UNIVERSITY OF
BIRMINGHAM

The need for Corpus Statistics: Corpus analysis and the identification of linguistically relevant patterns

Launching the Corpus Statistics Group
11th Feb. 2016

University of Birmingham

The Corpus Statistics group

□ Core members (not just speakers today)

□ Results and work-in-progress reports from projects (internally and externally funded)

□ Need for a group? Problems are often interpreted from different disciplinary perspectives. Aim to work collaboratively!

□ Impact and challenges of availability of resources and data, infrastructure

Aims for today:

- □ (Corpus) linguistically relevant patterns – what do we want to find?
- □ How do linguistic patterns relate to statistical problems?
- □ Finding a way of communication across disciplines

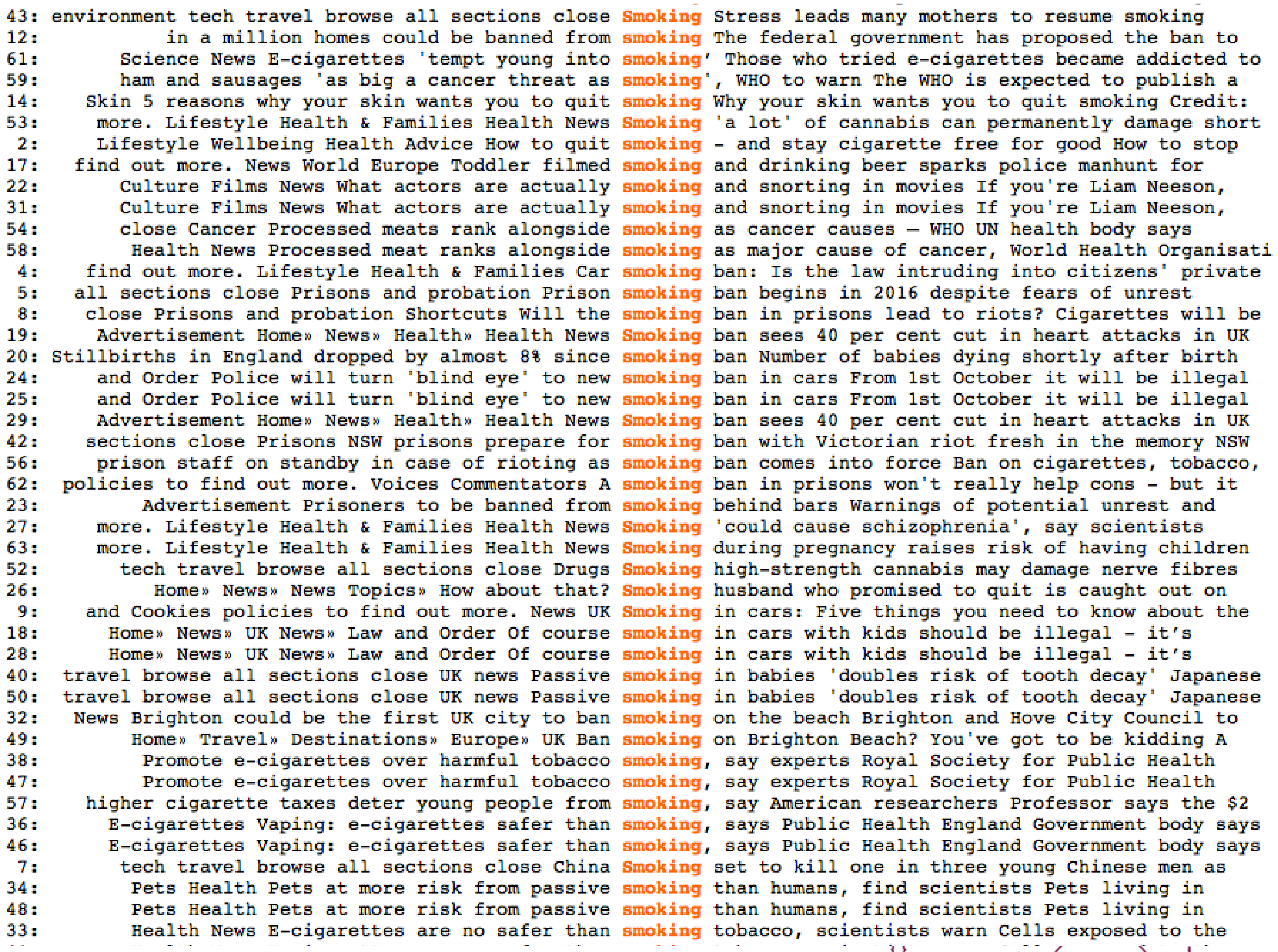
Patterns of language: 3 tenets of corpus linguistics

1) Language is a social phenomenon

2) Meaning and form are associated

3) Corpus linguistics prioritises lexis

1. Language is a social phenomenon



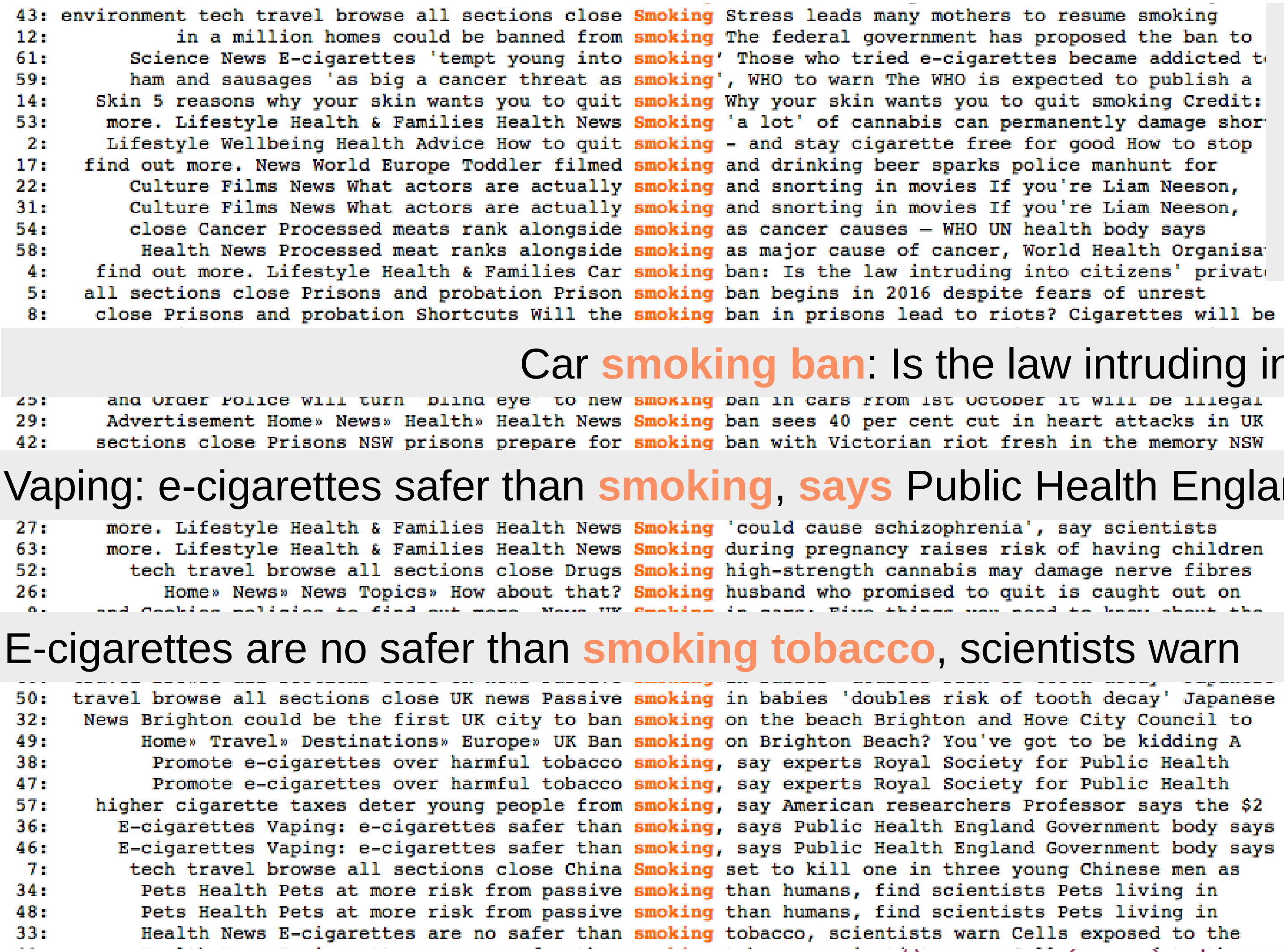
```
43: environment tech travel browse all sections close Smoking Stress leads many mothers to resume smoking
12:           in a million homes could be banned from smoking The federal government has proposed the ban to
61:       Science News E-cigarettes 'tempt young into smoking' Those who tried e-cigarettes became addicted to
59:       ham and sausages 'as big a cancer threat as smoking', WHO to warn The WHO is expected to publish a
14:    Skin 5 reasons why your skin wants you to quit smoking Why your skin wants you to quit smoking Credit:
53:     more. Lifestyle Health & Families Health News Smoking 'a lot' of cannabis can permanently damage short
 2:     Lifestyle Wellbeing Health Advice How to quit smoking - and stay cigarette free for good How to stop
17:   find out more. News World Europe Toddler filmed smoking and drinking beer sparks police manhunt for
22:       Culture Films News What actors are actually smoking and snorting in movies If you're Liam Neeson,
31:       Culture Films News What actors are actually smoking and snorting in movies If you're Liam Neeson,
54:       close Cancer Processed meats rank alongside smoking as cancer causes – WHO UN health body says
58:        Health News Processed meat ranks alongside smoking as major cause of cancer, World Health Organisati
 4:    find out more. Lifestyle Health & Families Car smoking ban: Is the law intruding into citizens' private
 5:   all sections close Prisons and probation Prison smoking ban begins in 2016 despite fears of unrest
 8:    close Prisons and probation Shortcuts Will the smoking ban in prisons lead to riots? Cigarettes will be
19:     Advertisement Home» News» Health» Health News Smoking ban sees 40 per cent cut in heart attacks in UK
20: Stillbirths in England dropped by almost 8% since smoking ban Number of babies dying shortly after birth
24:     and Order Police will turn 'blind eye' to new smoking ban in cars From 1st October it will be illegal
25:     and Order Police will turn 'blind eye' to new smoking ban in cars From 1st October it will be illegal
29:     Advertisement Home» News» Health» Health News Smoking ban sees 40 per cent cut in heart attacks in UK
42:    sections close Prisons NSW prisons prepare for smoking ban with Victorian riot fresh in the memory NSW
56:     prison staff on standby in case of rioting as smoking ban comes into force Ban on cigarettes, tobacco,
62:  policies to find out more. Voices Commentators A smoking ban in prisons won't really help cons - but it
23:         Advertisement Prisoners to be banned from smoking behind bars Warnings of potential unrest and
27:     more. Lifestyle Health & Families Health News Smoking 'could cause schizophrenia', say scientists
63:     more. Lifestyle Health & Families Health News Smoking during pregnancy raises risk of having children
52:       tech travel browse all sections close Drugs Smoking high-strength cannabis may damage nerve fibres
26:          Home» News» News Topics» How about that? Smoking husband who promised to quit is caught out on
 9:    and Cookies policies to find out more. News UK Smoking in cars: Five things you need to know about the
18:      Home» News» UK News» Law and Order Of course smoking in cars with kids should be illegal - it's
28:      Home» News» UK News» Law and Order Of course smoking in cars with kids should be illegal - it's
40:  travel browse all sections close UK news Passive smoking in babies 'doubles risk of tooth decay' Japanese
50:  travel browse all sections close UK news Passive smoking in babies 'doubles risk of tooth decay' Japanese
32:   News Brighton could be the first UK city to ban smoking on the beach Brighton and Hove City Council to
49:        Home» Travel» Destinations» Europe» UK Ban smoking on Brighton Beach? You've got to be kidding A
38:         Promote e-cigarettes over harmful tobacco smoking, say experts Royal Society for Public Health
47:         Promote e-cigarettes over harmful tobacco smoking, say experts Royal Society for Public Health
57:    higher cigarette taxes deter young people from smoking, say American researchers Professor says the $2
36:      E-cigarettes Vaping: e-cigarettes safer than smoking, says Public Health England Government body says
46:      E-cigarettes Vaping: e-cigarettes safer than smoking, says Public Health England Government body says
 7:       tech travel browse all sections close China Smoking set to kill one in three young Chinese men as
34:        Pets Health Pets at more risk from passive smoking than humans, find scientists Pets living in
48:        Pets Health Pets at more risk from passive smoking than humans, find scientists Pets living in
33:        Health News E-cigarettes are no safer than smoking tobacco, scientists warn Cells exposed to the
```

Retrieved with WebCorp – UK broadsheets

1. Language is a social phenomenon

Linguistic evidence of social interaction

Language is used to do things.



```
43: environment tech travel browse all sections close Smoking Stress leads many mothers to resume smoking
12:         in a million homes could be banned from smoking The federal government has proposed the ban to
61:      Science News E-cigarettes 'tempt young into smoking' Those who tried e-cigarettes became addicted t
59:      ham and sausages 'as big a cancer threat as smoking', WHO to warn The WHO is expected to publish a
14:   Skin 5 reasons why your skin wants you to quit smoking Why your skin wants you to quit smoking Credit:
53:     more. Lifestyle Health & Families Health News Smoking 'a lot' of cannabis can permanently damage shor
 2:     Lifestyle Wellbeing Health Advice How to quit smoking - and stay cigarette free for good How to stop
17:   find out more. News World Europe Toddler filmed smoking and drinking beer sparks police manhunt for
22:      Culture Films News What actors are actually smoking and snorting in movies If you're Liam Neeson,
31:      Culture Films News What actors are actually smoking and snorting in movies If you're Liam Neeson,
54:      close Cancer Processed meats rank alongside smoking as cancer causes – WHO UN health body says
58:       Health News Processed meat ranks alongside smoking as major cause of cancer, World Health Organisa
 4:    find out more. Lifestyle Health & Families Car smoking ban: Is the law intruding into citizens' privat
 5:   all sections close Prisons and probation Prison smoking ban begins in 2016 despite fears of unrest
 8:    close Prisons and probation Shortcuts Will the smoking ban in prisons lead to riots? Cigarettes will be
25:     and Order Police will turn 'blind eye' to new smoking ban in cars from 1st October it will be illegal
29:     Advertisement Home» News» Health» Health News Smoking ban sees 40 per cent cut in heart attacks in UK
42:    sections close Prisons NSW prisons prepare for smoking ban with Victorian riot fresh in the memory NSW
27:     more. Lifestyle Health & Families Health News Smoking 'could cause schizophrenia', say scientists
63:     more. Lifestyle Health & Families Health News Smoking during pregnancy raises risk of having children
52:      tech travel browse all sections close Drugs Smoking high-strength cannabis may damage nerve fibres
26:         Home» News» News Topics» How about that? Smoking husband who promised to quit is caught out on
 9:    and Cookies policies to find out more. News UK Smoking in cars: Five things you need to know about the
50:  travel browse all sections close UK news Passive smoking in babies 'doubles risk of tooth decay' Japanese
32:   News Brighton could be the first UK city to ban smoking on the beach Brighton and Hove City Council to
49:       Home» Travel» Destinations» Europe» UK Ban smoking on Brighton Beach? You've got to be kidding A
38:        Promote e-cigarettes over harmful tobacco smoking, say experts Royal Society for Public Health
47:        Promote e-cigarettes over harmful tobacco smoking, say experts Royal Society for Public Health
57:   higher cigarette taxes deter young people from smoking, say American researchers Professor says the $2
36:      E-cigarettes Vaping: e-cigarettes safer than smoking, says Public Health England Government body says
46:      E-cigarettes Vaping: e-cigarettes safer than smoking, says Public Health England Government body says
 7:      tech travel browse all sections close China Smoking set to kill one in three young Chinese men as
34:       Pets Health Pets at more risk from passive smoking than humans, find scientists Pets living in
48:       Pets Health Pets at more risk from passive smoking than humans, find scientists Pets living in
33:       Health News E-cigarettes are no safer than smoking tobacco, scientists warn Cells exposed to the
```

Car **smoking ban**: Is the law intruding into citizens' private

Vaping: e-cigarettes safer than **smoking**, **says** Public Health England

E-cigarettes are no safer than **smoking tobacco**, scientists warn

Retrieved with WebCorp – UK broadsheets

2. Meaning and form are associated

□ Lexico-grammatical: *smoking ban, quitting smoking, tobacco smoking, passive smoking*

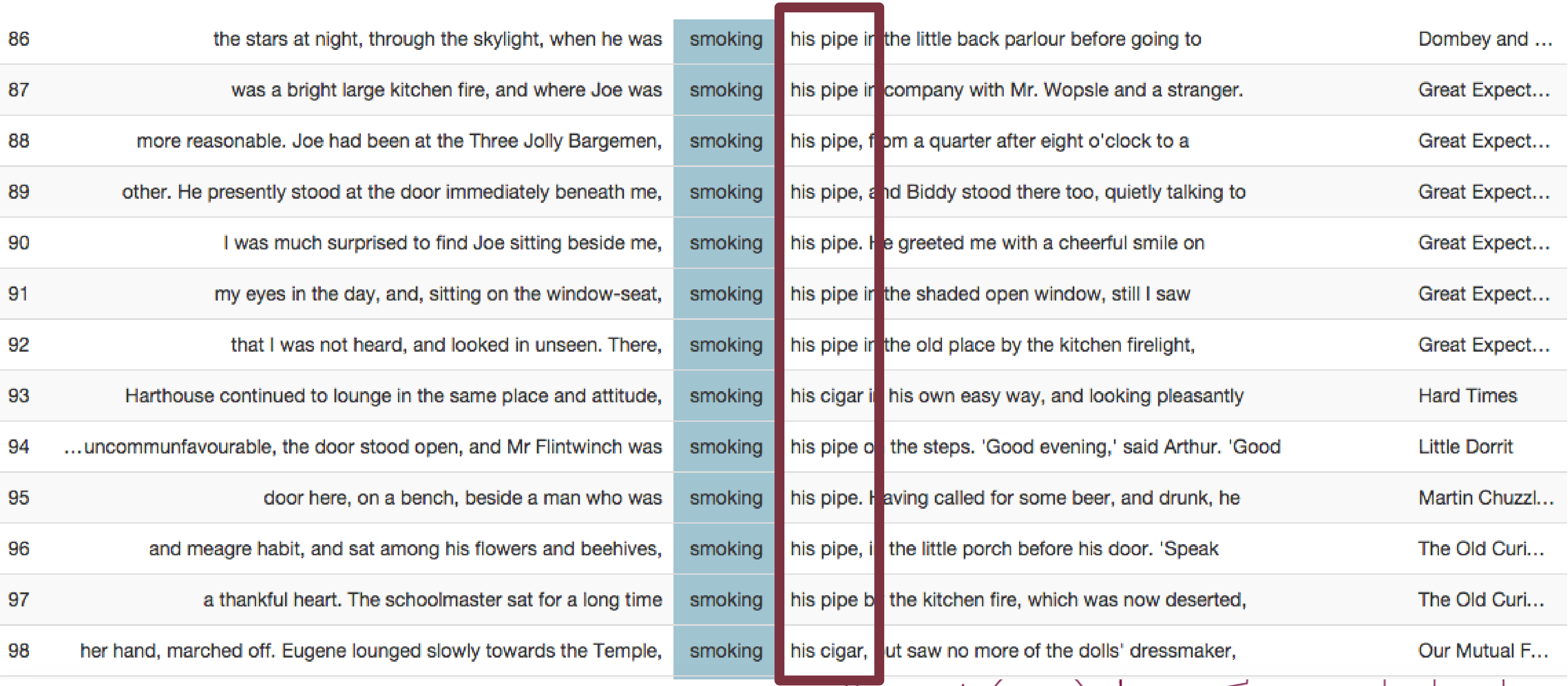
□ Text sections:

Vaping: e-cigarettes safer than smoking, says Public Health England

□ Types of texts:

2. Meaning and form are associated

- Types of texts: *smoke* as a verb



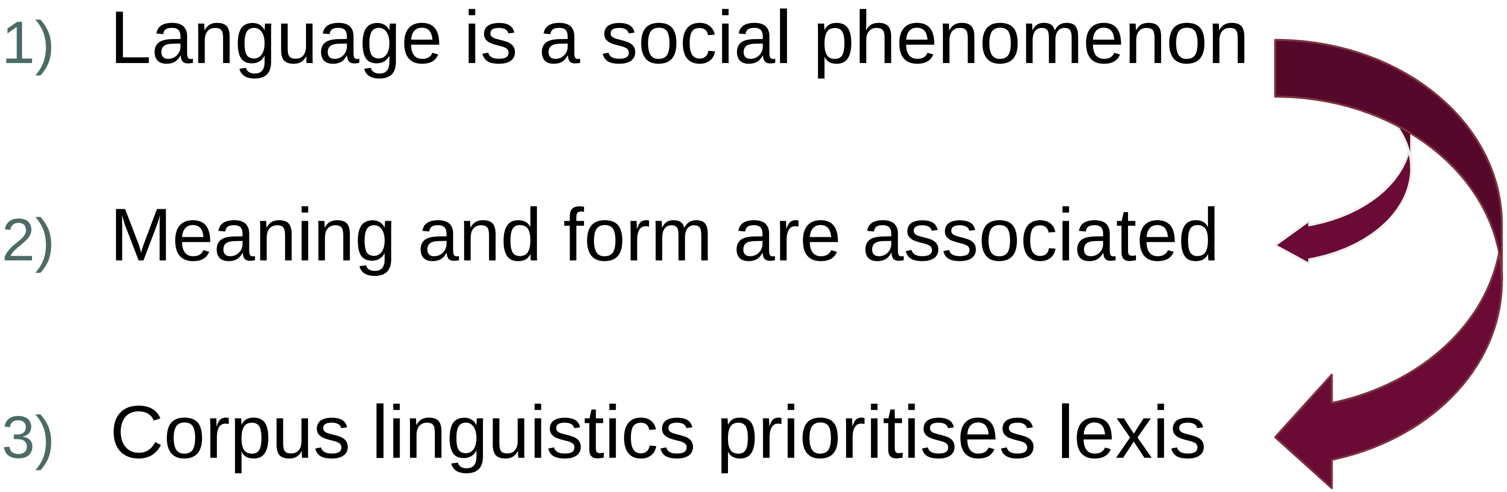
86	the stars at night, through the skylight, when he was	smoking	his pipe in the little back parlour before going to	Dombey and …
87	was a bright large kitchen fire, and where Joe was	smoking	his pipe in company with Mr. Wopsle and a stranger.	Great Expect…
88	more reasonable. Joe had been at the Three Jolly Bargemen,	smoking	his pipe, from a quarter after eight o'clock to a	Great Expect…
89	other. He presently stood at the door immediately beneath me,	smoking	his pipe, and Biddy stood there too, quietly talking to	Great Expect…
90	I was much surprised to find Joe sitting beside me,	smoking	his pipe. He greeted me with a cheerful smile on	Great Expect…
91	my eyes in the day, and, sitting on the window-seat,	smoking	his pipe in the shaded open window, still I saw	Great Expect…
92	that I was not heard, and looked in unseen. There,	smoking	his pipe in the old place by the kitchen firelight,	Great Expect…
93	Harthouse continued to lounge in the same place and attitude,	smoking	his cigar in his own easy way, and looking pleasantly	Hard Times
94	…uncommunfavourable, the door stood open, and Mr Flintwinch was	smoking	his pipe on the steps. 'Good evening,' said Arthur. 'Good	Little Dorrit
95	door here, on a bench, beside a man who was	smoking	his pipe. Having called for some beer, and drunk, he	Martin Chuzzl…
96	and meagre habit, and sat among his flowers and beehives,	smoking	his pipe, in the little porch before his door. 'Speak	The Old Curi…
97	a thankful heart. The schoolmaster sat for a long time	smoking	his pipe by the kitchen fire, which was now deserted,	The Old Curi…
98	her hand, marched off. Eugene lounged slowly towards the Temple,	smoking	his cigar, but saw no more of the dolls' dressmaker,	Our Mutual F…

Retrieved with CLiC – Dickens's novels

3. Corpus linguistics priorities lexis

□ Starting from the word to identify patterns and meanings: concordances, collocations, co-occurrence patterns, …

3 tenets of corpus linguistics (Mahlberg 2005)



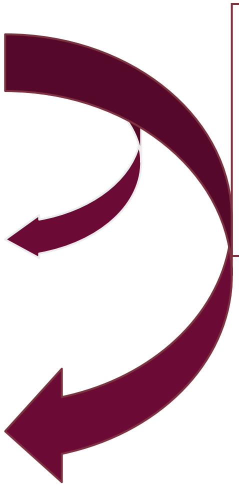
1) Language is a social phenomenon
2) Meaning and form are associated
3) Corpus linguistics prioritises lexis

3 tenets of corpus linguistics (Mahlberg 2005)

1) Language is a social phenomenon

2) Meaning and form are associated

3) Corpus linguistics prioritises lexis



Availability of data and methods

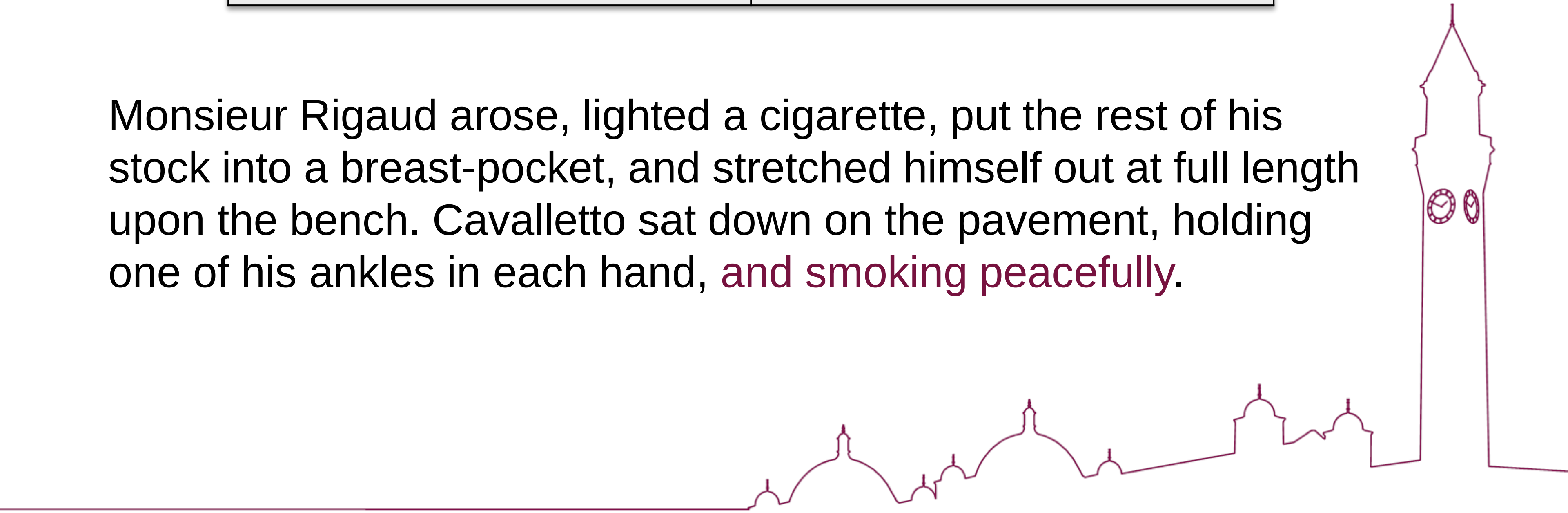
in texts and relationships between texts

Meaning based on evidence of interaction

- Is best studied in corpora with plenty of options for comparisons and the identification of textual relationships

***smoking* in Dickens**	
in quotes	in non-quotes
11 pmw	54 pmw

Monsieur Rigaud arose, lighted a cigarette, put the rest of his stock into a breast-pocket, and stretched himself out at full length upon the bench. Cavalletto sat down on the pavement, holding one of his ankles in each hand, and smoking peacefully.



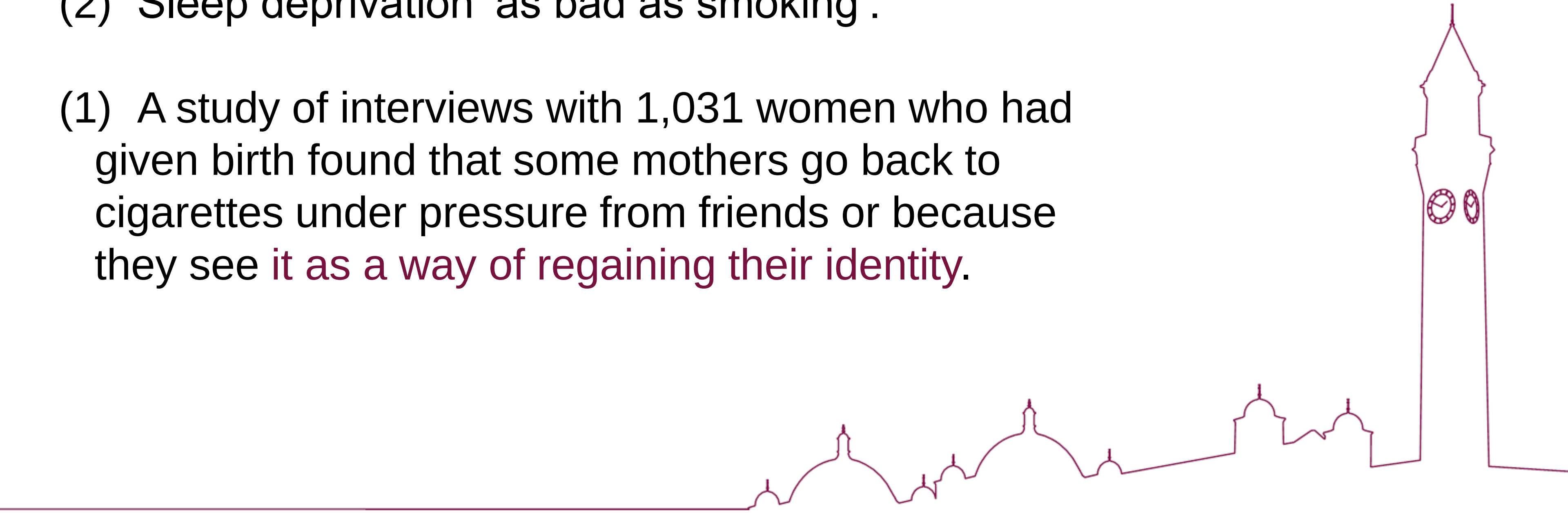
Meaning based on evidence of interaction

□ Is flexible and negotiated by the language users, it has a historical dimension (cf. e.g. Teubert 2015)

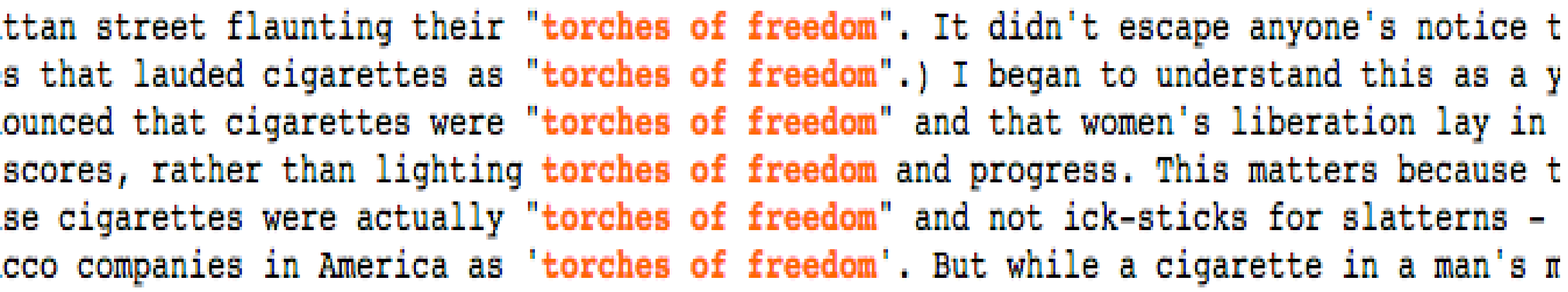
(1) The World Health Organisation is expected to issue new guidelines warning that processed meat products such as bacon and sausages are a cancer risk on the scale of smoking and asbestos.

(2) Sleep deprivation ‘as bad as smoking’.

(1) A study of interviews with 1,031 women who had given birth found that some mothers go back to cigarettes under pressure from friends or because they see it as a way of regaining their identity.



(4) Smoking and feminism: fallen women and prostitutes, from social taboo to *Torches of Freedom*



```
ttan street flaunting their "torches of freedom". It didn't escape anyone's notice t
s that lauded cigarettes as "torches of freedom".) I began to understand this as a y
ounced that cigarettes were "torches of freedom" and that women's liberation lay in
scores, rather than lighting torches of freedom and progress. This matters because t
se cigarettes were actually "torches of freedom" and not ick-sticks for slatterns -
cco companies in America as 'torches of freedom'. But while a cigarette in a man's m
```

WebCorp – Feb 2016 – 5 of the 6 references to historical events

An Ancient Prejudice Has Been Removed
"TOASTING DID IT"–
LUCKY STRIKE
CIGARETTES
"It's toasted"
No Throat Irritation–No Cough.

Meaning based on evidence of interaction

- Is multimodal

© AP/Warner Bros Pictures

Key semantic domain in Bond: Smoking and non-medical drugs

cigarette, smoked, cigarettes, tobacco, cigar, smokes, dope, smoking, cigarette-case, Marihuana

Meaning based on evidence of interaction

- Highlights that the description of meaning is not just a linguistic matter:
  - Medical research questions: smoking and cancer
  - “Scholars don't pay enough attention to what non-scholars think about the world” (Proctor 2012: 89)
  - Health issues in literature: e.g. Pickwickian syndrome

… mere boy of nineteen or twenty, who, though it was yet barely ten o’clock, was drinking gin and water, and smoking a cigar, amusements to which, judging from his inflamed countenance, he had devoted himself pretty constantly for the last year or two of his life. (PP)

Effects of alcohol, fetal alcohol syndrome, *gin – mother’s ruin*

Betsy Martin, widow, one child, and one eye. Goes out charing and washing, by the day; never had more than one eye, but knows her mother drank bottled stout, and shouldn't wonder if that caused it (immense cheering). Thinks it not impossible that if she had always abstained from spirits she might have had two eyes by this time (tremendous applause).

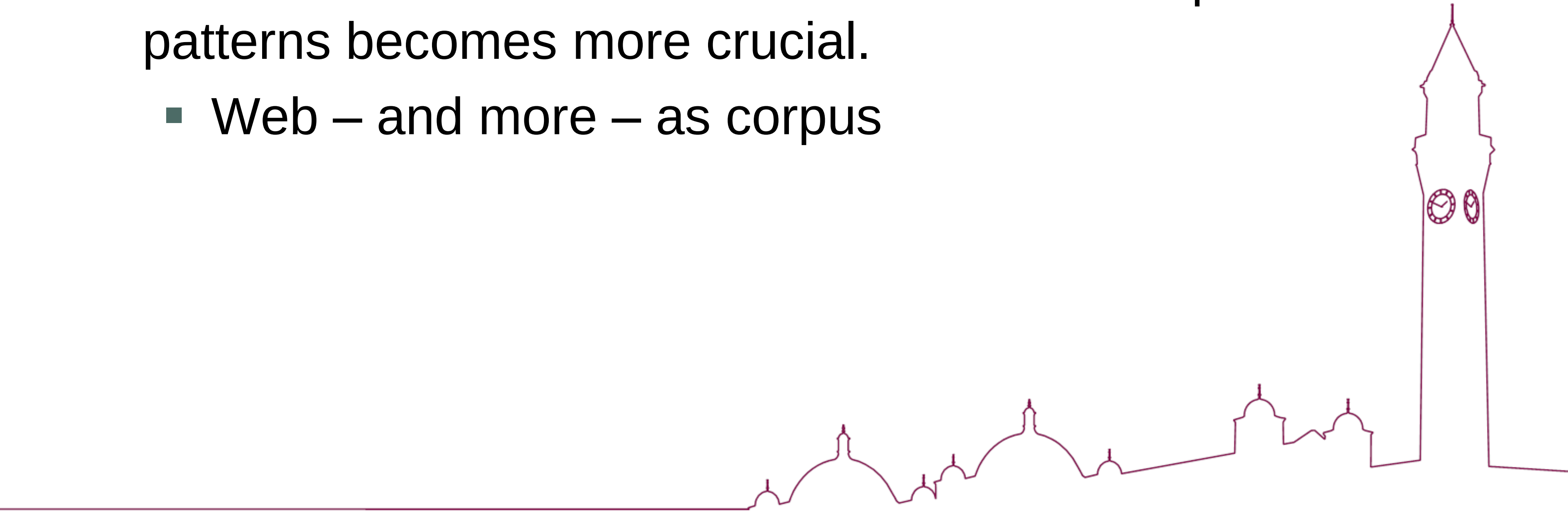
(*Pickwick Papers*)

Meaning based on evidence of interaction

- □ Calls for less ‘artificial / tidy / linguistic’ corpora
  - ▪ Not just a question of full texts vs text extracts.

    New sources of data through digitisation and data born digital.

- □ The selection of ‘candidates’ for detailed interpretation of patterns becomes more crucial.
  - ▪ Web – and more – as corpus



Meaning based on evidence of interaction

- Linguistically relevant patterns:
  - Collocations, co-occurrences, key words, topic modelling, network graphs
- Less ‘artificial / tidy / linguistic’ corpora:
  - Dickens and novels, TDA, journals
  - Multimodal (pictures in Times, films – with Andrew Salway)
- Not just linguistic or statistical:
  - work with Kate Fleming, Marnie Brennan
- RQs guide the search for candidates
- Ideally studied across disciplines, combining methods, data sets, tools and RQs:

  The Corpus Statistics Group