Key domain analysis: mining text in the humanities and social sciences

Paul Rayson
Department of Computing
Lancaster University

Dawn Archer
Department of Humanities
University of Central Lancashire
A talk of two halves …

• Motivation
  – Mismatch between tools developed and research questions
  – E.g. Manual classification of concordances, N-grams and Key words

• Key domains
  – Wmatrix tool demo
  – Extends key words to key semantic domains
  – Case study (NCSE)
Text analysis tools

- Corpus linguistics
  - WordSmith, AntConc, Wmatrix, MLCT, BNCweb
- Electronic text analysis
  - TACT
- CAQDAS (Computer-Assisted Qualitative Data Analysis Software)
  - Nvivo, Atlas.ti, HyperResearch
- Text mining
  - TerMine, Chesire

- Lack of awareness and duplication of effort?
Historical text mining

• Two workshops
  – Historical Text Mining (Lancaster, July 2006)
  – Text Mining for Historians (Glasgow, July 2007)
Historical text mining (HTM)

- Historical theory
- Natural language processing & Computational linguistics
- Linguistic theory

Corpus

Statistical and rule-based language models

Empirical evidence to inform theory

Linguistic theory

Historical theory

HTM
Tool-driven linguistics?

• C.f. corpus-driven or corpus-based linguistics
  – “So what?” problem

• Three examples
  – Manual categorisation of concordance lines
  – N-gram analysis
  – Key words
Manual coding of concordance lines in corpus linguistics

N-grams

• Terminology
  – Clusters (Scott)
  – Lexical bundles (Biber)
  – Recurrent combinations (Altenberg)

• Problems of analysis
  – Very large number of examples
  – Overlap between N and (N+1)-grams, (N+2)-grams etc
### 2-grams (top 10)
- 265 of the
- 174 in the
- 128 to the
- 94 had been
- 77 at the
- 72 and the
- 71 it is
- 67 it was
- 66 the russians

### 3-grams (top 10)
- 24 one of the
- 20 in order to
- 18 as a result
- 15 the fact that
- 13 the foreign office
- 13 is hard to
- 13 that he was
- 12 at the time
- 12 it is hard
- 12 a number of

### 4-grams (top 10)
- 11 it is hard to
- 6 at the end of
- 6 under the control of
- 6 mi6 and the cia
- 6 the portland spy ring
- 6 despite the fact that
- 6 at the same time
- 5 the control of the
- 5 the director general of
- 5 a member of the

### 5-grams (top 10)
- 5 under the control of the
- 4 it is hard to believe
- 4 is hard to believe that
- 3 will rid me of this
- 3 defence of the realm as
- 3 the defence of the realm
- 3 the director general of mi5
- 3 with the help of the
- 3 it is hard to think
- 3 of the portland spy ring

### 6-grams (top 10)
- 4 it is hard to believe that
- 3 the defence of the realm as
- 3 it is hard to think of
- 3 who will rid me of this
- 3 the end of world war ii
- 3 at the end of world war
If we compare text A with text B, we can discover the most significant items within text A and not only the frequent items.
Key words: problems

• Too many to examine
  – Filter by p-value (chi-squared critical value)

• Phrases missing
  – Key clusters (n-grams) WordSmith

• Manual classification (by grammar or semantics)
  – Wmatrix
Wmatrix demo

- Key words
- Key domains
  - Extends keywords to semantic fields
- Data-driven
  - Bridges quantitative and qualitative analyses
- 2005 general election
  - Liberal Democrat party manifesto
  - Labour party manifesto
Welcome to Wmatrix2

Messages of the day:
1. Volunteers are requested for testing the new My Tag Wizard feature. This allows the creation of personal dictionaries which can extend or override the existing semantic lexicon and MWE list. Please contact Paul if you are interested.
2. There is a new discussion forum for Wmatrix hosted at the Digital Arts and Humanities site.

Wmatrix is the web interface to the USAS and CLAWS corpus annotation tools. Functionality is provided by the toolbars at the top of each page. Wmatrix was initially developed in the REVERE project by Paul Rayson. New functions are still being added. Please send feedback to Paul.

The tag wizard guides you step-by-step through the process of uploading data, assign part-of-speech and semantic tags, then comparing your data to standard corpora, and finally seeing your data in context. Files are stored in folders which are like directories (previously called workareas) for one text and the analysis carried out on the text. You should use one folder per text.

See the help pages for information on how to use the tool and the part-of-speech and semantic tagsets. If you are used to the old Wmatrix system, then you can switch to the advanced user interface since this is fairly similar to the previous version with a new look and feel to the user interface in terms of menus and colour scheme.

Please reference Wmatrix as follows:
  http://ucrel.lancs.ac.uk/wmatrix/
  (abstract or full text)

If you publish a paper using Wmatrix, please acknowledge the tool as above and send a reference to Paul. Thanks!
You can see various views on this dataset:

1. **List of words and their frequencies**
   Sorted by: **Frequency or Word**

2. **Word search**
   Enter search term  Go

3. **Word clouds**
   Compared to:
   [Entrepreneurship Small Business Spoken (CESB)]  Go

4. **Tag clouds**
   Compared to:
   [Entrepreneurship Small Business Spoken (CESB)]  Go

©2000-8 UCREL, Lancaster University.
For technical queries please contact Paul Rayson: paul@comp.lancs.ac.uk
S 100 alternatives altogether am arms at_present bureaucracy car care chancellor class cleaner concentrate
congestion conservative cost council cut democrat democrats
dental despite diagnosis elderly elected energy environment environmental executive fairer farming food
for_example former government green guarantee hand high-quality i instead_of it job letters liberal long-term lords make_sure
manifesto ministers much my n't off oppose parliament pay
plans policies politicians pollution power principles prisoners promise propose really
red_tape replacing richest rural save Scotland scrap secretary seriously
shadow sizes spokesperson taught tax taxes taxpayers that these this top-up
tuition_fees unfair unnecessary waste what which whilst Whitehall worse wrong yet

Wmatrix
Key domain case studies

1. An exploration of the semantic field of ‘love’ in Shakespeare’s comedies and tragedies (Archer et al, forthcoming)


3. Analysis of interview transcripts in leadership and entrepreneurship studies (Doherty et al, 2006)

4. Child protection in online social networks: Adults masquerading as children (Isis project)
Nineteenth-Century Serials Edition

- Free, online edition of six nineteenth-century periodicals and newspapers, segmented to article level
  - Monthly Repository (1806-1837) and Unitarian Chronicle (1832-1833)
  - Northern Star (1838-1852)
  - Leader (1850-1860)
  - English Woman’s Journal (1858-1864)
  - Tomahawk (1867-1870)
  - Publishers’ Circular (1880-1890)
- Facsimile component
  - A repository of full-page facsimiles and textual transcripts generated through OCR
- Keyword component
  - An index of semantic keywords and person, place and institution names, generated using text mining and natural language processing techniques.
- Both components of the system are fully searchable and include rich, bibliographic metadata attached to titles, volumes, issues, departments and articles within the edition.
ncse nineteenth-century serials edition

Browse by Subject

A: General and Abstract Terms
B: The Body & the Individual
C: Arts & Crafts
D: Emotional Actions, States & Processes
E: Food & Farming
G: Government & the Public Domain
H: Architecture, Buildings, Houses & the Home
I: Money & Commerce in Industry
K: Entertainment, Sports and Games
L: Life and living things
M: Movement, Location, Travel & Transport
N: Numbers & Measurement
Q: Substances, Materials, Objects & Equipment
P: Education
R: Linguistic Actions, States & Processes: Communication
S: Social Actions, States & Processes
T: Time
W: The World & Our Environment
X: Psychological Actions, States & Processes
Y: Science & Technology
Summary

• Connect problem-based research questions to tools & methods available
  – Iterative development

• Key domain analysis
  – Incorporates phrases
  – Extends key words approach to key semantic fields
  – Supports content analysis
  – Bridges quantitative and qualitative analysis
Thanks for listening …

• Any questions?

• Paul Rayson (paul@comp.lancs.ac.uk)
• Dawn Archer (dearcher@uclan.ac.uk)