Wmatrix for forensic linguistics: a practical hands-on demo

Slides at http://ucrel.lancs.ac.uk/paul/

Dr Paul Rayson
@perayson
Director of UCREL research centre
School of Computing and Communications
Lancaster University, UK
Semantic field annotation has applications for conceptual or topic tagging:

- There’s been more violence in the Basque country in northern Spain: one policeman has been killed, and two have been injured in a grenade and machine-gun attack on their patrol car.

- E3 = emotional states; Z2 = geographical names; M7 = places; M6 = location and direction; G3 = warfare; M3 = land transportation.
The work of many hands ...

- Joint research with
  - Geoffrey Leech
  - Roger Garside
  - Jenny Thomas
  - Andrew Wilson
  - Dawn Archer
  - Scott Piao
  - Sheryl Prentice
UCREL Semantic Analysis System (USAS)

- Full text tagging, not just selected words (c.f. Diction, LIWC, RID)
- Tagging the coarse-grained sense in context, not just the word
- Not task specific categories
- Flexible category set with hierarchical structure
- Words and multi-word expressions (MWE) e.g. phrasal verbs (stubbed out), noun phrases (riding boots), proper names (United States of America), true idioms (living the life of Riley)
Semantic fields

• AKA concepts, semantic domains
• ‘groups together word senses that are related by virtue of their being connected at some level of generality with the same mental concept’
• Not only synonymy and antonymy but also hypernymy and hyponymy
• E.g. EDUCATION: academic, coaching, coursework, deputy head, exams, PhD, playschool, revision notes, studious, swot, viva
<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</td>
<td>The body and the individual</td>
<td>Arts and crafts</td>
<td>Emotion</td>
</tr>
<tr>
<td></td>
<td>terms</td>
<td></td>
<td></td>
<td></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></td>
<td>Government and public</td>
<td>Architecture, housing</td>
<td>Money and commerce in</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>and the home</td>
<td>industry</td>
</tr>
<tr>
<td>K</td>
<td>Entertainment,</td>
<td>L</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>sports and games</td>
<td>Life and living things</td>
<td>Movement, location,</td>
<td>Numbers and measurement</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>travel and transport</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>Substances, materials,</td>
<td>P</td>
<td>Q</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>objects and equipment</td>
<td>Education</td>
<td>Language and</td>
<td>Social actions, states</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>communication</td>
<td>and processes</td>
</tr>
<tr>
<td>T</td>
<td>Time</td>
<td>W</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>World and environment</td>
<td>Psychological actions,</td>
<td>Science and technology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>states and processes</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>Names and grammar</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lexical resources

- Lexicon of 56,316 items
  - presentation NN1 Q2.2 A8 S1.1.1 K4
- MWE list of 18,971 items
  - travel_NN1 card*_NN* M3/Q1.2
- A small wildcard lexicon
  - *kg NNU N3.5
- Unknown words using WordNet synonym lookup
Disambiguation methods (1)

1. POS tag
   - *spring* noun [season sense] [coil sense]
   - *spring* verb [jump sense]

2. General likelihood ranking for single-word and MWE tags
   - *green* referring to [colour] is generally more frequent than *green* meaning [inexperienced]

3. Overlapping MWE resolution
   - Heuristics applied: semantic MWEs override single word tagging, length and span of MWE also significant
Disambiguation methods (2)

4. Domain of discourse
   - adjective *battered*
     - [Violence] (e.g. battered person)
     - [Judgement of Appearance] (e.g. battered car)
     - [Food] (e.g. battered cod)

5. Text-based disambiguation
   - one sense per text

6. Template rules
   - *Auxiliary verbs (be/do/have)*
   - *account* of NP [narrative]
   - balance of xxx *account* [financial]
Evaluation (modern data)

- Hand tagged test corpus of 124,839 words
- Error rate of 8.95%
- Ambiguity ratio 47.73%
- Reduced to 17.06% by disambiguation
- Not all ambiguity is resolved, but 1st choice tag selection gives 91% accuracy.
KEY SEMANTIC DOMAINS
Significance and effect size

- Log-likelihood (LL) Wizard online at:
  - [http://ucrel.lancs.ac.uk/llwizard.html](http://ucrel.lancs.ac.uk/llwizard.html)

- Spreadsheet and code also available for download
  - [https://github.com/UCREL/SigEff](https://github.com/UCREL/SigEff)

- Very important to consider dispersion and effect size measures (depending on your corpus) – included in Wmatrix CrossTab feature and keyness measures
  - See the work of Hardie, Gabrielatos, Rayson and Potts (forthcoming)
Significance versus effect size

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

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

• Experiment 3
  – \( f(\text{hoot, corpus3}) = 1,000 \)
  – \( f(\text{hoot, corpus4}) = 824 \)
  – corpus 3 & 4 sizes = 100,000
  – \( \text{Sig_LL} = 17.01 \) Effect_LR = 0.28
Figure 1: keywords in LibDem 2010 manifesto

2020 2050 affordable allow banking banks believe better Britain budget businesses
carbon change child climate create crime cut deficit democrats developing_countries
economy education emissions energy ensure environment establish EU
every fair fairness finances financial for funding future give global government
health help homes improve increase infrastructure insulate introduce jobs justice liberal
local localAuthorities long-term manifesto money mutuals need NHS our over_time paid pay
people politics polluting power protect public reduce reducing reform reforming
renewable replace restore review savings schools scrap seek services
so that spending state_pension such_as support sustainability
sustainable system target targets tax taxes to UK UN unfair we will
Figure 2: key domains (semantic fields) in LibDem 2010 manifesto

- **Law and order**: law, prison(s, ers), loopholes, security, police (force, officer, station, services)...

---

**Key Domains:**
- Able/intelligent
- Alive
- Allowed
- Attentive
- Business
- Business: Generally
- Chance, luck
- Change
- Cheap
- Confident
- Constraint
- Crime
- Danger
- Degree
- Deserving
- Education in general
- Entire; maximum
- Ethical
- Evaluation: Good
- Evaluation: Authentic
- Exceed; waste
- Expensive
- Expensive
- General actions / making
- Getting and giving; possession
- Giving
- Government
- Green issues
- Green issues
- Health and disease
- Helping
- Hindering
- Important
- Inclusion
- Interested/excited/energetic
- Law and order
- Lawful
- Location and direction
- Long, tall and wide
- Medicines and medical treatment
- Mental object
- Means, method
- Money and payment
- Money: Affluence
- Money: Lack
- Money: Affluence
- No constraint
- No obligation or necessity
- Other proper names
- Participating
- People
- Places
- Politics
- Putting, pulling, pushing, transporting
- Quantities: little
- Quantities: many/much
- Relationship
- Residence
- Safe
- Safe
- Science and technology in general
- Social actions, states and processes
- Strong obligation or necessity
- Success
- The Media
- The universe
- Time period; long
- Time: Future
- Time: Ending
- Time: New and young
- Time: Beginning
- Tough/strong
- Tough/strong
- Unethical
- Wanted
- Weather
- Work and employment: Generally
Example applications and studies

• UK General Election Manifestos (Rayson 2008)
• Around 100 papers listed at http://ucrel.lancs.ac.uk/wmatrix/
• Metaphor in end-of-life care (MELC) http://ucrel.lancs.ac.uk/melc/
• Encyclopaedia of Shakespeare's Language http://wp.lancs.ac.uk/shakespearelang/
Example applications and studies


Motivations, attribution of blame, assumption of agency. 21/70 Biber categories, MDA, tagged using Wmatrix & ICE tagsets.

ML model based on keywords, geo-spatial analysis, frequencies, semantic & sentiment analysis, key semantic tag analysis.

Fraudulent vs genuine papers: key semantic tags. Caution: Not suitable for prediction!

CDA & CL. SketchEngine & Wmatrix: frequency, collocation, concordance & key semantic tag analysis.
Example applications and studies


Online child protection


Frequency, key words and key semantic tags, alongside a large number of other features & ML model.
Key points

• Web-based (c.f. BNCweb, CQPweb, SketchEngine)
• Dedicated server, Secure HTTPS access
• You can load your own data (English currently in v4, Multilingual coming soon)
• Incorporates main methods in corpus linguistics toolbox
  – frequency lists, concordances, key words, collocations, n-grams
• Adds two levels of linguistic annotation (NLP methods)
  – POS tagging, Semantic field tagging
• Novelty
  – key domain analysis, semantic collocations
Hands-on Practical

• 2005 UK general election
  – Liberal Democrat party manifesto
  – Labour party manifesto
• 2010 UK general election
  – manifestos for all three main parties
• 2015 & 2017 UK general elections
  – manifestos for seven parties
• Aims:
  – To help you understand the basic Wmatrix features and key domains method
  – To give you some awareness of the semantic tagset
Open two web-browser windows or tabs

- Both URLs linked from Wmatrix home page:
  - [http://ucrel.lancs.ac.uk/wmatrix/](http://ucrel.lancs.ac.uk/wmatrix/)

1. Wmatrix tutorial
   - [http://ucrel.lancs.ac.uk/wmatrix/tutorial/](http://ucrel.lancs.ac.uk/wmatrix/tutorial/)

2. Wmatrix tool:
   - [https://ucrel-wmatrix4.lancaster.ac.uk/](https://ucrel-wmatrix4.lancaster.ac.uk/)
   - Login details:
     - Username: forgeucrelX
       - (where X is the number on your handout)
     - Password:
• [http://ucrel.lancs.ac.uk/wmatrix/tutorial/](http://ucrel.lancs.ac.uk/wmatrix/tutorial/)
• On your own or in small groups ...

• **Read** tutorials A and B (the actions are already done for you)

• **Do** tutorial C (key words, key domains and concordances)

• For the keen ones amongst you, move on to the other tutorials
• You can use your own data if you wish
• Ask questions any time!
Thanks for listening!

• Questions and comments?

• Contact:
  – Email: p.rayson@lancaster.ac.uk
  – Twitter: @perayson
References …

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

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

Acknowledgements

- Wmatrix was initially developed within the REVERE project (REVerse Engineering of Requirements) funded by the EPSRC, project number GR/MO4846, 1998-2001. Collocation Network Explorer (CONE), developed by David Gullick, was partly funded by an EPSRC vacation bursary at Lancaster University in 2010, and incorporates a collocation library designed by Scott Piao. Ongoing maintenance of taggers (e.g. recent Linux porting work by Stephen Wattam), development of new components (e.g. L-gram developed by Eddie Bell and C-grams developed by Andrew Stone) and dictionary updates (e.g. by Sheryl Prentice) are funded by user licence fees. Semantic taggers for new languages are being developed by Scott Piao funded by UCREL. Metaphor extensions have been developed in the MELC project (Metaphor in end-of-life care) funded by the ESRC (grant reference ES/J007927/1). The Historical Thesaurus Semantic Tagger (HTST) was developed in the SAMUELS project (Semantic Annotation and Mark-Up for Enhancing Lexical Searches) funded by the AHRC in conjunction with the ESRC (grant reference AH/L010062/1).