LITERARY GEOGRAPHIES

The Construction of Space in English and German Literature 1790–1848: Geoparsing the Corvey Collection

Asko Nivala

University of Turku

Abstract:
In this article, I analyse named entity linking as a new method to study the construction of space in the English and German texts of *European Literature, 1790–1840: The Corvey Collection*. The Corvey Collection is among the most comprehensive datasets to have survived from the Romantic Era of literature. However, German-language documents in particular suffer from poor OCR scanning. To avoid noise caused by incorrectly digitized characters, I have re-OCRed the collection. In contrast to named entity recognition, named entity linking is able to disambiguate toponyms and find coordinates for them from linked open data sources such as DBpedia. I have then imported the geocoded places to geographic information systems, which enables comparing spaces imagined in British and German literature from the 1790s to the 1840s. To link spatial information to the semantic content of texts, I have applied topic modeling to find common themes shared by the works. Studying the spatial imagination of the popular texts published in the Romantic era discloses an alternative view to our present notion of Romanticism based on the close reading of a few canonized authors. The comparison of English and German corpora shows the way in which the spatial imagination reflected the asymmetrical relationship of center and periphery: the core of British literature was located in London, whereas no single center appears in the German-language data.

Keywords: geoparsing; named entity linking; geographic information systems; linked open data; Romantic studies; nineteenth century studies.

Author contact: aeniva@utu.fi

Introduction

When writing about the geography of the Gothic novel, David Punter (1999: 4) maintained that 'Gothic geography is an impossibility.' According to him, the geography of text can only have the status of a metaphor. In contrast to that, literary geography assumes that there is a referential relationship between place names and geographical space in fictional texts. For example, most readers remember only the monster from Mary Shelley's (1818) *Frankenstein*, but when it is read spatially, a striking set of locations emerges. The novel begins on the extreme periphery of the North Pole and visits Saint Petersburg, Archangel, the Rhine Valley, the Alps, and the Orkney Islands, among many other sites. As Adriana Craciun (2016: 83-98) identified, *Frankenstein* participated in the British debate about the possibility of colonizing the Polar Regions. There are good reasons to assume that the geographical references in *Frankenstein* and other Gothic novels are not only metaphors; they anchor fictional texts to their historical and political context.

Franco Moretti (1998) was among the first scholars to develop literary geography in its modern form. According to Barbara Piatti (2008: 20), literary geography asks where literature is located and why. In the field of Romantic studies, David Cooper, Christopher Donaldson, Ian Gregory, and their research groups developed a digital method called literary geographic information systems (GIS) technology. This approach uses natural language processing (NLP) methods to extract locational information automatically from full texts and import that to GIS (Cooper, Donaldson and Murrieta-Flores 2016; Donaldson, Gregory and Murrieta-Flores 2015; Donaldson, Gregory and Taylor 2017; Gregory et al. 2015; Gregory and Cooper 2009; Gregory and Hardie 2011). Michael Gavin and Eric Gidal called their similar approach geospatial semantics, which combines corpus linguistics with geospatial analysis (Gavin and Gidal 2016, 2017; Gidal and Gavin 2019).

This article studies the literary geography of the Romantic era literature (1790–1848) in Britain and Germany.[1] My aim is to explore the implied geography of fictional texts by using an algorithm to extract references to historical place names from full texts. Romanticism is usually studied by close reading of canonized works written by famous authors such as William Wordsworth, Lord Byron, or Friedrich Hölderlin (on the formation of the Romantic canon, see Gamer 2006). However, digital humanities methods allow the distant reading of thousands of texts (Jockers 2013; Moretti 2007, 2013). I will use the *European Literature, 1790–1840: The Corvey Collection* dataset to research the spatial references produced in the literary mass culture of the Romantic period. The Corvey Collection is part of the Nineteenth Century Collections Online (NCCO) digitized by Gale Cengage.

During my *Romantic Cartographies: Lived and Imagined Space in English and German Romantic Texts, 1790–1840* project, I have developed a computational methodology for the study of Romantic literary geography that applies named entity linking (NEL) to algorithmically recognize toponyms from English and German texts and study the construction of space by mapping the recognized spatial entities on historical maps. However, the Corvey Collection is based on optical character recognition (OCR) of old books, which poses a methodological problem related to noise, that is misidentified letters. I show that the quality of NEL results, especially in German, can be improved by re-OCR

scanning the whole text corpus. A more reliable OCR provides the foundation for using this corpus with my geoparsing pipeline.

The method used in this article is based on a geoparser, which combines NLP with GIS to build a literary GIS (see Gregory and Cooper 2009). However, for the recognition of geospatial entities (also known as 'geotagging' or 'toponym resolution'), I have used NEL software called DBpedia Spotlight. In contrast to named entity recognition (NER), NEL is able to disambiguate entities by linking them to DBpedia, which enables retrieval of their coordinates and plotting them on maps. However, the place names mentioned in the text alone do not include information about the thematic context in which they are used. For this reason, I have divided the texts into chunks of paragraphs and run topic model algorithm in both languages to the paragraphs. This resulted in 200 topics classifying the content of the texts in both languages, which have been geotagged with the place names mentioned in these paragraph.

Studying the spatial imagination of the popular texts published in the Romantic Era discloses an alternative for our present notion of Romanticism based on the close reading of a few canonized authors. Geospatial analysis of the Corvey Collection enables a comparison of English and German corpora and shows the way in which their spatial imagination reflected the asymmetrical relationship of center and periphery: The core of British literature was located in London, whereas no single center appears in the German-language data. On the other hand, this overview can be refined by examining some of the popular topics in the literature of the time, which have a spatial framework. I select two topics that appear in both English and German for further analysis: Romantic landscape aesthetics and sailing. Although many place names in the German-speaking corpus are in fact centers outside Germany, the German landscape topic has a more national bias. The sailing theme highlights how British Romantic literature was more colonial than German.

The Corvey Collection

Romanticism developed at the same time as the explosion of mass-produced literature. At the turn of the nineteenth century, the Stanhope press was able to double the output of the Gutenberg press, and then, in 1810, Friedrich Koenig patented the steam press. Printing technology became industrialized and the number of books published increased at the same time that the Romantic notion of literature was being formulated. Rolf Engelsing (1973) suggested that intensive reading of the same classics was replaced with extensive reading of newly published books. *Lesesucht* (reading mania) was a German term for the consumption of mass-published entertainment literature. Friedrich Schlegel, who was considered the figurehead of early Romantic movement in Germany, distinguished his concept of Romanticism from the popular fiction of the era and mentioned feminine reading addiction as a dangerous tendency of the time (Schlegel 1979: Vol. II, 330). As Michael Gamer (2006) described it, Gothic novels written and read by women influenced British Romanticism, although they were at the same time despised by Samuel Taylor Coleridge and many others.

The Corvey Collection was a product of reading addiction: it was collected to satisfy the *Lesesucht* of the Landgraves of Hesse-Rotenburg. Located in the Princely Abbey of

Corvey, near Höxter, in Germany, the library provides one of the largest collections of Romantic era literature in the world. It includes now rare popular works that were typically considered ephemera and not worthy of preservation. Because the literary taste of the original owners was biased toward English popular literature, the Corvey library provides an exceptionally good resource to study the Anglo-German cultural exchange and the Gothic novel. It is widely regarded as the most comprehensive collection of English popular fiction from the 1790s to the 1840s. It includes many very rare works and approximately 1,000 unique works that have not been found in any other libraries (Garside 1992; Steinecke 1992: 13-14).

When the Corvey library was discovered in the 1980s, its private collection impressed book historians, and its scholarly significance was immediately recognized. Since *Projekt Corvey* (1985–1999), supervised by Rainer Schöwerling at Paderborn University, the collection was used in many book history and 'humanities computing' projects during the 1990s, including the *Sheffield Hallam Corvey Project* founded in 1996, *The Corvey Novels Project* started in 1999 at the University of Nebraska, and the *British Fiction 1800–1829: A Database of Production, Circulation & Reception* project founded in 1999 at the University of Cardiff by Peter Garside. Garside also edited *Cardiff Corvey: Reading the Romantic Text* (1997–2005), which was a journal focusing on the Corvey studies.[2] The results of these projects were included in the two-volume collection *The English Novel 1770–1829: A Bibliographical Survey of Prose Fiction Published in the British Isles* (2000).

While the earlier projects were based on the Corvey Microfiche Edition, Gale Cengage digitized a significant part of its volumes and published them as a module of their commercial Nineteenth Century Collections Online (NCCO) dataset. The digitized Corvey Collection includes the full text of 3,355 English and 2,500 German titles. Moreover, the methods of the digital humanities and NLP developed rapidly during the 2010s, leading the research situation to change significantly in the 2020s. As I show next, NEL and GIS have the potential to disclose a completely new view of the Corvey Collection.

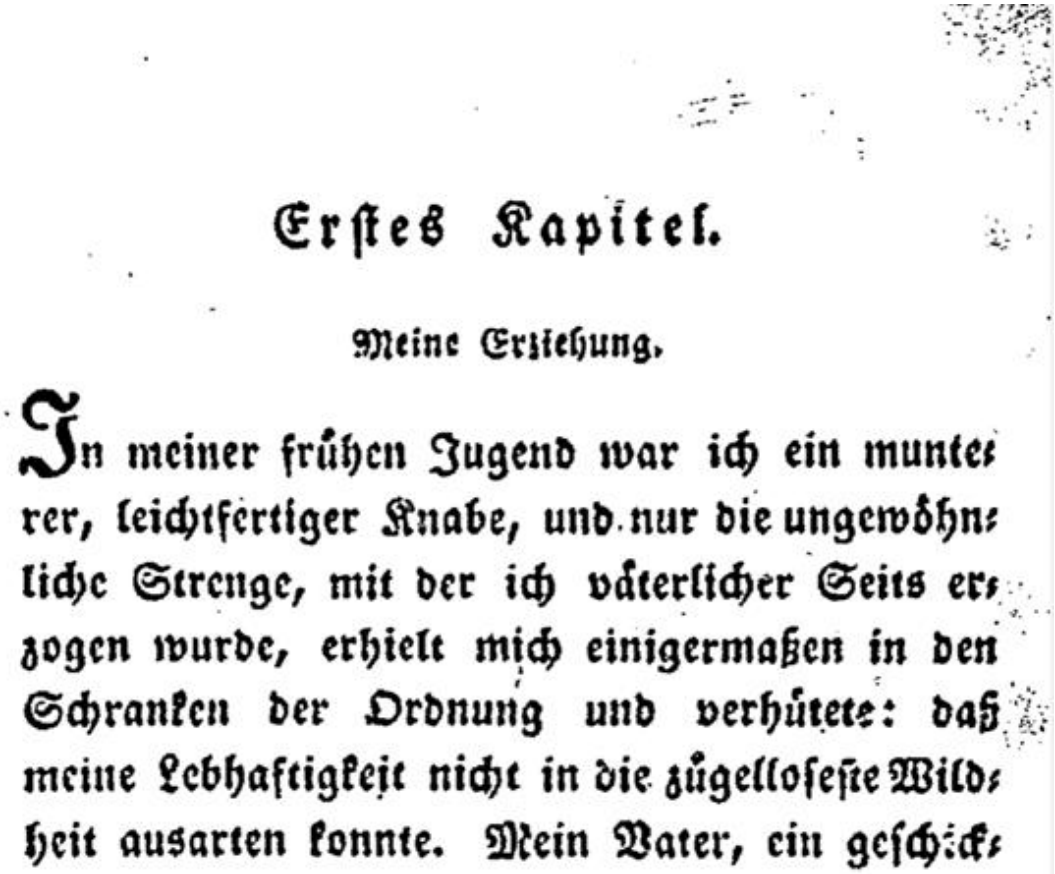
I have selected English and German documents from Corvey based on the metadata provided by Gale Cengage. It should be noted that there is a big difference in the number of texts written in these languages. A significant amount of especially the German documents in the Corvey Collection have been published before or after the 1790–1848 timeframe. Using the metadata, I have selected all titles from the time period: 3,303 English and 1,581 German titles in total. Moreover, works that have multiple volumes are divided into separate documents in the dataset. Perhaps because the three-volume novel was the standard form of publishing for British fiction during the nineteenth century, the English corpus has more multivolume titles, which raises the count of English documents even higher. Table 1 presents the number of documents (volumes) and word counts after NEL scanning and selecting only the documents that were recognized to include at least one geospatial entity.

Table 1. The number of documents (volumes) and words for each decade and language.

	documents		words	
years	English	German	English	German
1790–1799	704	16	167,461,583	8,144,781
1800–1809	1,778	88	431,381,414	45,389,007
1810–1819	1,661	364	431,045,111	152,991,332
1820–1829	2,336	882	733,650,417	357,594,746
1830–1839	958	505	338,705,482	219,749,008
1840–1848	19	151	13,511,405	95,089,348
TOTAL	**7,456**	**2,006**	**2,115,755,412**	**878,958,222**

We can see that the material in the Corvey Collection is biased toward the English language, but this can be corrected by normalizing the number of spatial entities in each corpus. What is more, the temporal series is biased toward the 1820s, which makes temporal comparisons problematic but gives a good overview of the most important decades of the Romantic era.

Re-OCRing the Corvey Collection

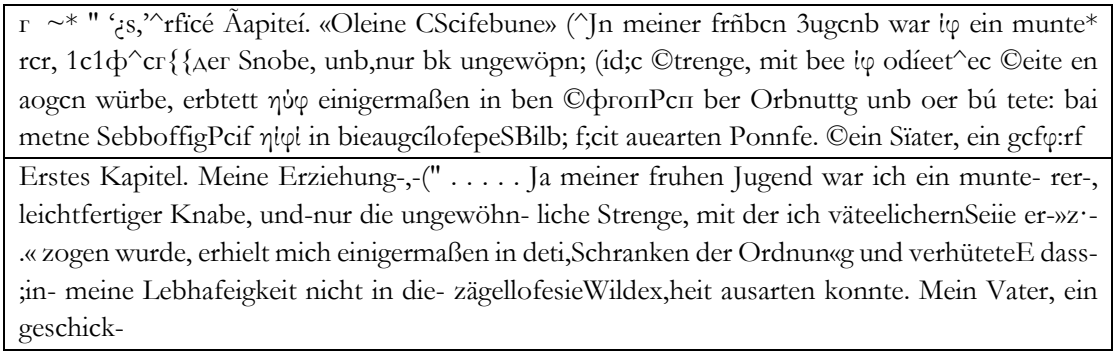


Erſtes Kapitel.

Meine Erziehung.

In meiner frühen Jugend war ich ein munte-
rer, leichtfertiger Knabe, und nur die ungewöhn-
liche Strenge, mit der ich väterlicher Seits er-
zogen wurde, erhielt mich einigermaßen in den
Schranken der Ordnung und verhütete: daß
meine Lebhaftigkeit nicht in die zügelloſeſte Wild-
heit ausarten konnte. Mein Vater, ein geſchick-

Figure 1. A sample German image file from the NCCO Corvey.



r ~* " ‘¿s,’^rficé Ãapiteí. «Oleine CScifebune» (^Jn meiner frñbcn 3ugcnb war ìφ ein munte* rcr, 1c1ф^cr{{ᴀer Snobe, unb,nur bk ungewöpn; (id;c ©trenge, mit bee ìφ odíeet^ec ©eite en aogcn würbe, erbtett ηὐφ einigermaßen in ben ©фroпPcп ber Orbnuttg unb oer bú tete: bai metne SebboffigPcif ηìφì in bieaugcílofepeSBilb; f;cit auearten Ponnfe. ©ein Sïater, ein gcfφ:rf

Erstes Kapitel. Meine Erziehung-,-(" . . . . . Ja meiner fruhen Jugend war ich ein munte- rer-, leichtfertiger Knabe, und-nur die ungewöhn- liche Strenge, mit der ich väteelichernSeiie er-»z·-.« zogen wurde, erhielt mich einigermaßen in deti,Schranken der Ordnun«g und verhüteteE dass-;in- meine Lebhafeigkeit nicht in die- zägellofesieWildex,heit ausarten konnte. Mein Vater, ein geschick-

Figure 2. A comparison of the original NCCO Corvey OCR (above) and the new OCR (below) implemented with the German Tesseract Fraktur library.

The primary problem related to the digitized Corvey Collection is the quality of OCR used to convert the scanned book pages into digital text. In general, OCR technology has improved in recent years, but it is still not perfect. First, OCR is always limited by the quality of source images. In the case of the Corvey Collection, they are scanned from microfilms that have noise and missing text. In some cases, individual characters or pages of the original books may have been damaged. Second, the Gothic (*Fraktur*) typeset used in German books needs a specially trained library for successful OCR results. In particular, the German OCR data were completely unusable at the level provided by Gale Cengage, so the first phase of the study required a new OCR scan of the entire data set in both languages.

To solve this major difficulty, all German texts were rescanned with Tesseract using the German Fraktur library. Figure 1 shows an example image file from NCCO Corvey (Bach 1812). Figure 2 compares the results of the original OCR text and my rescan. As the comparison demonstrates, the rescanning of the images improved the quality of German full texts substantially, although they still included much noise, to be similar to the English OCR texts that I also re-OCRed with Tesseract.

Table 2. Tesseract OCR confidence for each decade and language.

Years	English	German
1790–1799	81.32	71.98
1800–1809	83.55	68.64
1810–1819	85.20	65.28
1820–1829	85.53	66.35
1830–1839	86.94	70.46
1840–1848	86.05	78.36
MEAN	84.77	70.18

Table 2 shows the confidence values from the Tesseract 4 engine for English and German. We can see that English results are good, considering the low quality of the microfilm scan, while the German OCR data is now acceptable when compared with the unusably noisy starting point.[3] However, as Figures 3 and 4 show, the lower quality of German OCR will affect the comparison of NEL results, as English data is cleaner. The

situation may improve in the future, as German-language OCR libraries specializing in Gothic typefaces continue to develop. Since the number of entities found rises as the level of OCR increases, the re-OCR of both the English and German Corvey Collections has improved the results of the study, although it was a time-consuming step.

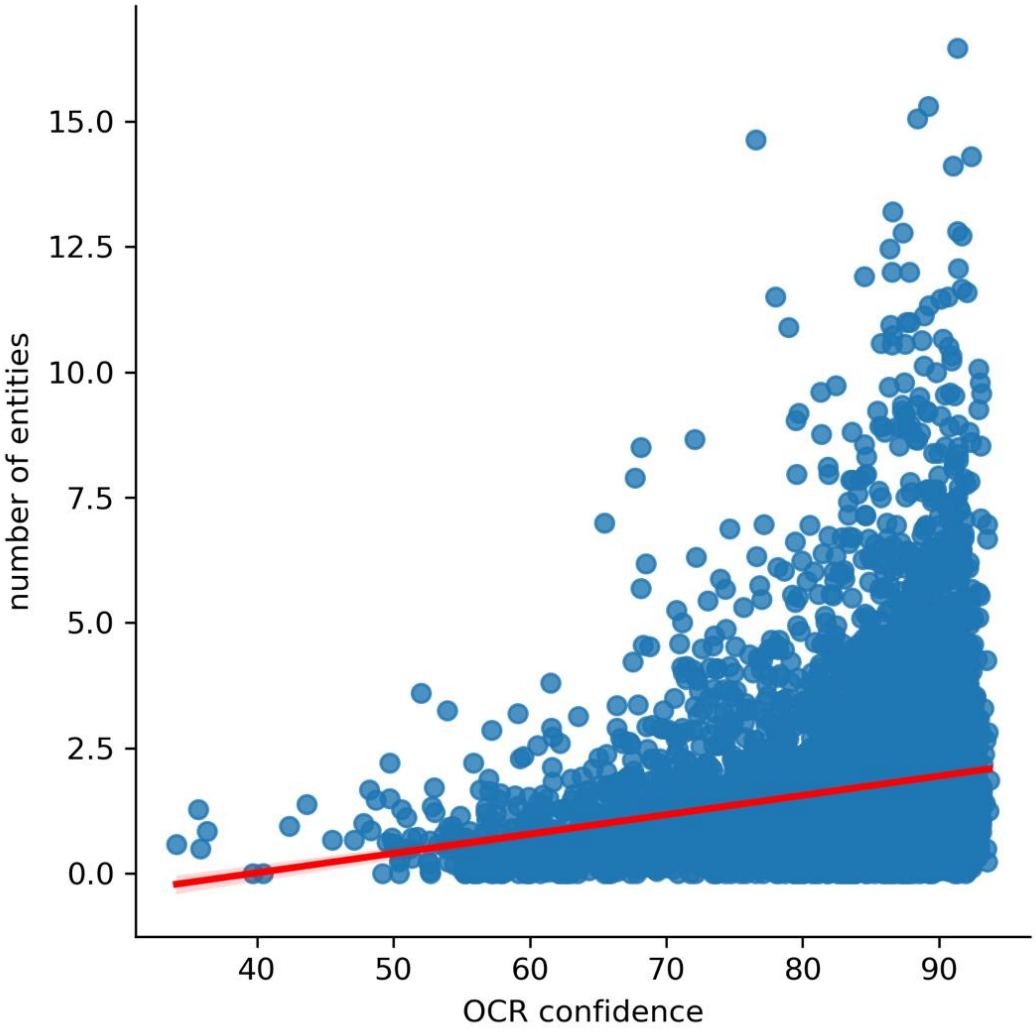


Figure 3. Scatterplot of correlation between the number of entities found and OCR confidence for English.

NEL Scanning

OCR scanning converts image files into text strings from which spatial entities can be extracted. I have used DBpedia Spotlight for this purpose (Daiber et al. 2013; Mendes et al. 2011). DBpedia Spotlight not only identifies the named entities but also links them to DBpedia. This is necessary in order to disambiguate place names and link them to geographic coordinates. NEL does not achieve as high accuracy as the state-of-the-art NER systems based on BERT (see Liu et al. 2021). On the other hand, the NER systems do not disambiguate entities, which is necessary for the geoparsing of the results. The scan

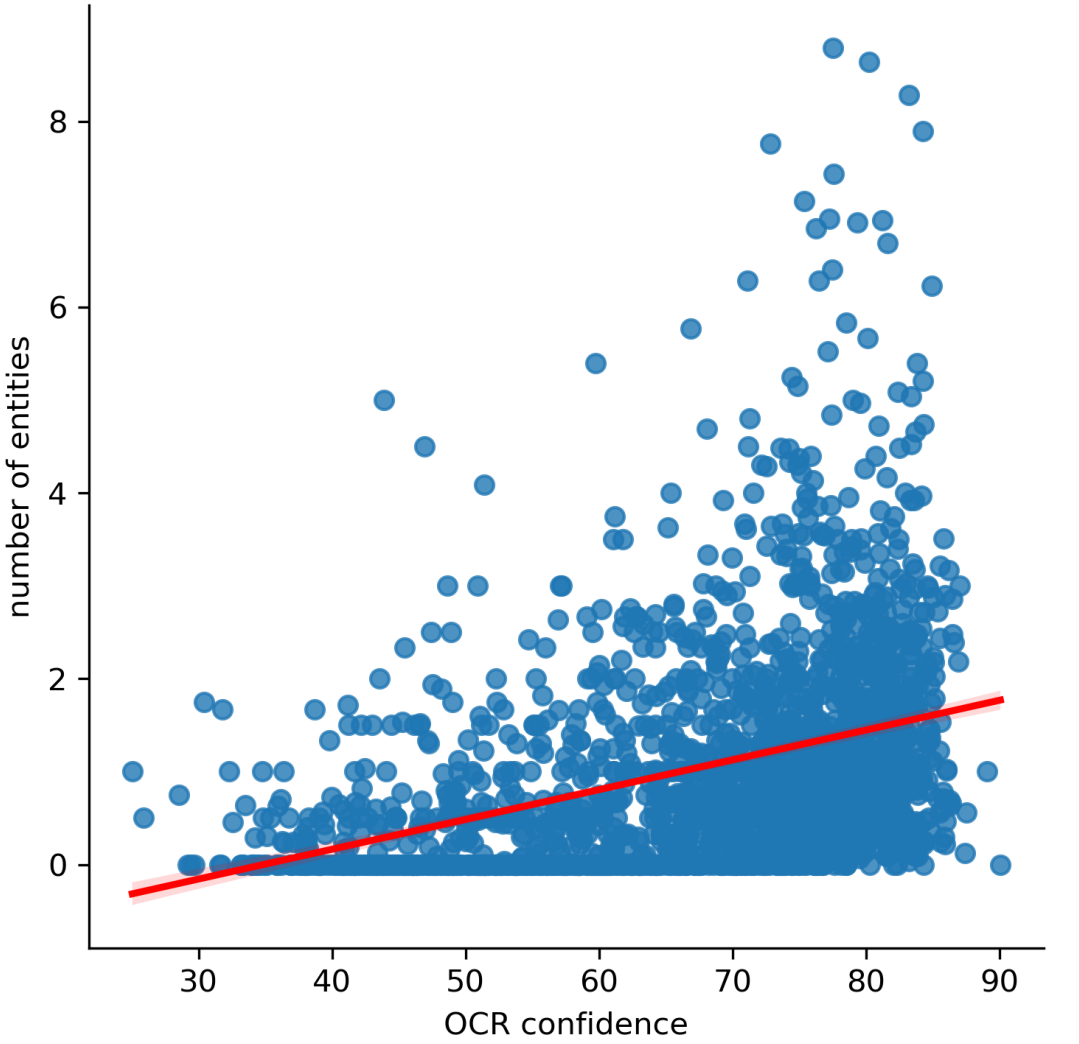


Figure 4. Scatterplot of correlation between the number of entities found and OCR confidence for German.

results contain some false positives, but DBpedia Spotlight estimates a support value for the hits, which enables filtering of the most obvious errors from the results. I have used a support value lower than 1,000 as a threshold. As Table 3 shows, the support values of English NEL results are slightly higher, but the difference is not significant.

Table 3. The mean of DBpedia Spotlight support values for each decade and language.

Years	English	German
1790–1799	64,680	37,870
1800–1809	54,825	52,314
1810–1819	49,160	55,604
1820–1829	49,822	47,558
1830–1839	49,320	51,447
1840–1848	69,124	56,670
MEAN	56,155	50,259

The geocoded texts were then exported into a non-SQL database (MongoDB). I developed a user-friendly search interface for browsing the database using the Python Dash library that is used for GIS functions and mapping in addition to QGIS (Hossain 2019; QGIS Development Team 2009). Dash allows visualization of spatial data on different map layers and also as a timeseries. I used these maps to collect a blacklist for false positives with a high support value (cf. Gregory and Hardie 2011: 304-5). Despite some individual errors, the results of the NEL scan provide a reasonably accurate overview of the imagined space in the Romantic era literature.

Centers and Peripheries in Britain and Germany

In this section, I analyse the spatial distribution of the NEL results in the cultural historical context of Romanticism, focusing on the tension between the center and periphery in British and German literature. Not all the texts in the collection are related to Romanticism, although it was the dominant literary and cultural paradigm between 1790 and 1840. The Romantic ideas and values are reflected in the Corvey Collection, although the titles in it were not selected based on a predefined canon of Romanticism and thus represent a much more comprehensive selection of popular fiction published and read during the Romantic era.

Romantic literary theory was formed in the 1790s in Britain and Germany. It was characterized by a focus on the natural landscape and on folk poetry, rather than the conventional rules of classicism. In Germany, Romanticism arose in response to the standardizing culture that came from the center of Europe. In 1797, the famous Romantic literary critic Friedrich Schlegel wrote polemically about the domination of Italian, French, and English culture in Europe:

> The Italian manner as well as French and English manner had their Golden Ages in which they despotically ruled over the taste of all the rest of cultivated Europe. Only Germany has until now experienced the most multifaceted foreign influence without a reciprocating effect. By means of this association, the harsh severity of the original national character is increasingly effaced and finally almost entirely destroyed. In its place steps a general European character, and the history of every national poetry of the moderns contains nothing else but the gradual transition from its original character to the subsequent character of an artificial culture. (Schlegel 1979 [1797]: Vol I, 226; trans. in Schlegel 2001: 23).

According to Schlegel, the modernization of Germany was belated and its culture reflected its peripheral position in economy and politics. The aim of Schlegel and other theorists of early Romantics was to produce a new Golden Age of German culture in the future (Nivala 2017). The new generation of German authors gained attention elsewhere in Europe: Johann Wolfgang von Goethe's and Friedrich Schiller's works were translated into English, Samuel Taylor Coleridge and Thomas De Quincey studied German philosophy, and Madame de Staël promoted German culture in France.[4]

Interestingly, we can form a hypothesis from Schlegel's contemporary assessment and use the NEL results to investigate whether the centers mentioned in the German texts of the Corvey Collection were outside the territory of present-day Germany? And were the places most frequently mentioned in the English corpus located in the British Isles? Moreover, Goethe declared that literature of the Romantic period should be cosmopolitan 'world literature' (*Weltliteratur*, see Moretti 2013: 39). We can use the results of the NEL scan to explore whether the spatiality construced in the dataset is Eurocentric or has a national bias.

The place names in the Corvey Collection are not evenly distributed across the map: the imaginary space is concentrated in Europe, and even there in certain regions. Table 4 shows the 25 most popular spatial entities in English and German for the period 1790–1848. It is notable that the places most referenced in the dataset correlate with urban centers. The largest clusters of mentions form centers, while areas with only few mentions are more peripheral. As a critique of Goethe's concept of world literature, Franco Moretti uses Immanuel Wallerstein's model of the world economy to analyse the structure of world literature (*Weltliteratur*), which is divided into the core and the periphery. Moretti proposes that England and France were the centers of European literature during the nineteenth century (Moretti 1998: 173; 2013: 38-9, 107-8, 114-15; Wallerstein 1974).

Table 4. 25 most frequent toponyms in English and German corpus with raw word counts (occurencies) and normalized word counts (occurencies per Million words in each corpus).

English			German		
entity	**occurencies**	**per Million**	**entity**	**occurencies**	**per Million**
London	50,218	23.74	Paris	6,796	7.73
England	22,838	10.79	Deutschland	5,535	6.30
France	22,618	10.69	Rom	5,506	6.26
Paris	19,002	8.98	Frankreich	5,216	5.93
Rome	10,913	5.16	Wien	3,651	4.15
Ireland	10,414	4.92	Berlin	3,489	3.97
Spain	9,330	4.41	Europa	3,391	3.86
Italy	9,257	4.38	Spanien	3,114	3.54
India	6,965	3.29	Türken	3,047	3.47
Naples	6,747	3.19	Neapel	2,982	3.39
Edinburgh	5,892	2.78	Polen	2,558	2.91
Florence	5,657	2.67	London	2,494	2.84
Venice	5,177	2.45	Leipzig	2,326	2.65
Dublin	4,800	2.27	Schweden	2,279	2.59
Oxford	4,238	2.00	Rhein	1,986	2.26
Madrid	3,580	1.69	Venedig	1,754	2.00
Germany	3,574	1.69	Ungarn	1,714	1.95
Scotland	3,346	1.58	Schweiz	1,644	1.87
Vienna	2,663	1.26	Prag	1,580	1.80

Warwick	2,586	1.22	Sachsen	1,565	1.78
Brighton	2,464	1.16	Hamburg	1,412	1.61
Switzerland	2,383	1.13	Italien	1,379	1.57
Bristol	2,296	1.09	Dresden	1,314	1.49
Holland	2,192	1.04	Bern	1,238	1.41
Lisbon	2,183	1.03	Alpen	1,131	1.29

The NEL results demonstrate how strongly London dominates the English corpus. Since the title page is a separate element in the XML files of the Corvey Collection, I have been able to filter them out of the NEL scan. Because the title page has been removed in the pre-processing of all the works, the importance of London as a printing location does not explain this result. As I will show in the next section, an analysis based on place-name density also confirms the centrality of London area in the British data. Furthermore, London's high ranking is not explained by just a few outliers, but is frequently mentioned throughout the corpus. This suggests that the English-language literature of the Romantic era was perhaps more urban than has been assumed, when looking not just at a few canonized works but at a larger sample.

After London and England, Paris and France rank high in the British results. What is interesting about France, however, is that it is mainly described through a single center that is Paris. In contrast to that, in the case of Italy, there are many other cities along the Grand Tour—such as Naples, Florence, and Venice—that are mentioned besides Rome. Ireland, India, Edinburgh, and Dublin are included in the most frequently mentioned spatial entities as parts of the British Empire. With the exception of India, all the top 25 places are in Europe, but as will be shown in discussion of the sailing topic, there are works in the English dataset that have a more global setting.

In the German results, there is no single center. Although Paris and France are the most mentioned toponyms in German data, the German results are more evenly distributed. What is striking about the German results is that Paris, Rome, France, and Europe have a high ranking in relation to the domestic toponyms Germany, Wien, and Berlin. This confirms Schlegel's contemporary observation that German culture was dominated from foreign centers. However, *Deutschland* has second the highest number of mentions in the German corpus, although there was no single country called Germany during the period researched. As Friedrich Schlegel proposed: 'Regarding the original image of Germanness that some great patriotic inventors have established, there is nothing to criticize except for its incorrect position. This Germanness is not behind us, but ahead of us' (Schlegel 1967 [1797]: Vol II, 151; my translation). 'Germany' was something that did not exist as a political entity but rather was produced by the German literature of the Romantic period.

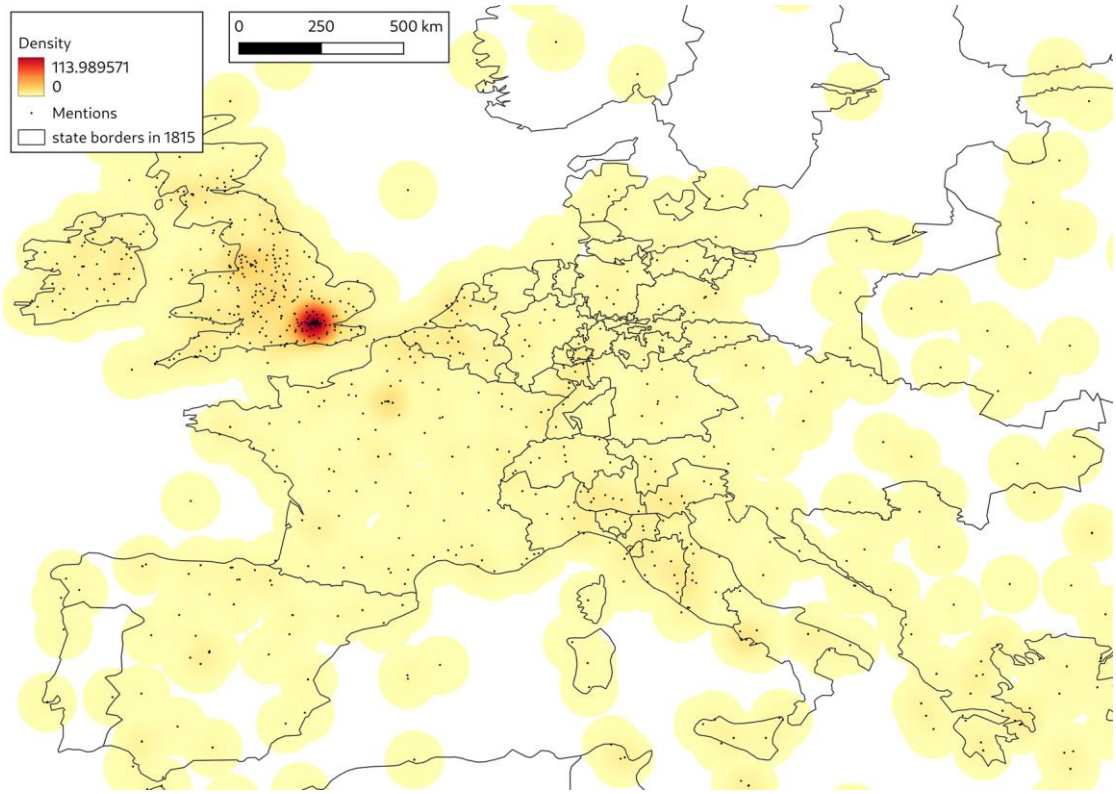
It is worth pointing out that, in addition to Vienna and Berlin, there are several local centers in the German region, such as Leipzig, Hamburg, and Dresden, between which the mentions are scattered. German place names do appear in the German Corvey, but Vienna or Berlin do not emerge as a similar domestic center as London in English corpus, to which significant portion of the references would be directed. Throughout its history, Germany

has been an administratively fragmented region with more than one capital. As in the British data, the Italian toponyms mentioned in German dataset include many cities that were on the Grand Tour. In contrast to the English corpus, the German language area is characterized by the high ranking of Eastern European place names including Poland, Hungary, and Prague. Unlike the English corpus, Europe is popular toponym in German corpus, but the German-language top 25 list does not include any place names outside Europe.

Density Maps

Examining the 25 most popular places highlights the centers of the dataset, but there may also be significant clusters of smaller places with mentions spread over several toponyms. For this reason, the data should also be examined using density maps. As already mentioned, the advantage of NEL over NER is that it identifies geospatial entities and links place names to coordinates. It is then easy to map the entities; however, point maps are difficult to interpret:

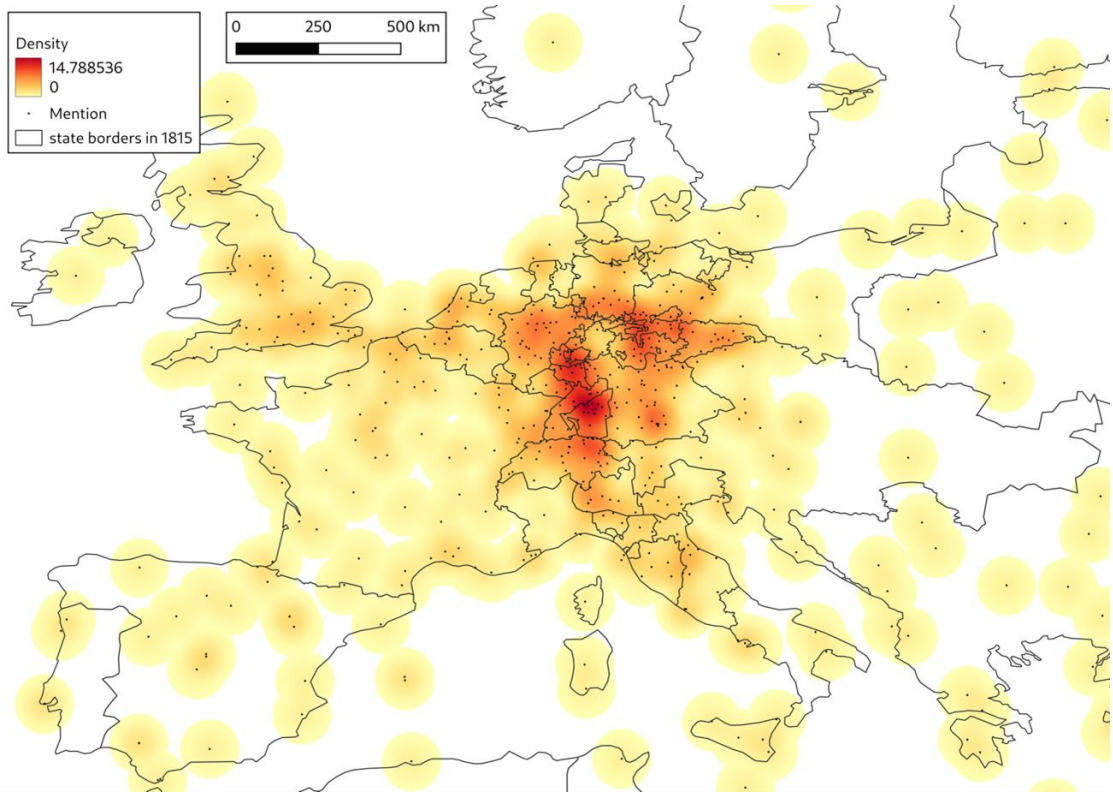
> It is well known in the cartographic literature that dot maps of this sort are difficult to interpret when they show a large amount of overlapping data […]. To simplify the pattern and make it more comprehensible a technique called density smoothing can be used. (Gregory and Hardie 2011: 302)



Map 1. A density smoothed map of toponyms mentioned in the English dataset.

Maps 1–2 show the density of mentions in English and German. The area of London also dominates the density map of the English corpus. The density map does not take into account the large number of mentions the toponym *London* receives, but the London area still has so many individual place names mentioned that it is the only center on Map 1. In contrast, the German data shown on Map 2 is much more evenly distributed. Although many of the individual place names most frequently mentioned in German literature are outside Germany, if one looks not at the number of mentions but at their density, the center of German literature is in Germany, but the mentions are quite widespread in several areas. I have included the state borders in 1815 for the map to illustrate how there was no state called Germany until the 1870s, but the German-speaking area consisted of a patchwork of many independent principalities. After the Congress of Vienna in 1815, Austria was regarded as the leading state in German-speaking Europe, but was challenged by Prussia during the nineteenth century.

To conclude, also the density map of the English corpus revolves around the London area while the German corpus is scattered around several local centers. Thus, the literature of the Romantic Era reflects the different center-periphery structure in these countries.



Map 2. A density smoothed map of toponyms mentioned in the German dataset.

Geospatial Topic Models

So far, I have only counted and mapped the number of place names. How does the spatiality of the Corvey Collection manifest itself when the semantic meanings of literary

texts are combined with geospatially collected data? Topic modeling makes it possible to isolate common themes among thousands of texts. By processing texts paragraph by paragraph, we can associate topic models with the spatial entities mentioned in each paragraph.

Topic modeling is well established as a robust way to cluster documents together by the 'topics' they share. This method has been used successfully in literary studies (see Jockers 2013: 118-53; Erlin 2014; Gavin and Gidal 2016). LDA (Latent Dirichlet Allocation) is a common method of topic modeling. However, similar results are obtained with non-negative matrix factorization (NMF), which I have used in this paper (see Pedregosa et al. 2011). The reason for preferring NMF is that I have segmented the texts by paragraph in order to maintain the link between the place name mentioned and the topic covered by the paragraph (see Gavin and Gidal 2016). My observations are that NMF often produces more coherent topics than LDA when the documents are relatively short. Splitting the works into short chunks also produces a better topic model in general (Erlin 2013: 61).

The English and German documents were preprocessed and modelled separately. In pre-processing the data, I have followed established research practices (see Jockers 2013: 131-3; Erlin 2013: 61). The most common stop words in each language, punctuation, and other special characters including numbers have been removed from each corpus. The paragraphs have then been run through a part of speech tagging (POS) algorithm (Honnibal et al. 2020) trained on the target language to select only nouns in the model. This step also tokenizes the text. The TF-IDF (term frequency-inverse document frequency) matrix is then computed for the documents. The TF-IDF is able to extract keywords from each paragraph. In order to be able to cluster the paragraphs into groups by the topics they share, the TF-IDF matrix is then processed with the NMF algorithm, which is given the parameter to search for 200 typical topics in each text corpus. If the number of topics is too low, the topics tend to be too abstract to include as many texts as possible. If the number of topics is too high, the topics easily become structured by works, i.e. the model overfits. (On the selection of the number of topics to be extracted, see Jockers 2013: 128.) After several test runs, 200 topics was the best value for this data.

The topic modeling produced two models, in English and German, which I combine with geotagged entities in each paragraph to explore spatial themes in the Corvey Collection. The NMF algorithm does not name the topics, but only lists the typical words mentioned in them. If we look, for example, at the most typical themes in Sir Walter Scott's novels included the Corvey Collection, they are topic 76 (*death*, *armour*, *squire*, *helmet*, *list*, *lance*, *arm*, *shield*, *sword*, *knight*) and topic 151 (*ain*, *sae*, *bit*, *thing*, *mair*, *ane*, *er*, *weel*, *hae*, *wi*). Topic 76 refers to medievalism, which was a typical topic in Scott's historical novels. The words in topic 151 are not OCR noise; they belong to the Scots language, which is spoken in Scotland. It is perfectly logical that paragraphs containing words from another language should form an independent topic. Furthermore, the five most common place names related to topic 151 are Edinburg (105), Glasgow (100), London (71), France (44), and Aberdeen (36). Most places are in Scotland, but London and France are on the list, as they are often mentioned in paragraphs building this topic.

In this article, I cannot discuss all 200 topics for two different language corpora. Therefore, I have chosen two topics related to the construction of core and periphery in the literature of the Romantic Period. The topics selected for further analysis are 'landscape' and 'sailing', the presence of which I compare in English and German datasets.

Landscape Aesthetics Topics in Corvey

Interest in Romantic ideas began in Britain in the 1760s, during which time the rise of picturesque landscape aesthetics also became popular. These new ideas had a geographical framing: Wales and the Lake District were the original sites of the picturesque and in Germany, the Rhine Valley with its ruined castles was the location of *Rheinromantik* (Cepl-Kaufmann and Johanning-Radzienė 2019). The rehabilitation of mountains played an import role in the rise of Romantic landscape aesthetics. Before the emergence of the concepts of the sublime and the picturesque, mountains and topographically hilly terrain were regarded as remote areas with land that was difficult to make productive (Rigby 2004: 131-72; Ziolkowski 1992: 18-63). In contrast to the Enlightenment, Romantic cartography created tension between the 'standardized' culture of urban centers (London, Paris, and Rome) and the local tradition of rural peripheries. How well do these generalizations apply to the Corvey Collection?

In both languages, there is a topic that seems to be related to landscape aesthetics. In English topic 127, the most common words are: *lake*, *scene*, *sky*, *air*, *sun*, *cloud*, *valley*, *hill*, *rock*, and *mountain*. In German, the corresponding topic is number 128: *Lied*, *Ferne*, *Ufer*, *Hütte*, *Höhe*, *Fels*, *Thal*, *Gegend*, *Wald*, and *Berg*. English-language texts whose paragraphs build on this theme include the following titles:

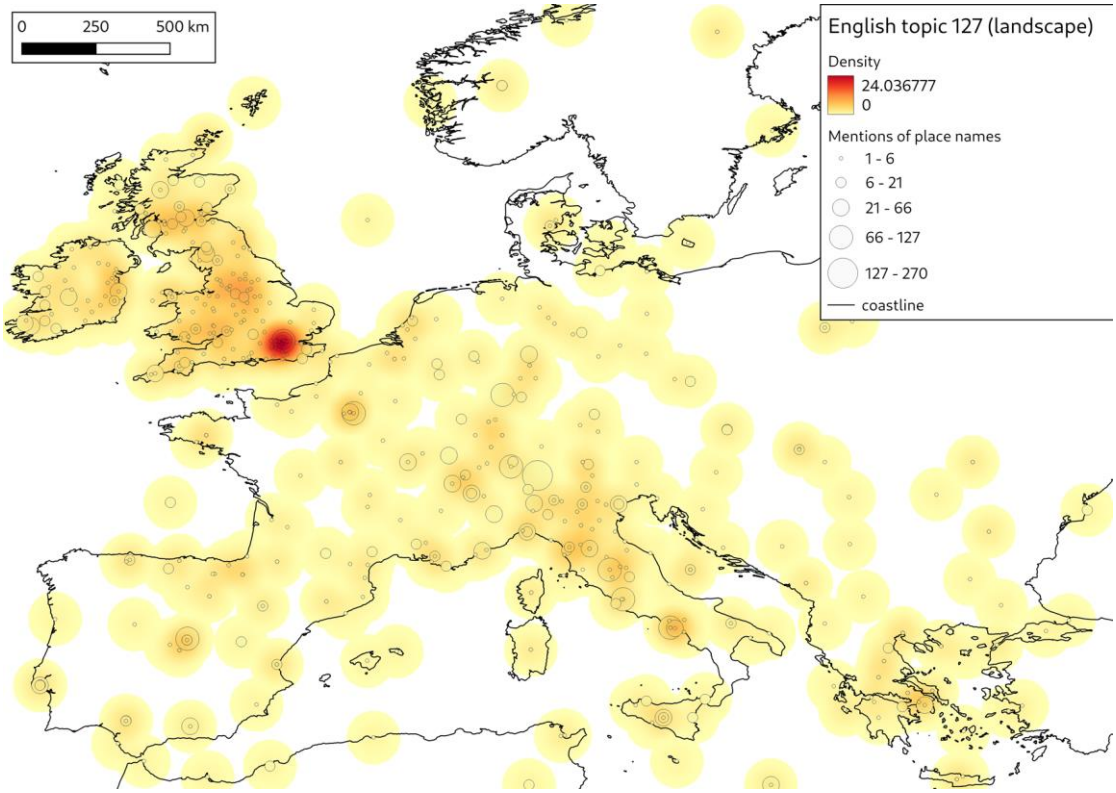
- *Continental Adventures* (1826) by Charlotte Eaton
- *The Pilgrim of the Hebrides* (1830) by Charles Hoyle
- *Look to the End: or, The Bennets Abroad* (1845) by Sarah Stickney Ellis
- *Florence Macarthy: an Irish tale* (1818) by Lady Sydney Morgan
- *Gaston de Blondeville* (1825) by Ann Radcliffe
- *Recollections of a Pedestrian* (1826) by Thomas Alexander Boswell
- *Italy, and Other Poems* (1828) by William Sotheby.

Most of these works do not belong to the canonical Romanticism, but they do contain landscape depictions with Romantic elements. As Map 6 shows, the main areas of the English-language material are the Alps, Italy, Scotland, Ireland, and the English countryside. Wales and the English Lake District are included, but do not carry as much weight as might be expected. Perhaps surprisingly, the London area has high presence in this map, but it is to be expected that there may also be urban place names mentioned in the passages with nature-related topic.

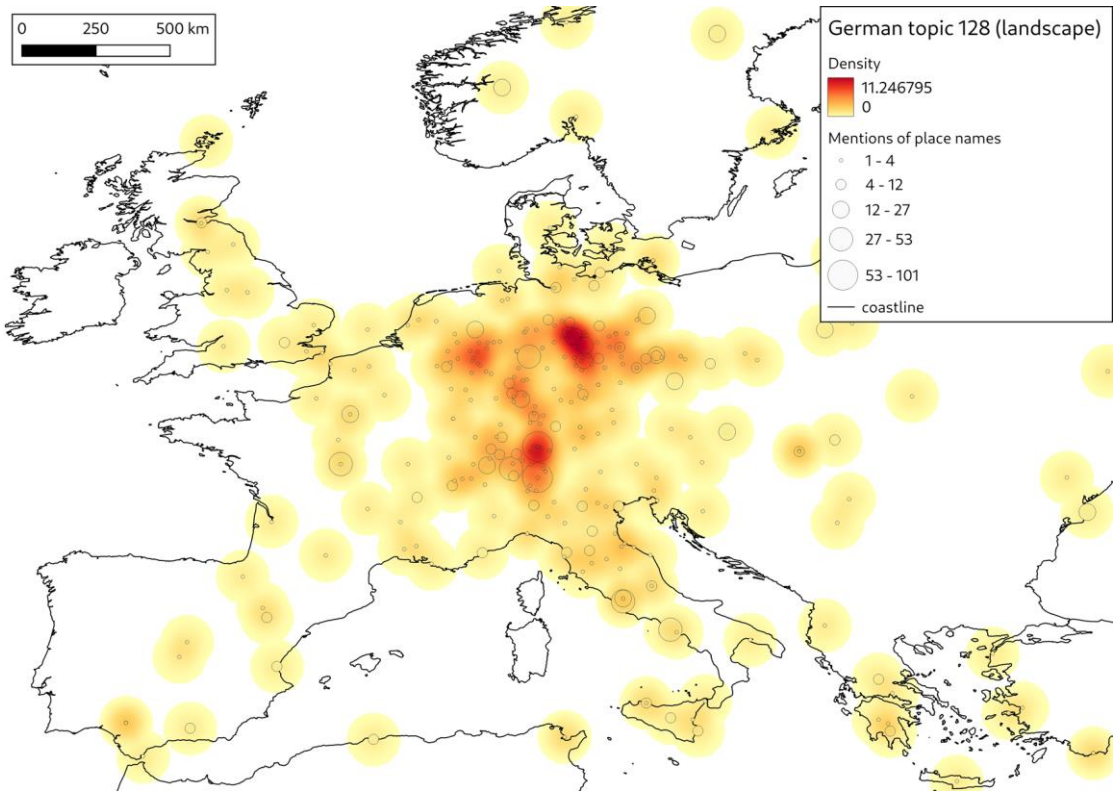
The German-language works that build on the theme of landscape are to a greater extent also written by famous Romantic authors. They include:

- *Ahnung und Gegenwart* (1815) by Joseph von Eichendorff
- *Die vier Norweger: ein Cyklus von Novellen* (1837) by Norwegian Romantic author Henrich Steffens
- *Altsächsischer Bildersaal* (1818) by Friedrich La Motte-Fouqué
- *Bilder aus der Schweiz* (1824) by originally German, later Swiss, author Heinrich Zschokke
- *Schottischer Robinson* (1827) by Heinrich Oswald.

Of these, Joseph von Eichendorff and Friedrich La Motte-Fouqué are very well known representatives of German Romanticism. Friedrich La Motte-Fouqué was of a family of French Huguenot origin, but born in Prussia. The presence of the Norwegian author Steffens and the Swiss author Zschokke highlights that the German-language corpus of the Corvey Collection does not in fact consist only of German authors. However, this alone does not explain why the Alps are the most frequently mentioned region and Switzerland ranks so high in both English and German dataset, because the references to Alps are spread over several texts. The density of mentions in German corpus is concentrated in the Swiss Alps and Italy, but also in the Harz mountains and near the Rhine Valley. Thus alongside the domestic landscapes such as the Rhine, the Alps and also Italy were the most important areas of the Romantic Picturesque.



Map 3. Place names mentioned in association with the English topic 127 (landscape).



Map 4. Place names mentioned in association with the German topic 128 (landscape).

The theme of the Romantic landscape puts my earlier observation that the English Corvey's focus is more on the British Isles and the German abroad in a whole new light. If you look only at the most popular place names, it now seems that the English-language corpus includes many references to the Alps and Italy–and hardly any domestic destinations (see Table 5). However, the density of place names is much higher in the British Isles (Map 3). This suggests that authors often write about other countries in an abstract way, using only the names of countries, regions, or capitals such *Alps*, *Switzerland*, *Italy*, or *Rome*. When they write about their home country, they describe areas in higher resolution and name many smaller places, perhaps more familiar to the authors and readers, which is reflected in the density maps.

While Paris and other foreign centers dominate the overall count in the German-language Corvey Collection, the Romantic Landscape topic has a stronger domestic focus on its map. In fact, the same feature can be seen in the German topic 185, which I interpret as medievalism (typical words: *Knappe*, *Ritter*, *Vest*, *Fels*, *Sturm*, *Gut*, *Prinzessin*, *Knecht*, *Mauer*, and *Burg*). This is very understandable considering the aims of Romanticism: Romantic authors sought to build a national literature that would challenge the influence of foreign cultural centers.

Topic of Sailing in Corvey Collection

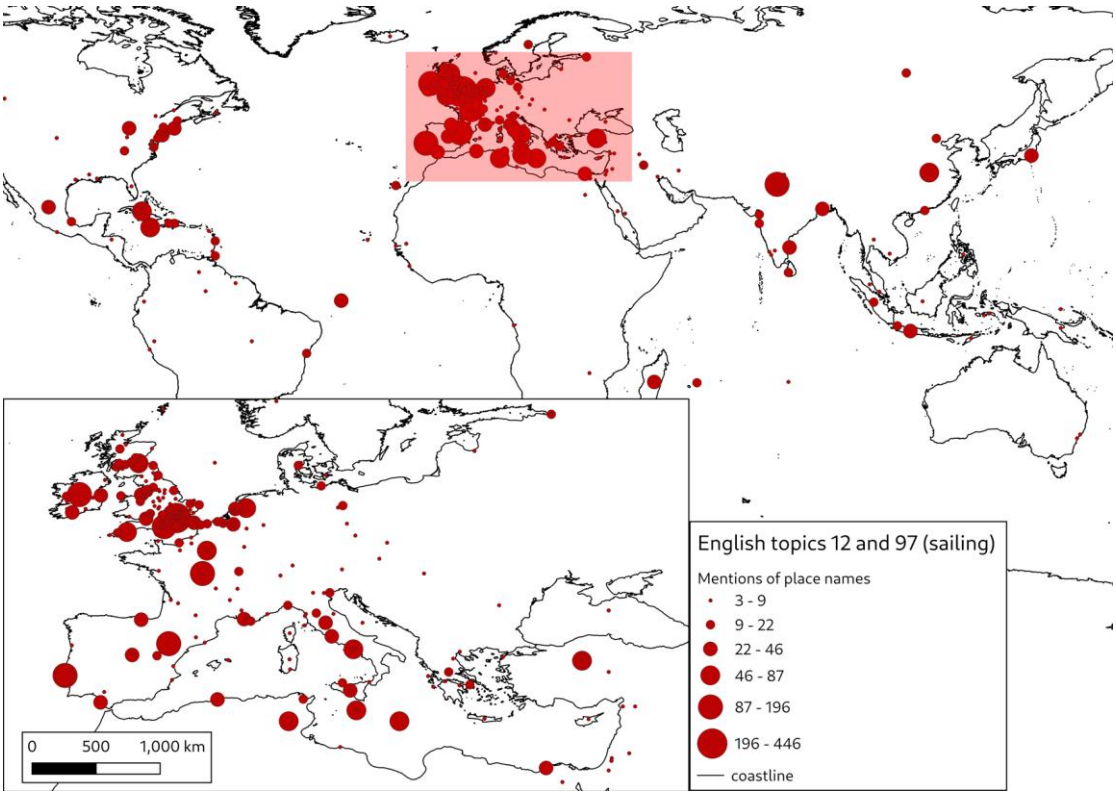
I have deduced that two topics in the English corpus are related to 'sailing', as their typical words are: *island*, *people*, *voyage*, *captain*, *boat*, *shore*, and *ship* (topic 12) and *sailor*, *vessel*, *cabin*,

board, *deck*, *ship*, and *captain* (topic 97). I have not noticed any substantial thematic or spatial difference between the two topics, but a closer reading of the texts might reveal one. Topic 106 in the German corpus corresponds this, since the words it lists are: *Segel*, *Küste*, *Boot*, *Hafen*, *Wind*, *Ufer*, *Welle*, *Sturm*, and *Meer*.

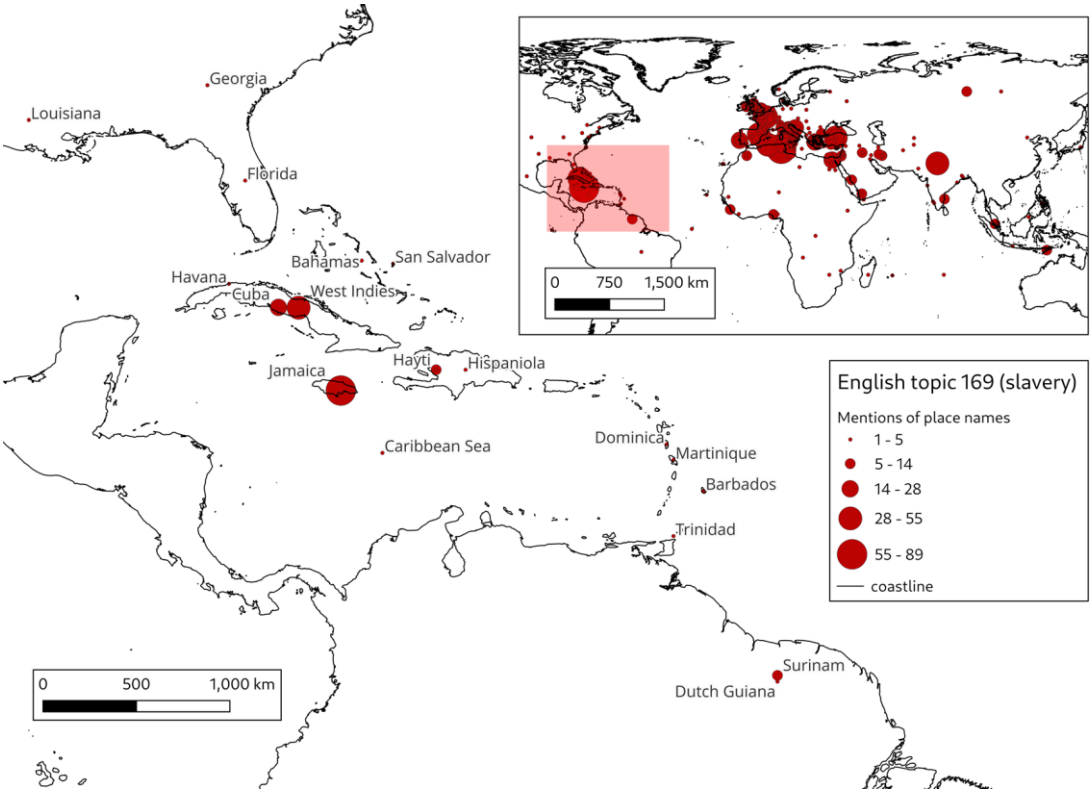
Table 5. The 20 most often mentioned spatial entities associated with the landscape topics in English and German corpora.

English topic no. 127 (landscape)		**German topic no. 128 (landscape)**	
Alps	270	Alpen	101
Switzerland	127	Rhein	74
Italy	123	Deutschland	53
Rome	121	Schweiz	49
Spain	116	Rom	38
Earth	113	Neapel	35
France	112	Frankreich	30
Naples	86	Gletscher	27
London	84	Europa	27
Rhine	79	Elbe	27
Ireland	66	Weser	26
Palestine	55	Donau	25
Geneva	52	Wien	24
Florence	52	Thüringen	23
Paris	51	Paris	23
Valais	51	Türken	22
Jerusalem	50	Prag	21
Savoy	46	Böhmen	20
Killarney	45	Schweden	19
Nile	42	Heidelberg	18

The interesting thing about these topics is that maps implied in them are not limited to Europe but have a global scope. Map 5 shows the English-language 'sailing' topics on a map, with the place names mentioned in concordance with this topic placed on maps of the world and Europe. Strikingly, London and England are by far the most mentioned place names in relation to this topic (see Table 6). However, the sailing topics also frequently refer to important smaller transport hubs such as Lisbon, Portsmouth, and Plymouth, which are port towns or cities. Map 5 shows sites in both the Caribbean and Indian Ocean, reflecting the role of colonialism in English texts. These colonized areas are the Wallersteinian periphery of world market, i.e. an area that is economically dependent on the center and from which raw materials are exported to its market. In the context of Romantic literature, faraway lands and sea voyages were romanticized exotic themes.

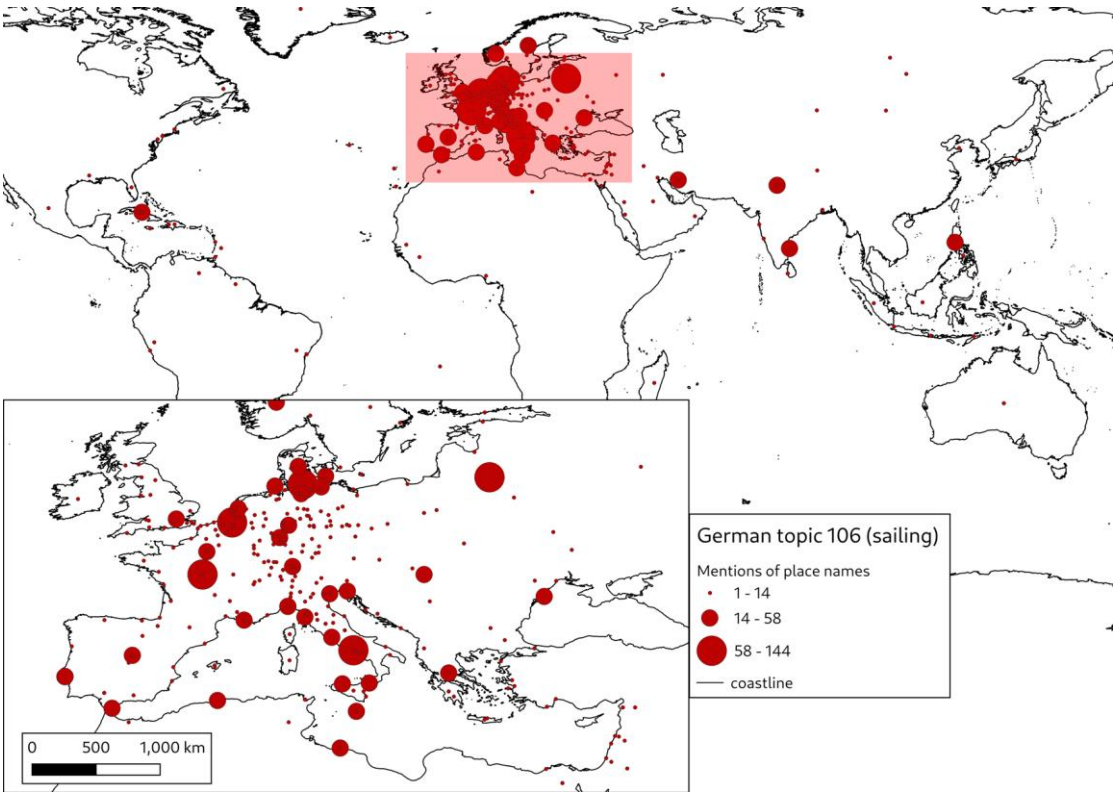


Map 5. English topics 12 and 97 (sailing).



Map 6. English topic 169 (slavery).

The English model contains another topic, which has quite similar geographic coverage to the 'sailing' topics, and is worth mentioning. I interpret topic 169 as 'slavery', since its most typical words are: *subject*, *power*, *plantation*, *church*, *slavery*, *system*, *estate*, *island*, *negro*, and *slave*. The map on this topic also contains many place names in Europe (see Map 6), but if we look at the titles of the works that mainly comprise the topic, we find that most of them are actually set in the Caribbean and Africa. Topic 169 includes works such as *Outalissi: a Tale of Dutch Guiana* (1826) by Christopher Edward Lefroy and anonymously published *Marly: or, A Planter's Life in Jamaica* (1828) and *The Koromantyn Slaves: or, West-Indian Sketches* (1823). There is no such overtly colonialist or slavery-related theme in the German model, which is understandable as Germany was not an imperialist power like the UK.



Map 7. German topic 76 (sailing).

If we compare English results with the places associated with the German sailing theme (see Map 7), they are very different. The most popular spatial entity in German results is in fact not a place name but a nationality *Türken* (Turkish). Yet, this is correct in that the Turks are associated with the sailing theme in many German historical novels in the Corvey Collection, including *Scipio Cicala* (1832) by Philipp Joseph von Rehfues and *Der Venetianer: historisch-romantische Gemälde* (1829) by Karl Herloßsohn. For example, the protagonist of the historical novel *Scipio Cicala* is Cığalazade Yusuf Sinan Pasha (1545–1605), who was the Grand Admiral of the Ottoman Navy.

After *Türken*, the most frequent spatial entities associated with sailing are port cities such as Antwerp and Kiel (see Table 6). All German ports are on the Baltic Sea and Kiel remains an important maritime hub in addition to Hamburg. Alongside them are Mediterranean port cities such as Naples, Marseille, and Algier. The German-language material shows a few colonies on the world map, but the sailing scenes in German works are much more frequently located in the ports on the Mediterranean or the Baltic Sea, or inland rivers such as the Rhine.

Table 6. The 20 most often mentioned spatial entities associated with the 'sailing' topics in English and German corpora and the 'slavery' topic in English corpus.

English topics nos. 12 and 97 (sailing)		**German topic no. 106 (sailing)**		**English topic no. 169 (slavery)**	
London	446	Türken	155	Africa	89
England	432	Antwerpen	144	Jamaica	80
France	196	Europa	114	England	73
India	155	Kiel	99	West Indies	55
Ireland	107	Frankreich	73	Rome	50
Portsmouth	106	Neapel	72	Algiers	49
Spain	100	Venedig	58	France	49
Lisbon	97	London	53	Spain	44
China	87	Spanien	43	India	42
Dutch	80	Genua	42	Turkish	37
Jamaica	76	Rhein	36	Turks	28
Portuguese	72	Paris	35	Tunis	27
Turks	71	Elbe	35	Athens	24
Plymouth	69	Marseille	34	Cairo	23
Paris	68	Perser	34	Zara	22
Africa	65	Deutschland	33	Italy	21
Naples	63	Algier	32	Virginia	21
Portugal	59	Hamburg	31	Paris	19
West Indies	58	Donau	30	London	18
Edinburgh	57	Schweden	30	Kingston	18

Conclusions

The Corvey Collection provides a rich resource for comparing the literature of the English and German Romantic periods from the perspective of literary geography. The named entity linking and topic modelling are able to present an alternative interpretation of the literary space of the Romantic period 1790–1840. Geospatial analysis shows that the London area dominates the English-language corpus in terms of both the most popular place name and the density map. For German dataset, on the other hand, no such single center can be identified. In terms of the most popular place names, foreign centers such

as Paris dominated German-language literature, which was also a concern for contemporaries. On the other hand, the density map showed that German-language literature was also rich in references to smaller domestic place names and that the German cultural sphere had many smaller regional centers in addition to Vienna and Berlin.

The use of topic models enriches this overview and allows thematic maps to be drawn from the data. When we look at the theme of the Romantic landscape in the data, the situation looks different. Indeed, the Alps are the most popular single toponym in both English and German texts. The Italian cities also come to the fore. However, the density maps reveal that the landscape topics were also associated with the presence of many lesser-known domestic places in close proximity to each other. German Romantic literature was related to nation-building; hence, it is natural that the Romantic landscape was set not only in the Alps and Italy, but also in the Harz Mountains and along the Rhine. The sailing theme, on the other hand, sheds light on the more global framework of Romantic literature, which was also clearly linked to themes such as colonialism and slavery in the English-language material. Germany, on the other hand, is a continental country whose literary sailing scenes were more often located in northern port cities, the Mediterranean, or inland bodies of water than in distant countries. Thus, what is narrated as a center in the literature depends on the context in which the space is constructed. In the English data it is very often London, but in the German data the center-periphery structure is more dispersed.

Notes

[1] For the periodisation and definition of Romanticism in the previous research, see Ziolkowski (2018: 1-2). By the broadest definition, the Romantic Era began with the French Revolution of 1789 and ended with the Revolutions of 1848. In the British context, Queen Victoria's accession on 20 June 1837 is sometimes used as the point of termination for the Romantic Era, but this does not work in the German context. Moreover, for such a broad trend as Romanticism, it is impossible to set an end point to within a year.

[2] See https://extra.shu.ac.uk/corvey; http://english.unl.edu/sbehrendt/Corvey/html/Projects/CorveyNovels/CorveyNovelsIndex.htm; http://www.british-fiction.cf.ac.uk; https://portal.issn.org/resource/ISSN/1471-5988.

[3] As Kettunen and Koistinen (2019: 272) point out, there is no unambiguous measure for assessing the quality of the re-OCR process. For example, because Tesseract reports its confidence score at the character level, it counts in errors such as punctuation and special characters, even though they do not usually impede a human reader's ability to read text as much as incorrect letters. The OCR errors also do not appear significantly in the topic models, suggesting that the data is usable.

[4] On cultural exchange between the UK and Germany, see, for example, Maertz (2017) and Raven (2004).

Works Cited

Bach, K. E. (1812) *Alberts Jugendjahre: Ein Komischer Roman: Von Karl Eduard Bach*. Schüppelschen Buchhandlung.

Cepl-Kaufmann, G. and Johanning-Radzieńé, A. (2019) *Mythos Rhein: Kulturgeschichte eines Stromes*. Darmstadt: WBG.

Cooper, D., Donaldson, C. and Murrieta-Flores, P. (2016) *Literary Mapping in the Digital Age*. London: Routledge.

Craciun, A. (2016) *Writing Arctic Disaster*. Cambridge: Cambridge University Press.

Daiber, J., Jacob, M., Hokamp, C. and Mendes, P. (2013) 'Improving Efficiency and Accuracy in Multilingual Entity Extraction.' *In I-semantics. Proceedings of the 9th International Conference on Semantic Systems*. Graz. [Online] [Accessed 14 July 2023] https://doi.org/10.1145/2506182.2506198

Donaldson, C., Gregory, I. N. and Murrieta-Flores, P. (2015) 'Mapping Wordsworthshire: A GIS Study of Literary Tourism in Victorian Lakeland.' *Journal of Victorian Culture*, 20(3), pp. 287-307.

Donaldson, C., Gregory, I. N. and Taylor, J. E. (2017) 'Locating the Beautiful, Picturesque, Sublime and Majestic: Spatially Analysing the Application of Aesthetic Terminology in Descriptions of the English Lake District.' *YJHGE Journal of Historical Geography*, 56, pp. 43-60.

Engelsing, R. (1973) *Zur Sozialgeschichte deutscher Mittel- und Unterschichten*. Göttingen: Vandenhoeck & Ruprecht.

Erlin, M. (2014) 'The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864.' *In* Erlin, M. and Tatlock, L. (eds) *Distant Readings*. Tatlock. Rochester, NY: Camden House, pp. 55-90.

Gamer, M. (2006) *Romanticism and the Gothic: Genre, Reception, and Canon Formation*. Cambridge: Cambridge University Press.

Garside, P. (1992) 'Collections of English Fiction in the Romantic Period: The Significance of Corvey.' *In* Schöwerling, R. and Steinecke, H. (eds) *Die fürstliche Bibliothek Corvey*. München: Fink, pp. 70-81.

Gavin, M. and Gidal, E. (2016) 'Topic Modeling for Historical Geography.' *Studies in Scottish Literature*, 42(2), pp. 185-197.

Gavin, M. and Gidal, E. (2017) 'Scotland's Poetics of Space: An Experiment in Geospatial Semantics.' *Journal of Cultural Analytics*, 2(1), pp. 1-36.

Gidal, E. and Gavin, M. (2019) 'Infrastructural Semantics: Postal Networks and Statistical Accounts in Scotland, 1790–1845.' *International Journal of Geographical Information Science*, 33(12), pp. 2523-2544.

Gregory, I. N. and Cooper, D. (2009) 'Thomas Gray, Samuel Taylor Coleridge and Geographical Information Systems: A Literary GIS of Two Lake District Tours.' *International Journal of Humanities and Arts Computing*, 3(1-2), pp. 61-84.

Gregory, I., Donaldson, C., Murrieta-Flores, P. and Rayson, P. (2015) 'Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research.' *International Journal of Humanities and Arts Computing*, 9(1), pp. 1-14.

Gregory, I. N. and Hardie, A. (2011) 'Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems.' *Literary and Linguistic Computing*, 26(3), pp. 297-314.

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A. (2020) 'spaCy: Industrial-strength Natural Language Processing in Python.' doi:10.5281/zenodo.1212303

Hossain, S. (2019) 'Visualization of Bioinformatics Data with Dash Bio.' *In* Calloway, C., Lippa, D., Niederhut, D. and Shupe, D. (eds) *Proceedings of the 18th Python in Science Conference*, pp. 126-133.

Jockers, M. L. (2013) *Macroanalysis: Digital Methods and Literary History*. Urbana, IL: University of Illinois Press.

Kettunen, K. and Koistinen J. (2019) 'Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals.' *In* Navarretta, C. M Agirrezabal, M. and Maegaard, B. (eds) *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, pp. 270-282.

Liu, Z, Jiang, F., Hu, Y., Shi, C. and Fung, P. (2021) 'NER-BERT: A Pre-Trained Model for Low-Resource Entity Tagging.' *arXiv*. [Online] [Accessed 14 July 2023] https://doi.org/10.48550/arXiv.2112.00405

Maertz, G. (2017) *Literature and the Cult of Personality: Essays on Goethe and His Influence*. Stuttgart: ibidem.

Mendes, P.; Jakob, M., García-Silva, A. and Bizer, C. (2011) 'DBpedia Spotlight: Shedding Light on the Web of Documents.' *In I-semantics. Proceedings of the 7th International Conference on Semantic Systems*. Graz. [Online] [Accessed 14 July 2023] https://doi.org/10.1145/2063518.2063519

Moretti, F. (1998) *Atlas of the European Novel, 1800–1900*. London: Verso.

Moretti, F. (2007) *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Moretti, F. (2013) *Distant Reading*. London: Verso.

Nivala, A. (2017) *The Romantic Idea of the Golden Age in Friedrich Schlegel's Philosophy of History*. New York: Routledge.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) 'Scikit-learn: Machine Learning in Python.' *Journal of Machine Learning Research*, 12(85), pp. 2825-2830.

Piatti, B. (2008) *Die Geographie der Literatur: Schauplätze, Handlungsräume, Raumphantasien*. Göttingen: Wallstein.

Punter, D. (1999) 'Introduction: Of Apparitions.' *In* Byron, G. and Punter, D. (eds) *Spectral Readings*. London: Palgrave, pp. 1-10.

QGIS Development Team (2009) *QGIS Geographic Information System*. Open Source Geospatial Foundation. [Online] [Accessed 14 July 2023] https://www.qgis.org

Raven, J. (2004) 'Cheap and Cheerless: English Novels in German Translation and German Novels in English Translation, 1770–1799.' *In* Huber, W. and Schöwerling, R. (eds) *The Corvey Library and Anglo-German Cultural Exchanges, 1770–1837*. München: Wilhelm Fink Verlag, pp. 1-34.

Rigby, K. (2004) *Topographies of the Sacred: The Poetics of Place in European Romanticism*. Charlottesville: University of Virginia Press.

Schlegel, F. (1967) *Kritische Friedrich-Schlegel-Ausgabe*. Band II. Ed. Eichner, H. München: Schöningh.

Schlegel, F. (1979) *Kritische Friedrich-Schlegel-Ausgabe*. Band I. Ed. Behler, E. München: Schöningh.

Schlegel, F. (2001) *On the Study of Greek Poetry*. Trans. Barnett, S. New York: SUNY Press.

Steinecke, H. (1992) 'Die fürstliche Bibliothek Corvey: Perspektiven ihrer wissenschaftliche Erschließung.' *In* Schöwerling, R. and Steinecke, H. (eds) *Die fürstliche Bibliothek Corvey*. München: Fink, pp. 13-20.

Wallerstein, I. (1974) *The Modern World-System*. Vol. I. New York: Academic Press.

Ziolkowski, T. (1992) *German Romanticism and Its Institutions*. Princeton, NJ: Princeton University Press.

Ziolkowski, T. (2018) *Stages of European Romanticism: Cultural Synchronicity Across the Arts, 1798–1848*. Rochester, NY: Camden House.