Text Analysis of Aberdeen Burgh Records 1530-1531

Adam Wyner[1], Jackson Armstrong[2], Andrew Mackillop[2], and Philip Astley[3]
[1]University of Aberdeen, Department of Computing Science, Aberdeen, Scotland
azwyner@abdn.ac.uk
[2]University of Aberdeen, Department of History, Aberdeen, Scotland
{j.armstrong,a.mackillop}@abdn.ac.uk
[3]Aberdeen City and Aberdeenshire Archives, Aberdeen, Scotland
PAstley@aberdeencity.gov.uk

Abstract

The paper outlines a text analytic project in progress on a corpus of entries in the historical burgh and council registers from Aberdeen, Scotland. Some preliminary output of the analysis is described. The registers run in a near-unbroken sequence form 1398 to the present day; the early volumes are a UNESCO UK listed cultural artefact. The study focusses on a set of transcribed pages from 1530-1531 originally hand written in a mixture of Latin and Middle Scots. We apply a text analytic tool to the corpus, providing deep semantic annotation and making the text amenable to linking to web-resources.

1 Introduction

The council registers of Aberdeen, Scotland are the earliest and most complete body of town (or burgh) council records in Scotland, running nearly continuously from 1398 to the present; they are hand written in Latin and (largely) Middle Scots. Few cities in the United Kingdom or in Western Europe rival Aberdeen's burgh registers in historical depth and completeness. In July 2013, UNESCO UK recognised the register volumes from 1398 to 1509 as being of outstanding historical importance to the UK. The registers offer a detailed *legal* view into one of Scotland's principal burghs , casting light on administrative, legal, and commercial activities as well as daily life. The registers include the elections of office bearers, property transfers, regulations of trade and prices, references to crimes and subsequent punishment, matters of public health, credit and debt, cargoes of foreign vessels, tax and rental of burgh lands, and woods and fishings. Thus the entries present the burgh's relationships with the countryside and countries around the North Sea.

To make this historical resource available to a wider audience, the National Records of Scotland and Aberdeen City and Aberdeenshire Archives collaborated to image the volumes digitally up to 1511 and made them (temporarily) available on the internet.[1] However, the images of scribal records are inaccessible to all but a few scholars. To address this, a pilot project at the University of Aberdeens Research Institute of Irish and Scottish Studies (RIISS) has transcribed 100 pages of the records from the period 1530-1531, translated the Latin and Middle Scots, and provided a web-accessible database application; the application allows users to query the database for locations and names of individuals, returning the textual portions that contain the names and locations.[2] However, the pilot project does not make use of text analytic or Semantic Web technologies to facilitate understanding of and access to the records.

In this paper, we outline a funded text analytic project in progress on this corpus of 100 pages and provide some preliminary output. The project *A Text Analytic Approach to Rural and Urban Legal Histories* has been funded by the dot.rural Resource Partnership at the University of Aberdeen.[3] We outline the project objectives, present the text analytic tool, provide some sample results, relate our work to other projects, and sketch future work. The paper and project contribute to the application of language technologies for cultural heritage and the humanities. We discuss deep semantic annotation of the documents as well as plans to address linguistic variation and linking of the annotated material to other digital, web-based resources.

[1] http://www.scotlandsplaces.gov.uk/digital-volumes/burgh-records/aberdeen-burgh-registers/
[2] http://www.abdn.ac.uk/riiss/Aberdeen-Burgh-Records-Project/connecting-projecting.shtml
[3] http://www.dotrural.ac.uk

2 Objectives

The project engages legal historians, council archivists, and computational linguists. For legal historians, the burgh registries are an opportunity to study source materials concerned with the law and community concerning questions as:

- What legal roles in jurisdictions do individuals perform?
- What are the social and legal networks?
- How do social and legal concepts evolve?
- What does the historical record say about resource management and conflict?

While traditional historical methodology applied to archival material has served well enough, it is costly, slow, and does not allow analysis of the volume and complexity of information. In particular, some of the questions above are *relational*, e.g. relations of individuals in legal roles, which are difficult to track across a large corpus. With text analytic support, legal historians can query a corpus and receive data either in context or extracted.

For council archivists, the agenda is to increase public access to archival materials for tourism, curriculum development, business, and research. This can be done, we believe, by making the rich content of the archives accessible by translation, semantic search, or link to the content of the archival materials or other web-accessible resources such as dictionaries, maps, DBPedia entries, other council archival material, and so on.

For computational linguists, the objective is to annotate, enrich, and link the burgh records in order to support semantic querying, extraction, and reuse. One challenge is to find or develop the range of necessary text analytic components to do so. For non-standardised historical languages, e.g. Middle Scots, the issues are orthographical variation, lack of electronic lexicons, and so on. A more substantive challenge is to develop the appropriate set of semantic annotations, tailored to the historical, legal context and the goals of historical legal analysis.

3 Text Analysis

To identify, query, and extract the textual elements from the source material with respect to semantic annotations, we use the GATE framework (Cunningham et al., 2002), which we briefly describe. We then discuss our approach to analysis, the representation of textual elements using GATE, the annotations we introduce to the text, and then provide the results of sample queries.

3.1 Components of a Tool

GATE is a framework for language engineering applications, which supports efficient and robust text processing (Cunningham et al., 2002); it is highly scalable and has been applied in many large text processing projects; it is an open source desktop application written in Java that provides a user interface for professional linguists and text engineers to bring together a wide variety of natural language processing tools and apply them to a set of documents. The tools are formed into a pipeline of natural language processors. Our approach to GATE tool development follows (Wyner and Peters, 2011), which is: bottom-up, rule-based, unweighted, modular, iterative, incremental, among others. Once a GATE pipeline has been applied, we can view the annotations either *in situ* or queried using GATE's ANNIC (ANNotations In Context) corpus indexing and querying tool.

For our purposes, we emphasise the role of *gazetteers* and *JAPE rules*, which form the *bottom level* of the analysis. A gazetteer is a list of words that are associated with a central concept as provided by an analyst. In the lookup phase of processing the text, textual passages in the corpus are matched with terms on the lists, then assigned an annotation, e.g. a token term *burgi* is annotated with *LegalBody*, for it is one of the legal bodies reported in the text. Similarly, tokens such as *common council*, *curia*, *guild court*, and others are all annotated *LegalBody*. The gazetteer thus annotates related terms (e.g. *burgi* and *guild court*) with the same annotation; in this way, annotations serve as *conceptual covers* for tokens. We have gazetteers that provide a range of semantic concepts for named entities as well as:

- LegalBody - burgi, common council, ...
- LegalConcept - gude faith, ...
- LegalRole - Archbishop, Bailie, ...
- Offence - barganyng, tulyheing, etc
- Office - alderman, burgess, preposito, ...
- RegisterEntry - Bailie Court, Ordinance, ...
- MiddleScot - The, said, day, bailyeis, ...

Alternative spellings of a word would be represented as different tokens in the gazetteer. The selection and content of the gazetteer lists is preliminary and will be the object of significant research

over the course of the project. However, they are sufficient to facilitate exercise of the tool.

JAPE rules are transductions that take annotations and regular expressions as input (based on the gazetteers) and produce annotations as output. The annotations produced by JAPE rules are visible as highlighted text and are easily searchable in ANNIC. Querying for an annotation, we retrieve all the terms with the annotation. The annotations can also be used in JAPE rules to create higher level annotations, though we have not developed these at this point.

3.2 Output and Queries

Once the corpus is annotated, we can view the annotations *in situ*. In Figure 1, we have a passage that has been highlighted with the indicated (checked) annotation types (differentiated by colour in the orginal). In this figure, we see where the annotations appear and in relation to other annotations within a particular textual passage. Observations at this point can be used to analyse the text further.

Alternatively, we can use the ANNIC tool to index and query a database of annotated text. Searching in the corpus for single annotations returns all those strings that are annotated with the search annotation along with their context and source document. Complex queries can also be formed. A query and a sample result appear in Figure 2, where the query finds all sequences of annotated text where the first string is annotated with *Name*, followed by zero to five other *Tokens*, followed by a string with an *Office* annotation. The search returned four candidate structures. The extract identifies a *relation* between an individual and their office. Similar relational queries can be made about other aspects of the text. With the query language, we can search for any number of the annotations in the corpus in any order; the tool allows incremental refinement of searches, allowing for a highly interactive way to examine the semantic content of the texts. Thus, a range of semantic patterns can be identified that would otherwise be very hard to detect or extract. Such an approach can ground multi-disciplinary investigations of historical societies in large-scale textual sources of information, providing interpretable material on topics such as elites and social practice, relations between social classes and land, urban and rural development, and natural resource management. The text analysis also makes applicable a range of social web-mining approaches on historical text.

4 Related Work

Our work is closely related to other projects that have applied text analytic methods to *mine* information from the cultural heritage objects, broadly Digital Humanties. Most recently, there has been an extensive n-gram study of Scottish legal records. This takes a very different, though nonetheless relevant approach to the study of these records ngrams (Kopaczyk, 2013). Several recent projects in the UK and Ireland have applied such tools in limited ways to historical legal documents, e.g. *1641 Depositions* (Sweetnam and Fennell, 2012), which analysed verbal patterns in the text [4]; *The Old Bailey*, which was largely manually annotated though some elements were automatically annotated [5]; and Trading Consequences, a text an-

[4] http://1641.tcd.ie
[5] http://www.oldbaileyonline.org

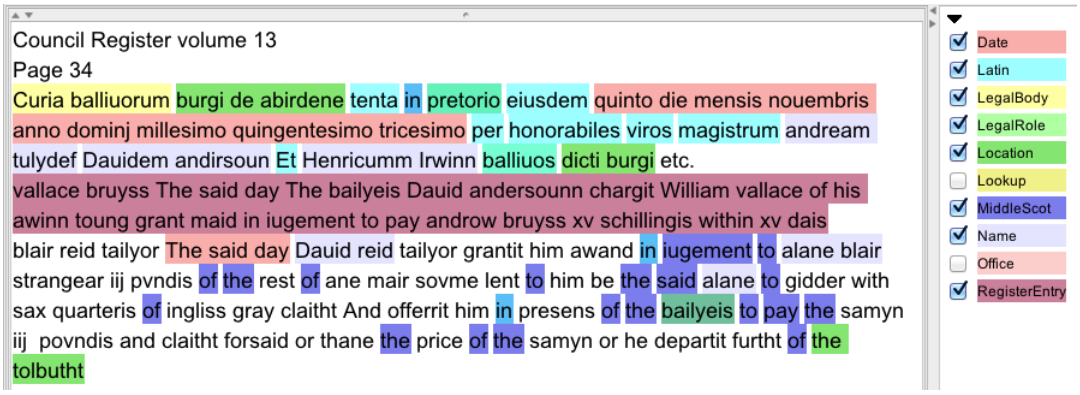

Figure 1: Highlighting Annotations in the Text

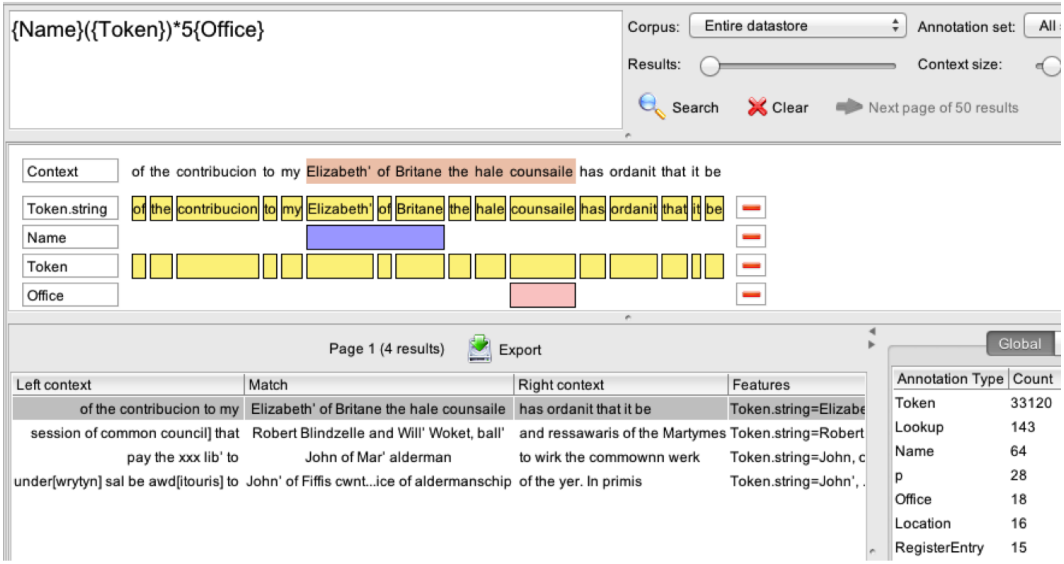

Figure 2: Searching for Relations in the Corpus

alytic study of British Empire records [6]. There are ongoing Semantic Web projects in the Humanities, e.g. the Curios Project [7], the CULTURA Project [8], projects at King's College London Centre for Digital Humanities [9].

5 Future Plans

Over the course of the project, we will develop and refine a useful set of annotations that reveal important information this is distributed over this corpus. Asides from the main categories previously discussed, there will be annotations to indicate orthographic variants, translations, and links to external material amongst other annotations to be developed. Legal historical questions will be posed with respect to the contents of the text, then the text will be queried using the annotations in complex patterns. In this way, the questions of legal historians are grounded in and tested against the textual substance. Another objective is to link the annotated material to other relevant material that is external to the corpus. For instance, locations could be associated with maps, names could be associated with DBPedia entries, words could be linked to Scottish and Latin dictionaries, and so on. This would not only further enrich the contents of the corpus, but also enrich these other materials by linking to the corpus. Similarly, these texts can be tied to other legal historical projects, focussing on the period c.1400 c.1800, that will inter-relate the council register source material with cognate collections held in Aberdeen (at the Aberdeen City and Aberdeenshire Archives, and at the University of Aberdeens Special Collections Centre, and elsewhere), in Scotland (in other local archives and in the National Records of Scotland), in the United Kingdom, or the European Union. This will foster both a comparative understanding of the city and its regions position regionally, nationally, and internationally, and over time.

Beyond the project, we look forward to enlarge the council register corpus and extend the text analysis. It would then be very attractive to create a web-based, interactive interface with which to interrogate the council register in complex and novel ways, not just by querying the text with semantic annotations, but also by following links to maps, recordings, images, related words, and so on. For example, the content could be linked to time series maps, showing development of social, legal, and political relationships over time and space.

Acknowledgments

The authors are grateful for support from RIISS, dot.rural, and the Aberdeen Council. We particularly thank Dr. Edda Frankot for her work on the pilot project at RIISS that produced the transcribed materials for this project.

[6] http://tradingconsequences.blogs.edina.ac.uk/

[7] http://www.dotrural.ac.uk/curios/

[8] www.cultura-strep.eu

[9] http://www.kcl.ac.uk/artshums/depts/ddh/index.aspx

References

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, pages 168–175.

Joanna Kopaczyk. 2013. *The Legal Language of Scottish Burghs*. Oxford University Press.

Mark S. Sweetnam and Barbara A. Fennell. 2012. Natural language processing and early-modern dirty data: applying IBM *languageware* to the 1641 depositions. *Literary and Linguistic Computing*, 27(1):39–54.

Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. In Katie Atkinson, editor, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pages 113–122. IOS Press.