CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Lancaster E-Prints

# Qualitative Geographies in Digital Texts: Representing historical spatial identities in the Lake District

**Rob Smail, Ian Gregory and Joanna E. Taylor**

This is an Author's Accepted Manuscript version of an article published by Edinburgh University Press in the *International Journal of Humanities and Arts Computing* (2019, vol. 13, pp. 28-38). The Version of Record is available online at: *https://www.euppublishing.com/doi/full/10.3366/ijhac.2019.0229).* This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License.

**Abstract:**

Techniques for extracting place names (toponyms) from texts and using them to conduct analyses of the geographies within the texts are becoming reasonably well established. These are generally referred to as Geographical Text Analysis (GTA) and allow us to ask questions about the geographies within a corpus. The problem with this approach is that the geographies that can be uncovered are solely associated with toponyms for which a coordinate-based location can be found. While this method is valuable, it is effectively a quantitative representation of the geographies associated with named places. Other representations of geography are ignored. To complement GTA, we need to develop techniques that are capable of representing the more qualitative representations of geography that are found within texts. Drawing on the Corpus of Lake District Writing, this paper presents some initial ideas about how this can be achieved, primarily by using techniques from corpus linguistics.

## Introduction:

As large corpora of digital texts have become widespread in their availability, so the challenge of analysing them in ways that go beyond simple keyword searching becomes ever more pressing. These approaches need to make use of the computer's ability to identify and summarise patterns in large bodies of content, and incorporate the human's ability to interpret and contextualise these patterns. It is only when this human element is embedded as a core part of computer-aided analysis that we can satisfactorily answer humanities-led research questions. One area that has shown considerable potential in this regard is the analysis of geographical information in textual sources.

Through the use of Geographical Text Analysis (GTA), we can ask three basic questions: what geographies are present in this corpus?; what geographies are associated with a particular theme within the corpus?; and, what themes are associated with this place or set of places? GTA allows us to conduct trans-historical analyses of spatial identities that are simultaneously extensive and nuanced – but it has relied, up until now, on the geographies associated with one particular type of location: named places.

GTA has become an increasingly widespread practice in the last five years, and has been applied to a wide range of textual sources and research questions including: nineteenth century public health reports (Murrieta-Flores et al 2015); nineteenth century newspapers (Gregory et al 2016; Porter et al 2015); Lake District writing (Donaldson et al 2017; Taylor et al 2018); Early Modern letters (Gregory et al 2019); and modern newspapers (Paterson & Gregory 2018). These projects all used approaches based broadly on Claire Grover et al.'s work on geoparsing unstructured texts (2010). Geoparsing is a two stage process. In the first stage, Named Entity Recognition (NER) techniques attempt to identify all of the toponyms (place names) within a corpus. In the second, these toponyms are compared to a gazetteer and a geographical co-ordinate is allocated to each toponym. In theory, the results of this allows every named place in the corpus to be read into GIS software, therefore opening it up to an array of geospatial mapping and analytical techniques that explore our first question – what geographies are present within the corpus?

Answering our second and third questions require us to associate locations with one or more themes. To do this we use Place Name Co-occurrences (PNCs) – which identify any toponyms within a given number of words (span) of a search term – to associate place names with a concept or theme that we might want to investigate. PNCs also provide us with a simpler, and often more accurate, way of geoparsing a text. By just geoparsing the text surrounding a search term we can restrict geoparsing to relevant parts of the corpus making the process quicker and easier to check for errors, a process known as concordance geoparsing (Rupp et al 2014).

**Quantitative geographies of tourists and travellers**

**[Figure 1 PNCs for 'tourist[s]' using (a) points and (b) density smoothing]**

Figure 1 shows an example of this approach in action. It is based on the Corpus of Lake District Writing (CLDW), a collection of eighty texts written about the English Lake District between the seventeenth and nineteenth centuries. It includes works by major figures such as Thomas West and William Wordsworth, as well as a range of lesser-known authors such as Harriet Martineau, Priscilla

Wakefield and Edwin Waugh.[1] Figure 1a shows the PNCs created where toponyms are found within ten words of the search terms "tourist" and "tourists". There are 405 instances (or occurrences) of "tourist[s]" in the corpus, and these form 332 PNCs (or 82 PNCs per 100 instances).[2] The maps in Figure 1 represent the toponyms that occur within ten words of these terms. They are, then, places that we are assuming that the corpus associates with "tourist[s]". Figure 1b simplifies this pattern to make it more understandable by using a technique known as density smoothing, which identifies the locations where points cluster near to each other (Lloyd 2011). A clear pattern emerges: "tourist[s]" are associated with particular places including, Bowness and Windermere in the south; Ambleside, Grasmere and Langdale in the central Lakes; Penrith in the north-east; Keswick in the north, and Buttermere in the north-west. Other parts of the Lake District seem not to be associated with "tourist[s]" to any great extent. What Figure 1b suggests is that Lake District tourism, as represented by the corpus, was a highly geographical experience concentrated on a set of well-defined locations.

**[Figure 2 PNCs for "traveller[s]" using density smoothing]**

**[Table 1: Instances and PNCs for tourist and traveller]**

Figure 2 begins to highlight some of the potential limitations with this method. Like Figure 1b, it is a density smoothed map, but this time the geography of the PNCs for "traveller[s]" is plotted This pattern is more dispersed across the Lake District than the geography for "tourist[s]". This is not surprising: as table 1 shows, there are 608 instances of "traveller[s]" in the corpus, or around 50% more than there are of "tourist[s]". However, these instances only result in 85 PNCs, a rate of 14 per 100. A PNC-led approach to GTA therefore suggests that travelling is a far less geographical experience than tourism. Our own experiences of the corpus tell us that this supposition is nonsensical. The difficulty is that, in the approaches used by the existing GTA methods, geography has been defined solely by named places – and, even more restrictively, by those named places for which a geographical coordinate can be prescribed during geoparsing. There is a further problem with this kind of representation: when geography is narrowly defined as places to which a precise geographical co-ordinate, the experiential geographies that make up each individual's daily lives are not well represented.

The reason for the differences found in table 1 is that geographies associated with tourism are closely associated with named places. "Tourist[s]" are told which places to visit and what they are likely to experience in them. Travelling, by contrast, is a more ephemeral experience in which the

[1] https://www.lancaster.ac.uk/fass/projects/spatialhum.wordpress/?page_id=64

[2] Note that an instance can lead to more than one PNC, for example, a sentence such as "The tourist goes to Ambleside and Grasmere" has two PNCs (Ambleside and Grasmere) for only one instance.

individual is expected to experience the landscape in a more personal way. Here precise locations are less important and more emphasis is placed on exploration and engagement. In short, using toponyms to represent geography, while useful in the study of certain location-specific themes, is both limited and highly quantitative. How we can explore more qualitative geographies, and include them to advance humanistic analysis, is one of the more pressing question facing the spatial humanities.

**Qualitative geographies of tourists and travellers**

One way to explore non-quantitative geographies is to reverse the process we outlined for GTA, and to start instead with the last question: what geographies are associated with a particular theme? In this case, though, we are not looking for place names; rather, we are looking for any geographical references. This might include toponyms for which we can allocate precise geographical co-ordinates (e.g. 'Grasmere', 'Wastwater'), but it might also comprise more general terms which are harder to situate geospatially including: nouns referring to physical features (such as 'the mountain', 'the road', 'the lake'); other geographical nouns (for instance, 'the view', 'to the north'); and geographical verbs (like 'walks', 'travels', or 'returns'). Of course, other nouns, verbs, adjectives and adverbs may also provide information relevant to geography.

There are two potential ways of identifying these additional terms. We might simply assess the frequency with which words occur in the co-text, or we can use a corpus linguistics technique (such as keyness) to compare how frequently these words occur in the co-text compared to what would be expected from the corpus as a whole (Adolphs 2006; Baker 2006). While both of these approaches work in identifying spatial references, there is an important difference between them: the first identifies what geographies are associated with a particular search-term, and the second identifies the spatial terms that occur more frequently with the search term than we would expect given the background language in the corpus.

**[Table 2: Numbers of keywords associated with tourist and traveller]**

**[Table 3: Keywords associated with tourist and traveller]**

Some basic results of this method, using the fifty most statistically significant keywords for each of the two search terms, are shown in tables 2 and 3. As before, these tables suggest that "tourist[s]" do seem to be written about in more explicitly geographical ways than "traveller[s]": 33 of the 50 most statistically significant keywords which co-occur with "tourist[s]" are one of the three types of geographical nouns or geographical verbs. As well as toponyms, the nouns referring to physical features typically refer to types of accommodation (quarters, inn[s], resort, etc) or ways of getting

around (road, bridge). Other geographical terms are typically transactional: for example, directions, routes, and distances measure where the tourist should begin and end their excursions, and additional information such as prices, ticket details and maps further guide the tourist in very deliberate ways around the region. It becomes clear that "tourist[s]" should be considered a geographical entity, because they need a lot of information just to get from one point to the next; the practicalities of tourism, in other words, are intrinsically geographical.

"Traveller[s]" are associated far less with precise geographical language, on account of their greater willingness to explore unfamiliar places in an unguided way. Only one toponym – Wythburn, a relatively obscure place when compared to Keswick, Ambleside or Buttermere – co-occurs with "traveller[s]", as well as one each of "inn", "accommodation" and "road". Instead, the language associated with travelling tends to be much more descriptive. The large number of adjectives – including "curious", "sylvan", "fastidious", "changeful", "antediluvian", and "superficial" – contrast with the only adjective ("useful") that co-occurs with "tourist[s]". Similarly, other nouns associated with "traveller[s]" are qualified by terms that signal degrees of appreciation– such as "proof", "beauties", "pleasure" and "sentimentalist" – rather than the transactional ones associated with tourists.

**Towards spatial representations of non-specific geographies**

Once we have identified non-geographical spatial terms, we can turn to our second question: what themes are associated with this place or set of places? In GTA, this approach refers only to the themes associated with the one or more toponyms that are found near the search term, and which are statistically significant when the co-text around the toponym is compared to the corpus as a whole (Paterson & Gregory 2018). An alternative approach is to identify one or more places (or types of place) and identify the words that co-occur with them – to reverse the process, in other words. We might identify these places through a pre-defined list based on our existing knowledge of the texts, or – to maintain a more quantitative approach – we might base these choices on the frequency with which they appear in the corpus. Based on frequency, the most common nouns referring to physical features in the corpus are: lakes[s], mountain[s], water[s], road[s] and house[s], while the most common other geographical nouns are: view[s] and scene[s, ry].

**[Table 4: Collocates associated with mountains and roads]**

Having identified these terms, we can add further nuance to our question. Now, we can ask: what themes are associated with non-toponymic geographical terms, and how do these differ? Table 4 shows an example of this approach, using a similar framework to tables 2 and 3. As with previous

GTAs, this method identifies collocates, or words that occur unusually often in the co-text around a search term (Huston 2002; McEnery & Hardie 2011). In this case, we have used a span of five word tokens around the search term, and measured the significance of the geographical term with t-scores. We can see that toponyms of popular places are closely associated with roads, but not with mountains. This phenomena is thanks to phrases such as "…the road from Penrith to Keswick…" (Robinson 1819) or "I left Keswick and took the Ambleside road…" (Anon 1852). Mountains, by contrast, are rarely named: the general impression is more important than their individual identities (Donaldson et al., 2017). However, the lack of toponyms associated with mountains might also suggest that there was more diversity in the language used to describe mountains compared to the relatively small number of places that roads travel between. Not surprisingly, roads are also associated with distances and directions ("miles[s]", "side", "left", "right"), and seem to have been a source of appreciation throughout the period: roads are labelled as "good" and "great", while mountains are dismissed as being merely "lofty". The relatively sparse language around mountains, compared to roads, may be a reflection of the corpus itself which contains a large number of texts that describe well-traversed routes around the valleys, lakes and towns of the Lake District, while few are interested in heading up into the mountains.

**Conclusions: Moving towards QSR**

Elsewhere in this volume, Stell argues for the use of Qualitative Spatial Representation and Reasoning (QSR) as a way of gaining a better understanding of geography in the humanities. In a conventional GIS database, each place is allocated to a precisely defined co-ordinate-based location that is usually represented as a point, line or polygon (Gregory & Ell 2007). The relationship between places in the database is subsequently defined by the differences in coordinates on a Euclidean plain, which provide the distances and directions used in both mapping and spatial analysis. In doing so, the GIS imposes a geographical order that the writer is highly unlikely to have perceived. However, where geospatial features cannot be allocated to a co-ordinate-based location, they cannot be included into a GIS database. That has meant that to- date non-toponymic geographies have been largely ignored in spatial analysis. We highlighted this problem using the search terms "tourist[s]" and "traveller[s]". We would expect both terms to be closely linked to the geography of the Lake District. Instead, the results for each term were very different. Only "tourist[s]" showed strong links with toponyms. In contrast, "traveller[s]" did not appear to have very strong geographical links when analysed in a GIS, as much of the experience of travelling is not tied to specific named places.

QSR takes a different approach. Places are defined as regions, but no geometric information is required about their locations. Instead, locations are defined using connections – the relationship between one region and another region. In this paper we have introduced how and why we can define places in qualitative ways that do not require coordinates. Whether these are “Keswick” or “the road”, these can be used as regions in a QSR analysis. We have also used corpus linguistic based methods to provide approaches that can help to reveal the characteristics of these regions as described by the writers themselves. The next challenge is to use the texts to identify connections between regions, so that we can develop an understanding of spatial representations that go beyond toponyms, and can thus avoid imposing a coordinate based geography that may not be particularly relevant to human experience or humanistic analysis. The approach we have outlined here updates previous GTAs to provide a text-led framework for qualitative representations of space that is allied to both the methodological requirements of literary studies and the spatial ones of geographical work.

**Acknowledgements:** This paper was produced using funding from Arts and Humanities Research Council grant “Space and narrative in the Digital Humanities: A research network” (AH/R006482/1). We would like to thank everyone who contributed to the meetings and discussions at that this network funded. The research builds on work done under the Leverhulme Trust funded project “Geospatial Innovations in the Digital Humanities: A Deep Map of the English Lake District.” (RPG-2015-230) This analysis makes extensive use of AntConc software (https://www.laurenceanthony.net/software/antconc)

**References:**


Adolphs A. (2006) *Introducing Electronic Text Analysis*. London: Routledge. Chap. 3

Anon (1852) *Keswick and its Neighbourhood: A hand-book for the use of visitors*. J. Garnett: Windermere.

Baker P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum. Chap. 6

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" *Journal of Historical Geography*, 56, pp. 43-60

Gregory I., Atkinson P., Hardie A., Joulain-Jay A., Kershaw D., Porter C., Rayson P. and Rupp C.J. (2016) "From digital resources to historical scholarship with the British Library 19th Century Newspaper Collection" *Journal of Siberian Federal University: Humanities and Social Sciences*, 9, pp. 994-1006.

Gregory I.N. and Ell P.S. (2007) *Historical GIS: Technologies, methodologies and scholarship.* Cambridge University Press: Cambridge. Chap. 2

Gregory I., Tessier A., Urbánek V., and Whelan R. with Grover C., Martins B., Moreau Y., Murrieta-Flores P., and Porter C. (2019) "Geographies of the Republic of Letters" in Hotson H. and Wallnig T. (eds.) *Reassembling the Republic of Letters: Systems, Standards, Scholarship.* Göttingen University Press: Göttingen

Grover C., Tobin R., Byrne K., Woollard M., Reid J., Dunn S. and Ball J. (2010) "Use of the Edinburgh geoparser for georeferencing digitized historical collections" *Philosophical Transactions of the Royal Society A,* 368, pp. 3875-3889.

Hunston S. (2002) *Corpora in Applied Linguistics.* Cambridge University Press: Cambridge. Chap. 4

McEnery A.M. and Hardie A. (2011) *Corpus Linguistics: Method, theory and practice*. Cambridge University Press: Cambridge. Chap. 6

Lloyd C.D. (2011) *Local Models for Spatial Analysis* (Second Edition). CRC Press: Boca Raton, FL. Chap. 8.

Murrieta-Flores P., Baron A., Gregory I., Hardie A. and Rayson P. (2015) "Automatically analysing large texts in a GIS environment: The Registrar General's reports and cholera in the nineteenth century" *Transactions in GIS*, 19, pp. 296-320

Paterson L.L. and Gregory I.N. (2018) *Representations of Poverty and Place: Using Geographical Text Analysis to understand discourse.* Palgrave MacMillan: Cham, Switzerland. Chap 4.

Porter C., Atkinson P. and Gregory I. (2015) "Geographical Text Analysis: A new key to nineteenth-century mortality" *Health and Place*, 36, pp. 25-34

Robinson J. (1819) *A Guide to the Lakes, in Cumberland, Westmorland, and Lancashire.* Lackington & Co.: London

Rupp C.J., Rayson P., Gregory I., Hardie A., Joulain A., and Hartmann D. (2014) “Dealing with heterogeneous big data when geoparsing historical corpora” *Proceedings of the 2014 IEEE Conference on Big Data.* pp. 80-83

Taylor J.E., Gregory I.N. and Donaldson C. (2018) “Moving beyond close and distant reading: A multiscalar analysis of the English Lake District’s historical soundscape” *International Journal of Humanities and Arts Computing*, 12, pp. 163-182

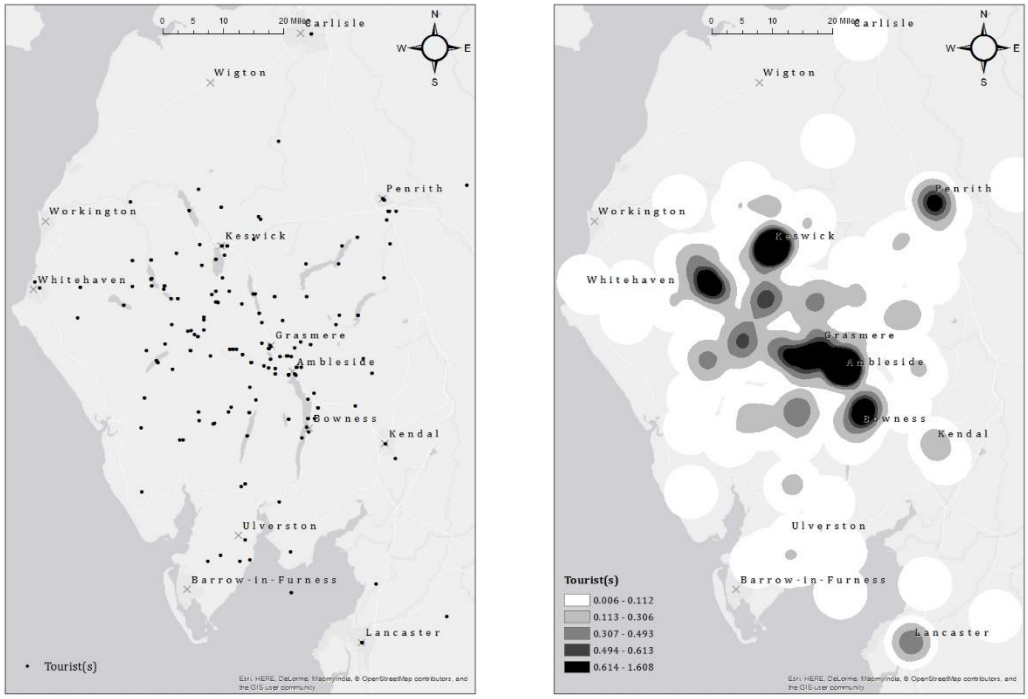


**Figure 1. The geography of tourists in the Lake District using (a) point symbols, and (b) density smoothing.** The maps use PNCs which show the locations of place names that occur within 10 words of the search terms “tourist” and “tourists.”

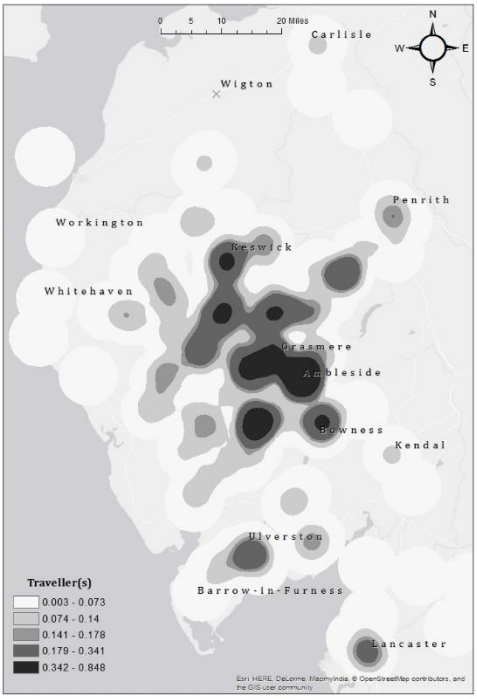


**Figure 2. The geography of travelling in the Lake District using density smoothing.** The maps use PNCs which show the locations of place names that occur within 10 words of the search terms "traveller" and "travellers."

| | Instances | PNCs | PNCs per 100 instances |
|---|---|---|---|
| Tourist[s] | 405 | 332 | 82 |
| Traveller[s] | 608 | 85 | 14 |

**Table 1: Numbers of instances and PNCs for “tourist[s]” and “traveller[s]”**

| Word type | Tourist[s] | Traveller[s] |
|---|---|---|
| Toponym | 11 | 2 |
| Noun referring to physical feature | 8 | 3 |
| Other geographical noun | 4 | 3 |
| Geographical verb | 9 | 7 |
| Preposition | 1 | 1 |
| Adjective | 1 | 7 |
| Adverb | 0 | 2 |
| Person's name | 0 | 1 |
| Other noun | 9 | 12 |
| Other verb | 6 | 10 |
| Pronoun | 1 | 2 |

**Table 2: Numbers of different words by type associated with "tourist[s]" and "traveller[s]" when compared to the corpus as a whole.** Associated refers to within ten word tokens of the search term. Comparison with the corpus done using keyness analysis that uses log-likelihood to identify statistically significant words. Only the top 50 key words are used for each search term.

| Word type | Tourist[s] | Traveller[s] |
|---|---|---|
| Toponym | Keswick, Ambleside, Buttermere, Langdale[s], Coniston, Hawes (Water), Waterhead, Royal (Hotel), Lancaster, Ara (Force), London. | Europe, Wythburn |
| Noun referring to physical feature | Quarters, inn[s], accommodation, road, lakes, bridge, resort, district | Inn, accommodation, road |
| Other geographical noun | Directions, route, distances, views | Spot, foot, lakewards |
| Geographical verb | Proceed[s], return[ing], pursue, find, attracts, alight, accommodated, leaving, explore | Explore[d], watched, visit[ing], comes, reaches, conducted, approach |
| Preposition | From | For |
| Adjective | Useful | Curious, sylvan, fastidious, changeful, antediluvian, superficial, worth |
| Adverb | - | Indisputably, leisurely |
| Person's name | - | Martineau |
| Other noun | Information, attention, tickets, leisure, artist, majority, guide, map, notice | Attention, notice, proof, beauties, curiosity, guides, fellow, objects, packets, pleasure, sentimentalist, tourists |
| Other verb | May, will, should, frequented, witnessing, hire | Will, may, should, be, wishes, recommend, prepares, has, must, gratified |
| Pronoun | The | Who, the |

**Table 3: Keywords by type associated with "tourist[s]" and "traveller[s]" when compared to the corpus as a whole.** Only the top 50 key words are used for each search term. Words in () are completions of two word place names.

| Word type | Mountain[s] | Road[s] |
|---|---|---|
| Toponym | - | Keswick, Ambleside, Penrith, Kendal |
| Noun referring to physical feature | Lake, rocks, vale | Bridge, lake |
| Other geographical noun | Side[s], head | Mile[s], side, left, right |
| Geographical verb | - | Leads, passes |
| Preposition | of, in, to, from, by, on, with, at, as, over, upon, into, down, up, for among | of, from, on, by, in, along, through, at, for, over, as, into, up, about, near, between |
| Adjective | High, other, lofty | High, little, good, great |
| Adverb | Where, not | Where |
| Person's name | - | - |
| Other noun | - | - |
| Other verb | Is, are, was, be, called, seen | Is, are, be |
| Pronoun | It, we, its their, I | It, we, our |

**Table 4: Collocates by type associated with "mountain[s]" and "road[s]" when compared to the corpus as a whole.** Only the top 50 collocates are used for each search term. Collocates identified using a span of +/-5 word tokens and using t-scores to assess significance.