

Report on GeoParser <http://bodach.ucs.ed.ac.uk:18080/parser/index.html>

There are two parts to this report. First, considering the output and accuracy of the parser itself. Second, on the interface and the usability of the site as a whole.

1. Accuracy of parser

1.1. Sample text: I used a text which covered the first 50 pages of the general report to the 1851 census. The sample had a total of almost 30,000 words, and contained (by my calculation) 318 place names. (Not knowing the terminological basis of the Parser I excluded the Roman and archaic Anglo-Saxon terminology which occurred in the report.)

Running the sample through the parser brought up a total of 945 places. Some of these were 'duplicate' names, relating to different types of geography. For the remainder of this section I shall only be considering unique place names identified by the parser. The Parser extracted a total of 295 place names. [See recommendation 1]. 234 of the parsers unique output occurred in my list of names, thus the parser collected 60 'places' which I had not identified as place names. These are listed below.

Table 1. Places selected by parser, but not identified as places from the text.

Barony	Husband	Royal
Book	Indian Navy	SECRETARY OF STATE
Borough	Isca	Sheriff of Linlithgow
Britannia	Islands	Son
County	Kemble	South Eastern
County of England	Kingdom	South Wales
County of Surrey	Law	St Kilda
Culloch	Lothian	Stonehouse
Deva	MACHINERY	Strang
East India	Malmesbury	Terras
Ella	Merchant	The Cheviot Hills
England.	Mona	Thule
European	Morecambe Bay	Tulloch
Florence	North Foreland	Tything
Frith	North Wales	Tythings
Greenwich Pensioners	Palgrave	Wallingford
Hamlet	PARISH	Wendover
Hamlets	PEOPLE	Wife
Harbour	Redgrave	
HOME DEPARTMENT	river Elbe	
Horsley	river Lea	

2 Analytical discussion of 'errors'

2.1. Errors of commission

2.1.1 Straight-forward errors

Of these 60 places some of these are outright incorrect, .e.g, HOME DEPARTMENT, MACHINERY, PARISH, PEOPLE, SECRETARY OF STATE. My guess is that the underlying algorithms selecting places has problems with capitalization.

2.1.2 Context sensitive errors

Some of these words identified are names of people (as well as names of places), e.g., Florence of Worcester, William of Malmesbury and Roger of Wendover (or people and books, e.g, Horsley's *Britannia Romana* and Palgrave's *English Commonwealth*). (Mr Redgrave produced some statistics about the numbers of people attending the Great Exhibition, Wallingford is found as an authority; Colonel Tulloch provided statistics on the numbers of Greenwich and Chelsea pensioners. The place Culloch returned is part of the economist M'Culloch's name.) I suspect that this is unusual in user texts, but it would be possible to (perhaps, iteratively) build up a list of indicators which would be used in a context specific manner, i.e., if a place is preceded by the a word representing a name and the word of, then the Parser would return a lower level of confidence for the hit. (Obviously Malmesbury, Wendover are places, but they are not places in this context, they are names.)

Also under this heading are other potential stop-words which are also places: Barony, Borough, County, Harbour and Hamlet. If one is supplying a list of places to the Parser then this is ok, but if one is supplying a free text, then there are problems. If the text reads "Portsmouth Harbour" then it should be clear that the place here is Portsmouth, and not a place called Harbour in the Isle of Man.

Finally under this heading is East India, which is only ever used in the sample text in the context of the East India Company. As with personal names, there are possibilities for increasing context sensitivity in identifying place names in texts.

2.1.3 Terminological errors

Some terms may have been picked up (presumably) because they are parts of the place names within the parsers set of key terms, e.g., County of England/Surrey, European, Lothian, Islands, Hamlets, Kingdom, Royal, South Eastern, Tything, etc. In the sample text these terms often appear in a context where there is no geographical term close by (but sometimes where there is...) These terms should generally be used as indicators and not returned in the results unless they are place names. However, they may sometimes become place names as in the aforementioned Hamlet or County.

2.1.4 Incomplete parsing errors

The next set of terms in my analysis are words which do refer to place names, but were incorrectly parsed: Frith and Lothian (see above). Frith occurs in the text as part of the term 'Frith of Solway', but neither the whole phrase or Solway separately are picked up in the parsing process. The term Stonehouse is an oddity falling into this category. It is of course a place name in itself, but actually only appears in the text as in the context of East Stonehouse.

2.1.5 Not errors

What looks like an error but isn't is the parser's correct identification of Deva (Chester). I excluded Roman and archaic places from my count but this one was picked up. (Other such places weren't though – I think that it would be useful for the user to know the base coverage which they are likely to receive hits on, i.e., there is no guarantee that foreign places will be correctly identified, etc.)

2.1.6 Other errors (capitalization?)

Finally there are some terms which may be place names, but I've never heard of: Book, Husband, Merchant, Son, and Wife. These all did return a lower level of confidence, but these should probably be stop words. Perhaps dealing with nineteenth-century capitalization is problematic? The other ambiguous term in this section is 'Law'. This is a place name, but is not used in the text as one. Perhaps, the place name dictionary could be compared with a standard dictionary, and all words which are used in common English could be given a different level of confidence, or dealt with in a more context sensitive fashion.

The uncommented on places above, e.g., Indian Navy and Greenwich Pensioner are just a problem.

2.2 Conclusion of errors of commission

All in all the identification of place names was relatively successful, and almost all of the 'rogue entries', i.e., those listed above were given lower levels of confidence in the demo version I looked at (though I'm told that these were simply place holders.)

It is also interesting to note that a number of place names occurred within this lower level of certainty which were certainly places. What I would be interested to know is what is happening to these? In this list (low confidence) are terms like Bethnal Green, Bristol Channel, Britain, Channel Islands, Crystal Palace, Guernsey, etc., which are clearly places. Should not these be added to the 'dictionary' if they are identified?, even though they don't have eastings and northings? Perhaps they are already in the dictionary which is why they are matched? I would recommend that there is a greater level of clarity on the meanings of the different confidence levels, so that the user does not speculate, as I do, on how these were made hits.

3.1 *Errors of omission*

On the other side of the coin are the places which were in the text (and I identified) which the parser did not.

3.1.1 (*My mistakes.*) First up was a spelling error in the original text which I didn't account for, Bannf. So forget about this one.

3.1.2 *Foreign places.* I wasn't really expecting the service to parse, but would probably be useful to highlight this for other users. In the text were: Belgium, Cairo, China, Drontheim, Elbe, Heligoland, Iceland, Lofoden Islands, Mediterranean, Mexico, Norway, Paris, Portugal, Tripoli, Turkey, Two Sicilies, and United States.) [In case its not clear Drontheim would seem to be a nineteenth-century spelling of Trondheim.] (However, some foreign places, like Africa, America, Bengal, etc., were picked up (at a low level of confidence though.) Everything or nothing should probably be the watch word here.)

3.1.3 *Cities* A group of reasonably important cities from England and Wales were not selected by the Parser. These were: Bath, Cambridge, Coventry, Derby, Gloucester, Hull, Lincoln, Newcastle-upon-Tyne, Plymouth, Rutland, Stockton-on-the-Tees, Sunderland-on-the-Wear, Wells, Wirrall, York. This is presumably because these are not places in current administrative geography. (I can understand how Sunderland-on-the-Wear was missed, but the others?)

3.1.4 *Counties.* A group of old English and Scottish counties were not picked up. These include the terms: Berks, Bute, Caithness, Hebrides, Linlithgow, Orkney, YORKSHIRE. (Interestingly the term Sheriff of Linlithgow was extracted, but Linlithgow alone was not.) Some effort should be made to include basic county names in the parser.

3.1.5 *Islands* Some islands in the British Seas, mostly around Scotland, but others around England too: Arran, Calf, Canna, Chapel, Coquet, Inch-Keith, Jethou, Lewis, Little Papa, Man, May, Muck, Mull, Raasay, Rona, Rum, Scalpa, Skye, St. Michael, Staples and Vaila were not found.

3.1.6 *Natural features* Some rivers and other natural features were excluded: Lea, Ribble, Severn, Tay, Tees, Thames, Tyne, Wash, Murray Frith, Morecombe Bay and Naze. On what basis does the parser attempt to identify these places? Can the 'dictionary' contain more terms which are likely to occur in this type of place to extract them. Note, however, that most of these rivers were mentioned in the text in contexts like "between the Wash and the Thames" where there is no direct indication that these are rivers.

3.1.7 *Others*. A rag-bag of other types were not identified: British Isles, Cheviot Hills, East Stonehouse, Greenwich, Hanover Square, Lowlands, May Fair (sic).

3.2 *Conclusion*

A fair proportion of place names were not picked up by the Parser. This proportion of misses is not large, but it is significant enough to cause concern, especially with some of the more 'basic' place names. It is also significant enough to cause concern for someone who is unwilling to check their results.

The different types highlighted in the above paragraphs should probably be addressed with different levels of priority. It would seem that a balance has been made to achieve the most correct matches while minimizing false positives. The impact of this level of prioritization may have caused a number of unmatched place names. (I can't be clear on this because I don't know exactly how the parser works.)

The user should also be made aware that the system is (so to speak) literal. If you include a place name in a submitted text with the word Barony in it, it will treat it as though it is a place name rather than a descriptor which it is likely to be 99 times out of 100.

4. *Usability*

I found the system easy to use, and reasonably intuitive especially given the lack of help currently available.

One method of preventing the plethora of different results for the user, i.e., when Avon comes up five times and only one is of value to the user, might be to include an intermediate stage in the process. Note that these five terms are all potential matches but it is likely that the user will only have one of these terms in mind, and will need to be able to distinguish between them.

The intermediate stage might work as follows: First, the user feeds in the text; second there are tabular results. Then, a new stage, prompted by a dialog box: "Some place names are related to more than one grid reference. Do you want to refine your results?" If the user answers yes, then they are taken to a grid for each non-unique place name which looks something like the ONS site below:

Match Type	LA	County	
Plymouth (Placename)	Plymouth	Plymouth	Select
Plymouth City Airport (Placename)	Plymouth	Plymouth	Select
Plymouth (LA)	Plymouth	Plymouth	Select
Plymouth (Ward)	Merthyr Tydfil	Merthyr Tydfil	Select
Plymouth (Ward)	The Vale of Glamorgan	The Vale of Glamorgan	Select

The user would be able to select which (or perhaps all) of the options, before moving to the next stage.

It would also be useful here to have the Match terms separated from their type, i.e., in a separate column, and perhaps the current administrative county, as in the ONS site (this could be calculated from the grid-reference?) In the example below, the third Aberdeen seems to be somewhere south of Keighley, and knowing that before clicking on the map view would mean we could discard it from our thoughts. (Not all users will look at the grid-references and see that the third Aberdeen is unlikely to be the one in Scotland.) Similarly in the test text, Alexandria refers to a place in Egypt and not in Scotland.

Selected Document: REPORT51_2.txt				
Place Name	Easting	Northing	Map	Certainty (%)
Aberdeen (cities)	392110	807979	View	50.3%
Aberdeen (parish admin)	392515	805509	View	50.3%
Aberdeen (undefined)	402499	434452	View	50.3%
Africa	none	none	-	20.1%
Alderney (ward)	405101	93981	View	50.3%
Alderney (village or hamlet)	404499	94483	View	50.3%
Alexandria (towns)	239484	680050	View	50.3%

Include locations with certainty greater than or equal to %, up to placenames per document. [Update](#)

Furthermore, at present the functionality of the dialog box below the results window (see above) seems not to be functioning. 'Seems', because if you alter the figures and click update, nothing happens, and I haven't had the patience to wait for the various forms of output which are available. It should be clear to the user that this is a chance to see fewer items on-screen and that clicking update actually updates what's visible. It was also unclear to me what function this was supposed to perform. By default this says 50% and 10 place names. But on first getting to this screen, items

with a certainty of less than 50% are already being displayed. Perhaps this element has not been implemented yet. When it implemented, there should be some link to some help which explains what the certainty percentage means, i.e., in the example above, 20.1% certainty means “we don’t have a grid-ref” and 50.3% means “we do have a grid-ref” (And, while we’re on certainties, is this supposed to mean (to the user) that we think that this grid-ref is x% likely to be the place mentioned in your text? – in which case how will it be calculated?)

I’ve not assessed any of the output formats because most times when I attempted to move from the summary page to the output the system seems to hang. It has worked on occasion, but I’ve never saved the results.

5. Overall conclusions

The overall results of this test was reasonably encouraging, especially as the sample text was historical and thus biased, and the return rate was acceptable.

What was not clear at the outset was whether the parser really wanted to have a list of place names submitted to it rather than a long text, which is what I expected. The user-requirements may need to be reassessed in this light.

The most worrying element of this exercise was the parser’s inability to identify obvious places such as Plymouth or York. For current material one might at least add parliamentary constituencies from the ONS IPN (Index of Place Names) or Local Authorities from the same source.

More stop words should probably be introduced, and a greater awareness of capitalization might prevent some of these false positives. The misses (and low frequency hits) will presumably be included within the base gazetteer at some later stage. (I guess that each file submitted to the Parser is kept somewhere at Edinburgh, along with the results. Someone could be kept busy analyzing these results, and systematically adding more material!)

What was probably not acceptable, was the large amount of variations within the results. As an example, the term Avon was returned five times. While there is an acceptable match by term there is not one for meaning. I have not calculated the number of these ‘incorrect’ matches, but it would be vital for the browser to contain information about what is being returned and what the base information is. Both naïve and professional users will find this (presumed) lack of accuracy tedious.

I am not suggesting that what is being returned is wrong per se, because it isn’t, but the usefulness to the user is downgraded by not taking into account the context of the material. For this particular source, Avon could not be a county or a ward, and

given its proximity in the text to another river one might assume that it is a river (this argument doesn't really wash because Thames is not identified by the parser either).

Some recommendations

1. Assess errors caused by capitalization in texts.
2. Add more stop words to the dictionary/gazetteer to prevent possible errors.
3. Integrate more 'fuzzy' stop or match rules for places which may also be names or for ambiguous entries in the gazetteer, i.e., the same name may refer to many different places.
4. Evaluate the possibility of excluding (again, perhaps, via stop-words) foreign places, and explicitly tell the user about what is being parsed here. (e.g., Alexandria.)
5. Increase the number of places in the gazetteer to include other smaller islands around the British Seas, names of 'historical', i.e., pre-1974 counties (and standard abbreviations), and some of the larger cities which are not currently within the named administrative geography of the country.
6. Document the processes which are at work here -- presumably for both "naïve" and "professional" user, i.e., make it clear what the system is able to do, for example, if countries are likely to be a problem to identify, e.g., Wales or Ireland then this should be explicit. (These are both identified with itty-bitty places.)
7. In the text version of the hits, it would be more convenient to see the place names and the distinguishing type for that place in a separate column of the table.
8. Implement a user-driven filter after the summary results before the "output" stage.
9. Fix the certainty filter and make the users aware of what it is for.
10. Fix the user-interface for the certainty filter.

Matthew Woollard

2 August 2004

Complete list of low confidence place name matches from this sample.

This list is included to highlight the range of places which while being identified as places have no grid-reference attached to them.

Africa	Merchant
Asia	Merionethshire
Australia	Merthyr Tydfil
Bethnal Green	Middlesex
Bombay	Montgomeryshire
Book	Morecambe Bay
Brecknockshire	North America
Brest	North Foreland
Bristol Channel	North Wales
Britain	Old Sarum
Bury St. Edmunds	PARISH
Cardiganshire	PEOPLE
Carnarvonshire	Persia
Channel Islands	Portsea Island
Clyde	Radnorshire
County of England	river Elbe
County of Surrey	river Lea
Crystal Palace	Royal
Cumberland	Russia
Dee	Rutlandshire
Dorsetshire	Sardinia
East India	Sark
East Indies	Saxony
England	SECRETARY OF STATE
England. ¹	Sheriff of Linlithgow
Europe	Somersetshire
European	Son
Frith of Solway	South Africa
Glamorganshire	South Eastern
Great Britain	South Shields
Greece	South Uist
Greenwich Pensioners	South Wales
Guernsey	St Kilda
Hamlets	St. Albans
Hants	St. Kilda
Holy Island	The Cheviot Hills
HOME DEPARTMENT	Tweed
Husband	Tything
Indian Navy	Tythings
Islands	Ulverstone
Isle of Man	West Derby
Isle of Wight	West Riding
Jersey	Westmorland
Lowestoft Ness	Wife
MACHINERY	Wisbeach
Madras	

¹ Query: why is this the only place which has been included with a full stop?