Implementing Corpus Analysis and GIS to Examine Historical Accounts of the English Lake District

Christopher Donaldson, Ian N. Gregory, and Joanna E. Taylor
Department of History
Lancaster University, Lancaster, UK
Contact address: c.e.donaldson@lancaster.ac.uk

Abstract: This paper reports on interdisciplinary research into the automated geographical analysis of historical text corpora. It provides an introduction to this research, which is being completed by two interrelated projects: the European Research Council-funded Spatial Humanities project and the Leverhulme Trust-funded Geospatial Innovation in the Digital Humanities project. In addition to contextualising the work of these projects, the paper introduces a case study that applies collocation analysis, automated geo-parsing, and Geographic Information Systems (GIS). The focus of this study is a 1.5 million-word corpus of writing about the English Lake District. This corpus comprises 80 works written between the years 1622 and 1900. In investigating this corpus, we demonstrate how a hybrid geographical and corpus-based methodology can be used to study the application of specific aesthetic terminology in historical writing about the Lakeland region.

Acknowledgements: The research leading to these findings has received funding from the Leverhulme Trust, as part of the Geospatial Innovation in the Digital Humanities research project, and from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007–2013): ERC grant Spatial Humanities: Texts, GIS, Places [agreement number 283850].

Introduction

This paper reports on the research of two funded projects: Spatial Humanities: Texts, GIS, Places (European Research Council: 2012–2016) and Geospatial Innovation in the Digital Humanities: A Deep Map of the English Lake District (Leverhulme Trust: 2015–2018) (*Spatial Humanities* 2012; *Geospatial Innovation in the Digital Humanities* 2015). These projects are both based at Lancaster University, in the United Kingdom. The ambition of both projects is to investigate how methods developed in Geographic Information Science (GISc) and corpus linguistics can be adapted to enrich the study of history, literature, and the arts. Principally, these investigations focus on a.) using techniques from corpus linguistics and natural language processing (NLP) to mine large corpora of historical texts and b.) analysing the data those corpora contain using Geographic Information Systems (GIS) technology (see Donaldson et al 2016; Gregory and Donaldson 2016; Donaldson et al 2015; Gregory et al 2015; Murrieta-Flores et al 2015; Porter et al 2015).

I. Geographic Information Systems (GIS)

GIS are computer systems that allow the user to integrate, arrange, and visualise the geographic information contained in any dataset. Geographic information broadly consists of two, interrelated components: a *spatial*

component (such as a place-name, a geographical feature, or a location) and a *thematic* component. A thematic component assigns attributes to the place-name, feature, or location specified by the spatial component. The distinction between the spatial and thematic components of geographic information can be usefully illustrated by means of a commonplace example, such as a census. A census is essentially a dataset that organises thematic content (such as rates of population and employment) in terms of discrete spatial units (such as villages, towns, and cities). GIS are computational tools for bringing these different types of information—the spatial and the thematic—together and for examining correlations between them. This process involves geo-referencing the dataset: in other words, using a gazetteer to assign geographic coordinates to each item of spatial information in the dataset. Once the dataset has been geo-referenced, the geographic information it contains can be displayed and analysed in a GIS environment.

One of the benefits of working with GIS technology is that it allows researchers a.) to explore the geography that underlies complex datasets, and, in doing so, b.) to identify more effectively spatial patterns and relationships in the information those datasets contain. It is on account of these affordances that GIS have often been implemented in the analysis of quantitative source data that can be linked to specific locations. In the fields of historical demography and historical epidemiology, for instance, GIS technology has proved an effective resource for compiling large amounts of vital registration data and for creating models of historical trends in disease and mortality (see Murrieta-Flores et al 2015; Porter et al 2015; Gregory and Marti-Henneberg 2010; Gregory et al 2010; Gregory 2009).

Significantly, over the past decade experimental research has begun to explore the application of GIS technology in the analysis of qualitative data (see Jang 2011; Bodenhamer et al 2010). Much of this research has focussed on applying GIS to investigate literary and artistic representations of specific places and landscapes (see Cooper et al 2016). Examples of such literary and qualitative historical GIS research projects are diverse. Collectively, though, these projects are both contributing to and consolidating a broader initiative in the digital humanities. Matthew Jockers has outlined this development in his 2013 study *Macroanalysis*. For Jockers, such research projects aspire to make fuller—and better—use of the resources available to scholars in the digital age. "Today", writes Jockers, "in the age of digital libraries and large-scale book-digitization projects, the nature of the evidence available to us has changed, radically"; this is "not to say", Jockers clarifies, "that we should no longer read [individual] books […], but rather to emphasize that massive digital corpora offer us unprecedented access to the [historical] record and invite, even demand, a new type of evidence gathering and meaning making" (Jockers 2013, 7–8). In other words, instead of focussing their research on a small number of examples, scholars should make use of the wealth of digitised data available to them. Developing appropriate methods for collecting and analysing that data, however, remains a challenge. Notably though, between 2007 and 2008, members of the Spatial Humanities and Geospatial Innovation projects made a crucial step in this direction by conducting a pilot study (entitled Mapping the Lakes) to test the interpretive possibilities of using GIS to analyse historical accounts of the English Lake District (*Mapping the Lakes* 2007).[1]

2. Initial Research: Mapping the Lakes

The Lake District is an ideal focus for an experiment in qualitative GIS because it is a region that has long been portrayed in poems, guidebooks, and paintings as a discrete area. As Norman Nicholson has observed, although the old counties that comprise the Lake District are topographically and geologically diverse, they are nevertheless "part of one geographical system, and their people have roots in a shared history and culture" (Nicholson 1955, 1; see also Cooper 2008, 813).[2] The Lake District is, consequently, a spatially unified area that is exceptionally rich in geo-specific thematic information. One thinks of cultural icons such as William Wordsworth, John Ruskin, and W. J. M. Turner, of course; but one can also think of an array of writers and artists—from Beatrix Potter and Arthur Ransome to Alfred Wainwright and Sheila Fell—who have shaped our perception of the Lake District and its landscapes.

The Mapping the Lakes pilot study sought to show how GIS technology could help researchers to analyse the accounts of early Lake District tourists and, in the process, to consider how representations of the region have changed over time. This study focused on two canonical historical sources: the poet and scholar Thomas Gray's epistolary account of his Lake District tour of 1769, and the letters and notebooks in which the poet and literary critic Samuel Taylor Coleridge documented his excursion into the western Lakeland fells in 1802.

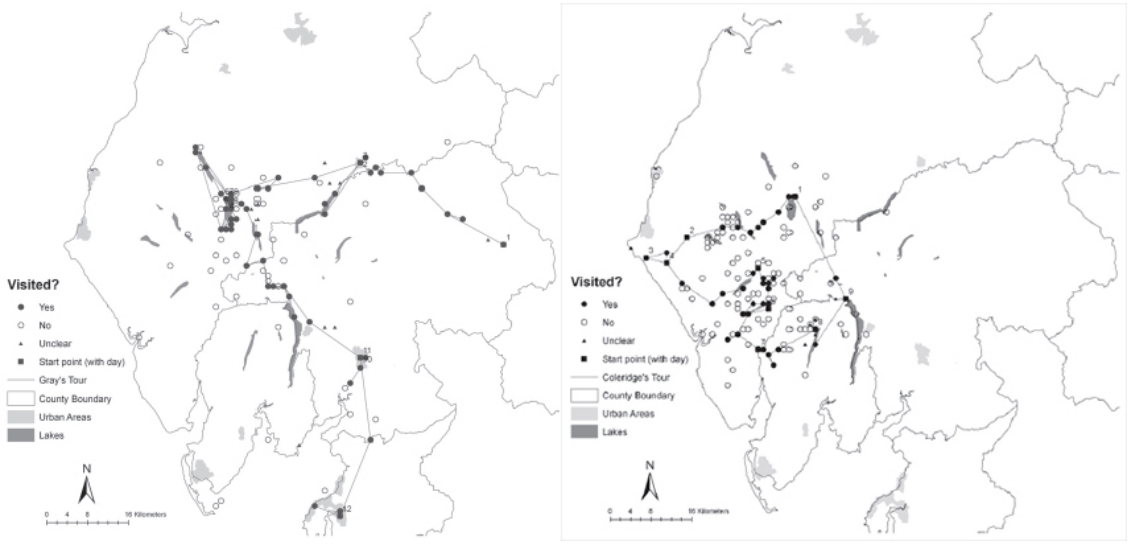


Figure 1. Dot maps of Gray's (left) and Coleridge's (right) accounts

Creating a GIS of these two accounts of the Lake District (which together comprise some 20,000 word tokens) involved a multi-step process. First of all, it was necessary to transcribe each text manually, and then to read through them to identify every place-name reference they contained. These place-names were then annotated using XML document mark-up, which allowed the research team to tag every place-name mentioned in Gray's and Coleridge's accounts. The tagged place-names were then paired with coordinates from a gazetteer to create a CSV file, which was then converted into a set of GIS-readable shapefiles using a GIS application, in this case ArcGIS. These shapefiles were then used to create a series of map layers displaying the underlying geography of the two writers' tours (Figure 1).

The maps displayed in Figure 1 shows us that Gray's and Coleridge's texts, though apparently accounts of the same region, are actually rather different. Gray's tour was a 15-day trip from Brough to Lancaster via Ullswater, Derwentwater, and Windermere. Coleridge's journey, by contrast, was a 9-day circuit from his house in Keswick to the village of St. Bees, and then from Scafell Pike to Eskdale, Coniston, and Grasmere before circling back to Keswick via Dunmail Raise (see Cooper and Gregory 2011). As we have observed elsewhere, these different routes are indicative of Gray's and Coleridge's differing attitudes towards the Lake District (see Murrieta-Flores et al 2014). The account provided in Gray's journal typifies the conventions of picturesque Lake District tourism (see Andrews 1989). Gray travelled by coach along the Lakeland's main roads, and he visited key sites near the region's main settlements. Coleridge, on the other hand, was residing in the Lake District when he went on his excursion, and the path of his journey suggests a desire to venture off the beaten track. The maps in Figure 1 confirm this general impression.

Dot maps, such as these, are useful in that they allow one quickly to discern general patterns. At the same time, however, dot maps can be difficult to interpret with accuracy. This is because dot maps generically assign each place the same symbol (a dot), and therefore tend to give the impression that every place marked on the map has an identical value. If either Gray or Coleridge mentions a single location more than once that location will appear multiple times in the data uploaded into the GIS application. But, to the naked eye, that location will only appear as a single place-mark on a dot map. This is because GIS applications tend to superimpose multiple place-marks over the same location.

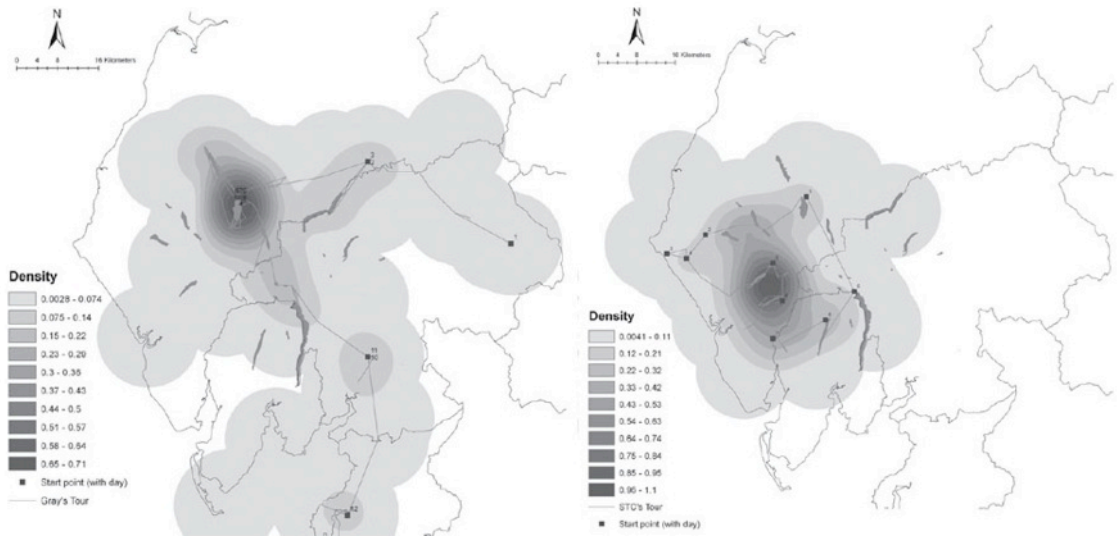


Figure 2. Density smoothing maps of Gray's (left) and Coleridge's (right) accounts

Analytical mapping techniques, such as density smoothing, can help us to interpret the geographies of Gray's and Coleridge's accounts more effectively. Density smoothing is a common method in GISc for identifying clusters of geographic events, such as place-name references (see Lloyd 2010). As the maps displayed in Figure 2 demonstrates, density smoothing works by generalising the distribution of the dot map to indicate the relative frequency of events (in this case, the locations associated with the place-names Gray and Coleridge mention) and the spatial proximity of those events to one another. Simply put, the more times a place-name is

mentioned, the denser the clustering (and the darker the colouration) around the location to which it refers. Consulting these maps helps us generally to discern the part of the Lakes region to which each writer paid more attention. For Gray, this is the area in the vicinity of the market town of Keswick, where he spent several nights during his tour; for Coleridge, this is the region around Scafell, which he famously ascended in early August 1802.

A full report of the outcomes of the Mapping the Lakes project has already been published (see Cooper and Gregory 2011). For the present purposes, it suffices to say that noticing these sorts of patterns has convinced us of two things: firstly, that working with GIS can help us begin to assess the experiences of individual travellers; secondly, that historical accounts of the Lake District can be surprisingly heterogeneous. These findings are what motivate the research of the Spatial Humanities and Geospatial Innovation in the Digital Humanities projects.

3. Current Research: The Corpus of Lake District Writing

Over the last five years the funding of the European Research Council and the Leverhulme Trust has facilitated the creation of a geo-referenced digital corpus of historical writing about the Lake District. To date this corpus comprises 80 individual works, which altogether contain more than 1.5 million word tokens. The earliest work in the corpus is the second edition of Michael Drayton's chorographical poem *Poly-Olbion* (1622); the latest work in the corpus is the twenty-second edition of *Black's Shilling Guide to the English Lakes* (1900), a bestselling Victorian Lake District guidebook. As these two titles imply, the corpus includes both canonical works of English literature as well as more ephemeral publications. Collectively, these works can be divided more-or-less equally into three historical sub-periods: the "long" eighteenth century (1688–1788); the Romantic period (1789–1836); and the Victorian era (1837–1901).[3] These classifications follow the system of periodisation commonly used in British History and English Literary Studies.

In addition to being reasonably large, the corpus of Lake District writing is also diverse. It includes a number of well-known works, such as the guidebooks of Lakeland luminaries like Wordsworth and the Victorian polymath Harriet Martineau. Crucially, however, it also features the works of dozens of lesser-known writers. This mixture of sources is significant. One of the main benefits of working with large corpora of historical texts, to reiterate Jockers's (2013) claims, is that it allows us to generate more historically nuanced interpretations. By situating canonical works, such as Wordsworth's *Guide*, alongside popular tourist publications, such as *Black's Shilling Guide*, we can perform far more complete analyses of how the Lake District was represented in the past.

Working with a corpus of more than a million words, however, poses challenges that were not encountered in the pilot study—specifically, scale. Given the number of words in our corpus, it would be optimal to develop techniques for automating the place-name identification and geo-referencing processes, instead of having to work through each text manually. Significantly, important steps have recently been made in this direction, particularly by the development of automated geo-parsing tools such as the Edinburgh Geoparser, which has been developed by researchers at the University of Edinburgh (see Grover et al 2010).

The Edinburgh Geoparser is a web-based geo-referencing tool that consists of two interlinked components. The first of these components is a *geo-tagger*, which uses a named entity recognition sub-component to identify the place-name entities in an HTML or plain-text input file. The tagger automatically inserts <enamex> elements around these place-names, and inputs the file to the *geo-resolver*, the second component, which resolves the tagged place-names by cross-referencing them with a preselected gazetteer, such as Geonames (<http://www.geonames.org/>). Grover et al (2010) provide a full report of the design and functions of the Edinburgh Geoparser system. What matters here is that the output of the system is a geo-referenced file that contains tagged place-names annotated with geographical coordinate information.

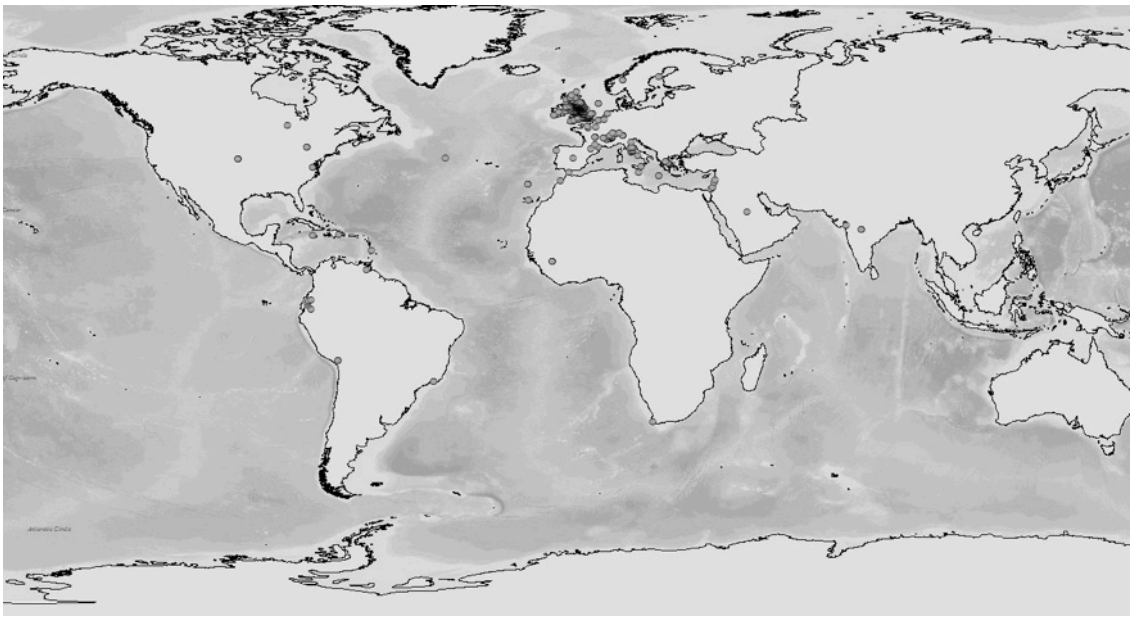


Figure 3. Distribution of the place-names in the gold-standard set

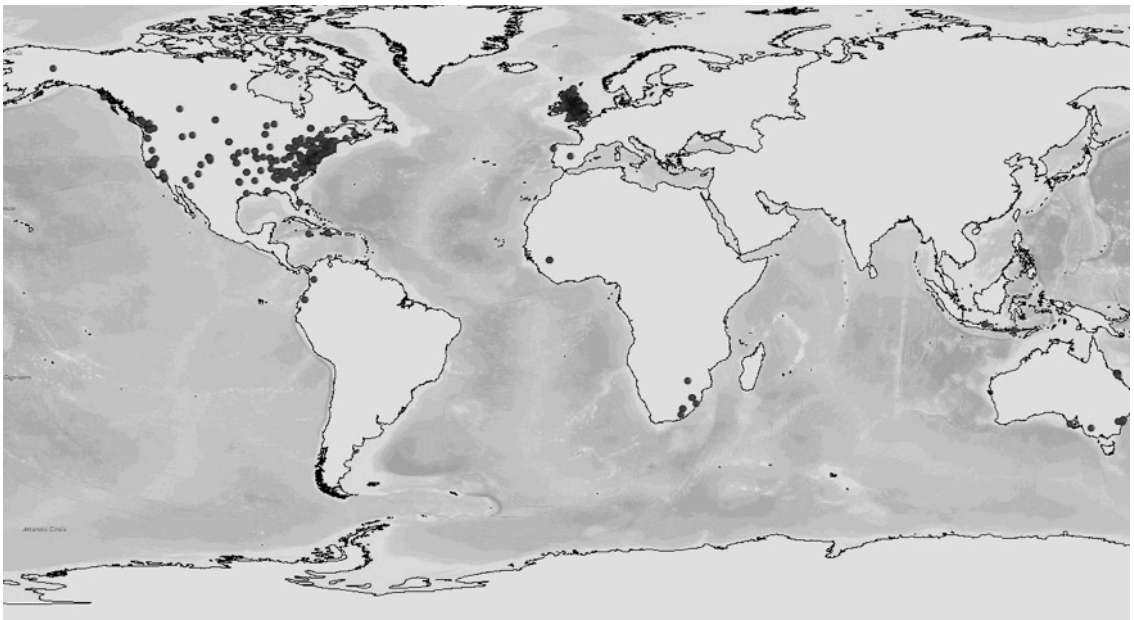


Figure 4. Distribution of the place-names tagged and resolved by the Edinburgh Geoparser

In order to test the effectiveness of implementing the Geoparser to examine our corpus of Lake District

Writing we conducted a case study using a gold-standard evaluation set of 28 texts, comprising 250,000 words: approximately one-sixth of our total corpus. These 28 texts were manually transcribed and coded, using the same procedure as in the pilot study, in order to ensure that all of the place-name references they contained were correctly identified (Figure 3). We then uploaded unstructured, plain-text versions of the same 28 texts to the Geoparser for automated geo-referencing. The results of the Geoparser output are displayed on the map in Figure 4. As a comparative assessment of the maps in Figures 3 and 4 illustrate, the Geoparser only recognised about 64% of the place-name references. Equally problematic, it assigned many of these place-names the wrong coordinates. The number of Lake District place-names geo-referenced to locations in North America (in Figure 3) alone indicates that although the Geoparser is a remarkable tool for corpus analysis it does have critical limitations.

One major limitation is that the Edinburgh Geoparser has difficulty in recognising variant spellings of place-names. Take a well-known Lake District place-name like Bowness, for example, which is commonly known as the port of Windermere. The works in our corpus contain at least four variant spellings of Bowness, including 'Boulness'. These variant spellings stem from the fact that the orthography of Lakeland toponyms was not established until the nineteenth century. Consequently, the writers of many of the earlier works in our corpus relied on oral sources for the names of the locations they visited (see Young 1770, 175). The Edinburgh Geoparser does not recognise variant spellings such as Boulness as place-names since they are not found in modern gazetteers. To further complicate matters, other variants of Bowness, such as 'Bonus', end up being assigned coordinates for other locations around the globe: notably, Bonus, TX, USA. With a place name like Bowness, moreover, there is another potential for error. After all, like many Lake District place names, Bowness can refer to more than one location: there is a place called Bowness near Windermere, as noted above; but there is another place called Bowness along the Solway Firth, which lies on the outskirts of the greater Lakeland region. The Edinburgh Geoparser is not well equipped to deal with this kind of ambiguity.

We have been able to mitigate these shortcomings by working with data from the historical gazetteer developed by the Digital Exposure of English Place-Names project (DEEP). Using the DEEP gazetteer, we have created our own customised Lake District gazetteer, and we have also made considerable headway by implementing NLP techniques to develop an iterative method of implementing, reviewing, and correcting the output of the Edinburgh Geoparser (see Rupp et al. 2014). This work has enabled us to geo-reference all 80 works in our corpus. The outcomes of this process confirm that our corpus contains 39,172 coordinate-based place-names. Of these place-names, 37,564 (96%) refer to locations in the British Isles; 34,530 (88%) refer to locations in northern England or southern Scotland. 23,459 (60%) of the place-names in the corpus refer to locations within the modern Lake District National Park. Having completed this foundational step, we have begun to utilise corpus analysis techniques to investigate what our corpus can tell us about historical responses to and descriptions of the Lake District.

4. Geographical Collocation Analysis: An Experiment

One of the chief aims of the research conducted by the Spatial Humanities and Geospatial Innovation in the

Digital Humanities projects is to study the different attributes that the works in our corpus historically assign to places in the Lake District. We are specifically interested to establish whether particular places are more strongly associated with certain attributes than others. We are also interested to determine whether the relationship between places and their attributes remains fixed or changes with the passage of time.

With these interests in mind, we have recently performed an experiment using collocation analysis to examine the relationship between the place-names in our corpus and a set of fourteen search terms—including keys words such as beautiful, picturesque, and sublime—all of which relate to particular kinds of emotional and aesthetic experiences. Collocation analysis is a corpus analysis method that allows one automatically to determine whether elements of lexis or annotation in a given text or corpus co-occur (or, in other words, *collocate*) more often than would be expected by chance alone.[4]

In order to perform collocation analysis it is necessary to determine the number of words on either side of the search term (here, each of our fourteen terms) that constitutes a position near the search term. We selected to use sentence boundaries as the bandwidth of proximity in this experiment. Despite the occasional errors of inclusion this bandwidth may introduce, we have discovered that in sentences containing multiple place-names the search term is frequently associated with each of the different locations. Consider, for example, the following sentence, from Edward Baines's 1829 guidebook *A Companion to the Lakes in Cumberland, Westmorland, and Lancashire*: "Again entering the boat, we passed up the channel between Lord's Island and the shore, from whence beautiful prospects are obtained of the majestic form of Skiddaw, with the woods of Castlehead and Cockshot Park in the foreground" (Baines 1829, 121). In this case, we are told that "beautiful prospects are obtained" not only of Lord's Island and Skiddaw, but also of Castlehead and Cockshot Wood. So a collocation with each named entity is not just acceptable, it is desirable.

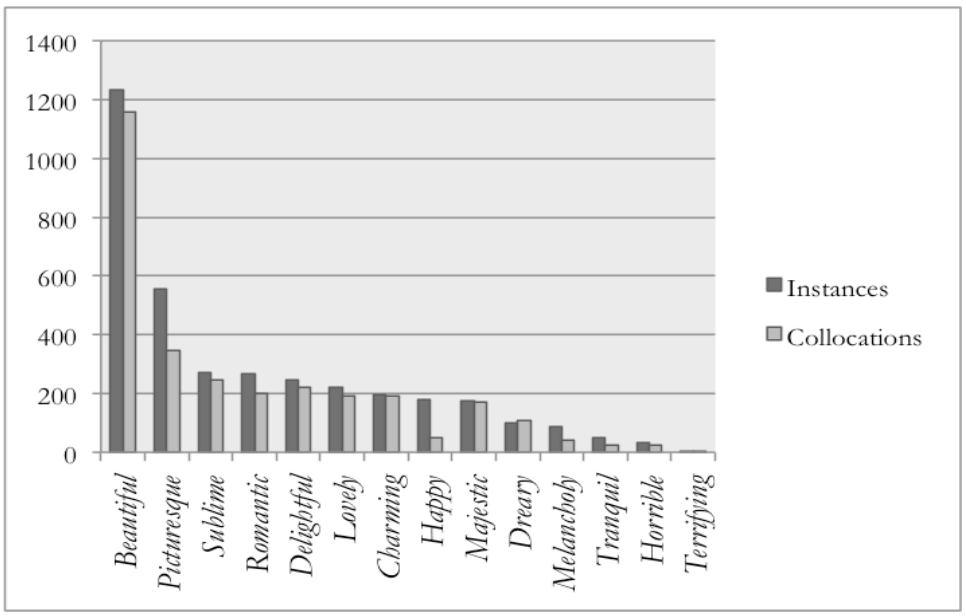


Table 1. Number of instances and collocations for the 14 search terms

We are still analysing the outcomes of this experiment, and we expect to publish a more detailed account of our findings soon. In what follows below we summarise some of our initial observations by focusing on the

results for the words "beautiful", "picturesque", and "sublime", which are three key aesthetic concepts in historical writing about the Lake District (see Andrews 1989).

Whilst studying the results of the collocation analysis, we have been especially intrigued by the number of occurrences of the word "beautiful" in our corpus. The word is used 1233 times, which is nearly half the sum of the occurrences of all the other search terms combined (Table 1). Even more intriguing than the frequency of the word "beautiful", however, is the rate at which it collocates with place-names in the corpus. As Table 1 indicates, "beautiful" collocates with an identified place-name 1157 times: in other words, in nearly 94% of its appearances. This suggests that beautiful is not only a word frequently used in our corpus, but also that it is a word that is often used in relation to locatable place-names. The results for the word "sublime", though it occurs far less often in the corpus, follow a similar trend. The word "sublime" collocates with a place-name in 247 of its 270 appearances: nearly 92% of the time. The word "picturesque", by comparison, occurs in the corpus roughly twice as often as "sublime" and nearly half as often as "beautiful", but it collocates with a place-name just 63% of the time: in 347 of its 554 occurrences.

For the most part, these findings can be said to make sense. Beautiful, picturesque, and sublime, as hinted above, are integral concepts to the development of the landscape aesthetics that influenced the appreciation of the Lake District during the eighteenth and nineteenth centuries. At the same time, however, it is notable that beautiful and sublime were words much more specifically defined in eighteenth- and nineteenth-century aesthetic philosophy. Picturesque, for its part, is a slightly more diffuse concept; it is a word one finds used both as a label for a cultural movement and as an aesthetic term. It therefore stands to reason that it might be used slightly less frequently in association with specific named locations.

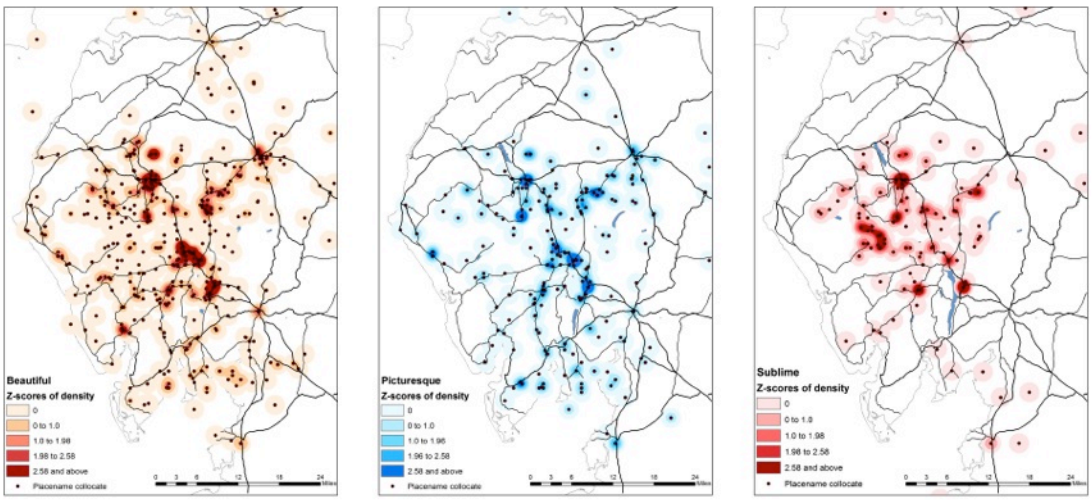


Figure 5. Distribution and density smoothing maps for beautiful, picturesque, and sublime

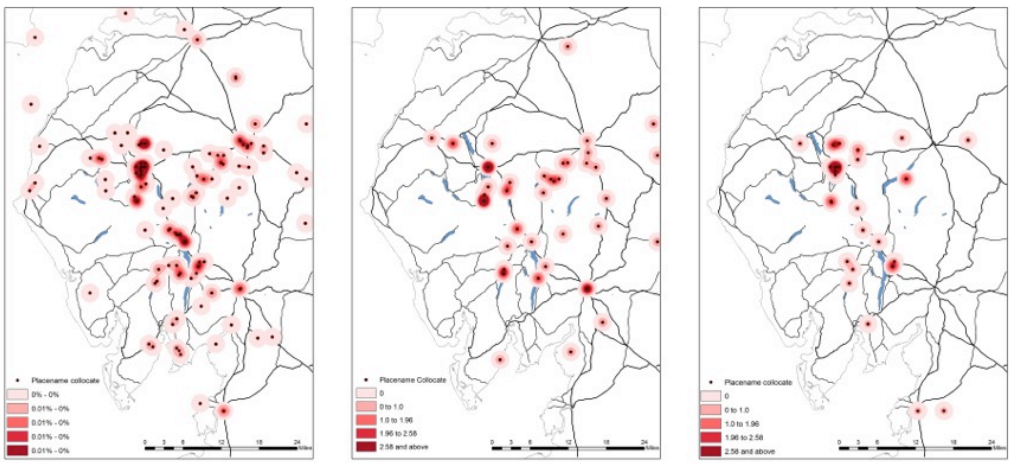
Such distinctions notwithstanding, it is surprising to see just how frequently the words "beautiful", "picturesque", and "sublime" are associated with locatable place-names in our corpus. This invites us to investigate the features and locations with which these three words were most commonly associated. In order to do this we need to use GIS to plot the distribution and density of the places with which they (Figure 5).

Studying the maps displayed in Figure 5 gives us a clearer sense of the areas throughout the Lake District with which "beautiful", "picturesque", and "sublime" are most commonly associated. What is most notable about these results is that whereas picturesque and beautiful most frequently occur in relation to locations near principle tourist centres (such as Grasmere, Ambleside, and Keswick), places that frequently collocate with sublime fall both near Keswick also in the region around Wasdale and Eskdale, which contains seven of the ten highest mountains in England. This finding suggests that the word sublime is most frequently associated with steep, mountainous terrain. In fact, if we tabulate the elevation of the locations associated with each search term—which can be done using a digital terrain model (see Gregory and Donaldson 2016)—we see that *sublime* is associated with places and features at high elevation more often than would be expected given the background geography of the corpus (Table 2).

% of place-name collocates			
	<300m	*300–600m*	*>600m*
Sublime	63.2	24.1	12.6
Corpus	78.1	13.9	8
% above expected			
Sublime	-19	73.5	57.9

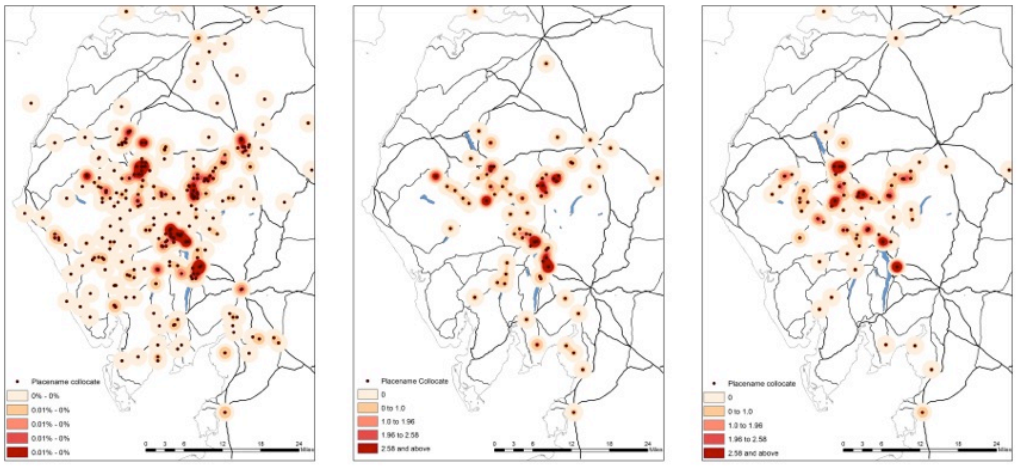
Table 2: Locations collocating with sublime, arranged by elevation

This impression is further reinforced when we consider the places that collocate with sublime in each of our historical sub-periods (Figures 6.1–6.3). As the figures displayed below indicate, it is not until the Romantic and, especially, the Victorian period that the word sublime begins to collocate with places in the western part of the Lakes. Although there are different potential explanations for this trend, it seems most likely that it is related to the fact that mountaineering did not emerge as a major tourist activity in the Lake District until the mid- to late-nineteenth century.[5]



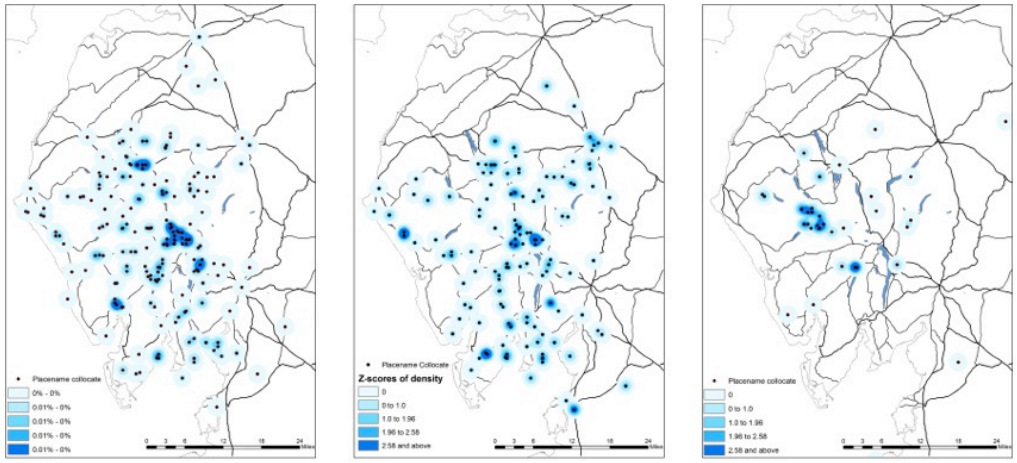
(a) (b) (c)

Figure 6.1: Maps for beautiful (a), picturesque (b), and sublime (c) in 'long' eighteenth century works



(a) (b) (c)

Figure 6.2: Maps for beautiful (a), picturesque (b), and sublime (c) in Romantic-period works



(a) (b) (c)

Figure 6.3: Maps for beautiful (a), picturesque (b), and sublime (c) in Vicotrian-era works

Conclusion

As indicated above, this research is still ongoing and we expect to publish a more detailed report of our findings soon. In the months to come, thanks to the funding received from the Leverhulme Trust, we will continue to expand the corpus of Lake District Writing, which will allow us to perform the sorts of analyses described above to an even larger collection of source materials. In addition, we will also be developing an exploratory environment for visualising and studying the geographic information in our corpus. For the present, the results summarised in this paper demonstrate the efficacy of combining geographical technologies and corpus-based approaches to study the qualitative features of historical text corpora.

Endnotes

[1] This project received funding under the British Academy's Small Grant scheme.