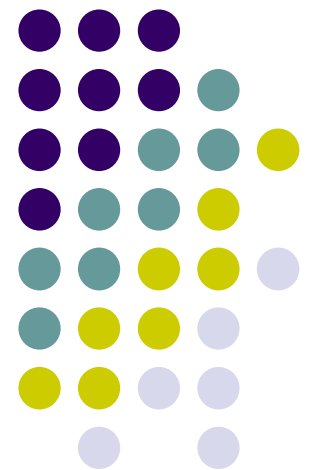


Exploring Text-initial Concgrams in a Newspaper Corpus

Matthew Brook O'Donnell,
Mike Scott & Michaela Mahlberg



Outline



- Background
 - AHRC funded project to investigation association of textual position and lexis in newspaper corpus
 - Summary of results so far
- Description of initial work using Concgrams
 - WordSmith 5.0 implementation of concgrams
 - Application to project research question

Background



- Hoey (2004, 2005) – Textual Colligation
 - words/items has particular associations with particular parts of a text, paragraph, sentence, e.g. *Once upon a time*
- AHRC Project – explore the extent and behaviour of text and paragraph initial words in hard news
 - M. Hoey, M. Scott, M. Mahlberg & M. O'Donnell
 - *The Guardian* Home News 1998-2004
 - 'initialness' at first defined in terms of significant frequency of occurrence in that position
 - Key Word procedure can be used for identification

Method: Classify sentences in corpus



- Classify sentences in each article according to where they occur:
 - TISC first sentence of first paragraph
 - PISC first sentence of subsequent paragraph
 - NISC any non-initial sentence

Method: Anatomy of a news article



Pope 'deeply sorry' but Muslim protests spread

→ **HISC**

(1) Italian police were yesterday ordered to tighten security at potential Catholic targets across the country as the leaders of the Roman Catholic church anxiously waited to see if a personal expression of regret by Pope Benedict would assuage Muslim fury over his remarks on Islam.

→ **TISC**
text initial sentence

(2) The Pope's speech in Germany last week, in which he quoted a medieval ruler who said Muhammad's innovations were "evil and inhuman", has led to widespread condemnation in the Muslim world.

→ **PISC**
paragraph initial sentence

(3) Last night the controversy seemed to have claimed its first victims when gunmen killed a 65-year-old Italian nun and her bodyguard at the entrance to a hospital where she worked in the Somalian capital, Mogadishu.

→ **NISC**
non-initial sentence

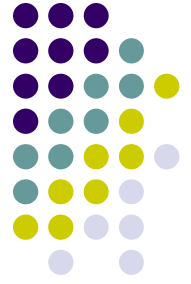
(4) A doctor said the nun, who was named as Sister Leonella Sgorbati, from Piacenza in northern Italy, had been shot four times in the back by two men with pistols.

→ **PISC**

(5) The attack was linked by some to the Pope's remarks.

→ **NISC**

Method: Key Word procedure to identify text-initial items



- Key Word procedure
 - Take a wordlist of items from TISC (Text-initial sentences)
 - And compare with one from NISC (Non-initial sentences)
- Some examples.....
 - *yesterday, last night*
 - *announced, suffered, stressed, according to a, it was {announced, revealed, disclosed, reported}*
 - *plans, report*
 - *over, a, under, after, against*
 - *fresh, branded*

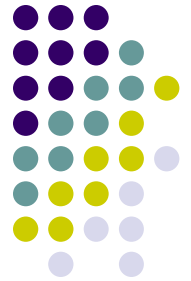


YESTERDAY: Sentence Position

- (1) George Bush **yesterday** suffered a blow to his argument...
(P1S1, Sept 27, 2006)
- (2) Bulgaria and Romania **yesterday** received the green light to join the EU in January
(P1S1, Sept 27, 2006)
- (3) **Yesterday**, the paper shelved its plan to give...
(P2S2, Apr 3, 1998)
- (4) **Yesterday** the prime minister was taunted by Iain Duncan Smith.
(P8S2, Apr 25, 2002)

	TISC	PISC	SISC	NISC
First word	0.2%	13.4%	10.9%	13.5%
Last word	25.7%	15.4%	9.8%	19.9%

It was V-ed...



		TISC	PISC	SISC	NISC
<i>it was</i>		3004	13469	14613	28236
	REVEALED	274	121	185	118
	ANNOUNCED	245	161	50	83
	CLAIMED	111	43	85	92
	DISCLOSED	111	21	21	22
	REPORTED	102	87	127	92
	CONFIRMED	78	28	51	25
	ALLEGED	25	36	40	65
		31.49%	3.60%	3.82%	1.76%

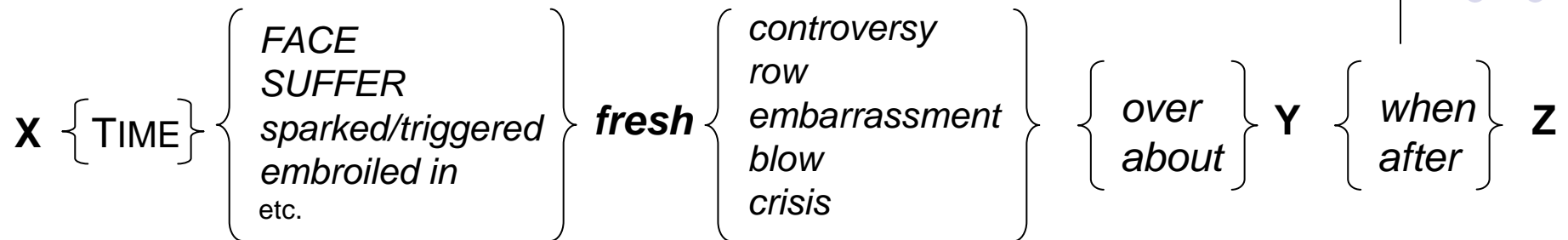
Method: Investigate Patterns



- *under* + NOUN + (to be) *announced* + TIME

	TISC	NISC
	78	10
under plans	24	1
under proposals	9	3
under a (X) scheme	9	1
under measures	5	
under a plan	3	
under a (x) deal	3	
under a programme	2	
under a procedure	2	
under guidelines	2	
under guidance	2	

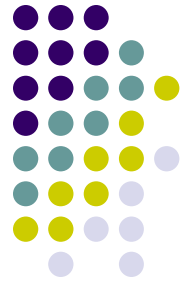
Method: Investigate Patterns



(1) *The government was yesterday embroiled in a fresh row over 'fat cat' pay rises after it emerged that Chris Woodhead...is to be reappointed...*

(2) *Labour faced fresh embarrassment over its funding yesterday when the independent electoral commission criticised it for continuing to break rules on donations*

Patterns of co-occurring Text-Initial Key Words



(1) *The government* was yesterday embroiled in a fresh row over 'fat cat' pay rises after it emerged that Chris Woodhead...is to be reappointed...

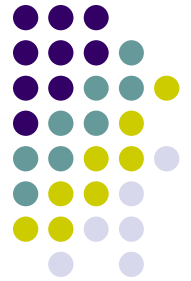
- text-initial key words
- Many of which upon examination of concordances are found to regularly co-occur in TISC sentences
- Is there a way to discover co-occurring key word sets more directly?

Concgrams



- **concgram**: ‘all the permutations of constituency and positional variation generated by the association of two or more words’ (Cheng, Greaves & Warren 2006)
- Attempt to move away from node centered investigation of word association
- **origin**: word(s) that form the basis of the search, can be of length 1, 2, 3, 4, 5, 6, etc.....
- constituent & positional variation
 - *announced plans, announced revised plans, announced that plans to...*
 - *announced plans, under plans announced...*

ConcGram©

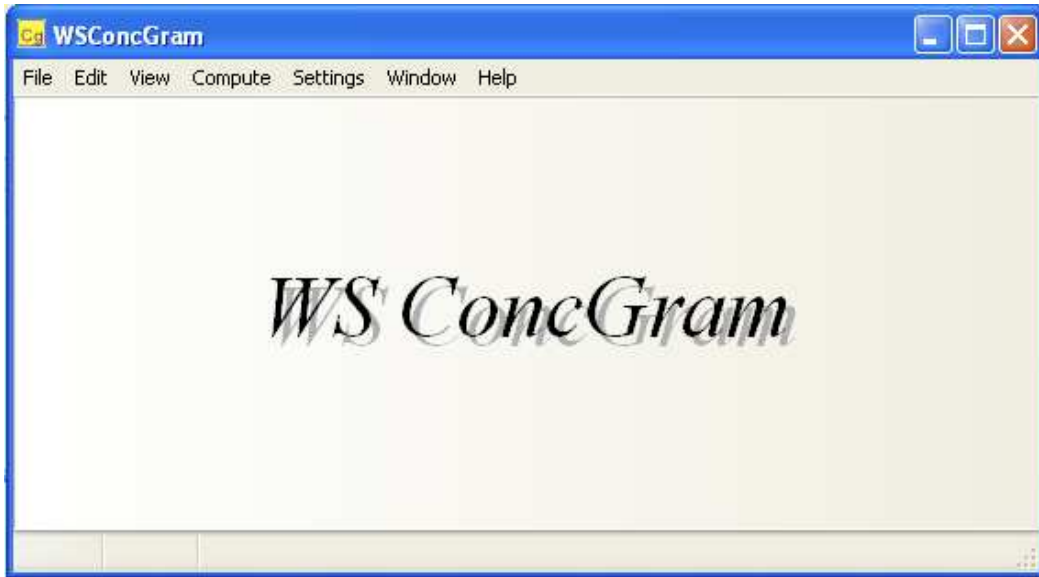
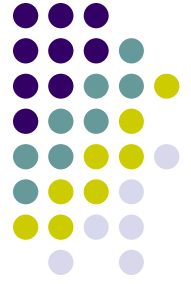


Example 5. 3-word concgram *challenges/facing/we*

1 at the moment **we** are **facing** tremendous **challenges** as our economy grapples with
2 at the moment **we're** **facing** tremendous **challenges** as our economy grapples with
3 based economy **we** are **facing** major **challenges** indeed difficulties may be with
4 **we** plan to overcome the numerous new **challenges** **facing** us but before I launch
5 I think that **we're** now looking at many **challenges** **facing** business community and
6 **we** are doing to tackle the difficulties and **challenges** **facing** us to lay the foundations
7 the one issue and that is the economic **challenges** **facing** Hong Kong and what **we're**
8 them what **we're** going to do about the **challenges** **we're** **facing** and where our
9 prolonged (pause) let me go back to the **challenges** **we** are **facing** in Hong Kong on
10 talking about I I think the the the most **challenges** that **we're** **facing** for FB
11 and shared pain to really resolve the **challenges** that **we** are **facing** these
12 er the both the opportunities and **challenges** that er **we** are **facing** er

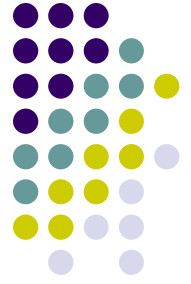
Cheng, Greaves & Warren (2006:426)

WSConcgram



- Evolving implementation in WordSmith 5.0
- Designed to find all concgrams in a corpus (10-20 mill. wds) in hours not days
- Builds on an index to find all pairs, triples, quadruplets, quintuplets, etc... within a specific span above a given threshold
- Allow user to browse and select concgrams from single, double, triple etc. 'origins'

WSConcgram procedure



- PAIRS
 - *fresh controversy, government fresh, fresh over, controversy, faced over, government faced, controversy after, fresh yesterday etc....*
- TRIPLES
 - *fresh controversy over, government fresh over, faced over after, faced fresh yesterday, etc..*
- QUADRUPLETS
 - *government faced fresh controversy, government fresh controversy after, etc..*
- QUINTRUPLETS
 - *government faced fresh controversy after, fresh controversy yesterday over after, etc....*
- etc.

Exploring some text-initial concgrams



- Index for TISC and a span of 10, stopping at sentence boundaries
 - Tokens: 3.12M Types: 58,432 Sentences: 113,288
- Results in 13,473 single origin items
- Current WS implementation allows export of resulting concgram strings but limited grouping around pairs, triples etc.
- Script to process concgram strings into tree representation with frequencies for various origins, e.g. *fresh + the + of* in 43 concgrams
- Nodes in tree imported into a wordlist

fresh concgrams in TISC

The screenshot shows the TISC software interface. At the top, the window title is 'TISC_10span.bas'. Below the menu bar (File, Edit, View, Compu), there is a search bar containing 'FRESH'. The main area is split into two panes. The left pane shows a table of word frequencies:

Word	Freq
FRESH	3,413
FAILED	3,400
RUN	3,399
FOOTBALL	3,399
INVESTIGATING	3,396
OFFICER	3,387
COMMONS	3,380
MANCHESTER	3,377
FACING	3,374
AL	3,367
BUSH	3,359
DRUGS	3,357
THOSE	3,314
PEACE	3,303
DEAL	3,296
EVER	3,255
WEEKEND	3,250

The right pane, titled 'Concgram', shows a list of phrases containing the word 'FRESH'. Each phrase is preceded by a checkbox. The phrases include:

- AMID A FRESH IN VIOLENCE WHICH AT
- AMID A FRESH THREAT OF AND
- AMID A OF FRESH ALLEGATIONS IN BY
- AMID AN OF FRESH IN THE
- AMID CLAIMS THAT FRESH EVIDENCE HAS TO
- AMID FRESH ALLEGATIONS BY FORMER THAT HE HAD
- AMID FRESH ALLEGATIONS OF THE PRIME
- AMID FRESH ALLEGATIONS THAT HE WAS UP BY HIS
- AMID FRESH CALLS FOR OF
- AMID FRESH CONCERNS ABOUT THE IN TO
- AMID FRESH EVIDENCE THAT IS
- AMID FRESH EVIDENCE THAT THE FOR
- AMID FRESH EVIDENCE THAT WAS TO
- AMID FRESH FEARS OF AMERICAN TROOPS YESTERDAY
- AMID FRESH FEARS THE PARTY THAT HE HAS
- AMID FRESH SIGNS OF OVER THE EURO AS
- AMID FRESH WARNINGS OF A IN THE NUMBER

At the bottom of the window, a status bar shows '13,473 OF 3,413'.

fresh + over concgrams in TISC

The screenshot shows the TISC software interface. The title bar reads "TISC_10span.base_pairs". The menu bar includes "File", "Edit", "View", "Compute", "Settings", and "Window". The main window is titled "OVER" and contains two panes. The left pane shows a list of words and their frequencies, with "OVER" selected. The right pane shows a list of concgrams for the word "FRESH".

Word	Freq
OVER	44,667
WHEN	43,626
YEAR	42,680
ITS	42,230
GOVERNMENT	42,042
NEW	41,437
INTO	39,820
WHICH	39,181
UP	38,639
THEY	37,613
POLICE	37,315
TWO	36,920
BRITISH	35,952
ONE	34,977
MORE	34,925
THAN	33,144
HER	32,561

Concgram

- FRESH
- PLUNGED INTO A FRESH ROW OVER AFTER IT EMERGED
- TONY BLAIR WAS LAST NIGHT IN FRESH CONTROVERSY OVER
- TONY BLAIR YESTERDAY A FRESH ROW OVER THE ISSUE
- UNDER FRESH PRESSURE YESTERDAY AS DEMANDED OVE...
- CHANCELLOR HAS BEEN INTO A FRESH ROW OVER
- FRESH PRESSURE YESTERDAY AS OVER CLAIMS THAT
- FRESH ROW OVER CRONYISM AFTER IT EMERGED THAT AT
- FRESH ROW OVER THE OF AS HE
- INTO A FRESH ROW OVER CRONYISM AFTER IT EMERGED TH...
- BEEN PLUNGED INTO A FRESH ROW OVER AFTER IT
- BLAIR YESTERDAY A FRESH ROW OVER THE ISSUE OF
- OVER BRITAIN'S MEMBERSHIP OF THE SINGLE CURRENCY IN...
- OVER BY MEDIA TYCOON RUPERT MURDOCH A FRESH DIPLO...
- OVER HIS RESHUFFLE AND A FRESH AS
- OVER KOSOVO WERE GIVEN FRESH AS THREE US
- OVER NHS TOOK A FRESH TWIST YESTERDAY WHEN THE
- OVER SECTION TOOK A FRESH TWIST

13,473 RUN 400

fresh + over + government concgrams in TISC

The screenshot shows the TISC (Text Search and Indexing) software interface. At the top, a green banner contains the text "fresh + over + government concgrams in TISC". Below this, the main window is titled "OVER" and contains two panes. The left pane displays a list of words and their frequencies, with "GOVERNMENT" selected. The right pane shows a list of concordance lines (Concgram) for the selected word, with the first line highlighted.

Word	Freq
OVER	44,667
WHEN	43,626
YEAR	42,680
ITS	42,230
GOVERNMENT	42,042
NEW	41,437
INTO	39,820
WHICH	39,181
UP	38,639
THEY	37,613
POLICE	37,315
TWO	36,920
BRITISH	35,952
ONE	34,977
MORE	34,925
THAN	33,144
HFR	32,561

Concgram

- FRESH
- GOVERNMENT FACES A FRESH REVOLT OVER TOMORROW'S E...
- GOVERNMENT FACES FRESH CRITICISM THIS WEEK OVER BRIT...
- GOVERNMENT IS BRACED FOR FRESH CRITICISM AND SOME RE...
- GOVERNMENT LAST NIGHT SUFFERED A FRESH SETBACK OVE...
- GOVERNMENT MADE A FRESH ATTEMPT TO FEARS OVER GENE...
- GOVERNMENT WAS LAST NIGHT EMBROILED IN FRESH CONTRO...
- GOVERNMENT WAS LAST NIGHT EMBROILED IN FRESH CONTRO...
- GOVERNMENT WAS LAST NIGHT STRUGGLING TO FRESH EMBA...
- GOVERNMENT WAS YESTERDAY EMBROILED IN A FRESH ROW ...
- HEALTH SELECT COMMITTEE THE GOVERNMENT FACES A FRES...
- FRESH ROW OVER POLITICAL YESTERDAY AFTER THE GOVERN...
- GOVERNMENT FACED FRESH EMBARRASSMENT OVER DEVELO...

13,473 GOVERNMENT 12

fresh + over + after concgrams in TISC

The screenshot shows the TISC software interface. At the top, a green banner contains the text "fresh + over + after concgrams in TISC". Below this, the word "AFTER" is entered in a search box. The main window is divided into two panes. The left pane displays a table of word frequencies, with "AFTER" selected. The right pane, titled "Concgram", shows a list of phrases containing the word "AFTER".

Word	Freq
AFTER	81,372
HAS	78,572
HIS	77,856
IT	76,546
HAVE	72,710
ARE	71,890
NIGHT	69,235
WHO	64,688
THEIR	60,188
WILL	55,937
WERE	53,613
HE	51,984
BEEN	50,572
POUNDS	46,606
OVER	44,667
WHEN	43,626
YEAR	42,680

Concgram

- PLUNGED INTO A FRESH ROW OVER AFTER IT EMERGED
- EMBROILED IN A FRESH ROW OVER POLITICAL YESTERDAY ...
- EMBROILED IN A FRESH ROW OVER YESTERDAY AFTER
- FACED FRESH EMBARRASSMENT OVER LAST NIGHT AFTER IT
- FACED FRESH EMBARRASSMENT OVER YESTERDAY AFTER ...
- FRESH CONTROVERSY OVER AMERICAN LAST NIGHT AFTER
- FRESH EMBARRASSMENT OVER FUNDING LAST NIGHT AFTE...
- FRESH EMBARRASSMENT OVER YESTERDAY AFTER THE CO...
- FRESH OVER AFTER BEING TO
- FRESH PRESSURE OVER HIS AFFAIRS LAST NIGHT AFTER IT
- FRESH ROW OVER POLITICAL YESTERDAY AFTER THE GOVE...
- FRESH ROW OVER WITHIN THE TORY PARTY YESTERDAY AF...
- FRESH ROW OVER YESTERDAY AFTER THE
- FRESH STRIKES WHEN THE WAR IN IRAQ IS OVER AFTER
- GOVERNMENT FACED FRESH EMBARRASSMENT OVER DEVE...
- INTO A FRESH OVER AFTER BEING FORCED TO
- LABOUR MANIFESTO FACED FRESH EMBARRASSMENT OVER...
- LAST NIGHT PLUNGED INTO A FRESH OVER IMMIGRATION AFF...

13,473 WHICH 31

[Expand All](#) | [Contract All](#)

FRESH

- 📁 THE 1325
- 📁 A 1267
- 📁 OF 870
- 📁 TO 806
- 📁 NIGHT 515
- 📁 YESTERDAY 503
- 📁 LAST 488
- 📁 IN 405
- 📁 OVER 389
 - 📁 FOR OVER 35
 - 📁 INTO OVER 34
 - 📁 OVER WHEN 36
 - 📁 AFTER OVER 27
 - 📁 AFTER OVER ROW 10
 - ⌘ AFTER IT OVER 8
 - ⌘ AFTER EMBARRASSMENT OVER 6
 - ⌘ AFTER OVER PLUNGED 5
 - 📁 OVER THAT 18
 - ⌘ AND OVER 15
 - 📁 OVER WAS 30
 - 📁 AS OVER 23
 - ⌘ BY OVER 9
 - 📁 HIS OVER 31
 - 📁 OVER WITH 22
 - 📁 OVER ROW 72
 - 📁 EVIDENCE OVER 8
 - ⌘ OVER WILL 11

BEEN PLUNGED INTO A FRESH ROW OVER AFTER IT
EMBROILED IN A FRESH ROW OVER POLITICAL YESTEI
EMBROILED IN A FRESH ROW OVER YESTERDAY AFTE
FACED FRESH EMBARRASSMENT OVER LAST NIGHT AI
FACED FRESH EMBARRASSMENT OVER YESTERDAY A
FRESH CONTROVERSY OVER AMERICAN LAST NIGHT /
FRESH EMBARRASSMENT OVER FUNDING LAST NIGHT
FRESH EMBARRASSMENT OVER YESTERDAY AFTER TI
FRESH OVER AFTER BEING TO
FRESH PRESSURE OVER HIS AFFAIRS LAST NIGHT AFTI
FRESH ROW OVER CRONYISM AFTER IT EMERGED THA
FRESH ROW OVER PAY AFTER IT EMERGED
FRESH ROW OVER POLITICAL YESTERDAY AFTER THE
FRESH ROW OVER WITHIN THE TORY PARTY YESTERD
FRESH ROW OVER YESTERDAY AFTER THE
FRESH STRIKES WHEN THE WAR IN IRAQ IS OVER AFTE
GOVERNMENT FACED FRESH EMBARRASSMENT OVER
INTO A FRESH OVER AFTER BEING FORCED TO
INTO A FRESH ROW OVER CRONYISM AFTER IT EMERG
LABOUR MANIFESTO FACED FRESH EMBARRASSMENT
LAST NIGHT PLUNGED INTO A FRESH OVER IMMIGRAT
NIGHT PLUNGED INTO A FRESH OVER IMMIGRATION A
OVER THE JACK STRAW FRESH CONTROVERSY YESTEI
OVER WERE THROWN INTO FRESH YESTERDAY AFTER
PLUNGED INTO A FRESH OVER AFTER
PLUNGED INTO A FRESH ROW OVER AFTER IT EMERGE
UNDER FRESH PRESSURE OVER HIS TAX AFFAIRS LAST

TISC

FRESH CONCESSIONS
FRESH CONTROVERSY IN NIGHT
FRESH CONTROVERSY NIGHT OVER
FRESH CONTROVERSY YESTERDAY
FRESH COOK
FRESH DOWNING STREET
FRESH EMBARRASSMENT THAT
FRESH EMBROILED IN OVER
FRESH EMBROILED OVER
FRESH EVIDENCE THAT THE
FRESH FACE TO
FRESH FACED WHEN YESTERDAY
FRESH FOR ITS
FRESH GOVERNMENT THAT THE
FRESH GOVERNMENT WAS
FRESH GOVERNMENT'S THE
FRESH HAVE THE
FRESH IN LAST WAS
FRESH INTO THROWN
FRESH IT LAST NIGHT THAT
FRESH IT LAST THAT
FRESH IT NIGHT THAT
FRESH IT OF
FRESH LAST THAT THE
FRESH MAKE TO

NISC

FRESH MAKE
FRESH MAY
FRESH NEED
FRESH NEW OF
FRESH OF SAID
FRESH OVER TO
FRESH PROVIDE
FRESH SOME THE
FRESH START UNDER
FRESH THEIR TO
FRESH TO WE
FRESH A ALSO
FRESH A AND NEW
FRESH A FOR START
FRESH AND IN THE
FRESH ARE IN
FRESH AT TO
FRESH BE OF
FRESH BEEN THERE
FRESH BLAIR
FRESH BRING
FRESH BRITAIN
FRESH FACE
FRESH FOR OF THE
FRESH FROM OF THE

Comparing Concgrams



- Key concgrams?
 - Take two lists of concgrams for a certain item (e.g. *fresh*) produced using the node grouping method with occurrences above a certain threshold
 - Use total number of concgrams found for item as list size
 - Apply Key Word procedure

Key Concgrams TISC against NISC

FRESH NIGHT

FRESH LAST NIGHT
FRESH YESTERDAY
FRESH LAST
FRESH A NIGHT
FRESH AFTER
FRESH OVER
FRESH A LAST NIGHT
FRESH THE YESTERDAY
FRESH WHEN
FRESH ROW
FRESH WHEN YESTERDAY
FRESH CONTROVERSY
FRESH FACING
FRESH TODAY
FRESH AFTER NIGHT
FRESH A YESTERDAY
FRESH A LAST
FRESH A ROW
FRESH AFTER LAST
FRESH NIGHT THE
FRESH AFTER LAST NIGHT
FRESH INTO

FRESH LAST NIGHT THE
FRESH BLAIR TONY
FRESH EMERGED
FRESH NIGHT OF
FRESH EMBARRASSMENT
FRESH TONY
FRESH A INTO
FRESH NIGHT WHEN
FRESH CRISIS
FRESH ROW THE
FRESH LAST WHEN
FRESH A TODAY
FRESH A OVER
FRESH LAST NIGHT WHEN
FRESH NIGHT WAS
FRESH OVER YESTERDAY
FRESH LAST NIGHT WAS
FRESH INTO NIGHT
FRESH IN NIGHT
FRESH A
FRESH OVER ROW
FRESH BLAIR
FRESH LAST NIGHT TO

Key Concgrams TISC against NISC

FRESH LAST NIGHT TO
FRESH NIGHT TO
FRESH BLOW
FRESH PRESSURE
FRESH OF YESTERDAY
FRESH TO YESTERDAY
FRESH THE WHEN YESTERDAY
FRESH INTO LAST NIGHT
FRESH A FACING
FRESH NIGHT OVER
FRESH ATTACK
FRESH LAST NIGHT OF
FRESH AMID
FRESH IN LAST NIGHT
FRESH NIGHT THAT
FRESH THE WHEN
FRESH AFTER YESTERDAY
FRESH SUFFERED
FRESH A NIGHT THE
FRESH A WHEN
FRESH FACING THE
FRESH AGAINST
FRESH A ROW THE

FRESH WAS YESTERDAY
FRESH THE TODAY
FRESH OVER THE
FRESH AFTER THE
FRESH LAST OVER
FRESH EMERGED IT
FRESH A THE YESTERDAY
FRESH CONTROVERSY NIGHT
FRESH A OVER ROW
FRESH LAST WAS
FRESH FACED
FRESH LAST THAT
FRESH A BLOW
FRESH LAST NIGHT THAT
FRESH AFTER INTO
FRESH A CRISIS
FRESH AS YESTERDAY
FRESH CONTROVERSY LAST
FRESH PRESSURE TO
FRESH POLITICAL
FRESH LAST NIGHT OVER
FRESH LAST THE
FRESH GOVERNMENT



Summary

- In our newspaper corpus we have found many items with a strong association with text position (especially text-initial)
- These associations are particularly complex and involve levels of nesting or combination
- Initial results suggest that the concgram procedure is able to discover these nested patterns

**A concgram we hope will become
a part of your mental corpus**

**! 2009 be conference corpus July
linguistics Liverpool there**

Corpus Linguistics Conference Liverpool July 2009 Be There!



UNIVERSITY OF
LIVERPOOL

Corpus Linguistics
Conference
20 - 23 July 2009

The fifth annual Corpus Linguistics Conference -
Monday 20 July: Pre-conference Workshops
Tuesday 21 to Thursday 23 July: Main Conference

The background of the poster is a photograph of the Liverpool city skyline, featuring the Royal Liver Building, the Port of Liverpool Building, and the Sea Tower, viewed from the water.