

***‘Milk, bread and toothpaste’:  
Adapting Data Mining  
techniques for the analysis of  
collocation at varying levels of  
discourse***

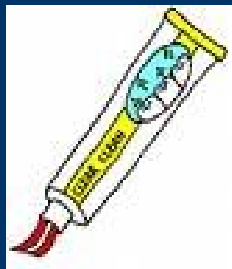
Rob Sanderson, Matthew Brook O’Donnell  
and Clare Llewellyn



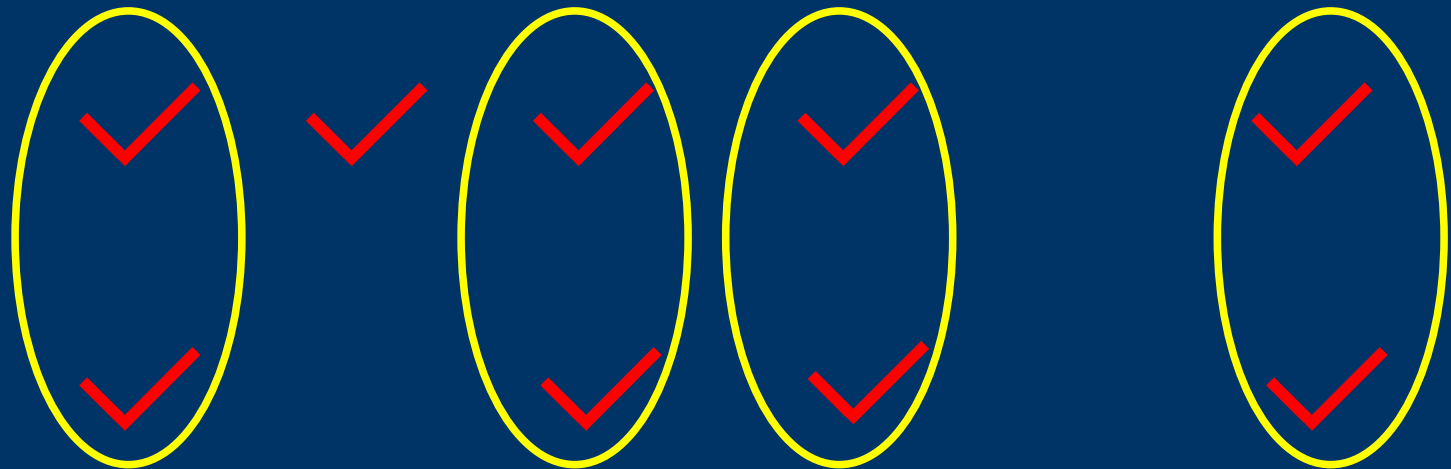
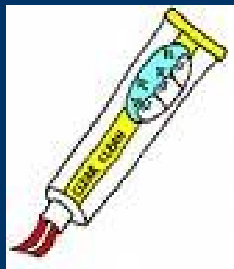
# *What happens with all that supermarket loyalty card data?*



# What happens with all that supermarket loyalty card data?



# What happens with all that supermarket loyalty card data?



bread & milk

4/6 (67%)

bread --> milk

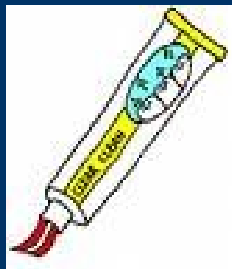
80% (4/5)

milk --> bread

100% (4/4)



# What happens with all that supermarket loyalty card data?



bread & toothpaste 2/6 (34%)

bread --> toothpaste 40% (2/5)

toothpaste --> bread 100% (2/2)



# *Corpus linguistic context: Collocation*

- Sinclair – discoverer of collocation
  - focus on pairs – node & (significant) collocate
  - 4 word optimum window size (OSTI etc.)
  - statistics well understood (Evert, Church etc.)
  - extensions/developments
    - filtering (per sentence, POS, constituents) – (Evert)
    - targeted analysis – collocation measures (Gries & Stefanowitch)
    - levels below and above the word (Baayen)
    - search for ngrams, frames and sets of more than two words
    - Concgrams (Cheng, Greaves & Warren)
- 
-

# *What can ARM offer to corpus linguistics?*

- base level – find collocate pairs
- reduce directionality & node focus
- find sets of 2+ co-occurring items
- find associations beyond the 9 word window at various levels of discourse



# *ARM overview: Terminology*

- ITEM SETS
  - ASSOCIATION RULES
  - SUPPORT
  - CONFIDENCE
- 
-

# *Association Rule Mining for Text*

- Requires Integers not strings -- one per type
  - No Order or Distance
  - No Frequency (but some algorithms have weight)
  - Apply to a range of discourse levels
    - Window of N words (sliding, discrete, centered on term)
    - Split at verb forms
    - Clause (as per shallow parser)
    - Sentence
    - Paragraph [subsection, section, chapter]
    - Full Text
- 
-

# *Desirables for Text ARM*

- No stoplist (or very minimal)
  - No *requirement* for PoS tagging
  - Flexible 'unit' (transaction) sizes
  - Not impossible amounts of data to look through
  - Order of results is important [cf google]
- 
-

# *Initial Experiments*

- Using 100,000 articles from the Home News section of *The Guardian* 1998-2004 (~45 mill. wds)
    1. 11 word span around a selected node within sentence & whole sentence search
    2. unfocused rule generation with an iterative mining process
    3. mining rules across an entire article from articles containing words from CONTROVERSY set
- 
-

# Example 1: the fact that

## Procedure

- a. using a flexible 11 word window, stopping at sentence breaks
- b. retrieving whole sentences containing *the fact that*

- stoplist: *a, an, the, and, fact, that*
- support threshold: 0.1% (i.e. item found in 5+ sentences)
- confidence threshold: 30% ( A --> B, B with A in 30+% of A instances)

## Results: Summary

- 5971 windows containing *the fact that*
  - 4480 sentences containing *the fact that*
- } *in 4198 records*
- 6253 item sets found
  - 12868 rules
- 
-

## *Example 1: 'the fact that' -- item sets*

waking to  
woken to  
of regardless  
of reminded  
of ignore  
of spite  
of cannabis  
of underlined  
spite in  
alerted to  
stated in  
comfort in

to ignore  
praised introducing  
of highlighted  
reflected in  
to resigned  
hide to  
of combined  
of proud  
of sight  
to referring  
lie in  
sharp in

of in  
drew to  
to reasonable  
of lies  
to reflect  
lies in  
reference to  
to respond  
of proportion  
of to  
glasgow in  
to in

*for 11 word window search*

# Example 1: 'the fact that' -- rules

the fact that	complicated	-->	by
	spite	-->	in
	compounded	-->	by
	may been	-->	have
	trying	-->	to
	according	-->	to
	referring	-->	to
	think there	-->	i
	secret i	-->	no of
	heightened	-->	by
	wake	-->	up to
	stems	-->	from
	stated	-->	in
	referred	-->	to
	much made	-->	of

all rules  
have 100%  
confidence

*for 11 word window search*

# Example 1: 'the fact that' -- rules

the fact that	due	-->	to
	trying	-->	to
	led	-->	to
	spite	-->	in of
	unable	-->	to
	none	-->	of
	attempt	-->	to
	according	-->	to
	kind	-->	of
	sort	-->	of
	number has	-->	of
	number been	-->	of
	proud	-->	of

all rules  
have 100%  
confidence

*for sentence search*

# *Example 1: trying to + the fact that*

## *the fact that X trying to Y*

- (1) But *the fact that* someone is trying to sell these weapons in the US is nothing new.
  - (2) ... and *the fact that* the department is trying to disown talk of targets at this late stage tells its own story.
  - (3) *The fact that* they are trying to ride the argument of inevitability demonstrates actually how weak their case is ...
- 
-

# *Example 1: trying to + the fact that*

## *trying to Z the fact that*

- (4) ... Thomas White, accused the Pentagon of trying to gloss over *the fact that* troops would remain for months.
  - (5) I am just trying to come to terms with *the fact that* I have seen a dead body for the first time in 10 years.
  - (6) Mr Riaz said his party was trying to address *the fact that* each time a white member of the party mentioned asylum or immigration, it was interpreted as racist.
- 
-

## *Example 2: sentence level rules*

- Uses every first sentence (TISC) from 100,000 articles in lemmatized form
  - stoplist: *the a an and be*
  - Strongest correlations appeared to be phrases  
e.g. *Sinn Fein, Tony Blair, Prime Minister*
  - Implemented merging procedure to capture these and treat them as single terms in a second iteration
- 
-

## Example 2: text-initial sentences

w1	w2	adj.	freq	%	ordering	
fein	sinn	265	265	100%	'sinn fein' 265	
alan	milburn	206	206	100%	'alan milburn' 206	
handling	of	160	162	98%	'handling of' 160	
accord	to	3443	3462	99%	'accord to' 3445 'to accord' 1	
cent	per	152	152	100%	'per cent' 167	
ministry	of	484	519	93%	'ministry of' 485 'of ministry' 1	
of	string	144	163	88%	'string of' 142 'of string' 2	
to	try	1256	1369	91%	'try to' 1220 'to try' 269	
for	responsible	301	318	94%	'responsible for' 302	
accuse	of	1094	2082	52%	-	
connection	in	with	404	416	97%	'in connection with' 404

## *Example 2: text-initial sentences – item sets*

foot mouth  
chancellor gordon\_brown  
student university  
united state  
pupil school  
right human  
nhs health  
secretary david\_blunkett  
foreign office  
hospital patient  
teacher school  
union european  
nhs hospital  
general election  
england wales

education school  
there no  
cabinet minister  
accord\_to survey  
child parent  
union leader  
more than  
find body  
conservative party  
party tory  
david\_blunkett home  
service health  
war iraq  
london ken\_livingstone  
court judge  
mayor london

## Example 2: text-initial sentences – rules

accuse after	-->	of	(97%)
accuse by	-->	of	(97%)
allow for	-->	to	(91%)
urge	-->	to	(89%)
time first	-->	for	(87%)
decision	-->	to	(86%)
plan on	-->	to	(85%)
part	-->	in	(84%)
end at	-->	of	(84%)
plan government	-->	to	(84%)
body find	-->	of	(84%)
pledge	-->	to	(83%)
give government	-->	to	(83%)
plan his	-->	to	(83%)
victim	-->	of	(83%)
convict	-->	of	(82%)
describe	-->	as	(81%)
plan new	-->	to	(81%)
murder on	-->	of	(81%)

---

---

## *Example 3: ARM at the article level*

- Retrieve articles containing at least one instance of *{controversy, row, embarrassment, blow}* in its first sentence (TISC)
  - POS filter for verbs, nouns, adjectives & adverbs
  - Stoplist: *be, have*
- 
-

# *Example 3: ARM at the article level*

resign

remark

blow

complaint

defend

adviser

dismiss

relation

row

leadership

criticism

shadow

document

smith

protest

declare

oppose

supporter

radio

criticise

chancellor

allegation

downing

investigate

william

urge

conduct

agreement

discuss

sunday

june

opposition

regard

reject

spokeswoman

brown

civil

threaten

ban

doubt

current

# Example 3: ARM at the article level

## 2 item sets

downing\_street prime\_minister

conservative tory

mr\_blair prime\_minister

office row

**issue row**

labour downing\_street

**secretary row**

minister row

secretary shadow

tony\_blair prime\_minister

labour cabinet

comment page

downing\_street minister

row claim

## 3 item sets

conservative party tory

**labour mp tory**

mr\_blair prime\_minister when

**labour party tory**

labour election party

labour mp party

one about row

labour mp minister

when make row

make about row

take make row

when about row

or about row

over make row

over when row

# Summary

- association rule mining is a useful tool to add to the corpus linguistic methodology
  - appears to reduce the effect of noise allowing the use of much wider windows, even up to the whole text level
  - has potential applications in
    - phraseology
    - 'pattern grammar'/ construction discovery
    - complex word associations (similar to concgrams)
  - need for refinements and wider evaluation (other corpora and comparison with range of collocational measures)
- 
-