



# Keywords are not enough

Paul Rayson


UCREL, Computing Department,  
Lancaster University, UK.



# Outline

- New kind of method and tool (Matrix) for the statistical analysis of corpora
- Standard corpus linguistic research process model identifies the research question (and the linguistic features) early in the study.
- Recent advances in annotation and size
- Matrix is a tool which assists in finding candidate research questions
- Matrix integrates part-of-speech tagging and semantic field tagging in a profiling tool
- Extends the keywords procedure to identify key grammatical categories and key concepts
- Comparison of UK 2001 general election manifestos of the Labour and Liberal Democratic parties

# Corpus Linguistic Research Process Model

- 
1. *Question*: A research question or model is devised
  2. *Build*: Corpus design and compilation
  3. *Annotate*: Computational analysis of the corpus
  4. *Retrieve*: Quantitative and qualitative analyses of the corpus
  5. *Interpret*: Manual interpretation of the results or confirmation of the accuracy of the model

# Data-driven versus corpus-driven

- ☞ Recent advances
  - Larger corpora
  - Linguistic annotation at multiple levels
- ☞ A tool which assists in finding candidate research questions
- ☞ Allows macroscopic analysis to inform microscopic analysis



# Matrix method (1)

	Corpus one	Corpus two	Total
Frequency of word	a	b	a+b
Frequency of word not occurring	c-a	d-b	c+d-a-b
<b>TOTAL</b>	c	d	c+d

$$E_i = \frac{N_i \sum O_i}{\sum_i N_i}$$

$$-2 \ln \lambda = 2 \sum_i O_i \ln \left( \frac{O_i}{E_i} \right)$$

$$E1 = c \times (a+b) / (c+d)$$

$$E2 = d \times (a+b) / (c+d)$$

$$LL = 2 \times ((a \times \ln (a/E1)) + (b \times \ln (b/E2)))$$

## Matrix method (2)

- Integrates POS tagging and semantic field annotation into a profiling tool
- Extends keywords procedure to identify key grammatical categories and key concepts
- Choice of log-likelihood statistic over chi-squared statistic

# Case studies

- Social differentiation in the use of English vocabulary (Rayson, Leech & Hodges, 1997)
- Profiling of learner English (Granger & Rayson, 1998)
- Semantic analysis of technical documents from the software engineering domain (Sawyer, Rayson & Garside, 2002)

# Worked example

- Comparison of UK 2001 General Election manifestos of the Labour and Liberal Democratic parties.



# Comparison at word level

LibDem manifesto		Labour manifesto	
Word	Frequency	Word	Frequency
the	1174	the	1482
and	794	to	1112
to	736	and	1100
of	632	of	719
will	461	we	669
we	428	in	546
a	345	will	515
in	320	a	506
for	308	for	491
by	196	is	330
on	166	our	272
are	128	with	242
that	123	are	226
is	119	have	209
be	109	by	194
more	107	on	185
with	107	be	173
have	97	new	165
this	94	more	162
their	93	people	160

Top 20 most significant differences at word level between Labour and LibDem manifestos

	Word	LibDem manifesto		Labour manifesto		O/U-use	LL
		Frequency	Rel. freq.	Frequency	Rel. freq.		
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

## Concordance of key word *would* from LibDem manifesto

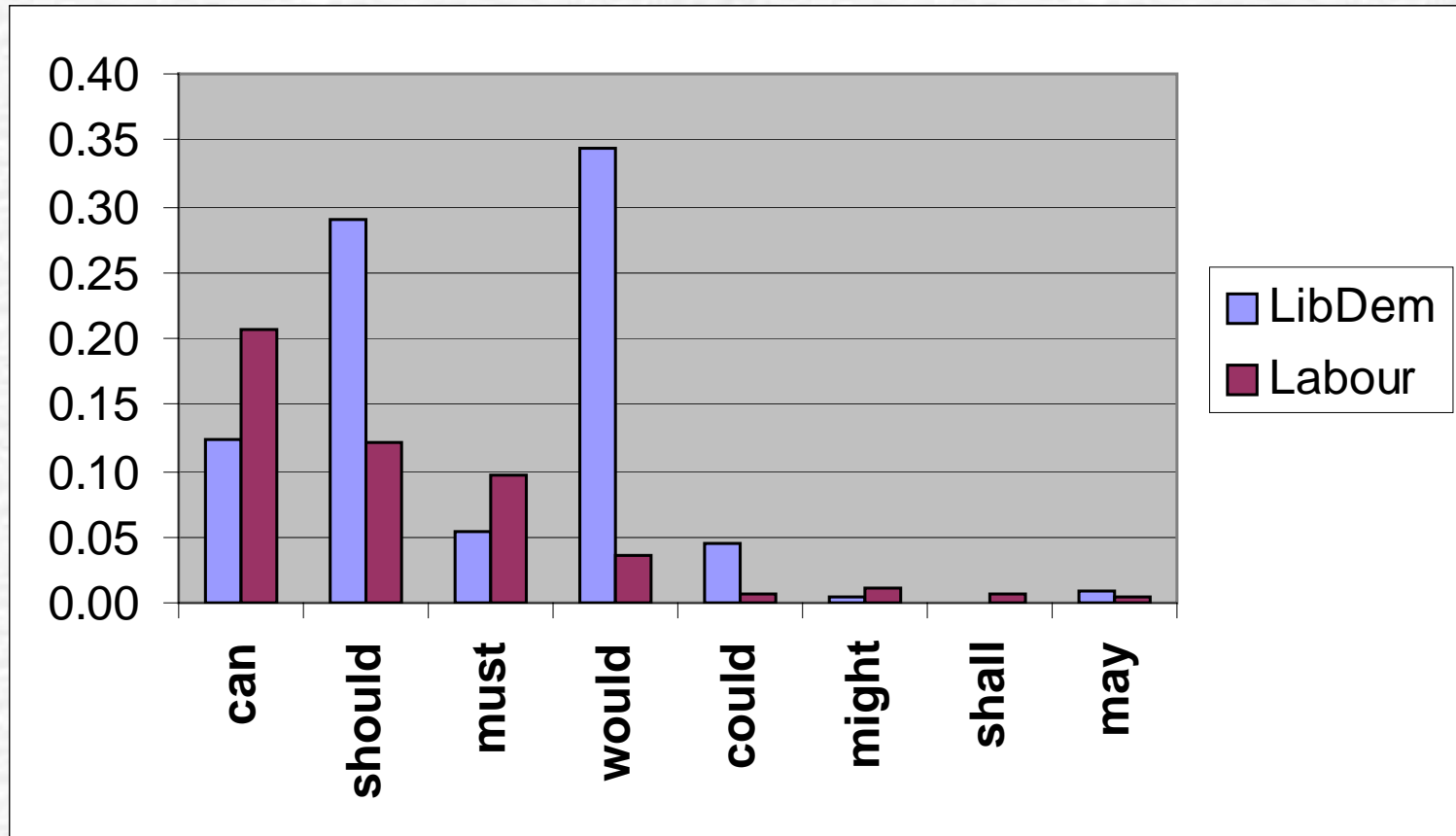
mitted to training programmes which would bring enormous benefits to the economic recovery. Companies eligible to claim R&D tax credits would include those working with Investors in Business. Companies eligible for the new Job Creation Tax Credit would like work. Companies eligible for the new Job Creation Tax Credit would include environmental assessment of buildings and promote the use of better insulation. Companies eligible for the new Job Creation Tax Credit would give the police more flexibility and retained police officers. This would only be released following an assessment of violent offenders, so that they respond not only to the problems caused by and violent offenders, so that they ensure that those young people who are not in trouble with the law are not limit the availability of places in Young Offender Institutions and lack; layout-grid-mode:line'> which would extend across the UK the scheme currently being brought in. This would be measured by a Quality of Life Index. This would include a statement of the standards we would ensure that a growing proportion of the population would no longer have to show a history of conviction and would eventually help around 3.4 million people on low income pay less tax even allowing for our 1.4 million people on low income who would give grants to support libraries, and ensure that these standards are maintained. We would then guarantee them the right to be automatically lapse. This would be set at 1500 and apply to all small businesses. This tax-free allowance

## Comparison at the POS level

	POS	LibDem manifesto		Labour manifesto		O/U- use	LL
		Frequency	Rel. freq.	Frequency	Rel. freq.		
1	MC	124	0.61	587	2.09	-	197.39
2	RT	13	0.06	105	0.37	-	55.26
3	VBZ	119	0.58	334	1.19	-	48.96
4	MD	22	0.11	122	0.43	-	48.15
5	NN2	1999	9.80	2271	8.08	+	39.30
6	DDQ	98	0.48	47	0.17	+	38.37
7	APPGE	199	0.98	438	1.56	-	31.61
8	VM	637	3.12	650	2.31	+	28.85
9	VV0	646	3.17	662	2.36	+	28.49
10	RR	379	1.86	368	1.31	+	22.77
11	GE	39	0.19	119	0.42	-	20.85
12	VH0	73	0.36	184	0.65	-	20.56
13	NNO	0	0.00	17	0.06	-	18.55
14	II21	68	0.33	41	0.15	+	18.19
15	IW	119	0.58	258	0.92	-	17.58
16	VVN	346	1.70	624	2.22	-	16.52
17	CSW	0	0.00	15	0.05	-	16.37
18	IO	633	3.10	718	2.55	+	12.64
19	NPM1	0	0.00	11	0.04	-	12.01
20	VVG	433	2.12	476	1.69	+	11.49



## Relative use of modal verbs in LibDem and Labour manifestos



# Comparison at the semantic tag level

	Semantic tag	LibDem manifesto		Labour manifesto		O/U-use	LL	Semantic category
		Freq.	Rel. freq.	Freq.	Rel. freq.			
1	N1	142	0.70	547	1.95	-	141.97	Numbers
2	S7.4+	131	0.64	47	0.17	+	72.72	Permission
3	T3-	139	0.68	375	1.33	-	50.05	Time: new and young
4	G1.1	362	1.77	293	1.04	+	46.13	Government etc.
5	I3.1	170	0.83	413	1.47	-	41.49	Work and employment
6	A1.7-	77	0.38	33	0.12	+	35.01	Constraint
7	M3	141	0.69	92	0.33	+	32.03	Vehicles and transport on land
8	A3+	236	1.16	490	1.74	-	27.95	Being
9	O4.3	30	0.15	6	0.02	+	26.08	Colour and colour patterns
10	N5	76	0.37	198	0.70	-	24.19	Quantities
11	A6.1-	99	0.49	63	0.22	+	23.74	Comparing: different
12	X2.4	93	0.46	59	0.21	+	22.45	Investigate, examine, test, search
13	W5	27	0.13	7	0.02	+	19.84	Green issues
14	T2++	38	0.19	114	0.41	-	19.30	Time: Continuing
15	T2-	58	0.28	32	0.11	+	18.25	Time: Stopping
16	A2.1+	156	0.76	321	1.14	-	17.60	Affect: Modify, change
17	N4	43	0.21	119	0.42	-	16.88	Linear order
18	O1	30	0.15	11	0.04	+	16.29	Substances and materials
19	N5-	110	0.54	88	0.31	+	14.56	Quantities
20	S4	40	0.20	108	0.38	-	14.44	Kin

## Concordance of key concept *permission* from LibDem manifesto

<p>n: yes"&gt; &lt;/span&gt; We will also &gt; &lt;/span&gt; We will extend the wers of Select Committees and s more say over the budget by te from the Finance Bill , to acerun: yes"&gt; &lt;/span&gt; We will allow the Welsh Assembly the cerun: yes"&gt; &lt;/span&gt; We would span&gt; They are essential to a black;layout-grid-mode:line' &gt; trong framework of individual by European law , so that the d personal relationship legal span style='color:black'&gt; The k'&gt; The Right to Know and the e individuals should have the eplace the system of warrants by Ministers with a system of r:black;font-style:normal'&gt; A :normal'&gt; We will protect the e that farm animals should be</p>	<p>allow right allow allowing allow allow right allow liberal Liberal rights rights rights Right Right right approved approval Right right entitled</p>	<p>people to stand for elected of to vote by post and investigat more pre-legislative scrutiny them to propose spending amend for greater consultation on ta the Welsh Assembly the right t to pass primary legislation an further devolution of powers a society in which people are en Democrats will : &lt;o:p&gt; &lt;/o:p&gt; , extending the protection alr of the individual outweigh the , such as next-of-kin arrangem to Know and the Right to Priva to Privacy &lt;o:p&gt; &lt;/o:p&gt; &lt;/span&gt; to know as much as possible ab by Ministers with a system of by judges to remove any confli to Environmental Information , to legal and peaceful protest to high welfare standards . &lt;s</p>
---	---	---

# Conclusion to the worked example

1. An investigation of the inclusive language of Labour, indicated by their manifesto having greater use of the word *our*
2. An investigation into the differing use of modal verbs between the LibDem and Labour manifestos, signposted by the overuse of *would* in the LibDem manifesto
3. An investigation into the relative use of *permission* and *freedom* concepts, highlighted by significantly greater use of these concepts in the LibDem manifesto
4. An investigation into the political renewal senses conveyed by overuse of terms such as *new*, *modern*, *reform*, and *change* in the Labour manifesto
5. An investigation into party policy differences between LibDem and Labour indicated by significant differences in the relative use of concepts related to environmental issues, family issues, work and employment, and transport



# Conclusion (1)

- Described the Matrix method and tool
- Frequency profiling of corpora, and comparison of those profiles across corpora.
- In order to suggest possible research questions for further investigation, the Matrix method uses the log-likelihood ratio statistic to compare frequencies and then rank them in terms of significant difference.

## Conclusion (2)

- ✔ Worked example of the method
- ✔ UK 2001 General Election manifestos
- ✔ Extends keywords approach to key grammatical classes and key concepts
- ✔ Key grammatical categories and semantic classes are used to group together lower frequency words and those words which would, by themselves, not be identified as key, and would otherwise be overlooked
- ✔ Comparison at the POS and semantic levels reduces the number of key categories that the researcher should examine

# Questions?

- ☛ **Rayson, P.** (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. *Ph.D. thesis*, Lancaster University.
- ☛ paul@comp.lancs.ac.uk
- ☛ [www.comp.lancs.ac.uk/computing/users/paul/](http://www.comp.lancs.ac.uk/computing/users/paul/)