Wmatrix corpus analysis and comparison tool: practical hands-on workshop

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

UCREL seminar
18th May 2017
Choose your own workshop

1. Newbies
   • if you have never used Wmatrix before or only started recently

2. Oldies
   • if you are an experienced user who has been to a Wmatrix workshop before

3. OR
   • make it up as you go along
Workshop Outline

3:00 – basic introduction

3:15 – hands-on
  – load your own data, key words and domains (newbies)
  – MWE extraction (oldies)

4:00 – new and future features intro
  – CrossTab, multilingual taggers

4:10 – hands-on
  – CrossTab, Collocations, semantic collocations, multilingual taggers etc
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 Wmatrix4)
• 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>G</td>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>Government and public</td>
<td>Architecture, housing and the home</td>
<td>Money and commerce in industry</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Entertainment, sports and games</td>
<td>L</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Life and living things</td>
<td>Movement, location, travel and transport</td>
<td>Numbers and measurement</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>Substances, materials, objects and equipment</td>
<td>P</td>
<td>Q</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>Language and communication</td>
<td>Social actions, states and processes</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>Time</td>
<td>W</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>World and environment</td>
<td>Psychological actions, states and processes</td>
<td>Science and technology</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>Names and grammar</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Key words

Text

Keywords

Text or reference corpus

Word frequency list

Word frequency list
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/
• Consider effect size measures, as well as range and dispersion (depending on your corpus)
Significance vs effect size

- **Experiment 1**
  - $f(\text{blah, corpus1}) = 100$
  - $f(\text{blah, corpus2}) = 50$
  - corpus 1 & 2 sizes = 10,000
  - Sig_LL = 16.99 Effect_LR = 1.00

- **Experiment 2**
  - $f(\text{ping, corpus3}) = 1,000$
  - $f(\text{ping, corpus4}) = 500$
  - corpus 3 & 4 sizes = 100,000
  - Sig_LL = 169.90 Effect_LR = 1.00

- **Experiment 3**
  - $f(\text{hoot, corpus3}) = 1,000$
  - $f(\text{hoot, corpus4}) = 824$
  - corpus 3 & 4 sizes = 100,000
  - Sig_LL = 17.01 Effect_LR = 0.28
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/

Newbies
- 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)

Oldies
- Tutorial D (advanced data analysis) on your own or in pairs
  - Suggested timings:
    - Steps D.3 and D.4 (10 minutes)
    - Spend most of your time from step D.5 onwards (remainder of the hour)

OR
- Tutorial E on your own or in pairs

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)
• Visualisations
  – Collocation Network Explorer (CONE)
  – GraphColl and #LancsBox
• Historical Thesaurus Semantic Tagger
• Semantic taggers for 14 languages ...
  – http://ucrel.lancs.ac.uk/usas/
New and planned features

• For linking direct to your own software, there is a Java API:
  – Jmatrix
  – Jane Ransom, Ricardo Gacitua
  – allows upload of data and download of tagged result plus frequency lists and key words

• Under development, a PHP API
  – PHPmatrix
  – Will Simm, VoiceYourView project
  – same functionality as above
Collocations and semantic collocations

<table>
<thead>
<tr>
<th>100 occurrences.</th>
<th>Extend context</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New Labour 's</strong> record: The contract delivered <strong>40;69;TOOLONG 10</strong> Chapter 1 Economy: <em>Rising</em> prosperity in an opp 1 More</td>
<td>1 More</td>
</tr>
<tr>
<td>: we showed we could run the economy well, cut crime, and stand 2 More</td>
<td>2 More</td>
</tr>
<tr>
<td><strong>40;69;TOOLONG 10</strong> Chapter 1 Economy: <em>Rising</em> prosperity in an opp well, cut crime, and stand 3 More</td>
<td>3 More</td>
</tr>
<tr>
<td>rue the modernisation of our economy, coupled with investment and spreading prosperity through 4 More</td>
<td>4 More</td>
</tr>
<tr>
<td>rd instead to an opportunity economy, spreading prosperity through 5 More</td>
<td>5 More</td>
</tr>
<tr>
<td>we will cement a new social contract with rights matched by respons 6 More</td>
<td>6 More</td>
</tr>
<tr>
<td>ts future. It is our social contract: we help you, you help your 7 More</td>
<td>7 More</td>
</tr>
<tr>
<td>9 New Labour 's record: The contract delivered Our country is chan 8 More</td>
<td>8 More</td>
</tr>
<tr>
<td>n now has a stable, growing economy with the lowest inflation sin 9 More</td>
<td>9 More</td>
</tr>
<tr>
<td>rum. There are 300,000 more businesses, providing jobs and increa 10 More</td>
<td>10 More</td>
</tr>
<tr>
<td>e New Labour 's record: The contract delivered There are record nu 11 More</td>
<td>11 More</td>
</tr>
<tr>
<td>first two years. Chapter 1 Economy: <em>Rising</em> prosperity in an opp 12 More</td>
<td>12 More</td>
</tr>
<tr>
<td>Britain, the fourth largest economy in the world, with the longe 13 More</td>
<td>13 More</td>
</tr>
<tr>
<td>ing a low debt/high employment economy which generates investment in 14 More</td>
<td>14 More</td>
</tr>
<tr>
<td>public services; supporting enterprise and wealth creation by making 15 More</td>
<td>15 More</td>
</tr>
<tr>
<td>Britain the best place to do business; and helping every part of B 16 More</td>
<td>16 More</td>
</tr>
<tr>
<td>in from the strength of our economy. And as we work globally to 17 More</td>
<td>17 More</td>
</tr>
<tr>
<td>could not be trusted with the economy. We are winning the argument 18 More</td>
<td>18 More</td>
</tr>
<tr>
<td>to economic stability. Our economy has <em>grown</em> in every quarter wi 19 More</td>
<td>19 More</td>
</tr>
<tr>
<td>and the needs of families, business and the environment. We will 20 More</td>
<td>20 More</td>
</tr>
<tr>
<td>nd VAT to food, children 's regime that supports British business 21 More</td>
<td>21 More</td>
</tr>
<tr>
<td>'s tax to its lowest ever level 22 More</td>
<td>22 More</td>
</tr>
<tr>
<td>seeking services. Supporting enterprise Government does not create we 23 More</td>
<td>23 More</td>
</tr>
<tr>
<td>planning, and transport. The economy of the future will be based o 24 More</td>
<td>24 More</td>
</tr>
<tr>
<td>25 More</td>
<td>25 More</td>
</tr>
</tbody>
</table>
Negative semantic prosody

1 a real downer of a word, likely to cause a lot of confusion. After
2 until I told her. What I did has caused a rift between me and my friend’
3 nd. His decision to oppose the war caused amazement in the ranks of SNP
4 quality in the borough”. The award caused anger among anti-racist groups a
5 of new media outlets and choice, causing audience fragmentation. To achi
6 een suffering from a kidney stone, caused by excess build-up of urine. No
7 s chronic unemployment problem is caused by featherbedding workers with
8 de, Isle of Wight, last June which caused damage estimated at £100,000.
9 in some carrots could cause dizziness and vomiting, but said
10 control fairly rapidly. Hepatitis causes inflammation of the liver and ca
11 minority of individuals intent on causing misery to their neighbours. The
12 ews of the government’s proposals caused outrage among medical associatio
13 issues, but it is starting to cause problems of supply. The traditio
14 of things in your life that are causing real concern. Don’t be complete
15 se that her disappearance did not cause serious concern. The social
16 against the gynaecologist who caused so much unnecessary pain and mis
17 Saddam, journalism so strong it caused the West to liberate northern Ir
18 No one knows quite what caused this frenzy. The setting up of
Semantic prosody / preference

- ‘a form of meaning which is established through the proximity of a consistent series of collocates’ (Louw 2000: 57)
- ‘The collocational meaning arising from the interaction between a given node and its typical collocates might be referred to as *semantic prosody*’ (Xiao and McEnery, 2006)
- *semantic preference* defined ‘by a lexical set of frequently occurring collocates [sharing] some semantic feature’ (Stubbs 2002: 449)
- *semantic prosody* ‘dictates the general environment which constrains the preferential choices of the node item’, *semantic preference* ‘contributes powerfully’ to building semantic prosody (Partington 2004: 151)
- Semantic preference can be viewed as a feature of the collocates while semantic prosody is a feature of the node word
<table>
<thead>
<tr>
<th>Author</th>
<th>Negative prosody</th>
<th>Positive prosody</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinclair (1991)</td>
<td>BREAK out</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HAPPEN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SET in</td>
<td></td>
</tr>
<tr>
<td>Louw (1993, 2000)</td>
<td>bent on</td>
<td>BUILD up a</td>
</tr>
<tr>
<td></td>
<td>build up of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>END up <em>verb</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GET oneself <em>verb</em>ed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a recipe for</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAUSE</td>
<td>career</td>
</tr>
<tr>
<td></td>
<td>FAN the flame</td>
<td></td>
</tr>
<tr>
<td></td>
<td>signs of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>underage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teenager(s)</td>
<td></td>
</tr>
<tr>
<td>Partington (1998)</td>
<td>COMMIT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PEDDLE/peddler</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dealings</td>
<td></td>
</tr>
<tr>
<td>Hunston (2002)</td>
<td>sit through</td>
<td></td>
</tr>
<tr>
<td>Schmitt and Carter (2004)</td>
<td>bordering on</td>
<td></td>
</tr>
</tbody>
</table>
## Multilingual semantic lexicons

<table>
<thead>
<tr>
<th>Language</th>
<th>Single word entries</th>
<th>Multiword expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>31,154</td>
<td>0</td>
</tr>
<tr>
<td>Chinese</td>
<td>64,541</td>
<td>19,048</td>
</tr>
<tr>
<td>Czech</td>
<td>28,161</td>
<td>0</td>
</tr>
<tr>
<td>Dutch</td>
<td>4,220</td>
<td>0</td>
</tr>
<tr>
<td>Finnish</td>
<td>46,225</td>
<td>4,422</td>
</tr>
<tr>
<td>Italian</td>
<td>13,098</td>
<td>5,622</td>
</tr>
<tr>
<td>Malay</td>
<td>64,863</td>
<td>0</td>
</tr>
<tr>
<td>Portuguese</td>
<td>13,499</td>
<td>1,781</td>
</tr>
<tr>
<td>Russian</td>
<td>17,443</td>
<td>713</td>
</tr>
<tr>
<td>Spanish</td>
<td>3,665</td>
<td>0</td>
</tr>
<tr>
<td>Urdu</td>
<td>1,765</td>
<td>235</td>
</tr>
<tr>
<td>Welsh</td>
<td>2,744</td>
<td>0</td>
</tr>
</tbody>
</table>
Practical two

• 2005 general election
  – Liberal Democrat party manifesto
  – Labour party manifesto
• 2010 general election
  – manifests for all three main parties
  – TV debates (need to be converted from PDF)

• Aims
  – To understand why you might want to use semantic collocations and how to calculate them using the Wmatrix interface

• (Option) Use your own data! Or experiment with the semantic taggers for other languages.
• http://ucrel.lancs.ac.uk/wmatrix/tutorial/

• Newbies
  – Can continue with Tutorials A, B, C or D if you wish

• Everyone else
  – Tutorial F on your own or in small groups

• Keep going until the end of the session
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).**