Customising Geoparsing and Georeferencing for Historical Texts

C.J. Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson, Ian Gregory, Andrew Hardie, Patricia Murrieta-Flores

@spatialhums
@perayson
http://www.lancaster.ac.uk/spatialhum/
Spatial Humanities:
texts, geographic information systems and places

- Five year project (2012-16)
- Funded by European Research Council under starting researcher grant to Prof Ian Gregory (agreement number 283850)
- Building on technical expertise in Digital Humanities, Corpus and Computational Linguistics and Historical Geographical Information Systems (HGIS)

Themes
- Methodologies (GIS + text analysis)
- Analysing qualitative sources (Lake District literature in 18th and 19th centuries; Histpop and Bopcris datasets; BL 19th century newspapers and books)
- Developing the skills base (training events and workshops)
Corpus of Lake District literature 1622-1900

- 80 texts, comprising around 1.5M words
- Mixture of canonical and non-canonical literature about the Lake District
- Mixture of genres, including guidebooks, travelogues, novels, poems, journals, and private letters
Geoparsing historical text

Out of the box results
Problems and customisations

- **Problems**
  - Spelling variation
  - Variant forms of placenames
  - Placenames that are specific to the region of the English Lake District

- **Solutions**
  - VARD variant spelling detector (pre-processing)
  - Gazetteer with the locations of attested historical placenames (DEEP)
  - Specialised georesolution in Geoparser giving preference to locations within the specified region
Hand correction georesolution
## Results

<table>
<thead>
<tr>
<th>Documents</th>
<th>Unlock</th>
<th></th>
<th>DEEP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Overall</td>
<td>91.58</td>
<td>63.68</td>
<td>86.06</td>
<td>74.40</td>
</tr>
<tr>
<td>1700-99</td>
<td>87.22</td>
<td>48.97</td>
<td>81.32</td>
<td>57.98</td>
</tr>
<tr>
<td>1700-69</td>
<td>93.15</td>
<td>57.30</td>
<td>85.92</td>
<td>66.85</td>
</tr>
<tr>
<td>1770-99</td>
<td>86.16</td>
<td>47.62</td>
<td>80.55</td>
<td>56.62</td>
</tr>
<tr>
<td>1800-99</td>
<td>94.70</td>
<td>72.25</td>
<td>88.70</td>
<td>72.25</td>
</tr>
<tr>
<td>1800-29</td>
<td>96.03</td>
<td>73.63</td>
<td>89.83</td>
<td>84.70</td>
</tr>
<tr>
<td>1830-99</td>
<td>92.75</td>
<td>70.23</td>
<td>87.07</td>
<td>83.26</td>
</tr>
</tbody>
</table>
Summary and conclusions

- Process intended to support close reading of historical texts on a much larger scale by highlighting using exploratory and data-driven approaches which parts of the corpus warrant further close analysis.
- Case study from a corpus of Lake District travel literature.
- Customisations to existing tools to extract place name information and visualise it on a map.
Thanks for listening!

• For more information about the project, see:
  – http://www.lancaster.ac.uk/spatialhum/

• @spatialhums

• @perayson

• Acknowledgements:
  – We would also like to thank Claire Grover, the geoparser developers and the Unlock Text support team in Edingburgh, Jayne Carroll and the DEEP project, and M. and J. Norgate for putting their work online as: http://www.geog.port.ac.uk/webmap/thelakes/html/lgaz/lgazfram.htm