





# Outline

- **Wmatrix:** *A new method and tool for the statistical analysis of corpora*
- **Introducing the system:** *Available tools in Wmatrix*
- **Hands-on demonstration:** *Newspaper genre analysis – broadsheets and tabloids*

# What it's all about

- Web-based tool
- You can load your own data
- Incorporates main tools in corpus linguistics:  
frequency lists, concordances, key words,  
collocations
- Adds two NLP tools:  
POS tagging and semantic field tagging

# How is it innovative?

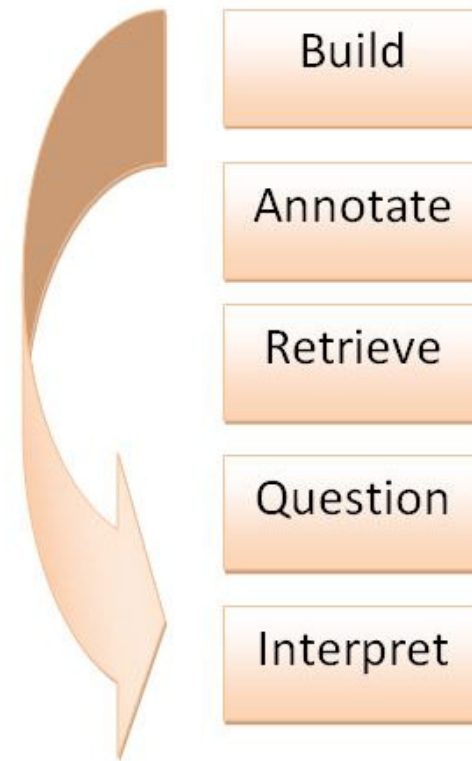
- Integrates POS tagging and semantic field tagging
- Extends keywords procedure to identify key grammatical categories and key semantic domains
- Allows macroscopic analysis to inform microscopic analysis

# How is it innovative?

Hypothesis-driven research model



Data-driven research model



Based on Rayson (2008)

# Tools: frequency lists

Word	Frequency	Relative Frequency	
the	28003	6.27	<a href="#">Concordance</a>
of	12754	2.86	<a href="#">Concordance</a>
to	12282	2.75	<a href="#">Concordance</a>
and	12104	2.71	<a href="#">Concordance</a>
a	10027	2.24	<a href="#">Concordance</a>
in	7715	1.73	<a href="#">Concordance</a>
that	5252	1.18	<a href="#">Concordance</a>
is	4879	1.09	<a href="#">Concordance</a>
it	4732	1.06	<a href="#">Concordance</a>
for	4071	0.91	<a href="#">Concordance</a>
on	4063	0.91	<a href="#">Concordance</a>
was	3288	0.74	<a href="#">Concordance</a>
's	3156	0.71	<a href="#">Concordance</a>
we	3060	0.69	<a href="#">Concordance</a>
with	2896	0.65	<a href="#">Concordance</a>
i	2890	0.65	<a href="#">Concordance</a>
be	2771	0.62	<a href="#">Concordance</a>
are	2689	0.60	<a href="#">Concordance</a>
as	2465	0.55	<a href="#">Concordance</a>
but	2366	0.53	<a href="#">Concordance</a>
from	2274	0.51	<a href="#">Concordance</a>
you	2207	0.49	<a href="#">Concordance</a>
have	2186	0.49	<a href="#">Concordance</a>
this	2121	0.47	<a href="#">Concordance</a>
media	2048	0.46	<a href="#">Concordance</a>
by	2010	0.45	<a href="#">Concordance</a>

(All examples taken from BBC blog corpus)



# Tools: concordances

blishers - traditionally mainstream	media	organisations - are well aware of th	376	<a href="#">More</a>		<a href="#">Full</a>
for the material of others . Social	media	was not part of the landscape in 200	377	<a href="#">More</a>		<a href="#">Full</a>
sion on day two of #bbcsms : Social	media	is increasingly driving traffic to n	378	<a href="#">More</a>		<a href="#">Full</a>
, said Nic Newman , the former BBC	media	strategist . Research shows that 7%	379	<a href="#">More</a>		<a href="#">Full</a>
research shows that 7% of mainstream	media	traffic is now coming from social me	380	<a href="#">More</a>		<a href="#">Full</a>
a traffic is now coming from social	media	. At the same time , the number of n	381	<a href="#">More</a>		<a href="#">Full</a>
etween big organisations and social	media	content is changing , but mainstream	382	<a href="#">More</a>		<a href="#">Full</a>
content is changing , but mainstream	media	is still setting the agenda on socia	383	<a href="#">More</a>		<a href="#">Full</a>
still setting the agenda on social	media	rather than the other way round . Ma	384	<a href="#">More</a>		<a href="#">Full</a>
the most successful users of social	media	are the stars of mainstream media :	385	<a href="#">More</a>		<a href="#">Full</a>
l media are the stars of mainstream	media	: so how do individuals become so-ca	386	<a href="#">More</a>		<a href="#">Full</a>
d what value does that add to their	media	organisation ? A research project on	387	<a href="#">More</a>		<a href="#">Full</a>
Newman was chairing the BBC Social	Media	Summit 's session on technology and	388	<a href="#">More</a>		<a href="#">Full</a>
and innovation . Liz Heron , Social	Media	Editor of the New York Times , descr	389	<a href="#">More</a>		<a href="#">Full</a>
cribed her role in a team of social	media	editors at the Times working closely	390	<a href="#">More</a>		<a href="#">Full</a>
ng closely with developers . Social	media	was originally seen as a way to reac	391	<a href="#">More</a>		<a href="#">Full</a>
e being told not to think of social	media	only in terms of alternative forms o	392	<a href="#">More</a>		<a href="#">Full</a>
But the paper does n't have social	media	guidelines : " We just tell people t	393	<a href="#">More</a>		<a href="#">Full</a>
stupid . " Journalists using social	media	are reminded they are still 'the Tim	394	<a href="#">More</a>		<a href="#">Full</a>
the first news agency of the social	media	age " , wanted to change attitudes t	395	<a href="#">More</a>		<a href="#">Full</a>
anted to change attitudes to social	media	. Its interactions have much in comm	396	<a href="#">More</a>		<a href="#">Full</a>

# Tools: collocations

Save this table as a [tab-delimited file](#)



Change statistic:

Log-Likelihood

MI  
MI2  
MI3  
Phi-Square  
**Log-Likelihood**  
Ochiai (OCH)  
McConnoughy Coefficient (MCC)  
Yule Coefficient (YUL)  
Fager and McGowan Coefficient (FAG)  
Kulczinsky Coefficient (KUC)  
Simple Matching Coefficient (SMC)

	Log-Likelihood	T-score	Collocation
1	6376.62	26.67	social media
2	2167.45	17.25	BBC News
3	1915.30	11.48	New York
4	1539.67	10.89	College Journalism
5	1149.14	9.63	World Service
6	1041.47	7.87	Middle East
7	979.80	8.54	Of course
8	925.99	9.35	years ago
9	846.83	10.42	news organisations
10	796.03	9.02	public interest
11	777.79	6.40	Human Rights
12	750.48	9.37	last year
13	739.49	9.37	around world
14	700.27	10.67	I think
15	665.94	10.00	BBC World
16	645.89	6.70	Social Summit



# Tools: key words

- Built in reference corpora (BNC samplers) or other uploaded files

	Item	O1	%1	O2	%2	LL
1	Concordance media	2048	0.46	111	0.01 +	3931.93
2	Concordance BBC	1886	0.42	100	0.01 +	3632.44
3	Concordance news	1900	0.43	217	0.02 +	3146.57
4	Concordance journalists	1426	0.32	27	0.00 +	3040.19
5	Concordance journalism	962	0.22	38	0.00 +	1924.17
6	Concordance US	797	0.18	0	0.00 +	1837.92
7	Concordance twitter	778	0.17	0	0.00 +	1794.10
8	Concordance social	1116	0.25	184	0.02 +	1653.00
9	Concordance facebook	527	0.12	0	0.00 +	1215.28
10	Concordance website	356	0.08	0	0.00 +	820.95
11	Concordance audience	524	0.12	73	0.01 +	820.25
12	Concordance online	509	0.11	65	0.01 +	817.57

# Tools

The **\_AT** lovers **\_NN2** , **\_**, whose **\_DDQGE** chief **\_JJ** scene **\_NN1** was **\_VBDZ** cut **\_VVN** at **\_II** the **\_AT** last **\_MD** moment **\_NN1** , **\_**, had **\_VHD** comparatively **\_RR** little **\_DA1** to **\_TO** sing **\_VVI** . **\_**.

CLAWS tagger

It **\_Z8** is **\_Z5** anticipated **\_X2.6+** that **\_Z5** the **\_Z5** system **\_X4.2** will **\_T1.1.3** be **\_Z5** administered **\_A9-** by **\_Z5** the **\_Z5** Library **\_Q4.1/H1** , **\_PUNC** but **\_Z5** this **\_Z8** will **\_T1.1.3** not **\_Z6** always **\_N6+++** be **\_the\_case** **\_A5.2+[i9.3** . **\_PUNC**

USAS tagger

# Tools: USAS tagger

<b>A</b> general and abstract terms	<b>B</b> the body and the individual	<b>C</b> arts and crafts	<b>E</b> emotion
<b>F</b> food and farming	<b>G</b> government and public	<b>H</b> architecture, housing and the home	<b>I</b> money and commerce in industry
<b>K</b> entertainment, sports and games	<b>L</b> life and living things	<b>M</b> movement, location, travel and transport	<b>N</b> numbers and measurement
<b>O</b> substances, materials, objects and equipment	<b>P</b> education	<b>Q</b> language and communication	<b>S</b> social actions, states and processes
<b>T</b> Time	<b>W</b> world and environment	<b>X</b> psychological actions, states and processes	<b>Y</b> science and technology
<b>Z</b> names and grammar			

# Tools: key parts of speech

	Item	O1	%1	O2	%2	LL
1	List1   Concordance NN	3995	0.89	3059	0.32 +	1879.04
2	List1   Concordance PPIS2	3060	0.69	3307	0.34 +	748.44
3	List1   Concordance VVG	8674	1.94	13564	1.40 +	548.84
4	List1   Concordance VV0	7208	1.61	11012	1.14 +	518.16
5	List1   Concordance RRQ	1761	0.39	2048	0.21 +	355.94
6	List1   Concordance NP2	436	0.10	231	0.02 +	320.07
7	List1   Concordance TO	8268	1.85	14041	1.45 +	302.96
8	List1   Concordance CST	3862	0.86	6409	0.66 +	167.83
9	List1   Concordance DAR	767	0.17	875	0.09 +	163.38
10	List1   Concordance PPH1	4729	1.06	8211	0.85 +	144.63
11	List1   Concordance VBZ	6112	1.37	11171	1.15 +	113.00
12	List1   Concordance VH0	1865	0.42	2971	0.31 +	105.84
13	List1   Concordance VVZ	4260	0.95	7602	0.79 +	101.33
14	List1   Concordance NNO2	183	0.04	139	0.01 +	87.10
15	List1   Concordance VHZ	1760	0.39	2901	0.30 +	80.17
16	List1   Concordance DA2	988	0.22	1477	0.15 +	79.35
17	List1   Concordance CSN	706	0.16	981	0.10 +	78.67
18	List1   Concordance RRR	754	0.17	1066	0.11 +	78.19
19	List1   Concordance VBR	3027	0.68	5435	0.56 +	67.70
20	List1   Concordance PN1	1349	0.30	2193	0.23 +	67.41

AT1 CCB<sub>cs</sub> CSA **CSN** **CST**<sub>csw</sub> **DA2** **DAR**<sub>DD DD1 DDQ EX I121 I133 JJT JK</sub> **NN**<sub>NN2 NNO</sub> **NNO2** **NNT2** **NP2**<sub>NPM1</sub> **PN1**  
 PNQS **PPH1** **PPI02** **PPIS2**<sub>PPX2 RA REX21</sub> **RGQ** **RGR** **RGT** **RR** **RRQ** **RRR** **RRT** **TO**<sub>VBG vbn</sub> **VBR** **VBZ** **VD0** **VDG** **VDI** **vdn** **VDZ**  
**VH0** **VHZ** **VV0** **VVG**<sub>VGK VVI</sub> **VVZ**<sub>XX</sub>

# Tools: key semantic domains

	Item	O1	%1	O2	%2	LL
1	List1   Concordance Q4.2	5176	1.16	828	0.09 +	7747.27 The Media: Newspapers etc.
2	List1   Concordance Q4	3321	0.74	740	0.08 +	4363.90 The Media
3	List1   Concordance Q4.3	3082	0.69	690	0.07 +	4041.25 The Media: TV, Radio and Cinema
4	List1   Concordance X2.2+	3936	0.88	2302	0.24 +	2608.38 Knowledgeable
5	List1   Concordance Z3	5699	1.28	4809	0.50 +	2298.84 Other proper names
6	List1   Concordance Y2	2598	0.58	1126	0.12 +	2280.83 Information technology and computing
7	List1   Concordance Q2.1	6160	1.38	7024	0.73 +	1313.84 Speech: Communicative
8	List1   Concordance L1+	507	0.11	93	0.01 +	722.18 Alive
9	List1   Concordance Q2.2	6689	1.50	9724	1.00 +	613.57 Speech acts
10	List1   Concordance A13	242	0.05	0	0.00 +	558.06 Degree
11	List1   Concordance Q1.2	3006	0.67	3691	0.38 +	518.37 Paper documents and writing
12	List1   Concordance S1.1.1	1969	0.44	2089	0.22 +	503.41 Social Actions, States And Processes
13	List1   Concordance I3.2	266	0.06	48	0.00 +	381.26 Work and employment: Professionalism
14	List1   Concordance A5.2+	870	0.19	779	0.08 +	316.28 Evaluation: True
15	List1   Concordance Q1.3	586	0.13	409	0.04 +	313.93 Telecommunications
16	List1   Concordance X2.2	257	0.06	87	0.01 +	269.59 Knowledge
17	List1   Concordance G2.1	1790	0.40	2418	0.25 +	222.81 Law and order
18	List1   Concordance W1	861	0.19	912	0.09 +	220.97 The universe
19	List1   Concordance W2	95	0.02	0	0.00 +	219.07 Light
20	List1   Concordance Y1	751	0.17	778	0.08 +	202.91 Science and technology in general

Attentive Belonging\_to\_a\_group Cause&Effect/Connection Change Crime Danger Degree: Non-specific Degree: Boosters Degree: Maximizers Easy Entertainment generally Ethical  
**Evaluation: True** Evaluation: Inaccurate Evaluation: Accurate Evaluation: Good Existing Expected Failure  
 General appearance and physical properties Government Helping Important  
**Information technology and computing** Interested/excited/energetic  
 Investigate, examine, test, search Knowledgeable Language, speech and grammar  
 Law and order Learning Location and direction Mental object: Conceptual object  
 No obligation or necessity No constraint Not part of a group **Other proper names**  
**Paper documents and writing** People Polite Pronouns Quantities: many/much Safe  
 Science and technology in general Sensory: Sound Sensory: Sight Size: Big

# Availability

- For a free one-month trial contact [p.rayson@lancaster.ac.uk](mailto:p.rayson@lancaster.ac.uk)
- Academic one-year license – for details see <http://ucrel.lancs.ac.uk/wmatrix/>





- **Twitter**

@The\_Moki

@usageguides

- **Blog**

[bridgingtheunbridgeable.com](http://bridgingtheunbridgeable.com)