Visual GISting: bringing together corpus linguistics and Geographical Information Systems

Ian N. Gregory
Department of History, Lancaster University, Lancaster, LA1 4YG, UK

Andrew Hardie
Department of Linguistics and English Language, Lancaster University, Lancaster, LA1 4YL, UK

Abstract

Corpus linguistics and Geographical Information Systems (GIS) are approaches exploiting computer-based methodologies in the study of, respectively, language and language usage, and spatial patterns in geographical databases. We present an approach that uses corpus methods to bridge the gap between the textual content of a corpus (and, thus, the typically textual concerns of many branches of the humanities) and the geo-referenced database at the heart of a GIS. Using part-of-speech tagging to extract instances of proper nouns from a corpus, and a gazetteer to limit these instances to those representing place–names, a database of the places mentioned in a corpus can be created, visualized, and analysed using GIS technology. It is also possible to visualize the *meanings* associated with particular place–names, by building GIS databases on the collocation of place–names with particular semantic categories in their immediate context. In this way, we can create maps that visualize the geographical distribution of mentions of concepts such as *war*, *government*, or *money* in a particular data set. The approach cannot be entirely automated and some manual intervention is required. Nevertheless, the method is clearly valuable for the interpretation of spatial phenomena in text corpora.

Correspondence:
Ian N. Gregory,
Department of History,
Lancaster University,
Lancaster, LA1 4YG, UK.
E-mail:
i.gregory@lancaster.ac.uk

1 Introduction

This article is concerned with bringing together two fields of research, corpus linguistics and Geographical Information Systems (GIS), to allow researchers to explore geographical patterns within large textual data sets. Corpus linguistics is concerned with using computer techniques to analyse large bodies of text known as *corpora*. The fundamental question asked is 'what is the nature of the usage of language within this corpus?', where the aspects of language to be investigated may encompass grammar, semantics, pragmatics, discourse structure, and any of a range of issues related to textual content (for overviews see Biber *et al.*, 1998; Adolphs, 2006; McEnery *et al.*, 2006; McEnery and Hardie, 2011). In contrast, the fundamental question asked of a GIS is 'what are the

Literary and Linguistic Computing, Vol. 26, No. 3, 2011.

doi:10.1093/llc/fqr022 Advance Access published on 20 May 2011

geographical patterns within this data set?' (Longley *et al.*, 2001; Schuurman, 2004). In a GIS database, each item of data is linked to a co-ordinate-based location and this is fundamental to the retrieval, analysis, and visualization of the data that it contains. Both corpus linguistics and GIS have proven highly successful approaches; however, the two have traditionally had very little to do with each other—corpus linguistics has largely ignored any spatial dimension within the texts that it studies, while GIS has largely ignored qualitative data, particularly textual data. Combining the two approaches in the investigation of the spatial patterns found in large bodies of text represents a methodological advance for both fields, adding the geographical to corpus analysis and the textual to GIS.

This article outlines how such a combined methodology can be implemented, and provides some examples of the potential that it has to shed new insights into the geographies inherent within a textual data set. In Section 2, we give brief background overviews of corpus linguistics and GIS. Section 3 provides a general account of the methodology we introduce and of the data set we use to exemplify this approach, the Lancaster Newsbooks Corpus. In Sections 4 and 5, we present case studies of two sets of analyses that operationalize the general approach.

2 Background: Corpus Linguistics and GIS

2.1 Corpus methodology: in linguistics, in the humanities, in geography

Although, there were earlier precursors (see McEnery and Wilson, 2001, p. 2–4), corpus linguistics may fairly be said to have emerged in the late 1950s and 1960s, with such pioneering work as the construction of the original Survey of English Usage,[1] the development of the Brown Corpus of American English (Francis and Kučera, 1964), and John Sinclair's early research on collocation (Sinclair *et al.*, 2004). Although initially a specialized sub-field, since the 1990s corpus linguistics has increasingly been adopted as a key component of the methodological toolbox by linguists of many stripes. The basis of all corpus methods is the computer-aided analysis of very large bodies of text. The most basic analyses that the computer can perform are: (i) frequency counts of items such as words or word sequences in the corpus and (ii) searches for all instances of a particular linguistic form, usually a word or phrase, in the corpus—the results of such searches typically being displayed as a concordance of each instance in context. However, much research in corpus linguistics has been devoted to the development of more complex methodologies. While corpus-based approaches often utilize statistical analysis—most notably, in dealing with corpus frequencies—it is nonetheless a critical aspect of corpus linguistics that it relies on both quantitative *and* qualitative techniques (Biber *et al.*, 1998, p. 4). For instance, while the computer can produce a concordance of a word, drawing meaningful conclusions about the word requires a careful qualitative analysis of the contents of that concordance.

Corpus-based approaches have many applications within linguistics. One of the most notable is lexicography (see Ooi, 1998), in which field corpus techniques are now all but indispensible. But corpora have also been exploited for the study of text-type variation (Biber, 1988) and dialect variation (Sampson, 2002); in the investigation of metaphor (Deignan, 2005); in literary stylistics (Culpeper 2002; Semino and Short, 2004; Mahlberg, 2007); and in Critical Discourse Analysis (Baker, 2006; Baker *et al.*, 2008). There presently exist two major schools of thought on the status of corpus linguistics as a discipline (Hardie and McEnery, 2010). From one perspective, corpus linguistics is seen as, in principle, an independent sub-field or theory of language (Tognini-Bonelli, 2001, p. 1; Teubert, 2005, p. 2); the other perspective sees corpus linguistics primarily as a methodology capable of being applied within a wide range of theoretical and analytical frameworks. From the latter perspective, it is natural to consider how corpus methods may be exploited in disciplines beyond linguistics.

Many branches of the humanities focus on textual evidence; whenever such evidence is considered on the large scale, corpus methods may be of use. However, to use corpus techniques in the field of

geography, a minimum requirement is that there be some link between the capabilities of corpus methods, and the centrally spatial concerns of geography. This link would seem to be most easily made in the context of GIS—an approach to geographical research which, like corpus linguistics, has exploited computer resources to produce new analytic methods.

2.2 GIS: in physical geography, in historical geography, in the humanities

Although GIS is often thought of as a mapping technology, it is more accurately seen as a form of database. A conventional relational database structures data in tabular form, where each data object is a row in the table, and the properties of the object are the columns. In the specialized database that is a GIS, data in this form are termed *attribute data* and each row of attribute data is also linked to a geographical reference termed the *spatial data*. In a vector GIS, spatial data can take the form of a point, a line, or a polygon (representing an area or zone); in a raster GIS the spatial data takes the form of a fine grid of pixels. The combination of attribute and spatial data allows the data set to be structured, queried, visualized, and analysed in ways that stress its geographical characteristics (Martin, 1996; Clarke, 1997; Chrisman, 2002). Visualizing data through mapping is clearly an important part of this type of analysis but the use of GIS goes far beyond this; location is inherent in all aspects of managing and researching the data within a GIS.

GIS originated in the 1960s, partly in the military and partly in the Earth sciences, and spread rapidly into academic disciplines such as the environmental sciences and physical geography (Coppock and Rhind, 1991). Take-up in human geography was slower because of debates over the quantitative and positivist nature of GIS (see Johnston, 1999; Pickles, 1999). Nevertheless, since the advent of desktop GIS software in the mid-1990s, GIS has become a standard tool within human geography. More recently, use of GIS has increasingly spread into various humanities disciplines. Significant progress has been made in history, to the extent that *Historical GIS* is now a recognized field (Gregory and Healey, 2007; Knowles, 2008). However, most progress in Historical GIS has been made in areas where *quantitative* source data is available and its use accepted, including environmental history (Cunfer, 2005), historical demography (Gregory, 2008), health and poverty (Dorling *et al.*, 2000; Gregory, 2009), medieval land-use (Campbell and Bartley, 2006), economic history (Knowles and Healey, 2006), and urban studies (Gordon, 2008). Thus the majority of the most advanced projects within Historical GIS are based primarily on numeric (statistical) and cartographic sources. This presents a major obstacle to the adoption of GIS in the humanities more generally, as text (rather than tables of statistics) is clearly the most widely used form of data across the humanities (Jessop, 2008; Bodenhamer *et al.*, 2010). While progress to date within Historical GIS is encouraging, if GIS is to become a widespread tool across the humanities, it must be able to incorporate and embrace text. Some progress has been made in developing techniques for doing this to allow geographical searching of certain corpora (Grover *et al.*, 2010); however we argue that the potential for using texts within GIS goes far beyond this: it must be possible to directly incorporate a text into a GIS and the GIS must then be demonstrably able to improve our understanding of a text or body of texts.

This article will show that this can be accomplished by using certain corpus linguistic techniques to mediate between the original, textual data and the tabular, spatially referenced form required by a GIS. The linguistic foundation of corpus-based methodologies means that, critically, GIS databases can be produced in which the *meaning* of the original text is reflected in the geographical outputs. Thus, qualitative content is preserved through the automatic processes that produce a GIS from a text corpus. In the remainder of the paper, we illustrate two ways in which this general procedure can be concretely operationalized. In Section 4, we outline a basic technique for mapping mentions of place–names found in a corpus within a GIS. In Section 5, we outline a more sophisticated methodology which exploits semantic tagging to produce GIS databases which visualize topics, domains, or themes from within a corpus.

3 Data and Methodology

3.1 A corpus-based approach to the generation of a GIS

Any data set explicitly or implicitly contains information about space and time. That is, as well as information that tells us *what* the data set is concerned with; there is in addition information about *space* telling us *where* the data set is concerned with; and information about *time* telling us *when* it is concerned with (Peuquet, 1994). The question of *what* information is within the data set is the traditional concern of corpus linguistics (in the broadest sense of 'information within', referring not only to the 'message' communicated but also the linguistic system that carries the message). However, several linguistic phenomena that may occur in a text are implicitly or explicitly concerned with space. These include, for instance, deixis (linguistic forms indicating distance from or proximity to a speaker or hearer), and prepositions and other markers of position or direction—and also, perhaps most straightforwardly, place–names, that is, proper nouns whose referent is a location. A simple geographical question that can then be asked of the corpus is then 'what places is the text discussing?', a more complicated question is 'what is the text saying about these places?'. If the text or its metadata also contains temporal information—for example, in a corpus of newspapers from different dates, diaries, or travelogues—then an additional question can be asked, namely '. . . and how does this change over time?'.[2]

To explore these types of questions effectively, it is necessary to develop a technique for generating a GIS on the basis of the place–names within a corpus. We define each individual token of a place–name in the corpus under consideration as a mention. To preserve both the quantitative and qualitative makeup of the corpus, each mention must map to a single data object in the GIS. Accomplishing this conversion requires four challenges to be overcome. First, all instances of place–names within the text must be reliably identified. Second, every occurrence of each relevant mention of a place–name must be *geo-referenced*, that is, associated with a co-ordinate that represents its location.[3] If a place–name is mentioned more than once, the geo-referenced data will include multiple (identical) data points to allow us to model these multiple mentions. Third, the co-ordinates must be converted into a GIS file format such as a Shapefile for ArcGIS,[4] or a Keyhole Markup Language (KML) file for Google Earth[5] or other virtual globes, and finally once this has been visualized, subsequent analysis needs to take place to explore and summarize the spatial patterns within the corpus. Thus, like corpus linguistics itself, the method we propose is inherently both qualitative and quantitative: frequency of mentions plays a role in the automatic extraction of data for the GIS, but the resulting GIS visualizations are analysed qualitatively and rely on the analyst's understanding of the socio-historical context from which the original textual data derived.

This methodology provides enough information to enable the researcher to explore *where* the corpus is talking about. In Section 4, we present an operationalization of this technique and consider what it can tell us about our example corpus, at a relatively broad-stroke level of description. Adding detailed information about *what* the text is saying requires an additional step of automated processing based on the application of semantic tagging. In corpus analysis, *tagging* (or *annotation*) is a well-established technique of labelling of all elements at some linguistic level within a corpus according to their category within a defined scheme of analysis (see Leech, 1997). In the particular case of semantic tagging, every word-token is labelled with a tag indicating its position within some ontology, that is, a schema which categorizes its meaning. The tagger and analytic schema used here is USAS, the UCREL Semantic Analysis System[6] (see Rayson, 2008). Our interest in the present context is in what semantic tags co-occur or *collocate* with particular place–names (see Sinclair, 1991, p. 109–22 for an overview of collocation)—or, to look at it another way, what place–names collocate with the concepts indicated by particular semantic tags (as we will illustrate in Section 5). The underlying assumption of this analysis is that if a word with a particular semantic tag occurs in proximity to a place–name, then the concept indicated by that tag is relevant to that place. This assumption will clearly

not be true in every case. However, since this analysis is conducted across the entirety of a large corpus, spurious instances will be outweighed by genuine examples of meaningful relationships that reflect *what* is being said about places in the corpus.

3.2 Example data set: the Lancaster newsbooks corpus

As a testbed data set, we utilize a subsection of the Lancaster Newsbooks Corpus.[7] This is a collection of all of the surviving news periodicals printed in London between mid-December 1653 and the end of May 1654 (Hardie and McEnery, 2009; Prentice and Hardie, 2009). The subsection of the corpus that we use consists of approximately 870,000 words of material that is highly relevant to the social and political history of a turbulent period. The Protectorate of Oliver Cromwell was instituted in December 1653; the corpus also covers the latter part of the Glencairn Uprising, a Royalist rebellion against Cromwell's rule in Scotland, named after its leader. Also during these 5 months, a peace treaty between England and Holland was being concluded, and Queen Christina of Sweden abdicated and converted to Catholicism. Though this data set is relatively small by the standards of corpus linguistics, it is large enough to make a hand-and-eye analysis prohibitively time-consuming. Moreover, the geographical spread of the events known to be of historical interest within this text collection makes it an excellent testbed data set. It should, however, be noted that these techniques are equally applicable to modern sources as they are to historic ones.

4 From a Corpus to a GIS: Mapping Mentions

The first stage in converting a corpus to a GIS is to identify and extract the place–names from within the text. The theory of doing this is relatively straightforward; place–names are, by definition, proper nouns and technologies to extract proper nouns are well established. In particular, part-of-speech (POS) tagging, another form of corpus annotation, can be undertaken automatically and will typically distinguish proper nouns from all other word classes, including common nouns, to a high degree of accuracy. We tagged the whole corpus using the CLAWS[8] software, and then extracted every instance of words tagged as proper nouns.[9] This list of proper-noun tokens is taken to include all potential place–names within the corpus.

The next stage of the procedure is to identify which proper nouns are actually place–names. This involves using a place–name gazetteer to allocate co-ordinates to each of the words found within the gazetteer, while at the same time removing from the list those proper nouns not found in the gazetteer and thus deemed not to be place–names. A place–name gazetteer, sometimes also known as a thesaurus, is basically a database table linking place–names to co-ordinates. The co-ordinates are typically in the form of latitude and longitude for international material, or Eastings and Northings for British or Irish places. Gazetteers often also hold additional information such as variant spellings; position within an administrative hierarchy, such as which county or state the place lies within; what type of feature the name refers to, such as town, river, mountain, and so on; and perhaps some statistics such as population or area. A number of broad-coverage place–name gazetteers are currently available, including the Ordnance Survey's (OS) 1:50,000 gazetteer, which provides Eastings and Northings for every place–name on the OS's Landranger maps of Britain,[10] and the GeoNames[11] and World-Gazetteer[12] websites which provide freely available and downloadable gazetteers for the whole world, giving latitudes and longitudes for each location. In this case World-Gazetteer was used; however, the choice of which gazetteer to use with a particular corpus requires careful consideration and a different gazetteer might well be more appropriate for other studies or applications.

The first stage in using a gazetteer to geo-reference a list of proper nouns is to filter both the list of proper nouns and the gazetteer to remove unnecessary information. The list of proper nouns can be filtered to remove obvious non-place–names, such as words preceded by *Mr*, *Ms*, *Duke of* and so on. The gazetteer can also be simplified to remove areas known in advance to be irrelevant to the corpus at

hand. For the 1654 newsbooks, we removed all references to places outside Europe and the Mediterranean as there were few, if any, mentions of places further afield. This prevents a token such as *Lancaster*, for instance, from generating a mention linked to Lancaster, California or Lancaster, Ontario. A relational join between the list of proper nouns and the gazetteer then gives a first pass at geo-referencing the data set, in the form of a table or view of the data that gives for each instance, at minimum, the place–name and the co-ordinates of its location. Crucially, the data objects in this table consist of *mentions*—specific instances of a place–name occurring at a particular point in the corpus—and thus the same place–name and associated co-ordinates can (and in fact do) occur many times in the table.

This is the most difficult stage of the operation, and one in which some user interaction in inevitable. The precise amount of manual intervention that is required will vary according to the number of place–names within the text and the level of accuracy required. There are many potential problems distinguishing place–names from other proper nouns. One problem is that many proper-noun word-forms can be names of people as well as names of places. For instance, relative to *Lancaster* as the name of a city, names such as *Roy Lancaster*, the *Duke of Lancaster*, and the *Lancaster Bomber* are potential problems. Another problem is the disambiguation of place–names that can refer to more than one place such as *Newcastle*, *Newton*, or *Richmond*. Our first-pass approach was to produce *multiple* place–name mentions from each instance of such an ambiguous word-form, that is, to generate one mention for each possible location, but then to allow subsequent manual disambiguation and removal of the spurious cases. An automated approach to disambiguation based on the type of place, its size, or its proximity to other places may be possible but is likely to be error prone. A third problem is that, even with modern material, the same place–name can be spelt in a number of ways, such as *Newcastle-on-Tyne* versus *Newcastle upon Tyne*, or *Saint Helen's* versus *St. Helens*. With historical material, where spelling variation is the rule rather than the exception (see Pilz *et al.*, 2008), or material where standards of data capture may be poor, this problem may be exacerbated.[13]

The table of mentions which results from the procedure outlined above can easily be converted into a GIS format. In a GIS the basic unit of storage combining attribute and spatial data is termed the *layer* (or sometimes *coverage* or *theme*). In this case, for each mention, the place–name is its attribute data and the co-ordinates are its spatial data. Most GIS software packages can convert tables with co-ordinate data into layers of point data; an alternative approach is to convert the table into KML format for use in Google Earth or other virtual globes.

Figure 1 shows a GIS produced from the first pass of extracting place–names from the Lancaster Newsbooks Corpus. No initial attempts to clean this data set have been made; a total of 8,430 provisional place–names were mapped and every mention is represented as a single point on the GIS. Although the spread of points is at first glance interesting, there are a number of features, such as the surprizingly large number of mentions of Eastern European places, which suggest errors or noise within the data set and alert us to the need to check these further.

It is well known in the cartographic literature that dot maps of this sort are difficult to interpret when they show a large amount of overlapping data (Robinson *et al.*, 1995). To simplify the pattern and make it more comprehensible a technique called *density smoothing* can be used. This was originally pioneered in disciplines such as epidemiology and criminology. The idea behind it is that, rather than measure discrete events at precisely defined locations, the varying intensity of events near each location is measured. Rather than using point data in a vector system, a raster system is used, in which the density of events near each pixel is measured, with nearer events having more influence than those further away (see for example, Bailey and Gatrell, 1995 or Lloyd, 2007). This is relatively simple to implement with GIS software. The result of applying the technique to the provisional data is shown in Fig. 2, where a clearer pattern emerges. The majority of mentions are concentrated across England and central Scotland. Beyond this, there are some distinct clusters, particularly around Paris, Bordeaux, what

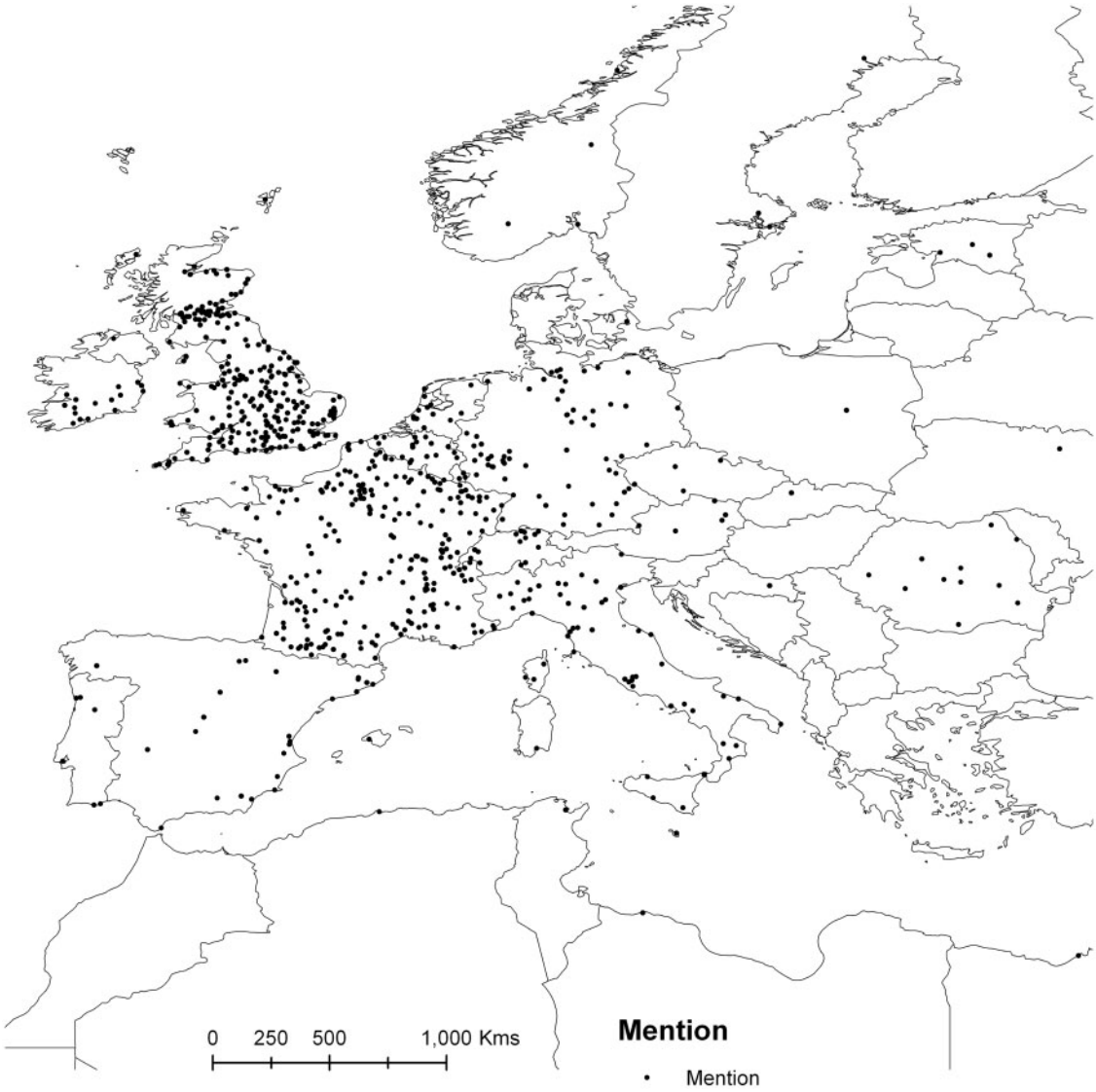

Fig. 1 A first pass of mapping the Lancaster Newsbooks Corpus using a dot map

is today the Netherlands, and Hamburg, and less dense clusters near Stockholm, Rome, and Naples.

That most news reported in the London press concerns England is not particularly surprizing. Equally, many locations where mentions cluster outside the British Isles can also be identified and are usually explicable in terms of the presentation of news in seventeenth century journalism. News from a particular correspondent was typically headed by a statement of where they wrote from and when their letter was dated, for instance 'From Hamburgh December 20' (in *Mercurius Politicus* issue 186). News reported from a given location might well originate elsewhere; in the issue cited, the news '[f]rom Hamburgh' included tidings from the Hague, Stockholm, and the Holy Roman Emperor's court. So while places discussed in the news attract mentions, so too do places that are

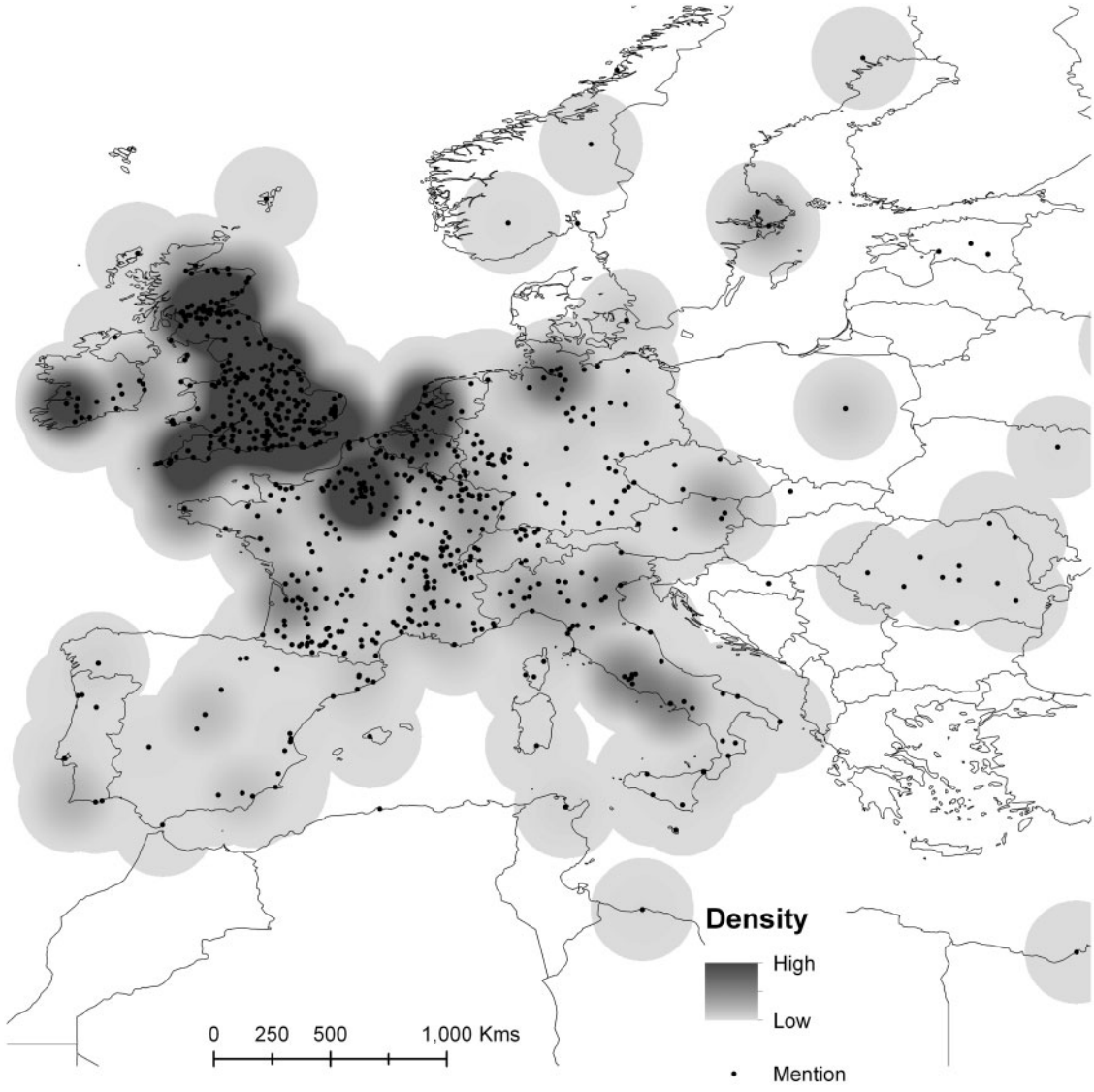

Fig. 2 A density smoothed map of provisional place–names mentioned in the Lancaster Newsbooks Corpus

conduits for news—such as Paris, Hamburg, and Rotterdam. On the other hand, as is often the case with mapping, absence can be as interesting as presence. Notably absent from the pattern in Fig. 2 are any major clusters in places including Spain, Portugal, Africa, and Greece. Tentatively, this may illustrate how marginal (most of) the Mediterranean is to the concerns of English consumers of news at this period, compared to the English Channel and the North Sea. The key point, however, is that we believe that this is the first time that techniques to demonstrate and quantify the geographical distributions within texts have been developed.

Figures 1 and 2 represent the data before any effort was made to filter out noise. But exploring maps of this sort is in fact an effective means for spotting errors—by querying clusters and distinct outliers, we can discover the reasons for consistent

distortions. For instance, there is a cluster of mentions near Manchester caused by the word *Middleton*. It turns out that this is almost always *General Middleton* or *Gen. Middleton*, who was a military leader in the Glencairn Uprising. This indicates the need for filtering of the list of potential place–names to remove words preceded by titles that indicate a name of a human being. Another type of error is where names are linked to the wrong place. An example of this is the cluster on the west coast of Ireland, which overwhelmingly consists of mentions of *Newcastle*. Checking the original text strongly suggests that the reference is in fact to *Newcastle upon Tyne* in the north–east of England so the database needs to be updated accordingly, making these mentions match only Newcastle upon Tyne and not Newcastle in Ireland. Likewise, among the outliers in Eastern Europe, five of the fourteen mentions in what is now Romania are to *Victoria* or *Alexandria*. Checking these back with the text shows that these too are people, not places, and they can thus be removed from the list.

Figure 3 shows the same map of the data once it has been cleaned. This reduces the number of provisional place–names from 8,430 to 6,297. The summary pattern shown by Fig. 3 is not, however, very different from the uncleaned pattern shown in Fig. 2. The most noticeable difference is the loss of the cluster around the Irish *Newcastle*. There has also been some downgrading of the clusters along the eastern border of France and the loss of some outliers in eastern Europe. It is probably fair to say that the cleaning renders the clusters that remain, in England, Paris, and the Dutch coast, more pronounced rather than less.

Figure 4 maps the mentions of place–names within Britain and Ireland. London is far and away the most mentioned place; otherwise there are prominent clusters in Edinburgh, Newcastle and Plymouth, and lesser clusters in Cornwall (primarily Falmouth), in Dover, along the Severn Estuary (Bristol and Gloucester), around the Glasgow area, and in Aberdeen and Inverness. The west coast of Scotland and much of Ireland, especially what is now Northern Ireland, are noticeable for their relative lack of mentions. The two clusters in central Scotland illustrate another advantage of this form of mapping: there are 173 mentions of Edinburgh, making it the third most-named place in Britain and Ireland; however, nearby Dalkeith provides an additional ninety-six mentions, greatly enhancing this cluster. The Glasgow cluster actually stretches from Falkirk in the east to Ayr in the south–west, taking in a total of 218 mentions. The place–names mentioned are Stirling (fifty mentions), Glasgow (45), Ayr (33), Renfrew (25), Dumbarton (23), and fourteen other places. If each location were considered in isolation, the newsbooks' interest in this part of the country would not be clear; however, once the data is visualized, its relevance becomes obvious.

5 From a Corpus to a GIS: Mapping Meaning

The previous section illustrated two key points: (i) it is possible to map a text and discern interesting patterns in the result; (ii) place–names are problematic, and moving from automated extraction of proper nouns to a definitive list of disambiguated place–name mentions requires a significant amount of user intervention, which while time-consuming, is achievable. This simple spatial visualization brings out some of the geographical structure of the corpus' content that is of interest to the researcher; however, a more interesting question is not simply *where* is being talked about but also *what* is being said about these places. It is in this context that the collocation of place–names with particular semantic tags, discussed in Section 3.1 above, becomes relevant.

Using the semantically tagged version of the corpus, we once again extracted all instances of proper nouns as candidate place–names. However, in this case, we extracted: (i) not only the instance of the word itself, but also the five words before and the five words after it and (ii) not only the actual words, but also their semantic tags. A five-word span before and after a word of interest is often used as the 'window' of collocation in computational linguistics (Seretan and Wehrli, 2007, p. 75). Some other researchers limit the study of

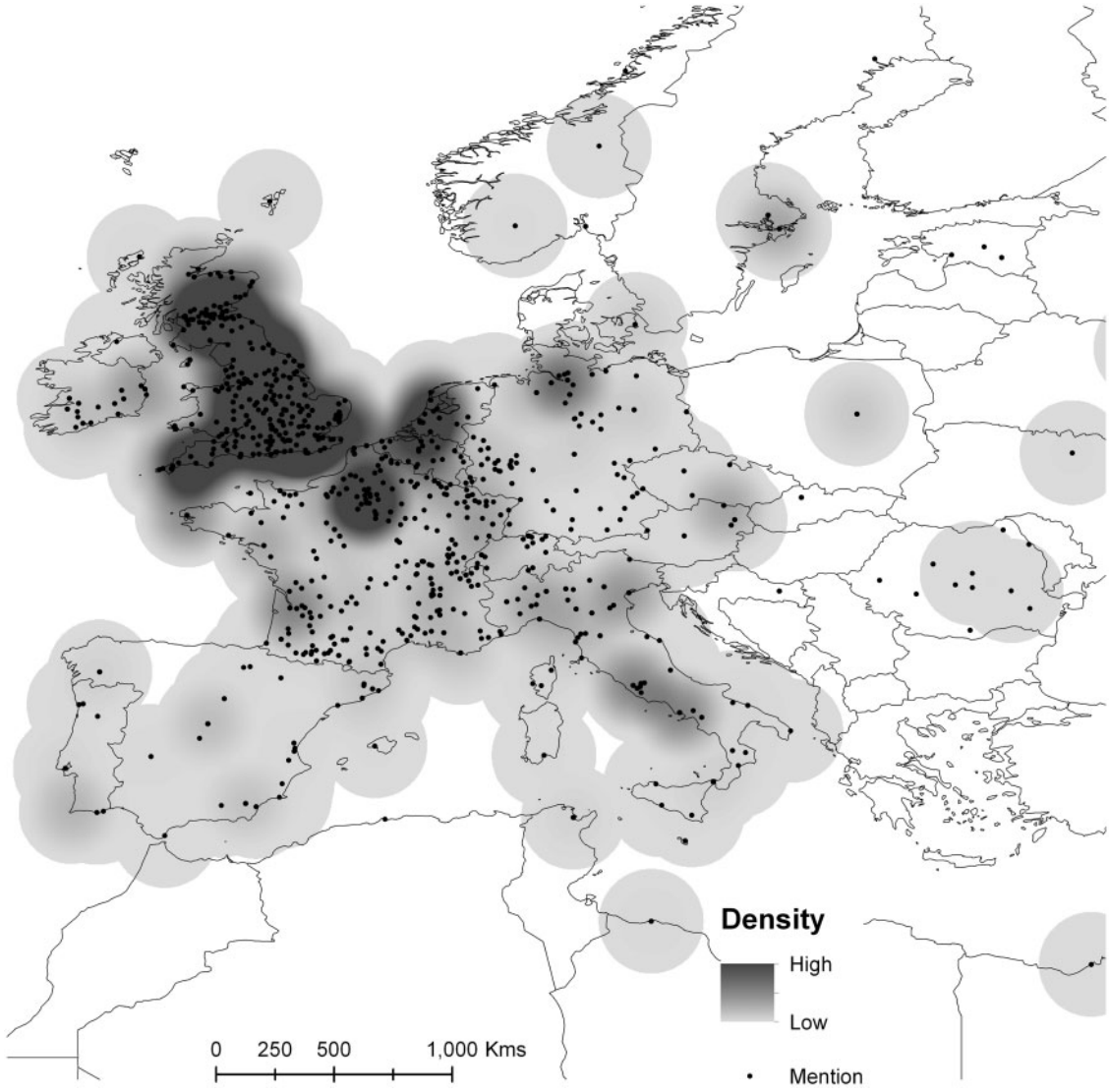

Fig. 3 A density smoothed map of cleaned place–names mentioned in the Lancaster Newsbooks Corpus

collocation to looking at syntactically related tokens (for instance, the object of a verb, the adjective modifying a noun: see Grefenstette, 1992; Evert, 2005). However, in this case our specific interest was in what was said *nearby* to the place–name mentions, not in any particular syntactic relationship, and so proximity-based collocation was a better fit to our research aims. The data was filtered using the gazetteer, and coordinates added, as before. The result was a database where each entry contains a great deal of information about the semantic context of a given instance of a place–name, and which could be queried in two ways: to extract all the semantic tags found in the vicinity of any mention of a given place–name, or to extract all instances of place–names that have a given semantic tag in their immediate neighbourhood. The first type of query can help us get a handle on what is said

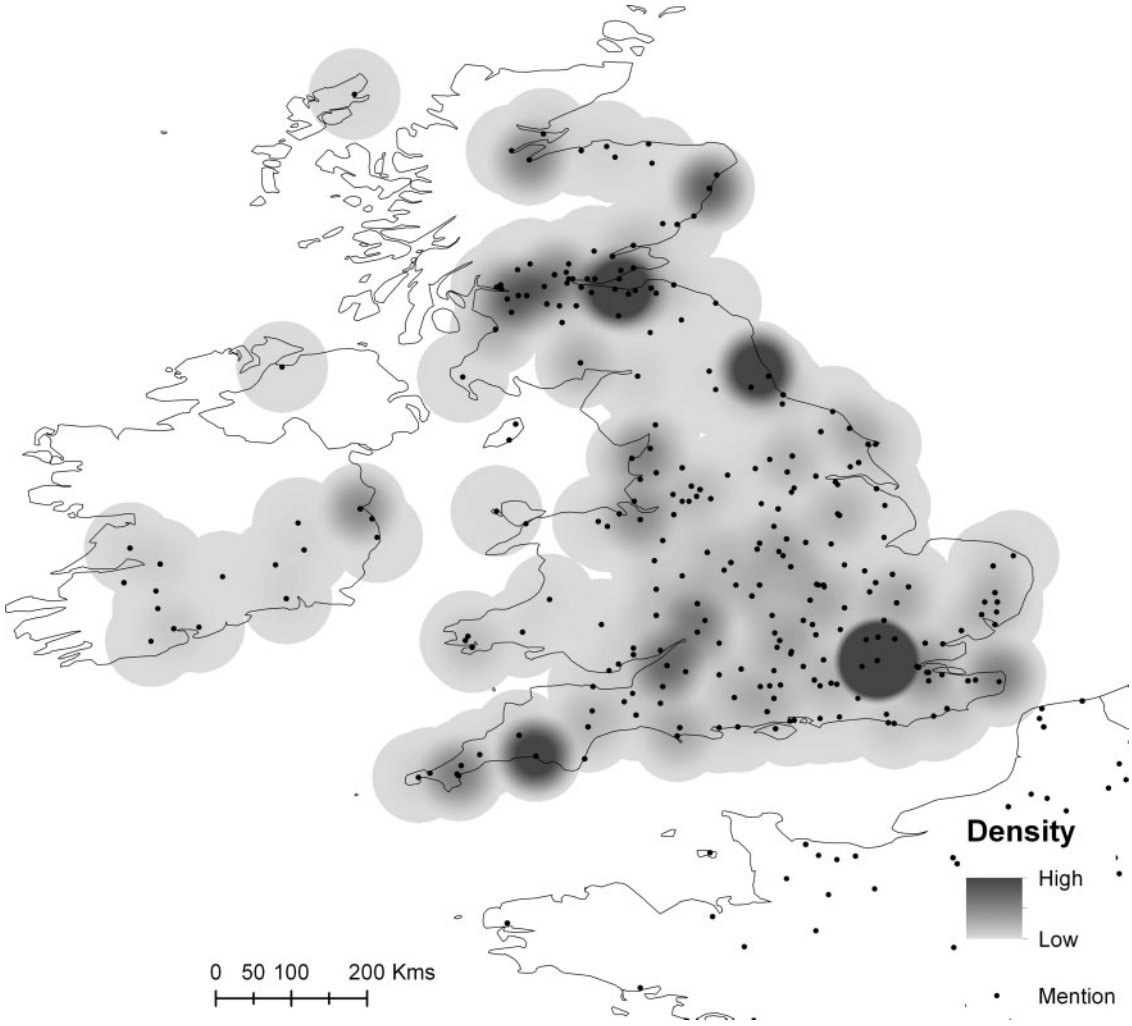

Fig. 4 A density smoothed map of place–names mentioned in Britain and Ireland

about a particular place. Although this is useful, our focus here is on the second type of query, because it generates a subset of geo-referenced mentions found to be associated with a particular semantic tag, and this subset can be mapped.

An example of how this works is shown in Table 1, which shows the entry for a mention of *Dunkirk* from fairly early in the corpus (a newsbook published on 10 January 1654). The ten words of context contains '. . . of a rich Fleet from Dunkirk, consisting of about forty . . .' (The scripts that generate this database treat the comma as a separate word-token.) Many of the tags are not of interest for our purposes, including *Z5* (grammar bin) and *N1* (a number). The tags *A1:8+* and *A13:4*, subcategories of category A, *general and abstract terms*, are potentially more interesting, although interpreting the relationship between concepts such as approximation and inclusion and a geographical location would be a challenge. More concrete, and therefore straightforwardly interpretable, is the association of *Dunkirk* with the word *rich*, tagged *I1:1+*, a subcategory of *I1* which covers all meanings related to money. If we query the database for mentions of

Table 1 Semantically tagged database entry for one instance of the word *Dunkirk*

Preceding context						Following context				
5 Left	4	3	2	1 Left		1 Right	2	3	4	5 Right
of	a	rich	Fleet	from	Dunkirk	,	consisting	of	about	forty
Z5	Z5	I1:1+	Z2	Z5	Z2		A1:8+	Z5	A13:4	N1

The tags are: *Z2* geographical names; *Z5* grammar bin; *I1:1+* money—affluence (positive); *A1:8+* inclusion/exclusion (positive); *A13:4* degree—approximators; *N1* numbers.

place–names associated with this tag for money, this instance will be retrieved, and can be plotted in a GIS. If there are many such instances in the corpus as a whole, the visualized GIS for 'money' will show a cluster at Dunkirk just as, in the overall visualizations discussed above, there are clusters at Paris and Hamburg. If this is the only association of Dunkirk with money, in contrast, it will be outweighed by the mass of other place–name mentions linked to money elsewhere. Given that much naval activity in the seventeenth century was concerned with capturing enemy ships for profit, it does not seem unlikely that this pattern should in fact be repeated.

Table 1 illustrates another important point. The word *Fleet* has been tagged as a geographical name; this is incorrect and it should have been tagged *M4* (for *travel by water*). However, this is a single error in a very large corpus. If the association of Dunkirk with matters naval is consistent, as we might expect, instances of Dunkirk in association with other *M4* words will outweigh the single instance here and produce a suitable cluster in the GIS.

An example of a query for a semantic category being mapped in a GIS is given in Fig. 5. *G3* indicates *war*. There are 360 mentions of place–names with words tagged G3 nearby; these mentions refer to a total of ninety-three distinct places. Figure 5 shows that references to war are particularly concentrated in central and eastern Scotland—a clear reflection of the Royalist uprising being given substantial coverage in the newsbooks. While *Aberdeen* is the most-mentioned place (thirty-four mentions), the cluster encompasses twenty-six separate places–names with a clear spatial distribution across the Lowlands and up the east coast. *Brest* attracts twenty-five mentions in relation to war, second only to Aberdeen. Unlike the Scottish distribution, this is an isolated point. Many of the mentions are of the *Brest men of war*, where Brest is simply mentioned as the home port of ships involved in naval conflict; this explains why only one place, and not an area, is mentioned here. London also attracts many mentions, as does northern France and the Netherlands.

Figure 6 shows a density smoothed map of place–names near a word-token tagged as *G1*, government. London is the most-mentioned place, followed by Rome. The Edinburgh and Newcastle area receives a large numbers of mentions, as does the Rotterdam and Amsterdam area. Bordeaux, Cologne, and Venice also have clusters of mentions. Figure 7 shows the pattern for *I1*, money. It is noticeably different to the pattern for government shown in Fig. 6. While London remains important, it is less dominant than for government. The whole east coast of Britain receives mentions based partly on Edinburgh and Newcastle, places of recurring importance, but also on Scarborough—driven by mentions relating to prize ships—and a cluster of places in East Anglia including Framlingham and Debenham—driven by discussions of work to relieve the poor. Paris, barely mentioned in relation to government, comes out strongly in relation to money, with Amsterdam and Rotterdam providing a lesser cluster. In North Africa, Tunis and Tripoli provide clusters that prove to be spurious. The emphasis on Tunis results from several instances of the report 'they shall call the Turks to an account at Tunis, or Algier, for the wrongs they have done'. The word *account* here is being used metaphorically, with the conceptual link to money indirect and thus, less relevant for our purposes. Polysemy is an issue that the semantic tagger unaided often struggles with. Tripoli, meanwhile, was itself (incorrectly) tagged as related to money, rather than as a geographical term.

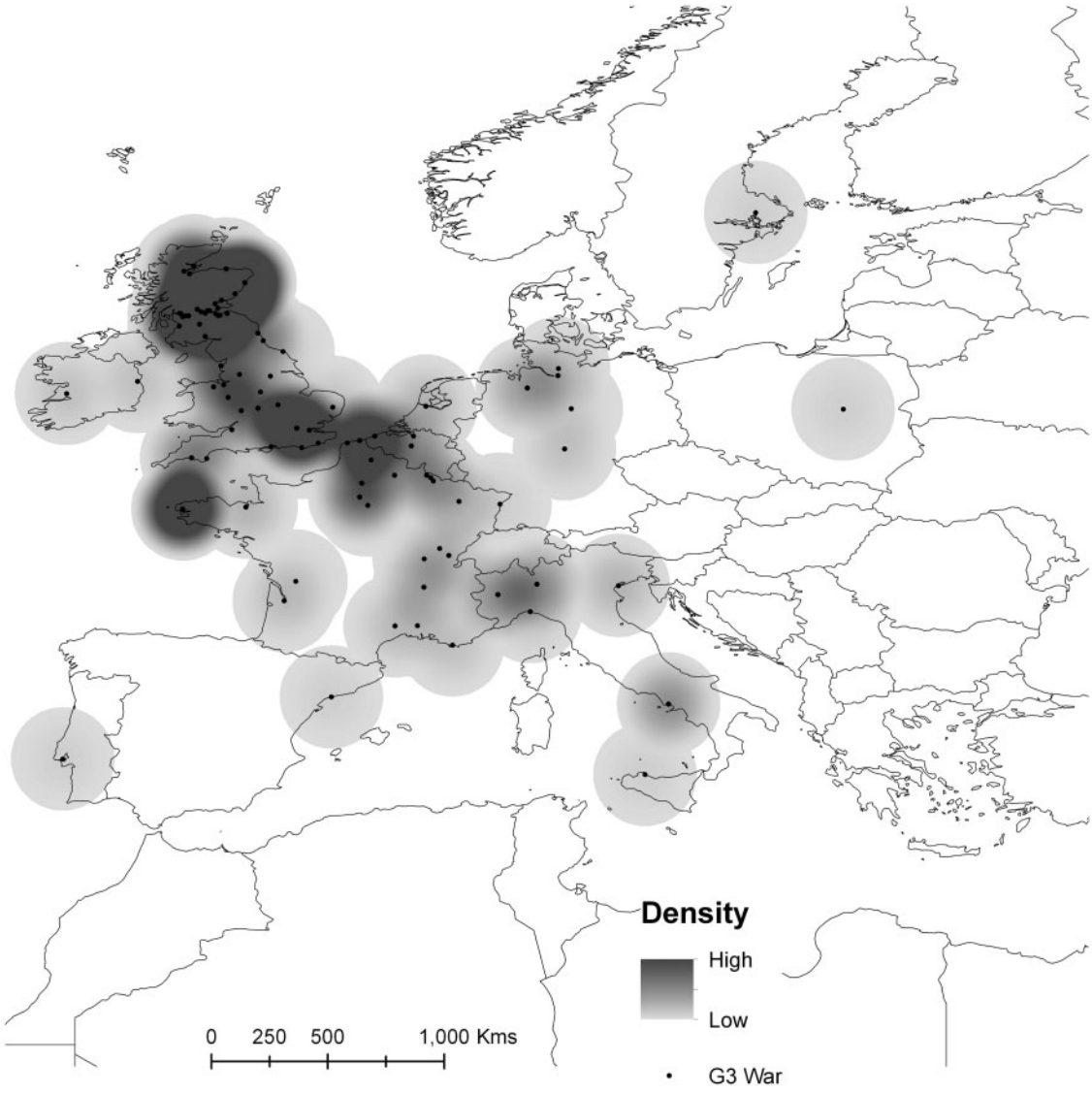

Fig. 5 Mapping war: references to places near a word tagged as *G3*

6 Conclusion

We propose the term *Visual GISting* for the joint corpus-linguistic/GIS method outlined above, since the abstractions created across many instances of semantic tags constitute in some respect the *gist* of the content of the corpus, and the use of GIS techniques allows these content summaries to be rendered visually. It is, we would argue, a technique that enables corpus analysts to think geographically just as it enables GIS specialists to exploit textual data. Our examples here focused on news text and news collection in a particular historical period; however, we would argue, there is no reason to think that it could not be generally applicable to a wide range of research questions in geography and in the humanities. The approach could be applied to any study that wants to ask geographical

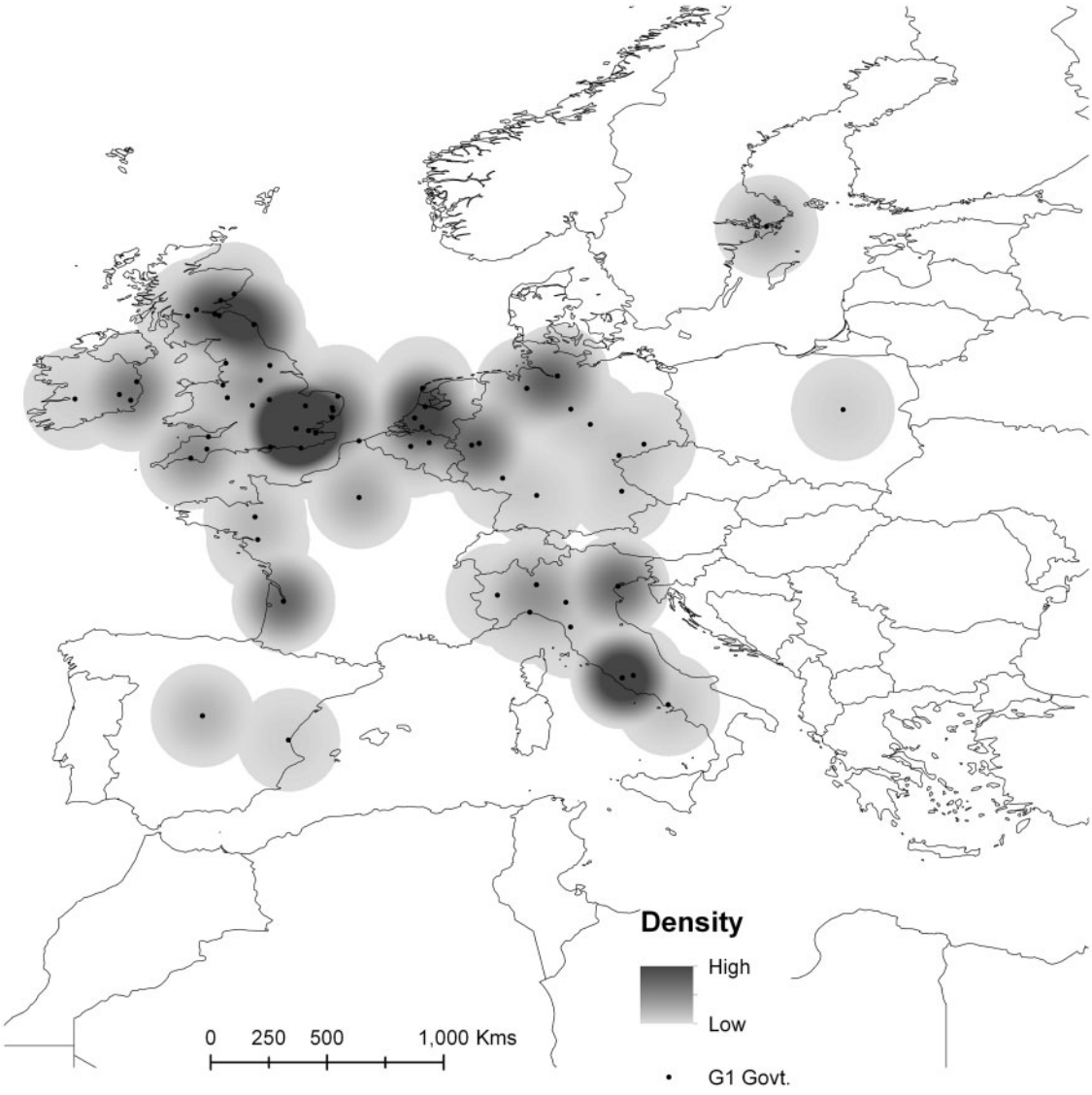

Fig. 6 Mapping government: references to places near a word tagged as *G1*

research questions of a corpus that contains place–names.

That said, the analyses presented above are no more than pilot studies. We have barely scraped the surface of the patterns within just the Lancaster Newsbooks Corpus. Moreover, there is clearly room for improvement to the procedure we have outlined. The data resources utilized, most critically the gazetteer, play a crucial role: the GIS output is only as good as the database that produces it. An improved gazetteer, for example one with information to assist in disambiguation of same-name places, would enhance the output. In a similar way, the incorporation into our method of Named Entity Extraction approaches could help automate the distinction between place–names and personal names. Another useful extension would be to assess the effect of applying

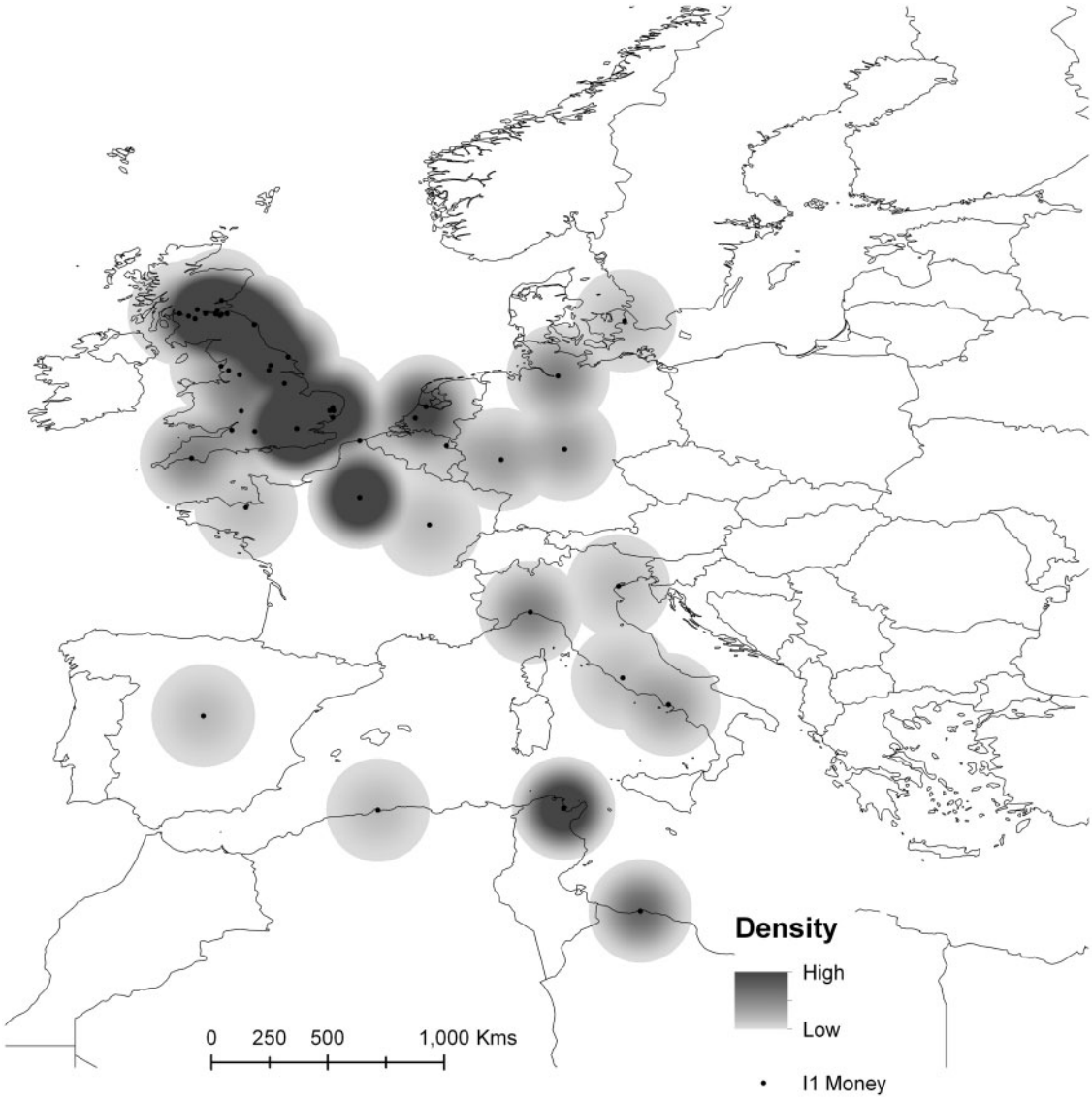

Fig. 7 Mapping finance: references to places near a word tagged as *I1*

different models of collocation when identifying co-occurrence of place–names and semantic tags. The proximity-based approach to collocation which we applied is very common in corpus linguistics (Sinclair, 1991; Sinclair *et al.*, 2004); however, other approaches exist and could be used within our method, such as considering words to collocate if they appear anywhere in the same document (or news story in our example) regardless of actual distance (Kim and Choi, 1999). The syntax-based approach to collocation mentioned in Section 5, although not adopted in the analysis presented here, is another possibility for further exploration.

However, while these and other aspects of the technique could be refined, it is our judgement that there will always need to be some user

intervention in the process of moving from a concordance of proper nouns, to a disambiguated and clean list of place–name mentions with coordinates. A wholly automated computation of a semantic visualization from a corpus would be possible, but probably unsatisfactory due to spurious place–name matches and the other problems illustrated in this article. In this way, Visual GISting inherits a key trait of corpus linguistics: that quantitative and qualitative analysis, both automatically generated outputs and hand-and-eye techniques, are indispensable.

References

Adolphs, S. (2006). *Introducing Electronic Text Analysis*. London: Routledge.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Harlow: Longman.

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

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T., and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society*, **19**: 273–306.

Baron, A., Rayson, P., and Archer, D. (2009). Automatic standardization of spelling for historical text mining. *Proceedings of Digital Humanities 2009*, University of Maryland, USA, 22–25 June 2009.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bodenhamer, D. J. (2010). The potential of spatial humanities. In Bodenhamer, D. J., Corrigan, J., and Harris, T. M. (eds), *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington: Indiana University Press, pp. 14–30.

Campbell, B. M. S. and Bartley, K. (2006). *England on the Eve of the Black Death: An Atlas of Lay Lordship, Land and Wealth, 1300-49*. Manchester: Manchester University Press.

Chrisman, N. R. (2002). *Exploring Geographic Information Systems, 2nd edn*. Chichester: John Wiley.

Clarke, K. C. (1997). *Getting Started with Geographic Information Systems*. Upper Saddle River, New Jersey: Prentice Hall.

Coppock, J. T. and Rhind, D. W. (1991). The History of GIS. In Maguire, D. J., Goodchild, M. F., and Rhind, D. W. (eds), *Geographical Information Systems: Principles and Applications. Volume I: Principles*. London: Longman Scientific and Technical, pp. 21–43.

Culpeper, J. (2002). Computers, language and characterisation: an analysis of six characters in Romeo and Juliet. In Melander-Marttala, U., Ostman, C., and Kyto, M. (eds), *Conversation in Life and in Literature: Papers from the ASLA Symposium, Association Suedoise de Linguistique Appliquee (ASLA), 15*. Uppsala: Universitetstryckeriet, pp. 11–30.

Cunfer, G. (2005). *On the Great Plains: Agriculture and Environment*. College Station, TX: Texas A&M University Press.

Deignan, A. (2005). *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Dorling, D., Mitchell, R., Shaw, M., Orford, S., and Davey Smith, G. (2000). The ghost of christmas past: health effects of poverty in London in 1896 and 1991. *British Medical Journal*, **321**: 1547–51.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart. http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/ (accessed 30 November 2010).

Francis, N. and Kučera, H. (1964). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for use with Digital Computers*, Department of Linguistics, Brown University, Providence. http://icame.uib.no/brown/bcm.html (accessed 12 October 2010).

Garside, R., Leech, G., and Sampson, G. (1987). *The Computational Analysis of English: a Corpus-Based Approach*. London: Longman.

Gordon, C. (2008). *Mapping Decline: St. Louis and the Fate of the American City*. Philadelphia: University of Pennsylvania Press.

Gregory, I. N. (2008). Different places, different stories: infant mortality decline in England & Wales, 1851-1911. *Annals of the Association of American Geographers*, **98**: 773–94.

Gregory, I. N. (2009). Comparisons between the geographies of mortality and deprivation from the 1900s to 2001: spatial analysis of census and mortality statistics. *British Medical Journal*, **339**(b3454): 676–9.

Gregory, I. N. and Healey, R. G. (2007). Historical GIS: structuring, mapping and analysing geographies of the past. *Progress in Human Geography*, **31**: 638–53.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval*. New York: ACM, pp. 98–7.

Grover, C., Tobin, R., Byrne, K. *et al.* (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **368**: 3875–89.

Hardie, A. and McEnery, T. (2009). Corpus linguistics and historical contexts: text reuse and the expression of bias in early modern English journalism. In Bowen, R., Mobärg, M., and Ohlander, S. (eds), *Corpora and Discourse – and Stuff: Papers in Honour of Karin Aijmer*. Gothenburg Studies in English 96. Göteborg: Acta Universitatis Gothoburgensis, pp. 59–92.

Hardie, A. and McEnery, T. (2010). On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics*, **15**: 384–94.

Jessop, M. (2008). The inhibition of geographical information in digital humanities scholarship. *Literary and Linguistic Computing*, **23**: 39–50.

Johnston, R. J. (1999). Geography and GIS. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (eds), *Geographical Information Systems: Principals, Techniques, Management and Applications, 2nd edn*. Chichester: John Wiley, pp. 39–47.

Kim, M.-C. and Choi, K.-S. (1999). A comparison of collocation-based similarity measures in query expansion. *Information Processing and Management*, **35**(1): 19–30.

Knowles, A. K. (2008). GIS and history. In Knowles, A. K. (ed.), *Placing History: How Maps, Spatial Data and GIS are Changing Historical Scholarship*. Redlands, CA: ESRI Press, pp. 1–26.

Knowles, A. K. and Healey, R. G. (2006). Geography, timing, and technology: a GIS-based analysis of Pennsylvania's iron industry, 1825-1875. *Journal of Economic History*, **66**: 608–34.

Leech, G. (1997). Introducing corpus annotation. In Garside, R., Leech, G., and McEnery, A. (eds), *Corpus Annotation*. London: Longman, pp. 1–18.

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (2001). *Geographical Information Systems and Science*. Chichester: John Wiley.

Lloyd, C. (2007). *Local Models for Spatial Analysis*. Boca Raton, FL: CRC Press.

Mahlberg, M. (2007). A corpus stylistic perspective on Dickens's great expectactions. In Lambrou, M. and Stockwell, P. (eds), *Contemporary Stylistics*. London: Routledge, pp. 19–31.

Martin, D. (1996). *Geographic Information Systems and their Socio-Economic Applications. 2nd edn*. Hampshire: Routledge.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics*. 2nd edn, Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, Z., and Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Ooi, V. B. Y. (1998). *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Peuquet, D. J. (1994). Its about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, **84**: 441–61.

Pickles, J. (1999). Arguments, debates, and dialogues: the GIS-social theory debate and the concern for alternatives. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (eds), *Geographical Information Systems: Principals, Techniques, Management and Applications*, 2nd edn. Chichester: John Wiley, pp. 49–60.

Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P., and Archer, D. (2008). The identification of spelling variants in English and German historical texts: manual or automatic? *Literary and Linguistic Computing*, **23**: 65–7.

Prentice, S. and Hardie, A. (2009). Empowerment and disempowerment in the glencairn uprising: a corpus-based critical analysis of early modern English news discourse. *Journal of Historical Pragmatics*, **10**: 23–55.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, **13**: 519–49.

Robinson, A. H., Morrison, J. L., Muehrcke, P. C., Kimerling, A. J., and Guptill, S. C. (1995). *Elements of Cartography*, 6th edn. New York: John Wiley and Sons.

Sampson, G. (2002). Regional variation in the english verb qualifier system. *English Language and Linguistics*, **6**: 17–30.

Schuurman, N. (2004). *GIS: A Short Introduction*. Oxford: Blackwell.

Semino, E. and Short, M. (2004). *Corpus Stylistics: Speech, Writing and thought Presentation in a Corpus of English Writing*. London: Routledge.

Seretan, V. and Wehrli, E. (2007). Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, **43**: 71–85.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J., Jones, S., Daley, R., and Krishnamurthy, R. (2004). *English Collocational Studies: The OSTI Report*. London: Continuum.

Teubert, W. (2005). My Version of Corpus Linguistics. *International Journal of Corpus Linguistics*, **10**: 1–13.

Tognini-Bonelli, E. (2001). *Corpus linguistics at Work*. Amsterdam: John Benjamins.

Notes

1 See http://www.ucl.ac.uk/english-usage/about/history.htm (last accessed 10 October 2010).

2 While we do not directly address the issue of time in the present paper, it should be noted that the methods we outline can be applied diachronically as easily as synchronically.

3 Some place–names may not be relevant, especially very aggregate places such as countries or continents which cannot be satisfactorily represented using point locations. These would be better geo-referenced using polygon-based approaches that are beyond the scope of this article.

4 ArcGIS is one of the standard GIS software packages. It is produced by Environmental Systems Research Institute (ESRI); see http://www.esri.com (last accessed 10 October 2010).

5 http://earth.google.com (last accessed 10 October 2010).

6 http://ucrel.lancs.ac.uk/usas.

7 http://www.ling.lancs.ac.uk/newsbooks.

8 Constituent Likelihood Automatic Word-tagging System; see http://ucrel.lancs.ac.uk/claws and Garside *et al*. (1987).

9 It should be noted that while the steps of POS tagging and semantic tagging of the data are *conceptually* separate in the method we propose, in our actual text-processing setup, CLAWS and USAS (the semantic tagger) were in practice run across the data as a single process (USAS is designed to take CLAWS output as its input). The tagged corpus was indexed and searched using Corpus Workbench (http://cwb.sourceforge.net) and a relational database system (MySQL) was used to handle the joining of the corpus query output and the gazetteer. Finally, a small number of custom scripts were used to manipulate the resulting XML data files, e.g. to extract only mentions close to a given semantic tag, or to convert the data into an appropriate input format for GIS software (ArcGIS) for visualization and spatial analysis. Google Earth can be used as a viewer for GIS data.

10 This is available to the UK HEI community through Edina; see: http://www.edina.ac.uk (last accessed 10 October 2010).

11 http://www.geonames.org (last accessed 10 October 2010).

12 http://www.world-gazetteer.com (last accessed 10 October 2010).

13 We were fortunate in our current data set in that most spelling variation in the Lancaster Newsbooks Corpus was annotated with normalized spelling equivalents during the transcription process; although the transcribers sometimes failed to normalize a place–name that was unfamiliar to them, the amount of noise in the data attributable to spelling variation was even so much less than it would have been in raw transcribed text. For this reason, our discussion later in this article focuses on other forms of noise in the data than spelling variation, although for other historical data spelling variation will clearly be a much more urgent issue, and one that may need some automatic preprocessing to address (see Baron *et al*., 2009).