Wmatrix

Using Wmatrix: corpus analysis and comparison tool

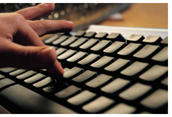
Paul Rayson
School of Computing and Communications
Lancaster University
p.rayson@lancaster.ac.uk
@perayson

MELC workshop
10th January 2014

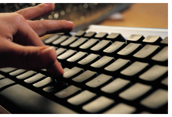
UCREL

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

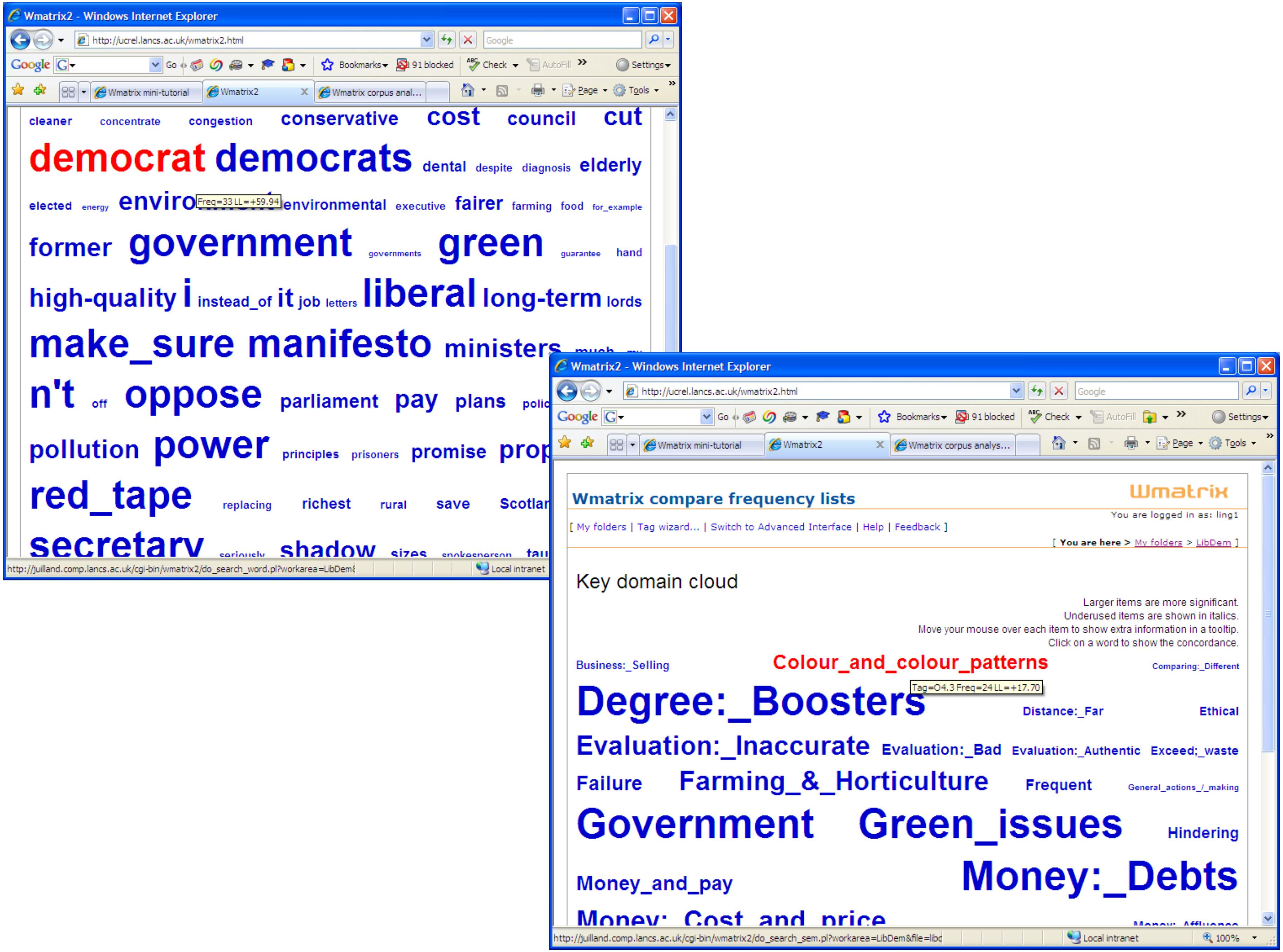
Wmatrix main points

- 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)

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			



cleaner concentrate congestion conservative cost council cut
democrat democrats dental despite diagnosis elderly
elected energy enviro Freq=33 LL=+59.94 environmental executive fairer farming food for_example
former government governments green guarantee hand
high-quality i instead_of it job letters liberal long-term lords
make_sure manifesto ministers much my
n't off oppose parliament pay plans polic
pollution power principles prisoners promise prop
red_tape replacing richest rural save Scotlan
secretary seriously shadow sizes spokesperson tau

http://juilland.comp.lancs.ac.uk/cgi-bin/wmatrix2/do_search_word.pl?workarea=LibDem&

Wmatrix compare frequency lists

Wmatrix

You are logged in as: ling1

[My folders | Tag wizard... | Switch to Advanced Interface | Help | Feedback]

[You are here > My folders > LibDem]

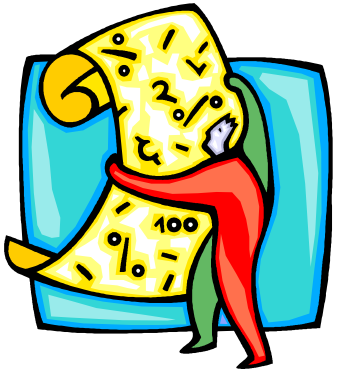
Key domain cloud

Larger items are more significant.
Underused items are shown in italics.
Move your mouse over each item to show extra information in a tooltip.
Click on a word to show the concordance.

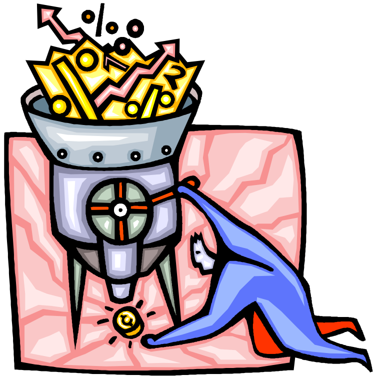
Business:_Selling Colour_and_colour_patterns Comparing:_Different
Degree:_Boosters Tag=O4.3 Freq=24 LL=+17.70 Distance:_Far Ethical
Evaluation:_Inaccurate Evaluation:_Bad Evaluation:_Authentic Exceed;_waste
Failure Farming_&_Horticulture Frequent General_actions_/_making
Government Green_issues Hindering
Money_and_pay Money:_Debts
Money:_Cost_and_price Money:_Affluence

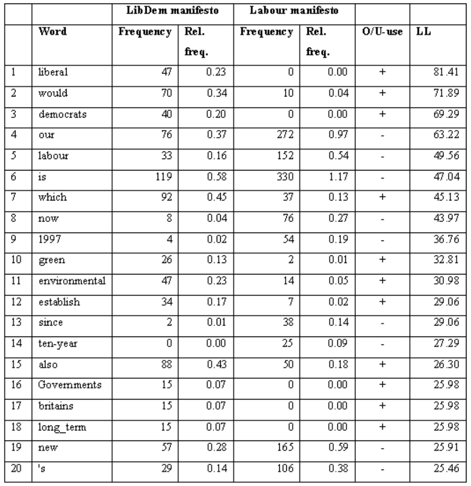
http://juilland.comp.lancs.ac.uk/cgi-bin/wmatrix2/do_search_sem.pl?workarea=LibDem&file=libd

Key words



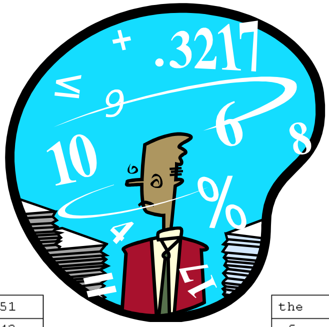
Text



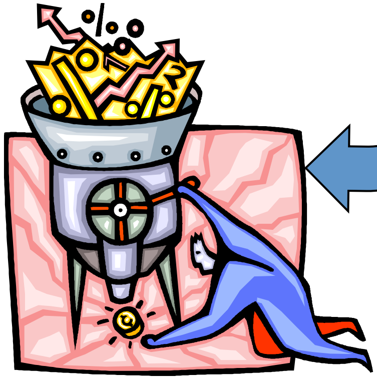


		LibDem manifesto		Labour manifesto			
	Word	Frequency	Rel. freq.	Frequency	Rel. freq.	O/U-use	LL
1	liberal	47	0.23	0	0.00	+	81.41
2	would	70	0.34	10	0.04	+	71.89
3	democrats	40	0.20	0	0.00	+	69.29
4	our	76	0.37	272	0.97	-	63.22
5	labour	33	0.16	152	0.54	-	49.56
6	is	119	0.58	330	1.17	-	47.04
7	which	92	0.45	37	0.13	+	45.13
8	now	8	0.04	76	0.27	-	43.97
9	1997	4	0.02	54	0.19	-	36.76
10	green	26	0.13	2	0.01	+	32.81
11	environmental	47	0.23	14	0.05	+	30.98
12	establish	34	0.17	7	0.02	+	29.06
13	since	2	0.01	38	0.14	-	29.06
14	ten-year	0	0.00	25	0.09	-	27.29
15	also	88	0.43	50	0.18	+	26.30
16	Governments	15	0.07	0	0.00	+	25.98
17	britains	15	0.07	0	0.00	+	25.98
18	long_term	15	0.07	0	0.00	+	25.98
19	new	57	0.28	165	0.59	-	25.91
20	's	29	0.14	106	0.38	-	25.46

Keywords



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

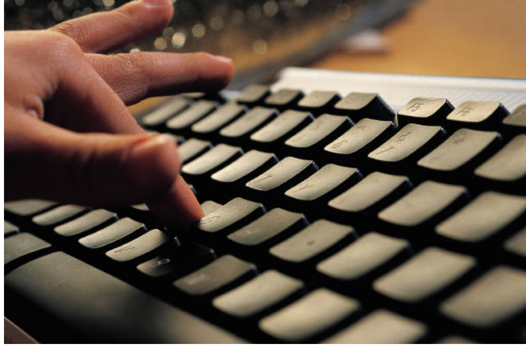
Log-likelihood (LL)

- 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/

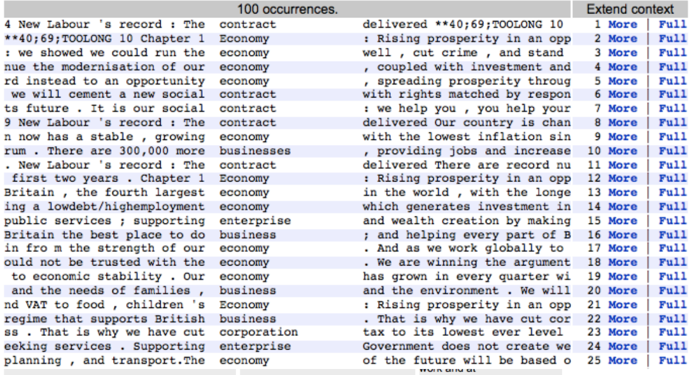
1. Wmatrix tutorial
   - http://ucrel.lancs.ac.uk/wmatrix/tutorial/

2. Wmatrix tool:
   - http://ucrel.lancs.ac.uk/wmatrix3.html
   - Login details:
     - Username:
     - Password:

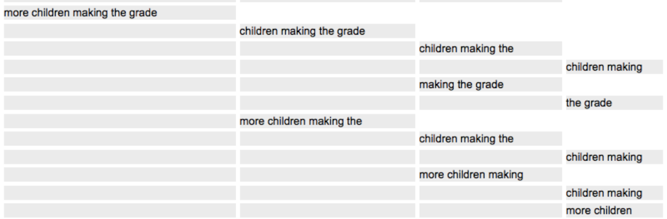
- http://ucrel.lancs.ac.uk/wmatrix/tutorial/
- On your own or in small groups:
  - **Read** tutorials A and B (the actions are already done)
  - **Do** tutorial C (key words, key domains and concordances)
- Advanced users:
  - Tutorial D (advanced data analysis) on your own or in small groups
  - Suggested 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

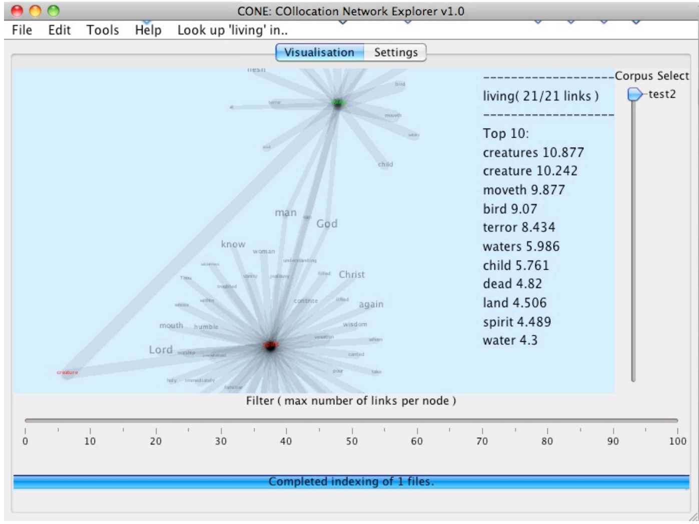
- CrossTabs
- Concordance
  - highlighting and filtering by context
  - concgrams-style
- Collocations and semantic collocations
- N-grams and C-grams
  - Aka clusters, lexical bundles
  - Faster implementation (L-gram)
  - http://code.google.com/p/lgram/
- Visualisations
  - Collocation Network Explorer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



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



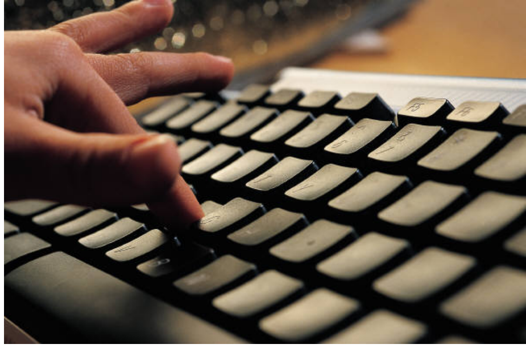
```
                                                        work and at
more children making the grade
                               children making the grade
                                                        children making the
                                                                              children making
                                                        making the grade
                                                                              the grade
                               more children making the
                                                        children making the
                                                                              children making
                                                        more children making
                                                                              children making
                                                                              more children
```



```
-------------------
living( 21/21 links )
-------------------
Top 10:
creatures 10.877
creature 10.242
moveth 9.877
bird 9.07
terror 8.434
waters 5.986
child 5.761
dead 4.82
land 4.506
spirit 4.489
water 4.3
```

Completed indexing of 1 files.

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):
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13:4, pp. 519-549.
- Wmatrix, CLAWS and USAS websites:
  - http://ucrel.lancs.ac.uk/wmatrix/
  - http://ucrel.lancs.ac.uk/claws/
  - http://ucrel.lancs.ac.uk/usas/
- 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

- **Further reading (mostly key words related).**
- 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
- 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 [Beware 20Mb download]
- 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