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

English Semantic Tagging

- Semantic field annotation has applications for conceptual or topic tagging:
  - There_Z5 ’s_Z5 been_A3+ more_N5++ violence_E3- in_Z5 the_Z5 Basque_Z2 country_M7 in_Z5 northern_M6 Spain_Z2 :_PUNC one_N1 policeman_G2.1/S2m has_Z5 been_Z5 killed_L1- ,_PUNC and_Z5 two_N1 have_Z5 been_Z5 injured_B2- in_Z5 a_Z5 grenade_G3 and_Z5 machine-gun_G3 attack_G3 on_Z5 their_Z8 patrol-car_M3/G2.1 ._PUNC
  - 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

A General and abstract terms	**B** The body and the individual	**C** Arts and crafts	**E** Emotion
F Food and farming	**G** Government and public	**H** Architecture, housing and the home	**I** Money and commerce in industry
K Entertainment, sports and games	**L** Life and living things	**M** Movement, location, travel and transport	**N** Numbers and measurement
O Substances, materials, objects and equipment	**P** Education	**Q** Language and communication	**S** Social actions, states and processes
T Time	**W** World and environment	**X** Psychological actions, states and processes	**Y** Science and technology
Z Names and grammar			

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

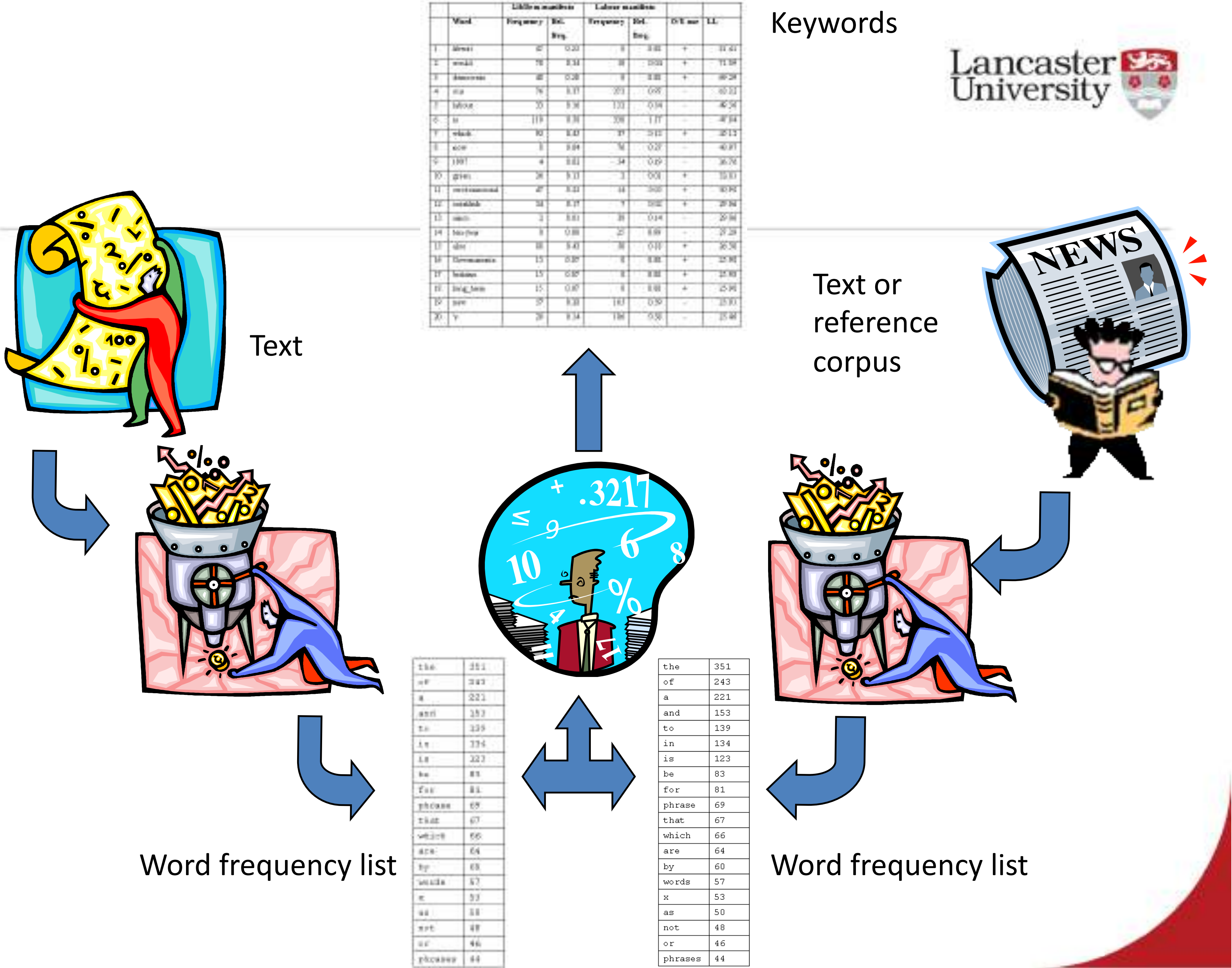
Disambiguation methods (2)

- 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



Keywords

		Lib Dem manifesto		Labour manifesto			
	Word	Frequency	Rel. freq.	Frequency	Rel. freq.	O/E use	LL
1	liberal	67	0.23	0	0.00	+	91.41
2	would	78	0.34	10	0.03	+	71.59
3	democrats	48	0.28	0	0.00	+	69.29
4	our	76	0.17	203	0.66	-	60.22
5	labour	33	0.38	133	0.34	-	42.56
6	is	119	0.38	308	1.17	-	41.04
7	which	90	0.43	37	0.12	+	40.12
8	now	8	0.04	76	0.27	-	40.87
9	1997	4	0.01	54	0.19	-	36.78
10	given	28	0.13	3	0.01	+	33.83
11	environmental	47	0.22	14	0.05	+	30.92
12	establish	38	0.17	7	0.02	+	29.94
13	much	2	0.01	39	0.14	-	29.06
14	[illegible]	8	0.08	25	0.09	-	27.29
15	also	88	0.40	38	0.15	+	26.58
16	Government	15	0.07	0	0.00	+	25.98
17	[illegible]	15	0.07	0	0.00	+	25.98
18	long_term	15	0.07	0	0.00	+	25.98
19	new	57	0.20	163	0.59	-	25.83
20	we	29	0.14	186	0.58	-	25.46

Text

Text or reference corpus

Word frequency list

the	351
of	243
a	221
and	153
to	139
in	134
is	123
be	83
for	81
phrase	69
that	67
which	66
are	64
by	60
words	57
x	53
as	50
not	48
or	46
phrases	44

the	351
of	243
a	221
and	153
to	139
in	134
is	123
be	83
for	81
phrase	69
that	67
which	66
are	64
by	60
words	57
x	53
as	50
not	48
or	46
phrases	44

Word frequency list

Significance and effect size

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

- Spreadsheet and code also available for download
  - 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(blah, corpus1) = 100
  - f(blah, corpus2) = 50
  - corpus 1 & 2 sizes = 10,000
  - Sig_LL = 16.99 Effect_LR = 1.00
- Experiment 2
  - f(ping, corpus3) = 1,000
  - f(ping, corpus4) = 500
  - corpus 3 & 4 sizes = 100,000
  - Sig_LL = 169.90 Effect_LR = 1.00
- Experiment 3
  - f(hoot, corpus3) = 1,000
  - f(hoot, corpus4) = 824
  - corpus 3 & 4 sizes = 100,000
  - Sig_LL = 17.01 Effect_LR = 0.28

Figure 1: keywords in LibDem 2010 manifesto

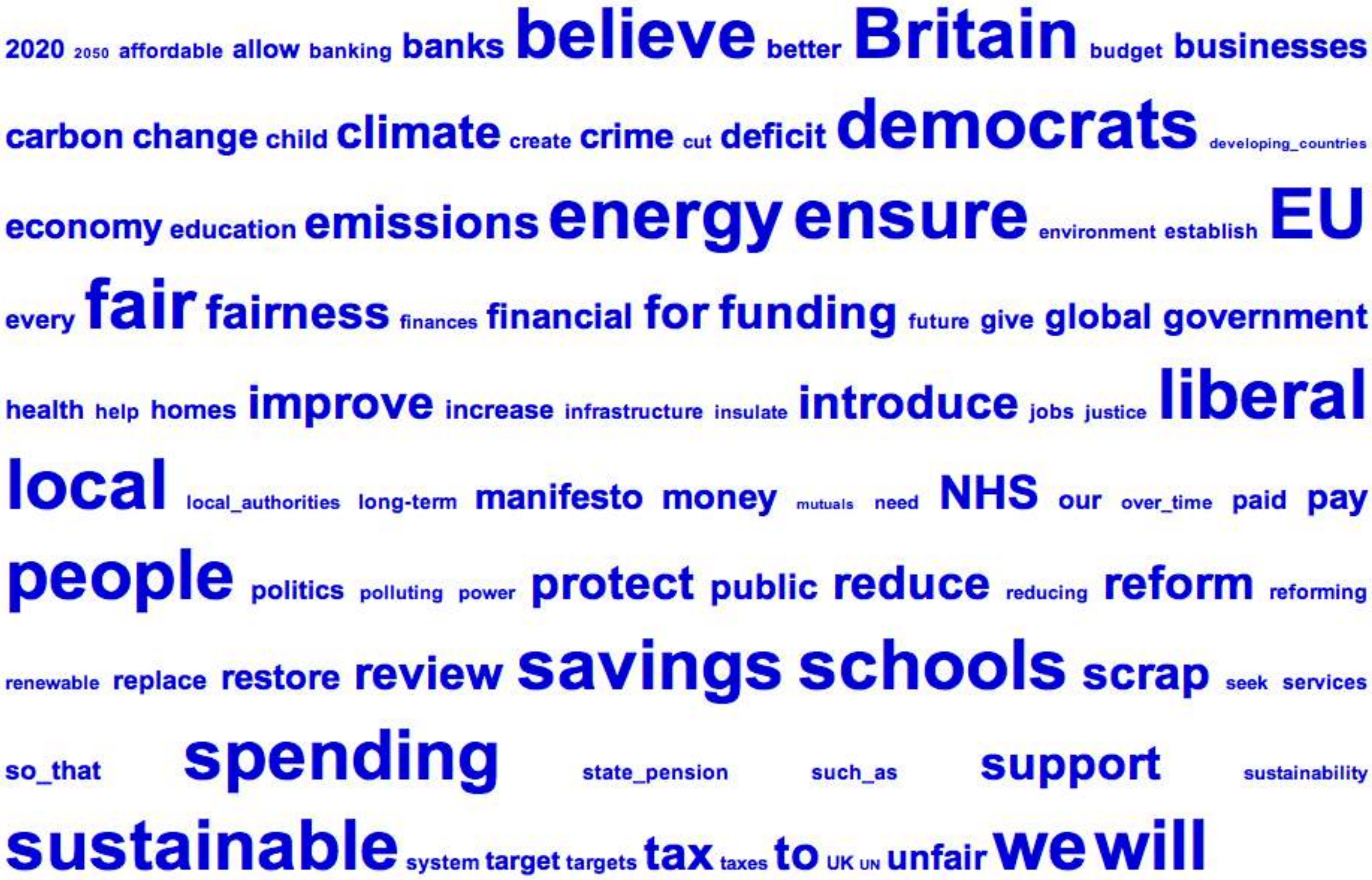
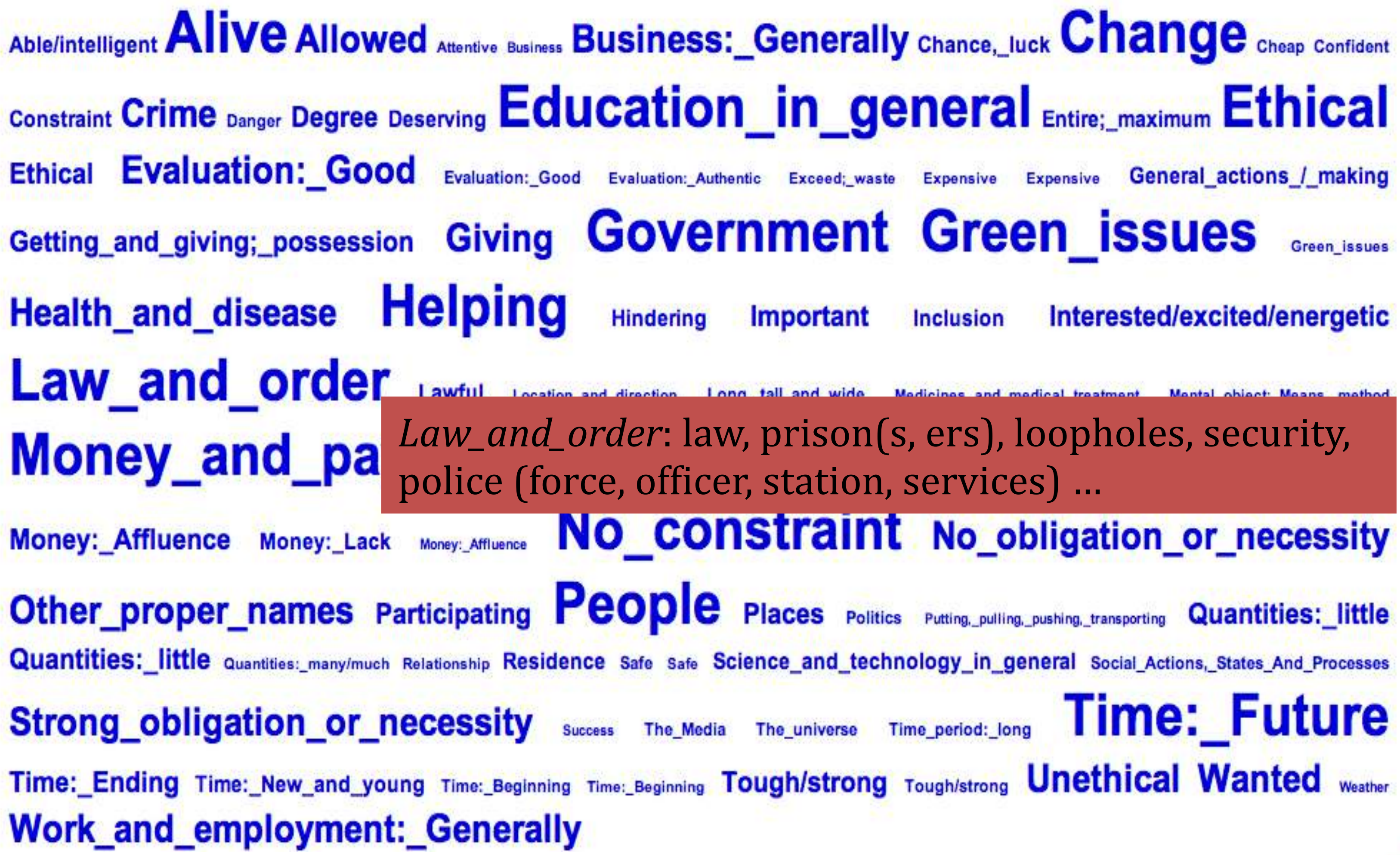


Figure 2: key domains (semantic fields) in LibDem 2010 manifesto



Law_and_order: law, prison(s, ers), loopholes, security, police (force, officer, station, services) …

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/

FORENSIC, LEGAL, POLICING APPLICATIONS

Example applications and studies

- Lord V, Davis B, Mason P. 2008. Stance shifting in language used by sex offenders. Psychology, Crime & Law 14, 357-379.
- Charitonidis C., Rashid A., Taylor P.J. (2017) Predicting Collective Action from Micro-Blog Data. In: Kawash J., Agarwal N., Özyer T. (eds) Prediction and Inference from Social Networks and Social Media. Lecture Notes in Social Networks.
- Markowitz DM, Hancock JT (2014) Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel. PLoS ONE 9(8): e105937. doi:10.1371/journal.pone.0105937
- Potts, A. and Kjær, A.L. (2015) Constructing Achievement in the International Criminal Tribunal for the Former Yugoslavia (ICTY): A Corpus-Based Critical Discourse Analysis. International Journal for the Semiotics of Law. doi: 10.1007/s11196-015-9440-y

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

- Jeffrey T. Hancock, Michael T. Woodworth and Stephen Porter (2013) Hungry like the wolf: A word-pattern analysis of the language of psychopaths. Legal and Criminological Psychology. Volume 18, Issue 1, pages 102-114. http://dx.doi.org/10.1111/j.2044-8333.2011.02025.x
- FBI Law Enforcement Bulletin (July 2012) The Language of Psychopaths: New Findings and Implications for Law Enforcement. By Michael Woodworth, Ph.D.; Jeffrey Hancock, Ph.D.; Stephen Porter, Ph.D.; Robert Hare, Ph.D.; Matt Logan, Ph.D.; Mary Ellen O'Toole, Ph.D.; and Sharon Smith, Ph.D. https://leb.fbi.gov/articles/featured-articles/the-language-of-psychopaths-new-findings-and-implications-for-law-enforcement
- Shapero, J. J. (2011). The Language of Suicide Notes. Unpublished Thesis. The University of Birmingham. http://etheses.bham.ac.uk/1525/
- Prentice, S, Rayson, P & Taylor, P 2012, 'The language of Islamic extremism: towards an automated identification of beliefs, motivations and justifications' International Journal of Corpus Linguistics, vol. 17, no. 2, pp. 259-286. DOI: 10.1075/ijcl.17.2.05pre
- Prentice, S, Taylor, P, Rayson, P & Giebels, E 2012, 'Differentiating act from ideology: evidence from messages for and against violent extremism' Negotiation and Conflict Management Research, vol. 5, no. 3, pp. 289-306. DOI: 10.1111/j.1750-4716.2012.00101.x

Psychopathy Checklist-Revised (PCL-R), Wmatrix & Dictionary of Affect and Language (DAL). Psychopaths vs control: key semantic tags.

Frequency, key words and key semantic tags.

Frequency, key words and key semantic tags, concordance analysis.

Counter-extremist vs extremist messages: Frequency, key words and key semantic tags, concordance analysis, collocation.

Online child protection

- Rashid, A, Baron, A, Rayson, P, May-Chahal, C, Greenwood, P & Walkerdine, J 2013, 'Who am I? Analysing Digital Personas in Cybercrime Investigations' Computer, vol. 46, no. 4, pp. 54-61. DOI: 10.1109/MC.2013.68
- May-Chahal, C, Mason, C, Rashid, A, Walkerdine, J, Rayson, P & Greenwood, P 2014, 'Safeguarding cyborg childhoods: incorporating the on/offline behaviour of children into everyday social work practices' British Journal of Social Work, vol. 44, no. 3, pp. 596-614. DOI: 10.1093/bjsw/bcs121

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

WMATRIX VERSION 4

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/

1. Wmatrix tutorial
   - http://ucrel.lancs.ac.uk/wmatrix/tutorial/
2. Wmatrix tool:
   - 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/
- 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):
  - Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12. http://www.lancaster.ac.uk/staff/rayson/publications/usas_lrec04ws.pdf
  - Rayson, P. (2008). From key words to key semantic domains. International Journal of Corpus Linguistics. 13:4, pp. 519-549.
  - Piao, S., Rayson, P., Archer, D., McEnery, T. (2005) Comparing and combining a semantic tagger and a statistical tool for MWE extraction. Computer Speech and Language, 19 (4), pp. 378 – 397 http://dx.doi.org/10.1016/j.csl.2004.11.002
  - Piao, S. (2002) Word alignment in English-Chinese parallel corpora. Literary and linguistic computing, 17 (2), 207-230. doi:10.1093/llc/17.2.207

Further reading …

- Baker, P. (2004) Querying keywords: questions of difference, frequency and sense in keywords analysis. Journal of English Linguistics. 32: 4, pp. 346-359. DOI: 10.1177/0075424204269894
- Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. Corpora 1(2), pp. 109-151. http://www.eupjournals.com/doi/abs/10.3366/cor.2006.1.2.109
- Gabrielatos, C. and Marchi, A. (2012) Keyness: Appropriate metrics and practical issues. CADS International Conference 2012. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?, 13-14 September, University of Bologna, Italy.
- Hardie, A. (2014) Log Ratio – an informal introduction. http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/
- Leech, G. and Fallon, R. (1992). Computer corpora - what do they tell us about culture? ICAME Journal, 16, pp. 29 - 50. http://icame.uib.no/archives/No_16_ICAME_Journal_index.pdf
- Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. Corpora 2 (1), pp. 1-31. http://www.eupjournals.com/doi/abs/10.3366/cor.2007.2.1.1
- Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. International Journal of Corpus Linguistics. 2 (1), pp 133 - 152. http://ucrel.lancs.ac.uk/papers/rlh97.html
- Scott, M. (1997). PC analysis of key words - and key key words. System 25 (2), pp. 233 - 245.
- Adam Kilgarriff (2005) Language is never ever ever random. Corpus Linguistics and Linguistic Theory 1 (2): 263-276. http://www.kilgarriff.co.uk/Publications/2005-K-lineer.pdf
- Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. Anglistik, 20(1), 41-67.
- Rayson, P. and Potts, A. (forthcoming) Analysing keyword lists. In Gries, S. Th. And Paquot, M. (eds.) A Practical Handbook of Corpus Linguistics. Springer.

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).