Geographical Text Analysis

Gregory, Ian; Donaldson, Christopher

CHAPTER 3

GEOGRAPHICAL TEXT ANALYSIS: DIGITAL CARTOGRAPHIES OF LAKE DISTRICT LITERATURE

Ian Gregory and Christopher Donaldson

1. Introduction

This chapter introduces an interdisciplinary approach to the geographical analysis of digital literary corpora. It does so by presenting a case study undertaken as part of Lancaster University's *Spatial Humanities: Texts, GIS, and Places* project. Combining corpus-based approaches, automated geo-parsing techniques and geographic information systems (hereafter GIS) technology, this study investigates literary responses to the landscape of the English Lake District (Figure 3.1). The focus of this investigation is a custom-built, 1,500,000-word georeferenced corpus of Lake District literature. This corpus consists of eighty digitised texts, ranging in date from 1622 to 1900. A historically representative sample of writing about the Lakes region, the corpus comprises a variety of canonical and non-canonical texts, including the works of Lakeland luminaries, such as those of William Wordsworth, as well as more ephemeral publications, such as *Black's Shilling Guide to the English Lakes*. In engaging with this resource, our aim is to exemplify how a hybrid corpus- and geographic-based methodology – which we label geographic text analysis – can be used in conjunction with more traditional forms of close reading and contextual analysis to understand how literary landscapes, such as the Lake District, were perceived and represented in the past.

[Insert here Figure 3.1 – Outline map]

Figure 3.1 The major geographical features of the Lakeland region

Source: The Authors; created using ArcGIS ©2015

2. GIS and the Geographies of Digital Literary Corpora

2.1 GIS and 'Macroanalysis'

GIS are a foundational form of digital geospatial technology that has been integral to innovation in the social sciences since the 1970s. In the humanities the application of GIS is a much more recent phenomenon, and is, in part, a result of the rapid growth of humanities computing and the proliferation of the digital humanities over the past twenty years. More broadly, the emergence of humanities GIS is also a consequence of the widespread adoption of geographical perspectives, approaches and techniques (or the 'spatial turn') across the arts and sciences (see Tally). Within literary studies, one of the more recent disciplines to take such a turn, the adoption of geographical principles and practices has chiefly been driven by pioneering research projects, such as ETH Zurich's *A Literary Atlas of Europe*, Trinity College Dublin's *Digital Literary Atlas of Ireland* and the University of Edinburgh's *Palimpsest: Literary Edinburgh* – to name only three. Taking inspiration from the ground-breaking work of Franco Moretti, Matthew L. Jockers and the Stanford Literary Lab, each of these projects has endeavoured to show how geospatial technologies can transform the way we engage with the geographical and spatial dimensions of individual literary works as well as those of large literary corpora.

Underpinning these endeavours is a conviction in the value not only of geographical thinking, but also – more specifically – of maps as 'analytical tools' for displaying information derived from literary works in ways which 'bring to light relations that would otherwise remain hidden' (Moretti, *Atlas* 3). Maps, in so many words, are valuable because they can serve the literary scholar both as instruments for generating abstract representations of particular aspects of specific works and, furthermore, as a means of compiling information from multiple works and of collectively assessing and comparing them. This latter process of aggregate analysis, which has variously been called 'distant reading' (Moretti, 'Conjectures')

and ‘macroanalysis’ (Jockers), represents a major advance for the discipline of literary studies, since, as Jockers explains, it enables literary scholars to make substantial use of the ‘massive digital-text collections’ now available to them and, in the process, to launch the discipline fully into the digital age:

> Today, in the age of digital libraries and large-scale book-digitization projects, the nature of the evidence available to us has changed, radically. Which is not to say that we should no longer read [individual] books … but rather to emphasize that massive digital corpora offer us unprecedented access to the literary record and invite, even demand, a new type of evidence gathering and meaning making. (Jockers 7-8)

Put simply, instead of engaging with only a handful of outstanding or exemplary works, literary scholars should create new knowledge about those works by studying them in relation to the larger corpora – whether construed historically (in terms of period), formally (in terms of genre, type or mode), tropologically (in terms of specific themes and motifs) or otherwise – into which they can be assembled. Geographical text analysis, the methodology introduced in the following pages, aims to facilitate just this sort of approach by augmenting the traditional methods of textual analysis employed in literary studies with techniques from geography and corpus linguistics (see Gregory et al.).

2.2 Geoparsing

The first step in performing geographical text analysis is geoparsing. Geoparsing involves identifying and extracting place-names from the corpus under analysis and assigning each place-name to a coordinate-based location. Geoparsing can, of course, be performed manually, but when working with a large corpus there are obvious incentives for automating

the procedure. In this case study we have used a customised version of the Edinburgh Geoparser: an open-source, automated georeferencing tool that consists of two interlinked components. The first of these components is a 'geo-tagger', which uses Named Entity Recognition (NER) technology to identify and extract place-names (including named settlements, landmarks and geological formations). The second component is a 'geo-resolver', which allocates coordinate data to the extracted place-names using digital gazetteers (Grover et al). Once the corpus has been geoparsed in this fashion, the georeferenced place-names it contains can be extracted, along with their relevant co-text (the text to the left and right of the place-name), and imported into a GIS application where it can be displayed and analysed.

It should, of course, be emphasised that automated geoparsing is not an error-proof process. Place-names are surprisingly complex for software to process automatically. Errors can occur for a variety of reasons. In some cases the software may simply fail to recognise that a word really is a place-name and this will result in it being omitted. There is also the potential for errors of inclusion stemming from the difficulty that even state-of-the-art software has in disambiguating between place-names, personal names and toponymic titles (such as the Bishop of Carlisle or the Duke of Devonshire). There is, moreover, the additional difficulty of disambiguating between places with the same name. As a result, when automatically geoparsing a historical text corpus, it is important to remember that the raw output produced will almost invariably contain oversights and inaccuracies that the researcher will need either to account for or, ideally, to identify and correct. In order to avoid these problems, which stem from an over-reliance on technology, we have developed an iterative method of implementing, reviewing and correcting the results from the geoparsing process (see Rupp et al.).

2.3 The Geoparsed Lake District Corpus: Initial Visualisations and Observations

In total, there are almost 40,000 instances of mappable place-names in the corrected Edinburgh Geoparser output for the Lake District corpus, which means that place-names account for roughly 2.6% of all the tokens in the corpus.[1] Of these nearly 40,000 place-names, 96% refer to locations in the UK, 88% of which are in northwestern England or southwestern Scotland: in other words, within and around the greater Lakeland region. Notably, only some 60% of these locations are within the boundaries of the modern Lake District National Park: a finding which reflects the fact that early literary accounts of the Lakes – such as John Dalton's influential *Descriptive Poem, Addressed to Two Ladies after their Return from Viewing the Mines at Whitehaven* (1755) – are often as concerned with the periphery of the region as its centre. This finding, furthermore, goes some way towards confirming the contention, expressed elsewhere, that the Lake District, though ostensibly bounded, is geographically extensive (see Cooper; Nicholson).

[Insert here Figure 3.2 – Dot map]

Figure 3.2 A dot-map of place-name instances taken from the Lake District corpus; only the area around the Lake District is shown.

Source: The Authors; created using ArcGIS ©2015

Once a corpus has been geoparsed and the geoparser output has been corrected and assembled in a GIS layer, one can begin performing a geographical text analysis by assessing the spatial dimensions of the geography that the corpus contains. The most elementary way to

[1] In corpus linguistics each occurrence of a particular word within a corpus is called an *instance*. The units that comprise a corpus are, moreover, called *tokens* rather than words. This is because a token can be a word, a numeral or a punctuation mark.

do this is by creating a dot-map in which the list of referenced place-names are converted into point-data and displayed as dots on the map-interface of the GIS. As shown in Figure 3.2, dot maps are a simple way of representing georeferenced data; however, they are of limited value for interpreting a large corpus of qualitative sources such as our collection of Lake District writing. This is in large part because they represent each place marked on the map in the same way and, as a result, tend to 'flatten-out' datasets instead of highlighting the variations they contain. Take a quality like the frequency of references to a specific place, for example. Dot-maps are inadequate as a means of representing frequency because, when displaying point data, GIS applications superimpose multiple place-marks in the same location (see Fotheringham et al.). Consequently, a dataset may contain dozens of references to a particular place and only one reference to another, but, on a dot-map, these places will appear in exactly the same way. An additional problem with dot-maps – and one of particular relevance for this study – is that they tend to imply an accuracy of location that may be misleading if the geographical entity in question extends over a significant extent of space. Think, for instance, of a mountain, lake or estuary.

In order to overcome these limitations it is useful to employ an analytical technique such as *density smoothing*, which is a common method for simplifying and displaying point patterns. Performing a density smoothed analysis involves calculating the number of points that occur near to each location on the study area (see Lloyd). The results can be presented using a 'heatmap' in which areas of the higher density are shown with darker shading. For the present purposes, it suffices to say that the application of this technique results in maps that are both easier to evaluate and that do not misleadingly imply that each place-name in the corpus corresponds to a single, precise location.

[Insert here Figure 3.3]

Figure 3.3 A density-smoothed distribution of place-names in the Lake District corpus; a z-score of 0 is the mean density of the dataset, 1.0 is one standard deviation above the mean, and 1.96 and 2.58 are expected to represent 5% and 1% of the values (two-tailed) in the dataset.

Source: The Authors; created using ArcGIS ©2015

To this end consider Figure 3.3, which displays the distribution and density of the places that are referenced in the corpus and which are located in and around the greater Lakeland region. Studying these maps enables us to draw a number of initial observations. The most significant of which being that whereas the dot-map (Figure 3.2) suggested that these places are spread around the Lake District, the darker shading on the density map (Figure 3.3) indicates a different underlying pattern. Instead of being more or less evenly distributed, the geography of the corpus is shown here to be marked by areas of greater and lesser density, with clusters of references forming in specific localities. These include the areas near Skiddaw and Keswick, as well as areas south of this including Borrowdale, Buttermere and, to a lesser extent, Scafell. Moving eastwards, one notices other significant clusters stretching from Thirlmere and Helvellyn, and Ullswater. Further south, there are the clusters around Grasmere, Ambleside, Langdale, Windermere and Coniston. There are also clusters centred on the larger settlements to the east of the Lake District including Lancaster, Kendal and Penrith (see Figure 3.1 for orientation).

That these localities are the ones most frequently mentioned in the corpus stands to reason. Each, after all, figures prominently enough in the literary and cultural history of the region to remain integral to our conception of Lake District heritage today. Yet, one must be mindful that density maps still need to be interpreted with care. Specifically, one must bear in mind that the density map depicts a generalisation of the pattern displayed in the dot-map,

and that, in certain cases, difficulties of disambiguation complicate our ability to distinguish where a cluster is centred precisely on the spot it should be. For instance, with locations such as Coniston and Windermere, towns which share their names with nearby lakes, one must be mindful that it is difficult, if not impossible, for an automated process, to determine which is being discussed. Notwithstanding such complications, density smoothing *does* allow us to identify that – collectively – the texts in the corpus contain a disproportionate number of references to places within specific parts of the Lake District. Concomitantly, it also allows us to perceive the existence of geographies of absence: areas which are either mentioned infrequently, such as the ones around Haweswater and Shap Fell, or which are ignored all together, such as the ones to the north of Bassenthwaite and to the south of Ennerdale. If, as seems reasonable, we are willing to use the frequency of place-name references within the corpus as an index for the amount of interest and attention received by any given area, these preliminary observations mean that even within the centre of the region, the texts in the corpus pay the most attention to a handful of key locations and pay far less attention – or altogether neglect – several others.

3. Comparing Descriptions of Different Locations

Viewing Figure 3.3 gives us a sense of the various places mentioned in our corpus of Lake District writing. It helps us, moreover, to discern the number of times these places are mentioned and thus, by extension, the amount of attention given to each. This, in turn, prompts us to investigate why certain places receive more attention than others. It also encourages us to find out what is being said about them. Examining these sorts of issues is integral to geographical text analysis. In order to do this, however, it is necessary to combine GIS-facilitated spatial analysis with complementary methods from corpus linguistics, such as collocation analysis: a basic approach for identifying words that are frequently paired with

specific named entities, such as place-names.

3.1 Skiddaw and Scafell

As a way of illustrating this process, let us compare the words that are frequently associated with two locations that are the sites of clusters in Figure 3.3: the area around Skiddaw and the one around Scafell. With peak elevations of 931m and 964m respectively, Skiddaw and Scafell are amongst the highest and most iconic mountains in the Lake District. This similarity aside, however, they are located in markedly different parts of the region and, as a result, have markedly different associations. Whereas the latter rises above the sloping, slate-rich fells in the north of the district, near the popular tourist town of Keswick, the former rises amongst the steep volcanic group in the centre of the region, near the head of the less-accessible and, notably, less-frequented valley of Wasdale. Performing collocation analysis of the place-names *Skiddaw* and *Scafell* enables us to ascertain whether or not – and if so, how – these differences are manifest in the accounts of the Lake District that comprise our corpus. In order to perform this analysis, however, one must first define what one means by Skiddaw and Scafell. This is not as simple a task as it may seem, since there are several named locations around each mountain that have names of their own but are nonetheless part of the same massif. Scafell Pike, the highest summit in England, which is one kilometre northeast from Scafell, but part of the same range, is one excellent example; as is Lingmell, which is an outlying shoulder of the mountain. In brief, to search for Scafell alone would be inadequate, as it would fail to take such contiguous formations into account.

One way of compensating for this would be to perform a collocation analysis not of Skiddaw or Scafell alone, but of all the place-names found within a defined radius of each summit. Arguably, however, this approach is unsatisfactory because the distance of the radius would be entirely arbitrary, and neither mountain range is circular in shape. A preferable

approach, and the one adopted here, is to perform density smoothing using a bandwidth determined by the formula presented by Fotheringham et al., and to define clusters using the resulting density smoothed pattern. Performing this analysis reveals a strong cluster of place-name references around Skiddaw and a weaker one around Scafell (shown in Figure 3.3). These clusters can be taken as indicative of the area of interest that corresponds to each mountain. For the Skiddaw cluster this gives us a list of nearly 700 place-names which contain the word Skiddaw, such as Skiddaw Fell, and a handful of variant spellings, such as Skiddow. The place-names that comprise the Scafell cluster are slightly more varied and include not only variant spellings, such as Scawfell, but also the names of other nearby landmarks, such as Great End, Styhead Pass, Styhead Tarn, Mickledore and Lingmell. Whether or not these places can be said collectively to constitute the location of Scafell depends on the nature of the research being undertaken. In this case study, we have decided to accept them as indicative of the extent of the Scafell range and to explore the consequences.

3.2 Initial Findings

In the first instance**,** simply counting and comparing the number of place-name references that occur in each cluster indicates key differences between the two. The cluster around Skiddaw contains 691 references; the one around Scafell includes only 533, which suggests that the former is a more prominent location in the corpus than the latter. This, in itself, may be unsurprising given that Skiddaw towers above Keswick, which is one of the more famous tourist resorts in the Lake District. At the same time, however, the difference between these two figures is not as great as one might expect it to be. (Given that more tourists flocked to Keswick than any other settlement in the region during the period represented in the corpus, one might fairly expect Skiddaw to be mentioned even more

frequently than it is.) Turning to the texts themselves, it is striking to note that whereas locations within the Skiddaw cluster are mentioned in sixty-one of the eighty works in the corpus, those within the Scafell cluster are only referenced in thirty-two.

All such inferences must be tentative, but these findings would seem to indicate that although fewer writers discuss the area around Scafell, those who do seem to mention the places that comprise it fairly frequently. Carrying this reasoning one step further, one might posit that although Scafell figures in fewer accounts than Skiddaw, the writers who mention the former devote a significant amount of attention to it. Intriguingly, consulting the corpus corroborates this contention, as it reveals that four texts – Edward Baines's *A Companion to the Lakes* (1829), C. N. Williamson's 'The Climbs of the English Lake District' (1884) and the 1853 and 1900 editions of *Black's Shilling Guide to the English Lakes*[2] – account for most of the 533 references to the area around Scafell. Each mentions place-names in the Scafell cluster more than fifty times. (By contrast, only one text in the corpus, Baines's *Companion*, references places in the Skiddaw cluster more than fifty times.) It, of course, might be argued that the significance of the number of references per text depends greatly on the length of the texts in question. Here, however, a statistical comparison of the number of place-names per million words in each text across the corpus reveals a similar pattern: namely, that whereas the Scafell cluster has four texts with over 250 place-name instances per million words, the Skiddaw cluster has only one. This finding suggests that although the area around Scafell is mentioned less frequently within the corpus than the area around Skiddaw, when mentioned it is discussed in more detail.

3.3 Collocation Analysis of Skiddaw and Scafell

[2] The 1900 edition of *Black's Shilling Guide* reproduces, with only slight modifications, the text of the 1853 edition.

Establishing the relative amount attention paid to locations within the Skiddaw and Scafell clusters is one thing; exploring how they are being portrayed is another. This is where collocation becomes relevant. As noted above, collocation analysis is a standard method within corpus linguistics for identifying words that appear unusually frequently or in close proximity to – that is to say, collocate with – one another. Evaluating collocation frequencies involves using statistical measurements (such as t-scores, the measurement we use here)[3] to compare how often each word occurs near the search-term in relation to the number of times it appears in the corpus as a whole.[4] Obviously, performing a collocation analysis on the basis of proximity requires the researcher to determine how many word tokens on either side of the search-term constitutes a position near the search-term. Here we have adopted a bandwidth of ten word tokens as a measure of proximity.

Table 3.1 A selection of statistically significant collocates around place-names in (and around) the Skiddaw and Scafell clusters; these collocates are words occurring within ten words of the search-terms (place-names). Only words with statistically significant t-scores and a minimum collocation frequency of five have been included.

Categories	**Skiddaw only**	**Both clusters**	**Scafell only**
Judgements of appearance	Picturesque, majesty, majestic,	Fine, lofty, magnificent, vast,	Rugged, steep

[3] A t-score is a statistical test that compares two samples. In this case allows us to compare the frequency of the word near to place-names with its frequency in the corpus as a whole

[4] For more on collocation see Barnbrook et al. The collocation analysis described here was performed using AntConc 3.2.4 (Anthony).

	awful, grand, beauty, beautiful, mighty, interesting	towering	
Physical features	Village, trees	Mountain(s), hill, top, summit(s), ridge, precipices, rock(s), crag(s),	Peak, chimney, chasm, cliffs
Weather	Sky, weather, sun, cloud(s), snow, mists		
Transport	Walk		Path, climbers, climbing

In order to determine the kind of language used to describe locations within the Skiddaw and Scawfell clusters, we used the names of the places contained in each as search-terms and recorded each of the statistically significant collocates (Table 1). For the Skiddaw cluster these included a range of complex aesthetic terms, such as *picturesque*, *beauty* and *beautiful*, *majesty* and *majestic* and *awful*. For the Scafell cluster, by contrast, the collocates comprise mainly words indicative of scale, size and physical appearance, such as *vast*, *lofty*, *steep*, *rugged* and *towering*. Intriguingly, whereas *vast*, *lofty* and *towering* are also significant collocates in the Skiddaw cluster, *rugged* and *steep* are not; nor for that matter are *cliff*, *chasm* and *peak*, all of which collocate with locations in the Scafell group. This, of course, can be said to make sense: the fells around Skiddaw, as noted above, are rounded and rise above a famous tourist resort, whereas those around Scafell are more escarped and remote. What is intriguing, however, is that this confirms what might be expected: that – at the level

of semantics – the texts in the corpus respond to this difference.

4. Keyword Querying the Geography of the Corpus

From the foregoing analyses we can draw two preliminary conclusions: firstly that the area around Scafell is described in more detail (but in fewer texts) than the area around Skiddaw; and secondly that, in general, the word most frequently associated with Skiddaw and Scafell can be seen to correspond to the physical and geographical differences between the two mountains. These findings affirm the merits of the hybrid corpus- and geographic-based approaches showcased in this chapter. Crucially, however, they do not exhaustively demonstrate their potential. Collocation analysis can, after all, work in the opposite direction. In additional to helping to determine the keywords that collocate with a given location, it can assist us in identifying the locations that collocate with any given search term. When we combine GIS and collocation, moreover, we can also explore the distributions of those locations and their relation to one another. With this in mind, consider the following examples, which are based on the distribution of the place-names that collocate with two of the key terms identified in the previous section: *awful* and *steep*.

4.1 Collocations with Awful

[Insert here Figure 3.44]

Figure 3.4 Place-name instances that collocate with *awful*

Source: The Authors; created using ArcGIS ©2015

Figure 3.4 displays the distribution of the places that collocate with *awful* within the

corpus.[5] As this map indicates, although the term *awful* collocates with Skiddaw, it is even more strongly associated with locations near Derwentwater and Borrowdale (the lake and the valley south of Keswick) and Ullswater and Patterdale (the lake and village southwest of Penrith). This pattern would seem to indicate that although the texts in the corpus occasionally describe mountains as awful, the term is more often associated with locations within some of the region's lower-lying valleys. Reading the texts that contain the word *awful* supports this impression. In her famous account of her journey through the Lakes in 1794, for example, the novelist Ann Radcliffe describes Ullswater as 'bounded on one side by the precipices of Place Fell, Martindale Fell, and several others equally rude and awful that rise from its edge' (258). Similarly, Charles Cooke's *Tourist's and Traveller's Companion* of 1827, which incorporates passages from Radcliffe's account, reports that the village of Patterdale is 'bounded by the precipices of Place Fell, Martindale Fell, and several others equally rude and awful, that rise from its edge' (74). Descriptions of Borrowdale, near Keswick, apply *awful* in a similar way. When describing a boat trip on Derwentwater, for instance, Edward Baines mentions passing 'Wallow Crag, whose awful precipice towers over the wood [that] spread around its base' (121). Radcliffe, for her part, makes note of 'the awful rocks, that rise over the fall of Lowdore' (349) when describing her tour through Borrowdale.

In each of these cases, we find writers using the word *awful* to describe mountains and cliffs viewed from a position of lower elevation. This suggests that, in the period represented by the corpus, the sense of awestruck wonder denoted by *awful* implied the appearance of something towering above the viewer. This impression is further supported by other instances the word within the corpus. Thus, in Thomas West's account of the vale of Keswick, we find

[5] *Awful*, in this context, should be understood in the eighteenth-century sense of the word: 'That which strikes with awe, or fills with reverence' (Johnson 139).

reference to ‘the skirts of Skiddaw, which raises here in awful majesty his purple front’ (97). Elsewhere, Cooke refers to ‘towering Skiddaw wrapped in awful shade’ (83). The list of examples goes on and on. In fact, the only notable exception to this trend is found in Baines’s *Companion*, which offers the following account of a journey above Langdale: ‘At a fearful depth beneath the summit lies the Stickle Tarn on one side, … and at the other side the more awful depth of Great Langdale … and descended an easier path to the level of Angle Tarn, which lies at an awful depth beneath the precipitous summit of Bowfell’ (247). Given that Baines is the only writer in the corpus who uses *awful* in this way, this exception can be viewed as useful in that it underscores a general rule.

In addition to implying a particular vantage point, *awful* also seems to be a word that only a particular subset of writers use. The word only turns up 155 times in the corpus, and more than half of these occurrences are found in five texts. Two of these texts (West’s *Guide to the Lakes* and Radcliffe’s *Journey*), were published in the late eighteenth century; the other three (Baines’s *Companion*, Cooke’s *Companion* and John Robinson’s *Views of the Lakes*) were published between 1827 and 1833. In each of these cases, except the latest text (Robinson’s *Views*), *awful* collocates with a place-name each time it occurs. In the other four texts which contain several instances of the word, it collocates with a place-name at least fifty per cent of the time. This suggest that, within the corpus *awful* mainly serves as a term for landscape description but only does so for a relatively limited period of time: approximately 50 years.

4.2 Collocations with Steep

Figure 3.5 indicates the distribution of the locations that collocate with *steep*. As noted above, this word is more strongly associated with the area around Scafell than around Skiddaw. Yet, as Figure 3.5 suggests, *steep* is more often used to describe the valleys than the

mountains of the Lake District. The localities that most commonly collocate with *steep* include Buttermere, Borrowdale, Keswick, Coniston, Ullswater and Patterdale, as well as the area around Grasmere, Ambleside and the Langdales. Examining the co-text surrounding these collocates reveals that *steep* most often occurs in the description of roads, routes and directions of travel. Thus, in Charles MacKay's *Poetry and Scenery of the English Lakes*, the tourist is advised that the 'road from Buttermere, through Newlands to Keswick, leads by a very steep ascent' (161). Likewise, Samuel Leigh's *Guide* tells us that '[l]eaving Coniston Waterhead, the tourist ascends a steep hill' (17). Harriet Martineau's *Complete Guide*, for its part, warns that the 'descent to all the Ambleside inns is steep' (42).

[Insert here Figure 3.5]

Source: The Authors; created using ArcGIS ©2015

Intriguingly, although roads are frequently described as *steep* in this manner, mountains are only rarely. Investigating the place-name collocates in the Scafell cluster illustrates this. Here, the only place-name within the cluster that collocates with *steep* is Sty Head Pass, the mountain road between Wasdale and Borrowdale. Surveying the whole of the Lakes region one notices the cluster of collocates near Helvellyn (the third highest peak in the district). Here again, however, inspection of the co-text reveals that these collocates occur in passages that describe journeys below the mountain. Hence, William Green's *New Tourist Guide* (1819) informs us that the road beside 'the lake passes … under the steep and shaggy brow of Helvellyn' (423). In sum, like *awful*, the word *steep* seems to imply a specific context. Whereas the former tends to imply a position in relation to a site of higher elevation, the latter tends to indicate a consideration or discussion of travel.

5. Calibrating Keyword Collocations against the Geography of the Corpus

One potential criticism of the maps in Figures 3.4 and 3.5 is that they fail to account for the fact that some place-names appear in the corpus more often than others and are, therefore, likely to collocate with keyword queries more frequently. As a result, one might claim that the foregoing maps tell us more about the number of times individual place-names are mentioned in the corpus than they do about the actual distribution of collocations with the search-term in question. When examining the cluster around Keswick in Figure 3.5, for instance, one might wonder whether this pattern appears because Keswick frequently collocates with *steep* or whether it occurs because Keswick is a place-name that recurs throughout the corpus and is, thus, more likely to collocate with any search term. Equally, one might wonder if those locations that appear to collocate with the search-terms only a few times – such as Workington and Whitehaven – appear to be less significant because their names are not mentioned as often in the corpus. In order to address these sorts of queries we need to augment the approaches to geographical text analysis outlined above by using more complex methods of spatial analysis. In effect, we need to use statistical scans to measure and compare the distributions from Figures 3.4 and 3.5 with the one displayed in Figure 3.3. In doing so, we need to identify both those locations where there are significantly more collocations in Figure 3.4 or 3.5 than would be expected, on the basis of Figure 3.3 (hereafter called hot-spots), *and* those locations where there are significantly fewer collocations in Figure 3.4 or 3.5 than would be expected on the basis of Figure 3.3 (hereafter called cold-spots). This can be done using a test called Kulldorf's Spatial Scan Statistic.[6]

[Insert here Figure 3.6]

Figure 3.6 Spatial scan statistic of place-name instances that collocate with *awful*; the symbols representing hot spots show all of the place-names from the corpus that lie within

[6] See Kulldorf; in this analysis the scan was implemented using SatScan (Kulldorf, *SatScan*).

the cluster regardless of whether or not they collocate with *awful*. Instances of *awful* are shown as smaller grey dots. The density smoothed surface, taken from Figure 3.3, is also shown.

Source: The Authors; created using ArcGIS ©2015

Figure 3.6 shows the results of applying Kulldorf's Spatial Scan Statistic to evaluate the distribution of place-name collocations with *awful*. The map displayed here is more complicated than the ones above; but, effectively, it shows us that the technique has identified two hot-spots: one on the eastern side of Derwentwater and another stretching along the eastern shore of Ullswater towards Haweswater and Kentmere. The first of these two clusters indicates that locations and landmarks along the precipitous south-eastern edge of Derwentwater do, in fact, collocate with *awful* more regularly than might be expected given the number of times they are mentioned in the corpus. In contrast, although other areas – such as the one around Skiddaw and the one extending from Borrowdale to the Langdales – also collocate with *awful*, the spatial scan suggests that this has more to do with the number of times they are mentioned in the corpus than it does with a particularly strong association with the search-term.

The second cluster identified in Figure 3.6 requires even more careful interpretation. As with the eastern side of Derwentwater, the area around Ullswater comprises a number of locations (including Patterdale, Place Fell and Martindale Fell) that are mentioned several times in the corpus. This might lead us to infer that these locations often collocate with *awful* simply because they appear so frequently. However, the spatial scan indicates that many of the names of these locations do, in fact, collocate with *awful* more regularly than might be expected. Thus, the observation that it is frequently described as *awful* is not simply a consequence of this place being commonly mentioned in the corpus.

[Insert Figure 3.7]

Figure 3.7 Spatial scan statistic of place-name instances that collocate with *steep*

Source: The Authors; created using ArcGIS ©2015

Figure 3.7 displays the output of the spatial scan of the place-name collocations with *steep*. Here the pattern is more complex than the one produced by the spatial scan of collocations with *awful*. One noticeable difference between the two is the presence of cold-spots (that is, places that collocate less frequently with the term than would be statistically expected). Clusters of cold-spots are found throughout the peripheries of the central Lakeland fells, from the Cumbrian coastal plain eastwards to the Eden Valley and southwards to the south Lakes. Upon reflection, this might be said to make sense: each of these areas are, after all, distinguished by relatively level terrain and thus, one suspects, would be unlikely to be regarded as steep. In contrast, the distribution of hot-spots indicate that only a handful of key localities in the upland centre of the region – including the areas around Patterdale, Borrowdale, Coniston and the Langdales – collocate with *steep* more often what one would statistically expect. Panning out, moreover, one notices an intriguing difference between the clusters identified by the density-smoothed surface and those identified by the spatial scan. These include the areas around Ambleside, Keswick and Buttermere. Each of these settlements is frequently mentioned in the corpus. Each, furthermore, is a popular tourist destination from which steep roads radiate. Yet, for all that, the spatial scan indicates that although they often collocate with *steep*, this has more to do with the number of times they are mentioned than with a significant association of with the word *steep*.

Comparing the patterns in Figures 3.6 and 3.7 furnishes us with some significant insights into the geography of the corpus, specifically because it shows us the place-names

associated with *awful* and *steep* do not overlap. They do occasionally come close to one another: the fells east of Derwentwater are *awful*, whereas Borrowdale, just to the south, is *steep*; and whereas Patterdale is *steep*, the fells east of Ullswater are *awful*. In general, however, the two terms are associated with different localities. This finding is intriguing since it illustrates how we can explore in detail the ways in which different parts of the Lake District are perceived and represented. Noticing this pattern in the descriptive semantics of the corpus invites us to assess whether or not it correlates with patterns in the physical geography of the Lakes region itself. In order to do this, however, it is first necessary to incorporate additional contextual data within the GIS environment.

6. Incorporating Contextual Data: Digital Terrain Models

One of the most useful and underrated features of GIS technology is its ability to integrate georeferenced data from different sources. For the purposes of this case study, this means that we can add extra contextual data – such as topographic relief data – to enhance our analyses. The standard was of incorporating relief data is by employing digital terrain models (DTMs), which use pixel-data to represent the surface of the earth (Burrough and McDonnell). Consider Figure 3.8, which displays a DTM for the Lake District on which the place-name collocates with *awful* have been superimposed.

[Insert here Figure 3.8]

Figure 3. 8 A Digital Terrain Model of the Lake District with *awful* place-name collocations; the coast line and the DTM do not match up completely because of differences in tidal definitions.

Source: The Authors; created using ArcGIS ©2015

One benefit of examining point-data, such as the collocates displayed in Figure 3.8, against the backdrop of a DTM is that it enables us to examine how their distribution maps onto the contours of the terrain. A key feature to consider here is elevation. Once a layer displaying point-data has been superimposed on a DTM, one can merge information from the two data tables and, in doing so, assign each item of point-data a height above sea level. One must, of course, remember that the accuracy of the location of individual points may vary. Nevertheless, using the DTM to assign each point an elevation is an efficient and reasonably reliable way of assessing whether items of data – such as collocations with *awful* and *steep* – occur at a range of different heights or whether they tend to cluster around particular altitudes. With this in mind, have a look at the graphs displayed in Figure 3.7, which document the different elevations of the places whose names collocate with *awful* and *steep*. In both cases the elevation of all the locations mentioned in the corpus (represented by a dotted line) have been included for comparison.

[Insert here Figure 3.9.1 and 3.9.2]

Figure 3.9 The heights of the instances of (1) *awful* and (2) *steep* instances compared to all instances in the corpus; 'cluster' refers to instances occurring within a cluster identified by Kulldorf's scan statistic. Positive clusters are hotspots, negative are coldspots. There are no coldspots for *awful*.

Source: The Authors

As Figure 3.9a indicates, whereas most of the places in the corpus are below 200m, those that collocate with *awful* tend to be located at moderate to high altitudes (between 200m and 600m), and are unusually common between 300 and 600m when compared to the corpus as a whole. Places that collocate with *steep* follow a similar pattern in that they

become unusually common above 300m. This suggests that whereas the highest parts of the Lake District (those above 600m) are sometimes associated with *steep*, they tend not to be associated with *awful*. Cold-spots for *steep* are, unsurprisingly, commonly found in low areas. In sum, this analysis confirms the pattern described in the previous sections: namely, that both *steep* and *awful* are associated with valleys and mountain passes. High mountains are sometimes described as *steep*, but this tends to be obscured because mountain places are mentioned less frequently than locations in the surrounding valleys.

The previous sections indicated that these two terms were used in different ways and that they were associated with different geographies. Although *awful* occurs relatively infrequently, when it does appear it is typically used to describe high landmarks viewed from a point of lower elevation. The eastern sides of the Derwentwater valley and of Ullswater are localities associated *awful* particularly often, as is Harter Fell at the southern end of Haweswater. Places in and around Borrowdale also frequently collocate with *awful*; however, this is at least in part because they are mentioned so frequently in the corpus. *Steep*, by contrast, tends to be particularly associated with roads and the places they connect. Once the popularity of these places is taken into account, however, passes such as Honister and Wrynose still stand out as significant collocates with *steep*. Some of the higher fells occasionally collocate with *steep*; although this is not immediately apparent because they are not mentioned very frequently in the corpus. As with *awful*, *steep* frequently occurs in passages that describe the high slopes of valleys and hillsides seen from a lower altitude. These valleys and hillsides are, however, in different places than the ones associated with *awful*.

7. Conclusions

In demonstrating the application of geographical text analysis, this chapter has

illustrated how corpus-based and geographical approaches can be integrated to facilitate the study of any corpus. Our discussion has focussed on how these approaches enable the researcher to explore the underlying geography of a corpus by identifying, extracting and displaying its geographical information; to apply collocation analysis to examine how the corpus thematises this geographical information; to assess how specific concepts and terminology from the corpus relate to its underlying geography; and to incorporate additional contextual data to enhance his or her analysis. In general, this chapter has indicated that the summaries produced by geographical text analysis – namely: maps, tables and graphs – are effective tools for identifying geographical patterns within the text. It should be noted, however, that these tools are not capable of providing definitive answers about those patterns. In effect, they pose questions and point the researcher towards the parts of the corpus that will likely contain the answers to them. In this chapter, these tools have been applied in an exploratory way and many potential questions they raise have not been explored. A more applied paper will focus more on comparing preselected places, search-terms, times, writers or genres to explore the basic question of how different language is associated with different places. This work remains to be done; however, the potential of geographical text analysis as a method for aiding our understanding of the geographies within texts should be clear in which it can provide a framework within which we can improve our understanding of the geographies within texts.

Acknowledgments:

The research leading to these results has received funding 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 no. 283850).

Works Cited

Anthony, Laurence. AntConc Homepage. *Laurence Anthony's Website*. 2014. Web. 20 March 2015.

Baines, Edward. *A Companion to the Lakes.* 2 ed. London: Hurst, Chance & Co., 1830.

Barnbrook, Geoff, Oliver Mason and Ramesh Krishnamurthy, *Collocation: Applications and Implications*. London: Palgrave Macmillan, 2013.

Burrough, Peter A., and Rachael A. McDonnell, *Principles of Geographical Information* Systems. 2nd ed. Oxford: Oxford University Press, 1998.

Cooke, Charles. *The Tourist's and Traveller's Companion to the Lakes.* London: Sherwood, Jones & Co., 1827.

Cooper, David. 'Poetics of Place and Space.' *Literature Compass* 5 (2008): 807-21.

Fotheringham, A. Stewart, Chris Brunsdon and Martin Chartlon, *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage, 2000.

Green, William. *The Tourist's New Guide*. Vol. 1. Kendal: Lough & Co., 1819.

Gregory, Ian**,** David Cooper, Andrew Hardie and Paul Rayson. 'Spatializing and Analysing Digital Texts: Corpora, GIS and Places.' *Spatial Narratives and Deep Maps*. Ed. David Bodenhamer, John Corrigan and Trevor M. Harris T. Bloomington: Indiana University Press, 2015. 150-178

Grover, Claire, et al. 'Use of the Edinburgh Geoparser for Georeferencing Digitized Historical Collections.' *Philosophical Transactions of the Royal Society A* 368 (2010): 3875-89

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Urbana-Champagne: University of Illinois Press, 2013.

Johnson, Samuel. *A Dictionary of the English Language*. Vol. 1. London: Strahan, 1755.

Kulldorf, Martin. ‘A spatial scan statistic.’ *Communications in Statistics: Theory and Methods* 26 (1997): 1481-96.

---. *SatScan*, 2015. Web. 20 March 2015.

Leigh, Samuel. *Leigh's Guide to the* Lakes. London: Leigh, 1830.

Lloyd, Christopher D. *Local Models for Spatial Analysis*. Boca Raton, FL: CRC Press, 2007.

MacKay, Charles. *The Scenery and Poetry of the English* Lakes. London: Longman, 1846.

Martineau, Harriet. *A Complete Guide to the English Lakes.* Windermere: J. Garnet, 1855.

Moretti, Franco. *Atlas of the European Novel, 1800-1900*. London: Verso, 1998.

---. ‘Conjectures on World Literature.’ *New Left Review* 1 (2000): 54–68.

Nicholson, Norman. *Greater Lakeland*. London: R. Hale, 1969.

Radcliffe, Ann. *A Journey Made in the Summer of 1794*. Vol. 2. 2nd ed. London: Robinson, 1795.

Robinson, John. *Views of the Lakes in the North of England*. London: Whittaker & Co., 1833.

Rupp, C.J., et al. ‘Dealing with heterogeneous big data when geoparsing historical corpora.’ *Proceedings of the 2014 IEEE Conference on Big Data* 2015: 80-83.

Tally Jr., Robert T. *Spatiality*. Abingdon: Routledge, 2013.

West, Thomas. *A Guide to the Lakes*. London: Richardson and Urquhart, 1778.