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

Launching the Corpus Statistics Group
11\textsuperscript{th} 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

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

Linguistic evidence of social interaction
Language is used to do things.

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

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

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 in texts and relationships between texts

Availability of data and methods
Meaning based on evidence of interaction

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

<table>
<thead>
<tr>
<th>smoking in Dickens</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>in quotes</td>
<td>in non-quotes</td>
</tr>
<tr>
<td>11 pmw</td>
<td>54 pmw</td>
</tr>
</tbody>
</table>

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*

"...flaunting their "torches of freedom". It didn't escape anyone's notice that lauded cigarettes as "torches of freedom".) I began to understand this as a young woman that cigarettes were "torches of freedom" and that women's liberation lay in the "torches of freedom". This matters because these cigarettes were actually "torches of freedom" and not ick-sticks for slatterns—coco companies in America as 'torches of freedom'. But while a cigarette in a man's."
Meaning based on evidence of interaction

- Is multimodal

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