Association measures

Mutual Information and Collocations

Simon Šuster, LCT
April 2011, Seminar in Statistics and Methodology @ RUG

Association measures

- Formulae based on expected and observed frequency counts for word co-occurence
- True statistical association or just chance co-occurrence?
- Different questions = different measures (~50) = different scores
- 3 criteria important
  - Frequency
    - Plain freq. yields function words, punctuation...
  - Significance
    - Observed freq. > expected freq.
  - Effect size
    - Ratio between O and E
- Word-word, word-construction strength

Collocations

- Any habitual co-occurrence between words, i.e. between node and the collocate
- “dangerous + and”
  - Very frequent, highly significant, but O not much higher than E
- “dangerous + driving/substances/situations...”
  - Also high effect size

(Pointwise) Mutual Information

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

- MI measures information shared by x and y
  - how much knowing one var. reduces uncertainty about the other.
- If x and y are independent, MI is 0.
- Pointwise = particular co-occurrence event
- Church & Hanks 1990:

$$PMI = \log \frac{p(x,y)}{p(x)\,p(y)} \qquad PMI = \log \frac{observed}{expected}$$

PMI

- Effect size (how much more often than by chance)
- Values theoretically between -inf and +inf, but in practice determined by N
- Known to attribute high scores to low freq. words, technical terms
  - Need for a frequency threshold
  - Heuristic versions of MI boost O:

$$PMI2 = \log \frac{p(x,y)^2}{p(x)\,p(y)} \qquad PMI3 = \log \frac{p(x,y)^3}{p(x)\,p(y)}$$

Calculation

$$PMI = \log \frac{p(x,y)}{p(x)\,p(y)} \quad \rightarrow \quad PMI(w_1, w_2) = \log \frac{\frac{f(w_1, w_2)}{N}}{\frac{f(w_1)}{N}\,\frac{f(w_2)}{N}}$$

$w_1 = \text{dangerous}$

$w_2 = \text{substance}$

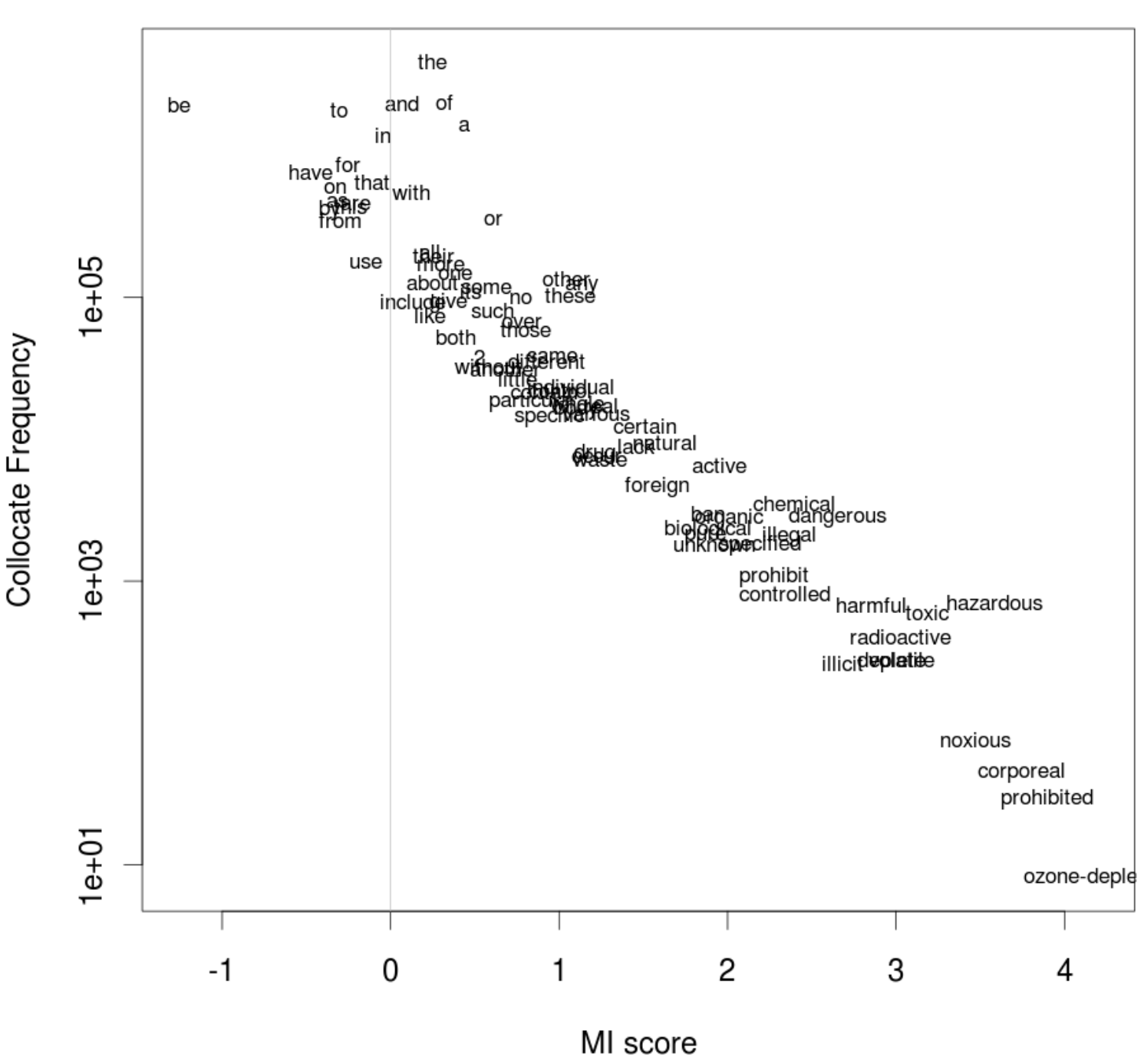
$N = 67063111$

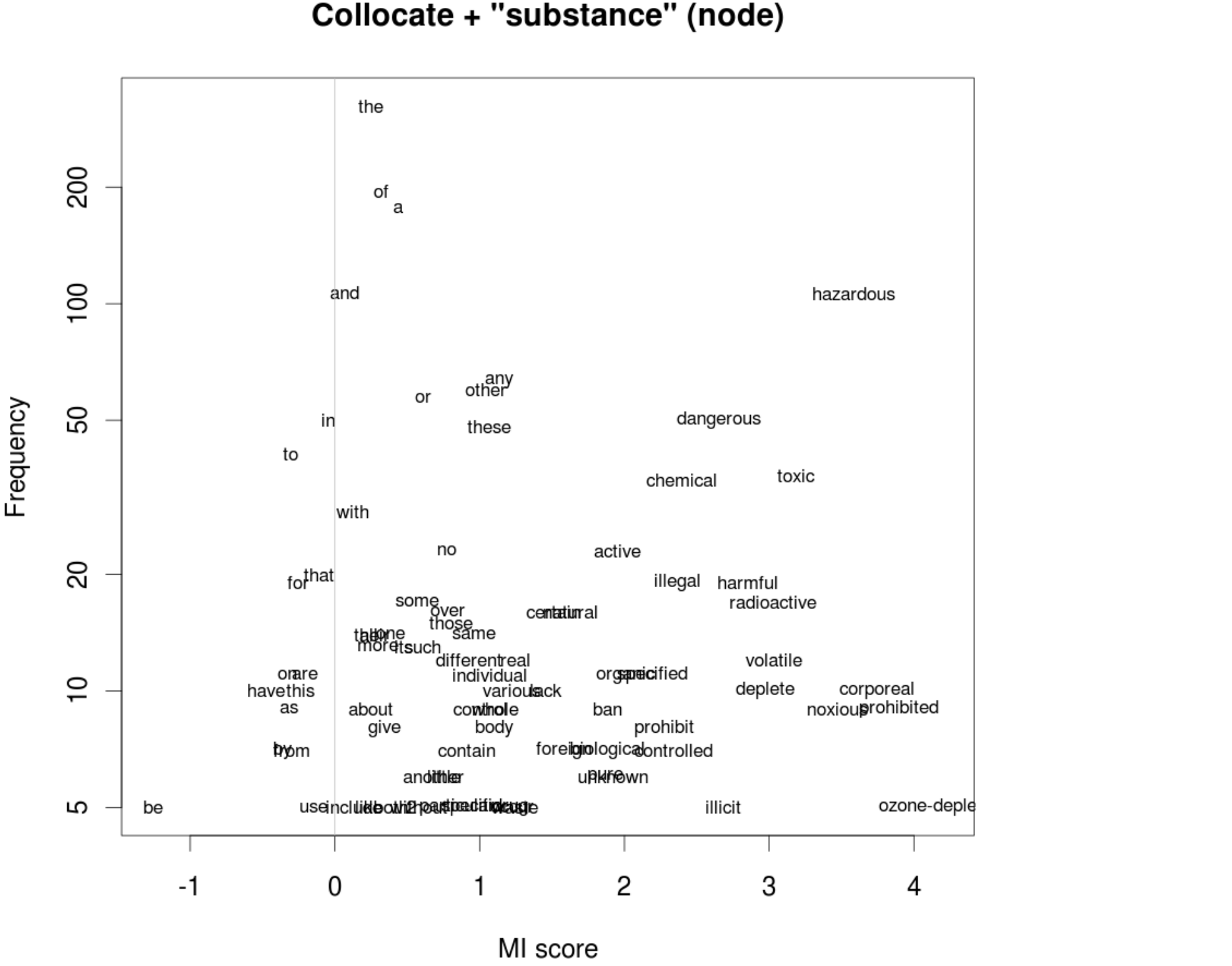
$$PMI(dangerous, substance) = \log \frac{\frac{50}{67063111}}{\frac{2825 \times 2657}{67063111^2}} = \log 447 = \mathbf{2.65}$$

$$(PMI(dangerous, and) = 0.36)$$

Task

- Measuring lexicographic appropriateness of automatically collected collocations
- Compare to Oxford Collocations Dictionary (2002)
  - BNC-based
  - Human-validated
- Ukwac, corpus of web texts
  - Total 2B tokens, here 67M random selection used
- Evert's CWB, UCS toolkits for processing
- Bigrams with freq. > 5


Collocate + "substance" (node)
Collocate Frequency
1e+01
1e+03
1e+05
MI score
-1
0
1
2
3
4
the
be
to
and
of
a
in
have
for
on
that
with
from
or
use
other
these
such
over
those
both
certain
natural
active
foreign
chemical
dangerous
illegal
prohibit
controlled
harmful
toxic
hazardous
radioactive
noxious
corporeal
prohibited
ozone-deple


Collocate + "substance" (node)
Frequency
MI score
200
100
50
20
10
5
-1
0
1
2
3
4
the
of
a
and
hazardous
any
other
or
in
these
dangerous
to
chemical
toxic
with
no
active
that
for
illegal
harmful
some
radioactive
over
those
same
more
such
different
real
volatile
individual
specified
deplete
corporeal
various
lack
about
whole
ban
noxious
prohibited
give
body
prohibit
from
contain
controlled
another
unknown
be
illicit
ozone-deple

Collocate + “substance”

OCD	CORPUS	GOOD			
34 total	**82** total	**18 (53%)**			
		MI	**MS**	**MI** cutoff 10	**MS** cutoff 10
	First **5**	**2 (40%)**	**5 (100%)**	4 (80%)	5 (100%)
	First **10**	**6 (60%)**	**9 (90%)**	8 (80%)	6 (60%)
	First **15**	**9 (60%)**	**10 (66%)**	10 (66%)	7 (47%)

Collocate + “charm”

OCD	CORPUS	GOOD			
17 total	**29** total	**8 (47%)**			
		MI	**MS**	**MI** cutoff 10	**MS** cutoff 10
	First **5**	**2 (40%)**	**2 (40%)**	/	/
	First **10**	**4 (40%)**	**4 (40%)**	/	/
	First **15**	**8 (53%)**	**8 (53%)**	/	/

Collocate + “network”

OCD	**CORPUS**	**GOOD**			
30 total	**391** total	**27 (90%)**			
		MI	**MS**	**MI** cutoff 10	**MS** cutoff 10
	First **5**	**0**	**3 (60%)**	/	3
	First **10**	**0**	**7 (70%)**	1 (10%)	7
	First **15**	**1 (7%)**	**9 (60%)**	3(20%)	9

Conclusions

- If possible, increase frequency threshold for MI
- MS outperforms MI in that relevant collocates are located on the top of the list

- Try other POS
- Bigger corpus
- Comparison of OCD vs. corpus collocates difficult
  - BNC vs. ukwac, criteria of dictionary editors
- OCD data old (BNC > 20 year-old texts)
  - Out-of-date collocates for “network”