Geoparsing Historical and Contemporary Literary Text set in the City of Edinburgh

Lang Resources & Evaluation
https://doi.org/10.1007/s10579-019-09443-x

ORIGINAL PAPER

Geoparsing historical and contemporary literary text set in the City of Edinburgh

Beatrice Alex[1,2] · **Claire Grover**[2] · **Richard Tobin**[2] · **Jon Oberlander**[2]

Abstract While a reasonable amount of work has gone into automatically geoparsing text at the city or higher levels of granularity for different types of texts in different domains, there is relatively little research on geoparsing fine-grained locations such as buildings, green spaces and street names in text. This paper reports on how the Edinburgh Geoparser performs on this task for different types of literary text set in Edinburgh, the first UNESCO City of Literature. The non-copyrighted gold standard datasets created for this purpose are released along with this article.

Keywords Geoparsing · Geotagging · Georesolution · Fine-grained place names · Mining literary text · Edinburgh gazetteer

This article is dedicated to our colleague and friend Professor Jon Oberlander.

✉ Beatrice Alex
balex@ed.ac.uk

[1] School of Literatures, Languages and Cultures, Edinburgh Futures Institute, 50 George Square, Edinburgh EH8 9LH, UK

[2] Institute for Language, Cognition and Computation, School of Informatics, Crichton Street, Edinburgh EH8 9AB, UK

Published online: 26 February 2019

1 Introduction

This article presents work on fine-grained geoparsing carried out as part of the Palimpsest project on mining historical and contemporary literary texts set in Edinburgh.[1] During this project we adapted the Edinburgh Geoparser,[2] a tool which is used to geoparse text, to literary text containing fine-grained place names located in and around Edinburgh. By fine-grained we mean locations at a lower level of granularity than the city or village level, such as names of streets, buildings, monuments or parks.

The output of the Palimpsest project is accessible in a web-based map interface called LitLong[3] as well as via the LitLong:Edinburgh iOS app[4] (Loxley et al. 2018). Both interfaces display the geoparsed literature by allowing users to browse literary excerpts containing Edinburgh-based locations, search by author, gender and date, and to create literary paths through the city (see Figs. 1 and 2).

The Edinburgh Geoparser has been applied to historical text in the past (Grover et al. 2010; Alex et al. 2015). However, up until Palimpsest it was set up to geoparse up to the city or village level, but did not attempt to locate places at a lower level of granularity. In Palimpsest the geoparser was extended to geoparse fine-grained place names like street names, names of buildings, parks or monuments within the Edinburgh area. As is common for other geoparsing tools and to clarify the terminology used throughout this paper, its geoparsing process is made up of two steps: geotagging and a georesolution. Geotagging involves identifying place names mentioned in the text and georesolution refers to resolving them to geographical coordinates. This paper presents results obtained in experiments evaluating the performance of the adapted geoparser for both steps.

After reporting on the background for Palimpsest, related projects and, in particular, related geoparsing evaluation work in Sect. 2, we will give an overview of the existing Edinburgh Geoparser and will explain how it was adapted to geoparse fine-grained place names (see Sect. 3). As a gold standard, we used three types of literary text for evaluating the adapted geoparser:

- historical, raw optically character recognised (OCRed) text,
- historical, manually crowd-corrected OCRed text and
- contemporary born electronic text.

Section 4 explains how these datasets were selected and prepared, provides counts on the geoparsing annotation and reports inter-annotator agreement (IAA). We will then explain in detail how the Edinburgh Geoparser preformed on this data both for identifying location mentions in text and for georesolving them to latitude and longitude coordinates in gazetteers (see Sect. 5). An additional contribution along

[1] https://www.ed.ac.uk/literatures-languages-cultures/english-literature/research/palimpsest.

[2] www.ltg.ed.ac.uk/software/geoparser.

[3] www.litlong.org.

[4] https://itunes.apple.com/gb/app/litlong-edinburgh/id1004433531?mt=8.

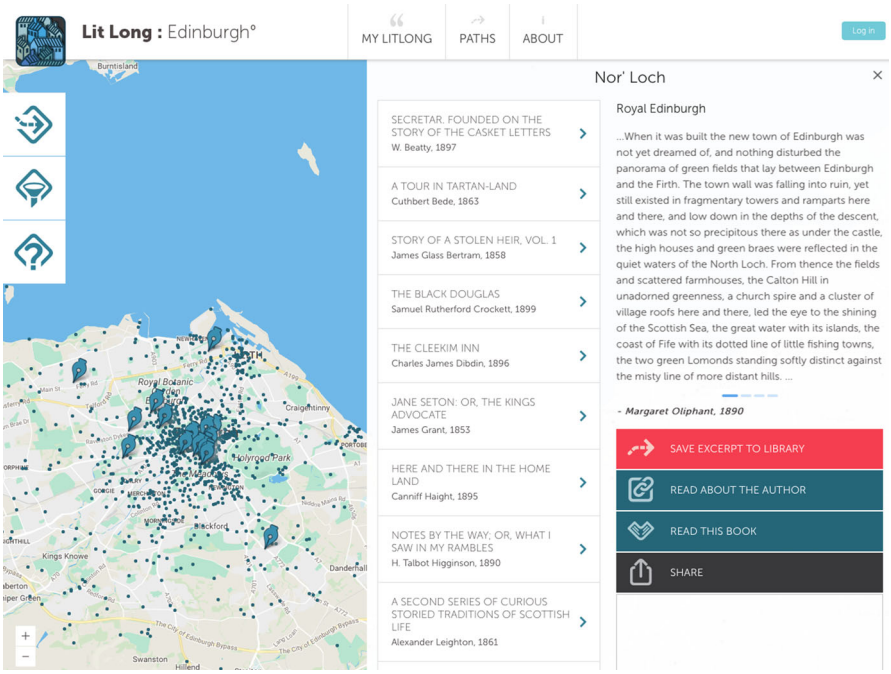

Fig. 1 The LitLong web interface at www.litlong.org

with this paper is the release of the non-copyrighted gold standard data used in the evaluation presented.

2 Related work

While there have already been many digital literary mapping projects, in this section we will present the most relevant ones as well as existing studies evaluating geoparsing literary text or comparing geoparsers. We refer readers interested in other literary mapping work to an extensive list of 74 projects recently reviewed by Luchetta (2017).

The idea for Palimpsest arose out of an initial prototype which visualises a small set of excerpts set in Edinburgh and manually collected by literary scholars at the University of Edinburgh, a project which was initiated by the literary scholar Dr. Miranda Anderson.[5] Related smaller-scale endeavours have relied on the collection of titles or passages by a few individuals or via crowd sourcing (e.g. Edinburgh Reads[6] created by Edinburgh Libraries. A larger crowd sourcing effort is that of the

[5] http://palimpsest-eng.appspot.com.

[6] https://www.google.com/maps/d/viewer?mid=1hkhLnpdhJEkiKa67NmkCTGxA7MY&hl=en_US&ll=55.94115059499049%2C-3.252396699999963&z=11.

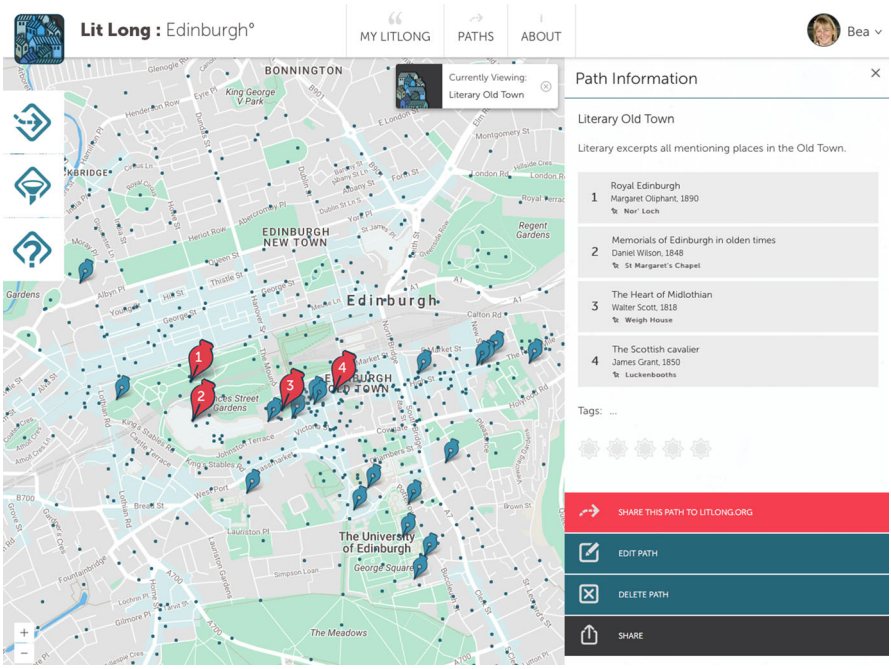

Fig. 2 A literary path created on LitLong

Global Book Map.[7] It currently contains 142,917 books for 18,103 locations worldwide and this data is being collected either by users adding and mapping books manually on the document level or by exploiting metadata from Open Library[8] or from LibraryThing.[9] The Global Book Map currently contains 393 entries for *Edinburgh* but does not map to fine-grained locations within the city not does it allow users to browse excerpts containing such locations. This functionality is made possible by Palimpsest as LitLong users are able to read through the context of location mentions and thereby immerse themselves in the literary landscape of Edinburgh.

There have been numerous literary city or area mapping projects involving fine-grained locations including the Mapping of St.Petersburg[10] primarily achieved via manual analysis combined with online mapping of places in St.Petersburg in works by Dostoevsky and Gogol. There is also the Literary City site[11] presenting a map of literature set in San Francisco. While there is no explanation on how the location

[7] http://www.mappit.net/bookmap.

[8] https://openlibrary.org.

[9] https://www.librarything.com.

[10] http://www.mappingpetersburg.org.

[11] http://www.sfchronicle.com/theliterarycity/.

information was derived for the latter site, we presume it was also created manually. The Literary Atlas of Europe[12] is an interdisciplinary research collaboration between literary scholars, cartography and visualisation experts mapping literature set in three distinct regions (Prague, Nordfriesland and Vierwaldstättersee). As far as we understand, its geoparsing work was also done manually involving literary experts identifying locations in text and resolving them to latitude and longitude coordinates. The advantage of such manual work is that it is very accurate. However, it is very time-consuming to create as individuals have to identify pieces of literature, mark up all the locations within them and disambiguate them by assigning latitude/longitude pairs or linking them to gazetteer entries. As a result, such efforts tend to focus on a few select pieces of literature or do not attempt to geoparse entire literary works. In comparison, the aim of Palimpsest was to geoparse the full text of a comprehensive set of literature set in Edinburgh which is why the Edinburgh Geoparser was employed to assist with this work even though it is not 100% accurate. In the LitLong interface users are now able to alert the team when they spot a geoparsing mistake and in future iterations such error notifications can be fed back into our tools to improve geoparsing accuracy. This goes beyond the argument made by Solina and Ravnik (2010) of employing automatic methods for geoparsing literature where possible and combining them with human selection and interpretation (Solina and Ravnik 2010) as human input can be exploited to optimise the technology as well.

One of the most well-known projects applying automatic geoparsing to literature is the Mapping the Lakes project led by Prof. Ian Gregory. His group adapted the Edinburgh Geoparser to do this work (Cooper and Gregory 2011) and have presented precision and recall scores for geotagging of 91.6 and 74.4 depending on the type of gazetteer used (Rupp et al. 2013). Alves and Queiroz (2015) report on the mapping of Portuguese literature as part of the project LITESCAPE.PT - Atlas of Literary Landscapes of Mainland Portugal (Alves and Queiroz 2015). Their paper refers to "distant reading" being employed as part of their methodology. They mention that they are exploring the use of Portuguese computational linguistics tools but do not go into detail on their exact methods and how well they perform.

The most relevant work in terms of evaluation of fine-graining geoparsing of literature is that of Moncla et al. (2017). Their paper evaluates the first step in the process (geotagging), and their analysis is limited to recognising Paris street names in French literary text. They report on two methods for identifying street names in 31 French novels and measure performance in terms of precision, recall and balanced F1-score, metrics which are also used in this paper. While both methods score high at 0.98 and 0.99 F1, boundary errors are reported but not included in their calculations. Recognising street names is a relatively easy task given that they are often signalled by the occurrence of the words *rue*, *avenue* or *boulevard* and these results support this claim. Furthermore, they carry out manual correction of OCR errors caused by the digitisation process. In this paper we do not differentiate between different types of fine-grained locations as the Edinburgh Geoparser aims to identify all of them. We report geotagging performance using a strict measure of

[12] http://www.literaturatlas.eu/en.

F1 (used in the CoNLL 2002 competition for recognising named entities (Tjong Kim Sang 2002), see Sect. 5) which includes all boundary errors and therefore counts them both as false positives and as false negatives. We also do not correct the original text nor any of the processing steps prior to geotagging.

Earlier work by Moncla and his collaborators involved geoparsing hiking descriptions containing fine-grained toponyms such as names of churches, cottages, hamlets and lakes (Moncla et al. 2014). Their NER tagging method is similar to that employed by the Edinburgh Geoparser in Palimpsest (combining rules with lexical lookup) and scores very high when applied to a gold standard which was hand-corrected for part-of-speech (POS) tags. Performance suffers by up to 15% when automatic POS tagging is used. Their main contribution in terms of georesolution is a method for resolving location mentions not found in the gazetteer to geographical areas instead of precise points. It lends itself well for hiking descriptions as location mentions tend to follow a path and are in close proximity to each other restricted to a relatively small area. This means that location mentions contained within the gazetteer can be used to constrain the area of those not found. This method would potentially work well for travel literature, for example, where the author describes a walk through the city, but would be less successful for resolving place names in other types of literary works where they are used to describe the location of a character or to set the scene of a plot.

There is also existing research on spacial uncertainty of locations mentioned in literature (Reuschel and Hurni 2011). It mainly focusses on the visualisation of vague place names as they do not tend to have concrete boundaries. When geoparsing text containing such names (e.g. the area *Leith* in Edinburgh) gazetteers often do not distinguish between them and locations with concrete latitude and longitude coordinates. The Edinburgh Geoparser processes vague locations (as long as they are named) in the same way as other place names and their georesolution is largely dependent on their gazetteer entries.

More recently researchers have started to publish comparisons of different geoparsers to determine the shortcomings of such systems and possible routes for future work in this area. For example, Gritta et al. (2017) compared five systems (GeoTxt[13] (Karimzadeh et al. 2013), Yahoo! PlaceSpotter[14], CLAVIN[15], Topocluster (DeLozier et al. 2015) and the Edinburgh Geoparser) on two contemporary English datasets (Wikipedia pages and news articles) in terms of their geotagging and georesolution performance as well as their speed. The authors argued that a geoparser must perform well on all three aspects and concluded that, albeit not performing perfectly, only the Edinburgh Geoparser managed to do so. They also provided an extensive error analysis and discussed ways in which geoparsers can be improved which is extremely useful for those working on this kind of technology.

[13] http://www.geotxt.org/api/.

[14] https://developer.yahoo.com/boss/geo/docs/key-concepts.html.

[15] https://clavin.bericotechnologies.com/about-clavin/.

3 Adapting the Edinburgh Geoparser

For Palimpsest, we adapted the existing default version of the Edinburgh Geoparser to process literary text set in and around the City of Edinburgh and geotag and georesolve local Edinburgh-specific place names. We already have experience in adapting this geoparser to historical text (Alex et al. 2015) but this was the first time we modified it to work with literary text, both historical and contemporary. The default geoparser contains the two main components shown in Fig. 3: a text mining pipeline for recognising place names (geotagging) and other entities in text and a geographic ambiguity resolution (georesolution) component which chooses between competing interpretations of place names (i.e. different geographic coordinates) given their textual context. Both components make extensive use of place name gazetteers, the selection of which depends on the geoparsing task at hand and the type of data to be processed.

3.1 The Edinburgh gazetteer

There was no freely available gazetteer which includes all the Edinburgh place names, the use of which we wanted to capture in Palimpsest. These place names have different granularity, ranging from area names (*Portobello*, *Cramond*), through street names (*The Royal Mile*, *Cockburn Street*) to open spaces (*The Meadows*, *Princes Street Gardens*), buildings (*Craigmillar Castle*, *Holyrood Palace*), statues and monuments (*Greyfriars Bobby*, *The Scott Monument*) etc. Therefore, a prerequisite for geoparsing was to create an Edinburgh gazetteer by aggregating information from a variety of different sources. For street names we used the Ordnance Survey's OS Locator (OSL) data,[16] for building-level information we used the Canmore site records database[17] from the Royal Commission on the Ancient and Historic Monuments of Scotland (RCAHMS) which is now part of Historic Environment Scotland.[18] For other information, ranging from area names through to pub names, we used an Edinburgh subset of Open Street Map (OSM).[19] The aim was to create a gazetteer which could be used both as a place name lexicon when identifying potential place names in text during geotagging and as a gazetteer for georesolution, i.e. assigning latitude/longitude coordinates to geotagged place names.

The aggregation process involved converting records from all three sources into one common XML format followed by a data clean-up stage to discard records which might trigger faulty geotagging of place names in text. For example, Canmore contains records for places with generic names such as *Station House* or *Barracks*, as well as records for residential houses with names such as *Bonny Views*. OSM contains records for numerous modern-day businesses such as *Bay of Bengal* (a restaurant) and *Blossom* (a guest house). We attempted to exclude records such as

[16] https://www.ordnancesurvey.co.uk/business-and-government/products/os-locator.html.

[17] https://canmore.org.uk.

[18] https://www.historicenvironment.scot.

[19] https://www.openstreetmap.org.

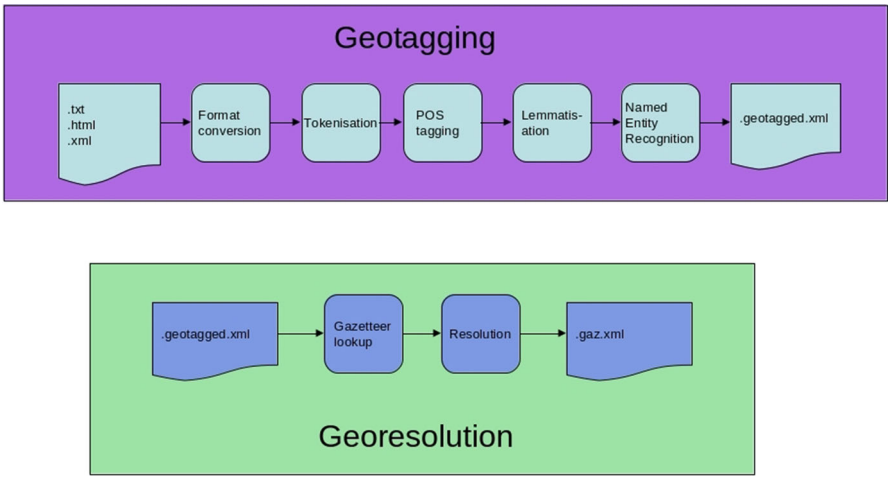

Fig. 3 Overview of the Edinburgh Geoparser pipeline made up of a geotagging and a georesolution component

these semi-automatically. The final gazetteer still contains many rather questionable Edinburgh place names, e.g. *Alpine Garden* (part of the Royal Botanic Gardens Edinburgh), *Beach House* (generic descriptor) or *The Waiting Room* (a pub). The presence of this kind of record in the gazetteer, however, does not seem to have had too deleterious an effect on geotagging and subsequent georesolution.

There are some place names which occur in the Palimpsest works for which none of the three sources has a record. These are mostly historical forms of modern place names or spelling variants (for example, *Cowgate-port*, *Nor' Loch* or *Edinboro*). For cases where such an omission has been observed, we have manually added appropriate records. We added 25 records by hand collected as a result of manual error analysis and 92 records as alternative names for locations which were suggested by literary scholars involved in the project.

The final version of the Edinburgh gazetteer contains 13,064 records corresponding to 10,204 unique place names. Listing 1 shows some example records in the gazetteer. The source of each record is stored in the `source` attribute (as either `rcahms`, `osm`, `osl`, `byhand` or `altnamelist`). Note, we did not eliminate duplicate entries of places from different sources with slightly different coordinate values (e.g. see *Oxford Bar*).

```
<gazetteer>
...
<place name="Nor' Loch" lat="55.950135" long="−3.200252" source="byhand"
id="pg5"/>
...
<place name="Adam House" lat="55.948109" long="−3.187424" source="rcahms"
id="pg144"/>
<place name="Adam Smith Statue" lat="55.9497628" long="−3.1900024" source="osm"
ptype="historic" id="pg145"/>
...
 <place altfor="Edinburgh" long="−3.192704" lat="55.949428" name="Edinboro"
source="altnamelist" id="pg4070"/>
...
<place name="Hunter's Close" lat="55.9476" long="−3.1945" source="osl" id="pg5825"/>
...
<place name="Oxford Bar" lat="55.9529618" long="−3.2047389" source="osm"
ptype="amenity" id="pg8518"/>
<place name="Oxford Bar" lat="55.952983" long="−3.204677" source="rcahms"
id="pg8519"/>
...
</gazetteer>
```

Listing 1 Example location entries in the Edinburgh gazetteer.

Efforts to build a community and infrastructure for linked open geo-data such as Pelagios Commons[20] and the working group Linked Pasts[21] advocate the use of semantic web technology to create linked open geo-historical gazetteers which can capture alternative names, vagueness and changing location of places. In practice we found that the existing gazetteer resources related to Edinburgh were all in different formats, only some of them openly available and none of them were linked to each other. Alternative names were not usually recorded and many questionable names were listed without context which confused the geoparser. This is why the clean-up stage was so important. Equally, the project only allowed a limited amount of time for this clean-up which is why we employed semi-automatic methods instead of carefully curating an Edinburgh gazetteer. Converting the final, combined Edinburgh gazetteer to one linked open dataset was unfortunately not an option as not all of the original sources were openly available.

3.2 Geoparsing

The Edinburgh Geoparer's text mining pipeline first converts an input text into a common XML format and then each stage of processing incrementally adds annotations to the mark-up (see Fig. 3). First the text is segmented into paragraphs which are tokenised to add word and sentence elements. Words are then part-of-speech tagged using the C&C POS-tagger (Curran and Clark 2003) and lemmatised using Morpha (Minnen et al. 2000). Subsequently, Named Entity Recognition (NER) is performed using hand-written rule sets combined with lexical look-up. For place name recognition (geotagging), extensive lexicons of place names both from

[20] http://commons.pelagios.org.

[21] http://linkedpasts.org.

the UK and the rest of the world are used. The choice of gazetteer depends on the particular type of textual data processed and the geoparsing task at hand.

In the Palimpsest system, this stage is augmented to include a lexicon of Edinburgh place names derived from the Edinburgh gazetteer. Look-up in the Edinburgh lexicon precedes look-up in the other place name lexicons that are included in the default Edinburgh Geoparser. Otherwise, the processes are the same as in the distributed default version where lexical look-up is combined with context-sensitive rules to identify entity mentions in the text and disambiguate entity types (Grover et al. 2010). This works by first performing lexical lookup against a series of lexicons and adding attributes to the XML elements of phrases matched in text. The lookup works by preferring longer matches over shorter ones (e.g. *Princes Street Gardens* is matched instead of just *Princes Street*). Ambiguities between entity types are resolved after lookup using rules (e.g. the preposition *in* before *Deacon Brodie's* suggest that it is a location, in this case a pub, rather than a person name).

The output of the text mining pipeline contains named entity annotations for person and place names as well as dates. This is the input to the georesolution step which looks up place names in one or more gazetteers. Candidate matches are ranked to arrive at the most probable interpretation given the context of the document. In Palimpsest, look-up in the Edinburgh gazetteer precedes look-up in more general Ordnance Survey[22] and GeoNames[23] gazetteers. Ranking uses heuristics combined with weighting of information such as geographic feature and size. We assume that a degree of geographic coherence holds within documents in that the relevant text is more likely to mention many places in a single area rather than a set of geographically unrelated places. To model this, proximity between gazetteer records for all the places mentioned in the document is strongly weighted to ensure that all locations mutually constrain one another to be as close together as possible. Thus the highest ranked interpretation of *Haymarket* will be the one in Edinburgh in a document containing many Edinburgh or Scottish place names and the one in London in a document with more London-based or English place names. More details on the ranking of location candidates can be found in Grover et al. (2010), Alex et al. (2015) as well as in the documentation of the Edinburgh Geoparser.[24]

The georesolution results are added as XML annotations along with their immediate context of each place name. This information is used for display in the Palimpsest interfaces. We call the context surrounding a georesolved Edinburgh place name mention a Palimpsest snippet. In the final system implementation, we set this context to be the sentence containing the location as well as the previous and the following sentence without crossing paragraph boundaries.

The Palimpsest snippets are also ranked by an 'interestingness'-score (i-score). This was inspired by work on automatic prediction of text aesthetics and interestingness (Ganguly et al. 2014). The aim is to rank snippets per document

[22] https://www.ordnancesurvey.co.uk.

[23] http://www.geonames.org.

[24] http://groups.inf.ed.ac.uk/geoparser/documentation/v1.0/html/index.html.

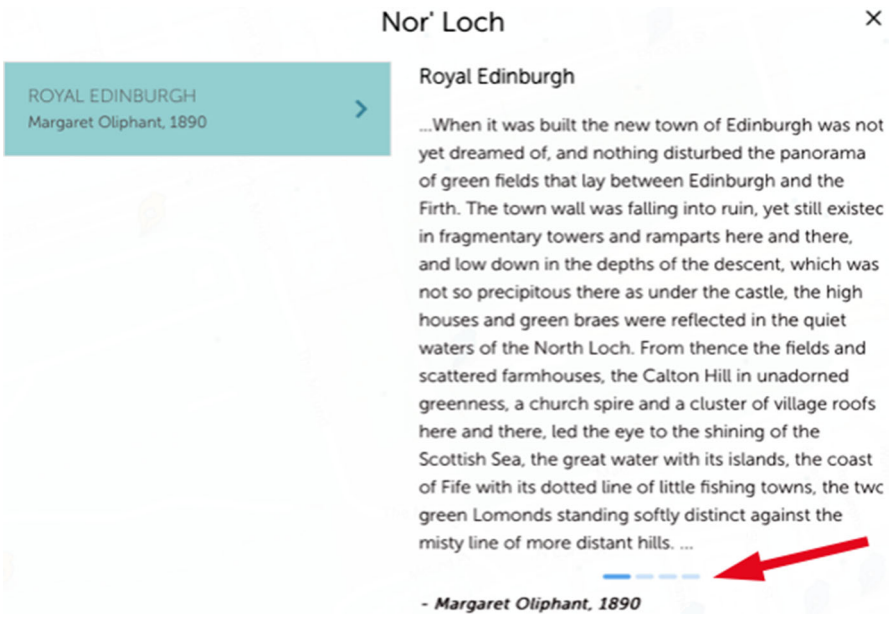

Fig. 4 One of four literary excerpts in Margaret Oliphant's *Royal Edinburgh* containing mentions of *North Loch*. The user can read them one by one by swiping left or right. The arrow points to bars which are used to visualise multiple excerpts

to give those snippets where the Edinburgh place name is not only a mention in passing more importance and therefore make them appear earlier on in the user interfaces. We compute this score by checking for a number of features, including snippet length, the presence of multiple Edinburgh-based locations in the snippet, the presence of at least one Edinburgh-based location (excluding variants of Edinburgh), an adjective or adverb appearing in the snippet, the presence of different forms of certain verbs (*be*, *do*, *say* or *go*) and word repetition within the snippet. The i-score is computed by treating each of the features equally and its value can range between 0 and 1 where 1 represents snippets for which all features apply and 0 those where none of the features apply. The idea is not to list snippets containing the georesolved locations in the order they appear in the literary work but to order them by 'interestingness'. The web-based user interface orders snippets by i-score but does not make this ordering apparent to the user. Figure 4 shows how multiple excerpts from the same work (in this case excerpts from Margaret Oliphant's *Royal Edinburgh* containing the place name *North Loch*) are displayed to the user (Oliphant 1890). The excerpt with the highest i-score is displayed first and the others can be browsed by swiping left or right. The i-score computation is preliminary work and still requires feature development and optimisation which is beyond the scope of this paper but it is mentioned here to provide some context.

The output of the Edinburgh Geoparser is fed into a database which serves as the input data for the user interfaces. It comprises of 546 literary works mentioning 1600 unique Edinburgh locations appearing in more than 47,000 literary excerpts. The literary source data which is part of this Palimpsest corpus is described in more detail in the next section.

We have described the creation of a historically-informed local-level gazetteer for Edinburgh, the product of significant amounts of processing of the source data, as well as the Edinburgh Geoparser and any changes that were made to it for the Palimpsest project. If we were to port our pipeline to a new city, we would need a new local-level gazetteer. We often get asked how much work would be involved in creating such a resource for a different city and how long it would take to port the Edinburgh Geoparser to process text centred around that city. These questions are difficult to answer because it depends largely on the available gazetteer resources. This will get easier as more local gazetteers are being made available. However, especially when combining different gazetteer resources for one city, a clean-up stage is unavoidable.

While not as extensive as the Edinburgh gazetteer, we have successfully created local gazetteers for Aberdeen and Dundee in a recent project where we applied a text mining pipeline similar to the one described here to Scottish historical newspapers from the nineteenth and early twentieth centuries. The basis for these local gazetteers was the open source resource, OS Open Names[25] which was made available after the Palimpsest research project was completed. It supplies "a comprehensive dataset of place names, roads numbers and postcodes for Great Britain". We were able to use initial characters of postcodes, i.e. *DD* for Dundee and *AB* for Aberdeen, to extract all the place names and their coordinates from the OS data covering the larger area around each of the cities. We then used an appropriate bounding square to identify those place names which are actually within the city. The programming involved in accessing relevant entries was quite simple and only required some format conversion to create gazetteer entries suitable for use with our pipeline. This technique has allowed quite robust local-level geoparsing of historical newspapers and can be replicated for document collections relating to places anywhere within Great Britain. Rapid porting to place names in other countries would require resources similar to OS Open Names.

Creating a new gazetteer and making changes to the geoparser is feasible by other researchers external to the core development team. In fact, the tool is already used in several research collaborations in the area of DH and very detailed documentation on its various components and their usage is available. The first author of this paper has also published a Programming Historian lesson on how to get started using the Edinburgh Geoparser (Alex 2017) on behalf of the team who are also available for further technical support in cases where that is needed.

[25] https://www.ordnancesurvey.co.uk/opendatadownload/products.html#OPNAME.

4 Data and annotation

The Palimpsest corpus is made up of 503 historical, out-of-copyright works of literature and 43 works written by contemporary authors, all of which were geoparsed to identify fine-grained Edinburgh-specific locations and other place name mentions and their latitude and longitude coordinates.

4.1 Literary source data

The pool of historical Edinburgh-specific documents was collected via a semi-automatic information retrieval method where automatic location-based information retrieval was combined with a manual curation cycle, a process we refer to as assisted curation (Alex et al. 2017). The information retrieval was run over large literary document collections from different data providers (including all worldwide public domain material from HathiTrust,[26] the British Library Nineteenth Century Books collection,[27] all public domain English Project Gutenberg books,[28] the Oxford Text Archive data[29] and works obtained from the National Library of Scotland[30]). The contemporary works were specifically chosen by literary scholars as they are known to be good examples of literature set in Edinburgh. For the processing of the latter, we obtained permission from authors and publishers.

To give an example, Fig. 5 provides an excerpt from a book in the British Library Nineteenth Century Books collection which illustrates how Edinburgh locations can be used to set the scene (Hibbert-Ware 1883, p.248). Some place names are used to describe merely where something takes place (i.e. *Tam Neil had left the tavern in Libberton Wynd, he, in company with his apprentice Jock, was walking at a smart pace along the road to Duddingston village, ...*). Libberton Wynd was a steep street in Edinburgh which no longer exists today, and had in fact disappeared when this story was published. It got demolished around the time when George the IV Bridge was built at the beginning of the 1830s. Other place names are used when describing the mood of a story (e.g. *The moon had risen, and in its pure, pale light Salisbury Crags, a bold and lofty amphitheater of precipitous rocks, stood out clearly defined against the starlit sky ...*).

4.2 Creating the gold standard

To examine the effect which text quality and publication date have on geoparsing performance, the experiments described in the next section were carried out for books from three of the collections that were processed in Palimpsest, two of the historical text collections as well as the contemporary books:

[26] https://www.hathitrust.org.

[27] http://www.bbk.ac.uk/lib/elib/databases/arts/nineteenth-century-books.

[28] http://www.gutenberg.org.

[29] https://ota.ox.ac.uk.

[30] http://nls.uk.

CHAPTER XIV.

THE BASS FIDDLE.

About twenty minutes after Tam Neil had left the tavern in Libberton Wynd, he, in company with his apprentice Jock, was walking at a smart pace along the road to Duddingston village, the former carrying on his shoulder some object in a bag, which, though somewhat bulky, was of no great weight, and the form of which appeared to be oblong.

The moon had risen, and in its pure, pale light Salisbury Crags, a bold and lofty amphitheatre of precipitous rocks, stood out clearly defined against the starlit sky, Tam and his companion, striding along in the King's Park, being the only moving objects

Fig. 5 Literary excerpt from Mary Clementina Hibbert-Ware's book called *His Dearest Wish* (1883, vol. 2, p. 248)

- OCRED: historical OCRed texts from the British Library Nineteenth Century Books collection containing OCR errors,
- CORRECTED: historical OCRed texts from the Project Gutenberg collection which were crowd-corrected by hand, and
- MODERN: contemporary (born digital) pieces of text set in Edinburgh, including works by authors like Muriel Spark, Irvine Welsh, Alexander McCall Smith and Doug Johnstone.

We carried out the same experiments (see Sect. 5) for all three sets to show how geoparsing performance varies across these data types. To do this type of evaluation we required a gold standard for each dataset, i.e. a sample manually annotated for

place names and their latitude and longitude coordinates. As the gold standard was created after the tuning of the Edinburgh Geoparser was completed and not used for system development, they can be considered as unseen test data.[31]

To prepare the gold standard, we selected a small sub-set of each of the three collections (approximately 2.5% per collection). We did that by splitting each document into small 5000-byte chunks and selecting a sub-set of chunks per document, without guaranteeing that a chunk will contain a location. This type of random sampling assures that the distribution of locations in the gold standard remains representative of each collection.

Each collection of text chunks was then manually annotated for location mentions and latitude/longitude coordinates to create a gold standard used in the experiments described in Sect. 5. This was done by linguistically trained annotators in two stages. Firstly, all location mentions were annotated using the Brat annotation tool (Stenetorp et al. 2012).[32] Annotators were instructed to mark up all place names (fine-grained or not), including vernacular or made up names. Each annotated location mention could be given an optional "Edinburgh-specific" attribute to distinguish it as being a place located in Edinburgh or in its close surroundings. The annotation guidelines specified to mark up a location as Edinburgh-specific if it occurred within the Lothian area as defined on Wikipedia.[33]

The gold location mentions were then annotated further with latitude and longitude coordinate information. This was done using the Edinburgh Geo-annotator, a web-based georesolution annotation and evaluation tool which we developed in-house (Alex et al. 2014). This tool includes a map-based annotation interface which lets annotators select candidate pins on a Google map. For each location mention appearing in the text, annotators were able to choose between different competing candidates occurring in the gazetteers used by the Edinburgh Geoparser for georesolving the Palimpsest datasets. Locations not present in the gazetteer were marked as "not found".

Some figures regarding the number of document chunks and locations in each gold standard collection can be found in Tables 1 and 2. The OCRED gold data, the largest gold dataset with 250 document chunks, contains an average of 12.2 locations per chunk. The CORRECTED and the MODERN texts contain fewer but roughly the same number of document chunks (78 and 80, respectively). The CORRECTED chunks contain on average 9.9 locations, whereas the MODERN text data contains only on average 5.5 locations per chunk. When examining the Edinburgh-specific locations only, the MODERN texts contain on average 1.7 locations per chunk, whereas the historical collections contain only 1.1 or 1.2 locations per chunk. While the historical text collections are more dense in location mentions overall, the modern works contain by far the largest percentage of Edinburgh-specific locations (31.7% of location mentions are based in Edinburgh). This is unsurprising as they

[31] The OCRED and CORRECTED gold data sub-sets are made available at https://github.com/LitPalimpsest/Palimpsest. The MODERN data is under copyright restrictions and we only have permission to make its geoparser output available via the LitLong interfaces.

[32] http://brat.nlplab.org.

[33] https://upload.wikimedia.org/wikipedia/commons/8/83/The_Lothians.png.

Table 1 Number of document chunks (5000 bytes each), place names and Edinburgh-specific place names in each gold standard collection

Dataset	Chunks	All place names		Edinburgh place names		
		Total	Avg.	Total	% of all	Avg.
OCRED	250	3039	12.2	283	9.3	1.1
CORRECTED	78	770	9.9	92	12.0	1.2
MODERN	80	438	5.5	139	31.7	1.7

Avg. refers to average per chunk, *% of all* means percentage of Edinburgh-specific place names over all place names

Table 2 Location ratio (LR) of unique over all location mentions for all and Edinburgh-specific place names in each gold standard collection

Dataset	Place names	Unique place names	LR
All place names			
OCRED	3039	1473	0.48
CORRECTED	770	425	0.55
MODERN	438	231	0.53
Edinburgh place names			
OCRED	283	125	0.44
CORRECTED	92	42	0.46
MODERN	139	81	0.58

are well-known examples of Edinburgh-specific literature. The historical texts were retrieved in a semi-automatic fashion which resulted in the discovery of less well-known works, and in ones containing on average less Edinburgh-specific locations (see Table 1).

Table 3 lists some examples of Edinburgh-specific place names in the gold standard. The most frequent mentions are well-known locations today, including one location nickname (*Auld Reekie*, another name for Edinburgh). Examples of infrequently used mentions of place names include names of areas, streets, buildings and establishments which still exist today (e.g. *Drummond Street*, *Rutland Square* or *Balerno*). Others are names of places which have either disappeared altogether (e.g. *Calton Gaol*, an old prison on Carlton Hill now the site of St Andrew's House), which have changed their purpose (e.g. *House of Bruntsfield*, now part of James Gillespie's High School), which have had multiple locations (e.g. *Physic-garden*) or are now known under a different name (e.g. *Empire Theatre*, now called the Festival Theatre). There is also one example of a name with a spelling variation (e.g. *auld Toun* referring to Old Town) and two names containing OCR errors (*Sahsbury Crags* is referring to Salisbury Crags and *SdvermiUs* which should be Silvermills). Georesolution of such locations can be difficult, especially if names contain errors or spelling variants or if they have moved or disappeared over time.

Table 3 Most frequent and infrequent Edinburgh place names in the gold standard data

Most frequent place names		Least frequent place names	
Count	Place name	Count	Place name
134	Edinburgh	1	Drummond Street
11	Dalkeith	1	Rankeillor Street
9	George Street	1	Empire Theatre
8	Moray Place	1	Tron Kirk
8	Forth	1	Rutland Square
7	Holyrood	1	Potterrow
7	Canongate	1	Greyfriars Church
7	Arthur's Seat	1	St Cecilia's Hall
6	Scotland Street	1	Sandy Bell's Bar
6	Queen Street	1	House of Bruntsfield
6	Princes Street	1	Balerno
6	Leith	1	auld Toun
6	High Street	1	Calton Gaol
5	Edinburgh Castle	1	Physic-garden
5	Cowgate	1	Sahsbury Crags
5	Auld Reekie	1	SdvermiUs

4.3 Inter-annotator agreement

A small sub-part (10%) of each gold standard dataset was doubly annotated to determine inter-annotator agreement (IAA). IAA is measured for geotagging and georesolution by comparing the annotations of one annotator to those of another. This is done to gage how difficult it is for a person to geoparse a piece of text. It also helps to understand if the annotation guidelines are clear and gives us an idea of how well a machine might be expected to geoparse text automatically if it worked at human capacity.

4.3.1 Geotagging

The IAA of the location entity annotation is measured in precision, recall and balanced F1-score (see Table 4).[34] What figure constitutes precision or recall when computing IAA depends on which order the annotators are compared to each other and is therefore marked as `P/R` in our tables. The results show that IAA F1-scores for geotagging locations in text are high across all three collections (ranging between 0.96 and 0.98 in F1 for all locations and between 0.91 and 1 in F1 for Edinburgh-specific locations). Agreement is lowest (F1 = 0.91) for OCRED text for Edinburgh-specific place names.

Some of the location tagging disagreements are due to mismatching boundary annotations, e.g. *St. Provincial's* versus *St. Provincial*, *bay of Lochnannagh* versus

[34] We do not compute IAA for named entity annotations as Cohen's kappa scores because this metric was found to be inappropriate for this type of annotation as discussed in detail by Deleger et al. (2012).

Table 4 IAA for location mention annotation for all place names and Edinburgh-specific place names

We report number of true positives (TP), false positives (FP), false negatives (FN), precision or recall (P/R) and balanced F1-score (F1)

IAA for geotagging						
Dataset	TP	FP	FN	P/R	P/R	F1
All place names						
OCRED	283	12	12	0.96	0.96	0.96
CORRECTED	79	5	1	0.94	0.99	0.96
MODERN	31	0	1	1.00	0.97	0.98
Edinburgh place names						
OCRED	21	3	1	0.88	0.96	0.91
CORRECTED	7	0	0	1.00	1.00	1.00
MODERN	17	1	0	0.94	1.00	0.97

Lochnannagh or *Forth bridge* versus *Forth*. In some of these cases the boundary decision affects the georesolution coordinates so it is important to get the entity annotation as correctly as possible to avoid cascading errors. In a few rare cases, one or the other annotator forgot to annotate a location. In one of those cases, we noticed that the place name contains an OCR error (*Loch Raiiza* for *Loch Ranza*) and it is possible that this error contributed to the oversight. Overall, however, we can conclude that the annotation of location names is a relatively easy task for human beings to perform consistently but that it is marginally more difficult when annotating OCRED text.

4.3.2 Georesolution

We also measured IAA accuracy for georesolution by taking the gold location markup for 10% of each gold dataset and letting two annotators georesolve it independently to pins on the map corresponding to candidate entries matched in the gazetteers. Table 5 lists how many locations were resolved manually to latitude/longitude coordinates by both annotators (Pin selected), how many locations were not found in the gazetteers (Not in gaz) and for how many either annotator decided that none of the suggested pins were appropriate (None selected). Note that IAA accuracy scores are computed only for those locations for which a pin was selected. The other figures provide an insight into gazetteer coverage.

Table 6 shows the georesolution IAA measured in exact accuracy (Acc.) as a strict measure but also in accuracy at different distances in kilometers (A@n). Accuracy scores are determined by matching coordinate pairs at different distances (0 to 5 km). The reason for providing the relaxed accuracy scores is that in some cases the gazetteers contain duplicate entries for the same location. For example, the Edinburgh gazetteer contains two candidates for *Oxford Bar* (see Sect. 3.1). Each annotator can only select one of them as the gold annotation in the same way as the Geoparser only chooses a top candidate as its georesolution prediction. While the annotation guidelines say to choose the most central pin for a location (e.g. the most mid-way point for a street, or the pin closest to the middle of a park) sometimes it can be difficult to choose between competing candidates. This is why we believe

Table 5 Georesolution IAA counts for all locations and Edinburgh-specific locations

IAA georesolution counts			
Dataset	Pin selected	Not in gaz	None selected
All locations			
OCRED	295	77	35
CORRECTED	80	17	15
MODERN	32	8	0
Edinburgh locations			
OCRED	17	3	2
CORRECTED	5	0	2
MODERN	14	3	0

We report number of locations for which both annotators selected a pin (pin selected), number of locations not found in gazetteer (not in gaz) and number of locations for which either annotator did not select any of the pins on the map (none selected)

Table 6 Georesolution IAA figures for all locations and Edinburgh-specific locations

IAA georesolution accuracies					
Dataset	Acc. (%)	A@0.1 (%)	A@0.5 (%)	A@1 (%)	A@5 (%)
All locations					
OCRED	92.3	94.5	96.7	97.3	98.9
CORRECTED	93.8	93.8	95.8	100.0	100.0
MODERN	91.7	95.8	95.8	95.8	95.8
Edinburgh locations					
OCRED	94.1	94.1	94.1	100.0	100.0
CORRECTED	100.0	100.0	100.0	100.0	100.0
MODERN	85.7	85.7	85.7	85.7	100.0

We report exact accuracy in terms of lat/long coordinates (Acc.), accuracy at 0.1 km (A@0.1), at 0.5 km (A@0.5), etc. up to accuracy at 5 km (A@5)

exact accuracy at small distances (e.g. up to 1 km) to be a reasonable metric to consider when evaluating georesolution of fine-grained locations. We report accuracy at up to 5 km for information purposes.

The results show that IAA scores for georesolution of all place names are reasonably high, ranging between 91.7 and 93.8% exact accuracy across the three datasets and, for example, between 95.8% and 96.7% accuracy at a distance of 0.5 km. IAA figures are even higher for Edinburgh-specific place names for the historical data (94.1% and 100% exact accuracy). These findings support the hypothesis that on a more granular level within a city it is easier to disambiguate different place names as there is less ambiguity at least within the boundaries of a city. However, given that the number of Edinburgh-specific locations in the doubly

annotated data is fairly small, it is difficult to draw conclusions that would definitely hold true for larger datasets. The reason for the lower 85.7% agreement for the MODERN data for Edinburgh-specific locations is that one annotator chose the pin for *Edinburgh* (the populated place) from the Edinburgh gazetteer whereas the other chose the pin for *Edinburgh* (the populated place) from the more general Ordnance Survey gazetteer. Technically both are correct which is why at a distance of more than 1 km accuracy increases to 100%.

Overall, IAA figures for geotagging and georesolution are suggesting that they are both relatively easy tasks for human beings to perform given clear annotation guidelines and instructions. However, the figures also show that doing these tasks manually can lead to disagreements. This helps us to put the performance of a computer doing the same tasks automatically into perspective.

5 Automatic geoparsing

In this section we report the results for automatic geotagging and georesolution using the adapted Edinburgh Geoparser. We use the same evaluation metrics as those reported for the IAA measurements (precision, recall and F1-score as well as accuracy).

5.1 Geotagging

We firstly examine the geoparser's geotagging performance to understand how well it is able to recognise location mentions in text. We first compare a baseline, the performance of the default Edinburgh Geoparser[35] when used in combination with the GeoNames gazetteer (see Table 7) to that of the Palimpsest adapted geoparser described in Sect. 3 (see Table 8) for all place names present in the gold standard.

It is difficult to draw any meaningful conclusions from the baseline scores because the default geoparser was not designed for geoparsing literary text nor for geotagging and georesolving fine-grained place names. The slightly lower performance for the MODERN data (F1 = 0.61) is caused in part by its higher frequency of Edinburgh-specific locations. What is apparent however is that the process of adapting the Edinburgh Geoparser has paid off. The results show an increase in F1-score across all three datasets, ranging between 0.05 and 0.14 with the biggest improvement obtained for the MODERN data.

Geotagging performance of the adapted geoparser varies across the three gold standard sets. When looking at all location mentions occurring in the text, the tagger performs best on MODERN text ($F1 = 0.75$), worst on historical OCRED text (F = 0.68) and roughly in-between on historical CORRECTED text (F1 = 0.72). While precision scores are very similar across all three collections, the difference in F1-score is caused by the fact that recall scores vary considerably. This finding is partly in line with previous observations and experiments which have found that OCRed text has a negative cascading effect on natural language processing tasks (Kolak and Resnik

[35] This distribution can be downloaded at www.ltg.ed.ac.uk/software/geoparser.

Table 7 Baseline geotagging results for all place names using the default Edinburgh Geoparser

Default geoparser NER results						
Dataset	TP	FP	FN	P	R	F1
All place names						
OCRED	1795	854	1244	0.68	0.59	0.63
CORRECTED	471	199	299	0.70	0.61	0.65
MODERN	267	167	171	0.62	0.61	0.61

We report number of true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and balanced F1-score (F1)

Table 8 Geotagging results for all place names and Edinburgh-specific place names using the adapted geoparser

Adapted geoparser NER results						
Dataset	TP	FP	FN	P	R	F1
All place names						
OCRED	1780	410	1259	0.81	0.59	0.68
CORRECTED	487	106	283	0.82	0.63	0.72
MODERN	302	68	136	0.82	0.69	0.75
Edinburgh place names						
OCRED	164	49	119	0.77	0.58	0.66
CORRECTED	52	5	40	0.91	0.57	0.70
MODERN	103	8	36	0.93	0.74	0.82

We report number of true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and balanced F1-score (F1)

2005; Lopresti 2005, 2008b; Alex et al. 2012) and information retrieval (Hauser et al. 2007; Lopresti 2008a; Reynaert 2008; Gotscharek et al. 2011).

Overall, the geotagging scores seem low even for modern text. One major reason is that most of the geoparser adaptation effort was targeted towards aggregating an Edinburgh-specific location gazetteer, which means that there was less focus on tagging and resolving other locations correctly. When geotagging Edinburgh-specific locations, performance is much higher for the MODERN texts than for both historical text collections. Aside from OCR quality, historical language variations and the higher percentage of Edinburgh-specific locations in the MODERN text, the reason for this difference is also the fact that the Edinburgh-specific location gazetteer is made up of a series of modern gazetteer resources (including street names from OS Locator, buildings and monuments from RCAHMS and locations from OSM). Recall for MODERN text is higher as the gazetteer contains modern locations occurring in the Edinburgh area. A similar but more reduced effect can also be seen when examining the recall scores obtained for identifying all location mentions.

Table 9 Georesolution results for all locations and Edinburgh-specific locations

We report number of locations georesolved, number of locations not found in gazetteer (not in gaz) and number of locations for which the annotator was unable to select any of the pins on the map (none selected)

Adapted geoparser resolution stats			
Dataset	Georesolved	Not in gaz	None selected
All locations			
OCRED	1718	1001	320
CORRECTED	484	190	96
MODERN	316	96	26
Edinburgh locations			
OCRED	205	59	19
CORRECTED	59	20	13
MODERN	110	23	6

Table 10 Georesolution results for all locations and Edinburgh-specific locations

Adapted geoparser resolution accuracies					
Dataset	Acc. (%)	A@0.1 (%)	A@0.5 (%)	A@1 (%)	A@5 (%)
All locations					
OCRED	65.9	66.5	69.0	72.8	85.9
CORRECTED	66.3	69.0	73.3	76.4	86.2
MODERN	70.3	75.0	79.7	81.3	86.1
Edinburgh locations					
OCRED	84.9	88.8	93.2	94.6	96.1
CORRECTED	78.0	83.1	94.9	96.6	96.6
MODERN	70.9	82.7	90.9	95.5	98.2

We report exact accuracy in terms of lat/long coordinates (Acc.), accuracy at 0.1 km (A@0.1), at 0.5 km (A@0.5), etc. up to accuracy at 5 km (A@5)

5.2 Georesolution

We also wanted to examine the georesolution performance of the Edinburgh Geoparser to understand how well it is able to assign latitude and longitude coordinates to location mentions and distinguish between multiple candidates in the case of ambiguous place names. Table 10 presents the georesolution scores for all three gold standard sets, both for all and Edinburgh-specific locations.

In this case, accuracy scores are reported only for locations which are contained in the gazetteer and for which the annotator was able to select a pin on the map when creating the gold standard annotation. Numbers of locations where this was not the case are also listed in Table 9. The figures show that a fair number of location mentions do not occur in the gazetteer. Most of them are fine-grained locations outside of the Edinburgh area, locations in other cities (e.g. *India House* in London) or made-up place names (e.g. *Wrinkly Scaurs*) which were annotated in the gold standard.

Georesolution accuracy scores are presented in Table 10. Exact accuracy scores for all locations increase when moving from historical and lower-quality to modern and high-quality text. The biggest difference (of over 10%) occurs for accuracy measured at 0.5 km. For Edinburgh-specific locations only, exact accuracy is lowest for MODERN text and considerably so (70.9%). One reason could be the fact that the location ratio within that set is largest (see location ratio figures in Table 2). When evaluating accuracy at increasing (but small) distances, this performance difference decreases. So the reason for the low exact accuracy for MODERN text is also caused by the system choosing a correct duplicate candidate in the vicinity of the annotated gold candidate. At 1 km, georesolution accuracy for MODERN text is similar to the results obtained for the historical datasets.

Accuracy scores for resolving Edinburgh-specific locations only are considerably higher than those obtained for georesolving all locations (e.g. 21.8% higher for historical OCRED text at a distance of 1 km). A lot of our work in Palimpsest was spent on aggregating and cleaning the Edinburgh gazetteer which was necessary to map literature set in Edinburgh. It is encouraging to see that this effort has paid off.

6 Summary and conclusion

This article has presented extensive evaluation of the Edinburgh Geoparser for geoparsing fine-grained location names in literary text using three manually annotated gold standard sets. The non-copyrighted gold standard test sets are made available for future research. Our evaluation was done both for the geotagging and the georesolution steps of the Edinburgh Geoparser using different types of literary data manually annotated for comparison (historical versus modern text, clean text versus text containing OCR errors). We also computed inter-annotator agreement scores as an upper bound to system performance and have shown that both tasks are relatively easy to perform manually even if not completely consistently.

We have shown that the historical text containing errors is more difficult to geotag automatically and that the availability of a suitable gazetteer is essential for geotagging and georesolution. Name variations in the text can throw the system as not all of them might be recorded in the gazetteer and some place names might be missing altogether, for example because they have disappeared over time. While we have shown that putting effort into developing the Edinburgh gazetteer paid off in terms of geoparsing fine-grained locations, a lot more work can be done to improve performance. In future work, we are hoping to integrate errors spotted by users of the LitLong interfaces in a feedback mechanism to increase geoparsing performance overall.

Acknowledgements Palimpsest was funded by the AHRC (Digital Transformations in the Arts and Humanities—Big Data, PI: Professor James Loxley). We thank the entire Palimpsest team made up of collaborators at the University of Edinburgh's School of Informatics, School of Literatures, Languages and Cultures and Edina and at the SACHI group at the University of St. Andrews. We thank all the data providers, including The British Library, Project Gutenberg, HathiTrust, the Oxford Text Archive and the National Library of Scotland for their support as well as all contemporary authors who gave us permission to geoparse their works. We also thank Dr. Kate Byrne and Vasilis Karaiskos who prepared the gold

standard data and helped to annotate it. Finally, we thank the reviewers and their critical and helpful comments during the preparation of this article.

Funding This work was funded by the Arts and Humanities Research Council in the UK (AH/L009935/1) as part of the Digital Transformations theme.

References

Alex, B. (2017). Geoparsing English-language text with the Edinburgh Geoparser. *The Programming Historian*. https://programminghistorian.org/en/lessons/geoparsing-text-with-edinburgh.

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2014). A web-based geo-resolution annotation and evaluation tool. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII), COLING 2014* (pp. 59–63).

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh Geoparser for historical georeferencing. *International Journal for Humanities and Arts Computing*, *9*(1), 15–35.

Alex, B., Grover, C., Klein, E., & Tobin, R. (2012). Digitised historical text: Does it have to be mediOCRe? In *Proceedings of KONVENS 2012 (LThist 2012 workshop)* (pp. 401–409).

Alex, B., Grover, C., Oberlander, J., Thomson, T., Anderson, M., Loxley, J., et al. (2017). Palimpsest: Improving assisted curation of loco-specific literature. *Digital Scholarship in the Humanities*, *32*(Suppl. 1), 4–16. https://doi.org/10.1093/llc/fqw050.

Alves, D., & Queiroz, A. I. (2015). Exploring literary landscapes: From texts to spatiotemporal analysis through collaborative work and GIS. *International Journal of Humanities and Arts Computing*, *9*(1), 57–73. https://doi.org/10.3366/ijhac.2015.0138.

Cooper, D., & Gregory, I. (2011). Mapping the English Lake District: A literary GIS. *Transactions of the Institute of British Geographers*, *36*(1), 89–108.

Curran, J., & Clark, S. (2003). Language independent NER using a maximum entropy tagger. *Proceedings of CoNLL, 2003* (pp. 164–167).

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K., & Solti, I. (2012). Building gold standard corpora for medical natural language processing tasks. In: *American Medical Informatics Association 2012 annual symposium* (pp. 144–153).

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence, AAAI'15*, (pp. 2382–2388). AAAI Press. http://dl.acm.org/citation.cfm?id=2886521.2886652.

Ganguly, D., Leveling, J., & Jones, G. J. F. (2014). Automatic prediction of text aesthetics and interestingness. In: *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)* (pp. 23–29).

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K. U., & Neumann, A. (2011). Towards information retrieval on historical document collections: The role of matching procedures and special lexica. *IJDAR*, *14*(2), 159–171.

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2017). What's missing in geographical parsing? *Language Resources and Evaluation.*, *52*, 603–623. https://doi.org/10.1007/s10579-017-9385-8.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., et al. (2010). Use of the Edinburgh Geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875–3889.

Hauser, A., Heller, M., Leiss, E., Schulz, K. U., & Wanzeck, C. (2007). Information access to historical documents from the Early New High German period. In L. Burnard, M. Dobreva, N. Fuhr, & A.

Lüdeling (Eds.), *Digital historical corpora—Architecture, annotation, and retrieval*. Germany: Dagstuhl.

Hibbert-Ware, M.C. (1883). His Dearest Wish, vol. 2. F.V. White and Co. https://archive.org/details/hisdearestwish02waregoog.

Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J., Hardisty, F., Pezanowski, S., Mitra, P., & MacEachren, A. (2013). GeoTxT: a web API to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval, GIR 2013,* (pp. 72–73). Association for Computing Machinery. https://doi.org/10.1145/2533888.2533942.

Kolak, O., & Resnik, P. (2005). OCR post-processing for low density languages. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 867–874).

Lopresti, D. (2005). Performance evaluation for text processing of noisy inputs. In *Proceedings of the Symposium on Applied Computing* (pp. 759–763).

Lopresti, D. (2008). Measuring the impact of character recognition errors on downstream text analysis. In: A. B. Yanikoglu, & K. Berkner (Eds.), Document recognition and retrieval (Vol. 6815). SPIE.

Lopresti, D. (2008). Optical character recognition errors and their effects on natural language processing. In *Proceedings of the second workshop on analytics for noisy unstructured text data* (pp. 9–16).

Loxley, J., Alex, B., Anderson, M., Hinrichs, U., Grover, C., Thomson, T., et al. (2018). Multiplicity embarrasses the eye: The digital mapping of literary Edinburgh. In I. Gregory, D. Debats, & D. Lafreniere (Eds.), *The Routledge Companion to Spatial History (Routledge companions)* (pp. 604–628). London: Routledge. https://doi.org/10.4324/9781315099781.

Luchetta, S. (2017). Exploring the literary map: An analytical review of online literary mapping projects. *Geography Compass*, *11*(1), e12303.

Minnen, G., Carroll, J., & Pearce, D. (2000). Robust, applied morphological generation. *Proceedings of INLG*, *2000*, 201–208.

Moncla, L., Gaio, M., Joliveau, T., & Le Lay, Y. F. (2017). Automated geoparsing of Paris street names in 19th century novels. In *Proceedings of the 1st ACM SIGSPATIAL workshop on geospatial humanities*.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., & Gaio, M. (2014). Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14* (pp. 183–192). https://doi.org/10.1145/2666310.2666386.

Oliphant, M. (1890). *Royal Edinburgh: her saints, kings, prophets and poets*. Siegel-Cooper. https://catalog.hathitrust.org/Record/004962573

Reuschel, A. K., & Hurni, L. (2011). Mapping literature: Visualisation of spatial uncertainty in fiction. *The Cartographic Journal*, *48*(4), 293–308. https://doi.org/10.1179/1743277411Y.0000000023.

Reynaert, M. (2008). Non-interactive OCR post-correction for giga-scale digitization projects. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 617–630).

Rupp, C. J., Rayson, P., Baron, A., Donaldson, C., Gregory, I. N., Hardie, A., & Murrieta-Flores, P. (2013). Customising geoparsing and georeferencing for historical texts. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6–9 October 2013, Santa Clara, CA, USA* (pp 59–62). https://doi.org/10.1109/BigData.2013.6691671.

Solina, F., & Ravnik, R. (2010). Georeferencing works of literature. In *Proceedings of ITI 2010, the 32rd International Conference on Information Technology Interfaces* (pp. 249–253). Cavtat, Croatia.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). Brat: A web-based tool for NLP-assisted text annotation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, Avignon, France* (pp. 102–107). http://dl.acm.org/citation.cfm?id=2380921.2380942

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning, CoNLL-2002, COLING-02, Taipei, Taiwan* (pp. 1–4). https://doi.org/10.3115/1118853.1118877.