

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)

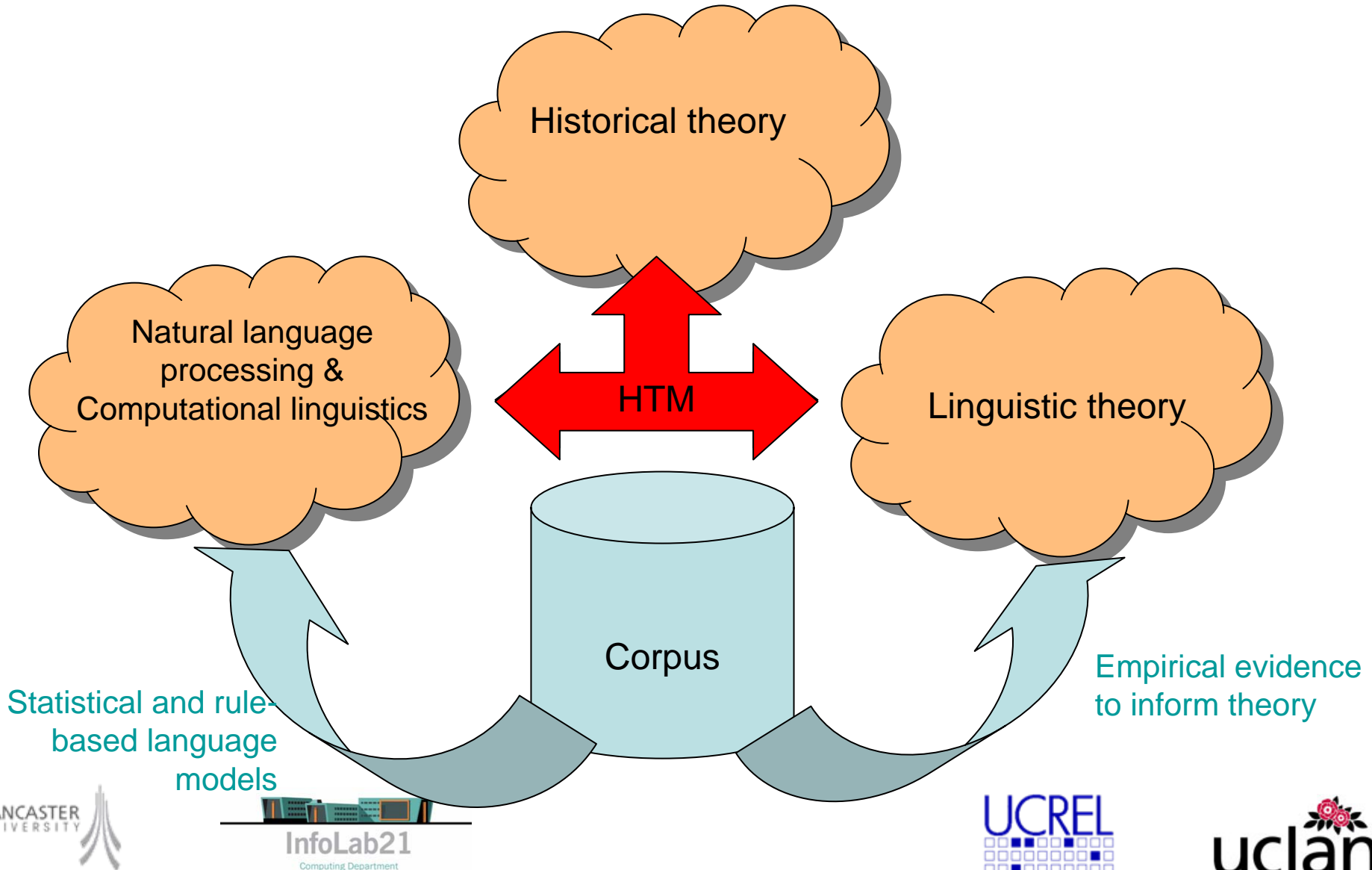
Text analysis tools

- 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)



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

- Smith, N., Hoffmann, S. and Rayson, P. (2008). Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations. *Literary and Linguistic Computing*, 23 (2), pp. 163-180. doi: 10.1093/lc/fqn004

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
71 by the
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

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

Key words

If we compare
text A

... with text B

... we can discover the most
significant items within text A

Item	O1	%1	O2	%2	LL
liberal	106	0.69	0	0.00 +	192.54
democrats	66	0.43	0	0.00 +	119.88
i	64	0.41	12	0.05 +	62.35
democrat	33	0.21	0	0.00 +	59.94
power	20	0.13	0	0.00 +	36.33
n't	24	0.16	1	0.00 +	36.23
green	21	0.14	1	0.00 +	31.04
taxes	17	0.11	0	0.00 +	30.88
's	100	0.65	62	0.27 +	30.08
make_sure	22	0.14	2	0.01 +	28.26
tax	50	0.32	20	0.09 +	27.71
secretary	14	0.09	0	0.00 +	25.43
manifesto	20	0.13	2	0.01 +	24.99
government	84	0.54	53	0.23 +	24.46
oppose	13	0.08	0	0.00 +	23.61
red_tape	13	0.08	0	0.00 +	23.61
shadow	12	0.08	0	0.00 +	21.80
unfair	15	0.10	1	0.00 +	20.80
pollution	19	0.12	3	0.01 +	20.08
long-term	11	0.07	0	0.00 +	19.98
ministers	11	0.07	0	0.00 +	19.98
it	90	0.58	65	0.28 +	19.77
these	28	0.18	9	0.04 +	19.10
pay	36	0.23	15	0.07 +	19.09

... 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

Wmatrix

[[My folders](#) | [Tag wizard...](#) | [Help contents](#)]

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, to 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:

- **Rayson, P.** (2007) Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University.
<http://ucrel.lancs.ac.uk/wmatrix/>
- **Rayson, P.** (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. *Ph.D. thesis*, Lancaster University.
([abstract](#) or full text  )

If you publish a paper using Wmatrix, please acknowledge the tool as above and send a reference to Paul. Thanks!

Wmatrix

©2000-7 UCREL, Lancaster University.
For technical queries please contact Paul Rayson : paul@comp.lancs.ac.uk

My folders



























Wmatrix

You are logged in as: paul

[[Admin](#) > [Admin](#)]

[[My folders](#) | [Tag wizard...](#) | [Switch to Advanced Interface](#) | [Help](#) | [Feedback](#)]

[**You are here** > [My folders](#)]

 A1NA1P	 A1N_fragment09	 A1X_fragment05	 Amsterdam_aca
 Amsterdam_dem	 Amsterdam_fic	 Amsterdam_news	 BusMagWM
 CCID_Edited102	 CCID_Edited89	 CCID_Translated	 CCID_Translated102
 CCID_Translated89	 CuckooAll	 CuckooPost	 CuckooPre
 Dutch	 EWJ	 FujitsuPreCourseFeb08	 LabourManifesto2001
 LabourManifesto2005	 Lamp70	 LampManChk	 LampeterSample
 LibDemManifesto2001	 LibdemManifesto2005	 NORMDATA	 PWE-ALT-001

Wmatrix2 - Windows Internet Explorer

http://ucrel.lancs.ac.uk/wmatrix2.html

Google

Language and Spe Go Bookmarks 120 blocked Check AutoFill Send to Settings

Home page - Lancaster Univ... Wmatrix2 Centre for Research in Appli...

You are here > My folders > LibdemManifesto2005

You can see various views on this dataset:

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

major	6
majority	4
make	37
make_a_difference	1
make_sure	22
makes	3
makes_sure	1

- 2. Word search**
Enter search term

have been the real oppositi
ID cards . The challenge and
ilence and the opportunity i
to provide the real alternat
p . And at the heart of our
i delivers the social priori
ple want . The mark of a dec

- 3. Word clouds**
Compared to:
Entrepreneurship Small Business Spoken (CESB)

democracy
nent green i it lik
o ministers n't oppose pa
secretary shadow tax t:

- 4. Tag clouds**
Compared to:
Entrepreneurship Small Business Spoken (CESB)

Frequent **Govern**
Money:_Debt
Politics Time:_Ending

Wmatrix

©2000-8 UCREL, Lancaster University.
For technical queries please contact Paul Rayson : paul@comp.lancs.ac.uk

Done Local intranet 100%

Underused items are shown in italics.
Move your mouse over each item to show extra information in a tooltip.
Click on a word to show the concordance.

'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 governments 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

Freq=13 LL=+23.61

Wmatrix

©2000-8 UCREL, Lancaster University.
For technical queries please contact Paul Rayson : paul@comp.lancs.ac.uk

Wmatrix2 - Windows Internet Explorer

http://ucrel.lancs.ac.uk/wmatrix2.html

Google Language and Spe Go 120 blocked Check AutoFill Send to Settings

Home page - Lancaster Univ... Wmatrix2 Centre for Research in Appli...

[Admin > Admin]
[My folders | Tag wizard... | Switch to Advanced Interface | Help | Feedback]

[You are here > My folders > LibdemManifesto2005]

Key domain cloud

Larger items are more significant.
Underused items are shown in italics.
Move your mouse over each item to show extra information in a tooltip.
Click on a word to show the concordance.

Allowed Business:_Selling Business:_Generally Caution **Colour_and_colour_patterns** Comparing:_Different

Degree:_Boosters Distance:_Far Ethical **Evaluation:_Bad**

Evaluation:_Inaccurate Evaluation:_Authentic Evaluation:_Bad Evaluation:_True Exceed:_waste

Failure **Farming_&_Horticulture** Frequent **Government**

Green_issues Hindering Measurement:_Size **Money:_Debts**

Money:_Cost_and_price Money:_Affluence **Other_proper_names**

Paper_documents_and_writing Personal_names **Politics** Pronouns Strong_obligation_or_necessity

Thought,_belief **Time:_Ending** Time:_Late Time:_Old,_grown-up **Unethical** Unwanted Using

Vehicles_and_transport_on_land Worry

Wmatrix

©2000-8 UCREL, Lancaster University.
For technical queries please contact Paul Rayson : paul@comp.lancs.ac.uk

LANCASTER UNIVERSITY

Local intranet 100%

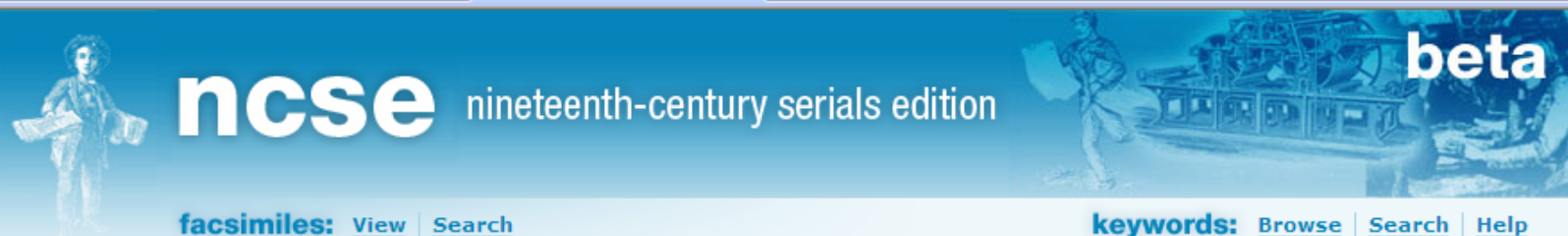
Key domain case studies

1. An exploration of the semantic field of 'love' in Shakespeare's comedies and tragedies (Archer et al, forthcoming)
2. Novel browsing indexes for the Nineteenth-Century Serials Edition: a free, online edition of six nineteenth-century periodicals and newspapers (www.ncse.kcl.ac.uk)
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.





Home

About ncse

Editorial Commentary

History of the Project

Reference

ABOUT THE TITLES:

THE
MONTHLY REPOSITORY
OF
Theology and General Literature.

The Northern Star,
AND LEEDS GENERAL ADVERTISER.

Leader.
A POLITICAL AND LITERARY REVIEW, MERCANTILE JOURNAL.

THE
ENGLISH WOMAN'S JOURNAL.
PUBLISHED MONTHLY.

THE TOMAHAWK:
A SATURDAY JOURNAL OF SATIRE.

THE PUBLISHERS' CIRCULAR

Browse by Subject

[A: General and Abstract Terms](#)
[B: The Body & the Individual](#)
[C: Arts & Crafts](#)
[E: Emotional Actions, States & Processes](#)
[F: 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](#)
[O: Substances, Materials, Objects & Equipment](#)
[P: Education](#)
[Q: 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](#)

facsimiles: View Search

keywords: Browse Search Help

Home

About ncse

Editorial Commentary

History of the Project

Reference

ABOUT THE TITLES:

THE
MONTHLY REPOSITORY
OF
Theology and General Literature.

The Northern Star,
AND LEEDS GENERAL ADVERTISER.

Leader.
A POLITICAL AND LITERARY REVIEW, MERCANTILE JOURNAL.

THE
ENGLISH WOMAN'S JOURNAL.
PUBLISHED MONTHLY.

THE TOMAHAWK:
A SATURDAY JOURNAL OF SATIRE.

THE PUBLISHERS' CIRCULAR

Browse by Subject: 1 - 10 (of 6806)

Filtering by: Entertainment, Sports and Games
 Entertainment, Sports and Games > Drama,
 the theatre and show business
 Sorted by: Relevance

<< first < prev 1 2 3 4 next > last >>

Leader and Saturday Analyst

Leader - Town Edition (14/01/1860) Vol. second series 1 No.
512 Page 30

9 x 14"; 24 x 35cm. Price 5d::6d.

View facsimile: [item](#) | [page](#) | [issue](#) View extracted keywords

Leader. A Political, Literary, Commercial, and Family Weekly Newspaper, and Record

Leader - Country Edition (31/12/1859) Vol. 10 No. 510 Page
1416

9 x 14"; 24 x 35cm. Price 5d::6d.

View facsimile: [item](#) | [page](#) | [issue](#) View extracted keywords

Leader. A Political and Literary Review

Leader - Country Edition (07/06/1856) Vol. 7 No. 324 Page 549
9 x 14"; 24 x 35cm. Price 5d::6d.

View facsimile: [item](#) | [page](#) | [issue](#) View extracted keywords

Leader

Leader - Town Edition (24/04/1852) Vol. 3 No. 109 Page 403
9 x 14"; 24 x 35cm. Price 6d.

View facsimile: [item](#) | [page](#) | [issue](#) View extracted keywords

Leader. A Political and Literary Review

Leader - Town Edition (27/06/1857) Vol. 8 No. 379 Page 623
9 x 14"; 24 x 35cm. Price 6d::5d.

View facsimile: [item](#) | [page](#) | [issue](#) View extracted keywords

Northern Star, and National Trades' Journal

Northern Star - Edition 1 (13/06/1846) Vol. 10 No. 448 Page 4

CATEGORIES

- * General and Abstract Terms
- * The Body & the Individual
- * Arts & Crafts
- * Emotional Actions, States & Processes
- * Food & Farming
- * Government & the Public Domain
- * Architecture, Buildings, Houses & the Home
- * Money & Commerce in Industry
- * Entertainment, Sports and Games
 - ▶ Entertainment generally
 - ▶ Music and related activities
 - ▶ Recorded sound
 - ▶ Drama, the theatre and show business
 - ▶ Sports and games generally
 - ▶ Sports
 - ▶ Games
 - ▶ Children's games and toys
- * Life and living things
- * Movement, Location, Travel & Transport
- * Numbers & Measurement
- * Substances, Materials, Objects & Equipment
- * Education
- * Linguistic Actions, States & Processes: Communication
- * Social Actions, States & Processes
- * Time
- * The World & Our Environment
- * Psychological Actions, States & Processes
- * 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)