A Deeply Annotated Testbed for Geographical Text Analysis

THE CORPUS OF LAKE DISTRICT WRITING

Paul Rayson
Alex Reinhold
James Butler
Chris Donaldson
Ian Gregory
Joanna Taylor

Lancaster University

Data

- The Corpus of Lake District Writing (CLDW) was assembled for the Lancaster University Spatial Humanities: Text, GIS & Places project between 2012 and 2016
- The corpus contains 80 digitized texts about the Lake District.
- These texts date from between the years 1622 and 1900, and they include a number of different genres, such as travel journals, essays, novels and poetry.
- The 80 texts in the corpus contain over 1.5 million word tokens in total.

Gold Standard Corpus

- This gold standard subset included a representative sample of 28 texts, which were selected from the different genres and historical periods included in the corpus
- Approx. 242,000 word tokens: about one-sixth of the entire corpus.

Table 1: Gold Standard Sample

Author	Title	Date	Word tokens
Lt. Hammond	A Relation of a Short Survey of 26 Counties	1634	2,492
Daniel Defoe	A Tour thro' the Whole Island of Great Britain	1725	3,019
George Smith	'Survey of the Northwest Coast of England'	1746	2,172
George Smith	'A Journey up to Cross-Fell Mountain'	1747	2,774
George Smith	'A Journey to Caudbec Fells'	1747	6,955
John Brown	Description of the Lake and Vale at Keswick	1751	1,107
John Wesley	An Extract of the Rev. Mr. John Wesley's Journal	1759	1,783
Thomas Gray	Journal of A Visit to the Lake District in 1769	1769	6,388
Thomas Pennant	A Tour in Scotland. MDCCLXIX	1769	1,069
Arthur Young	Six Months' Tour Through the North of England	1770	13,388
Thomas Pennant	A Tour in Scotland, and Voyage to the Hebrides	1772	13,994
Thomas West	A Guide to the Lakes	1778	36,793
William Cockin	Ode to the Genius of the Lakes. A Poem	1780	4,793
R. J. Sullivan	Observations Made During a Tour	1780	2,799
Stebbing Shaw	A Tour in 1787	1787	7,073
J. H. Manners	Journal of a Tour	1796	1,311
S. T. Coleridge	Letters and Notebooks	1802	9,237
Priscilla Wakefield	A Family Tour through the British Empire	1804	5,459
Anonymous	Gleanings of a Wanderer	1805	1,802
John Keats	Letters	1818	3,794
William Wordsworth	The River Duddon: A Series of Sonnets	1820	10,938
Jonathan Otley	A Concise Description of the English Lakes	1823	29,704
John Ruskin	Iteriad; or, Three Weeks Among the Lakes	1830	21,346
William Wordsworth	Guide through the District of the Lakes	1835	5,464
Anonymous	Keswick and its Neighbourhood: A Hand-book	1852	29,494
James F. Clarke	Eleven Weeks in Europe	1852	1,606
Anonymous	The English Lakes	1859	10,126
Herbert Rix	'Down the Duddon with Wordsworth'	1893	5,412
		Total	242,292

Gold Standard Corpus

- The corpus was run through the Edinburgh geoparser, which performed Named Entity Recognition (NER) on the texts and matched locations to the Unlock gazetteer.
- This subset was hand checked and coded using XML tags in order to mark every place-name entity it contained.

<p> But I must not forget <cdplace>Winander Meer</cdplace>, which makes the utmost northern bounds of this shire, which is famous for the char fish found here and hereabout, and no where else in <cdplace>England </cdplace>; it is found indeed in some of the rivers or lakes in <cdplace>Swisserland </ cdplace> among the <cdplace> Alps </ cdplace>, and some say in <cdplace> NorthWales </cdplace>; but I question the last. It is a curious fish, and, as a dainty, is potted, and sent far and near, as presents to the best friends. </ p>

Deep analysis

- Through the delineation of locational type, place-names with a common element (such as Coniston [lake], Coniston Old Man [height], Coniston Beck [waterway], Coniston Fell [height], Coniston [settlement] or Coniston Hall [house]) could be rendered as distinct entities, allowing an unprecedented range of geospatial analysis without any ambiguity as to point of reference.
- This process is similar to that found in the Pelagios Project, linking multiple variants and place classification within a text.

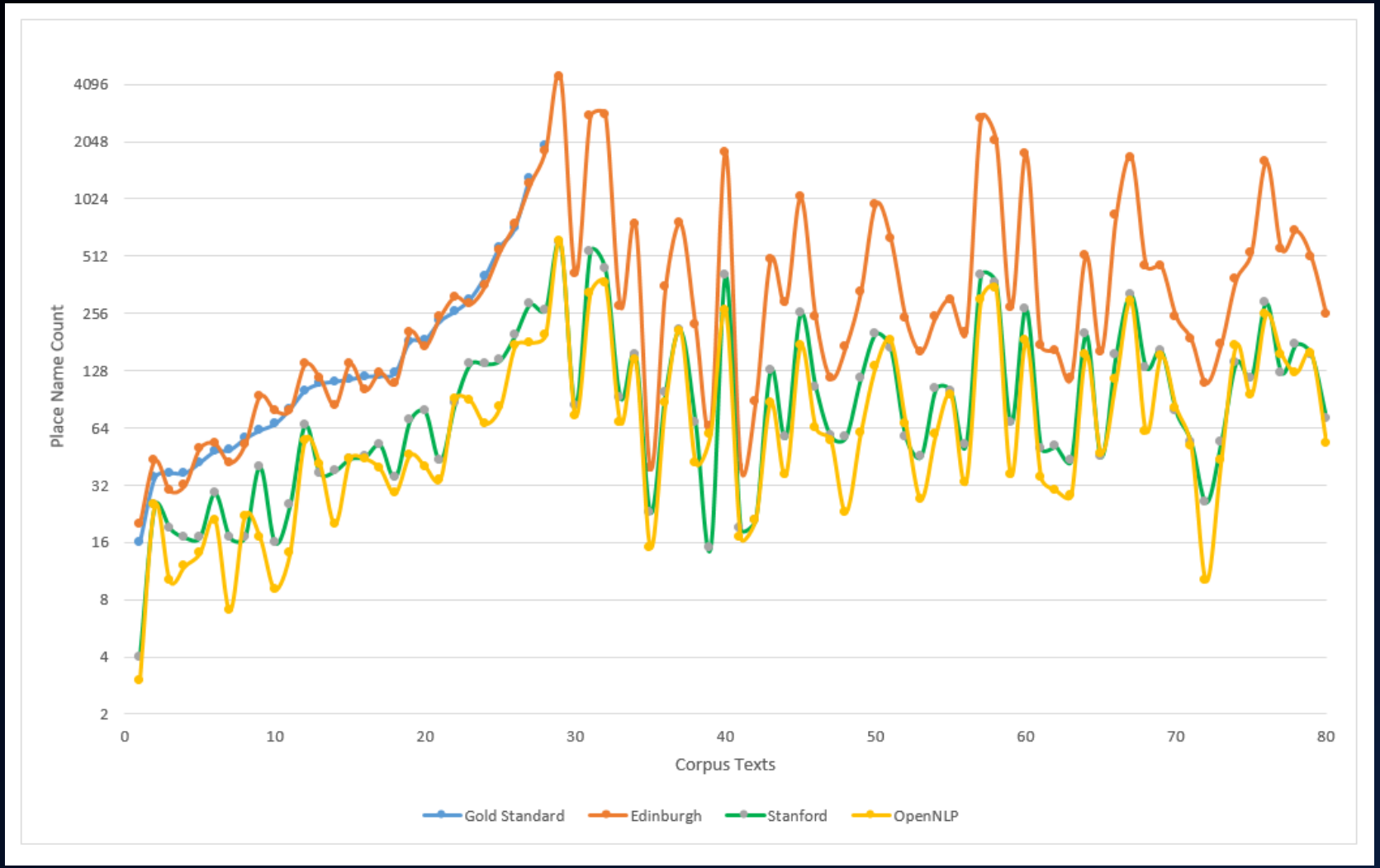
CASE STUDY: NER EVALUATION

- To test the potential of the corpus as a deeply annotated testbed we used two further well-known NLP libraries: Stanford CoreNLP and OpenNLP.

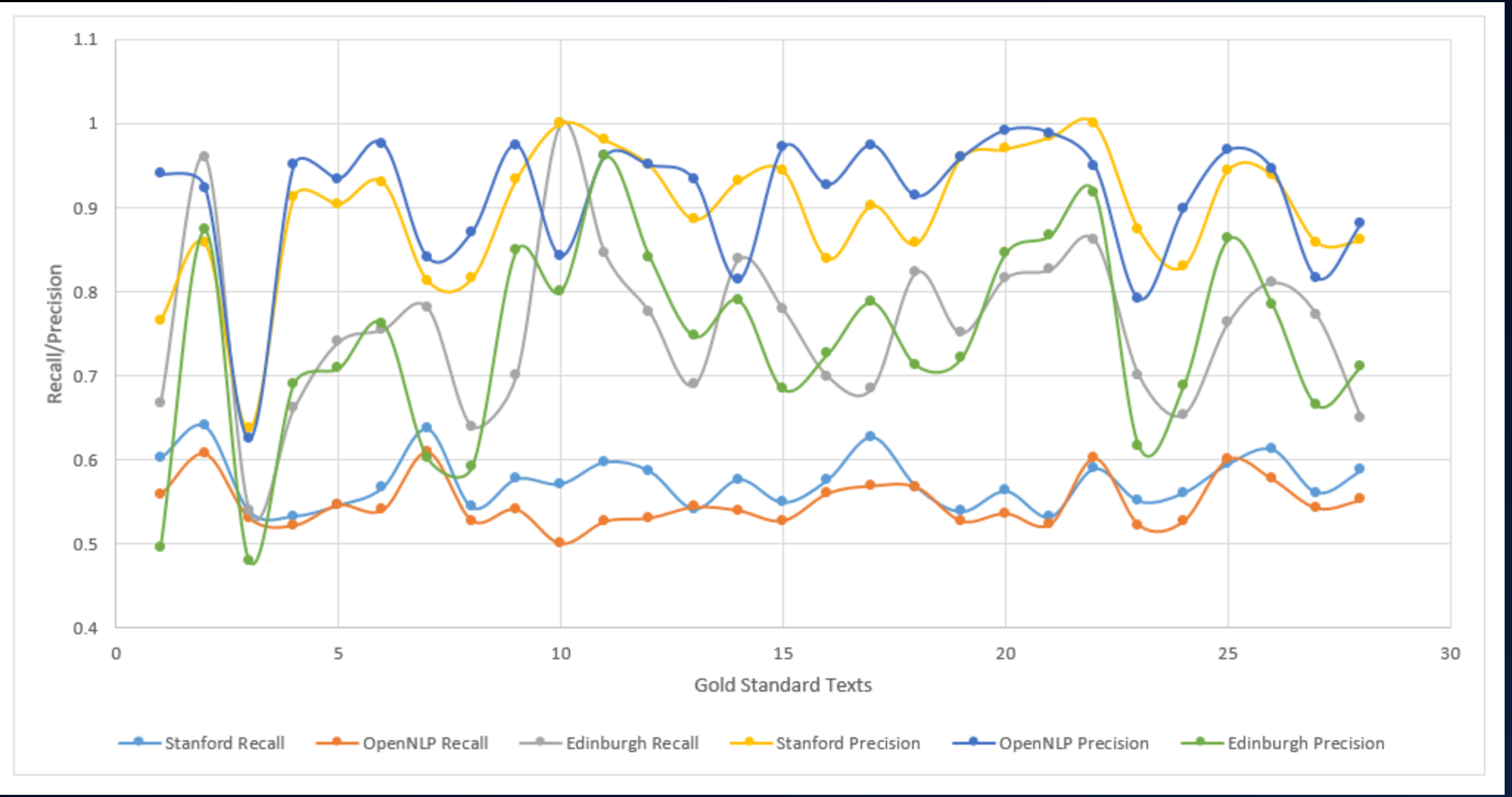
Table 2: Average NER Precision and Recall results for 28 files

	Average Precision	Average Recall
Edinburgh	0.74	0.76
Stanford	0.90	0.57
OpenNLP	0.91	0.55

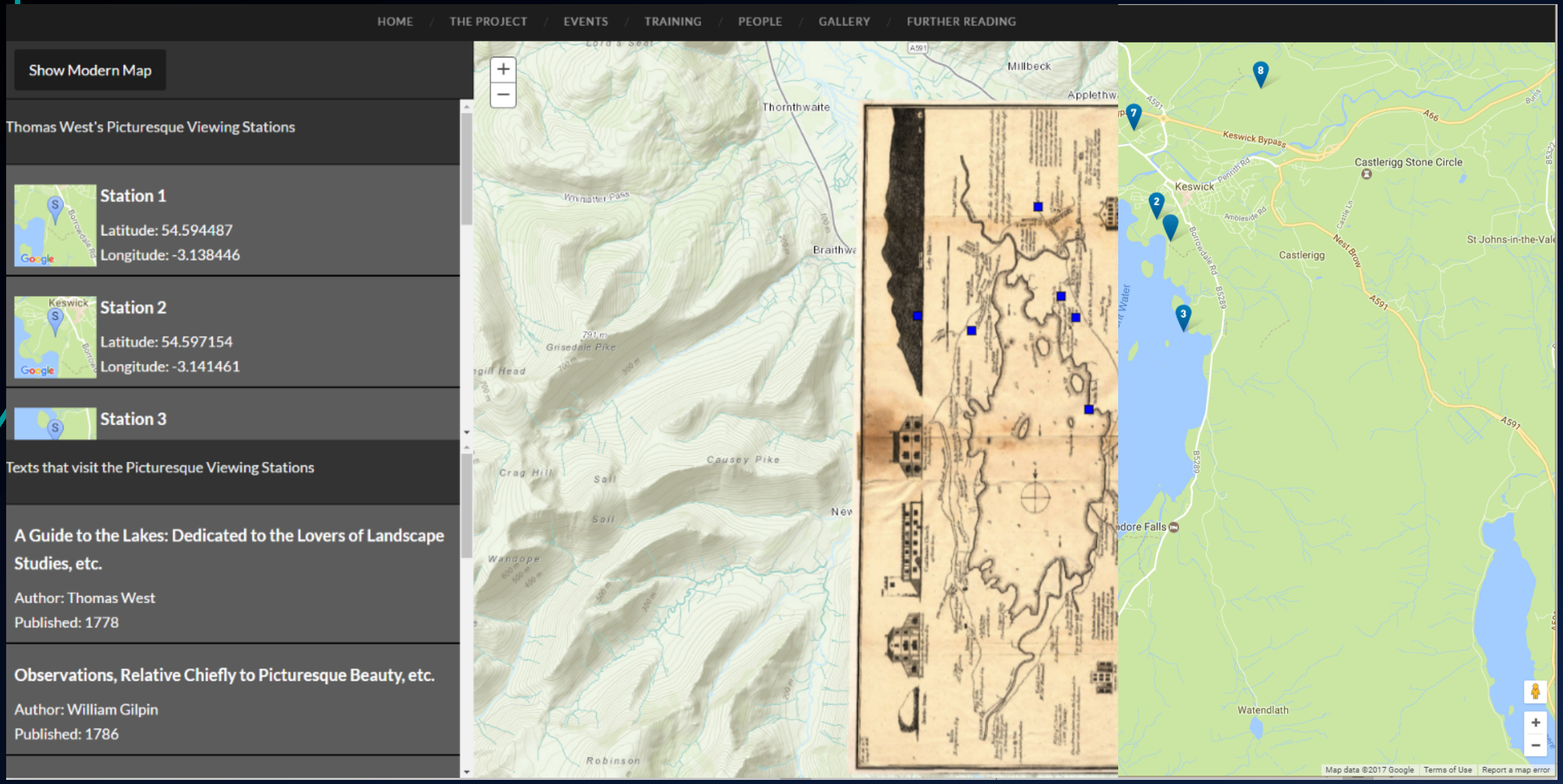
Comparison of NER tools with gold standard.



NER Precision and Recall results for 28 files.

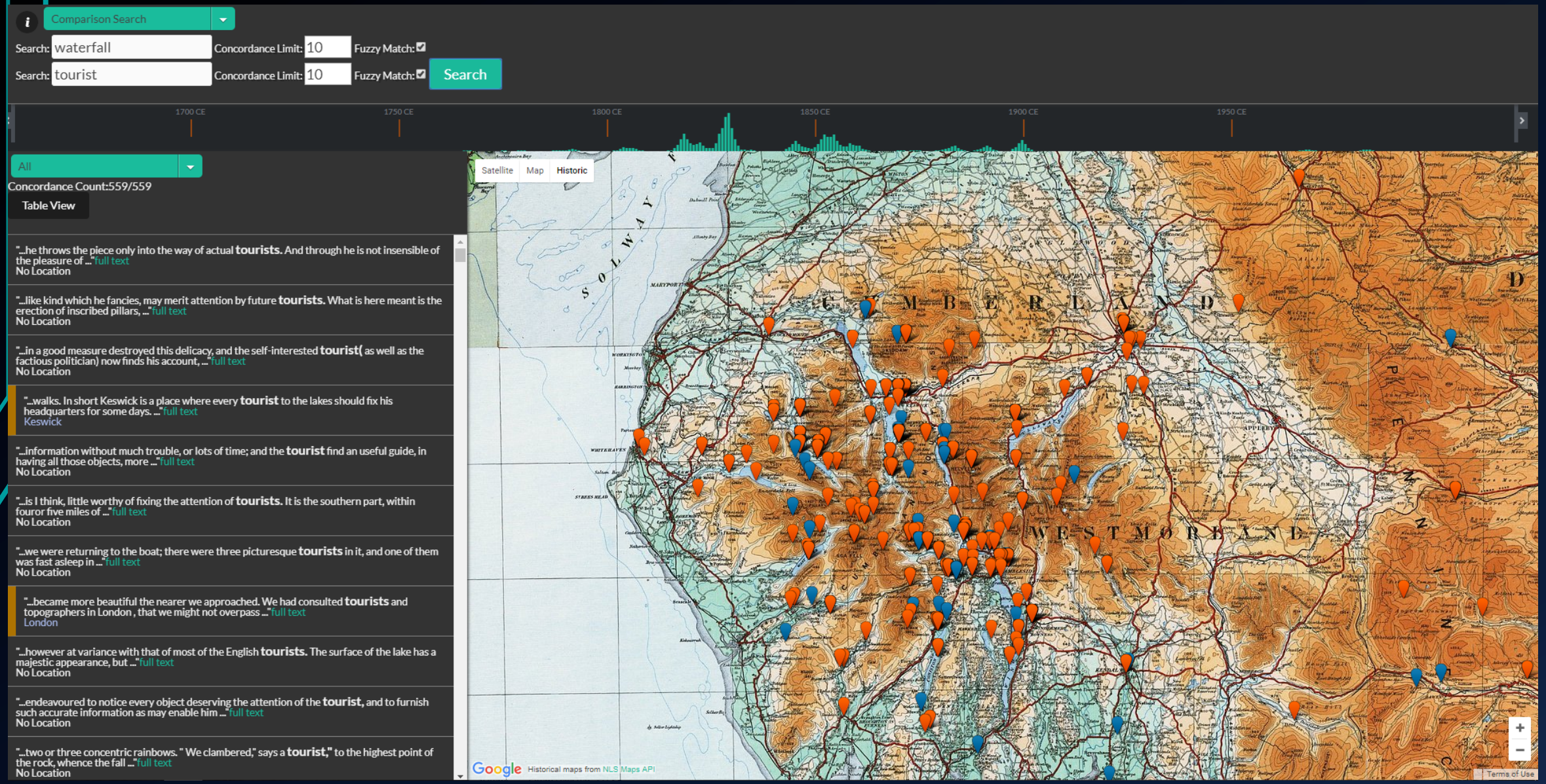


Crosthwaite Deep Map – Map View



Crosthwaite Deep Map – Station View

NLP Deep Map


Comparison Search
Search: waterfall
Concordance Limit: 10
Fuzzy Match:
Search: tourist
Concordance Limit: 10
Fuzzy Match:
Search
1700 CE
1750 CE
1800 CE
1850 CE
1900 CE
1950 CE
All
Concordance Count:559/559
Table View
"...he throws the piece only into the way of actual tourists. And through he is not insensible of the pleasure of ..."full text
No Location
"...like kind which he fancies, may merit attention by future tourists. What is here meant is the erection of inscribed pillars, ..."full text
No Location
"...in a good measure destroyed this delicacy, and the self-interested tourist(as well as the factious politician) now finds his account, ..."full text
No Location
"...walks. In short Keswick is a place where every tourist to the lakes should fix his headquarters for some days. ..."full text
Keswick
"...information without much trouble, or lots of time; and the tourist find an useful guide, in having all those objects, more ..."full text
No Location
"...is I think, little worthy of fixing the attention of tourists. It is the southern part, within fouror five miles of ..."full text
No Location
"...we were returning to the boat; there were three picturesque tourists in it, and one of them was fast asleep in ..."full text
No Location
"...became more beautiful the nearer we approached. We had consulted tourists and topographers in London , that we might not overpass ..."full text
London
"...however at variance with that of most of the English tourists. The surface of the lake has a majestic appearance, but ..."full text
No Location
"...endeavoured to notice every object deserving the attention of the tourist, and to furnish such accurate information as may enable him ..."full text
No Location
"...two or three concentric rainbows. " We clambered," says a tourist," to the highest point of the rock, whence the fall ..."full text
No Location
Satellite
Map
Historic
SOLWAY F.
CUMBERLAND
WESTMORELAND
Google
Historical maps from NLS Maps API
Terms of Use

CONCLUSION

- We have described the Corpus of Lake District Writing which goes some way to addressing humanities issues and provides a novel more deeply annotated testbed than has been available before.
- Much more research remains as future challenges for geographical text analysis, not least related to metonymy which occurs where a place-name is used as a substitute for something else.
- Our Lake District corpus is available to download from our GitHub repository https://github.com/UCREL/LakeDistrictCorpus.