Historical Text Mining and the Nineteenth-Century Serials Edition

[slide]

1. Intro

a. Volume and format of new digitisations

In the move to digital publishing nineteenth-century materials have been a key growth area. The venture has attracted a diverse set of participants from large commercial publishers like Thompson-Gale and Pro-quest, keen to be key players in the market for subscription-only services, to JISC, and the British Library and a whole host of smaller-scale BA and AHRC projects like **ncse**. An ever increasing proportion of nineteenth-century print is in desperate need of preservation beyond the budgets of holding repositories and digitisation is seen *the* remedy to this situation. As a result millions of pages of OCR and digital images of the pages are currently being generated by libraries, archives, publishers and academic projects. However, as we all know digitisation can do much more than preserve this material. Tools for exploration and analysis could enable users to engage with these texts not just as mines of OCR-based information but as the **rich historical objects** they are. Reciprocally, the unprecedented volume and diversity of printed matter in the nineteenth century also makes it a good source of experimental and diagnostic materials useful in developing processing techniques which might be utilised on other corpora. In conclusion we will outline the particular challenges faced by those applying such techniques to nineteenth-century serial literature.

2. Placement of ncse in relation to these projects

Our experience of these issues comes from our work on the nineteenth-century serials edition.

[slide]

ncse is a three year ahrc project to digitise six nineteenth century periodicals and newspapers and make them freely available on the web. **ncse** aims at completeness (multiples, supplements, advertisements etc) and will provide an edition of 98,565 very diverse pages (words, images, different genres etc) with complex searching and indexing functions. The edition and accompanying tools will foreground the periodical as a genre and enable users to see these materials not just as mines of information, but to engage with them critically as historical material objects. The way we want to do this is by taking care to contextualise the OCR text - the primary material with which users will interact. **ncse** will do this by bringing out the complex relationships which operate **within** the content, **between** the *form* and the content, and **across** the genre of nineteenth-century serial publication. We hope to use text mining of our OCR to reveal some of these relationships.

At the most basic level we want to bring out relationships operating at the level of content conceived of solely as the words on the page.

3. Textual relationships

a. Content

I. Thematic

Perhaps one of the most obvious ways to draw out relationships in our texts is to seek some way of mapping content on equivalent themes across the life of a publication and between publications. This would be particularly useful where publications used distinct vocabularies reflecting their attempt to circumscribe specific audiences.

[slide]

Here we have two articles from two very different publications about the same popular topic - the problems associated with working class drinking habits. Both of these articles implicitly support the argument for temperance. However, the two do not share common keywords or vocabulary the first does not even mention drunkenness though the implication that the attackers were inebriated is clear. Their address is also significantly different. Whilst the sensation of the *Northern Star*'s reproduced police reports encourage their respectable reformist upper working class readers to tutt and curse at the behavior of their less elevated fellows, the EWJ encourages its readers to abstractly fret from the safety of their well upholstered armchairs about the ‘condition of the great unwashed’.

[slide]

Associating concept mapping terms like ‘temperance’ - which neither article uses - with a thesaurus of keywords generated by text mining to map content across the edition would bring out these important relationships and help users to see how debates about such issues played out across different publications and over the century. Such work also has further intellectual purpose – serving as an evidence based mapping of modern analytical categories onto actor categories – which in turn would serve to orientate none expert users in the edition.

[slide]

II. Generic: linguistic analyses to study genre - address / language, metrics (sentence / word lengths etc)

III. People, places, events, publications etc

> Existing techniques of indexical extraction and linguistic analyses of historical texts will also be of huge value in this context. Linguistic metrics could for example help us to detect changes in address between publications and see how aspects of the genre develop across the century. Extraction of indices of people, places, etc cross referenced with extant databases of contemporary data would not only enable greater precision in searching but illuminate prosopographies of the period generally and nineteenth-century print culture specifically.

However, what **ncse** aims to do goes way beyond this content based remit. One of the key challenges of the project is to break out of the tendency of OCR based editions to over-privilege the text – narrowly conceived of as the OCR-ed words on the page – and instead to contextualise this data in ways which do not estrange the words from the historical artefacts in which they are enmeshed. We want to do this by making apparent the relationships which operate between the *form* and the content, and *across the genre* of nineteenth-century serial publication.

b. Form

I. Relationships between text and context

[slide]

Northern Star example: This slide shows the contextual relevance of items: this is by O'Connor, the proprietor of the paper, and is in the most important textual position.

[slide]

BL *Penny Illustrated Paper* example: privileges text-based content, conceiving it as located in stand-alone articles that can be separated from their contexts. This bias in inscribed in the way users access the title:

- This collection is designed for general users and has a simple, easy to use interface but it estranges users from the context of the material they are reading.
- You cannot browse, or even search by date. You have to supply a keyword. This implies all users want to find articles *about* something, and when hits are returned they takes users straight to the segment. Users have to choose to select full page view in order to see what else is on the page.
- In this example the whole page view reveals that a part of the segment is missing, and that this is an advertisement accompanying one for lemonade, a prize competition, the end of a story, and a news report about a retired army officer on trial for embezzlement. Also, as users

can only zoom into segments, they cannot see the page numbers, which do not correspond to those given in the display.

[slide]

II. Structural / hierarchical relationships

- Relationship between titles: *MR* family tree (it becomes more secular after *Unitarian Chronicle* – you need to know where this material goes)
- Edition is structured in a hierarchy: edition; title; series; volume; number; department; item. Each of these could be considered a “unit” of analysis.
- This is important: a review appearing in a review section has a different status than one in the news section.
- Volumes and numbers often have supplements such as advertising wrappers, indices, prefaces etc.. These can complicate the relationships between numbers, making serials less of a linear sequence

[slide]

- There might also be multiple editions – and we have just under 3 per number of *NS* and 1½ per number of the *Leader*. Show NS white space example.

III. Formal/generic relationships

Diversity in terms of appearance …etc

[slide]

But also certain common genres to C19th periodicals:

- o Leading articles – and these tend to be in the middle of weeklies.
- o News
- o Correspondence (and two of our titles call this “Open Council”)
- o Obituaries – in the *PC* they change the name from “Obituary” to “In Memoriam” in 1889 – so if users wanted to search, they would have to be aware of this.

And there are generic aspects connected to things like typography, layout, and paper size:

[slide]

- o e.g. *NS* becomes less newspaper-like and more periodical like, finishing up looking a lot like the Leader, which has a very different class of reader: on slide: digital images obscure the size change – just before it becomes the *Star of Freedom* (p.15 no.753 17 April 1852), the editor, G. Julian Harney, announces his return with a price cut (from 5d to 4½d). He reveals it cannot be any cheaper without having to cut the size in half, or filling it with police reports to maintain the necessary circulation of 30,000 a week. From 14 August 1852 it undergoes an even more radical change, 3 cols and smaller, but twice as long (16pp) – the editor suggests it will encourage people to buy it in ½ yearly volumes. In other words it is no longer a newspaper, and its format – even to the extent of publishing a town and country edition – resembles the *Leader*.

IV. Verbal and the visual- again relationships are important and OCR over-privileges text

[slide]

The Tomahawk Disraeli example – the relationship is signalled in the caption and the title: but how do we capture visual information?

4. Challenges for text mining technologies

Two categories of challenges which make our material a great test bed for developing text mining techniques.

Firstly those associated with peculiarities in content and how it is reproduced:

[slide]

a. Satire, humour.

When a text does not mean what it says it means. Show *Tomahawk* fake advert. Very important in Tomahawk as large portions of the text rely on what is not said, whether relying on the knowledge of the reader or through elaborate puns and allusions.

b. Poor (uncorrected) OCR

Reproduction problems and unconventional typography etc – forces us to attend to more than just the words on the page if we are to get the most out of it.

c. Fragments

Particular problem with our material is fragmentary content – serial literature is necessarily heterogeneous: it is literally made up of fragments. As such, it often is inappropriate as a corpus as its scope is too wide, and the number of words too small.

There are ways around this: for instance genres (identified by keyword extraction) might constitute corpora across titles. Intertextual references.

The second is to do with placing content in context:

We recognize the value of text mining in content retrieval, but in encouraging users to engage with these texts not just as mines of OCR based information but as **rich historical objects** these tools to be contextual. For instance, we would like to see some way of reflecting the relationship between form and content and those operating at the level of genre – e.g. perhaps by ranking based on physical indicators of form?

At the moment text mining techniques operate on what we acknowledge is a *necessarily* narrow conception of text, allowing sophisticated identification of content from large corpora. However, in digital publishing we need to recognize that there are other orders of information – relational, contextual and visual – that should not be ignored as non-textual. Text mining techniques already recognize the connection between word use and textual meaning: it is when we begin to move beyond the occurrence of words to consider where they occur, and what that they look like that these techniques will realise their full potential as analytical and exploratory tools for scholarly research into historical materials.