information like street names or city districts which could be used to determine the city (for example, Berlin, Jena, Bad Frankenhausen). Almost three-quarters of the photographs included in our analysis could be connected with a geographical location. In the last step, we automatically geocoded these locations, finding latitude/longitude coordinates for all recognised locations for purposes of geo-spatial analysis. Figure 13.3 shows the locations of the photographs of individuals associated with Jena; the size of the circles represents the number of photos at that location. The green circles represent photos taken in East Germany and the red circles represent photos taken in West Berlin.

It is apparent in Figure 13.3 that Berlin and Jena are the most important regions and are strongly linked. We know from analysing the temporal distribution of these photographs that Jena remains the most important region until 1984, thus confirming previous studies stressing the importance of the Jena region for the East German dissident community. From 1985 onwards, Berlin gains in importance and becomes the most frequently referenced region in our data. This change is well in line with the overall course of events during the second half of the 1980s. A good example of how this change is connected to specific places are photographs referring to the *Umweltbibliothek* (Environment Library) in East Berlin. The library was founded in 1986 in the cellar rooms of the *Zionsgemeinde* and rapidly became one of the central communities of the East German dissident movement. What is not so clearly visible in the network graph (Figure 13.2) or the map (Figure 13.3) but evidenced by the dataset itself is that the core social network of East German opposition was rather small, revolving around certain key figures who were rather mobile. These central figures account for connections to some of the smaller towns. A good example of this phenomenon are the towns Fürstenwalde (Spree) and Grüneheide, both loosely connected to Robert Havemann (1910–1982), an intellectual and dissident sentenced to house arrest in 1976. A remarkable portion of photographs taken between 1975 and the early 1980s document Havemann's life under house arrest and the people visiting him.

Freezing Networks: Snapshots in Time and Local Networks

Just as giving researchers a view of the whole is important, it is helpful to visualise the network in specific places and at specific times. We can do that easily by filtering by place in Palladio and visualising only the nodes and edges associated with one place. Figure 13.4 shows the Jena-associated people who appear in photographs taken in West Berlin. West Berlin is a locus for individuals who were deported from East Germany but continued to be of interest to the East German government, such as Jürgen Fuchs and Roland Jahn. We can see that the network in West Berlin is much smaller, but still contains very important figures.

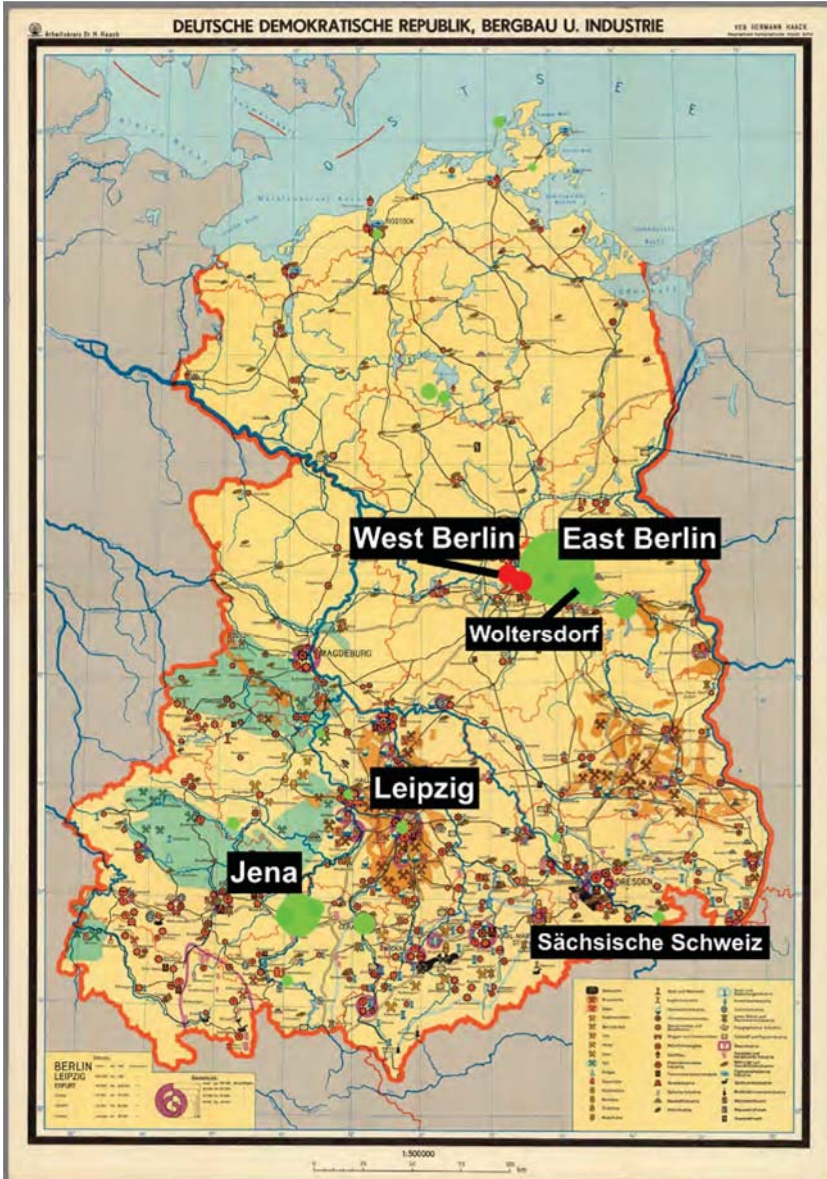


Figure 13.3: Map of photographs of GDR dissidents linked to Jena, 1975–1990, basemap. Courtesy of the David Rumsey Collection. Source: Authors; basemap Haack 1965.

Roland Jahn was an active member in the Jena dissident community, who engaged himself as a young university student in protest actions from the mid-1970s onwards. Jahn was ex-matriculated from the University of Jena in



Figure 13.4: Network graph of GDR dissidents linked to Jena in photographs taken in West Berlin, 1975–1990. Source: Authors; data from the Robert-Havemann-Gesellschaft.

1977, arrested again in the early 1980s, and finally expelled to West Germany in May 1983. However, even after his expulsion, Jahn continued to support the East German opposition movement. Jürgen Fuchs, in turn, had already been expelled to West Germany in 1977, but he remained involved in the East German dissident community until 1990, especially via Lutz Rathenow, another strong figure in the network visualised in Figure 13.4.¹⁷ Figure 13.4 shows that the networks of Fuchs and Jahn were, indeed, interconnected and formed the core of the network in West Berlin. There are other sub-networks (for example, the sub-network without a clear hub) of which Lilo Fuchs, Wolfgang Diete and Lutz Leibner were members. Despite appearing in a similar number of photographs as Jahn, Fuchs connects more individuals and more disparate parts of the network.

Similarly, making network graphs of specific moments, or ‘snapshots’, of the network at various times can be highly informative in understanding the evolution of the network and disentangling connections made in different periods.¹⁸ Whereas Figure 13.2 is a static network that displays all of the nodes and connections in the period from 1975 to 1990, Figure 13.5 shows only the nodes and edges present in 1981. Tom Sello, the construction worker from Großenhain, is central to one network here, despite his relative youth at 24 years of age.

The dataset was filtered to the year 1981 to produce Figure 13.5; the graph shows two clusters of dissidents and their associates who appear in photos together. The first cluster is centred on Tom Sello and six other people in Sächsische Schweiz. The second is a larger set of individuals in a tightly clustered network located in Woltersdorf and centred on the trio of Reinhard Weißhuhn,

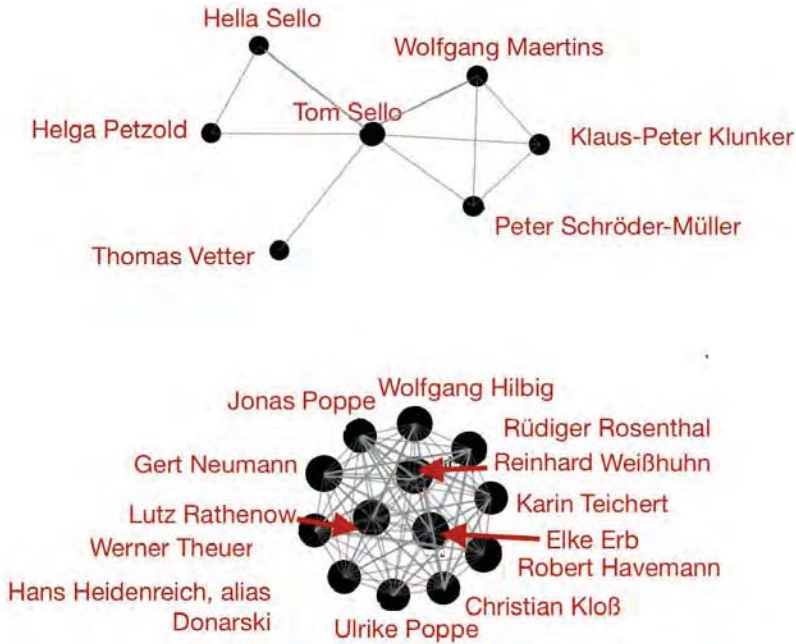


Figure 13.5: Network graph of GDR dissidents linked to Jena in photographs taken in 1981. Source: Authors; data from the Robert-Havemann-Gesellschaft.

Lutz Rathenow and Elke Erb. The high clustering shows that this group often appeared in many photos together. In fact, you may recognise this horizontal network of people with many connections to one another from the photo of the reading at Ulrike and Gerd Poppe's house on 27 June 1981 (Figure 13.1). This group appears in multiple photos together at the Poppe residence, which accounts for their high degree of association in the data. Unlike the photos of Tom Sello and his associates, these photos document a large and important gathering of leading figures from the movement in the same space.

Clarifying and Simplifying Network Graphs

Another way of cleaning up network graphs to make them more legible is to remove nodes that fall below a certain threshold in connection to the core of the network. In the case of Figure 13.2, we observed that the giant component comprised the majority of the nodes and edges, but was difficult to analyse due to its density. In order to focus on this giant component, which forms the core of the network, we removed the nodes which were not connected to the giant component and then graphed that network using Gephi. Figure 13.6 shows this network core.

By focusing only on the central component of the network, we can make out more of the network structure. For instance, we can see which nodes function

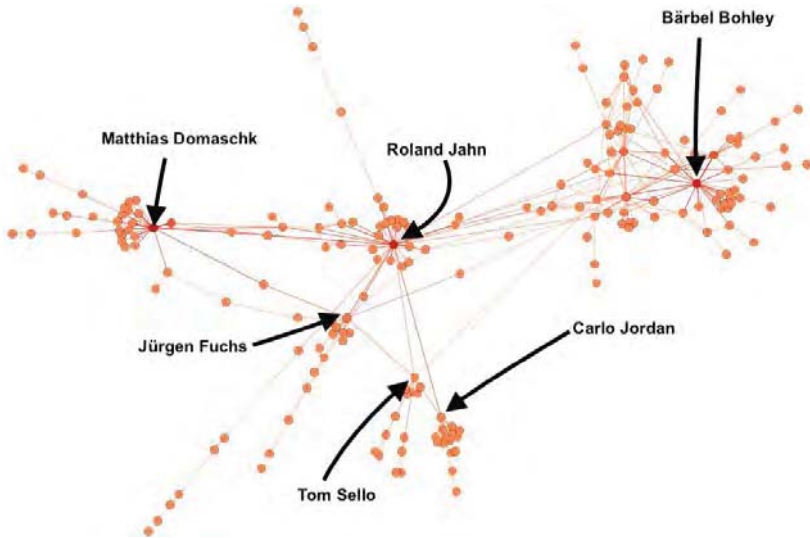


Figure 13.6: Network graph of GDR dissidents linked to Jena in photographs, 1975–1990. Source: Authors; data from the Robert-Havemann-Gesellschaft.

as hubs, having a number of connections that far exceeds the average number per node; we can also see which nodes function as connectors, tying together different sub-networks. Within this network, the centrality of Roland Jahn derives from the high number of individuals with whom he appears in photos, which makes him a hub; it is also significant that he connects the various communities, or cliques, we see in this graph. Other hubs include Bärbel Bohley, Matthias Domaschk and Carlo Jordan, all of whom appear in a large number of photos with other individuals. Many of the individuals who appear in photos with Domaschk and Jordan do not appear in photos with many (or any) others. Matthias Domaschk was active in Jena, including in the *Junge Gemeinde Jena-Stadtmitte*, but his contacts mostly only appeared in photos with him. Jordan was primarily active in the Berlin region, where he went on to become a leader in the green movement. Jürgen Fuchs and Tom Sello both have significant numbers of connections with otherwise isolated nodes, but they also have more connections to the broader network.

Comparing these observations to our lists of the presumed core members of the Jena group, especially Matthias Domaschk, Jürgen Fuchs, Bärbel Bohley and Tom Sello (see above), we can see the group was strongly connected through Roland Jahn, at least in terms of co-appearances in photographs. We can also readily observe that the women in the group, specifically Katja Havemann, Bettina Wegner and Petra Falkenberg, play less of a central role in tying together the network of photographic co-occurrences than do the major male figures. The relative absence of female figures from the core of the network makes the centrality of the artist Bärbel Bohley that much more striking,

especially in the latter years when she is highly central. Figures like Mattias Domaschk were photographed with others in the opposition movement who did not appear elsewhere in photos, much as we saw earlier with Tom Sello, suggesting that he was bringing many people into the movement or associating with non-opponents of the regime. On the other hand, Bärbel Bohley is part of a sub-network that has many more co-occurrences among its members and has no clear hub. Again, it is worth investigating whether this photographic pattern has any reflection in real-world relationships or if it is an artifact of the photographic record, as was the case with the clique centred on the Poppe residence in Woltersdorf. Jürgen Fuchs, Carlo Jordan and Tom Sello are each connected to their own small 'clique' of people together with whom they appear in photos, although these cliques are more cut off from other sub-networks, as well as being smaller. The most significant mystery remains why the sub-network containing Bärbel Bohley is more diffuse and 'leaderless' than the rest of the sub-networks. When we examine the photos in which she appears, we see that she appears in many photos with crowds and figures who are not otherwise in the database, including, for example, the Dalai Lama. It would appear that Bohley brought many new people into the movement and connected dissidents to a broader world of activists than earlier 'hubs'.

Conclusions

HNA can make use of the tools of SNA to understand the structure of social relationships, whether in smaller networks or at scale. While some of the tried and tested techniques, like large 'hairball' graphs, may be limited in their use value for historians, network graphs which reflect real research questions, such as the shape and size of a network in a particular year or in a particular place, can easily be created when one is familiar with how information has been captured and structured.¹⁹ These graphs are not mere illustrations of previously known relations, but a way of exploring segments of a large dataset to find new patterns and new questions: how, for example, the GDR opposition movement changed in relation to events like deportation to Berlin or the release of a member from detention. We have seen that the changes in the network following such events are not always predictable, such as when the movement held a very large and semi-public reading at Woltersdorf following Havemann's release or when Bärbel Bohley's very public events sometimes led to photographs depicting fewer documented members of the movement. For this reason, it is fundamental to consider what the graph is depicting (in this case appearances together in photos) and not to use the graphs in a naïve fashion to represent all real connections between individuals.

This chapter sought to exemplify how HNA could be used to explore and analyse personal and geospatial connections and ties behind a real historical phenomenon, the East German dissident movement. The networks reconstructed

and analysed in this chapter fit quite well with the historical facts about the East German opposition, although we need to note that we cannot estimate how much data is missing or whether a larger sample would have affected our results differently. Nor can we know to what extent patterns in the photographic co-occurrences reflect relationships outside of photographs. Network analysis is known to be quite sensitive to missing data,²⁰ but applying network analysis can reveal discrepancies, such as when a group gains members or loses them, or when the structure of a group changes from tighter to looser, as the Jena-linked dissidents often did when they were deported or arrested. Network analysis can also reveal when group members like Tom Sello or Bärbel Bohley introduce many new people to the network. Despite these pitfalls, we are hopeful that our chapter could convince its reader that HNA could help scholars to gain new insights into the way in which social networks reflect political pressures in sometimes unexpected ways (for example, by growing or becoming more public in response to crackdowns). Although not impossible, such insights can be difficult to gain through the traditional methods of historical research.

Notes

- ¹ Conroy 2019.
- ² Robert-Havemann-Gesellschaft e.V., Archiv der DDR-Opposition, Bildarchiv.
- ³ The security authorities' operation 'Counter-strike' is documented in BStU 2013.
- ⁴ See, e.g., Schroeder 1998: 233ff; Gieseke 2008.
- ⁵ Veen 2000: 27–29.
- ⁶ Neubert 1998: 488.
- ⁷ Matthias Domaschk was a young political activist, who died on 12 April 1981 in Gera in a pre-trial detention of the East German security service after 13 hours of continuous interrogations. Tom Sello, in turn, engaged himself in several dissident groups in the GDR, especially in the 1980s. He also wrote for several underground publications (*Samisdät*) and was repeatedly attacked by the security service. Bettina Wegner was an East German songwriter and lyricist. In 1983, she was threatened with prison and forced to leave the GDR for West Berlin. Gerd Poppe was a political activist who fought for human rights in the GDR. He was also actively engaged in the publication and dissemination of several illegal underground publications (*Samisdät*). Poppe was subject to the Stasi's intensive observation and repressive activities. Bärbel Bohley was an East German opposition activist and artist. She was one of the co-founders of the Initiative for Peace and Human Rights (1985) and of *Neues Forum* (1989). For a detailed description of the other people mentioned, see Elo 2018.
- ⁸ For a history of the development of SNA within the social sciences, see Prell 2012: 19–50.

- ⁹ On how HNA differs from SNA more generally, see Düring and von Keyserlingk 2015.
- ¹⁰ Lemerrier 2015.
- ¹¹ Ibid.
- ¹² Morrissey 2015: 69–70.
- ¹³ Nocaj, Ortmann & Brandes 2015.
- ¹⁴ Some of the most commonly used are Gephi, Cytoscape and R.
- ¹⁵ Conroy et al. 2020.
- ¹⁶ Conroy 2019.
- ¹⁷ See further Elo 2018; Neubert 1998.
- ¹⁸ Conroy et al. 2020.
- ¹⁹ Drucker 2011.
- ²⁰ Wetherell 1998: 125.

References

- BStU.** (2013). *Aktion 'Gegenschlag': die Zerschlagung der Jenaer Opposition 1983*. Berlin: BStU.
- Conroy, M.** (2019). *Visualizing historical networks with Palladio*. White paper, Humanities Commons. DOI: <http://dx.doi.org/10.17613/785h-s253>
- Conroy, M., et al.** (2020). *Visualizing historical networks using graphs and matrices*. White paper, Humanities Commons. DOI: <http://dx.doi.org/10.17613/jcd0-vt59>
- Drucker, J.** (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, 5(1), 1–21.
- Düring, M., & von Keyserlingk, L.** (2015). *Netzwerkanalyse in den Geschichtswissenschaften: Historische Netzwerkanalyse als Methode für die Erforschung von historischen Prozessen*. Berlin: Springer.
- Elo, K.** (2018). Geospatial social networks of East German opposition (1975–1989/90). *Journal of Historical Network Research*, 2(1), 143–165. DOI: <https://doi.org/10.25517/jhnr.v2i1.45>
- Gieseke, J.** (2008). East German espionage in the era of detente. *Journal of Strategic Studies*, 31(3), 395–424.
- Haack, H.** (1965). *Deutsche Demokratische Republik, Bergbau u. industrie*. Arbeitskreis VEB Hermann Haack, Geographisch Kartographische Anstalt Gotha (Wirtschaftskarte: Deutsche Demokratische Republik. MDI d. DDR nr. 7579/62).
- Lemerrier, C.** (2015). Formal network methods in history: why and how? In G. Fertig (Ed.), *Social networks, political institutions, and rural Societies* (pp. 281–310). Turnhout: Brepols Publishers.
- Morrissey, R. M.** (2015). Archives of connection. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48(1), 67–79.

- Neubert, E.** (1998). *Geschichte der Opposition in der DDR 1949–1989: Forschungen zur DDR-Gesellschaft*, Berlin: Ch. Links.
- Nocaj, A., Ortmann, M., & Brandes, U.** (2015). Untangling the hairballs of multi-centered, small-world online social media networks. *Journal of Graph Algorithms and Applications: JGAA*, 19(2), 595–618.
- Prell, C.** (2012). *Social network analysis: history, theory and methodology*. London: Sage.
- Robert-Havemann-Gesellschaft e.V., Archiv der DDR-Opposition.** *Bildarchiv*. Retrieved from <http://www.havemann-gesellschaft.de/archiv-der-ddr-opposition/bildarchiv/>
- Schroeder, K.** (1998). *Der SED-Staat: Geschichte und Strukturen der DDR*. Munich: Bayerische Landeszentrale für politische Bildungsarbeit.
- Veen, H.-J.** (Ed.) (2000). *Lexikon Opposition und Widerstand in der SED-Diktatur*. Munich: Propyläen.
- Wetherell, C.** (1998). Historical social network analysis. *International Review of Social History*, 43(1), 125–144.

Transit of Venues: Discourses of Astronomy

Reetta Sippola

A Popular Astronomical Event

In the 1760s, one of astronomy's rarest predictable phenomena, the so-called Transit of Venus, was calculated to take place twice: in 1761 and in 1769. This phenomenon, when the planet Venus passes across the Sun, from the Earth's vantage point, was not only extremely rare, as the previous transit had taken place in 1639 and the next was to follow in 1874, but also very valuable scientifically, as observing this kind of transit would make it possible to determine the distance between the Earth and the Sun more accurately than before. This could in turn make it easier to improve a number of practical issues relying on astronomical knowledge, foremost among them to improve the accuracy of calculating locations at sea, which at this time was at best inaccurate, often resulting in costly and deadly accidents. Thus, the two Transit of Venus events and the astronomical information that could be derived from observing them enjoyed wide interest among both scientific professionals and the general

public. The scientific interest in the transits during the 18th century was represented through a large number of news items and scientific reports in the scientific literature, especially in scientific periodicals, such as the *Philosophical Transactions of the Royal Society of London*. In short, there was a large and varied scientific discourse talking about the Transits of Venus. This chapter explores what new historical knowledge about science in the 18th century we can derive from using digital methods to study such scientific discourse related to a particular scientific phenomenon.

However, alongside the natural philosophical news and reports, many broader perspectives towards early modern knowledge and the early modern world itself were also communicated in the scientific journals. By using digital history methodologies to analyse the qualitative meanings and quantitative amounts of the common topics and themes in a scientific periodical, I suggest in this chapter that there were simultaneously nine different ways of talking, or discourses, about astronomy and the two Transits of Venus of the 1760s.

Astronomy has been one of the main ways for scientists to explore explanations about our place in the universe, but astronomical knowledge has also been used for other more practical applications in economy, politics and transportation. This combination of pursuing natural philosophy endeavours both for knowledge and for practical applications was also central for the Royal Society of London, established in 1660 in England for improving the knowledge of nature and mankind, as articulated in their full official name. To serve astronomical practice, the Society funded two large expeditions to make observations of the 1760s Transits of Venus, but at the same time and partly through these expeditions they also took part in the larger and wider transformation of how the natural sciences were understood in the 17th and 18th centuries, by creating a communal and public space for circulating the new knowledge. The Society circulated a newsletter that disseminated and shared the new scientific information coming from collecting and observing abroad on scientific voyages and commercial encounters, as well as through experiments at home. The letter soon turned into the form of a periodical, entitled *Philosophical Transactions Giving some Account of the present Undertakings, Studies, and Labours of the Ingenious in many considerable parts of the World*, established in 1665 and published as the Society's official journal from 1752. From the mid-18th century onwards, *Philosophical Transactions* was a journal publishing scientific correspondence that had been selected and reviewed by the fellows of the Society. The journal also had a much wider readership than just natural philosophers.¹ Its topics included news about the latest innovations and discoveries and reports on geography, natural scientific specimens and natural and man-made phenomena, such as weather and electricity. Even the work of amateur experimenters was published, whereas many professional reports sent to the Royal Society were rejected and put to one side with a note 'not to be printed'.²

In the mid-1700s, many nations participated in a race for the knowledge, prestige and power that could be gained through successful observations of the two

predicted Transits of Venus. National and international networks of knowledge invested a significant amount in astronomy research, and the importance of succeeding in observing the Transits of Venus has often been compared to the Cold War space race of the 20th century. Central in this were the observation instruments which were carefully collected and arranged for transportation to judiciously planned observation locations.³ The Transits of Venus were then widely discussed in various contemporary printed periodicals, as well as in the general press. It was understandably a prominent topic in periodicals such as *Philosophical Transactions* reporting on the latest news within natural philosophy. The emphasis was on the conducted experiments and the reporting and theorising of the events. However, at the same time, the Royal Society's public communications also conveyed other more elusive and implied perspectives on the philosophical inquiries to their wide group of readers. In this light, this chapter uses the reporting of the Transits of Venus in *Philosophical Transactions* to critically discuss what kind of view of 'scientific' inquiry was offered to its readers: What was actually discussed about astronomy in particular and the new science in general among the texts that succeeded in being printed in *Philosophical Transactions* during the late 18th century? The guiding hypotheses behind the study is that the general values of 18th-century British society and the philosophical environment have been repeated directly and indirectly in the texts. By re-reading this discourse with the digital methods, we can access the underlying thoughts, paradigms (such as a shared model of the universe) and practices related to experimental science that existed at this time.

Topic Modelling the Publication of New Science

The historical meanings of the editing and peer reviews of *Philosophical Transactions* has been studied by Ellen Valle in particular, and subsequently by Julie McDougall-Waters, Noah Moxham and Aileen Fyfe.⁴ My research continues their work by applying the use of 'machine reading' and the method of topic modelling for finding the underlying patterns and values that have influenced the public writing about new scientific knowledge. By comparing statistical amounts of various topics in *Philosophical Transactions* at different times, topic modelling reveals temporal changes in scientific approaches in the many ways of talking about the process and of the degrees of certainty within the emergent sciences. This could be a major contribution to the history of early scientific communications because this kind of temporal change has not been discussed in the earlier scholarship, which has only applied manual reading of these resources. The digital methodology of topic modelling provided a way to examine the large amount of texts in *Philosophical Transactions* and to reveal the changing patterns in the ways in which the two transits were discussed. It located the internal relations inside the 'big data', comprising all the words in the texts, and teased out their shared and underlying meanings.

I have focused on 1753–1777 as a period of 25 years containing the important astronomical events (the two Transits of Venus in 1761 and 1769). This period also covers the time after the Society had formally taken over the journal in 1752 and begun editing the articles communally via the so-called Committee of Papers, and before it changed into a new period of knowledge circulation in 1778 when Joseph Banks took on the presidency of the Royal Society. The time frame also contains the collection of new data during world-changing voyages of exploration, such as those of Captain Wallis in 1765–1768 and Captain Cook in 1768–1780, both supervised by the Royal Society.

The selected corpus of texts from *Philosophical Transactions* consists of a set of 1,421 documents relating to the transits, which represents a collective public discourse on scientific discoveries, innovations and experimental routine (a typical genre of texts published in the journal). The texts of the corpus were generated by optical character recognition (OCR) of a digitised version of *Philosophical Transactions* provided for research purposes by JStor, and I used a temporal selection of the *Royal Society Corpus*⁵ which had been collected and pre-processed in a previous linguistics project headed by Elke Teich.⁶ Their pre-processing included the transformation of data into a standardised format, cleaning of data (for example, OCR errors) and derivation and annotation of metadata.⁷

The research process consisted of three stages, which combined statistical and computational quantitative methods with qualitative analysis of the texts. During the first stage, I applied topic modelling to the corpus to create lists of probable keywords describing the themes existing within the data corpus. To operate with the topic modelling algorithm, I prepared a so-called ‘stop list’ of very common words (such as *the, is, as, etc.*, as well as prepositions and conjunctions) to be filtered out and excluded in the analysis of the corpus. I also arranged the processing of the corpus temporally by dividing it into five-year sets (1753–1757, 1758–1762, 1763–1767, 1768–1772, 1773–1777) and then ran these through the topic modelling algorithm implemented in the MALLET software application. The output was a list of various topics or themes of interconnected keywords co-occurring throughout the articles. As a preliminary analysis, I used MALLET to produce a varying number of topics from the corpus and concluded that the best and most realistic fit, in terms of what from the keywords appeared to be relevant and meaningful topics that were not too general or too narrow, was when the number of topics was 50. Following this, I grouped the topics along semantic similarities in their keywords in order to locate their relations to the scientific contexts in which the original texts of the corpus were created. This was a crucial part of the research; while a ‘topic’ to the computer is merely a list of words that occur together in statistically meaningful ways,⁸ in the following manual analyses by the researcher these lists are shown as a semantically meaningful string of keywords which need to be sewn together with a larger thread of historical context and interpreted in order to provide a meaningful label that refers back to the historical context. Therefore,

the historical work of contextualising the English 1700s science discourse was ongoing throughout the entire research process.

In summary of the above, I programmed MALLET to create a list of 50 topics, each consisting of 20 keywords, and entered these into a spreadsheet which showed the number of these topics during the chosen years. The keywords represent co-occurrence patterns in the data corpus and probabilistic appearances of a particular theme. I located keywords that signified general trends and particular themes, such as those that concerned instruments in observation and experiments. After examining and labelling all the topics which were present in 2% or more of the text mass, I grouped the topics according to their semantic similarities. This revealed that the common topics in the data could be roughly grouped according to 'polite correspondence style', 'astronomy', 'chemistry', 'weather' and 'instruments and their use'. In actual fact, it was not until this stage that the processing of the data highlighted the most fruitful research questions warranting further investigation.

As mentioned above, Teich's linguistics team had already carried out research that applied topic modelling to the first two centuries of *Philosophical Transactions*. Their visualisations enable us to see topical trends in the corpus, in particular in terms of discipline formation and specialisation, that there is a growing separation of individual scientific disciplines over time and that the discourse became increasingly specific over time.⁹ Aiming at different perspectives on the proposed material and research questions, I thus designed the commands and continued to fine-tune my topic modelling less towards general themes stretching over large time periods, and more towards more specific themes connected to my more focused period and scientific developments. As a result, this is an example not only of applying a fresh perspective on an already known resource, but, importantly, also of a collaboration among experts which is characteristic for contemporary digital humanities.¹⁰ Teich's team identified 24 different topics in *Philosophical Transactions* which indicated the development of scientific sub-disciplines (including Chemistry, Mechanics and Reproduction, among others). This research was an excellent starting point for my inquiries, in which I used an alternative approach to examine the periodical and which found 25 topics that were theoretically visible, as they each referred to more than 1% of the amount of text. Our two approaches to the same data differ in the close reading perspectives, and the comparison has led me to notice other wider perspectives. In the following section, I will problematise what else can be depicted from the results of a computer-assisted distant reading of the periodical from the perspective of the cultural history of science.

Thematic Trends of the Discussion

When beginning to discuss the quantitative results of the topic modelling, it is indeed no surprise that the keywords and their frequencies show a nota-

ble rise in astronomical topics around both transits. They are visible during most years as a relatively large number of texts on astronomy, with the text mining revealing keywords such as ‘transit’, ‘June’ and ‘Venus’, where ‘June’ was the month when the 1761 transit occurred. Astronomy and its related practices and techniques were commonly mentioned, but in the researched time span, *Philosophical Transactions* also discussed a few non-astronomical topics. However, these also helped to locate relevant aspects of the scientific ways of talking also within astronomy, as I will demonstrate below when discussing the case of systematising observation results. In particular, these non-astronomical topics have a specific value as they can vividly illuminate what is *not* part of other discussions.

Shared discourses, split viewpoints

The remainder of this chapter will analyse the astronomical ways of talking by contextualising and connecting the prevalent topics in the published scientific communications (for example, in *Philosophical Transactions*) to the making of early modern knowledge. I begin with a thematic analysis of the text-mining results and then conclude with a summary of the approaches and temporal appearances of the nine ways of talking about science. Keywords of all of the discussed topics are listed in Appendix 14.1.¹¹

The data shows that the discourse on the first transit (1761) circulated around various viewpoints concerning reliability and the structure of the solar system. The common themes consisted of exactness, measuring, relating and belonging (to the universe), which also reflect the excitement of using the enthusiastically approved, state-of-the-art instruments in making observations of the first transit. Instead of describing causal chains, the observations were communicated as indicating the values of various connecting sets of rules and theories. The observations were thus not neutral, but became charged with rules and theories.

The lists of observation data were apparently published as they had been recorded, and were then consequently explained through the use of mathematical algebraic reasoning (the topic ‘Astronomical distance’ with 11%), as a system of heavens that could be revealed (‘Rules of the stars’ topic with 5%) and how it related to the solar system (‘Appliances’ topic with 8%).¹² The first two topics were only present in 1758–1762, while ‘Solar system’ was an ongoing topic, although notably only at the beginning, as it almost disappeared (1% to 2%) around the second transit. In 1763–1767, the theme seems to split into two topics, as a similar topic ‘Heavenly bodies’ shortly appears at this point. However, it differs somewhat, as this topic focuses on the planetary system and its parts (which are described in a similar manner as the reports would talk about the human body, possibly in a ‘plain’ style), while the topic ‘Solar system’ discusses what can be achieved on Earth by using those heavenly bodies:

for example, 'latitude' and 'degree' are among its keywords, which could often signify the calculation of the location at sea. These distinctions already show that the discourse on the transits addressed many separate views on the topic, varying talk which not only promoted causal theories and mathematical reasoning or arranging raw data 'plainly', but also offered discussions on more social factors, such as the extent of trust and belief that could be placed in astronomical calculations.

There might be some skewing of this topic, as a large number of its keywords might be brought about by one single lengthy paper (23 pages in print) in *Philosophical Transactions* by the Astronomer Royal Nevil Maskelyne, on observations and methodological rules he developed following the 1761 transit. By collecting and comparing the data received from various observations at St. Helena about the Transit of Venus, Maskelyne formulated a set of theoretical rules about the skies and the use of astronomical instruments.¹³ During this time, if the transaction discourse is representative, apparently a rather small part ('Rules of the stars' topic with 5%) of the astronomical interest concerned astronomy's theoretical system. This theoretical discourse topic also existed at the time of the second transit in 1769, meaning that Maskelyne's ideas continued to be referred to or discussed after his paper's initial 1761 publication. However, as the topic decreased from 5% to 2%, this indicates an increasing emphasis on other perspectives, such as a dispute on the relative importance of causal versus mathematical evidence. That Maskelyne's article constituted a dominant part of the topic could easily be identified through a keyword search in the dataset, and it is worth emphasising that it is both important as well as useful to also apply close reading to enable closer study and control the content of themes that arise from the distant reading of the text-mining results.

One alternative to using theories and mathematics to explain astronomical measurements was through empirical means (namely, repeated experiments or observations). This is demonstrated by the topic 'Distance measuring process' being continuously very strong (9–12%) throughout the period, thus representing scientists' interest in the general process of experimentation, while the keywords of the 'Astronomical distance' topic point towards explanations and mechanisms based on the laws of physics, physical circumstances and entities, such as velocity and distance. Apart from 'glass' (a reference to a material most likely indicating the lens of a telescope or an experiment tube), the keywords of the 'Distance measurement process' topic ('distance', 'degrees', 'line' and 'places') primarily point towards intangible concepts connected to very tangible experiments, of measuring and of practically producing new observational data. Interestingly, its keywords of 'appearance', 'means' and 'method' give this discussion various kinds of unsure, unsettled and dynamic connotations and, furthermore, the topic contained discussions about particular location-specific, probable or 'apparent' aspects.

The topic 'Astronomical distance' was discussed only once, in the 1758–1762 segment, with a sudden 11% spike, in which the observations of the heavens

(and the astronomical signified keywords of [moon/sun's] limb, sun, star and wheel) were linked with the data in the tables (measure, feet, foot, inches) and algebra (cos, distance, wheel). This and the 'Distance measurement experiments' topic both mention the word 'distance', but the other keywords indicate a clear difference in meaning between the two, as the topics reveal how 'distance' was to be observed, depending on whether it addressed celestial distances or distances on Earth. In comparing these two topics, I claim that the topic modelling algorithm has revealed a very important temporal change: it has located a mainly shared discourse about the practices of deciding on the distance, but it has split this theme into two. Their separate keyword lists initially differ between practical experimentality and understanding the physical theories of the phenomena, but after this, both topics also share similar passages about the use of instruments and theorising the measurements.¹⁴ Both topics discuss various means to measure distance, and the shared keywords of *measure*, *distance* and [the method of heaven's] *wheel*, which all appear in both of the discussed topics, capture a broader, shared discussion. This is the value of topic modelling: by using machine-reading and locating shared but slightly variable topics, the nuances of the discussions can better come to light.

The examination of these topics indicates that there were two profoundly different ways of what was considered to be the reliable way to find data. The first transit in particular allowed plenty of interest (11%) in making arguments through calculations, but for the second transit the frequencies indicate that the mathematics were no longer a primary worry. There is only a short temporal change within the statistics that could, however, exemplify a dramatic change in the paradigm. It appears that a more general perspective about measurements was developing and that means other than mathematical results had become increasingly relevant in the early modern search for reliability and 'truth'.

Following Newton's formulation of mathematical laws of nature, in the 18th century it was usually considered sufficient to settle an astronomical dispute if one could arrive at a successful set of mathematical calculations.¹⁵ At a time when logical causalities became less interesting to the scientists, the large number of words indicating probability in my data raises questions about reliability, the meanings of eye-witnessing and the capabilities of the human mind and senses. These topics can also indicate that there was a split discourse according to the two schools of realists: one consisting of mathematical realists, including those who thought mathematics could provide the real motions of celestial bodies; the other comprising the group of physical realists who held that a mathematical model of the real structure of the heavens should be based on physics.¹⁶

Another example of transitions within the paradigm can be noted in the texts written in Latin. A few data and observation reports were received and published in Latin, including the transit observations from Uppsala in Sweden.¹⁷ However, this kind of discourse was narrow and concentrated into two topics: 'Solar system Latin' and 'Transit and Latin.' The two topics were both most

popular around the first transit, and while they both circulated the calculations for longitude, they differed in their perspectives. The first described the observations plainly, while the other seems to have emphasised the materiality of the observation process. This means that in Latin there were at the time in question two ways of discussing astronomy: first, as descriptions of the solar system; and, second, in the form of the actual process of making observations of the solar system. It also appears that during this period both Latin topics started to disappear. This decline of Latin had also been noted in passing by the previously mentioned linguistics researchers when stating that some non-thematic topics, including texts written in Latin, reached their peak in the early 18th century.¹⁸ However, as we can observe in the keyword list, Latin was still used in the late 18th century in publishing accounts concerning the heavens and the solar system. The topic ‘Solar system Latin’ was popular between 1758 and 1767, with a 6% to 8% share of all discussions, but then almost disappeared (1% to 2% in the subsequent decade), while the ‘Transit and Latin’ topic was visible throughout the entire transit decade, but only with a 2% to 5% share.

The host of Latin words in ‘Transit and Latin’, ‘vero’, ‘inter’, ‘hoc’, ‘enim’ and ‘hujus’, are all abstract, frequent words. ‘Vero’ is the fascinating one, as it means ‘truly, indeed, to be sure, certainly’. In fact, among the 500 keywords in the other topics, there are many that indicate a continuous interest towards the ‘true’, such as in addition to ‘true’, also ‘purpose’, ‘order’, ‘error’, ‘anemones’, ‘effect’, ‘probability’ and ‘rules’. And deriving from the keywords ‘apparent’ and ‘error’, the ‘Telescope observations’ topic can be seen as promoting an openness towards the meanings of the results and how changing circumstances could affect the observations, although the task in question seemed to be regarding some precise measurements, including ‘minutes’ and ‘seconds’.

Taken together, these topics discuss a contemporary science culture that was seen as being open-ended and constantly recreated by various actors.¹⁹ In particular, the certainty of probabilities and the transparency about practical choices when making observations seem to be continuing themes in the various discourses.

Talking about the weather

The external conditions of the observations were and are a central issue to astronomers. It especially appears to be the main way to talk about the astronomical observations in one topic which, due to its huge number of weather keywords, has been labelled ‘Astronomer’s weather’. This was a strong topic during its time: although it only existed during the 1768–1772 period, the topic arrived with a notable 14% of the volume of text published in *Philosophical Transactions*. The first words in this topic are ‘ditto’, ‘air’ and ‘limb [of the moon or the sun]’ which apparently refer to astronomical observation tables. These keywords were also accompanied by ‘June’ and by as many as five words describing weather.

Throughout the years in question, the weather was continuously discussed, but there is some increase in the latter 10 years of the studied period, during which the second Transit of Venus took place. This sudden rise in the topic modelling chart is understandable in the light of the unfortunate event of not having been able to make successful observations of the transit on the first attempt in 1761 due to difficult weather. It is therefore no surprise that we can find some discussion about the conditions already before and after the 1761 transit: it was anticipated that clouds could obscure visibility at the vital moment, as they had so often done when Edmond Halley had been observing a similar transit almost a century earlier.²⁰ The unpredictability of weather was just one of the several reasons why there was a need to simultaneously observe the Transit of Venus from many different and varied places on Earth. Beside cloud cover, Halley had already in 1716 pointed out in his observation advice the importance of arranging coverage from northern and southern latitudes, as both the ingress and the egress could not be observed from all the locations. Thus, to avoid a similar fate, a large-scale operation was connected to the Transits of Venus in order to coordinate several international observations and, as had been suggested by Halley, to arrange multiple observatory tents to be set up around the globe by British, Austrian and French observers.²¹ The more observations that were made from many different and widely separated vantage points, the more accurate the ultimate results were likely to be.²²

Weather was also employed in other discourses. In the topics ‘Environmental circumstances’ and ‘Travel narrative’, weather figures in the form of the blame and complaints that often figured in early modern observation accounts. These often served very practical purposes, as a failure to conduct an experiment could often be more easily accepted due to bad weather. Simultaneously, these topics address the practice and materiality of the observation process in the form of broken instruments, successful delivery of the instruments and other aspects affected by bad weather. In this, the topics provide valuable tangible insights into early modern scientific practice.

The topics concerning the weather were used in particular around the second transit (1769). The topic ‘Telescope observations’ appears to mention the weather during both transits, but only moderately, with a few percent (2% to 4% during the 1758–1772 period), while the main emphasis is on discussing the way in which the new more advanced instruments could create or diminish the observation errors. Nevil Maskelyne wrote a letter to the Royal Society in 1761 which discussed at length various contextual influences on the use of the instruments. Before listing the transit observation results, Maskelyne wrote about ‘the observations themselves, and mention[ed] some cautions concerning them,’²³ which among others included the exact practical adjustment of the quadrant (a navigational instrument used for angle measurements).^{24,25} The context in which the instruments were used, then, was important to the narrative and the manner in which they were discussed.

On the other hand, the failure to produce consistent and certain results during the scientific voyages of observation connected to the 1760s transits has been blamed not just on bad weather, but also on other external or material conditions that affected the voyage, such as war, illness and, above all, inexperience in observing the phenomenon in question.²⁶ The quantitative distant reading of results through topic modelling provides a new opportunity to investigate how strong the pattern has been to make arguments about the failure (and at times success) of experiments through referring to external circumstances. Such talk about instruments and circumstances that affected the observation process attracted a wider group of readers among scientists besides astronomers. In the texts, therefore, such talk can also be seen as functioning as a way of elevating the experimenter's status as a scientist, or as mirroring the scientist's thoughts or search for reinforcement from other scientists; namely, to write down the observations in a punctual manner was entangled with a deeper purpose that made visible the discourse of complimenting others or showing evidence of their mistakes and failures. Reid even called this kind of performative behaviour 'playing the astronomer', as the astronomer was in a precarious position on an expensive voyage as he was likely the first one among the crew to be seen as less important and as excess weight.²⁷

In this light, it is interesting to examine how instruments and bodies were referred to, as this reflects the general values that kept being repeated in the accounts. Simon Schaffer has suggested that the states of disrepair in observation reports or travel narratives refer simultaneously to the tools and to the humans that interact with them and with one another.²⁸ As he wrote, the states of disrepair helped distribute responsibility across cultures and spaces, offering resources to defend some reputations and damn others.²⁹ From this perspective, the referral to the external conditions and the others who contributed in the making of the observations were part of the pattern of this way of distributing the responsibility for scientific success and failure.

Referring to the sources and collaborators was also a matter of reliability. According to the semantic arrangement of the keywords in the topics 'Astronomer's weather', 'Environmental circumstances' and 'Travel narrative', the materiality of objects was closely linked with the circumstances of observing. Maskelyne's previously mentioned 1761 lengthy report was careful to mention and honour the builders and senders of the equipment,³⁰ and in a similar fashion while on a state-sponsored exploration voyage, another astronomer in his journal listed the people who had built or fixed the instruments he used.³¹ If no observation was possible due to fog or a broken instrument, their role in the experiment or its possible failure was at least taken into consideration in the description of the event. Hence, the keyword 'wood' is interesting. This might indicate the 'honouring' often offered the makers of the instruments used in the observations, as such compliments were written with full description of the types of materials used in building them and thus the

quantitative data also indicates the epistemological importance of the astronomical equipment.

Talking about the weather could also be an instrument of politics, as the making of science was, and is, very political. In the corpus, the weather was at times discussed in a text with what can be described as politeness or plain style. This structured a discourse which proposed openness towards the various possibilities and active influencing on the observer's circumstances. The impact of weather could not be controlled, but it could be understood by systematisation;³² and as it affected the success of observations and experiments, it was used as both an explanation and a weapon by scientists in the political game of making knowledge and acquiring more funding. Royal Society funding affected the discourse and the attempts to increase the natural history knowledge, and while only a nominal part of the Society's funding came from the king, the results and communications of scientific results needed to please the funders. John Henry has pointed out that this meant that the funders would wish to gain practical applications from the experiments and developments of science. As a result, the Society had to be very apologetic and have its value clearly propagandised in its attempts to demonstrate the usefulness of science to the state.³³ In order to maximise the attention within the administration, the importance and rarity of the observations of the singular 1761 and 1769 Transits of Venus would have given a reason to report the events with such care. It was a good opportunity to emphasise the relevance of the expensive astronomical observations, which was brought up wherever and whenever the transits were mentioned.

The final topic featuring the weather was 'Travel Narrative', consisting of plain or technical descriptions of weather, geography and exploration events. The topic was very frequent and steadily present by 6% to 9% throughout the period, but was at its largest at the turn of the 1770s. As weather has been proved to be an important part of spatial descriptions of 'new' regions, which were especially discussed around this time, it is most likely that there was a correlation between talks of weather and such new areas with promises for imperial and scientific explorations.

Having so many obviously different kinds of weather topics at the same time means that, simultaneously, many and varied aspects were considered. If some discussion has been very small (or unimportant) in its scale, it would not be visible in this kind of topic modelling of 1,421 articles.

Talking about instruments

Regarding the importance of various disciplines within astronomy, before 1758 and between the transits, there was according to the statistical results more room for showing interest in chemistry, whereas at the end of the 25-year period, the data indicates a turn towards widening discussions about the use of experiments and optical instruments. This was connected to experiments

of measurable phenomena such as air pressure, distance and microscopic life forms, and was influenced by technological developments (besides telescopes) that also led to increased use of three observation instruments in the form of the thermometer, the barometer and the hygrometer.

These were notable topics. The talk about the instruments and their particular use each received 6% to 10%, respectively, of the space of the published texts. However, the discussion was not fully diverted from astronomical topics, as the development of the instruments is linked to weather discourses. In fact, as the three new instruments became available to more observers, it was possible to collect much larger amounts of data of local and global weather patterns, so they could be systematised and understood not just as weather, but as climates. The influence of the political interests involved in producing the 'facts' of far-afield weather should not be underestimated either. Morgan Vanek has suggested that the actual reason why 18th-century literature is saturated with the rhetoric of meteorological science lies in seeing the topic as a prominent and productive term in the public debate about Britain's imperial obligations. She claims that 18th-century writers amplified the threat of environmental influence to justify a British right to govern all over the world, as with their governance the British could improve living conditions in the 'new' harsh regions.³⁴

Finally, the last approach to astronomy is quite different from the others as it addresses a rather different mindset. Interestingly, together with the perspectives on the influences of various circumstances, at the time of the 1769 transit there was also the topic 'Ancient tradition' (with 6% of the text mass), which linked the new observational data with ancient calculations and beliefs. This might indicate the astronomers' certainty of the coming success in observing the second transit and their enthusiasm in comparing their innovations with other foundational theories. This could also discuss the many stages in which the rational thinking could be influenced when the collection of new knowledge in 17th- and 18th-century Europe was largely a cooperation among various specialised actors, where some collected field data and others depicted and analysed it. According to Francis Bacon, the different investigators (observers, experimenters and theoreticians) were in this required to be on guard against 'the idols', the profound and sometimes erroneous ideas created in one's mind, as well as the defective sensations of any particular individual.³⁵ Indeed, it sounds like a difficult business to create reliable knowledge navigating the circumstances of observation, various investigators and possible mindsets.

Conclusion: Nine Ways of Talking about Astronomy

In summarising the critical analysis of the topic modelling results, I suggest that during the third quarter of the 18th century *Philosophical Transactions* communicated astronomical topics in at least nine different ways. The actual events surrounding the two significant Transits of Venus and other aspects of astronomy naturally dominated the discussions in the years 1758–1762 and

1768–1772 and, connected to this, the periodical contained 24 topics that were common enough to be meaningful in the text-mining results. I located five ways of talking about astronomy that existed as temporal trends, emphasised only during the transit of 1761 or 1769.

The first two ways were especially common around the 1761 transit. These sub-themes explored *the meanings given to the systematisation of data*: they denoted the practical ways of describing raw data and the systematisation of results (and in fact ended up publishing the long lists of various observation results), or attempted *to reveal a system of the universe by theorising the heavenly bodies*. This reveals the emerging new natural philosophical ways of thinking about the universe.

The next three ways of talking in astronomical topics were particularly common around the 1769 transit. In the 1770s, both the instruments and circumstances of experimentation seem to have been included in a wide way of talking about observation, such as *how the surroundings and weather can influence the observing process*. The third way of talking then points to the lived experience of the field practices of astronomical observation. The fourth astronomical discourse addressed weather and geographical locations *as evidence for the technological advances and greatness of the empire that used them despite the environment and conditions*. The material conditions offered a strong rhetorical explanation to success and failure: honouring and complaining was linked to collaboration between the makers and users of the materials and also to accepted rules of polite communication. These two ways of talking also connect to Steven Shapin's research on the role of actorship in early scientific practice and my data suggests that it would be relevant to further explore the language and style of how one communicated one's research in early modern scientific networks. The fifth discourse linked events to their *tradition* and could be seen to guard the human senses against erroneous input.

The above five ways of talking were temporal trends that existed only during a part of the research period in question. There were also four continuous sub-themes that were observed in the data. According to the statistical appearance of topics, these four discourses were dynamic and their perspective might have changed over time, but nevertheless they continued in some notable form. These aspects were pointed out first and foremost via the digital method, although in retrospect it is easier to see and contextualise them as a part of the data.

The search for *reliable means of observation* was another theme that was characteristic in the entire period around the two Transits of Venus. The sixth astronomical way of talking concerned *the materiality of instruments, their use, and the practice of experiments*. The processes of observing and measuring would in this discourse be linked to trusting the results made through material objects, including the instruments, senses and the human body.

The seventh discourse also concerned the search for reliability and indicates that there continued to be some difference in the ways of making arguments

through *mathematical or physical causal reasoning*. While generally the quest between the causal and mathematical evidence for new knowledge is visible in many topics, the nuances or sub-themes appeared in temporal turns. For example, the topic 'Rules of the skies' lessened towards the end of the studied period, whereas experimentality seems to have become more common. In the two topics that concerned the measuring of distance, the discourse was split into two subsequent practices, with a different focus on the means of obtaining reliable measurements. In other words, this study revealed a temporal variability in what means of observation were seen as reliable.

The eighth perspective suggests that the Royal Society's interaction with its public in the form of its periodical typically proposed some flexibility and openness for developing knowledge, primarily through further experiments. The *openness to probabilities* indicates a structurally designed outlook on knowledge and aligns with the Society's values that everyone should themselves test the reliability of accepted truths through using thorough experiments. In 1758–1762, the astronomical topics tackled reliability issues with algebra and by systematising of results: the way to do this was through collecting large amounts of data and systematising them, which according to the results was a popular method regarding the first transit. At the end of the decade, in 1768–1773, the process of measuring and observing seems to have been more cautious in that it was giving more attention to probabilities, which indicates easier acceptance towards new open-ended scientific hypotheses and continuously *ongoing experimentation*. During the second transit, the tables and the raw observation data had still been published, but now with less importance than before, as the data lists are nowhere near as visible when compared to earlier observations. This means that the mechanical approach to systematise raw data into thematic lists was no longer a remarkable approach. Instead, the natural laws and mathematics were emerging as a more reliable explanation for proving the validity of the experiments.

And, finally, the ninth way reveals the publication of the scientific reports to have been entangled with *accepted social manners*. It illuminates the importance of the language in communicating about one's research.

Regarding the various individual topics, eight were discussed continuously across the whole period and only one of these (the 'Distance measuring process' topic) concerned astronomy. These topics displayed more general talk about the making of scientific knowledge through collecting and observing in the form of various 'plain' descriptions: the travel location ('Travel narrative'), the events and collected specimens ('Events, history, specimens'), events in a polite style ('Events supposing politeness'), 'new' species ('Species') and observation reports ('Reports'), the last two accounting for up to 5% of the temporal change, although usually being less. The descriptions of these 'standard' observations continued to be presented in mostly stable amounts throughout the period. Among the continuous topics, the relationally biggest temporal change from a large amount (6% or 8%) to virtually nothing took place with regard to

the most specific topics (relating to the states of substances and environmental circumstances). These findings about the ways of talking are also supported by the findings of Teich's linguistics team, who noticed, interestingly, that some of the major changes occur for non-thematic topics, as the method brought out 'the hows' of the discourses.³⁶ However, the non-thematic 'hows', by which I mean the values and themes inside the discussions, are the most interesting results. These 'hows' describe the ways of doing or speaking (for example, through polite language or by expressing caution towards various probabilities).

Many results in this study were 'exactly what one would expect'³⁷ according to the research literature. Hence, it was pleasing to begin the analysis with those results as they seemed to confirm that the topic modelling had been carried out correctly. While the results in this way often confirm an earlier hypothesis, the real value of topic modelling is, however, in generating new ways of examining our materials, to deform them.³⁸ I see this as an opportunity to remove all the expected details and to reveal what surprising new findings are left. It is in this way that this study has shown how a well-known but manually too large research material can present different historical insights when interrogated through computational distant readings of digitised sources, thus illuminating the underused methodological opportunities for historians. The power of topic modelling really emerges when we examine change of the thematic trends across the entire text collection. This means that the temporal changes become visible as the topics are listed at intervals in a longer period.³⁹ As the case of the talk of 'distance' demonstrated, the comparison of two topics of the same theme resulted in noticing both a general discourse and located particular themes and also different ways to talk about them.

The temporal change of the paradigm was not seen directly in the results sheet produced by running the data corpus through MALLETT, but became visible when interpreting the results through their contextualisation as a part of the well-known events within the history of science. As a list of words, composed statistically by an algorithm which knows nothing about the context of the study, the keywords per se are not the conclusion to our research questions. Rather, the key is the joint participations of the results and the researchers' insights and contextualisation, which is the approach that produces the hidden underlying meanings of the texts in *Philosophical Transactions*. Examining my familiar sources with this new method has, in other words, been beneficial to help me better understand the big phenomenon of the new emerging practice of science, which is a result that is much more important than just creating a general understanding of how scientists were talking in different ways about Venus.

Appendix 14.1: The Topic Labels and Keywords

Environmental circumstances: observations side air light difference lower years rain observation circumstances acid diameter set table latitude height scale electric force highest

Heavenly bodies: water sun parallax comet eclipse event it's wheel earth green-wich crystals cos axis chance coins square moon's nerves orbit lungs

Astronomical distance: feet limb observations sun foot cos inches velocity star measure est distance meteor modes cum electrified mercury black diapason wheel

Ancient tradition: birds venus transit clear blood ancient quadrant spot satellite vesuvius highest morning density roman horizon dist rest instrument charcoal observatory

Transit and Latin: longitude light contact weight vero transit clouds inter hoc contained servant primarytopic enim tab south city happen sky vapours hujus

Solar system: side quae moon vel stone observation clear degree read fpage earth sed height sin solis corpusbuild letter lpage latitude quod

Rules of the stars: venus ditto clock center moon's system veneris vertical solar miles column atque wood maskelyne mons sun's horizontal egress meridian quidem

Astronomer's weather: ditto air limb wind fluid sun time cloudy amp part clock salt june lowest contact bird matter clouds fixed north

Distance measuring process: distance experiment made experiments glass case great appears give proper state means degrees small method observation general appearance line places

Ancient manuscripts: observations inscription greek appears word fig hours inscriptions pag urine birch cliff emperor lead spindle steam tumor cells things father

Telescope observations: sun's parallax apparent cape telescope pro paris error wire power west etiam rev minutes issn etruscan rest centers april seconds

Travel narrative: made equal water observed body called half surface good kind sur place weather fpage received inches force white ground plate

Events, history, specimens: water large likewise issn days manner strong thing william inches red applied increase place easily larger numbers discovered cut history

Events supposing politeness: time found great day letter part parts matter title amp;c bodies proportion difference present mentioned till vol suppose power fire

Species: number quantity small fig air philosophical heat sun feet appeared considerable fla iron species corpusbuild lpage sea true end long

Reports: part point amp;c greater manner parts sir put end primarytopic years london left observed june substance observe order lightning animal

Hygrometer measurements French: temperature qui une proportion inch conductor form lgs hygrometer mountains luc's cette nous anemones deux open general aug sides temperatures

Thermometer, electricity and effects: heat des thermometer degree electric-ity column surface series instrument effect inches heights birds electrical torpedo boiling hill columns top common

Barometer measurements French: air les quicksilver barometer height tube experiments dans hath pour metal sur point fish ball density expansion shock ces par

States of substances with jargon: motion colour common form earth read kind sir learned bottom stones mercury spirit james measure england head lib miles proved

Notes

- ¹ Moxham 2016: 469.
- ² Byrch 1761; Dunn 1761, *The L&P V*: 102–103.
- ³ See, e.g., the correspondence regarding the transport and arrival of the instruments by Maskelyne in 1760.
- ⁴ See Valle 2006; McDougall-Waters, Moxham & Fyfe 2015; Moxham & Fyfe 2018.
- ⁵ Kermes, Degaetano, Khamis, Knappen & Teich. 2016a.
- ⁶ I wish to express my gratitude to them for generously allowing me to use the data corpus that they had pre-processed in their project ‘Information Density and Scientific Literacy in English: Synchronic and Diachronic Perspectives of the Collaborative Research Centre SFB 1102’.
- ⁷ Kermes, Degaetano, Khamis, Knappen & Teich 2016b: 1929.
- ⁸ Shawn, Weingart & Milligan 2013.
- ⁹ Fankhauser, Knappen & Teich 2016: 498.
- ¹⁰ Their project has already been finished and I have not collaborated in the research, but I was later permitted to use their open data. On another view upon collaboration, Matthew Kirschenbaum described the cooperation as also being ‘a social undertaking. It harbors networks of people who have been working together, sharing research, arguing, competing, and collaborating for many years.’ See Kirschenbaum 2016.
- ¹¹ In addition to the 20 discussed topics (see Appendix 14.1), there were five topics (making 25 topics in total) that were unrelated to astronomical issues, and were therefore excluded from this study. The excluded six topics focused on medical knowledge, nutrition, mortality rates, paleontological finds and geographical events such as earthquakes.
- ¹² In this topic occurs the keywords *fpage*, *corpusbuild* and *lpage*, which derive from terms used for the metadata included in the JStor text files and not from the original articles. This error originates in the pre-processing of the data: had it been noticed earlier in the research process, the corruption of the topic could have been avoided by including these terms in the stop word list. The topic has, however, been accepted as a part of the analysis in its partly erroneous form, as it was not considered to skew it considerably based on the reason that the keywords did not appear in the earliest part of the topic but as the 10th, 16th and 18th in the order of appearance. In topic modelling with MALLETT, the earliest keywords have more significance in the analyses of the topic’s meaning than the later ones in the list.

- ¹³ Maskelyne 1761.
- ¹⁴ My depiction was guided by a similar analysis as in Mohr & Bogdanov 2013: 554.
- ¹⁵ Henry 2017: 168.
- ¹⁶ See Çimen 2018.
- ¹⁷ Topic ‘Transit and Latin’ uses keywords like ‘servant’, which indicates polite letter writing and, in the case of the Royal Society, very probably a set of observations sent to them.
- ¹⁸ Fankhauser, Knappen & Teich 2016: 498.
- ¹⁹ Golinski 2017: 181.
- ²⁰ Jardine 1999: 141.
- ²¹ Halley 1716: 246; Leverington 2003: 143.
- ²² Gascoigne 2000.
- ²³ Maskelyne 1761: 562.
- ²⁴ Turner & Turner 1998: 30.
- ²⁵ Maskelyne 1761: 563.
- ²⁶ Galison & Daston 2008: 316.
- ²⁷ Reid 2008: 172.
- ²⁸ Schaffer 2011: 708.
- ²⁹ Ibid.: 709–710.
- ³⁰ Maskelyne 1761: 560.
- ³¹ Wales 1777: 73.
- ³² Hooke 1663; cf. Daston 2015.
- ³³ Henry 2017: 174.
- ³⁴ Vanek 2016: ii–iii.
- ³⁵ Sargent 2012, 83.
- ³⁶ Fankhauser, Knappen & Teich 2016: 498.
- ³⁷ Blevins 2010.
- ³⁸ Shawn, Milligan & Weingart 2013.
- ³⁹ Blevins 2010.

References

Primary sources

- Byrch, W. D.** (1761). *Of the late transit of Venus, with obsns or help from Mssrs Sic, Ridout and Dean*. The Letters and Papers V, p. 102.
- Dunn, S.** (1761). *Of the late transit of Venus*. 1-page letter to Dr. Morton, read 8 June. Not printed. The Letters and Papers V, p. 103.
- Halley, E.** (1716). A new method of determining the Parallax of the Sun, or his distance from the Earth. *Philosophical Transactions*, XXIX, 243–249.
- Hooke, R.** (1663). *For the making a more accurate history of the changes of weather*. Royal Society, Cl.P/20/24.
- Kermes, H., Degaetano, S., Khamis, A., Knappen, J., & Teich, E.** (2016a). The Royal Society Corpus.

- Maskelyne, N.** (1761). A letter from the Rev. Nevil Maskelyne, M. A. F. R. S. to the Rev. Thomas Birch, D. D. Secretary to the Royal Society: containing the results of observations of the distance of the moon from the Sun and fixed stars, made in a voyage from England to the Island of St. Helena, in order to determine the longitude of the ship, from time to time; together with the whole process of computation used on this occasion. *Philosophical Transactions*, 52, 559–581.
- Wales, W.** (1777). *The original astronomical observations, made in the course of a voyage towards the South Pole, and round the world, in H.M.S. Resolution and Adventure, in the years MDCCLXXII, III, IV, and V.* London: printed by W. and A. Strahan.

Secondary sources

- Blevins, C.** (2010). Topic modeling Martha Ballard's diary. Retrieved from <http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>
- Çimen, Ü.** (2018). On saving the astronomical phenomena: physical realism in struggle with mathematical realism in Francis Bacon, Al-Bitruji, and Averroes. *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, 9(1), 135–151. DOI: <https://doi.org/10.1086/701058>
- Daston, L.** (2015). Supervision: weather watching and table reading in the early modern royal society and Académie Royale des Sciences. *Huntington Library Quarterly*, 78(2), 187–215.
- Fankhauser, P., Knappen, J., & Teich, E.** (2016). *Topical diversification over time in the Royal Society Corpus.* Digital Humanities 2016: Conference Abstracts (pp. 496–500). Kraków: Jagiellonian University and Pedagogical University.
- Galison, P., & Daston, L.** (2008). Scientific coordination as ethos and epistemology. In H. Schramm, L. Schwarte & J. Lazardzig (Eds.), *Theatrum scientiarum—English edition*, Vol. 2: *Instruments in art and science.* Berlin and Boston, MA: De Gruyter.
- Gascoigne, J.** (2000). Exploration, enlightenment and enterprise: the motives for Pacific exploration in the late eighteenth and early nineteenth centuries. Retrieved from <https://web.viu.ca/black/amrc/index.htm?home.htm&2>
- Golinski, J.** (2017). Enlightenment science. In I. R. Morus (Ed.), *The Oxford illustrated history of science 1* (pp. 180–210). Oxford: Oxford University Press.
- Henry, J.** (2017). The scientific revolution. In I. R. Morus (Ed.), *The Oxford illustrated history of science 1* (pp. 143–179). Oxford: Oxford University Press.
- Jardine, L.** (1999). *Ingenious pursuits: building the scientific revolution.* New York, NY: Doubleday.

- Kermes, H., Degaetano, S., Khamis, A., Knappen, J., & Teich, E.** (2016b). *The Royal Society Corpus: from uncharted data to corpus*. Abstract book. Proceedings of the LREC 2016. Portoroz, Slovenia. Retrieved from <http://www.lrec-conf.org/proceedings/lrec2016/summaries/792.html>
- Kirschenbaum, M.** (2016). *What is digital humanities and what's it doing in English departments?* Retrieved from <http://dhdebates.gc.cuny.edu/debates/text/38>
- Leverington, D.** (2003). *Babylon to Voyager and beyond: a history of planetary astronomy*. Cambridge: Cambridge University Press.
- McDougall-Waters, J., Moxham, N., & Fyfe, A.** (2015). *Philosophical Transactions: 350 years of publishing at the Royal Society (1665–2015)*. Retrieved from <https://royalsociety.org/~media/publishing350/publishing350-exhibition-catalogue.pdf>
- Mohr, J. W., & Bogdanov, P.** (2013). Introduction—topic models: what they are and why they matter. *Poetics*, 41(6), 545–569.
- Moxham, N.** (2016). Authors, editors and newsmongers: form and genre in the philosophical transactions under Henry Oldenburg. In J. Raymond & N. Moxham (Eds.), *News networks in early modern Europe* (pp. 445–492). Leiden and Boston, MA: Brill.
- Moxham, N., & Fyfe, A.** (2018). The Royal Society and the prehistory of peer review 1665–1965. *Historical Journal*, 61(4), 863–889.
- Reid, F. L.** (2008). William Wales (ca. 1734–1798): playing the astronomer. *Studies in History and Philosophy of Science*, 39, 170–175.
- Sargent, R.-M.** (2012). From Bacon to Banks: the vision and the realities of pursuing science for the common good. *Studies in History and Philosophy of Science*, 43(1), 82–90.
- Schaffer, S.** (2011). Easily cracked: scientific instruments in states of disrepair. *Isis*, 102(4), 706–717.
- Shawn, G., Milligan, I., & Weingart, S.** (2013). The historian's macroscope: big digital history. An experiment in writing in public, one page at a time. In *The historian's macroscope—working title*. Under contract with Imperial College Press. Open draft version. Retrieved from <http://themascope.org>
- Shawn, G., Weingart, S., & Milligan, I.** (2013). Getting started with topic modeling and MALLETT. In *The historian's macroscope—working title*. Under contract with Imperial College Press. Open draft version. Retrieved from <http://themascope.org>
- Turner, G. L'E., & Turner, A.** *Scientific Instruments, 1500-1900: An Introduction*. University of California Press, 1998.
- Valle, E.** (2006). 'Reporting the doings of the curious': authors and editors in the *Philosophical Transactions* of the Royal Society of London. In B. Nicholas (Ed.), *News discourse in early modern Britain: selected papers of CHINED 2004* (pp. 71–90). Bern: Peter Lang.
- Vanek, M. E.** (2016). *Changeable conditions: British writing about the weather in Canada, 1700–1775* (PhD thesis, University of Toronto).

Topic Modelling German Humanism with DTM and MALLET

Heidi Hakkarainen and Zuhair Iftikhar

Introduction

Topic modelling is often described as a text-mining tool for conducting a study of hidden semantic structures of a text or a text corpus by extracting topics from a document or a collection of documents.¹ Yet, instead of one singular method, there are various tools for topic modelling that can be utilised for historical research. Dynamic topic models, for example, are often constructed temporally year by year, which makes it possible to track and analyse the ways in which topics change over time.² This chapter provides a case example on topic modelling historical primary sources. We are using two tools to carry out topic modelling, MALLET and Dynamic Topic Model (DTM), in one dataset, containing texts from the early 19th-century German-language press which have been subjected to optical character recognition (OCR). All of these texts were discussing humanism, which was a newly emerging concept before mid-century, gaining various meanings in the public discourse before, during and after the 1848–1849 revolutions. Yet, these multiple themes and early interpretations of humanism in the press have been previously under-studied. By

analysing the evolution of the topics between 1829 and 1850, this chapter aims to shed light on the change of the discourse surrounding humanism in the early 19th-century German-speaking Europe.

The concept of humanism (*Humanismus*) was first coined by Friedrich Immanuel Niethammer (1766–1848) in German-speaking Europe in 1808.³ The concept was originally used in the pedagogical debate concerning education, especially in the *Gymnasium*. This pedagogical debate between humanist and philanthropist (realist) education was related to 19th-century educational reforms and especially to the school reform in Bavaria, which preceded the Prussian school reform between 1809 and 1819.⁴ However, in addition to these pedagogical debates, the concept of humanism spread more widely in the 1830s and 1840s, and in this gained new meanings and interpretations.

However, as the previous studies have focused on the early 19th-century pedagogical debates, this wider dissemination and popularisation of the new concept in the printed press has not been under an extensive close study.⁵ There exist a large number of printed publications discussing humanism already in the first part of the 19th century, which makes an inquiry on press debates a challenging task for historians.⁶ In order to tackle this challenge of the vast size of potential source material, this chapter uses the quantitative method of computer-based ‘topic modelling’ to assist the qualitative analysis.

By topic modelling a set containing almost 100 key texts from the years between 1829 and 1850, this chapter recovers several of those multiple discourses connected to humanism before, during and after the outburst of the 1848–1849 revolutions. By combining and comparing topic modelling with MALLETT with Dynamic Topic Modelling (DTM), this chapter seeks to map and analyse what kinds of topics were related to humanism before 1850 and how these topics changed and evolved over time. During 21 years, humanism appeared in various contexts from education to philosophy, religion and politics. Where the MALLETT, as the most well-established topic modelling tool within the field of digital history,⁷ is used in detecting the most prominent themes in the discussion on humanism, DTM makes available a finer look into the topics at the temporal level and, in this case study, provides a new kind of insight into the growing importance of temporality within the German-language humanism discourse between the early 1800s and the mid-century.⁸

In contrast to temporally ambitious research on huge corpuses, this chapter focuses on a rather small text corpus, which allows more exploration of the possibilities of cross-reading the material with methods of close and distant reading. This study of the discussion surrounding humanism before 1850 thus provides a reasonably manageable but rich investigation of some of the ways in which newspapers and periodicals addressed topical issues and transferred concepts and new ideas across political borders within the lands of the German Confederation.

At the same time, we seek to explore what kinds of methodological benefits and risks are involved in the topic modelling of historical sources. The technique of topic modelling decides what constitutes a topic on an algorithm

that creates a statistical model of word clusters. It is thus not a fixed schema, but a variable probabilistic model that should also be treated as such. We will demonstrate how various forms of cleaning and filtering of the data can have drastic results on the output of the topic model. We also present and compare outputs from different methods of topic modelling, using the MALLET application and DTM, and address various methodological concerns related to topic modelling.

Topic Modelling with MALLET

The first essential step in describing the 19th-century German-language press discourse on humanism was to identify its various individual themes or topics using the quantitative method of topic modelling. Topic modelling has its roots in information retrieval, natural language processing and machine learning. This probabilistic tool has attracted attention among historians, because it enables detecting underlying thematic structures behind a large corpus of documents, as well as surprising connections between individual texts. A topic comprises a distribution of words. A single document is assumed to contain words about multiple topics within the whole dataset. Each word is drawn from those dataset topics. The study used two topic modelling tools where the first is called MALLET (Machine Learning for Language Toolkit, version 2.0.8.), which is an open source Java-based software package for natural language processing using Latent Dirichlet Allocation technique (LDA).⁹

Before using MALLET and in pre-preparation, the machined encoded OCR German-language press texts were cleaned and corrected manually (especially the recurring problem with some Unicode characters). In some cases, this included shortening the texts by excluding clearly irrelevant sections.

The model was then made with the 'optimise-interval' command, which sets each topic's probabilistic Dirichlet parameter that indicates the topic's proportion in the whole dataset, and gives a better fit to the data by allowing some topics to be more prominent. In addition, the number of topics to be identified by MALLET is set beforehand as there is no 'natural' number of topics in a corpus, but this part requires manual evaluation and iteration by the researchers.¹⁰ Both MALLET and the DTM tool only mechanically detect topics and assign them numeric values, whereas identifying and naming the topics (that is, determining and labelling the thematic categories found by the machine reading) is something the human researcher has to carry out using manual reading. And this is an act of interpretation.

Topic Modelling with DTM

Within probabilistic topic modelling, LDA is a frequently used technique and its MALLET implementation has traditionally been the most popular tool to analyse historical corporuses. Ever since topic modelling was first introduced in the early 2000s, there have been new extensions that help to model temporal

relationships. One shortcoming of the LDA method is that it assumes that the order of documents is irrelevant. But if we – as historians are often prone to – want to discover the evolution of topics over time, then we have to take the time sequence into account. DTM attempts to overcome this shortcoming and captures the dynamics of how topics emerge and change over time.¹¹

DTM is designed to explicitly model the ways in which topics evolve over time and to give qualitative insights into the changing composition of the source material. However, it is not the only such tool available and it has also been subjected to critique for penalising large changes from year to year.¹² The DTM is a probabilistic time series model, which is designed to track and analyse the ways in which latent topics change over time within a large set of documents. For example, David M. Blei and John D. Lafferty demonstrated the functioning of DTM by investigating topics of the journal *Science* between 1880 and 2000.¹³ Our case study is based on a small source corpus, which, as we will soon see, was one important factor in the output from the dynamic topic modelling. Because of the small size of the dataset, cleaning and filtering the data had a major impact on DTM's output. The more historical sources were pre-processed, the more stable the model became.

As mentioned above, few but not all text files were reviewed for common mistakes and in a few instances some mistakes were manually corrected. Python's Natural Language Toolkit (nltk) library was used for the pre-processing and filtering of the texts. Prior to passing on the text data to DTM tools, the text was processed using the following pre-processing pipeline:

1. **Punctuation and numbers removal.** Punctuation characters within and around all the words were deleted and all the other characters except alphabetic characters were removed.
2. **Stop words removal.** This is a common operation when processing text in any domain. The list of German stop words was initially taken from the nltk library and MALLET tool. This list was extended by reviewing the texts and some words deemed to be useless were then added to the list. Any words in the stop words list were removed in pre-processing.
3. **Stemming and lemmatising.** Stemming is the process by which a word is reduced to its base form and all the inflectional forms of a word are reduced to a single base stem. Using language dictionaries, lemmatisation converts a word to its base lemma. This is the word from which all the inflectional forms are derived. The base stem is then used by the lemmatisers to find the base lemma, which is then kept in the text.
4. **Classification.** The words were then classified into different parts-of-speech with the goal being to keep various nouns and verbs identified in the input texts. Words which belong to other parts-of-speech were removed.
5. **Rare words.** As a final step, the words which appear only once in the whole input corpus were also removed.

We then used Gensim (Python library) to run the DTM tool. After creating various outputs of models with 5 to 20 topics, as for the previous analysis using MALLET, we decided to limit the number of topics to 10. Like MALLET, DTM also gives keywords (that is, a cluster of words relevant to the topic), which help to identify the topic.

Source Material

The source material used in this study is a sub-dataset from the digital corpus Austrian Newspapers Online (ANNO), provided by the Austrian National Library (at <http://anno.onb.ac.at>). The digital ANNO collection contains around 20 million pages of German-speaking newspapers and periodicals that are available for full text searches.¹⁴ The Austrian National Library at their ANNO-portal provides an OCR tool for machine encoded optically recognised text which, although not totally reliable and contains errors, can be used for the digital analysis.

According to the full text search engine of the ANNO portal, the word *Humanismus* (humanism) was mentioned 326 times in the press between 1808 and 1850.¹⁵ Because the old German *Fraktur* typeface is challenging for OCR, the results should not be interpreted as entirely reliable, but as giving an indication of the scale of use, how much this word was circulated in the press. In some texts, humanism appeared only once in passing, while in others it was mentioned several times and discussed explicitly. Based on their relevance, length and readability, we have selected 95 key texts for topic modelling analysis (see Appendix 15.1). These texts include book reviews, articles, news, feuilleton writings and political reports, while reprints, short notices, adverts and obituaries have been excluded.

Figure 15.1 illustrates the publishing centres and various publications that make up the dataset. The graph is made with the Gephi visualisation application and it aims to depict the source material in a visually conceivable way. Moreover, Gephi is a frequently used software tool for network analysis, because it enables the portrayal and analysis of relationships or interaction between persons, entities and objects, such as geographical places or publications.¹⁶ The objects (nodes) and their relationships (edges) can be presented in many different ways. In this case, the layout was made manually instead of choosing one of the most popular layout algorithms such as Force Atlas or Fruchterman Reingold. The nodes and edges tables were imported to Gephi as CSV files and in the edges table the connection between a publication and its place of publishing gained 'weight' in accordance to the amount of texts discussing humanism in that particular publication during the period between 1829 and 1850. The more humanism was mentioned, the thicker the line between a newspaper or a magazine and the city in which it was published. Accordingly, the strength of connections indicates which were the most important publishing centres

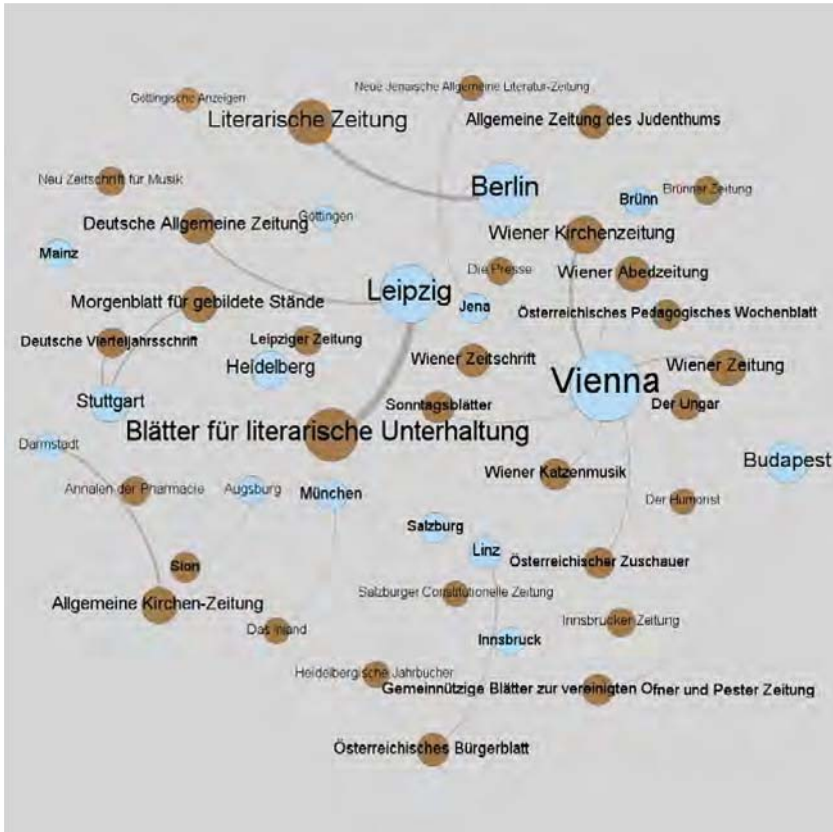


Figure 15.1: Network diagram of publications and publishing centres of the dataset. Source: Authors.

and highlights publications that most extensively dealt with humanism in this dataset. Even though the ANNO source corpus is partial and dominated by Austrian newspapers and magazines, Figure 15.1 shows that the early 19th-century discussion on humanism surpassed political borders within the German Confederation spreading in the area of fragmented German lands and German-speaking parts of the Habsburg Empire. Vienna, Leipzig and Berlin were the most important publishing centres and the literary journals *Blätter für literarische Unterhaltung* and *Literarische Zeitung* dealt with the topic most extensively, although the publications dealing with humanism ranged from daily newspapers to religious magazines and satirical journals.

German Humanism According to MALLET Topic Modelling

Initial details about MALLET are summarised in the previous section. Below are the eight topics in order of prevalence with their top words as discovered

by MALLET when asked to determine the 10 most prevalent topics and as labelled (education, reformation, etc.) by us. The number of topics was chosen after experimenting with different kinds of models and 10 topics were chosen as a best way for modelling the source corpus, which was small and fragmented. Topic modelling usually involves filtering away so-called stop words, non-informative frequently appearing words such as articles, particles and pronouns. However, especially when it comes to creating a model with a small number of topics, pre-processing the data has a danger of compromising the results as the researcher makes decisions on removal of stop words according to her or his pre-understanding, thus projecting into the data certain presuppositions regarding what is important in the corpus.¹⁷ Accordingly, in this model, no pre-filtering of stop words was carried out before the analysis, but two topics that contained only stop words were filtered out after creating the model. See Appendix 15.2 for the whole model.

Religion: fich menschen gott religion find juden zukunft religiösen gottes humanismus mensch christenthum christliche niht darum demokratie humanität christlichen christen theorie

Education: erziehung schulen lehrer sprache bildung seyn gymnasien unterricht realismus sprachen realschulen schüler jugend individuum wissenshaften anstalten schrift realschule

Revolution: wurde freiheit volk stadt wurden berlin revolution kammer bald volkes völker waren heute republik straßen preußen fast macht bürgerwehr haufen

Philosophy: fich philosophie ruge find nationalismus princip paris jahrbücher literatur preußen geschrieben briefe socialismus anficthen brief patriotismus rage artikel principien staatsanwalt

Reformation: kirche fich universitäten luther reformation staat lehre staatsreform gemeinden schottischen glaubens bloß kirchen verfassung staate theologen lehrer wissenschaft hervor

Death penalty: todesstrafe sei abg verbrechen strafe habe amendement antrag könne man dieß gesetze redner verbrecher abgeschafft jury wolle abschaffung angenommen gegen

Press debate: dafs christlichen philologie gegner muss zeitung liberalismus sache sinne bedeutung gesinnung jedenfalls artikel presse giebt philologen meinung klassischenmonarchischen christliche

Social issues: fie the fich hamburg euch gesehen zigeuner habt bey ift diese wissen feine sprachen stadt armen glück schüler jhr their

The output from MALLET provides eight topics with different keywords. In the 'Education' topic, words like *Erziehung* (education/upbringing), *Schulen* (schools), *Lehrer* (teacher) and *Sprache* (language) are clustered together with such difficult-to-translate German concepts like *Bildung* and *Gymnasien*, which indicate that this topic is related to the educational debates about the role of humanism in the modern schooling system that were a very important issue in the era of comprehensive school reforms. After all, the concept of

Humanismus (humanism) was, as mentioned above, first coined as a pedagogical concept, fostering classical education and the study of classical languages.¹⁸ The 'Reformation' topic, on the other hand, contains words like *Kirche* (church), *Universitäten* (universities), *Luther* (Luther) and *Reformation* (reformation), which give reason to believe that this topic deals with humanism historically in relation to Martin Luther and the reformation era.

However, in addition to these highly obvious and clear results, there are also topical word clusters which show a completely different kind of interpretation of humanism. The topic 'Philosophy', for instance, contains words like *Philosophie* (philosophy), *Ruge* (Ruge), *Nationalismus* (nationalism), *Princip* (principle) and *Paris* (Paris). All of these words are connected to the philosopher Arnold Ruge (1802–1880), who was also a political writer, associated with the Young Hegelians and Karl Marx, and known for his radical ideas that religion should be separated from politics and intellectual thinking. Ruge was one of the main figures who in the 1840s introduced a new interpretation of humanism as a political concept and his ideas were highly debated in the press.¹⁹ For Ruge, humanism meant political emancipation from the old *ancien régime*. He incorporated humanism in democratic-republican ideology, which combined social critique with critique towards religion and growing nationalism. Humanism meant political, religious and social freedom, which was universal for the whole of mankind and superseded national borders. Accordingly, in *Geschichtliche Grundbegriffe*, Ruge's interpretation of humanism is called *kosmopolitischer Humanismus* (cosmopolitan humanism).²⁰

This radical new political meaning of the concept of humanism is also visible in topics that dealt with social problems and political issues like the death sentence and the 1848–1849 revolution. For example, the topic labelled 'Social issues' contains keywords like *Zigouner* (gypsies), *Armen* (the poor), *Stadt* (city) and *Glück* (happiness). Again, the topic 'Death penalty' is clustering together words like *Todesstrafe* (death penalty), *Verbrechen* (felony), *Strafe* (punishment) and *Amendement* (amendment), which are all related to the debates around abolishment of the death penalty, which was a topical issue especially in Austria around 1849. Moreover, topic modelling of the dataset reveals a topic relating explicitly to the European revolutions in 1848–1849. This topic labelled with the title 'Revolution' contains the following keywords: *wurde* (came), *Freiheit* (freedom), *Volk* (people), *Stadt* (city), *Berlin* (Berlin), *Revolution* (revolution), *bald* (soon), *heute* (today), *Republik* (republic) *Straßen* (streets), *Macht* (power), *Bürgerwehr* (militia) and *Haufen* (pile). This topic, especially, indicates how humanism became a political concept in the 1840s when both early socialists and liberals adopted humanism in their political language as they demanded political emancipation from the old regime.²¹

This result demonstrates the diversity of the meanings given to humanism in the early 19th-century press. In addition to educational debates, humanism also appeared in the discussions surrounding social and moral issues, law and politics. In fact, the extremely diverse topics of humanism indicate a pervasive

reorganising of ideas related to the human being and his or her place in the universe in the post-Napoleonic era, in which the liberal bourgeoisie was gaining a new foothold in society at the same time that the Church and absolutist power were challenged in the aftermath of the French Revolution. This transformative era created new interpretations on how politics, religion, education and philosophical thinking should be organised in modern secularising society, and, despite the practices of censorship especially in Prussia and Austria,²² the press played a major role in circulating these ideas among a growing readership.

Consequently, the vast processes of secularisation and modernisation help us to understand why the 'Religion' topic was the most dominant theme in the early 19th-century press discussion on humanism. This most prevalent topic contains many interesting keywords indicating how discourses surrounding religion, morality and politics were actually significantly entangled in the early 19th-century discussion on humanism. The clustering of words like *Menschen* (human being), *Gott* (god), *Religion* (religion), *Juden* (Jews), *Zukunft* (future), *Humanismus* (humanism), *Christenthum* (Christianity), *Demokratie* (democracy), *Humanität* (humanity) and *Theorie* (theory) is a good example of the interpretative challenges that take place when identifying and labelling topics that are not cohesive but multifaceted and extremely complex. We will examine the 'Religion' topic closer below using DTM. But first, we will locate which years this topic emerged most dominantly between 1829 and 1850.

Following the task of identifying topics, it is vital to also explore them and their meanings in the historical context in which they came to life. In other words, it is essential to acknowledge the temporality of the topics and study them from in a dynamic historical perspective. For example, the volume of the press was very different in 1829 and in 1850. Furthermore, the new Young Hegelian philosophical ideas and growing interest in social issues was part of the intellectual and social landscape of the 1840s and it goes without saying that the outbreak of the revolutions in 1848 was clearly a major historical event that impacted on the public discourse surrounding humanism.

Without additional programming, MALLET does not present the topics in relation to time. Yet, it is possible to inspect the dynamic temporal aspect of the topics by organising the dataset chronologically.²³ Accordingly, the files of the dataset were numbered from the oldest, in this case 1829, to the youngest, here 1850. This means that it is now possible to study how topics emerged and changed over time (Figure 15.2). In Figure 15.2, the two stop word topics are filtered out, presenting only the eight relevant topics.

We can now see the thematic trends and how the topic patterns change over time. The figure above indicates that before 1840 'Education' and 'Social issues' were important topics in relation to humanism, but in 1848 the topic 'Revolution' became dominant. In 1849, it was replaced as the leading topic by 'Death penalty', with 'Religion' following in prevalence. The 'Religion' topic gained importance especially immediately after the revolution, which could indicate a reaction to the turbulence and violence in 1848–1849. Yet, despite

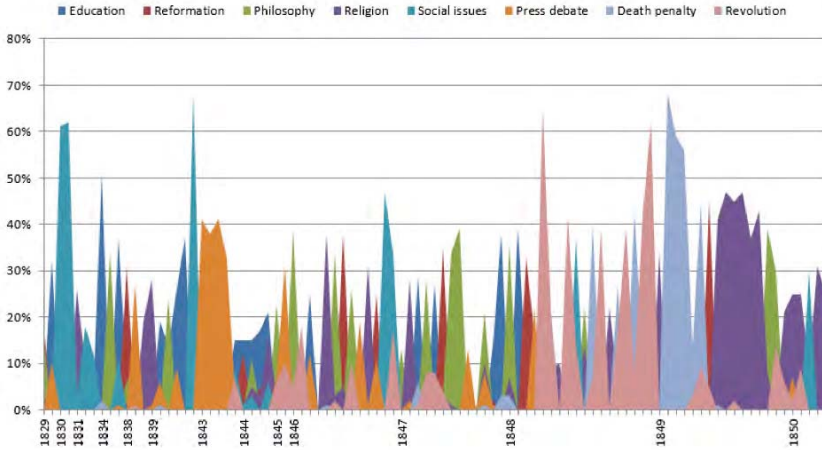


Figure 15.2: Annual allocation of topics. Source: Authors.

the chronological aspect, MALLET's results are always compressed and cannot give any further insight into the dynamics within the topics that have been discovered. In the next section, we will further analyse how dynamic topic modeling (DTM) can make it possible to gain insight about the dynamics within one singular topic.

Discovering Temporalisation of German Humanism with DTM

Preliminary details about DTM and text pre-processing details are mentioned above. The cleaning of the data had a major impact on the output of the model. At first, the results were very similar to the MALLET analysis and many topics seen before persisted. For example, humanism continued to emerge in relation to the topic of 'Religion' ['menschen', 'humanismus', 'humanität', 'zukunft', 'ste', 'stch', 'religion', 'wahrheit', 'bloß', 'demokratie', 'christenthum', 'recht', 'gegenwart', 'fich', 'wohl']. Also, the topics 'Education' ['bildung', 'mehr', 'erziehung', 'zeit', 'lehrer', 'jugend', 'realschulen', 'find', 'wissenschaft', 'sache', 'neuen', 'immer', 'gymnasien', 'zweck', 'mittel'], 'Death penalty' ['sei', 'verbrechen', 'abg', 'könne', 'redner', 'schon', 'antrag', 'amendement', 'angenommen', 'staat', 'dieß', 'abgeschafft', 'ab', 'be', 'abschaffung'] and 'Revolution' ['wurde', 'macht', 'völker', 'volk', 'geschichte', 'bald', 'freiheit', 'berlin', 'volkes', 'revolution', 'werk', 'wurden', 'je', 'regierung', 'tage'] were remarkably similar. However, there were also changes. Social issues and debates around Ruge's interpretation of humanism were more in the background and there was more than one category relating to religion and education.

Furthermore, with DTM, we had more fine-tuned results as the source corpus was divided into different time frames and keywords were arranged year by year. As the keywords appeared in a list from most important to least

1829	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1830	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1831	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1832	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1833	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1834	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1835	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1836	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1837	'menschen,'	'humanismus,'	'humanität,'	'zukunft'
1838	'menschen,'	'humanismus,'	'zukunft,'	'humanität'
1839	'menschen,'	'humanismus,'	'zukunft,'	'humanität'
1840	'menschen,'	'humanismus,'	'zukunft,'	'humanität'
1841	'menschen,'	'zukunft,'	'humanismus,'	'humanität'
1842	'menschen,'	'zukunft,'	'humanität,'	'humanismus'
1843	'menschen,'	'zukunft,'	'humanität,'	'humanismus'
1844	'menschen,'	'zukunft,'	'humanität,'	'humanismus'
1845	'menschen,'	'zukunft,'	'humanität,'	'humanismus'
1846	'menschen,'	'zukunft,'	'humanität,'	'humanismus'
1847	'zukunft,'	'menschen,'	'humanität,'	'humanismus'
1848	'zukunft,'	'menschen,'	'humanismus,'	'humanität'
1849	'zukunft,'	'menschen,'	'humanismus,'	'humanität'
1850	'zukunft,'	'menschen,'	'humanismus,'	'humanität'

Figure 15.3: Output from the DTM before data cleaning, including the four first keywords. Source: Authors.

important, it was possible to detect the ways in which the order of these keywords changed within one singular topic. The most striking new discovery with DTM was that there were cases in which words with temporal meaning such as *Zeit* (time) or *Zukunft* (future) became increasingly important towards mid-century. This discovery resonates strongly with the conceptual historian Reinhart Koselleck's famous argument that the early 19th century was a *Sattelzeit*, a period in which the notion of time changed radically and concepts became increasingly abstract and more future-oriented. Koselleck suggested that as modern concepts became more entangled with historical time, being associated increasingly with the past, the present and the future, the phenomena which previously were seen as static and unchanging became conceived as dynamic processes.²⁴

To give an example, in Figure 15.3 we have the four most important words for the topic 'Religion', containing words like *Menschen* (human being), *Humanismus* (humanism), *Zukunft* (future), *Humanität* (humanity), *Religion* (religion), *Wahrheit* (truth), *Demokratie* (democracy), *Christentum* (Christianity), *Recht* (justice) and *Gegenwart* (present), which are very similar to those words seen in the most prevalent 'Religion' topic in the MALLETT results.

However, here the topic seems to be relating more to human beings and morality rather than religion. In addition, the meaning of the word *Zukunft* (future) is of special interest here, as its position changes radically between

1829	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1830	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1831	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1832	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1833	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1834	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1835	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1836	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1837	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1838	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1839	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1840	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1841	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1842	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1843	'menschen,	'gott,	'religion,	'humanismus,	'zukunft,
1844	'menschen,	'gott,	'religion,	'zukunft,	'humanismus,
1845	'menschen,	'gott,	'religion,	'zukunft,	'humanismus,
1846	'menschen,	'gott,	'religion,	'zukunft,	'humanismus,
1847	'menschen,	'gott,	'religion,	'zukunft,	'humanismus,
1848	'menschen,	'gott,	'religion,	'zukunft,	'humanismus,
1849	'menschen,	'gott,	'religion,	'zukunft,	'humanismus,
1850	'menschen,	'gott,	'religion,	'zukunft,	'humanismus,

Figure 15.4: Output from the DTM after data cleaning, including the five first keywords. Source: Authors.

1829 and 1850. Figure 15.3 shows the output from the DTM before data cleaning, including the four first keywords. In post-cleaning, the letters 'ste' were filtered out.

However, this striking change did not appear in all the outputs, but the more we removed stop words and filtered the data for better results, the more stable the topic appeared (Figure 15.4). In addition, the word God (*Gott*), which is missing in the first output together with religion (*Religion*), is now continuously the second most important word after human being (*Mensch*). The information about the proposition of each word within the topic indicates that changes were so minor that altering the script by removing stop words and removing words that appeared only once changed and stabilised the model to the extent that changes could no longer be seen in the order of the keywords.²⁵

Yet, to give another example, the word *Zeit* (time) became increasingly important in another topic that included keywords such as *Wissenschaft* (science/knowledge) and *Erziehung* (education/upbringing). The change is visible both before and after filtering stop words. The Dirichlet parameter indicates that the weight of the word *Zeit* did not increase, but the growing importance resulted from the fact that the importance of the word *Wissenschaft* decreased radically around 1846.²⁶ This was a modest change, but it persisted in the outputs made before and after removing the stop words and carrying out other data filtering, such as removing words that appeared only once.

1829–1850
Philology
Church history
Philosophical tendencies
Revolution & Distribution of power
Political debate
Philosophy of science
Education of the Jews
Education
Religion, Morality & Relationship to God
Study of languages

Figure 15.5: Topics detected by the DTM tool after data cleaning. Source: Authors.

In the end, after data cleaning and filtering the historical sources, the DTM tool provided a list of the 10 most prevalent topics in the early 19th-century press (Figure 15.5). Yet, because of the short timeline and small size of the source corpus, the final output provided very static results and only very small changes within these topics were able to be discovered.

However, it is important to bear in mind that the dataset used in this case study was small. A larger dataset together with a potentially longer timeline would probably make it possible to detect and analyse more drastic changes over time. In any case, both of these examples illustrate that topic models are first and foremost probabilistic models providing estimates of the most salient discourse topics. Semantic changes are related to probabilistic proportional changes (in topic word list) and examining the probability distribution parameters (values associated with topic words in the output) is vital for understanding how these models work in practice.

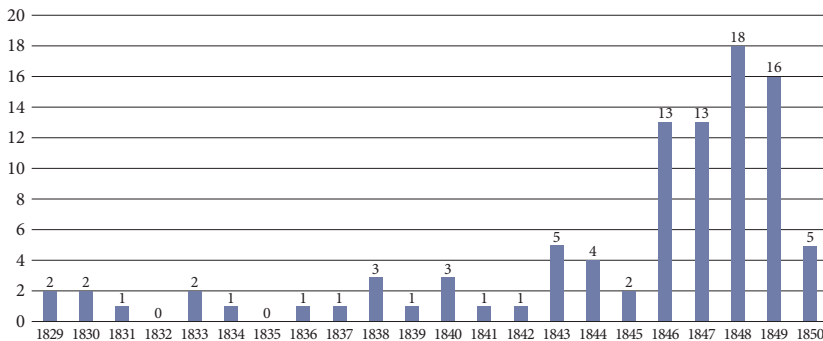
Conclusions

This study has investigated the early 19th-century German press discourse on humanism, which has been an under-researched area to date. In this chapter, we have modelled the topics of humanism in the early 19th-century German-language press with MALLETT and DTM. By analysing the evolution of the topics between 1829 and 1850, this chapter has explored the change of the discourse surrounding humanism in early 19th-century German-speaking Europe. Both topic modelling applications detected different topics among the text corpus and recognised different semantic categories in the early 19th-century German-language source material without any understanding of the substance or context of these texts.

Topic modelling contains various methods, which can be used for different purposes. As we have shown, topic modelling can provide assistance for historical research as a tool for analysis and interpretation. In this study, we created different topic models of a dataset that was relatively small and could be closely read in addition to distant reading. Both MALLET and the DTM tool not only enable us to identify thematic categories (that is, topics within the dataset), but they also make it possible to trace these topics back to file level. The outputs produced detailed results on how each topic appeared in each of the 95 texts of the dataset, which makes it possible to trace topics back to the level of individual articles for close reading analysis. If one is especially interested in, say, ‘Revolution’ as a press topic, one could select and read all the news articles and other texts in which this topic appeared during the time frame of 1829 to 1850. This kind of assistance is invaluable for mapping and assessing sources, which is often laborious and time-consuming.

At the same time, our study also sheds light on the potential benefits and risks of topic modelling within historical research. From a methodological perspective, it is important to bear in mind that although topic modelling might produce highly compelling results, the analysis of these results demands time, skills and caution. One has to remember that results can vary depending upon the input topic number, size of the dataset, specific tool used for topic modelling, data cleaning and methods of filtering. Topic modelling provides assistance for historical research as a tool for analysis and interpretation, but the output of a topic modelling process is not a result in itself and needs to be studied further for reliable conclusions. Topic modelling results can answer a historian’s intuitive questions by providing focus and direction to the analysis of historical corpuses through traditional methods of historical inquiry, source criticism, close reading and contextualisation. Perhaps even more importantly, topic modelling has the potential to challenge established patterns of thought and underlying presumptions by providing a completely different angle on historical sources.

Appendix 15.1: Dataset



Appendix 15.2: MALLET Tool

Topic ID	Topic label	Dirichlet parameter	Keywords
0	[EDUCATION]	0,12609	erziehung schulen lehrer sprache bildung seyn gymnasien unterricht realismus sprachen realschulen schüler jugend individuum wissenschaften anstalten schrift realschule
1	[REFORMATION]	0,07934	kirche fich universitäten luther reformation staat lehre staats reform gemeinden schottischen glaubens bloß kirchen verfassung staate theologen lehrer wissenschaft hervor
2	[PHILOSOPHY]	0,09784	fich philosophie ruge find nationalismus princip paris jahrbücher literatur preußen geschrieben briefe socialis- mus anfichten brief patriotismus rage artikel principien staatsanwalt
3	[RELIGION]	0,18337	fich menschen gott religion find juden zukunft religiösen gottes humanismus mensch christenthum christliche niht darum demokratie humanität christlichen christen theorie
4	[SOCIAL ISSUES]	0,05893	fie the fich hamburg euch gesehen zigeuner habt bey ift diefe wiffen feine sprachen stadt armen glück schüler jhr their
5	[STOP WORDS]	1,43466	mit aber nur man hat noch diese zeit welche haben mehr gegen denn selbst uns alle ohne ihm sondern leben
6	[PRESS DEBATE / CONTROVERSY]	0,06722	dafs christlichen philologie gegner muss zeitung liberalismus sache sinne bedeutung gesinnung jedenfalls artikel presse giebt philologen meinung klassischen monarchischen christliche
7	[LAW / DEATH PENALTY]	0,07172	todesstrafe sei abg verbrechen strafe habe amendement antrag könne man dieß gesetze redner verbrecher abgeschafft jury wolle abschaffung angenommen gegen
8	[REVOLUTION]	0,11352	wurde freiheit volk stadt wurden berlin revolution kammer bald volkes völker waren heute republik straßen preußen fast macht bürgerwehr haufen
9	[STOP WORDS]	0,07725	fich ift find feine fein diefe diefer feiner fei fondern nnd felbfi ihm nichts zwifchen diefem fchon lehre fehr wol

Appendix 15.3: DTM Tool

10 topics found by DTM after data cleaning:

Philology: philologie alterthums wissenschaft studien zeit bedeutung artikel klassischen richtung gymnasien weise zeigt damals gewinnen darauf

Church history: kirche deutschen humanismus ganz zeit Deutschland große ruge macht geschichte staat bald reformation freiheit princip hätte jahrhunderts

Philosophical tendencies: tendenz so deutfchen bey welt bildet herr wissen vermögen gerade briefe feuerbach menfchen diese

Revolution & Distribution of power: schon geht volk bleibt freiheit berlin verbrechen hand viele ersten fall davon gut

Political debate: mittel freien entwicklung liberalismus gemacht monarchischen indessen ganz bedeutung zeit glaubens weder regierung

Philosophy of science: geist denen idee lebens welt einzelnen vielmehr leben einzelnen philosophie schule recht staat partei

Education of the Jews: schon juden wenig sinne schule allgemeinen mag allerdings beziehung irgend sagen christlichen öffentlichen wenigstens

Education: bildung zeit erziehung schon humanismus schüler leben lehrer immer seyn kraft besonders deutsche ganz wohl allein aufgabe

Religion, Morality & Relationship to God: got zukunft humanität menschlichenleben wahre christenthum mensch freiheit demokratie wahrheit religiösen sagen gewalt welt politik

Study of languages: sprache sprachen zeit realschulen gesehen welt bildung amburg jugend schulen find gymnasien erfahrung habt werke neuen element

Notes

¹ See further Erling 2014: esp. 58–59, and Jacobi, Atteveldt & Welbers 2015.

² Blei & Lafferty 2006.

³ The concept of *Humanismus* was coined in 1808 when Niethammer used it in his book *Der Streit des Philanthropinismus und Humanismus in der Theorie des Erziehungs-Unterrichts unsrer Zeit*. However, the tradition of German humanism dates back to the 15th and 16th centuries, when the ideas of Italian renaissance humanism spread across Europe. Accordingly, such concepts as *humanitas* and *studia humanitatis* are much older origin, dating back to antiquity.

⁴ See further Bollenbeck 1994: 142–155.

⁵ In addition to Georg Bollenbeck's book, the most extensive studies discussing 19th-century German humanism are by Landfester 1988 and van Bommel 2015.

⁶ The book industry and the press were both growing in volume in the first part of the century, expanding even more dramatically as the 19th century neared its close. See further Erling & Tatlock 2014. Cf. St. Clair 2004.

- ⁷ See further Brauer & Fridlund 2013: 159.
- ⁸ Cf. Steinmetz & Freedden 2017: 2, 5.
- ⁹ LDA (Latent Dirichlet allocation) was developed by David Blei and others in 2003 and MALLET (MACHINE Learning for Language Toolkit) was written by Andrew McCallum. For more information, see <http://mallet.cs.umass.edu/about.php>. See also *The Programming Historian* tutorial on MALLET, Graham, Weingart & Milligan 2012.
- ¹⁰ The model provides as output three different files: topic ‘state’ assigning each word in the text to a topic, ‘topic keys’ consisting of the top words for each identified topic and the topic ‘composition’ consisting of allocation of percentage of every topic in each of the 95 files that were included in the analysis.
- ¹¹ DTM, Blei & Lafferty 2006.
- ¹² Hall, Jurafsky & Manning 2008: 364.
- ¹³ Blei & Lafferty 2006.
- ¹⁴ See ANNO webpage.
- ¹⁵ See further, Hakkarainen 2020: 27–28.
- ¹⁶ See further, e.g., Cordell 2015.
- ¹⁷ Cf. Schonfield, Magnusson & Mimno 2017.
- ¹⁸ See further Bollenbeck 1994: 142–155; van Bommel 2015.
- ¹⁹ See, e.g., Anon. 1846; Anon. 1 & 4 August 1847; Anon. 1848.
- ²⁰ Bödeker 1982: 1123–1124.
- ²¹ Ibid.:1121–1126. See also Hansson 1999: 77–106.
- ²² See further Dussel 2011: 25–34; Stöber 2014: 141–142.
- ²³ Cf. Blevins 2010.
- ²⁴ Koselleck 1985. See also Steinmetz & Freedden 2017: 2, 5.
- ²⁵ However, even after filtering there was a minor increase in the percentage. In 1829, the proportional number for the word ‘Zukunft’ was 0.010673; in 1850, it was 0.016537.
- ²⁶ In 1829, the proportional number for the word ‘Wissenschaft’ was 0.01164-2238616473554; in 1850, it was 0.006096759758556382.

References

- Anon.** (1846, 7 January). Arnold Ruge und sein neuester Standpunkt. *Blätter für literarische Unterhaltung*.
- Anon.** (1847, 1 August). Arnold Ruge: Politischer Bilder aus der Zeit. *Blätter für literarische Unterhaltung*.
- Anon.** (1847, 4 August). Arnold Ruge: Politischer Bilder aus der Zeit. *Blätter für literarische Unterhaltung*.
- Anon.** (1848, 11 January). Polemische Briefe von Arnold Ruge: Reihe von ‘Offenen Briefen zur Vertheidigung des Humanismus’. *Blätter für literarische Unterhaltung*.
- ANNO** – Austrian Newspapers Online (Austrian National Library). Retrieved from <http://anno.onb.ac.at/faq.htm>

- Blei, D. M., & Lafferty, J. D.** (2006, 25–29 June). *Dynamic Topic Models*. Paper presented at the ICML '06 Proceedings of the 23rd international conference on machine learning (pp. 113–120). Pittsburgh. Retrieved from https://mimno.infosci.cornell.edu/info6150/readings/dynamic_topic_models.pdf
- Blevins, C.** (2010). *Topic modeling Martha Ballard's diary*. Retrieved from <https://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>
- Bollenbeck, G.** (1994). *Bildung und Kultur: Glanz und Elend eines deutschen Deutungsmusters*. Frankfurt am Main: Insel.
- Brauer, R., & Fridlund, M.** (2013). Historizing topic models: a distant reading of topic modeling texts within historical studies. In L. V. Nikiforova & N. V. Nikiforova (Eds.), *Cultural research in the context of 'digital humanities': proceedings of international conference 3–5 October 2013, St Petersburg* (pp. 152–163). St. Petersburg: Herzen State Pedagogical University and Publishing House Asterion. Retrieved from <https://matsfridlund.files.wordpress.com/2014/04/publ2013brauerfridlundconf.pdf>
- Bödeker, H. E.** (1982). Menschheit, Humanismus, Humanität. In O. Brunner, W. Conze & R. Koselleck (Eds.), *Geschichtliche Grundbegriffe: Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*, Vol. 3: *H-Me* (pp. 1063–1128). Stuttgart: Klett-Cotta.
- Cordell, R.** (2015). Reprinting, circulation, and the network author in antebellum newspapers. *American Literary History*, 3(27), 417–445. DOI: <https://doi.org/10.1093/alh/ajv028>
- Dussel, K.** (2011). *Deutsche Tagespresse im 19. und 20. Jahrhundert*. 2nd edn. Berlin: Lit Verlag.
- Erling, M.** (2014). The location of literary history: topic modelling, network analysis, and the German novel, 1731–1864. In M. Erling & L. Tatlock (Eds.), *Distant readings: topologies of German culture in the long nineteenth century* (pp. 55–90). New York, NY: Camden House.
- Erling, M., & Tatlock, L.** (2014). Introduction: 'distant reading' and the historiography of nineteenth-century German literature. In M. Erling & L. Tatlock (Eds.), *Distant readings: topologies of German culture in the long nineteenth century* (pp. 1–25). New York, NY: Camden House.
- Graham S., Weingart, S., & Milligan, I.** (2012). Getting started with topic modeling and MALLET. *The Programming Historian* tutorial on MALLET. *The Programming Historian*, 1. Retrieved from <https://programminghistorian.org/lessons/topic-modeling-and-mallet>
- Hakkarainen, H.** (2020). Contagious humanism in early nineteenth-century German-language press. *Contributions to the History of Concepts* 3(15), 22–46. DOI: <https://doi.org/10.3167/choc.2020.150102>
- Hall, D., Jurafsky, D., & Manning, C. D.** (2008). Studying the history of ideas using topic models. In *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 363–371). Honolulu, Hawaii. Retrieved from <https://dl.acm.org/citation.cfm?id=1613763>

- Hansson, J.** (1999). *Humanismens kris: Bildningsideal och kulturkritik i Sverige 1848–1933*. Stockholm: Brutus Östlings Bokförelag Symposion.
- Jacobi, C., van Atteveldt, W., & Welbers, K.** (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 1(4), 89–106. DOI: <https://doi.org/10.1080/21670811.2015.1093271>
- Koselleck, R.** (1985). *Futures past: on the semantics of historical time*. Translated by K. Tribe. Cambridge, MA: MIT Press.
- Landfester, M.** (1988). *Humanismus und Gesellschaft im 19. Jahrhundert*. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Schonfield, A., Magnusson, M., & Mimno, D.** (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*, Vol. 2: *Short papers*. Retrieved from <http://aclweb.org/anthology/E17-2069>
- St. Clair, W.** (2004). *The reading nation in the Romantic period*. Cambridge: Cambridge University Press.
- Steinmetz, W., & Freeden, M.** (2017). Introduction: conceptual history. In W. Steinmetz, M. Freeden & J. Fernández-Sebastián (Eds.), *Conceptual history in the European space* (pp. 1–46). New York, NY: Berghahn Books.
- Stöber, R.** (2014). *Deutsche Pressegeschichte*. 3rd edn. Munich: UVK Verlagsgesellschaft Konstanz.
- van Bommel, B.** (2015). *Classical humanism and the challenge of modernity: debates on classical education in 19th century Germany*. Berlin: De Gruyter.

Computational Authorship Attribution in Manuscripts and Other Texts

Reima Välimäki, Alekski Vesanto, Anni Hella,
Adam Poznański and Filip Ginter

Introduction

Authorship and originality were tricky things in medieval literature and documents. They were written in a culture of imitation rather than originality. The Latin word *auctoritas* could mean both an author and their authority, usually both combined. *Auctoritas* had initially meant the quality by which a person can be trusted. Consequently, it came to mean the authoritative status of a person and further that of their writings.¹ So, the ‘author’ was not just any writer whose texts were read, but the modern equivalent to their status would be something like that of Judith Butler in gender studies or Max Weber in sociology. Often, only writers meriting the status of *auctoritas* were explicitly cited, others

silently borrowed and, in modern terms, plagiarised. Thus, it is not uncommon to find late medieval theological treatises where long passages are copied from other high and late medieval works, but only patristic sources and the most important medieval theologians such as Bernard of Clairvaux are named. As the 'authorship' in medieval discourse was more related to responsibility over content than style or form, we can find very original literary works written under the term 'compilation'.² Some of these compilations circulated under the name of an authoritative figure, such as Augustine, and were generally considered to convey his thoughts, even if the actual content contained very little of his original works. Furthermore, scribes and secretaries were often employed in the actual composition of the final work, creating a further layer of stylistic authorship in a text.³ Due to these characteristics, the scholarship of medieval literature has for a long time recognised that the role of a compiler and even copyist was often comparable and at times surpassed that of an *auctor*.⁴ Finally, many texts circulated anonymously or under an early modern misattribution.

In this chapter, we discuss one such complicated case of medieval authorship, an anti-heretical treatise known as the *Refutatio errorum*, written in the 1390s in German-speaking Europe. It has many of the characteristics described above. It is of compilatory nature, containing passages from different sources, very few of which are named in the text itself. It has no prologue or comparable section, where someone would claim their authorship over the text. Instead, the whole treatise is very practical, intended to provide information, but not to flaunt with rhetorical abilities of its composer. For a long time, the treatise was considered anonymous, until R. Välimäki provided contentual, structural and codicological evidence linking the treatise to the inquisitor Petrus Zwicker, who also authored a more famous anti-heretical work entitled *Cum dormirent homines*.⁵ As is usually the case with reattribution of a medieval work, the conclusions are not based on single evidence, but on a combination of mutually enforcing pieces of evidence (described below). The purpose of this chapter is to add a new element to the analysis: computational authorship attribution using a Support Vector Machine (SVM). We discuss the results of the computer classification in relation to qualitative analysis of the text. The aim is to find out if computational methods provide added value to conventional authorship attribution of a medieval text. Or, could one claim that the computational methods are to be regarded as superior to qualitative interpretation by an expert human reader?

Computational authorship attribution can be considered a sub-category of style-based document authentication (*Echtheitskritik*),⁶ and the first attempts to apply computational methods in the attribution tasks of Latin literature were in the 1970s.⁷ After that, there was a lull in computational study of classical and medieval literature with a few exceptions,⁸ but since the late 1990s and especially in the past five years several studies have demonstrated that computational authorship attribution can be a powerful tool in the recognition of classical and medieval authors.⁹ As perhaps symptomatic to the whole field

of digital humanities, the first publications of this new wave of computational authorship studies have concentrated on developing the methodology itself, and results have been published mainly in digital humanities journals. At the same time, the attribution of new texts to classical and medieval authors goes on with little regard to the results of computational stylistics,¹⁰ and some recent publications even claim that statistical stylometry has fallen out of favour.¹¹ Although such a claim betrays lacking knowledge of one's research field, digital humanities scholars are not entirely blameless. Very few publications have tried to bridge the gap between discussions on authorship in the fields of literature studies and history, and in the computational linguistics respectively. Remarkable exceptions are Jeroen de Gussem's recent article on trails of Nicholas of Montiéramey's secretarial style in Bernard of Clairvaux's writings, as well as Mike Kestemont and colleagues' study on collaborative authorship of Hildegard of Bingen and Guibert of Gembloux.¹²

Consequently, computational analysis can raise suspicion among humanities scholars trained in qualitative methods. Machine learning or other branches of computational text classification may appear as radically new ways of analysing sources that bypass the human expertise (and are therefore terrifying). This they, however, are not. Although utilising computational capacity and handling amounts of data that far surpass the abilities of any human individual, the computational authorship attribution uses stylistic features that have been long since recognised as marks of authorship. A. Mutzenbecher prepared a new edition of Maximus of Turin's sermons at the beginning of the 1960s and defined 16 criteria (some with several sub-categories), which he divided into four slightly overlapping groups: (1) external evidence, (2) biblical quotations and their exposition, (3) style and (4) sources. Some of his criteria were primary, some secondary. An authentic sermon had to fulfil two primary criteria and several secondary criteria.¹³

For the purposes of this chapter, it is not necessary to explain what all of these were. It is sufficient to note that many of Mutzenbecher's criteria were purely qualitative, such as the theological topics Maximus typically discussed, but especially criteria for the introduction and exposition of the biblical citations (numbers 6 to 8) and criterion 13, linguistic-stylistic characteristics, include features that are similar to stylistic features used in computational authorship attribution: word uni- and bigrams formed of function words and other very common expressions (for example, *enim, ex quo, hoc est, quanto magis, sed dicit, ego dico, mirum est*).¹⁴ Mutzenbecher was well aware that these stylistic features appeared in almost all other authors in addition to Maximus, and that none of them could individually constitute authorship, 'but if several of them support each other reciprocally, their relationship might express something typical.'¹⁵ Computational authorship attribution does precisely that: it uses features that appear in almost all authors, but with different emphasis. To put it simply: it is the combination of all the significant stylistic features in comparison to their combination in other authors that determines authorship. A computer,

however, is not limited to a few obvious stylistic features of an author, but can handle thousands and millions of these in a systematic and repeatable way.

The *Refutatio Errorum* and Its Redactions

The test case in this study is a text known as the *Refutatio errorum*. It is a polemical description of the Waldensians, a religious group persecuted as heretics by the Catholic Church in the Middle Ages and early modern period. In the 1390s, a series of inquisitions and other trials were directed against the group in German-speaking Europe,¹⁶ and the *Refutatio* was written as part of the literary polemics accompanying the persecution. The treatise gives a view of Waldensianism very similar to that of the better known polemical treatise against the Waldensians, *Cum dormirent homines* (henceforth, CDH), written by one of the most important inquisitors of the late 14th century, the Celestine provincial Petrus Zwicker. The *Refutatio* is clearly a representative of the same era and state of knowledge about the Waldensians as the CDH. It has been commented on by scholars much less than the CDH, quite likely because the only available printed version, edited by Jacob Gretser together with the CDH (1613/1677), is obviously incomplete. It has 10 chapters, but the text stops abruptly in the middle of the tenth chapter.¹⁷

Among the scholars, there has been confusion rather than actual disagreement about the *Refutatio*'s authorship. For a long time, everyone was reluctant to make definite claims about its authors. In his groundbreaking studies on the CDH, P. Biller did not suggest any author or dating for the *Refutatio*, but seems to have held the view that the two treatises were *not* written by the same author, that is Zwicker. In fact, Biller uses the common manuscript tradition of the *Refutatio* and CDH as an argument against the attribution of the CDH to Peter von Pillichsdorf, the author suggested by Gretser in his 17th-century edition. The argument runs as follows: Gretser's misattribution was based on the now lost Tegernsee manuscript, which included the CDH and a short anti-Waldensian treatise by Pillichsdorf, who is the only author mentioned in the manuscript. This consequently led Gretser to propose Pillichsdorf as the author of both these treatises treating the same topic. According to Biller, this is a parallel case to that of the several manuscripts, including the CDH and the *Refutatio*. These too were two different treatises on the same subject, but were treated as one by both medieval scribes and modern compilers of manuscript catalogues. Biller did not state anything explicit concerning the authorship of the *Refutatio*, calling it and Zwicker's CDH only 'two tracts on similar material'.¹⁸ They do indeed cover very much the same material, and because of this P. Segl has tentatively proposed that these two treatises originated from the same hand.¹⁹ E. Cameron describes the treatise very vaguely, but evidently treats it as a product of the 1390s, at one point calling it 'a third treatise from Zwicker's

circle.²⁰ A. Patschovsky has also associated the *Refutatio* loosely with Zwicker, without making any definite claims about its authorship.²¹ In other words, there has been a vague suspicion that Petrus Zwicker, or someone close to him, wrote the treatise.

To further complicate the study of this text, the only available printed editions are based on a text that is anything but representative of the manuscript tradition of the *Refutatio*. As noted, Jacob Gretser printed the tract in the 17th century from a manuscript that ends abruptly in the middle of Chapter 10. I. von Döllinger's 19th-century edition from the same manuscripts does not help, but adds further confusion, as the order of the chapters is mixed in the edition, and material not belonging to the *Refutatio* is inserted among the text.²² An analysis of all the preserved 19 manuscripts of the work by Välimäki has demonstrated that the edited version of the texts does not concur with the main manuscript tradition, that is the most common and widely circulated medieval text. All in all, Välimäki found four different redactions of the *Refutatio errorum*. Of these, Redaction 1 is by far the most common, with 13 manuscripts. It is also the only redaction accompanying Zwicker's better known and more popular treatise, the *Cum dormirent homines*. The two texts appear together in eight manuscripts. In comparison, the text printed by Gretser in the 17th century is a late and incomplete redaction (Välimäki's Redaction 4) represented by only two medieval manuscripts.²³

In addition to collation of the *Refutatio*'s manuscript tradition, Välimäki has also proposed that the treatise can be attributed to Petrus Zwicker. The two works present a very similar view on the Waldensians; they both follow similar structure of polemical refutation by presenting heretical propositions and Catholic counter-arguments, mainly based on biblical quotations. The most important pieces of evidence for the common authorship are the sources cited in these two works. In the CDH, Zwicker quotes almost exclusively the Bible in support of his arguments. The single exception to the rule is a reference to Boethius' *Consolation of Philosophy*. The same quote can be found in almost the exact same form in the *Refutatio errorum*. In addition, the author of the *Refutatio* had direct access to Moneta of Cremona's 13th-century anti-heretical treatise *Adversus Catharos et Valdenses*. The treatise was very rare north of the Alps, but Petrus Zwicker used it when composing the CDH. The final rare source implying the authorship of Zwicker is a misquotation of Ezekiel 33.12 in the *Refutatio*. The exact form of this quotation comes from the legal consultations on the case against the goldsmith Heynuš Lugner in the late 1330s or early 1340s, transmitted in two manuscripts, a Bohemian inquisitor's manual Linz MS 177, and another, St. Florian, MS XI 234, which is copied from the first manuscript. The Linz manual was once owned by Petrus Zwicker and the St. Florian manual was copied from his own inquisitor's manual. Ergo, the author of the *Refutatio* had access to a rare text, which has certain manuscript circulation only in connection to Petrus Zwicker.²⁴

Texts for the Analysis and Pre-processing

Next, we analyse the two most important redactions of the *Refutatio* with computational classification in order to verify Zwicker's authorship. The redactions selected for the classification are the most common and longest Redaction 1 and Redaction 4 representing the version in Gretser's edition. The text of Redaction 4 is taken from a manuscript Augsburg, Universitätsbibliothek MS 338 (TEST1) as well as Gretser's edition (TEST2). Redaction 1 is transcribed from the manuscript Vienna, Österreichische Nationalbibliothek (ÖNB), MS 1588 (TEST3). All texts are long enough for a reliable authorship attribution, from around 5,500 words in TEST2 to over 9,000 words in TEST3. We excluded Redactions 2 and 3, both extant in a single manuscript and not close to the original text. Neither of these redactions is representative of the medieval or modern reception of the work.

We trained the classifier with Petrus Zwicker's CDH (around 23,000 words). The text we used comes from the same Gretser's edition as one of the tested versions of the *Refutatio*. The reference corpus for training our classifier consisted of late ancient and medieval anti-heretical polemical treatises, which is the genre of both Zwicker's CDH and the *Refutatio*. In total, this training data has around 600,000 words. The emphasis is on medieval texts, and the corpus includes three works that are almost contemporary to Zwicker's texts: Wasmud von Homburg's *Tractatus contra hereticos*, an anonymous *Attendite a falsis prophetis* and the already mentioned Peter von Pillichsdorf's *Contra Pauperes de Lugduno*. In addition, the most important source and stylistic model for Zwicker's CDH, Moneta of Cremona's *Adversus Catharos et Valdenses*, is included. From Moneta's very long treatise, we selected only Book 5, where many of the anti-Waldensian arguments are presented. Alone, Book 5 has over 120,000 words, and including all 400,000 words from the whole treatise would have created an imbalanced reference corpus. The complete corpus with bibliographical information is in Appendix 16.1. The data is available at our GitHub page in masked form only to protect copyrights of recent editions used in the corpus.²⁵

The dataset we use is far from easy and common in authorship attribution tasks. It is a mixed corpus of different edition and transcription standards, which is a problem for feature selection. Even though character n-grams are widely used as features in text classification, recent computational studies on the authorship of classical and medieval texts have preferred lemma-level approach and function word analysis over character n-grams or plain text.²⁶ This is partly due to the orthographical variation in medieval Latin. The effects of orthographical variation are more marked when the features used are a few dozen function words. However, as our classifications are based on a much more complex set of features, the effect of single 'bad' features for the end result is minimal. Using word uni- and bi-grams from plain text, as well as character n-grams, also has significant benefits in Latin. It gives access to stylistic

solutions below the word level, such as the author's decision to use the subjunctive instead of the indicative.²⁷

We solved the most common issues of different editorial principles and orthographical variation with simple normalisation rules:

u → v
 j → i
 y → i
 ae → e
 oe → e
 char → car (to solve variation *charitas* vs. *caritas*)
 wa → va (to solve variation *ewangelium* / *evangelium* and *waldenses* / *valdenses*)

These solve the majority of orthographical variation caused by editorial and scribal conventions and the differences of medieval and classical Latin without masking potentially significant stylistic features. In addition to orthographical normalisation, in the pre-processing phase we cleaned the texts from editorial additions such as page numbers and chapter titles (unless part of the original). Punctuation, numerals and single characters were removed. From early medieval texts, we naturally cleaned the references to bible books and verses (which were added by later editors), but in late medieval texts, most notably Zwicker's own treatise, these are part of the original and were thus preserved. The pre-processing was done automatically, but confirmed with sanity checks.

However, the transcripts from medieval manuscripts have much more variation than edited texts. While orthographical habits and grammatical mistakes of individuals are excellent stylistic features when one is dealing with autographs, in medieval manuscript culture such variation is noise in data. We are usually not interested in writing conventions of an individual scribe, but those of the author or compiler of the work. Even the usual orthographical variation of late medieval manuscripts is challenging to normalise without also masking potentially significant stylistic features.²⁸ Thus, in addition to solving the question of Zwicker's authorship, we experimented with the data in order to find a relatively effortless way to pre-process and analyse such a corpus with a computer. The expected results from our dataset are as follows:

1. If the pre-processing and feature selection are able to overcome the orthographical challenges, all test cases of the *Refutatio* should be classified in a similar way. We expect that they are classified as Zwicker's works together with the CDH (values over 0).
2. All other works should get values below 0 in the classification.
3. If Peter von Pillichsdorf's treatise from Gretser's edition is classified together with Zwicker's works, the early modern editorial solutions have more weight more than medieval authorship.

Computational Authorship Attribution: Methods

The puzzle we set out to solve is: Did Petrus Zwicker write the *Refutatio errorum*? In authorship attribution, this is called a verification problem: we do not have a closed set of candidates, but one suspected author.²⁹ We constructed the verification problem as a simple binary classification, where Zwicker's treatise forms one class and all other authors in the training material a second class. The classifier was trained with this material, and the versions of the *Refutatio* were presented as a test case. We use the two corpora combined as training data for the classifier, while the test cases form the test data. The different redactions of the *Refutatio* are each treated as a separate test case.

Here, we present an overview of the methods. For technical details and code, please consult our project repository.³⁰ For the classification, we use a linear SVM, which is a simple yet effective classifier and has traditionally been applied in text classification tasks.³¹ The SVM works by learning a weight for every feature from the training data, so as to maximise the decision margin between the two classes. The weight being positive or negative is an indication of which class the feature is potentially associated with, although one needs to exercise caution when comparing features in isolation based on their weight. The features we use with the SVM are word unigrams and bigrams. In other words, we train the classifier with the training data to recognise the features typical and atypical of Petrus Zwicker's style. After that, the test cases are classified, and the output is a value indicating how much (positive) or how little (negative) the sum of weighted features in each test case resembles the class (Zwicker). The values are represented on a scale between -1 and 1.

The value and the decision are largely useless in isolation if we cannot be certain that the classifications are valid overall. Here, we apply the standard technique of cross-validation using the training data, which provides us with an estimate of the classification accuracy and therefore the reliability of our results on the actual test documents.³²

The classifier we use is by nature indiscriminating when it comes to the features. It does not care which features are used, as long as they increase the training accuracy. In authorship attribution tasks, this would ideally be features that describe the author's way of writing, such as the usage of function words. Even within a single genre as in our training data, however, the particular topic of each text affects the results. We run the classification to unmasked data, and among the 10 strongest positive features five included 'Waldensians' in some form.³³ A classification from such features is based only partly on an author's style, and the topic of the texts heavily distorts the results. Therefore, we must mask topic words so as to not let the classifier focus purely on the topic of the texts instead of the author's style. To this end, we calculated the thousand most common words in post-classical (Christian) Latin.³⁴ Any word not in the calculated word list will be masked. This has been shown to drastically increase the accuracy of cross-genre classifications, as it forces the classifier to

learn author-specific rather than topic-specific features.³⁵ This method does not completely remove topic words, but it only leaves the ones that appear regularly across different genres. In the following, we concentrate on results from the classification in the masked data.

Comparing the Results

The classification from the SVM using masked data is presented in Table 16.1 and in Figure 16.1.

The results were both expected and unexpected. First of all, the classification confirms that also from a stylistic perspective Petrus Zwicker is the author of the *Refutatio errorum*. All redactions, whether transcripts from manuscripts or the text from Gretser's edition, were classified as Zwicker's texts with a clear margin to other works. The exception here is the short treatise *Attendite a falsis*

Table 16.1: SVM Classification, masked data.

TEST2	Refutatio errorum R4b (Gretser)	1.0
Anon.	Attendite a falsis prophetis	0.953
Petrus Zwicker	Cum dormirent homines	0.926
TEST1	Refutatio errorum R4a	0.745
TEST3	Refutatio errorum R1	0.662
Durand of Huesca	Liber antiheresis	-0.062
Anon. of Passau	Tractatus de erroribus hereticorum	-0.212
Berthold von Regensburg	Sermones	-0.254
Wasmud von Homburg	Tractatus contra hereticos	-0.352
Anon.	Disputatio inter Catholicum et Paterinum hereticum	-0.357
Durand of Huesca	Liber contra Manicheos	-0.573
Petrus de Pillichsdorf	Contra Pauperes de Ludguno	-0.574
Moneta Cremonensis	Adversus Catharos et Valdenses, Book 5	-0.717
Alanus de Insulis	Contra haereticos	-0.807
Petrus Veronensis	Summa contra haereticos	-0.854
Johannes Cassianus	De incarnatione Domini contra Nestorium	-0.889
Hermannus de Scildis	Tractatus contra haereticos negantes immunitatem Ecclesiae	-0.953
Augustinus Hipponensis	Contra Faustum Manichaeum	-0.986
Augustinus Hipponensis	Contra epistulam Fundamenti	-1.0

Source: Authors.

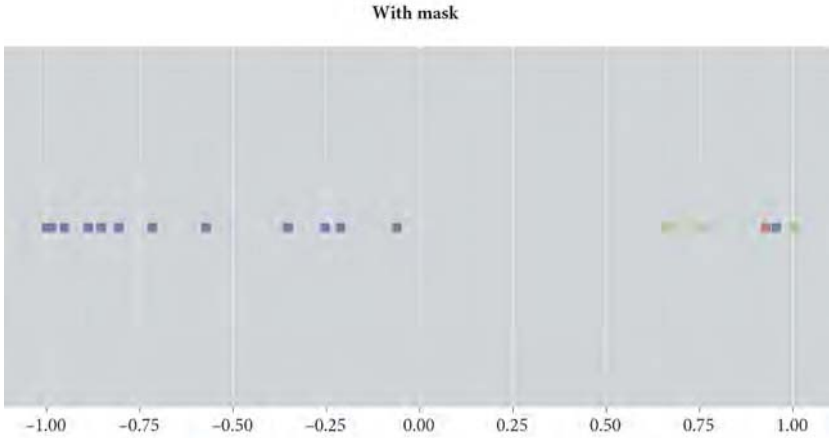


Figure 16.1: Green dots represent the test cases (*Refutatio*), red dots Zwicker's CDH and blue dots texts by other authors. Source: Authors

prophetis, discussed below. But if we exclude it, all other works from the reference corpus got values below 0, and Zwicker's texts were neatly classified between 0.662 and 1.0. Not surprisingly, the text from Gretser's edition got the highest value (1.0), in fact higher than the CDH. This appears contradictory at first, but the explanation is simple: the classifier first learns the weight of features from the whole text, but in cross-validation the text is divided into slices of 1,000 words, and the final value is the average of all the slices. Some of these got values below 1, weighting down the average. In other words, the *Refutatio*'s style is indistinguishable from Zwicker's style in the *Cum dormirent homines* in comparison to the reference corpus.

After pre-processing and masking, the features on which the SVM bases its decision pass the sanity check. In Appendix 16.2, there is a list of the 50 strongest positive and negative features. In both positive and negative class, these are function words or common content words, or bi-grams combining such common words with masked words. Among these, only one positive feature ('imo' 6.344) results from orthographical variation (imo vs. immo). All in all, a classification based on these features can be deemed reliable and non-dependable from topics.

The classifier was also able to distinguish authorial signature from both editions and manuscripts so that the editorial solutions or orthographical variation do not completely distort an author's style. This is confirmed not only by the consistent classification of the different versions of the *Refutatio*, but also by the value acquired by Peter von Pillichsdorf's *Contra Pauperes de Ludgungo*. Despite being a tract on the same topic (Waldensians) as the CDH and the *Refutatio*, and from the same edition (Gretser) as the CDH and TEST2, it got a clearly negative value of -0.574 . Six other texts got values nearer to the

threshold, so Pillichsdorf's tract is very far from Zwicker's texts. The edition, of course, has an effect, as we can see from the very strong value TEST2 got.

The unexpected result was the *Attendite a falsis prophetis*. It got a very high value (0.953), and in the classification in the unmasked data, not presented here in detail, the result was consistent (0.559). This cannot be explained by the same topic, as the extremely high value is based on masked data. How should we interpret this? Do we have a new text attributed to Petrus Zwicker? This is a possibility, but the SVM's classification must be considered against the historical context, manuscript tradition and the contents of the text.

First, very little in the contents of the text contradicts Zwicker's views in the *Refutatio* or the CDH. In fact, the *Attendite* presents similar Waldensian propositions and Catholic counter-arguments to those of Zwicker. For example, the CDH, *Refutatio* and *Attendite* all begin by refuting the Waldensian claim of a legitimate lay ministry and proceed then to treat individual points of doctrine such as denial of Purgatory and oath-taking. P. Biller has already pointed out a certain similarity between the *Attendite* and the CDH.³⁶ There is a minor detail: the *Attendite* states that the Waldensians do not accept the books of Maccabees as parts of the biblical canon.³⁷ In the CDH, Zwicker stays silent about this and in fact uses the Maccabees to prove that the intercession on behalf of the dead had its foundation in the Bible.³⁸ This small divergence, however, can be explained by the development of Zwicker's argumentation. He desperately needed the Maccabees in order to maintain the principle of finding the foundation of Catholic doctrine and practices solely in the Bible, a principle that was only fully developed in his main work, the CDH. The author of the *Attendite* did not follow these guidelines: some of the arguments are supported by patristic quotes. Yet, this does not automatically deny Zwicker's authorship. Although Zwicker got rid of extra-biblical quotes almost completely in writing the CDH, he refers to patristic *auctoritates* several times in the *Refutatio*.³⁹ Solely based on the contents, the *Attendite* could be an early work of Petrus Zwicker. He was, after all, a man obsessed about the Waldensians and the threat they posed to the Church, and it is not out of the question that he wrote a third treatise against them.

The main doubt comes from the dating of the work. This is remarkably difficult, because the text is very general and does not refer to any specific persons or incidents. Nor does the author use any particular or rare sources. In principle, any late medieval author with access to anti-heretical treatises commonly circulating in Central Europe could have written the text. There have been two propositions about the author, one obviously mistaken, and another probably due to confusion with another text. Based on one manuscript (Wrocław, University library MS I F 230), R. Cegna misdated the text to the year 1399 and misattributed it to the Silesian inquisitor Johannes of Gliwice.⁴⁰ There are no grounds whatsoever for either the dating or the attribution,⁴¹ and a few manuscripts predate the one used by Cegna. Older research attributes the treatise to the Bohemian reform preacher and troublemaker Conrad of Waldhausen,

which would date the text to the 1360s.⁴² The attribution might have resulted from confusion of this short tract with Conrad's sermon on the same Bible verse, given at some point in 1363 to 1369.⁴³ The manuscript transmission history points to Austria, Southern Germany, Bohemia and Silesia. P. Biller has proposed that the earliest possible dated manuscript of the *Attendite* is St. Paul im Lavanttal, MS 71/4, which has the year 1373 at folio 160va, referring to the composition date of a copy of a polemical letter from a converted Austrian Waldensian to Lombardian Waldensian Brethren.⁴⁴ Although the part with the *Attendite* (folios 144ra–146vb) belongs to the same fascicule with the letter, it is uncertain if 1373 is the production date of this particular exemplar. The manuscript MS 71/4 is a compilation with fascicules produced at different times in the late 14th and early 15th centuries.⁴⁵ The dating can only be confirmed through codicological analysis of the physical object itself, which is not possible within this study. The more secure dating comes from Klosterneuburg, MS CC 826, datable to 1391 and described by P. Biller.⁴⁶ With absence of a systematic study on the manuscript circulation of the *Attendite*, this is the most credible *terminus ante quem*. It means that the geographical distribution and dating of the manuscripts overlaps with the beginning of Petrus Zwicker's career as inquisitor of heresy, not excluding his authorship.

The final caveat comes from the credibility of the attribution itself. The text is only around 2,500 words long, making the attribution unreliable, as we are dealing with data with noise. In addition, the *Attendite* and the CDH (which is the material we used to train the computer for the class Zwicker) quote the same Bible verses. Although the quotations are not word-to-word identical, there is shared material in these two works. In the attribution of such a short text, it necessarily has an impact. Finally, we used a version of the text from a single manuscript, which we had in machine-readable format. There is a critical edition of the text by R. Cegna, but it too is mainly based on a single manuscript with variant readings in endnotes.⁴⁷ The final attribution of the *Attendite* is only possible when further study reveals the earliest redaction of the text and the manuscript dates are confirmed. From the earlier proposed authors, texts from Conrad of Waldhausen must be included in the classification as a possible author. At this point, we must be content to say that the *Attendite a falsis prophetis* is possibly attributable to Petrus Zwicker, but the attribution needs corroborating evidence from the manuscript tradition.

Conclusion: Additional Value of the Computational Analysis?

In the future, the computational authorship attribution should be taken into the toolbox of historians and philologists, who work with anonymous, pseudo and dubious texts. The classifiers developed for the analysis of modern literature or forensic purposes have been proved to be effective also in the study of ancient and medieval texts.

In our case study, the authorship of the *Refutatio errorum*, the computational methods produced both corroborating evidence and expected results, as well as radically new insights. The authorship of the *Refutatio* was confirmed as Petrus Zwicker through computational stylistics. Although there were previous, convincing pieces of evidence in support of this, the analysis is not without added value. A computer's decision is based on a completely different set of features than contents analysis and contextual evidence presented in the previous studies. Another important result was the classification of Peter von Pillichsdorf's treatise as clearly non-Zwicker. This not only confirms the earlier qualitative attribution, but demonstrates that our classifier can bypass the stylistic conventions of an early modern editor and detect the medieval author signature below.

The greatest added value of computational authorship attributions comes, however, from the unexpected results, from texts behaving in an anomalous way. In this classification, the *Attendite a falsis prophetis* did precisely this. Up until this point, nobody has really considered Zwicker's authorship, because the manuscript tradition points to a somewhat earlier treatise. Yet, when the classification gave a strong attribution to Zwicker, it forced us to reconsider the qualitative evidence. This, in turn, was revealed to be indecisive as well. Although we are not ready to declare the case closed and a new text attributed to Zwicker, the example demonstrates the true power of computational methods: it breaks the existing patterns of thought and demands re-evaluation of previous presuppositions.

Our chapter demonstrates that computational history cannot progress in isolation from the more conventional study of history, particularly the very basic archival study of sources. The attribution of the *Attendite a falsis prophetis* is to remain ambiguous until the existing manuscripts are surveyed in detail. The study of history depends on source criticism, and in order to date, attribute and localise sources with digital methods we have to take care that our metadata is up to standard.

Acknowledgements

This study has been funded by the Academy of Finland, Academy Programme DIGIHUM 2015-2019: consortium Profiling Premodern Authors (PRO-PREAU), project number 293024.

Appendix 16.1: Text Corpus

TEST1: *Refutatio errorum*, Redaction 4a

Source: Augsburg, Staats- und Stadtbibliothek, MS 338, fols. 159r–170r.

TEST2: *Refutatio errorum*, Redaction 4b

Source: Gretser, J. (Ed.). (1677). *Lucae Tvdensis episcopi, Scriptores aliquot*

svccedanei contra sectam waldensivm. Maxima Bibliotheca Veterum Patrum, Et Antiquorum Scriptorum Ecclesiasticorum. Tom. XXV. Lvgdvni: Anissonios, 302G–307F.

TEST3: Refutatio errorum, Redaction 1

Source: Vienna, Österreichische Nationalbibliothek, MS 1588, fols. 191r–211v.

TRAINING DATA

Suspected author: Petrus Zwicker:

Text: Cum dormirent homines (CDH)

Source: Gretser, J. (Ed.). (1677). *Lucae Tvdensis episcopi, Scriptores aliquot svccedanei contra sectam waldensivm*. Maxima Bibliotheca Veterum Patrum, Et Antiquorum Scriptorum Ecclesiasticorum. Tom. XXV. Lvgdvni: Anissonios, 277F–299G.

Other authors:

Author: Alanus de Insulis (Alain of Lille)

Text: Contra haereticos

Source: Patrologia Latina 210. Text from *Corpus Corporum*: <http://mlat.uzh.ch/?c=2&w=AlDeIn.ConHae>

Author: Anonymous

Text: Attendite a falsis prophetis

Source: St. Florian, MS XI 152, fols. 48v–50v.

Author: Anonymous

Text: Disputatio inter Catholicum et Paterinum hereticum

Source: Hoécker, C. (Ed.). (2001). *Disputatio inter catholicum et paterinum hereticum: die Auseinandersetzung der katholischen Kirche mit den italienischen Katharern im Spiegel einer kontroverstheologischen Streitschrift des 13. Jahrhunderts*. Tavarnuzze (Firenze): SISMELE edizioni del Galluzzo, 3–80.

Author: Anonymous of Passau

Text: Tractatus de erroribus hereticorum

Source: Nickson, M. A. E. (1962). *A critical edition of the treatise on heresy ascribed to Pseudo-Reinerius, with an historical introduction*. Queen Mary, University of London, 1–154.

Author: Augustinus of Hippo

Text: Contra Faustum Manichaeum

Source: Patrologia Latina 42. Text from *Corpus Corporum*: <http://mlat.uzh.ch/?c=2&w=AugHip.CoFaMa>

Author: Augustinus of Hippo

Text: Contra epistulam Fundamenti

Source: Corpus Scriptorum Ecclesiasticorum Latinorum (CSEL) 25.1. Text from *Corpus corporum*:
<http://mlat.uzh.ch/?c=19&w=August.CoEpFunCSEL>

Author: Berthold von Regensburg
 Text: Sermones [XXIII, XXVIII, XXVIII, 'Sancti pre Fidem' and 'Dominica Duodecima']
 Source: Czerwon, A. (2011). *Predigt gegen Ketzer: Studien zu den lateinischen Sermones Bertholds von Regensburg*. Tübingen: Mohr Siebeck, 203–233.

Author: Durand of Huesca
 Text: Liber contra manicheos
 Source: Thouzellier, C. (1964). *Une somme anti-cathare: le Liber contra Manicheos de Durand de Huesca*. Louvain: Spicilegium sacrum Lovaniense, 67–336.

Author: Durand of Huesca
 Text: Liber Antiheresis
 Source: Selge, K.-V. (Ed.) (1967). *Die ersten Waldenser: mit Edition des Liber antiheresis des Durandus von Osca* (Vol. 2). Berlin: De Gruyter, 3–257.

Author: Hermannus of Scildis
 Text: Tractatus contra haereticos
 Source: Zumkeller, A. (1970). *Hermannus de Scildis O.S.A.: tractatus contra haereticos negantes immunitatem et iurisdictionem sanctae Ecclesiae et Tractatus de conceptione gloriosae virginis Mariae*. Würzburg: Augustinus-Verl., 3–108.

Author: Johannes Cassianus
 Text: De incarnatione Domini contra Nestorium
 Source: Corpus Scriptorum Ecclesiasticorum Latinorum (CSEL) 17. Text from *Corpus Corporum*: <http://mlat.uzh.ch/?c=19&w=Cassia.ConNesCSEL>

Author: Moneta Cremonensis (Moneta of Cremona)
 Text: Adversus Catharos et Valdenses, Liber V
 Source: Moneta (Cremonensis). (1743). *Monetae Adversus Catharos et Valdenses: libri quinque*. T. A. Ricchini (Ed.). Romae: Ex Typographia Palladis, 389–560.

Author: Petrus de Pillichsdorf (Peter von Pillichsdorf)
 Text: Contra Pauperes de Lugduno
 Source: Gretser, J. (Ed.) (1677). *Lucae Tydensis episcopi, Scriptores aliquot svecedanei contra sectam waldensivm*. Maxima Bibliotheca Veterum Patrum, Et Antiquorum Scriptorum Ecclesiasticorum. Tom. XXV. Lvgdvni: Anissonios, 299E–302F.

Author: Petrus Veronensis (?)

Text: Summa contra haereticos

Source: Kaeppli, T. (1947). Une somme contre les hérétiques de S. Pierre Martyr (?). *Archivum Fratrum Praedicatorum*, 17, 295–335

Author: Wasmud von Homburg

Text: Tractatus contra hereticos Beckardos, Lulhardos et Swestriones

Source: Schmidt, A. (Ed.) (1962). Tractatus contra hereticos Beckardos, Lulhardos et Swestriones des Wasmud von Homburg. *Archiv für mittelhochdeutsche Kirchengeschichte*, 14, 336–386.

Appendix 16.2: The 50 Strongest Positive and Negative Features from the SVM Classification on Masked Data

Positive features:

tv xxxxxxxxxxx	6.895
imo	6.344
xxx item	5.824
sed dicis	5.234
xxx dixit	5.04
item xxxx	4.864
item xxxxx	4.389
item xxx	4.385
xxxxxxxx imo	3.262
semper xxxxx	3.081
nostris xxxx	2.793
non solvm	2.562
ecce xxxxxxxxxxx	2.517
xxxxxx ecce	2.459
svvm xxxxx	2.225
sanctorvm dei	1.919
dixit xxxxxx	1.833
dicis xxxxxxxxxxx	1.729
dicentes xxxxxx	1.691
dicis	1.648
xxxxxx item	1.608
xxxxxxx item	1.493
ecce	1.435
xxxxxx mea	1.397
adhvc in	1.346
item xxxxxx	1.303
vbi xxxxxxxxxxx	1.295

item xxxxxxxxx	1.287
nec qvidem	1.285
solv	1.235
qvod angeli	1.143
sibi ipsis	1.129
domine xxxxx	0.953
mevm item	0.893
domini nostri	0.819
habes	0.812
discipuli	0.752
xxxxx domine	0.746
dominus xxxxxxxx	0.744
nvqvm	0.742
qvi venit	0.729
noster xxxxx	0.692
tva	0.672
vt videlicet	0.658
privs per	0.65
velvt	0.634
xxxxx nolite	0.614
habere xxxxxxxx	0.614
xxxxx ecce	0.606
xxxxx pro	0.595

Negative features:

ait	-11.246
apostolvs	-6.224
tantvm	-5.863
nec	-5.201
vt	-3.964
ei	-2.957
idest	-2.747
xxxxxxx vt	-2.545
hvivs	-2.221
enim	-2.133
xxxxxxx vt	-1.95
qvod	-1.93
xxxxxxxxxxxxx xxxxxxxx	-1.705
deo	-1.642
ac	-1.619
qvomodo	-1.477
de	-1.415

dicitvr	-1.334
dicit	-1.321
nvllys	-1.277
hoc	-1.275
est	-1.251
libro	-1.198
dicvnt	-1.187
qva	-1.177
cavsa	-1.173
xxxxxxx qvod	-1.166
si avtem	-1.125
xx	-1.105
secvndvm	-1.066
se	-1.052
sic	-0.999
ista	-0.929
non est	-0.905
dictvm	-0.894
facit	-0.825
xxxxxxx qvia	-0.813
ab	-0.787
si	-0.739
xxxxxxxxxxx	-0.718
xxxxxxxxx	-0.714
potivs	-0.69
carnem	-0.684
itervm	-0.677
xxxxxxxxxxx	-0.661
et	-0.651
qvod xxxxxxxx	-0.63
legitvr	-0.629
aliqva	-0.629
xxxxxxx	-0.623

Notes

- ¹ See, e.g., Levy 2012: 23–24.
- ² Minnis 1988: 192–193, 196 and *passim*.
- ³ Kestemont, Moens & Deploige 2015; De Gussem 2017.
- ⁴ See, e.g., Johnson 1991; Williams-Krapp 2000; Minnis 2006; Conti 2012.
- ⁵ Välimäki 2016: 45–76; Välimäki 2019: 38–39, 48–49, 56–58, 61–64, 102.
- ⁶ Stover & Kestemont 2016a: 144.
- ⁷ Marriott 1979; one should also remember the pioneering work on statistical stylistics that precedes the computer era, Yule 1944.

- ⁸ See, e.g., Clark 1987.
- ⁹ Gurney & Gurney 1998; Tse, Tweedie & Frischer 1998; Tweedie, Holmes & Corns 1998; Forsyth, Holmes & Tse 1999; Kestemont 2012; Kestemont, Moens & Deploige 2015; Kestemont et al. 2016; Stover & Kestemont 2016a; Stover & Kestemont 2016b; De Gussem 2017.
- ¹⁰ See, e.g., Weidmann 2015.
- ¹¹ Adams 2016: 202.
- ¹² Kestemont, Moens & Deploige 2015; De Gussem 2017.
- ¹³ Mutzenbecher 1961: 202–219.
- ¹⁴ Ibid.: 203–204, 207–209.
- ¹⁵ Ibid.: 202.
- ¹⁶ Kolpacoff 2000: 247–261; Modestin 2007: 1–12; Välimäki 2019: 30–37.
- ¹⁷ Gretser 1677.
- ¹⁸ Biller 2001: 252–253.
- ¹⁹ Segl 2006: 185 n. 102.
- ²⁰ Cameron 2000: 140, 142–143.
- ²¹ Patschovsky 1979: 27 n. 42.
- ²² Döllinger 1890: 331–344.
- ²³ Välimäki 2019: 39–48.
- ²⁴ Välimäki 2016: 57–76; Välimäki 2019: 38–39, 48–49, 56–58, 61–64, 102.
- ²⁵ See <https://github.com/propreau/zwicker>.
- ²⁶ Kestemont, Moens & Deploige 2015; De Gussem 2017.
- ²⁷ Cf. Kestemont, Moens & Deploige 2015: 9–10.
- ²⁸ Such as variations *-ci-/-ti-* ; *-mq-/-nq-* ; *-dq-/-cq-* ; *-mp- / -mn-*, which cannot be normalised with simple replace rules without losing significant features in the text.
- ²⁹ Koppel, Schler & Argamon 2009; Koppel et al. 2012.
- ³⁰ See <https://github.com/avjves/AuthAttHelper>.
- ³¹ Cortes & Vapnik 1995; Chang & Lin 2011. In particular, we used the scikit-learn implementation of SVM with L2 penalty and squared hinge as loss. The C-parameter of the classifier was set using cross-validation so as to avoid overfitting on the test data.
- ³² In cross-validation, we only focus on our training data, ignoring the actual test texts. We remove one document at a time from the training data and consider it as a new test case. Our current training data now consists of all texts but the new test case, and using it we subsequently train the classifier and let it give a class and a value for the new test case. Since we know the actual authors of the texts included in the training data, these results show how accurately the classifier classifies data which it has not seen.
- ³³ ‘heretici valdenses’; ‘valdensis’; ‘valdensis heretice’; ‘valdensivm’; ‘tv valdensis’.
- ³⁴ We calculated the word list from a corpus of 15 million words compiled for an attribution task of Augustine of Hippo’s works. It is available at <https://github.com/propreau/zwicker>.
- ³⁵ Stamatatos 2017.

- ³⁶ Biller 1974: 365.
- ³⁷ St. Florian, MS XI 152, fol. 49v.
- ³⁸ Zwicker 1677: 288D–288E.
- ³⁹ Välimäki 2016: 77–114; Välimäki 2019: 61–62, 68–71, 73–85, 90, 94–98, 102–103.
- ⁴⁰ Cegna 1982.
- ⁴¹ Patschovsky 1994: n. 15.
- ⁴² Bartoš 1932: 32–33; Molnár 1989: 158 n. 29.
- ⁴³ Cf. Patschovsky 1979: 125–126.
- ⁴⁴ Biller 1974: 221. Biller cites MS 92/4, fol. 161va, but it is a mistake, the manuscript in question is MS 71/4. For the best overview of the manuscript tradition, see Biller 1974: 365–366.
- ⁴⁵ Glaßner 2002: Cod. 71/4 (olim 28.4.23).
- ⁴⁶ Biller 1974: 216–217, 365.
- ⁴⁷ Cegna 1982: 53–65.

References

- Adams, J.** (2016). *The revelations of St Birgitta: a study and edition of the Birgittine-Norwegian texts, Swedish National Archives, E8092*. Leiden: Brill.
- Bartoš, F. M.** (1932). Husitika a bohemika několika knihoven německých a švýcarských. *Vestník královské české společnosti nauk: Trída filosoficko-historicko-jazykozpytná*, 5, 1–92.
- Biller, P.** (1974). *Aspects of the Waldenses in the fourteenth century, including an edition of their correspondence* (doctoral dissertation, University of Oxford).
- Biller, P.** (2001). *The Waldenses, 1170–1530: between a religious order and a church*. Aldershot: Ashgate.
- Cameron, E.** (2000). *Waldenses: rejections of holy church in medieval Europe*. Oxford: Blackwell.
- Cegna, R.** (1982). La condizione del valdismo secondo l'inedito 'Tractatus bonus contra haereticos' del 1399, attribuibile all'inquisitore della Silesia Giovanni di Gliwice. In *I Valdesi e l'Europa* (pp. 39–66). Torre Pellice: Società di studi valdesi.
- Chang, C.-C., & Lin, C.-J.** (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 1–27.
- Clark, F.** (1987). *The pseudo-Gregorian dialogues*, Vol. 1. Leiden: Brill.
- Conti, A.** (2012). Scribes as authors, transmission as composition: towards a science of copying. In S. Ranković, I. B. Budal, A. Conti, L. Melve & E. Mundal (Eds.), *Modes of authorship in the Middle Ages* (pp. 267–288). Toronto: Pontifical Institute of Mediaeval Studies.
- Cortes, C., & Vapnik, V.** (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

- De Gussem, J.** (2017). Bernard of Clairvaux and Nicholas of Montiéramey: tracing the secretarial trail with computational stylistics. *Speculum*, 92(S1), S190–S225.
- Döllinger, J. J. I. von** (1890). *Beiträge zur Sektengeschichte des Mittelalters: Zweiter Theil. Dokumente vornehmlich zur Geschichte der Valdesier und Katharer*. Munich: C. H. Beck.
- Forsyth, R. S., Holmes, D. I., & Tse, E. K.** (1999). Cicero, Sigonio, and Burrows: investigating the authenticity of the *Consolatio*. *Literary & Linguistic Computing*, 14(3), 375–400.
- Glaßner, C.** (2002). *Inventar der Handschriften des Benediktinerstiftes St. Paul im Lavanttal*. Retrieved from <http://www.ksbm.oeaw.ac.at/stpaul/inv/index.htm>
- Greter, J.** (Ed.) (1677). Refvtatio Errorvm, Quibus Waldenses distinentur, incerto auctore. In *Lucae Tvdensis episcopi, Scriptores aliquot svccedanei contra sectam waldensivm, Maxima Bibliotheca Veterum Patrum, Et Antiquorum Scriptorum Ecclesiasticorum*, Vol. 25 (pp. 302G–307F). Lvgdvni: Anissonios.
- Gurney, P. J., & Gurney, L. W.** (1998). Authorship attribution of the *Scriptores Historiae Augustae*. *Literary and Linguistic Computing*, 13(3), 119–131.
- Johnson, L. S.** (1991). The trope of the scribe and the question of literary authority in the works of Julian of Norwich and Margery Kempe. *Speculum*, 66(4), 820–838.
- Kestemont, M.** (2012). Stylometry for medieval authorship studies: an application to rhyme words. *Digital Philology: A Journal of Medieval Cultures*, 1(1), 42–72.
- Kestemont, M., Moens, S., & Deploige, J.** (2015). Collaborative authorship in the twelfth century: a stylometric study of Hildegard of Bingen and Guibert of Gembloux. *Literary and Linguistic Computing*, 30(2), 199–224.
- Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W.** (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, 63, 86–96.
- Kolpacoff, J. M.** (2000). *Papal schism, archiepiscopal politics and Waldensian persecution (1378–1396): the ecclesio-political landscape of late fourteenth-century Mainz* (doctoral dissertation, Northwestern University, Illinois).
- Koppel, M., Schler, J., & Argamon, S.** (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9–26.
- Koppel, M., Schler, J., Argamon, S., & Winter, Y.** (2012). The ‘fundamental problem’ of authorship attribution. *English Studies*, 93(3), 284–291.
- Levy, I. C.** (2012). *Holy scripture and the quest for authority at the end of the Middle Ages*. Notre Dame, IN: University of Notre Dame Press.
- Marriott, I.** (1979). The authorship of the *Historia Augusta*: two computer studies. *Journal of Roman Studies*, 69, 65–77.

- Minnis, A. J.** (1988). *Medieval theory of authorship: scholastic literary attitudes in the later Middle Ages*. 2nd edn. Aldershot: Scolar Press.
- Minnis, A. J.** (2006). Nolens auctor sed compilator reputari: the late-medieval discourse of compilation. In M. Chazan & G. Dahan (Eds.), *La méthode critique au Moyen Âge* (pp. 47–63). Turnhout: Brepols.
- Modestin, G.** (2007). *Ketzer in der Stadt: der Prozess gegen die Strassburger Waldenser von 1400*. Monumenta Germaniae Historica, Studien und Texte 41. Hannover: Hahnsche Buchhandlung.
- Molnár, A.** (1989). *Storia dei valdesi 1. Dalle origini all'adesione alla Riforma (1176–1532)*. 2nd edn. Turin: Claudiana.
- Mutzenbecher, A.** (1961). Bestimmung der echten Sermones des Maximus Taurinensis. *Sacris Erudiri: Jaarboek voor Godsdienstwetenschappen*, 12, 197–293.
- Patschovsky, A.** (Ed.) (1979). *Quellen zur böhmischen Inquisition im 14. Jahrhundert*. Monumenta Germaniae Historica, Quellen zur Geistesgeschichte des Mittelalters 11. Weimar: Hermann Böhlau Nachf.
- Patschovsky, A.** (1994). Ablaßkritik auf dem Basler Konzil: der Widerruf Siegfried Wanners aus Nördlingen. In J. Pánek, M. Polívka & N. Rejchrtová (Eds.), *Husitství—reformace—Renesance: Sborník k 60. narozeninám Františka Šmahela* (pp. 537–548). Prague: Historický Ústav.
- Segl, P.** (2006). Die Waldenser in Österreich um 1400: Lehren, Organisationsform, Verbreitung und Bekämpfung. In A. de Lange & K. Utz Tremp (Eds.), *Friedrich Reiser und die 'waldensisch-hussitische Internationale' im 15. Jahrhundert: Akten der Tagung Ötisheim-Schönenberg, 2. bis 4. Oktober 2003* (pp. 161–188). Heidelberg, Ubstadt-Weiher and Basel: Verlag Regionalkultur.
- Stamatatos, E.** (2017). Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1: Long papers (pp. 1138–1149). Valencia: Association for Computational Linguistics. Retrieved from <http://aclweb.org/anthology/E17-1107>
- Stover, J. A., & Kestemont, M.** (2016a). The authorship of the Historia Augusta: two new computational studies. *Bulletin of the Institute of Classical Studies*, 59(2), 140–157.
- Stover, J. A., & Kestemont, M.** (2016b). Reassessing the Apuleian corpus: a computational approach to authenticity. *Classical Quarterly*, 66(2), 645–672.
- Tse, E. K., Tweedie, F. J., & Frischer, B. D.** (1998). Unravelling the purple thread: function word variability and the Scriptorum Historiae Augustae. *Literary and Linguistic Computing*, 13(3), 141–149.
- Tweedie, F. J., Holmes, D. I., & Corns, T. N.** (1998). The provenance of *De Doctrina Christiana*, attributed to John Milton: a statistical investigation. *Literary & Linguistic Computing*, 13(2), 77–87.
- Välímäki, R.** (2016). *The awakener of sleeping men: inquisitor Petrus Zwicker, the Waldenses, and the retheologisation of heresy in late medieval Germany* (doctoral dissertation, University of Turku, Turku).

- Välimäki, R.** (2019). *Heresy in late medieval Germany: the inquisitor Petrus Zwicker and the Waldensians*. York: York Medieval Press.
- Weidmann, C.** (Ed.) (2015). *Augustinus: sermones selecti*. Corpus scriptorum ecclesiasticorum Latinorum 101. Berlin and Boston, MA: De Gruyter.
- Williams-Krapp, W.** (2000). Die überlieferungsgeschichtliche methode: Rückblick und ausblick. *Internationales Archiv für Sozialgeschichte der Deutschen Literatur*, 25(2), 1–21.
- Yule, G. U.** (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.
- Zwicker, P.** (1677). [Pseudo]-Petri de Pilichdorf contra Haeresin Waldensium Tractatus [Cum dormirent homines]. In J. Gretser (Ed.), *Lucae Tydensis episcopi, scriptores aliquot svccedanei contra sectam waldensivm, Maxima Bibliotheca Veterum Patrum, Et Antiquorum Scriptorum Ecclesiasticorum*, Vol. 25 (pp. 277F–299G). Lvgdvni: Anissonios.

Towards the Sun of Socialism: Turning our Macrosopes

Risto Turunen

Introduction

But in spite of all indifference, the sun of socialism has cast its first rays there. Even there, great and clear thoughts on the injustices of the present system are silently smouldering.¹

The optimistic quote above was written in 1903, by a labour journalist outlining the preconditions of socialism in the eastern periphery of the Grand Duchy of Finland. Characteristic of the socialist discourse of the time, he used the phrase ‘the sun of socialism’. It was one of the most important symbols of the Finnish labour movement in the early 20th century, figuring not only in newspapers, but also in poetry and red banners. Without doubt, there was something in the red sun of socialism that captured the contemporary proletarian imagination.

Many studies in social and cultural history have proven that symbols acting as ‘simplified objectifications of ideologies’ play a crucial role in the making

of political movements.² The sun is the starting point for this chapter, for we believe that this symbol carries rich temporal information from a century ago. Thus, it can be used as a symbolic key to unlock socialist perceptions of the imagined past, present and future. The breakthrough of Finnish socialism has been analysed from a variety of perspectives,³ but the focus has not been on 'temporality', that is, the way human beings experience time. There are some occasional comments on the socialist temporality in the previous research, mainly concentrating on the Marxian interpretation of history or on the future expectations in the form of socialist utopianism and eschatology.⁴ The third dimension of time, the present, has largely escaped scholarly attention. For example, the sun of socialism has been seen in the context of the future, as a symbol for a better tomorrow and freedom.⁵ The future-oriented meaning certainly existed, but we can add more interpretative depth to the investigation of the sun by also including the present in our analysis.

According to Reinhart Koselleck's thesis on temporality, the emergence of modernity, especially the unexpected rupture of the French Revolution of 1789, diminished the value of experience in forecasting the future.⁶ While Koselleck's argument concerned the German-speaking world, we argue that the General Strike of 1905 had a similar effect in the Finnish context, expanding the gap between the experiences (of the past) and the expectations (towards the future) and, simultaneously, creating a new understanding of the present. The General Strike from 30 October to 6 November in 1905 was not only a direct result but rather an active extension of the 1905 Russian Revolution to the Grand Duchy Finland.⁷ For the first time in Finnish history, workers momentarily seized a great part of political power, and this brief moment, a mere one week of imagined proletarian rule, meant that neither the old rules of politics nor old temporalities applied to the new situation. The General Strike led to a set of parliamentary reforms and to universal suffrage in 1906, and finally in 1907, just four years after the quote at the beginning of this chapter, Finland had the largest socialist party with parliamentary representation in Europe.⁸

This chapter has a threefold goal. First, regarding historical content, it constitutes a case study that tries to decipher the intriguing symbol of the rising sun and, thus, to broaden our understanding of the socialist temporality in Finland. The focus lies on the relation between the sun and the present, or more precisely, on how the sun illuminates the proletarian perception of their reality at the turn of the century. Second, methodologically speaking, we introduce 'macroscopic' approaches that allow historians to see something in the sources that is unavailable to the naked eye.⁹ In practice, this means quantifying comparable word frequencies, collocates and key collocates. Third, we describe what it means to write digital history, by sketching a simple theoretical model, which sheds a new light on the intellectual journey the scholar undertakes on her way from original sources to historical wisdom.

Relative Word Frequencies: Counting the Heartbeats of Finnish Politics?

We begin our journey to the core of the socialist sun with an already well-established practice in digital history, that is, counting relative word frequencies over time. First, we download the dataset from the National Library of Finland: the raw text files of the biggest socialist (*Työmies*, ‘*The Working Man*’), conservative-nationalist (*Uusi Suometar*, ‘*New Finland*’) and liberal-nationalist (*Helsingin Sanomat*, ‘*Helsinki News*’, and before 1904 *Päivälehti*, ‘*The Daily Paper*’) newspapers from 1900 to 1917.¹⁰ Then, we find the words referring to the present in each year by using the search string ‘nyky*’, which covers the most common Finnish words denoting the present moment: ‘nykyinen’ (‘present’ as an adjective), ‘nykyisyys’ (‘present’ as a noun) and ‘nykyisin’ / ‘nykyään’ (adverbs for the present moment).¹¹

Figure 17.1 shows a trend. The socialist newspaper *Työmies* has the highest frequency of ‘the present’ in 1900, but by the year 1904 the references to

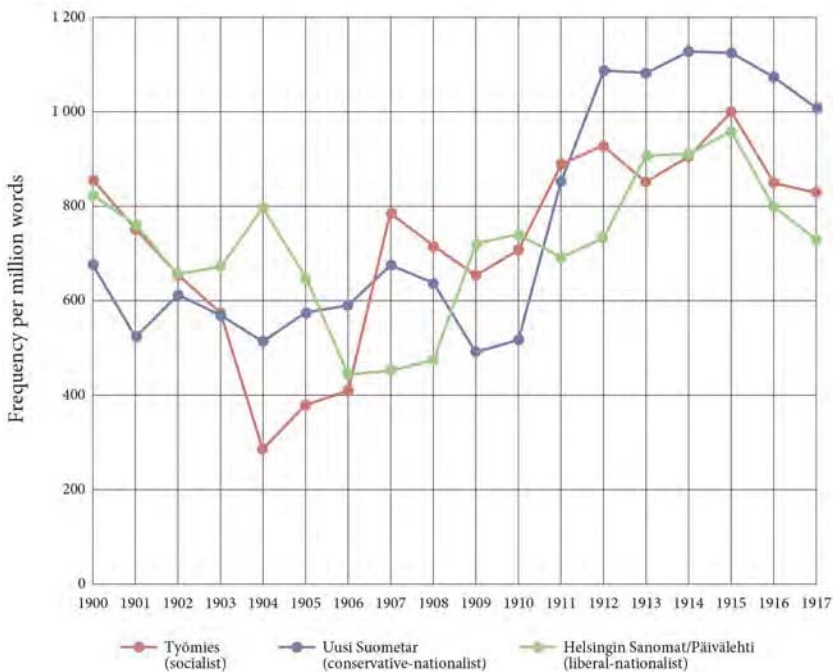


Figure 17.1: The relative amount of the present (‘nyky*’) in three newspapers, 1900–1917. Source: Raw text files of *Työmies*, *Uusi Suometar*, and *Helsingin Sanomat / Päivälehti*, distributed by the National Library of Finland, <https://digi.kansalliskirjasto.fi/opendata>.

the present have dropped far below the two bourgeois newspapers. What does this trend, this piece of information, mean? Based on previous research, censorship might lead us to the correct explanation. According to Antti Kujala, the censorship of the labour press in Finland became considerably tighter in 1903 when the Finnish Social Democratic Party officially adopted Marxism. Bans and warnings succeeded in silencing the most radical socialist discourse, for during the next year political commentary disappeared ‘almost completely’ from *Työmies*.¹² However, the General Strike of 1905 broke the silence on political matters as preventive censorship ended temporarily.

The renowned Finnish socialist poet Kössi Kaatra summarised the dramatic new temporality summoned by the strike in his poem *Suurina päivänä* (*During the Great Days*) from 1906:

It is great to be alive,
when in a single day, in a night
now we create more new things than in the work of many centuries.¹³

Reading Kaatra’s words and focusing especially on the temporal marker ‘now’, it is not a surprise that we see a sharp rise in the socialist present, especially during 1907, which happens to be the year when Finland held the first parliamentary elections. Could the new political situation (electoral speculation, campaigns and aftermath) explain the peak of 1907? Based on both close and distant reading of *Työmies*, this seems to be the case. The words such as ‘strike’, ‘government’, ‘nation, land’, ‘Duma’ and ‘senate’ increase greatly in close proximity to the present after the General Strike.¹⁴ Thus, the rise of the present means, in fact, the rise of the *political* present.

We could explain this finding in the light of Benedict Anderson’s theory of ‘imagined communities’ which argues that between 1500–1800 technological innovations and the advance of print-capitalism profoundly changed our experience of time and space.¹⁵ In the case of Finnish working people, these changes probably took place much later, beginning approximately from the mid-19th century onwards.¹⁶ When looking at the date on the front page of the daily socialist newspaper, a Finnish worker could see with her own eyes that time was moving linearly forwards day after day. In addition, she could imagine that *meanwhile* there were thousands of other workers like just like her reading the very same edition, although she had never actually met them.¹⁷ Using Anderson’s theory to explain the dissemination of socialism instead of nationalism, as it has usually been applied, guides our analysis towards the close connection between temporalities and print media, or in our case, between the socialist interpretation of the present and the Finnish labour press. Because temporalities are always constructed, they can be manipulated. The leading socialist newspaper reacted to the changing political conditions after the General Strike by accelerating the flow of time, by repeating an imperative temporal message: the time to act is *now*.

One should not forget that there might be other alternative explanations for the peak of 1907. For example, there could have been more adverts in the socialist newspaper in 1907 than before, as the adverts of the early 20th century often referred to the present in order to sell their products better. The lack of information on what constitutes the peaks and valleys of word frequencies is not a trivial problem, but rather characteristic of word frequency charts in digital humanities. Far too often, they neglect variation inside a given corpus. Figure 17.1 is slightly better than the usual combination of a relative word frequency (y-axis) and time (x-axis) in the sense that it contains the extra dimension of political affiliation. However, the figure would be even better if it showed the frequency of 'the present' in different newspaper genres (editorials, foreign section, adverts, poems, letters to the editor, etc.) for each newspaper under investigation. The distribution of genres would show us in which journalistic context the present is discussed in each major political language of the time.

Despite the weaknesses, simple word frequency charts can reveal useful, low-level information to historians. In this case, it revealed above all that the amount of the socialist present varies strongly over time. The valley of 1904 is probably due to censorship, whereas the peak of 1907 is explained by the heated political situation. However, the general trend is that all the newspapers increase their references to the present with the passage of time. Does the trend reflect the increasing heartbeat of Finnish politics, or the rise of present-intensive advertising in all newspapers, or perhaps something completely different? We do not want to get entangled in that question in the context of this chapter, but we do want to highlight the importance of keeping an open mind when attaching meanings to the figures. As doctors know, if the heart beats faster than normal, the possible causes are many and varied.

Collocation: Mining the Semantic Structure of the Socialist Present

Historians inspired by conceptual history, discourse studies or the Cambridge school of intellectual history have for long been interested in the linguistic contexts in which their historical objects of interest (concepts, discourses, intellects) figure.¹⁸ Nowadays, it is possible to quantify such linguistic contexts, given that the textual sources are in a machine-readable form. One approach to operationalise 'the linguistic context' is to define the context as all the words appearing in a window of x words to the left or right of the studied word. Since we are dealing with a highly inflected language, Finnish, it is important to lemmatise all the words in the text files, that is, to replace all word variations with their base form, before the actual analysis in order to get more reliable results.¹⁹ In our case study, we could quantify all the words that exist in proximity of five words from the words referring to the present in the three biggest socialist newspapers. Why five words? There is no magic formula for defining the

Table 17.1: The most frequent words in a window of five words from the ‘present’ (search strings ‘nykyi*’ and ‘nykyä*’) in three socialist newspapers, 1895–1917.

RANK	WORD	TRANSLATION	FREQUENCY
1	olla	to be	60,941
2	ja	and	25,893
3	se	it, that	19,693
4	että	that	14,741
5	joka	which	14,673
6	ei	no	12,174
7	tämä	this	8,618
8	ne	they	6,348
9	kun	when	4,779
10	saada	to get	4,721

Source: Lemmatised raw text files of *Työmies*, *Kansan Lehti* and *Länsisuomen Työmies / Sosialisti*, 1895–1917, distributed by the National Library of Finland, <https://digi.kansalliskirjasto.fi/opensdata>.

perfect window size. Historians must decide, usually through trial and error, which is the most appropriate selection for their own research questions.

As we can see in Table 17.1, the problem with this approach is that the most frequent words connected to the present are common words which do not reveal anything relevant from the historian’s perspective. Fortunately, corpus linguists have developed a statistically more sophisticated method in collocation analysis that produces more meaningful raw information for historians to contemplate. Collocates are words that appear *more frequently than statistically expected* in close proximity to the search word.²⁰

When looking at Table 17.2, one can immediately see that it contains useful information inviting a further human analysis. After close reading of the concordances, the list of examples of ‘the present’ as they occur in the socialist newspaper texts, we found three categories of technical errors: some collocates had, not surprisingly, OCR errors; some were lemmatised into a wrong base form; and some suffered from an incorrect word segmentation. We also increased the minimal frequency of collocation to a relatively high cut-off point of 200 instances, in order to filter out advertisements that plague all quantitative analyses of the Finnish newspaper corpus.²¹ Then, after cleaning up Table 17.2 for errors and function words, we created a simple visualisation that is hopefully easier to understand for most historians.²² Apart from absolute frequencies and Finnish originals, Figure 17.2 contains the same information as that in Table 17.2, but in a more accessible and user-friendly form. Figure 17.2 shows what could be poetically defined as ‘the architecture of the concept’.²³ It is based on the principle that the human brain can intuitively understand: the closer the word is to the centre, the more strongly it is connected to the socialist present.

Table 17.2: The collocates of the present ('nykyä', 'nykyi*') in three socialist newspapers, 1895–1917, min. frequency 200.

RANK	WORD	TRANSLATION	FREQUENCY	MI-VALUE
1	yhteiskuntajärjestelmä	societal system	354	10.29
2	tilanne	situation	1,460	9.91
3	järjestelmä	system	1,506	8.58
4	mallita	to prevail (OCR error)	448	8.44
5	vallita	to prevail	1,464	8.14
6	oleskella	to stay	404	8.11
7	yhteiskunta	society	2,026	8.11
8	tätä	this	282	7.56
9	olosuhteet	conditions	488	7.35
10	kapitalistinen	capitalist	394	7.32
11	kallis	expensive	587	6.95
12	pula	shortage	293	6.87
13	kurja	miserable	322	6.86
14	politiikka	politics	299	6.84
15	valtiollinen	state-	500	6.79
16	työttömyys	unemployment	452	6.76
17	muoto	form	801	6.68
18	asema	position, condition	3,547	6.64
19	olo	condition	1,731	6.63
20	kanta	view	897	6.56
21	kunnallinen	municipal	487	6.55
22	sota	war	1,415	6.55
23	taloudellinen	economic	658	6.53
24	vaikea	hard, difficult	295	6.52
25	käytäntö	practice	594	6.51
26	kehitys	progress, development	555	6.46
27	kurjuus	misery	280	6.42
28	säilyttää	to preserve, to save	323	6.39
29	epäkohta	grievance, shortcoming	307	6.36
30	mahdoton	impossible	485	6.35

Source: Lemmatised raw text files of *Työmies*, *Kansan Lehti* and *Länsisuomen Työmies / Sosialisti*, 1895–1917, distributed by the National Library of Finland, <https://digi.kansalliskirjasto.fi/opendata>.

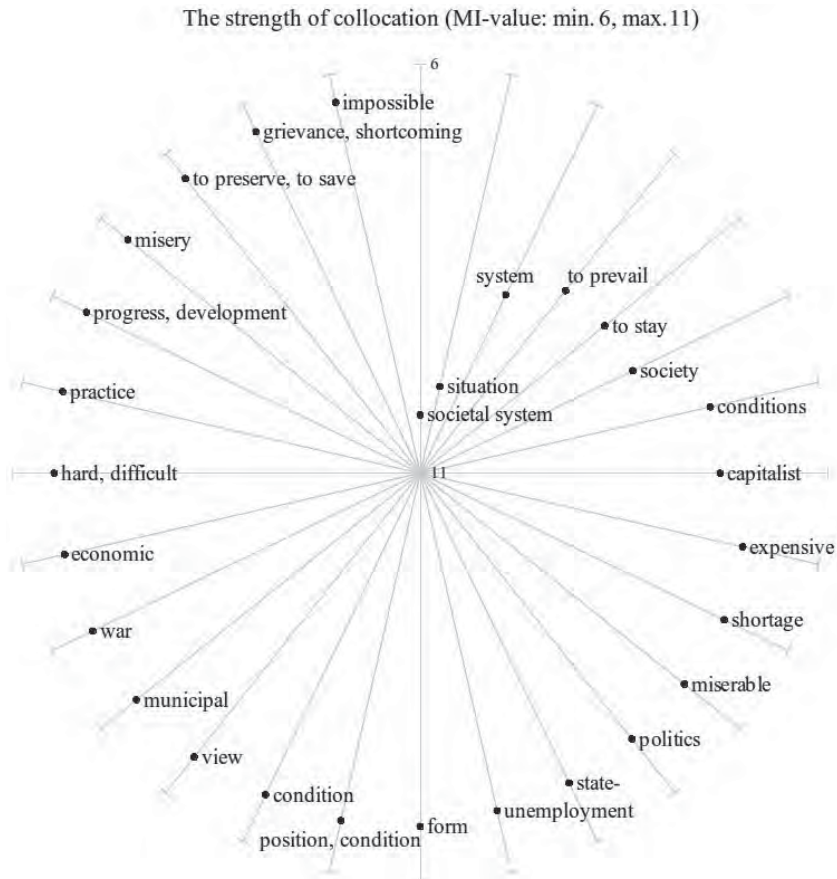


Figure 17.2: The collocates of the present ('nykyi*', 'nykyä*') in three socialist newspapers, 1895–1917. Source: Raw text files of *Työmies*, *Kansan Lehti* and *Länsisuomen Työmies / Sosialisti*, 1895–1917, distributed by the National Library of Finland, <https://digi.kansalliskirjasto.fi/opendata>.

It is relatively easy to find patterns in the figure for a historian with prior knowledge on the topic. First, the socialist present seems to attract phenomena that are considered to be negative, if not universally, at least as perceived by most people. In addition to the abstract concept of misery, the readers of the socialist newspapers were frequently introduced with the more concrete evils of 'shortage', 'unemployment' and 'war'. Negativity was enforced with the adjectives 'miserable' and 'hard, difficult', especially when talking about current 'conditions'. A close reading also reveals that current 'politics', referring to both tsarist repression and domestic bourgeois oppression, belong to the same negative semantic field.²⁴ The critique of 'politics' is understandable since the socialists, despite polling the most votes in every election, were not represented in the

government until in the year 1917.²⁵ Of course, in the socialist understanding, their interpretation of the present was not negative in an exaggerated sense, but rather a realistic portrayal of the inherent problems in capitalism.

This leads us to the second feature worth highlighting in Figure 17.2: the misery of the socialist present was not accidental, but *systematic*. The words ‘societal system’, ‘system’ and ‘society’ in the figure hint at this socialist pattern of conceptualising the present. The everyday had a structure, and those great ‘capitalist’, ‘state’, ‘economic’ and ‘municipal’ forces shaping the present were not mystical nor divine, but explainable through a rational analysis. Let us take a quote from each of the three main socialist newspapers to illustrate the logic:

The eyes of many workers have opened even here [in Eura] to see the misery into which the present system has brought our society.²⁶

The curse of the present system is precisely that the more satisfied the capitalists can be with their lives, the more miserable the life of the workers has become under the constraining conditions.²⁷

Everyday experience shows us that it is not possible to achieve sufficient improvements for the condition of the majority of the people on the grounds of the present bourgeois system ...²⁸

Thus, socialists not only disapproved of the present with negative words, but they also tried to explain the causal mechanism behind it, by arguing that the present system was the root of all evil.

Although the present was in essence systematically bad, there was hope, or as the grand old man of Finnish labour history Hannu Soikkanen has argued in his seminal study *Arrival of Socialism in Finland* (1961), ‘the present and future conditions were contrasted as starkly as possible.’²⁹ According to Soikkanen, this contrast was one of the main features of socialism that made it psychologically so attractive for the working people. The third pattern in Figure 17.2, visible to an experienced eye, refers to this connection between the present and the future, that is, to the words that imply changes and movement. ‘The present situation’ indicates a way of thinking that is not limited by eternal conditions created by God in the beginning of time. In addition, the phrase ‘the prevailing conditions’ is used regularly in the political language of Finnish socialism, in order to undermine the foundations of the present status quo:

Modern socialism is thus a product of the prevailing economic and social conditions in the present.³⁰

Class struggle is rooted in the unsolvable conflicts between employers and employees. A collective agreement cannot remove the conflict, and neither can it abolish the hegemony of the capital over work in the prevailing conditions of the present.³¹

Capital is accumulating in the hands of fewer and fewer people while the propertied class is not growing. Marx taught that the natural result is that the conditions themselves must change, there will be a fall of the prevailing system.³²

Finally, the concept of ‘progress’ is fundamental to understanding the socialist temporality, for it tied together the past, present and future. If the present capitalist society was indeed a temporary product of a historical process, then society would surely change in the future too, or like one socialist journalist foretold: ‘By the force of historical progress, the present system of oppression shall be once wiped from the stage, and its wretched henchmen shall get the reward they deserve.’³³

We have seen that the collocation method, combined with different visualisation techniques (tables, figures), can produce a massive amount of low-level information on the semantic content of a concept in historical texts, in this case on the concept of the present in three leading socialist newspapers. It would have taken several years, perhaps a decade, to manually close read the more than 180 million words printed in these newspapers. However, a computational distant reading helped us to discover that the socialist present was (1) negative, (2) systematic and (3) changeable.

In the end, it always depends on the skills of a historian whether or not elementary quantitative information is successfully transformed into historical knowledge. Here, instead of limiting our critical thinking only to the meanings of preliminary ‘results’, we should also inquire into the presuppositions embedded in each quantitative method. For example, from a historian’s point of view, the collocation method is lacking comparative contexts for it operates only within the political language of socialism. How do we know if these found features of the socialist present are unique, or if they belong to a more general discourse of the time?

Key Collocation: Placing the Socialist Present into the Contemporary Context

The collocation method shows the strength of mutual relation between two words. Another useful method historians interested in language could borrow from corpus linguistics is the keyness method, which can be used to show differences between two discourses. Keyness detects the words which appear more frequently than expected by pure chance in the text collection A (‘target corpus’) compared to the text collection B (‘reference corpus’).³⁴ Next, we combine the main ideas behind the two methods under the concept of key collocation, which aims to reveal semantic differences in the use of a certain historical concept.

First, as previously, we collect all the words appearing in a window of five words of the present (using search strings ‘nykyä*’ and ‘nykyi*’) in the socialist newspapers, and combine these words into one unified corpus. Then, we

do exactly the same for the bourgeois newspapers, in this case for the biggest liberal-nationalist newspaper *Helsingin Sanomat / Päivälehti* and the biggest conservative-nationalist newspaper *Uusi Suometar*. Now we have two corpora, and we can utilise the keyness method in order to see which words appear more frequently in the socialist discourse on the present compared with the bourgeois discourse on the present.

Table 17.3 places our findings on the socialist present in the previous section into the wider context of the early 20th-century newspaper discourse. The only clearly negative word in the top 20 key collocates is ‘unemployment’.

Table 17.3: The key collocates of the socialist present (‘nykyä*’, ‘nykyi’), compared with bourgeois newspapers, 1895–1917.

RANK	WORD	TRANSLATION	KEYNESS	EFFECT SIZE
1	työläinen	worker	1,917	19.8
2	yhteiskunta	society	1,586	5.8
3	oma	own (OCR error)	900	2.1
4	kapitalistinen	capitalist	598	38.5
5	köyhälistö	proletariat, the poor	587	28.7
6	yhteiskuntajärjestelmä	societal system	546	53.8
7	liitto	union	539	3.1
8	torppari	crofter	524	5.6
9	työmäki	working people (OCR error)	505	13.6
10	työ	work	504	2.0
11	työväki	working people	448	4.2
12	tilanne	situation	432	2.3
13	järjestelmä	system	422	2.3
14	työttömyys	unemployment	420	6.8
15	työnantaja	employer	414	5.8
16	työmies	working man	390	3.2
17	palkka	wage	368	2.3
18	moida	can (OCR error)	366	2.8
19	järjestö	organisation	357	5.3
20	ammattillinen	occupational, vocational	321	33.7

Source: Lemmatised raw text files of *Työmies*, *Kansan Lehti*, *Länsisuomen Työmies / Sosialisti*, *Helsingin Sanomat / Päivälehti* and *Uusi Suometar*, 1895–1917, distributed by the National Library of Finland, <https://digi.kansalliskirjasto.fi/opensdata>.

However, scrolling further down the list shows that socialists indeed use words such as ‘misery’ (ranked 24 by keyness value), ‘miserable’ (27), oppression (87) and ‘hunger’ (111) in close proximity to the present much more often than their contemporaries, leading us to the conclusion that the level of negativity in the socialist discourse on the present was extraordinary. Socialists imagined the worst possible present.

What about the systematic nature of the present we encountered when quantifying collocates? It exists also in Table 17.3 in the form of ‘society’, ‘capitalist’, ‘societal system’ and ‘system’. In addition, we have strong supporting evidence of ‘social order’ (76), ‘economic system’ (80), ‘production system’ (86), ‘class society’ (205) and ‘system of oppression’ (359).

The third feature of the socialist present, changeability, seems to be the least unique to labour newspapers. Apart from ‘situation’, words referring to the *changing* and *changeable* nature of the present are missing from Table 17.3. Some of them can be found in the key collocation list: for example, ‘reaction’ (70) and its counter-concept ‘progress / development’ (900). However, looking at the whole list of key collocates, it seems that this feature of the socialist present does not stand out in the context of Finnish newspaper discourse. Perhaps all the major political languages of the time—from liberalism, conservatism and socialism to Lutheran Christianity—believed that the world was changing, but they had different interpretations of what exactly was changing, how fast and, above all, if these changes happening in the present were leading to a better or worse society in the future.

It can be intellectually satisfying to find confirmation of prior interpretations, but nothing compares to finding something new. What is new is that the socialist protagonists of the present differ starkly from the bourgeois ones. The socialist version is based on the antagonism between good (‘worker’, ‘working people’, ‘working man’, ‘proletariat / the poor’, ‘crofter’) and evil (‘employer’) actors. This fundamental feature of the present was not found in the traditional collocation analysis, for the antagonism is so deeply rooted in the overall socialist discourse that it does not specifically stand out in the context of the socialist present. This socialist tendency to construct political agency through a vigorous repetition of collective singulars, especially ‘the working people’ and ‘proletariat / the poor’, can only be invoked when comparing socialism to other political languages of the time, in this case with the help of the key collocation method. Correspondingly, the trade union jargon (‘union’, ‘organisation’) escapes the collocation analysis, but it is clearly visible in the list of key collocations.

While collocation concentrates on the architecture of a concept in isolation, within only one discourse, key collocation can reveal the uniqueness or generality of these historical conceptual architectures. In the case of the socialist present, the latter method seems to confirm most of the findings achieved in the collocation analysis. Nevertheless, we should also respect the fact that an opposite result was possible. For example, we know that a negative present is not a feature confined to the socialist discourse of the early 20th century (old

people complained about present children and manners already in the days of Plato³⁵), but the point is that key collocation gives us a comparative empirical context, against which we can measure how much (for example, ‘negativity’) is much. Constructing comparative contexts through traditional close readings is a labour-intensive task. Perhaps this is one reason why historical temporalities have been analysed based on a rather limited amount of sources.³⁶

From Sources to Wisdom: DIKW for Digital Historians

In this final part before the concluding remarks and return to the sun of socialism, we rise from the empirical case study to a more abstract level, by providing a theoretical account of our intellectual journey so far with the help of the so-called DIKW pyramid, a concept that has been influential in the information sciences, knowledge management and systems theory for decades.³⁷ The pyramid describes the hierarchical relations between Data, Information, Knowledge and Wisdom. Although the pyramid has received criticism from several directions,³⁸ we believe that a slightly revised version of the pyramid can be useful for explaining not only the analytical process of this chapter, but also, more broadly, the idea and promise of distant reading in the context of digital history.

The vertical axis in the model represents what is usually described as ‘connectedness’. Connectedness increases as we climb up the ladder towards wisdom.³⁹ The idea is not completely unfamiliar to historians, but we traditionally prefer the word ‘context’ when describing the process of historical analysis. In fact, the etymological root of context means weaving or joining together.⁴⁰ Thus, we can replace ‘connectedness’ with ‘context’ in the model without a bad conscience.

Then, we should add one layer below data, that is, historical sources.⁴¹ The central difference between sources and data to digital historians is that the latter is machine-readable, and currently only a small part of historical sources is available for computational analysis as data. In the context of this chapter, physical historical newspapers are sources, whereas their digital representations—the PDF images and text files we downloaded from the National Library—belong to the category of data.

What is information, then? Here, we differ from general definitions of information as ‘data + context’, or ‘data + meaning’,⁴² for the words ‘context’ and ‘meaning’ carry too much historical weight in the humanities. ‘Information as processed data’ is a more suitable definition for our purposes.⁴³ Examples of information would be simple word frequency time series (for example, Figure 17.1), word frequency tables (for example, Tables 17.1–17.3) or visualisations of words appearing close to one another (for example, Figure 17.2). In each of these examples, raw data has been computationally re-organised into low-level information.

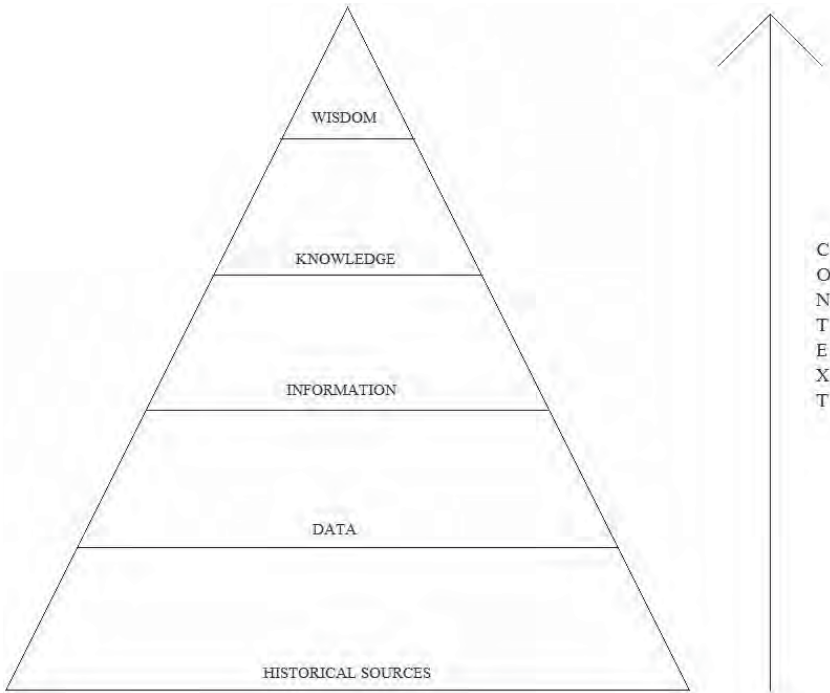


Figure 17.3: The revised version of the DIKW pyramid. Source: Author.

In our model, information does not include the historian's interpretation of information, which is located one step higher in the pyramid. Knowledge is information that is interpreted and contextualised by a human scholar. In this chapter, knowledge refers to assigning meaning to individual tables and figures, and then connecting these meanings not only to one another, but also to previous research—a difficult enterprise we undertook in the preceding pages.

The top of the pyramid is called wisdom, and it is the most controversial layer, for it escapes a clear definition.⁴⁴ Thus, it might be best to demolish it entirely. Russell Ackoff, often acknowledged as the founder of the pyramid, defined wisdom as evaluated understanding.⁴⁵ A historian could perhaps imagine wisdom as an ability to see which parts of the specific knowledge she has produced is relevant in answering the most complex questions of history. If knowledge means deciphering the meaning of the socialist sun, wisdom requires that a historian understands the meaning of this meaning under the aspect of eternity. (We have not reached an understanding this deep in this chapter.)

Now that we have reconstructed the pyramid, we can explain the point of distant reading in digital history. Distant reading aims at making the foundations of our historical explanations more solid, by piling up more stuff in the bottom of the pyramid. In other words, we want to increase the scale of historical sources, and this is possible as long as our sources are in a machine-readable

form. A historian can hardly read 100,000 words per day, whereas a computer can 'read' more than billions of words a minute. Of course, such distant reading is not reading with an understanding, but rather finding connections between words on a more primitive level.

However, we should not underestimate the fact that distant reading techniques, already in their current premature phase, can produce information that cannot be produced by human cognitive abilities alone. It is then up to a scholar to make sense of this elementary information. When a seasoned historian criticises your digital history research for 'lacking context', he or she probably means that you have not reached the level of historical knowledge, in other words, you have not paid enough attention to connecting your preliminary findings to one another and to previous research. Data nor information interprets itself. Nowadays, it is already a cliché that distant and close reading methods are complementary. With the visual aid of the pyramid, we could rephrase this idea: in digital history, the goal is to shift our limited cognitive energy upwards.⁴⁶ We distribute the most monotonous part of our historian's craft to the machines, in order that they could find patterns and trends that need human explanations. Machines are fast, precise and tireless, in other words, good at refining data to information, but, at least by now, only humans are able to refine information to knowledge.

Conclusions

What have we learned from our distant readings of the present, from turning our microscope towards the sun of socialism? First of all, as expected, the General Strike of 1905 seems to be a pivotal moment, a measurable rupture, in the history of socialist temporality. The words referring to the present increase rapidly after the strike. Metaphorically, the strike meant a mental earthquake that had long-term consequences. The strike reshaped the political environment so dramatically that old ideological maps lost much of their ability to explain contemporary reality. In this new situation, the political language of Finnish socialism turned out to be a temporary winner, gaining most votes in each of the post-strike elections. According to Hannu Soikkanen, socialism gave working people a coherent and solid world view which stood in sharp contrast to their unstable conditions.⁴⁷ We could specify this argument from the perspective of temporality by adding that a new understanding of time formed one important part in the breakthrough of the socialist world view. In those turbulent times, the political language of Finnish socialism offered the most believable interpretation of the present for the working people. As demonstrated by our collocation analyses, labour newspapers convinced their readers that the present misery was caused by the system, and this system could and should be changed.

Thus, based on our distant readings, we argue that the meaning of the socialist sun in the early Finnish labour movement was not limited to the temporal

dimension of the future. The socialist sun affected the present too, by making it appear in a new, bad light. When one saw the red sun in the future, simultaneously the shackles of capitalism came out in the broad daylight of the present. We also learned that socialists did not see the present system as eternal or divine, but as historical and man-made. In fact, we could contrast the sun of socialism with the biblical sun which was the same for everyone, or in Jesus' words: 'He (Father in heaven) causes his sun to rise on the evil and good, and sends rain on the righteous and on the unrighteous.'⁴⁸ Our distant readings, especially the key collocation analysis, revealed that the socialist sun was shining exclusively for the 'working people', 'workers' and 'the proletariat', but not for 'the capitalist employers'. Thus, unlike the biblical sun that directed people's attention towards the hereafter, the red sun of socialism highlighted earthly problems.

If we wanted an even deeper understanding of the socialist sun and temporality, in other words, if we wanted to get closer to the top level of historical wisdom in the DIKW pyramid, these same analyses (comparable relative word frequencies, collocation, key collocation) should be performed for each dimension of the time: the past, present and future. In addition, we could broaden our quite narrow focus from word frequencies to richer forms of linguistic information. For example, experimenting with verb tenses (the socialist use of past, present or conditional forms) sounds reasonable when solving questions related to historical perceptions of time.

In the end, we could speculate for a moment: if the telescope and microscope changed the fields of astronomy and biology, will the macroscope, or computational methods in general, change our historical research?⁴⁹ We believe the answer is positive in the long term, but we are not quite there yet. According to Max Weber, a new 'science' emerges where new problems are pursued by new methods,⁵⁰ but in this chapter we have mainly answered old questions by using novel tools developed outside the community of historians. While not revolutionising the field, these tools can help us to improve our craft, in our everlasting quest towards historical wisdom.

Notes

- ¹ 'Työväestön olosuhteista Karjalassa', *Työmies*, 21 August 1903, p. 1, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/728854?page=1>.
- ² The quote is from Korff 1993: 124. On political symbols in the context of socialist movements, see, e.g., Steinberg 2002: esp. 224–246; Hake 2017: esp. 100–119.
- ³ See, e.g., Soikkanen 1961; Haapala 1986; Ehrnrooth 1992; Suodenjoki 2010; Rajavuori 2017.
- ⁴ On Marxian interpretation of history, see, e.g., Soikkanen 1961: 30, 91–92, 231–232; on socialist utopianism, see Ehrnrooth 1992: 169–177; on socialist eschatology, see Huttunen 2010: 57–65.

- ⁵ Kaihovaara 1986: 34–35; Kaihovaara 1991: 52.
- ⁶ Koselleck 2004: 263–267.
- ⁷ Tikka 2017.
- ⁸ Alapuro 1988: 115–117; Eley 2002: 66, table 4.2.
- ⁹ *Le Macroscope* was the title of Joël de Rosnay’s classic introduction to systems theory (1975). Katy Börner reintroduced the concept in her often-cited article on plug-and-play macroscopes. See Börner 2011.
- ¹⁰ Available at <https://digi.kansalliskirjasto.fi/opendata?language=en>. For more information on the historical newspaper collection, see Pääkkönen et al. 2016.
- ¹¹ The absolute numbers for the search string ‘nyky*’ in 1900–1917: 72,874 hits in 96,354,158 total words in *Työmies*; 121,473 times in 157,438,611 total words in *Uusi Suometar*; and 96,923 times in 135,166,576 total words in *Helsingin Sanomat / Päivälehti*. All the calculations in the chapter were performed using AntConc (3.5.7.), available from <http://www.laurenceanthony.net/software>.
- ¹² Kujala 1995: 42–53.
- ¹³ Kaatra 1906: 4. Translation by the author.
- ¹⁴ This analysis is based on comparing the words appearing in close proximity to the present before (1.1.1904–30.10.1905) and after (7.11.1905–31.12.1907) the General Strike of 1905 in *Työmies*. In practice, all the words appearing in a window of five words to the left or right of the search string ‘nyky*’ were collected into one ‘post-revolutionary’ mini-corpus, which was then statistically compared with all the ‘pre-revolutionary’ neighbouring words of the same search string. The method is explained more carefully in the section entitled ‘Key Collocation.’
- ¹⁵ Anderson 2006.
- ¹⁶ Kokko 2016: 24–25, 44–45, 413–425.
- ¹⁷ Anderson 2006: 24–26, 33.
- ¹⁸ On conceptual history, see, e.g., Koselleck 2004; on discourse analysis, Baker 2006; on the Cambridge school, Skinner 2002.
- ¹⁹ Finnish text files were lemmatised with the LAS command-line tool. See Mäkelä 2016.
- ²⁰ There are different collocation metrics. We have used mutual information (MI) which measures the probability of whether or not the relationship between the search word and its neighbouring word is likely to exist by mere chance. The higher the MI value, the stronger the link between two words.
- ²¹ With the high cut-off point, many commercial adverts that contain words referring to the present ‘drown’ in the sheer mass of newspapers texts since not many of them circulate in each of the three chosen socialist newspapers. Thus, the ideological patterns of the socialist language use become more visible when the cut-off point is high.
- ²² The visualisation was achieved with LibreOffice (5.1.6.2), available from <https://www.libreoffice.org/>.

- ²³ de Bolla 2013.
- ²⁴ On tsarist repression in the socialist newspapers, see, e.g., ‘Suomen suhde Venäjään’, *Sosialisti*, 5 December 1908, p. 2, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/717172?page=2>; ‘Wenäjä ja me’, *Kansan Lehti*, 13 December 1909, p. 1, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/649887?page=1>; ‘Wirkaryssät uusia eduskuntawaaleja hommaa massa’, *Työmies*, 23 October 1912, p. 1, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/1187961?page=1>. On domestic bourgeois oppression, see, e.g., ‘Suhteemme hallitukseen’, *Kansan Lehti*, 4 July 1907, p. 1, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/645861?page=1>; ‘Työväen suo jeluslait’, *Sosialisti*, 6 April 1911, p. 2, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/1177250?page=2>; ‘Budjettikeskustelu eduskunnassa’, *Työmies*, 23 April 1913, p. 2, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/1187515?page=2>.
- ²⁵ Soikkanen 1975: 120–197.
- ²⁶ ‘Euran Pohjoispäästä’, *Sosialisti*, 27 April 1907, p. 3, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/702923?page=3>.
- ²⁷ ‘Länsi-Teiskosta’, *Kansan Lehti*, 12 January 1912, p. 3, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/1238591?page=3>.
- ²⁸ ‘Miksi me vaadimme järjestelmän muutosta’, *Työmies*, 23 December 1902, p. 2, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/739481?page=2>.
- ²⁹ Soikkanen 1961: 29.
- ³⁰ ‘Hiukan sosialismista’, *Kansan Lehti*, 22 June 1911, p. 1, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/1238279?page=1>.
- ³¹ ‘Kollektiiviset työ- ja tariffisopimukset Englannissa ja Amerikassa’, *Työmies*, 4 August 1909, p. 2, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/730033?page=2>.
- ³² ‘Anarkia, pakkoluowutukset ja sosialidemokratit’, *Sosialisti*, 28 September 1907, p. 2, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/703031?page=2>.
- ³³ ‘Sosialistin kirjapaino-osuuskunnan kirjapainon pakkokappaleasia’, *Sosialisti*, 25 November 1913, p. 2, <https://digi.kansalliskirjasto.fi/sanomalehti/binding/1178786?page=2>.
- ³⁴ Baker 2006: 125–128. There are different keyness metrics. We have used the log-likelihood (four-term) test in order to measure statistical significance. In addition, we have measured effect size using ratio of relative frequencies.
- ³⁵ Quintelier 2007: 165.
- ³⁶ For example, the only monograph focusing on Finnish temporality in the late 19th century is based on a correspondence of one educated family in eastern Finland. See Ollila 2000.
- ³⁷ Rowley 2007.
- ³⁸ See, e.g., Tuomi 1999; Frické 2009; Jennex 2009.
- ³⁹ See, e.g., Bellinger, Castro & Mills 2003.
- ⁴⁰ Hyrkkänen 2009: 260.

- ⁴¹ Perhaps we should add even one more layer to the very bottom of the pyramid, that is, life itself. People's rich lives leave only fragmentary sources for historians (and computers) to analyse.
- ⁴² See, e.g., Worboys & Duckham 2004: 5; Floridi 2010: 20.
- ⁴³ Ackoff 1989.
- ⁴⁴ Hoppe et al. 2011.
- ⁴⁵ Bellinger, Castro & Mills 2003.
- ⁴⁶ This idea is taken from a very different academic field: plasma physics. See Carpenter & Cannady 2004: 4–6.
- ⁴⁷ Soikkanen 1961: 30.
- ⁴⁸ Matt. 5:45. On the many meanings of the sun in the Bible, see Patterson 2011.
- ⁴⁹ Graham, Milligan & Weingart 2015: 1–2.
- ⁵⁰ Franco Moretti made Weber's quote famous for digital humanists in his pioneering article on distant reading; see Moretti 2000.

References

- Ackoff, R.** (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16, 3–9.
- Alapuro, R.** (1988). *State and revolution in Finland*. Berkeley, CA: University of California Press.
- Anderson, B.** (2006). *Imagined communities: reflections on the origin and spread of nationalism*. London: Verso.
- Baker, P.** (2006). *Using corpora in discourse analysis*. London and New York, NY: Continuum.
- Bellinger, G., Castro, D., & Mills, A.** (2003, 7 May). *Data, information, knowledge and wisdom*. Retrieved from <http://www.systems-thinking.org/dikw/dikw.htm>
- Börner, K.** (2011). Plug-and-play macroscopes. *Communications of the ACM*, 54(3), 60–69.
- Carpenter, S., & Cannady, J.** (2004, July). *Tool for sharing and assessing models of fusion-based space transportation systems*. 40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Fort Lauderdale, Florida. DOI: <https://doi.org/10.2514/6.2004-3535>
- de Bolla, P.** (2013). *The architecture of concepts: the historical formation of human rights*. New York, NY: Fordham University Press.
- Ehrnrooth, J.** (1992). *Sanan vallassa, vihan voimalla: Sosialistiset vallankumousopit ja niiden vaikutus Suomen työväenliikkeessä 1905–1914*. Helsinki: Finnish Historical Society.
- Eley, G.** (2002). *Forging democracy: the history of the Left in Europe, 1850–2000*. Oxford: Oxford University Press.
- Floridi, L.** (2010). *Information—a very short introduction*. Oxford: Oxford University Press.

- Frické, M.** (2009). The knowledge pyramid: a critique of the DIKW hierarchy. *Journal of Information Science*, 35(2), 131–142. DOI: <https://doi.org/10.1177/0165551508094050>
- Graham, S., Milligan, I., & Weingart, S.** (2015). *Exploring big historical data: the historian's macroscope*. London: Imperial College Press.
- Haapala, P.** (1986). *Tehtaan valossa: Teollistuminen ja työväestön muodostuminen Tampereella 1820–1920*. Helsinki: Vastapaino and Finnish Historical Society.
- Hake, S.** (2017). *The proletarian dream: socialism, culture, and emotion in Germany, 1863–1933*. Berlin: De Gruyter.
- Hoppe, A., Seising, R., Nürnberg, A., & Wenzel, C.** (2011, July). *Wisdom—the blurry top of human cognition in the DIKW-model?* Proceedings of the EUSFLAT conference, Aix-les-Bains, France.
- Huttunen, N.** (2010). *Raamatullinen sota: Raamatun käyttö ja vaikutus vuoden 1918 sisällissodan tulkinnoissa*. Helsinki: Finnish Literature Society & Finnish Society of Church History.
- Hyrkkänen, M.** (2009). All history is, more or less, intellectual history: R. G. Collingwood's contribution to the theory and methodology of intellectual history. *Intellectual History Review*, 19(2), 251–263. DOI: <https://doi.org/10.1080/17496970902722932>
- Jennex, M.** (2009, January). *Re-visiting the knowledge pyramid*. Proceedings of the 42nd Hawaii International Conference on Systems Sciences. Waikoloa, Hawaii. DOI: <https://doi.org/10.1109/HICSS.2009.876>
- Kaatra, K.** (1906). *'Suurlakkokuvia' ym. työväenlauluja*. Hämeenlinna: Boman & Karlsson.
- Kaihoavaara, P.** (1986). *Juurella lipun punaisen: Työväenjärjestöjen liput—osa työväenkulttuuria*. Helsinki: Yhteiskunnallinen Arkistosäätiö.
- Kaihoavaara, P.** (1991). Työväenliikkeen lippukulttuurin symboleista. In U. Peltonen & K. Stenvall (Eds.), *Myytit ja symbolit: Kirjoituksia suomalaisista kulttuuritulkinnosta* (pp. 47–61). Väki Voimakas 5. Tampere: Finnish Society for Labour History.
- Kokko, H.** (2016). *Kuviteltu minuu: Ihmiskäsityksen murros suomenkielisen kansanosan kulttuurissa 1800-luvun puolivälissä* (doctoral dissertation, University of Tampere, Tampere). Retrieved from <http://urn.fi/URN:ISBN:978-952-03-0282-5>
- Korff, G.** (1993). History of symbols as social history? Translated by H. Drost. *International Review of Social History*, 38(Supp. 1), 105–125.
- Koselleck, R.** (2004). *Futures past: on the semantics of historical time*. Translated by and with an Introduction by K. Tribe. New York, NY: Columbia University Press.
- Kujala, A.** (1995). *Venäjän hallitus ja Suomen työväenliike 1899–1905*. Helsinki: Finnish Historical Society.
- Moretti, F.** (2000). Conjectures on world literature. *New Left Review*, 1(January–February), 54–68.

- Mäkelä, E.** (2016). LAS: an integrated language analysis tool for multiple languages. *Journal of Open Source Software*, 1(6), 35. DOI: <https://doi.org/10.1109/HICSS.2009.876> 10.21105/joss.00035
- Ollila, A.** (2000). *Aika ja elämä: Aikakäsitys 1800-luvun lopussa*. Helsinki: Finnish Literature Society.
- Patterson, R.** (2011, 23 May). *Sun, light, and the Son*. Bible.org. Retrieved from <https://bible.org/article/sun-light-and-son>
- Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., & Mäkelä, E.** (2016). Exporting Finnish digitized historical newspaper contents for offline use. *D-Lib Magazine*, 22(7–8). DOI: <https://doi.org/10.1045/july2016-paakkonen>
- Quintelier, E.** (2007). Differences in political participation between young and old people. *Contemporary Politics*, 13(2), 165–180. DOI: <https://doi.org/10.1080/13569770701562658>
- Rajavuori, A.** (2017). *Esityksen politiikka: Sosialistinen agitaatio keskisuomalaisella maaseudulla 1906–1908*. Helsinki: Finnish Society for Labour History.
- Rowley, J.** (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 33(2), 163–180. DOI: <https://doi.org/10.1177/0165551506070706>
- Skinner, Q.** (2002). *Visions of politics: regarding method*, Vol. 1. Cambridge: Cambridge University Press.
- Suodenjoki, S.** (2010). *Kuriton suutari ja kiistämisen rajat: Työväenliikkeen läpimurto hämäläisessä maalaisyhteisössä 1899–1909*. Helsinki: Finnish Literature Society.
- Soikkanen, H.** (1961). *Sosialismin tulo Suomeen: Ensimmäisiin yksikamarisen eduskunnan vaaleihin asti*. Porvoo: WSOY. DOI: <http://doi.org/10.31885/2018.00017>
- Soikkanen, H.** (1975). *Kohti kansanvaltaa 1. 1899–1937: Suomen Sosialidemokraattinen Puolue 75 vuotta*. Helsinki: Suomen Sosialidemokraattinen Puolue, Puoluetoimikunta.
- Steinberg, M.** (2002). *Proletarian imagination: self, modernity, and the sacred in Russia, 1910–1925*. New York, NY: Cornell University Press.
- Tikka, M.** (2017). Strike in Finland, revolution in Russia: the role of workers in the 1905 General Strike in the Grand Duchy of Finland. In M. Hilson, S. Neunsinger & I. Vyff (Eds.), *Labour, unions and politics under the Northern Star: the Nordic countries, 1700–2000* (pp. 197–217). New York, NY: Berghahn Books.
- Tuomi, I.** (1999). Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *Journal of Management Information Systems*, 16(3), 103–117.
- Worboys, M., & Duckham, M.** (2004). *GIS: a computing perspective*. 2nd edn. Boca Raton, FL: CRC Press.

Part 4
Epilogue

Concluding Remarks: The Universal Borderlands of the Digital Humanities

Jo Guldi

In old-fashioned social history, the study of the ‘common landscape’ used to serve an index of social and cultural difference.¹ Take two regions, two neighbourhoods or two houses, side by side: their differences illuminate cultural ideas about hierarchy, the reality of divergent incomes, and separate relationships with the material world. Just so, the current volume invites us to conduct a survey of the ‘common landscape’ of digital history in all its variation.

‘Field’ though it might be in name, the domain of history as practised by scholars of different methodological and political orientations, geographical and temporal subjects of study, and institutions around the globe is really more of a patchwork of different fields and sub-fields, connected by an infrastructure of main-travelled roads and divergent footpaths that only precariously serve the whole. Some of these fields are closely guarded by an embattled elite, others plowed by an army of workers, still others remote provinces known only to a handful of toilers. Here and there, social historians and

business historians have already harvested crops for generations. The bumper harvest of the future promised by new technology frightens some with its scale and lack of care, a question addressed by at least two chapters in this volume. In other places, however, tinkerers deploy the new algorithms to cottage-sized gardens of their own liking. Whether they garden with medieval heretics or parliamentary discourse, these cottage-gardeners are increasingly dabbling in some new technology—be it topic models, vectors or other tools found in this volume. Their work is just the garden. The landscape itself is changing as a result of these multiple efforts: it is a changing ecosystem, rebalancing in reaction to human labour, sometimes enclosed or guarded.

The landscape metaphor, to be sure, obscures strong forces of historical change. In Chapter 4 above, Mats Fridlund offers an oceanic image for history and its trends. Forces whence we know not where transform the discipline, moving everything with them. For Fridlund, technology is the wave: in the modern era, technology is the reigning factor. Even scholars who eschew quantitative methods depend upon the word processor, JSTOR articles and newspaper indexes that self-proclaimed non-digital historians depend upon. Using the analytical insights of the history of technology, Fridlund deconstructs the digital, and argues that, like nature, technology is always already with us.

Indeed, there is something elemental in the transformation of scholarship in the modern era. New work with algorithms potentially participates in such a tide, and what is thrilling about it is the sense that any scholar, anywhere in the world, might contribute to its movement. Approaches to the study of history developed by one cadre of researchers working on 19th-century Finland may rapidly translate to studies of 19th-century Britain, of the 20th-century United States, or of medieval China.

Yet, the diversity of different specialisations, periods and interests persists, despite the waves and winds that blow through from time to time. As the rich and diverse studies of this volume have demonstrated, digital history is not so much a field or sub-field in this rich and varied landscape, as a universal approach to history. The practices and methods of digital history are transforming plots here and there across the entire landscape: here a neural net, there a topic model, elsewhere a map or a social network, spanning the entire range of periods, geographies, orientations and institutions, such that no field of historical scholarship, however remote, is today too far away from some garden bed where some scholar has applied a computer-aided practice to their labours.

For proof of this intermingling diversity, one has only to look at the rich set of different algorithms and questions in this volume. Each chapter contributes a different orientation and algorithm, such that the volume as a whole surveys a wide variety of different times, places and orientations. Nonetheless, the neural nets used to study medieval heresy have been used elsewhere to analyse the history of photography, and the topic models used here to analyse German humanist discourse have been used elsewhere to study the history of British parliamentary debates about infrastructure and 20th-century American newspaper coverage of cities.

Surveying digital history thus implies looking over the whole of the rich landscape of spaces that constitute traditional history. All of its periods, geographies and intellectual orientations are being reworked according to collective investments, new approaches, and the questions and problematics opened up thereby.

But is a common set of approaches, shared between knowledge-workers, necessarily the same as a hegemonic project of empire? What is the difference between a ‘universalising’ set of common practices and a ‘universal’ history, a single narrative that is supposed to provide a final answer for all times and places? Will the oceanic wave of digital history drown out the lovingly cultivated diversity of the cottage farmer?

In this chapter, I will set out to answer these questions, exploring how the methods of the digital historians exemplified in this volume overlap with the universally applicable discourses of critical theory, how far the two are exclusive and where they are now collaborating together to forge a new set of critical approaches to the study of time. I will use the examples in this volume to test and refine earlier assertions about the trajectory of digital history with regard to the *longue durée*, microhistory, critical theory and politics. I will conclude that the imprint of the *longue durée* is clear; the political implications of practising digital history less so. Along the way towards answering those issues, this chapter will remark on several major features of the common landscape of history as it is now evolving.

This chapter will therefore raise questions about the institutional, national and geographic alignments that make ‘doing digital history’ possible, reflecting further on some of the themes of institutional investment covered in this book. It will raise questions about the future institutional geography of scholarship, drawing on the implications raised by this volume—one of the most methodologically and forward-thinking volumes on digital history at present—coming from a European nation whose sometime borderland status flags important changes in the geography of scholarship under the digital turn.

Finally, this chapter will look ahead to new trends in scholarship visible in the contributions from this volume, notably: a rise of methodological articles that model the ‘bridge’ between close and distant readings, the rising importance of scholarship that targets the ‘fit’ between questions and algorithms, and the increasing importance of international institutional collaboration. As the chapters in this volume demonstrate, digital history is well on its way to establishing theories and methodologies that satisfy these most critical criteria of scholarly investigation; in the years that follow, scholars who pay attention to these themes will have even more to look forward to.

The Universalism or Common Space of the Digital Humanities

The discussion of methods herein, one might say, offers a truly ‘universal’ tendency for scholarly exchange, and it is worthwhile pausing to understand what we mean by that. In another volume, the reader who specialises in medieval

history might flip only to the chapter written by the medievalist. But it is a truism that in digital history, and the digital humanities more generally, to pass over the other sections would be an error. The methods developed for one sub-field will be relevant to the next sub-field tomorrow if they aren't already today.

We see in these critical reflections the shape of new standards for historical work as scholars puzzle over the fit between particular algorithms and questions. In Chapter 16, Välimäki and his co-authors (Reima Välimäki, Aleksis Vesanto, Anni Hella, Adam Poznański and Filip Ginter) used neural nets to confirm the authorship of the *Refutatio Errorum*, an anti-heretical treatise from Germany in the 1390s, establishing, in the process, a new standard for author detection. In Chapter 15, Heidi Hakkarainen and Zuhair İftikhar prove that topic models are well-suited to engaging Koselleck's idea of concept history, linking concepts with temporality. The chapter describes a vast culture of experimentation and discovery, as scholars try out competing algorithms, testing the fit of each method to the scholarly questions independently identified as problematic in each field.

To dub such a capacity for common meeting 'universalism' is to underscore that anyone might play with any of the ideas at stake, even while an enormous pluralism existed of period, subject and political orientation. It would certainly not imply the universal applicability of any one fact or conclusion reached by the research, which is subject, as all historical research is, to the revisions of new discoveries, new archives and new approaches. Perhaps an even better term than 'universalism' for what we are investigating would be the 'common space' of the city that Hannah Arendt identified as a metaphor for the best strivings of both Enlightenment and democracy: that they could be accessed by anyone, that they enlivened and illuminated all lives that touched them.²

By employing the image of 'common space' in the city as a metaphor for a certain aspiration in discursive activity, Arendt underscored the question of access: since the European Enlightenment, the modern city was defined by spaces of equal access, spaces that didn't shut people out. We might say the same about digital history: whatever the critique from the outside, the practitioners of digital history have taken pains not to shut out any practitioners whomsoever, and many of them have worked at length to convert digital tools into material for critically inspecting empire, gender and race.³

What about the contention that digital history is itself imperial and *universalising* in nature, threatening to draw all history practitioners into a single method, problem of study and macrohistorical overview of the *longue durée*? A notorious example of universalising claims by biologist interlopers relates to a decade ago, and the haunting claim, made in the national media of the United States, that historians would be obviated by the coming of the computer.⁴ As digital history is actually practised, we see very little of this. More relevant is a portrait of individual scholars or scholars in small units working together to execute some new perspectival opening onto their sub-field—medieval heresy, the Finnish Parliament, the Finnish borderlands or European humanists.

Not one chapter in this volume gestures towards a universalising macrohistory of the *longue durée* that would eliminate the perspective of workers or feminists beneath the triumph of the nation or empire. Not one chapter in the whole book even claims to dispense with or obviate the field of social history, biography, intellectual, cultural or political history, although most chapters build on aspects of other forms of history in some respect. To paint the common landscape of scholarship, the artist would truly have to recognise a thousand *digital histories*, not one digital history, stretching along the plain, informed by exchanges, building on works already in progress across the land.

Another version of the complaint against digital history as *universalising* and therefore coercive borrows an image from Lawrence Stone, who in the 1970s warned against quantitative enterprises of all kinds as commandeering graduate students into massive, pyramidal projects where their intellectual inquiry is dictated from above and individual initiative is squashed.⁵ This, too, seems not to be the case. In the quarters of digital humanities where graduate students are enlisted, they are often featured as first-authors on projects that they were invited to craft with the skill and support of leaders in the field.⁶

Graduate students and early-career faculty in this volume, including Reetta Sippola and Matti La Mela, are particularly distinguished as early adopters and explorers of new technology, who have been willing to extend their training in some sub-fields to adopt new technologies, test a theory or explore a time period. The borrowing—from computer science, statistics, the digital humanities, linguistics or library science—means that these early adopters are fast at work in building up the discursive commons that Arendt so praised.

Whatever we call it, exchange between sub-fields has quite a bit going for it. Discursive commons, technological borrowing and other such scholarly programmes of exchange represent an instinct for common space or the universal that, according to science, has been waning of late in the academy. This tendency to narrow the study of the universal back to readings in one's sub-field, we learn, is not merely a habit of the humanities, but can be quantitatively identified in the social sciences and sciences as well. According to the research of sociologists of knowledge such as James Evans, scholars today generally cite fewer readings directly outside their realm of concern than did scholars a generation ago. It seems to be the case that internet-enabled web catalogues have restricted universal reading habits of borrowing from nearby disciplines, whether for critical theory or for other inspiration. The sheer overwhelming scale of available knowledge left scholars paralysed by the task of keeping up with their nearest cohort.

History seems to indicate that information economies more generally are marked by a pulse of broadening the process of collecting information and refining the information thereby collected. This pulse of information analysis has been studied across decades-long exchanges about early-modern botanical knowledge, as well as on the personal scale of Darwin's notebooks.⁷ What we are calling 'universal' moments or moments of 'common spaces' seem to be

moments of expansion, when scholars looked to particular discussions as relevant for scholars who worked across a broad variety of periods and places.

Today, the digital humanities partake of a similar moment of universalism, in which scholars of 1980s online culture, 19th-century novels and Chinese medical texts regularly meet and compare notes, finding algorithms to borrow from one another. As a result, a discussion of methods offers a meeting ground on a broad scale, as well as an opportunity to compare notes across different sub-fields and disciplinary orientations.

There have been other moments of expansion on a theoretical level, where humanities disciplines are reforged through insights from without. Such, for instance, was the impact of the steady importation of continental philosophy and critical theory into the humanities since the 1970s. A Freudian reading of Augustine could be interesting to those contemplating the nature of the biographical subject during the American Revolution precisely because the *method* could be so easily transported and applied to other fields and subjects. By the same token, a Foucauldian reading of colonial India might, in theory, interest readers from the social history of industrialisation.

Indeed, it is possible that methodological moments of unification offer a necessary antidote to the paradigm of modern specialisation of knowledge, with its tendencies to mince fields into sub-fields and further sub-fields, with the concomitant risk of knowing more and more about less and less. The arrival of critical theory in the 1970s meant that the scholar of Virginia Woolf and of Classical Athens could find both a common meeting ground and a common language in terms of inclusiveness, femininity, the knowledge of the state and the construction of the individual.

At the same time, however, the digital humanities are beginning to see a moment of critical inward inspection, of the refinement of processes and pipelines. The ‘universal’ impulse in the digital humanities is thus giving way to another phase of information analysis—one predicated upon the close inspection and comparison of algorithms, the attention to metadata, and the examination of OCR errors and named-entity detection. As Kimmo Elo points out, this domain of attention is a necessary part of the process of refinement if the garden of earthly methodologies is to bear fruit. The labour of refinement and inspection will almost certainly be a domain of work that requires the labour of historians.

If building interoperable tools and applying them to great questions of history represents the universal access of the city square—liberating with its sense of wide access and possible exchange—the work of refining metadata and inspecting tools is more like the garden plot of digital history, a place that requires focused attention and hard work to produce useful results. The metaphor is complete if we imagine that the garden of tool- and data-refinement produces useful results that can be taken back to the universal exchange of the city square. As Johan Jarlbrink points out, inspecting the results of metadata analysis through simple methods such as the ‘tally’ stands to help us unpack the ‘black box’ of digital learning.

A volume of the present kind offers a rare glimpse into how the entire breadth of history is shaped—in its temporal range, from studies in the history of Finnish feminism to medieval heretics to labour politics, and in its methodological range, from the prosecution of new frontiers with existing tools like topic modelling to the close attention to algorithms, metadata and tallies of OCR errors. The scholar who represents each period, place, theme and method has a separate body of texts and different methods, to be sure; indeed, the scholarship in question was in part selected so as to adequately represent the diversity of possible methods, approaches and statistical rigour practised across history as a whole.

In offering a meeting ground of methods and periods, a methodological volume offers an important service to the discipline as a whole. The scholar of medieval heresy may wind up later borrowing the tools from the scholar of humanism or labour politics, even if their data for the moment looks entirely different. Digital practices tend to lend themselves across formats, political interests, historiographical orientations, periods or geographies. Thus, digital tools draw historians back to a certain methodological universalism, even in the face of other kinds of plurality, insofar as they encourage practices of reading-across-boundaries, in the form of conferences or volumes, like this one, that reward the practice of rich learning in new directions.

The Pedagogical Role of Discourse

In order to enjoy knowledge as a commons, one mandate is that the experts of the commons must be motivated with an eagerness to explain, to render accessible the more difficult concepts that they have assembled for the use of others. In the era of critical theory, difficult writers from Marx to Heidegger got interpreters who translated their concepts into ready-to-wear essays: ‘Benjamin for Historians’, the beginner’s guide.⁸ Digital humanities can only claim to be a ‘commons’ accessible to all sub-fields of history insofar as it too has been equipped with multiple translation projects, rendering difficult statistics and algorithms within the grasp of the total novice.

The present volume is a monument to the pedagogical impulse of digital history. The writers assembled here have taken pains to draw down the abstractions of algorithms, statistics and databases into language of period, inquiry and method familiar (or at least tractable) for traditional historians. Each chapter introduces its method, algorithm and dataset in careful detail, presuming little prior acquaintance with historical method. Writers test and explore the possibility of ‘false positives’ raised by misinterpreting topic models. They ask about the bridge between macroscopic ‘overviews’ of the material and ‘close reading’, and how an overview can or should guide the reader back to individual texts or episodes. They seek to open up the ‘black box’ of digital analysis, to unpack and critique its workings, and to thus devise a new machine for critical analysis of the past. The result is a series of essays that

are pedagogically precise: an instructive model about how to describe a new process for the use of other scholars. These are the tactics that historians of the future should recognise as a new standard for how openly, precisely and clearly we write—for what a truly ‘universal’ or ‘common’ practice of historical reasoning looks like.

The authors of the individual chapters of this volume have taken the editors’ direction for clarity, instruction and critical thinking to heart. In Chapter 9 on feminist history in Finland, the question under investigation is the correct approach to take, and Heidi Kurvinen playfully offers her own experiences of success and failure as an example for other scholars venturing onto this unpredictable frontier of knowledge, where the rules of adequate preparation are not entirely spelled out as yet. In Matti La Mela’s Chapter 11 on the freedom to roam in Finnish parliamentary debates, he walks the reader through a step-by-step recreation of the process, suitable to educating historians with no prior use of the algorithms or techniques at stake. In Chapter 16 by Välimäki and colleagues, the authors critically explore the use of each part of their data (the examples of anti-heretical texts with known authors used to train the neural net, the authors proposed by other scholars) and each algorithm (the Support Vector Machines used to detect authorship, the vector machine used to clean the data). Computational author detection, they urge, should become the new currency of the discipline, amplifying other criteria of authorship detection such as structure, manuscript tradition and argumentation.

The writers in this volume have also shouldered the burden of offering their own interpretation of the new approaches necessary to digital history. From the ‘tensor history’ of Timo Honkela’s foreword to the ‘resource criticism’ of Mats Fridlund’s Chapter 4, their inquiries point to the importance, as history adopts new methodologies, of adequately spelling out the limits of an inquiry, the sources of the data, the silences and limits of each inquiry and the possibilities opened by particular algorithms.

I read the theoretical trajectory of this volume as a powerful demand that each historical encounter explicitly renders obvious its limits, both in terms of sources and in terms of methods. And if that seems like something that historians have already done in carefully describing their paper archives and theoretical baggage, consider this: what if every historical study that leans, in part or in whole, on secondary sources, newspapers, parliamentary debates or other digitised corpora considered not only the limits of the micro-archive, but also the limits of the macro-archive? We would have to choose forms of analysis that allow us to form an overview of the archive as a whole, for the purposes of both *longue-durée* analysis and acknowledging the historical situatedness and inherent bias of each archive. In essence, we are being invited into a new age of historical criticism, one that brings the ‘capital/periphery’ critique of post-colonial studies home in the sense of recognising the limits of data and critique, acknowledging the way in which our source-base and view of history has been shaped by power and limited all along.

The Universal Borderlands of the Digital Humanities

One token of the universalism of the digital humanities is that the contributions of a group of historians stationed mainly in Finland (rather than Britain or the United States) could dare to claim for itself so vast a title as *Digital Histories*, as if aiming to define the new field. Nor is the volume bound by an exclusive orientation to Finnish history: the subjects of this volume range across Europe. Perhaps because of the ‘universal’ tendency of digital humanities research outlined above, digital humanities research in Finland is, indeed, as diverse as the historical discipline in Finland. The present volume represents an earnest attempt to define the historiographical range of questions presented by historians working with digital techniques, to cover the range of algorithmic and metadata practices used by our colleagues, and to introduce new scholars to best practices and techniques that merit attention from scholars of any time period or geography.

The digital humanities provide an arena of study that is ‘universal’ in that it incorporates broad engagement with international currents in intellectual and cultural history, feminist history and the history of science. The topics herein contained range from medieval studies of authorship among Waldensian heretical texts to the popular reception of scientific astronomy in the 18th century. Far from being rigid and inflexible, the chapters here show off an astonishing variety of methods, each the mirror of a historical problem with its own historiographical legacy stretching back over decades. The research projects in question demonstrate something of the expansiveness of interest and time period that could be found within most national traditions.

International sharing of methods and data has been intense over the last decade, among historical practitioners, and this international intensity has raised a number of new international capitals for digital humanities research—among them digital humanities centres and nodes of excellence in Umeå, Uppsala, Venice, the Max Planck Institute Berlin, the Dutch national research infrastructure CLARIAH, the University of Sussex, the Language Bank of Finland, the University of Helsinki and the University of Turku. On a global level, digital humanities researchers in Singapore, Taiwan, China and Latin America are producing significant demonstrations of new methods. The map of digital history practitioners on the avant-garde of methodology is both more international and marked by the presence of younger universities and research institutes than a more traditional map of excellence in the humanities. With all respect to the digital humanities summer institutes that introduce many scholars to the techniques of DH, this is simply not a field of which one gains mastery by a single visit to a great master at Oxford, Cambridge or Harvard: the field is moving too quickly, with nodes of excellence developing in seemingly improbable places.

The conditions for international participation in digital history are set by trajectories that have something to do with the existence of national traditions

of history in all of those places, as well as the rise of an international information economy over the lifetime of the authors of the present volume. As Paju's Chapter 2 explains, historical researchers in Finland have experimented with digital techniques since the 1960s, much as American researchers trace their roots back to experimentations with punch cards and Latin codices in the 1950s. Like universities around the world, Finland has a national tradition of historical research with centres of excellence of teaching and learning. Finnish scholars, building on generations of institutional development, have had the opportunity to theorise important questions about statistical measures and AI, close reading and distant reading, and the role of learned societies in building and maintaining the intellectual infrastructures of today.

As such, the volume represents an event horizon within the global practice of historical scholarship which marks the rise of institutional 'borderlands' that are less well-established than the Ivy League or the ancient universities of Western Europe. Finland's distinctiveness within the digital humanities thus offers a paradigmatic path for other national bodies of researchers who wish to vie for distinction on the frontiers of interdisciplinary knowledge-making. Like many other economies in the developed world, and many in the developing world as well, Finland is heir to the information economy with all of its perquisites. Like many privileged departments in Europe, Canada, North America and Australia, Finland's history departments, libraries and language banks have benefited from an aggressive institutional programme of digitisation and support. These three preconditions—a tradition of historical study, participation in the international information economy and institutional development funded on a national or international level—make it possible for scholars and universities to mark themselves out for distinction within the space of digital history research.

The 'universal' power of the digital humanities has thus established an arena where newer institutions and national traditions of historical research can play, on equal terms, with the oldest and most distinguished universities in the world. The present volume demonstrates how scholars from a European borderland have harnessed this power to demonstrate their engagement with new methods, tools and critiques.

Digital Directions: The *Longue Durée*, Identity, etc.

Critical readers will want to know whether the digital histories of this volume are closing out or displacing other kinds of inquiry. By implication, they wonder whether departments that choose to invest in digital research are necessarily thereby foreclosing on other kinds of research strategy. The rationale behind fears such as these are located in the real, historical experience of intellectual 'turns' in the academy: one meeting ground sometimes displaces another, and this was true of critical theory. Topics of study inherited by the 20th-century humanities from the 19th century and classical precedents included the

existence of 'genres' in fiction, the search for ideal character in the genre of biography, the ideal of succeeding generations of 'reform' in the study of political regimes and the history of progress through a cascading series of perfections (whether in the form of intellectual history, the history of science or the history of technology).

As scholars came to understand the past with the aid of critical theory, old categories of research (genre, character, reform and progress) each held within them a doctrine that was itself a historical construct. What critical theory did for the liberal conscience of the university was to create a series of substitutions where a former doctrine was broken open in favour of a series of new research questions. Many of those questions were informed by politics (for example, jettisoning the doctrine of empire's beneficence in favour of a series of critical questions). Rather than taking the 19th-century agenda of matters for historical investigation as a given, it was possible to subject each of them to deconstruction and historical analysis. In the process, a new set of research questions emerged: instead of character, there was agency, and with it the question of under what conditions it became possible for an individual or a certain group to exert change over their fate or the course of collective experience. Instead of technology as the progress of inventions from one generation to the next, the history of science and technology was reborn as a series of questions about the ideology of science and technology; their affiliation with empire, masculinity and capital; the institutions that support them; and the illusion of forward progress. Critical theory buried the naïve liberal scholarship that came before it, replacing it with a series of new research questions.

Just as critical theory pushed out the set of uncritical liberal targets of research that came before it, so, it might be expected, will the new goals of digital history displace some of the focus of the scholarly record before them. It is too soon to tell what the subjects of replacement will be; the relationship between digital history and earlier generations of history is still in formation. Moreover, the number of digital history papers that directly counters an extant historical theory is very small, in comparison with those in digital literature, where scholars such as Ted Underwood and Andrew Piper have explicitly taken on some of the mainstream conclusions of the field and shown how digitally produced knowledge overturns received wisdom about, say, the idiosyncrasy of Flaubert.

While digital history remains, as yet, immature, the field as a whole is guided by theories of when, how and whether digital history will call for a revision of lasting tropes in the discipline. In *The history manifesto*, my co-author and I pushed the strongest possible case for a revolution in critical thinking abetted by access to digital tools, and we sought to describe what that might look like: in brief, we conjectured that *longue-durée* timescales of 100 years or more would displace microhistory on the scale of the human life or shorter. We advanced some related claims about the political crises of the present and the new *longue-durée* inquiries that they might provoke (shifting attention to climate change

and economic inequality over identity politics). How have those prophecies held up when it comes to the writing of ‘digital history’ five years later?

One of our guesses was the renewed importance of *longue-durée* perspectives, given the fact that digital research made trivial the repetition of cultural-studies-type analytics on long time scales using scales of material that would be impossible for the traditional reader. Just such an approach is represented, in the current volume, by several uses of newspaper corpora. In Chapter 15, Heidi Hakkarainen and Zuhair Iftikhar prove that, over 21 years, the language of Austrian newspapers demonstrates the growing importance of historicist thought. They chart the succession of disciplines through which ideas of humanism spread, a modelling of ideas like a contagion, a transition from education to the reformation to revolution and later the death penalty. They reveal how newspapers gradually connected the human, the spiritual, democracy and the future, a new language evidenced by the increasing importance of the terms *Zukunft* (future) and *Zeit* (time) to the definition of humanism. In Chapter 11, Matti La Mela uses the Finnish parliamentary debates to investigate the freedom to roam over a century. On an even longer time span, in Chapter 12, Pasi Ihalainen (with the help of Aleksii Sahala) identifies several kinds of discourse about internationalism since the founding of the League of Nations, including internationalism typically opposed to (but sometimes arising from) *nationalism*; variants of *party* and *labour* internationalism; the *spirit* of nationalism linked to ideas and values; and the promise of *peace* and *democracy* arising from internationalism, especially after 1930. A revised chronology of internationalism shows a peak in the 1970s, followed by a period of frustration.

New findings that draw from long timescales can border on the breathtaking. Perhaps the most surprising finding in Ihalainen and Sahala’s investigation of long-term discourses of internationalism was the longevity of ideas. Ihalainen found that down to the ‘leave’ side of Brexit, many of the ways of describing the promise and threat of internationalism remained unchanged since discussions of the League of Nations in the interwar period, on both the left and the right.

Much of the work that was once carried out by the ‘close reading’ of the cultural turn in the 1980s and 1990s can now be achieved with greater precision and efficacy by algorithms designed to discern, and to measure, similarities and differences of expression and sentiment, allowing the tight comparison of decades, institutions, political parties and individuals. The work in this volume aptly demonstrates that the work of the cultural and linguistic turns—concerned with the shift of lexicons and the insight this provides about historical identities and communities—are now, on their cutting edge, digital in method.

There are real challenges facing the prosecution of the *longue durée* as well. In Chapter 3, Jari Eloranta, Pasi Nevalainen and Jari Ojala point to the overwhelming scale of the archives that still await digitisation in their domain, that of the modern business history of Finland. They cite the archives of Finland’s 20th-century government administration since the 1970s, now held in the National Archives: ‘roughly 200 shelf-kilometres’ of documents essential to

understanding the history of the welfare state and neoliberalism. These archival materials would have to be scanned first before they were analysed by digital means, and they are not scheduled for digitisation. A more practical approach, for the moment, is the kind of sampling recommended by Claire Lemerrier and Claire Zalc in their *Quantitative methods in the humanities*.⁹ The facts of this dynamic raise challenges to archival institutions, grant-writing bodies in the humanities and the institutions of democracy themselves. A survey *could* be conducted by sampling documents out of that 200 kilometres of shelves, but what would be left out? Surely citizens of each region in Finland deserve the tools to monitor the history of how changes in government organisation affected their own landscape. Serving citizens requires an infrastructure and the building of tools. If such an infrastructure needs to be built for other purposes, and the archives need to be digitised, then historians can rise to the challenge of asking questions about the *longue durée* of that 200 kilometres of shelves and its analysis. Indeed, it is possible, as Ted Underwood has lately argued, that trends otherwise invisible on the short *durée* would emerge from such an analysis.¹⁰

Some of the guesses hazarded by *The history manifesto* went against the actual course of scholarship, in particular the continuing importance of research into gender, race and class. Prognostications that identity politics would be displaced by a larger, shared concern over economic inequality proved short-sighted in the face of the rise of right-wing movements around the world and the renewed relevance of identity-based activism. To counter those movements, scholars, journalists, lawyers, artists and ordinary citizens have returned to the *longue durée* of injustice in a powerful way; for instance, through the demands for reparations for slavery in Britain and America, or in the public controversy over monuments to confederate generals across the American South. Indeed, until racism, sexism and nationalism are abolished, historians are bound to ask questions about where they came from.

Thus, what we see in the new practice of digital history is not so much the displacement of critical theory by digital history, as the integration of the questions posed by critical theory on longer time spans, addressed with methods that allow the historian to fully integrate the methods of the cultural, social and linguistic turns. The natural outgrowth of these dynamics is a kind of digital history that fixes on identity and empire as its subject, exemplified by the many projects gathered and reviewed in Roopika Risam's *Postcolonial digital humanities* (2019). The present volume gathers studies in the history of Karelian borderlands and Finnish feminism. It is unsurprising that any cultural or political event could be traced at scale and in depth by digital means.

Implications and Future Directions

What are the implications of such a study for other scholars, if not to mark out a definitive set of algorithms or revisions for others? I would argue that

the output of such a volume as this is significant in that it sketches out the computational best practices of a moment. It also speaks frankly to some of the challenges ahead. It is to the work of coming challenges that I will write for the remainder of this conclusion.

The need for theorising the bridge between distant and close readings

The possibility of informing close reading with the power of synthetic tools is one of the major promises of digital history. Of his century-long examination of collocation in parliamentary discourse, Ihalainen explains, 'distant reading reveals peculiar political points that might have gone unnoticed in mere close reading or full-text keyword searches of the same documents.'

While a few of the chapters in this book begin with distant reading and end by examining particular documents, there are few examples today of historical or literary practice that moves from the distant overview down to the level of authors or other categories and into particular passages in the book, critically examining the results of a search based on the close reading of the page. Like much of the work in the contemporary digital humanities, results of distant reading are frequently given in a summary chart or single finding. Two notable exceptions offer a meditation on the 'bridge' between the distant and the close, and provide historians with a way forward. In a rich meditation on the history of close reading among historians of women, in Chapter 9, Heidi Kurvinen draws a contrast between the tools she studied in her training and the topic modelling she applied to a study of Finnish suffragettes; in the process, she offers her reflections on using the topic model as an index of different episodes, and compares some of the findings of close and distant reckoning. Similarly, in Chapter 7, Johan Jarlbrink descends from a project that surveys the effects of media on cities to critically examine the sample of cities recognised by the computer. Bridging close and distant reading becomes, for Jarlbrink, an opportunity to recognise problems in the measurements supplied by algorithmic tools on 'dirty' data. As we can see from the two examples above, the practice of distant and close reading is evolving, and new hybrids are being forged that unlock insights in the archives and highlight shortcomings in the technology.

Future approaches to the bridge between close and distant reading may do well to follow the pattern set by Andrew Piper in his recent survey of modern literature, *Enumerations*, which proceeds from distant readings of the themes and trends across poetry and the novel, down to particular authors, poems and passages, as guided by the tools of distant reading.¹¹ After all, the same tools that draw our attention to words can be used to compare individual speakers as well as parties, and indeed to draw attention to the particular paragraphs and sentences that the computer discerns to be the most exemplary cases of a particular concept and collocation pattern. That is, where a scholar learns that an important collocate of 'internationalism' is 'nation', it would be useful to learn

next which individual speaker in Parliament pairs the two words together most frequently, and the individual speech in which those concepts are juxtaposed the most. Additional measures such as these may lend confidence that particular sentences given as examples of collocation are not merely cherry-picked for their familiarity, but actually offered foundational instances of the making or re-use of the concept.

More to the point, as historians engage in moving from the small example to the big question, and from the big overview back to individual speech acts, the process of movement itself is open to methodological argument, questions of interpretation and over-interpretation. It is important that historians begin to describe their choice of exemplary passages. Thick description of the process of extraction affords an opportunity to articulate the work of human interpretation and machine contextualisation. As we examine this frontier, we will begin to understand better the application of distant reading to work on different scales, including the scale of the corpus, the author, the work, the paragraph and the word.

Theorising the difference between AI and statistical measures

In the work here, the terms ‘mutual information’ and ‘neural nets’ appear on the same page. Their basis could not be more different, however. The former is statistical and can be described, mathematically, at every step; the latter has been developed mainly from computer-science departments aiming to mirror human processes, and essentially represents a black box of pattern recognition. Some scholars in the computer sciences herald a day where autonomous intelligence will obviate human supervision in most domains, including education. Colleagues in other parts of the university are more skeptical, arguing that AI, in most cases, relies on the labour-intensive hand-tooling of research questions to algorithms. More precise and transparent answers, they suggest, are to be had from old-fashioned statistics.

Humanists are far from the centre of these debates, but our testing of algorithms and our successes and failures have implications beyond our own discipline. As Eloranta and his colleagues point out, business and economic historians have a long history of critical engagement with statistical measures such as regression and event analysis that could easily contribute to a rigorous comparison between statistical measures and unsupervised machine learning. Elsewhere, in Chapter 16, Välimäki and his colleagues rely on neural networks (the epitome of unsupervised, black-box AI), while another advances a preference for mutual information: the epitome of advanced statistics, where equally useful clusters are formed on the basis of a relatively transparent clustering formula. Because we know our textual corpora and their historical context so well, as the scholars’ research on authorship illustrates, historians are often in a better position to ‘train’ and ‘test’ the scripts of AI and to comment on how

well they work. These findings could usefully be published in the *Papers of the National Academy of Sciences* or *Science*, to the edification of other disciplines and the credit of our own.

To the degree to which historical methods border on questions of the success of AI or a preference for transparent statistics, historians have an opportunity to theorise about the stakes of one choice over another. As the discipline of digital history becomes more advanced, we should expect more work of this kind.

*Transparent documentation of the choice of algorithm, text and result
in the practice of critical search*

Elsewhere, I have argued that scholarly engagement, both traditional and digital, in general tends towards the critical examination of the choices that inform a research project.¹² Whether the choices made in an archival visit are informed by the reading of critical theory, or whether political interest in the *longue durée* drives a scholar towards particular algorithms, critical thinking about the motives and limits of particular kinds of research is always being called upon to inform the constraints of the research process. I generalised the digital research process into a pattern I called ‘critical search’, and I characterised major opportunities for critical thinking inherent in any research process. Instances of critical search in an article might include discussing the choice of keywords and algorithms and how initial choices reveal and conceal aspects of a corpus later disclosed by adjustments to the initial search. The point, in any case, is that these are forms of critical reflection that scholars are prone to in general, and through these critical reflections, the entire community of readers comes to consensus about the uses of particular algorithms, the multiple dimensions of digital archives and the interpretative questions that govern digital research.

The chapters in this volume offer freely evolved examples of critical search, in the sense that they reflect upon the process of creating knowledge with digital tools. In their search for how 19th-century Austrian newspapers described the emergent principle of humanism, Hakkarainen and Iftikhar describe multiple iterations of different kinds of topic models, configurations, with or without stop words, and how they ultimately decided on a combination of topic models with a corpus little prepared for analysis. The process of curation, decision-making and interpretation is shown to be at the heart of scholarly digital work.

In Chapter 12, Ihalainen and Sahala explain their persuasive use of information theory’s concept of ‘mutual information’ to examine lexical change over time. They use pointwise mutual information (PMI) to identify the most regular combinations of words used to describe the international, the global and the transnational in British parliamentary debates in the 20th century. Their method begins with computational work, but traces back digital findings to the text, excerpting compelling passages illustrating the rising tide of interwar optimism about international relations.

We should expect more critical engagement with the search process in the future. As a community, we are learning how to better highlight the distance between *interpretive* work and *computational* work in each research process. In making the choices behind an algorithmic deployment transparent, digital scholars acknowledge that an algorithm isn't a toaster oven, into which a neophyte puts texts in order to achieve an automatic result. Rather, the process of curation, critical inquiry, secondary reading and interpretation remain at the heart of scholarly inquiry.

Engagement with new standards of scholarship from the institutions of historical and cultural knowledge-making

Several of the chapters in this volume implicitly call for a deeper level of participation by historians and their national and international societies in following standards of data preservation, sharing and transparency. In Chapter 10, Maiju Kannisto and Pekka Kauppinen offer a detailed description of copyright issues that had to be overcome for data analysis and sharing in the domain of contemporary media analysis. In Chapter 5, Jessica Parland-von Essen argues that the institutions of cultural analysis should take care to preserve, describe and make accessible multiple layers of data, including its mark-up, analyses, tools, descriptive metadata, consent, rights and attributions of labour. In the process, she describes a potential mountain of cultural labour to be executed by the libraries, archives and IT centres of the world.

Parland-von Essen's chapter suggests a precise charge to the meetings of national historical associations and other learned societies. All of our meetings should have not merely panels for presenting new work in the digital humanities, but also panels for discussing the standards of data presentation, annotation and interoperability.

Cultural institutions (for instance, the Swedish and Finnish literary societies) have a particular role to play in setting out standards for data that is transparent and accessible. Were they to engage the questions raised in this book, in meetings, pamphlets, conferences and hiring, they would have the opportunity to shape how the caretakers of data document the many kinds of labour that have shaped the collection, as well as how practising scholars indicate that they have used data with origins elsewhere. National and international historical meetings offer an important opportunity for inviting cultural institutions and providers of data, from our museums to the private Elseviers of the world, to cooperate with scholars in following these mandates.

On the one hand, directives from national historical associations and conferences are profoundly needed. Even high-profile infrastructure initiatives such as Europeana do not currently provide workflows suited to historians' needs. On the other hand, groups of historians concerned with these issues have already assembled over generations. Petri Paju's Chapter 2 discovers a long-standing tradition for digital-historical institutions and collaborations: for example, the

Helsinki Corpus of English Texts with its origins in the 1980s, the Association for History and Computing, and the Electronic Center for History Research (later Agricola), launched in 1996 by scholars, librarians and archivists; and later the Helsinki Centre for Digital Humanities, or HELDIG, was established at the University of Helsinki in 2016. If these institutions have historically had little influence over mainstream practice, then a new generation of national historical associations and other learned societies could usefully mine the affiliates list of the digital centres for faculties already invested in the important issues of data preservation, documentation and accessibility.

International collaborations and tool sharing

The beauty of computational research is its interoperability: once a technique has been discovered for author attribution in medieval Latin, it should work nearly as well on any early Latin texts whatsoever. The same is true for parliamentary discourse: the collocates machine described by Ihalainen and Sahala should apply to any modern body of text, and the particular tactics for engaging party-political differences of lexicon should be immediately applicable to digitised records of the French *débats*, the debates of the European Union, the Canadian and Australian Hansards and the debates of the City Council of New York City, to name a few active projects. Despite these opportunities, however, there are relatively few examples of international collaborations that take advantage of the astonishing interoperability of algorithms by drawing together scholars working on similar genres of texts.

One initiative that successfully crosses these boundaries is the Oceanic Exchanges project,¹³ to which six nations (including Finland, represented by Hannu Salmi at Turun Yliopisto) contribute their historical newspapers and their technical expertise. In the case of newspapers, text-recognition technology applied in one nation can rapidly be adopted to newspapers elsewhere, so the international collaboration represents a massive virtuous cycle of exchanges. All of these efforts contribute to making the international infrastructure of future research, such that one day soon, we should expect that high-school students of history in Finland, America and Mexico will be able to keyword search the newspapers of their national traditions and compare debates about democracy and markets across nations over 200 years.

It is harder to explain why there is not such an international collaboration for the novel, for debates of democratic bodies, for the records of courts of law, for religious texts, for plays, for musical notation, for stylometric authorship attribution and for other genre-specific questions where scholars have similar questions related to form: comparing authors and chapters of the novel, for instance; comparing speakers, parties and constituencies in democratic debate; comparing kinds of charges, prosecution and defence; or judges, juries and defendants in the courts of law. In each of these genres, an interna-

tional pooling of technical expertise could result in the rapid creation of new knowledge about best methods for turning raw visual images into readable text, for turning text into data annotated with appropriate metadata categories and for deriving meaning over time. The only bar against such collaborations is one of organisation, effort and collegiality: creating useful partnerships depends on planning for international visits, sharing plans of work and the willingness to engage in a process of mutual discernment about where grants and research projects overlap.

The benefits of such collaborations are, of course, tremendous: they can result in the pooling of common goals and strategies, the most efficient use of grant money, the joint discovery of new methodologies and even the joint funding of new infrastructure that makes clean, accessible data available to all.

Notes

- ¹ Stilgoe 1982.
- ² Arendt 1973 [1958].
- ³ Risam 2019.
- ⁴ Michel et al. 2011.
- ⁵ Stone 1979.
- ⁶ Klingenstein, Hitchcock & DeDeo 2014; Barron et al. 2018; Kraicer & Piper 2019.
- ⁷ Blair 2010; Murdock, Allen & DeDeo 2017.
- ⁸ Weeks 1982; Gay 1986; Schwartz 2001; Brown 2013.
- ⁹ Lemercier & Zalc 2019.
- ¹⁰ Underwood 2019.
- ¹¹ Piper 2019.
- ¹² Guldi 2018.
- ¹³ The website of the project is <https://oceanicexchanges.org/>.

References

- Arendt, H.** (1973 [1958]). *The human condition*. Chicago, IL: University of Chicago Press.
- Barron, A. T. J., Huang, J., Spang, R. L., & DeDeo, S.** (2018). Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences*, 115(18), 4607–4612. DOI: <https://doi.org/10.1073/pnas.1717729115>
- Blair, A.** (2010). *Too much to know: managing scholarly information before the modern age*. New Haven, CT: Yale University Press.
- Brown, C. G.** (2013). *Postmodernism for historians*. London: Routledge. DOI: <https://doi.org/10.4324/9781315836102>

- Gay, P.** (1986). *Freud for historians*. Oxford: Oxford University Press.
- Guldi, J.** (2018). Critical search: a procedure for guided reading in large-scale textual corpora. *Journal of Cultural Analytics*. DOI: <https://doi.org/10.22148/16.030>
- Klingenstein, S., Hitchcock, T., & DeDeo, S.** (2014). The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences*, 111(26), 9419–9424. DOI: <https://doi.org/10.1073/pnas.1405984111>
- Kraicer, E., & Piper, A.** (2019). Social characters: the hierarchy of gender in contemporary English-language fiction. *Journal of Cultural Analytics*. DOI: <https://doi.org/10.22148/16.032>
- Lemercier, C., & Zalc, C.** (2019). *Quantitative methods in the humanities: an introduction*. Charlottesville, VA: University of Virginia Press.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. & Lieberman Aiden, E.** (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182. DOI: <https://doi.org/10.1126/science.1199644>
- Murdock, J., Allen, C., & DeDeo, S.** (2017). Exploration and exploitation of Victorian science in Darwin's reading notebooks. *Cognition*, 159, 117–126. DOI: <https://doi.org/10.1016/j.cognition.2016.11.012>
- Piper, A.** (2019). *Enumerations: data and literary study*. Chicago, IL: University of Chicago Press.
- Risam, R.** (2019). *New digital worlds: postcolonial digital humanities in theory, praxis, and pedagogy*. Evanston, IL: Northwestern University Press. Retrieved from <http://www.jstor.org/stable/10.2307/j.ctv7tq4hg>
- Schwartz, V. R.** (2001). Walter Benjamin for historians. *American Historical Review*, 106(5), 1721–1743.
- Stilgoe, J. R.** (1982). *Common landscape of America, 1580 to 1845*. New Haven, CT: Yale University Press.
- Stone, L.** (1979). The revival of narrative: reflections on a new old history. *Past & Present*, 85[11], 3–24.
- Underwood, T.** (2019). *Distant horizons: digital evidence and literary change*. Chicago, IL: University of Chicago Press.
- Weeks, J.** (1982). Foucault for historians. *History Workshop*, 14[3], 106–119.