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)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>General and abstract terms</td>
<td>B</td>
<td>The body and the individual</td>
<td>C</td>
</tr>
<tr>
<td>F</td>
<td>Food and farming</td>
<td>G</td>
<td>Government and public</td>
<td>H</td>
</tr>
<tr>
<td>K</td>
<td>Entertainment, sports and games</td>
<td>L</td>
<td>Life and living things</td>
<td>M</td>
</tr>
<tr>
<td>O</td>
<td>Substances, materials, objects and equipment</td>
<td>P</td>
<td>Education</td>
<td>Q</td>
</tr>
<tr>
<td>T</td>
<td>Time</td>
<td>W</td>
<td>World and environment</td>
<td>X</td>
</tr>
<tr>
<td>Z</td>
<td>Names and grammar</td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>the</td>
<td>351</td>
<td>1444</td>
<td>of</td>
<td>343</td>
</tr>
<tr>
<td>a</td>
<td>339</td>
<td>2181</td>
<td>a</td>
<td>343</td>
</tr>
<tr>
<td>and</td>
<td>153</td>
<td>134</td>
<td>and</td>
<td>153</td>
</tr>
<tr>
<td>to</td>
<td>139</td>
<td>123</td>
<td>to</td>
<td>139</td>
</tr>
<tr>
<td>in</td>
<td>139</td>
<td>123</td>
<td>in</td>
<td>139</td>
</tr>
<tr>
<td>is</td>
<td>123</td>
<td>123</td>
<td>is</td>
<td>123</td>
</tr>
<tr>
<td>be</td>
<td>83</td>
<td>83</td>
<td>be</td>
<td>83</td>
</tr>
<tr>
<td>for</td>
<td>83</td>
<td>83</td>
<td>for</td>
<td>83</td>
</tr>
<tr>
<td>that</td>
<td>67</td>
<td>67</td>
<td>that</td>
<td>67</td>
</tr>
<tr>
<td>which</td>
<td>66</td>
<td>66</td>
<td>which</td>
<td>66</td>
</tr>
<tr>
<td>are</td>
<td>64</td>
<td>64</td>
<td>are</td>
<td>64</td>
</tr>
<tr>
<td>by</td>
<td>60</td>
<td>60</td>
<td>by</td>
<td>60</td>
</tr>
<tr>
<td>words</td>
<td>57</td>
<td>57</td>
<td>words</td>
<td>57</td>
</tr>
<tr>
<td>a</td>
<td>53</td>
<td>53</td>
<td>a</td>
<td>53</td>
</tr>
<tr>
<td>at</td>
<td>50</td>
<td>50</td>
<td>at</td>
<td>50</td>
</tr>
<tr>
<td>not</td>
<td>48</td>
<td>48</td>
<td>not</td>
<td>48</td>
</tr>
<tr>
<td>or</td>
<td>48</td>
<td>48</td>
<td>or</td>
<td>48</td>
</tr>
<tr>
<td>phrases</td>
<td>44</td>
<td>44</td>
<td>phrases</td>
<td>44</td>
</tr>
</tbody>
</table>

**Keywords**

**Text or reference corpus**

**Word frequency list**

**Word frequency list**

**Key words**

**Text**
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

To Do: 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).**