Using corpora to map language: Geographical Text Analysis of UK poverty

Laura L. Paterson (The Open University, UK) and Ian Gregory (Lancaster University, UK)

This paper demonstrates the viability of using techniques from Geographical Text Analysis (GTA) on multimillion word corpora. It emphasises the fruitfulness of including geography as a variable within corpus-based discourse analysis. The paper centres on a case study comparing media discussions of UK poverty with official deprivation statistics. Using corpora of the *Guardian* and the *Daily Mail* from 2010-2015, the analysis shows that media reports of UK poverty are London-centric and tend to systematically exclude rural areas. There is also evidence to suggest that poverty becomes particularly newsworthy – thus generating more hits in a corpus – when it occurs in places that would not be expected, such as relatively high-income areas. By comparing the geographical spread of media discussions of *poverty* (and 85 related query terms such as *unemployment*, *benefits*, *pay*, *austerity*) we can see how each newspaper locates discussions of poverty in geographical space. The analysis also contrasts the corpus-generated results with existing official statistics of poverty to determine whether the linguistic data and the statistical data tally.

To explain the method more fully, Geographical Text Analysis (GTA) is way of analysing language which focuses on the identifiable geographical locations referred to in texts. It allows the researcher to visualise their corpus cartographically. Once a suitable corpus or corpora have been compiled, query terms are selected to reflect the topic under investigation. For the current study these were generated by combining the top 50 lexical collocates of **poverty** in both corpora, with the addition of further terms based on wider reading (for example, the names of particular UK welfare payments like Jobseeker's Allowance (JSA) were included in the analysis). This led to a total of 86 queries which were run on both corpora. Concordance lines were generated, downloaded, and fed through a software programme known as a geoparser. In particular, the Lancaster University adaptation of the Edinburgh geoparser (Grover et al. 2010) was used for this research.

Geoparsing involves the automatic detection of place names occurring within a set span L/R of the concordance node (+/-10 for this study). Place names can be detected at country level down to district level for locations all over the world. Thus, although the focus here is restricted to UK place names only, GTA is a versatile method which can be used to analyse the use of locations globally. Once place names have been identified in the concordance lines, these Place Name Co-occurrences (PNCs) are tagged with the geographical coordinates corresponding to their location on the Earth's surface. The results of the geoparsing process are then manually analysed to spot errors and address any ambiguities where a place name could refer to more than one geographical location (e.g. Boston, Lincolnshire or Boston, Massachusetts). Once any erroneous hits have been eliminated, maps of the locations occurring in the cotext of the query nodes can be plotted (Figure 1).

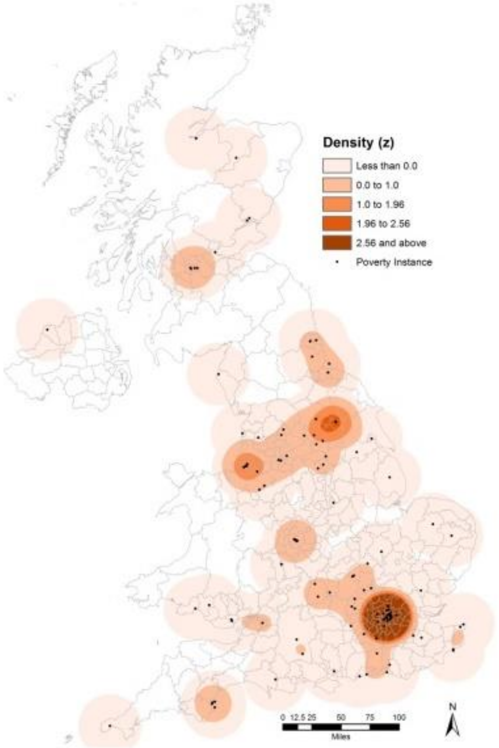

Figure 1: Map of **poverty** in the Guardian newspaper

There were 70,397 hits for all query terms in the *Guardian* and 463,278 in the *Daily Mail*, which were condensed during manual analysis to 6794 and 6605 hits respectively. The large differences between the geoparser output and the final hits can be explained by three factors:

1) erroneous geoparsing hits either for places outside of the UK (Albania, Toronto, Morocco) or for fictional places (Hundred Acre Wood, Downton Abbey),
2) erroneous query hits, such as *Chavez* being picked up by the query *chav**, and
3) hits which upon closer inspection of the surrounding discourse were not related to poverty (e.g. the query *expens** generated hits relating to the MPs' expenses scandal in 2010).

It is possible to train the geoparser and, as such, the results of the manual analysis of the *Daily Mail* and *Guardian* corpora will be used to decrease the rate of Type 1) errors in future work. Type 2) errors are easily removed from the geoparser output, but can be decreased pre-geoparsing by adapting queries based on an initial analysis of concordance lines. Finally, Type 3) errors are more difficult to detect pre-geoparsing, especially given the size of the two corpora, but could potentially be decreased by restricting query terms to bi-directional collocates and/or selecting or rejecting query terms based on analysis of a thinned sample of the geoparsed output. Testing these methods of reducing errors is designated for future work.

Forms of GTA have been successfully used to provide novel interpretations of literary texts and travel guides (Donaldson et al. 2015) as well as historical documents (Murrieta-Flores et al. 2015; Porter et al. 2016a), and work is ongoing with early UK newspapers (Porter et al. 2016b). The present paper is part of a project which uses GTA in combination with CDA to analyse the discursive construction and use of poverty discourses in twenty-first century mass media texts. Our choice to focus on poverty addresses the fact that statistical information alone cannot lead to in-depth understandings of social phenomena. The primary measurements of poverty and statistical deprivation used to set government policies are quantitative. However, poverty is not merely an economic circumstance. Jo (2013: 522) argues that the relationship between poverty and shame is constructed 'from the dominant discourse' and cultural norms, which are 'collectively assembled by multiple institutions which are governed by those with power and influence'. Thus, the measurement of poverty is not the measurement of an objective phenomenon. Linguistic interpretations of the term 'poverty' include Kress' (1994:28) argument that poverty is 'a characteristic which acts as a description of a person, a classification' (1994:29). By conceptualising poverty as a label for a particular group we can begin to see how its use in discourse could be a powerful move; to label someone as 'poor' or 'in poverty' is to imbue them with a set of (negatively-loaded) characteristics that, presumably, they cannot escape. Linguistic analysis of poverty – in the form of GTA – demonstrates that, although similar lexical resources may be used by each newspaper, their use of place names locates their discussions differently, and, furthermore, the places named in each newspaper do not necessarily correspond to the statistical data.

References

Donaldson, C., Gregory, I.N., & Murrieta-Flores, P. (2015). 'Mapping Wordsworthshire': A GIS study of literary tourism in Victorian England" *Journal of Victorian Culture*, 20, 287-307.

Grover, C, Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S. & Ball, J. (2010). Use of the Edinburgh Geoparser for georeferencing digitised historical collections. *Philosophical Transactions of the Royal Society A,* 368(1925), 3875-3889.

Jo, Y.N. (2013). Psycho-social dimensions of poverty: When poverty becomes shameful. *Critical Social Policy,* 33(3), 514-531.

Kress, G. (1994). Text and grammar as interpretation. In U.H. Meinhof & K. Richardson (Eds.), *Text, Discourse and Context: Representations of Poverty in Britain* (pp. 24-47). London: Longman.

Murrieta-Flores, P., Baron, A., Gregory, I.N., Hardie, A. & 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, 296-320.

Porter, C., Atkinson, P., & Gregory, I.N. (2016a). Geographical Text Analysis: A new approach to understanding nineteenth-century mortality. *Health & Place*, 36, 25-34.

Porter, C., Gregory, I.N., & Atkinson, P. (2016b). Investigating the temporal and spatial representations of disease in nineteenth-century British newspapers through text analysis and GIS. *American Association of Geographers 2016*, San Francisco, USA, 30/03/2016.