Using Wmatrix: corpus analysis and comparison tool

Paul Rayson
School of Computing and Communications
Lancaster University
p.rayson@lancaster.ac.uk
@perayson

MELC workshop
10th January 2014
Session Outline

11:00 – basic introduction

11:15 – hands-on
   – explore manifesto data, key words and domains

12:00 – hands-on with MELC data
   – Patients and Professionals interviews

12:30 – Break for lunch
   – Room A87
Wmatrix main points

• Web-based (c.f. BNCweb, CQPweb)
• You can load your own (English) data
• Incorporates main methods in corpus linguistics toolbox
  – frequency lists, concordances, key words, collocations, n-grams (coming back in 2014)
• Adds two levels of linguistic annotation (NLP or computational linguistics methods)
  – POS tagging, Semantic field tagging
• Novelty
  – key domain analysis, semantic collocations
## Semantic tags
(aka domains, fields, categories)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General and abstract terms</td>
<td>The body and the individual</td>
<td>Arts and crafts</td>
<td>Emotion</td>
</tr>
<tr>
<td>F</td>
<td>Food and farming</td>
<td>Government and public</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Entertainment, sports and games</td>
<td>Life and living things</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>Substances, materials, objects and equipment</td>
<td>Education</td>
<td>Language and communication</td>
<td>Social actions, states and processes</td>
</tr>
<tr>
<td>T</td>
<td>Time</td>
<td>World and environment</td>
<td>Psychological actions, states and processes</td>
<td>Science and technology</td>
</tr>
<tr>
<td>Z</td>
<td>Names and grammar</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Key words

## Text

## Keywords

<table>
<thead>
<tr>
<th>Word</th>
<th>LitDocs</th>
<th>US</th>
<th>Text</th>
<th>O/U-use</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. to</td>
<td>105.33</td>
<td>0.00</td>
<td>0.00</td>
<td>+</td>
<td>81.40</td>
<td></td>
</tr>
<tr>
<td>2. the</td>
<td>42.14</td>
<td>0.85</td>
<td>0.85</td>
<td>+</td>
<td>70.59</td>
<td></td>
</tr>
<tr>
<td>3. in</td>
<td>38.67</td>
<td>0.85</td>
<td>0.85</td>
<td>+</td>
<td>69.29</td>
<td></td>
</tr>
<tr>
<td>4. is</td>
<td>38.67</td>
<td>0.85</td>
<td>0.85</td>
<td>+</td>
<td>69.29</td>
<td></td>
</tr>
<tr>
<td>5. be</td>
<td>38.67</td>
<td>0.85</td>
<td>0.85</td>
<td>+</td>
<td>69.29</td>
<td></td>
</tr>
<tr>
<td>6. a</td>
<td>38.67</td>
<td>0.85</td>
<td>0.85</td>
<td>+</td>
<td>69.29</td>
<td></td>
</tr>
</tbody>
</table>

## Word frequency list

- the: 351
- of: 243
- a: 225
- and: 163
- to: 139
- in: 134
- is: 123
- be: 83
- for: 81
- that: 77
- which: 66
- are: 64
- by: 60
- words: 57
- a: 56
- in: 50
- not: 48
- or: 46
- phrases: 44

## Text or reference corpus

## Word frequency list

- the: 351
- of: 243
- a: 225
- and: 163
- to: 139
- in: 134
- is: 123
- be: 83
- for: 81
- that: 77
- which: 66
- are: 64
- by: 60
- words: 57
- a: 56
- in: 50
- not: 48
- or: 46
- phrases: 44
Log-likelihood (LL)

• Wizard online at:
  • http://ucrel.lancs.ac.uk/llwizard.html
• Spreadsheet also available for download

• Also see:
  • http://corpora.lancs.ac.uk/sigtest/
Wmatrix version 3
Practical one

- 2005 general election
  - Liberal Democrat party manifesto
  - Labour party manifesto
- 2010 general election
  - manifestos for all three main parties
  - TV debates (need to be converted from PDF)
- Aims:
  - To help you understand the basic Wmatrix features
  - To give you some awareness of the semantic tagset
- (Option) Use your own data!
Open two web-browser windows

• Both URLs linked from Wmatrix home page:
  – http://ucrel.lancs.ac.uk/wmatrix/

1. Wmatrix tutorial
  – http://ucrel.lancs.ac.uk/wmatrix/tutorial/

2. Wmatrix tool:
  – http://ucrel.lancs.ac.uk/wmatrix3.html
  – Login details:
    • Username:
    • Password:
• http://ucrel.lancs.ac.uk/wmatrix/tutorial/

• On your own or in small groups:
  – **Read** tutorials A and B (the actions are already done)
  – **Do** tutorial C (key words, key domains and concordances)

• Advanced users:
  – Tutorial D (advanced data analysis) on your own or in small groups
  – Suggested timings:
    • Steps D.3 and D.4 (10 minutes)
    • Spend most of your time from step D.5 onwards (remainder of the hour)

• Notes:
  – you can use your own data and your own username if you have them
  – Ask questions anytime
  – Keep going until the end of the hour
New and planned features

• CrossTabs
• Concordance
  – highlighting and filtering by context
  – concgrams-style
• Collocations and semantic collocations
• N-grams and C-grams
  – Aka clusters, lexical bundles
  – Faster implementation (L-gram)
  – http://code.google.com/p/lgram/
• Visualisations
  – Collocation Network Explorer (CONE)
• Replace indexing system
  – much larger corpora
• Other languages …
Practical two

• MELC data
  – MELC_CC_PatientsInterview
  – MELC_CC_ProfessionalsInterview

• Aims
  – To explore and compare the two datasets using the techniques that you have learnt so far

Switch to the advanced interface and compare the texts using key words and key domains methods
References

• Useful background reading (keyness, annotation and MWE):


• Wmatrix, CLAWS and USAS websites:
  – [http://ucrel.lancs.ac.uk/wmatrix/](http://ucrel.lancs.ac.uk/wmatrix/)
  – [http://ucrel.lancs.ac.uk/claws/](http://ucrel.lancs.ac.uk/claws/)
  – [http://ucrel.lancs.ac.uk/usas/](http://ucrel.lancs.ac.uk/usas/)


Further reading

- **Further reading (mostly key words related).**