Corpus Protocols: digital transformations of commercial newspaper collections for text and data mining to support academic research

IFLA 2014 Pre-Conference: Digital Transformation and the Changing Role of News Media in the 21st Century
August 13-14, 2014
International Telecommunication Union, Geneva, Switzerland
Working together: building research relationships
Corpus Protocols Project Team

- Seth Cayley: Head of Research Solutions, Cengage Learning EMEA
- Mike Gardner: Analyst and Web Developer in Web Technologies, University of Nottingham
- Kat Gupta: Corpus Protocols Researcher, University of Nottingham
- Michaela Mahlberg: Professor of English Language and Linguistics, University of Nottingham
- Neil Smyth: Faculty Team Leader – Arts, University of Nottingham
- Stella Wisdom: Digital Curator at the British Library
Context

- Focusing on textual data (corpus) we review relationships between institutional infra-structure, external business partners and research questions and methods that deal with textual data.
- Exploring opportunities to future developments (protocols).
Problem (practical and research)

• Some University of Nottingham owned data sets are not available for research for several reasons, including:
  – licence agreements restrict copying for data analytics;
  – additional charges for additional licenses;
  – technical issues: copying storage, accessibility of data;
  – and, a lack of shared understanding of research trends in applied linguistics, corpus linguistics and corpus stylistics.
Aims

• To articulate the relationship between data and tools for text and data mining
• Prototype a methodology for using locally stored data using existing corpus linguistic tools
• Use an existing corpus linguist tool to analyse a University owned data set that will be made available on the University network.
Corpus Protocols: didn’t start with discourse research questions

• Declassified Documents as opportunistic data

• The Declassified Documents corpus
  – 197,837,328 words in total
  – divided into 30 yearly subcorpora (between 3.3 to 8.3 million words each)
UK Copyright legislation: 2014

"the making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that (a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and (b) the copy is accompanied by a sufficient acknowledgement."

(The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014.)
Corpus methods:

• Discourse as representation: labeling, naming, creating reality (e.g. *weapons of mass destruction*).
• Lexis in context
application of IAEA or similar international safeguards on all their peaceful nuclear activities. This document consists of 3 pages.

FOR MANUFACTURE INCLUDING INFORMATION FOR PEACEFUL NUCLEAR ACTIVITIES. SAID WE THOUGHT PROBLEM BETTER.

accepts the application of Euratom safeguards upon all its peaceful nuclear activities. The US considers Euratom safeguards to be fully treaty of IAEA or similar international safeguards on all its peaceful nuclear activities as suggested in alternative paragraph 1 of Article II would object to "all source or special fissionable materials in all peaceful nuclear activities." We believed that the treaty must permit the continued facilitation of the application of IAEA or equivalent safeguards to their peaceful nuclear activities. There would be a review conference after a specified West Germany argued that this was discriminatory because the peaceful nuclear activities of the nuclear-weapon nations were not under safeguards. states should accept IAEA safeguards to their peaceful nuclear activities and that all transfers for peaceful purposes should take place among all nuclear-weapon states. Some of these considerations had emerged in the context of discussing the treaty of its peaceful nuclear activities as soon as practicable. Each State Party to this Treaty agreed in principle that all nuclear materials previously transferred to each other under the terms of the Treaty of Peace with Germany in 1979 should be returned to the country of origin. IAEA safeguards to peaceful nuclear activities. The safeguards material entered into force on May 22, 1995. The agreement requires that the safeguarded material be applied on all source or special fissionable material in all peaceful nuclear activities within the territory of States, under its jurisdiction or control. Article II requires that the safeguards in all peaceful nuclear activities within the territory of States, under its jurisdiction or control shall include mandatory IAEA safeguards for the peaceful nuclear activities of non-nuclear states. It anticipated that the non-nuclear states, including international co-operation in the field of peaceful nuclear activities. The article on control provides for the establishment of the Parties of international cooperation in the field of peaceful nuclear activities, including the international exchange of nuclear material of significant quantities (table) of nuclear material from peaceful nuclear activities. The Department of Safeguards is headed by the

The application of IAEA safeguards to all peaceful nuclear activities within the territory of States, under its jurisdiction or control, is the responsibility of the Department of Safeguards. All actions would be subject to a review conference.
<table>
<thead>
<tr>
<th>Year</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>AA</th>
<th>AB</th>
<th>AC</th>
<th>AD</th>
<th>AE</th>
<th>AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>NBSP</td>
<td>BR</td>
<td>TEXT</td>
<td>ILLEGIBLE</td>
<td>LIBRARY</td>
<td>ILLEGIBLE</td>
<td>ILLEGIBLE</td>
<td>ILLEGIBLE</td>
<td>ILLEGIBLE</td>
<td>ILLEGIBLE</td>
</tr>
<tr>
<td>1997</td>
<td>PHOTOCOPY</td>
<td>DWIGHT</td>
<td>ILLEGIBLE</td>
<td>WE</td>
<td>PHOTOCOPIED</td>
<td>WE</td>
<td>ARCHIVES</td>
<td>NARA</td>
<td>REPRODUCED</td>
<td>EO</td>
</tr>
<tr>
<td>1998</td>
<td>CARTER</td>
<td>EISENHOWER</td>
<td>TO</td>
<td>WE</td>
<td>TO</td>
<td>I</td>
<td>REPRODUCED</td>
<td>EO</td>
<td>NARA</td>
<td>SENSITIVE</td>
</tr>
<tr>
<td>1999</td>
<td>THIEU</td>
<td>TOP</td>
<td>TO</td>
<td>WE</td>
<td>REPRODUCED</td>
<td>EO</td>
<td>SENSITIVE</td>
<td>NARA</td>
<td>SENSITIVE</td>
<td>NARA</td>
</tr>
<tr>
<td>2000</td>
<td>GERALD</td>
<td>CHINESE</td>
<td>TO</td>
<td>WE</td>
<td>REPRODUCED</td>
<td>EO</td>
<td>SENSITIVE</td>
<td>NARA</td>
<td>SENSITIVE</td>
<td>NARA</td>
</tr>
<tr>
<td>2001</td>
<td>FORD</td>
<td>LBJ</td>
<td>WE</td>
<td>WE</td>
<td>REPRODUCED</td>
<td>EO</td>
<td>SENSITIVE</td>
<td>NARA</td>
<td>SENSITIVE</td>
<td>NARA</td>
</tr>
<tr>
<td>2002</td>
<td>LBJ</td>
<td>EISENHOWER</td>
<td>OUR</td>
<td>WE</td>
<td>WE</td>
<td>WE</td>
<td>REPRODUCED</td>
<td>EO</td>
<td>SENSITIVE</td>
<td>NARA</td>
</tr>
<tr>
<td>2003</td>
<td>FORD</td>
<td>EISENHOWER</td>
<td>WE</td>
<td>WE</td>
<td>WE</td>
<td>WE</td>
<td>REPRODUCED</td>
<td>EO</td>
<td>SENSITIVE</td>
<td>NARA</td>
</tr>
</tbody>
</table>

- **NUCLEAR**
TIDAL: Times Data Archive Lab project

- Continuing partnership with Cengage
- Times Digital Archive
- University of Michigan
- Investigating social reality in 19th century based on news discourse (to complement picture of Dickensian London)
British Library News Data: Future Plans

The British Library has one of the world's greatest news archives. Our collection of UK, Irish and world newspapers numbers over 60 million issues, from the 17th century to the present day, and we have growing collections of television, radio and web news.

http://britishlibrary.typepad.co.uk/thensnewsroom/
https://twitter.com/BL_newsroom
British Library News Data: Future Plans

Print newspapers from Colindale moved to a new newspaper storage building in Boston Spa in Yorkshire
British Library: Newsroom Opened 7th April 2014
British Library News Data: Future Plans

Broadcast News:
Daily television and radio news programmes broadcast in the UK since May 2010 available through an instant access service in the Reading Rooms, new programmes available within hours of broadcast. Over 60 hours of news are recorded every day from 22 channels including BBC, ITV, Channel 4, Sky News, Al Jazeera English and CNN.
Global Top 100 University

BROADCAST NEWS
Television and radio news programmes

- View and listen to television and radio news programmes broadcast in the UK since May 2010
- Available from twenty-two UK and international news channels, with more programmes added daily
- The channels from which we currently select are:
  - Radio: BBC London, BBC Radio 1, BBC Radio 4, BBC 5 Live, BBC World Service, LBC, talkSPORT
- Advanced word-searching by subtitles available for some television channels
- We welcome any feedback on this service. Please contact us broadcastnews@bl.uk
British Library News Data: Future Plans

Many recorded news programmes come with subtitles

A corpus of data and metadata with potential for innovative research
PROTECTING THE NHS FOR TOMORROW

David Cameron has delivered a speech on the future of the NHS setting out his five personal "guarantees" on reform of the health service

READ NEWS STORY
Conservative party deletes archive of speeches from internet
Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh and Alex Hern
The Guardian, Wednesday 13 November 2013 15.40 GMT
Jump to comments (1284)
British Library Labs
Experiment with our digital collections

http://labs.bl.uk
Chicago Woman: How much do you charge for a divorce?

Chicago Lawyer: One hundred dollars, ma'am, or six for 500 dollars.
Corpus Protocols:
digital transformations of commercial newspaper collections for text and data mining to support academic research

IFLA 2014 Pre-Conference: Digital Transformation and the Changing Role of News Media in the 21st Century
August 13-14, 2014
International Telecommunication Union, Geneva, Switzerland