language · research · ideology · power · L:I:P

UCREL Corpus Research Seminar
University Centre for Computer Corpus Research on Language
Lancaster University Departments of Linguistics and Computing

The presentation of Islam and Muslims in the UK press, 1998-2009
Identifying topics through detailed wordlist analysis

Costas Gabrielatos

(Lancaster University)

21 June 2010

Abstract

Corpus-based approaches to critical discourse analysis usually move from establishing large-scale patterns and trends to examination of keywords and/or collocates to close concordance analysis. This presentation will demonstrate how ...

- detailed concordance analysis can provide the motivation for large-scale analysis;

- the examination of high-frequency content words (through multi-sorted concordances) can provide strong indications of the main topics in a specialised corpus of newspaper articles;

- extending the analysis by examining mid-frequency words can provide a more comprehensive picture by establishing groups of words which a) indicate topics, b) specify contextual elements, and c) provide the co-text essential for the discussion of topics.

Project

Title: *The representation of Islam and Muslims in the UK press, 1998-2009*

Aims: Differences between newspapers

Diachronic changes

Funding body: ESRC

Principal investigator: Paul Baker

Co-investigator: Tony McEnery

Researcher: Costas Gabrielatos

Data: source and query

Online database: *Nexis UK*

Query

- *Alah OR Allah OR ayatolah OR burka! OR burqa! OR chador! OR fatwa! OR hejab! OR imam! OR Islam! OR Koran OR Mecca OR Medina OR Mohammedan! OR Moslem! OR Muslim! OR mosque OR mufti! OR mujaheddin! OR mujahedin! OR mullah! OR muslim! OR Prophet Mohammed OR Q'uran OR rupoush OR rupush OR sharia OR shari'a OR shia! OR shi-ite! OR Shi'ite! OR sunni! OR the Prophet OR wahabi OR yashmak! AND NOT Islamabad AND NOT shiatsu AND NOT sunnily*

Corpus

Articles: *200,000*

Words: 143 *million*

Spelling normalisation

Sub-corpora:
- per newspaper
- per year (1998-2009)
- broadsheets/tabloids
- political orientation

Business
Daily Express + Sunday Express
Daily Mail + Mail on Sunday
Daily Mirror + Sunday Mirror
Guardian + Observer
Independent + Independent on Sunday
People
Daily Star + Daily Star Sunday
Sun
Telegraph + Sunday Telegraph
Times + Sunday Times

Corpus tools and methodology

Tools

- WordSmith 5, Sketch Engine

Methodology

- Keywords, word sketches, detailed wordlist analysis (multi-sorted concordances)

Word Sketch

- Collocates of a word within a grammatical construction
    - *Muslim*_ADJ + Noun
    - *Muslim*_NN as Subject / Object (→ Agent / Patient)

Motivation (1)

Expanded concordance analysis of sketches:

NOUN *against/on Islam*

n+against-i+N, 905, 35.4		
war	340	5.78
crusade	64	9.02
blasphemy	28	8.49
prejudice	23	6.53
conspiracy	20	5.87
fight	20	4.68
struggle	12	4.32
hatred	12	4.16
West	11	7.88
offence	10	4.11

n+on-i+N, 1238, 9.1		
war	243	5.29
attack	157	4.62
expert	74	6.32
view	61	5.02
book	46	4.24
comment	36	5.72
lecture	29	7.67
assault	29	5.59
remark	26	6.16
conference	23	4.80
back	19	4.88
slur	10	7.34

This is the most salient sketch for Islam_N

Motivation (2)

Recurring juxtaposition of two frames:

war etc. *against/on* *Islam*
terror* / Osama bin Laden etc.

But Israel's best interests will not be served by poking sticks in the Palestinians' eye at a time when the West is straining every muscle to persuade the Muslim world that the **war against Osama bin Laden** is not a **war against Islam**. The Americans may be very close to Israel, but Mr Sharon will be making a mistake of historic dimensions if he forces them to choose between maintaining their alliance with Israel and winning the wider **war against terror**.

[*The Independent*, 26.09.2001]

Izzadeen told listeners: 'Everyone knows, Muslims and non-Muslims, that the **war on terror** is a **war against Islam**, and I'm telling you something, if they don't stop this there is going to be a very strong reaction from the community.'
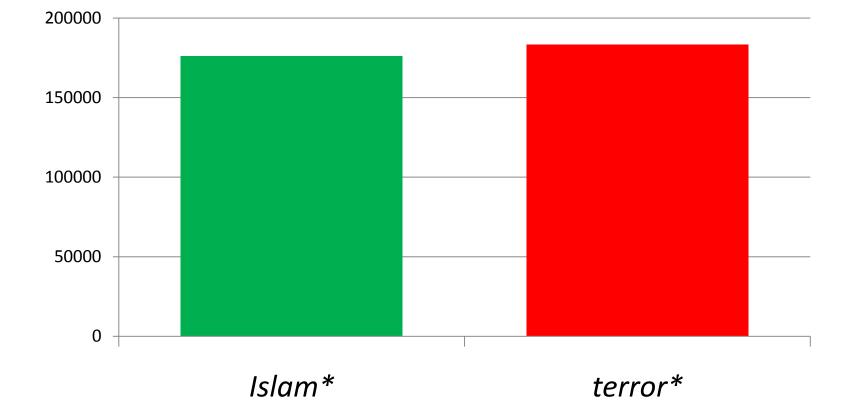
[*Daily Mail*, 23.09.2006]

Motivation (3)

- Frequent juxtaposition of frames → frequent co-occurrence of *Islam** and *terror** within the same text.

Are the frequent references to terrorism a characteristic of the juxtaposition of these frames, or a more general attraction?

Frequency comparison of *Islam** and *terror** in the corpus.

200000

150000

100000

50000

0

Islam*   *terror**

Corpus texts containing *Islam** tend to also contain *terror**

However

Islam* is a query term - **terror*** is not!

terrorist	45818
terror	44315
terrorism	43564
terrorists	34923
terrorised	646
terrorise	514
terrorising	426
terrors	218
terrorist-related	132
terrorism-related	119
terror-related	77
terrorises	32

The four clearly most frequent forms have an average frequency of around 40,000.

Examination of all content types with a frequency of at least 40,000 (40k content types).

40k content types

- Nouns, adjectives, verbs

  - 147 types

  - 0.03% of corpus content types

  - 15.1% of corpus content tokens

- 85 (58%) are clear indicators of topics or contextual elements in corpus articles (40k indicators)

  - 0.02% of content types

  - 9.5% of content tokens

40k indicators

Reference to ...	Types
... armed or violent conflict, and the attendant issue of death ... police / security	13
... countries or regions involved in armed conflict of some description - in which religion is, or is perceived to be, directly or indirectly, one of the main causes of that conflict.	29
... governance and leadership, human aspects (e.g. age and sex), time, money, and quantification. → These types usually refer to the (willing or unwilling) participants in the conflict, their representatives/leaders, or the place/time of conflict.	42

40k indicators directly or indirectly refer to issues of conflict ...

... although the query consisted of terms relating to a religion, its believers and attendant customs/practices.

Tip of the iceberg?

Can 85 very frequent content types – representing only 10% of the corpus content tokens – be indicative of the main corpus topics? **Yes**
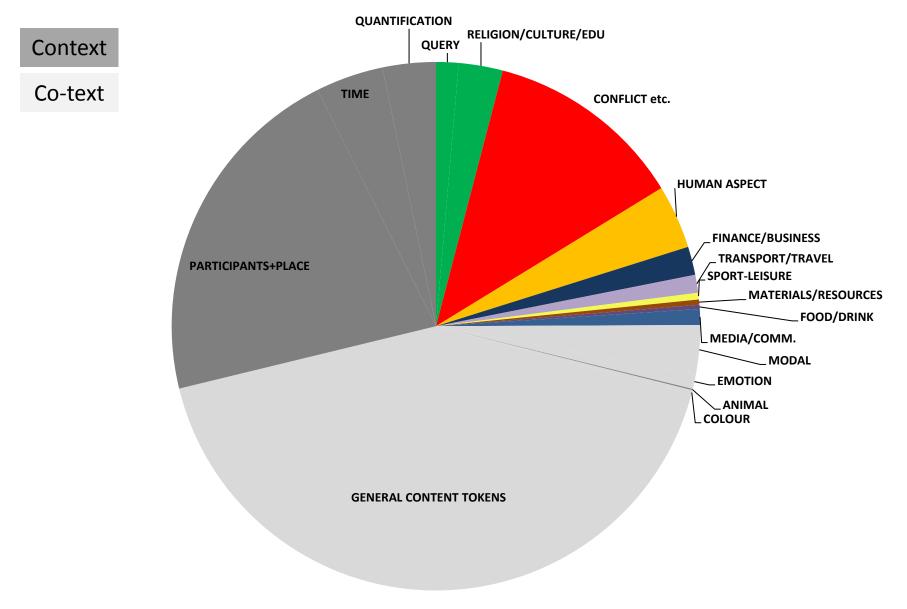
Examination of 4k content types

2,981 types
0.5% of content types
61% of content tokens

4k content types

Three categories:

- Types indicating *topics* in the corpus articles.

- Types indicating *contextual elements*
  - e.g. participants, place, time.

- Types with *general meaning / use*
  - Essential for discussing the topics, but not clear topic-indicators themselves (e.g. general adjectives, modality/attitude markers).
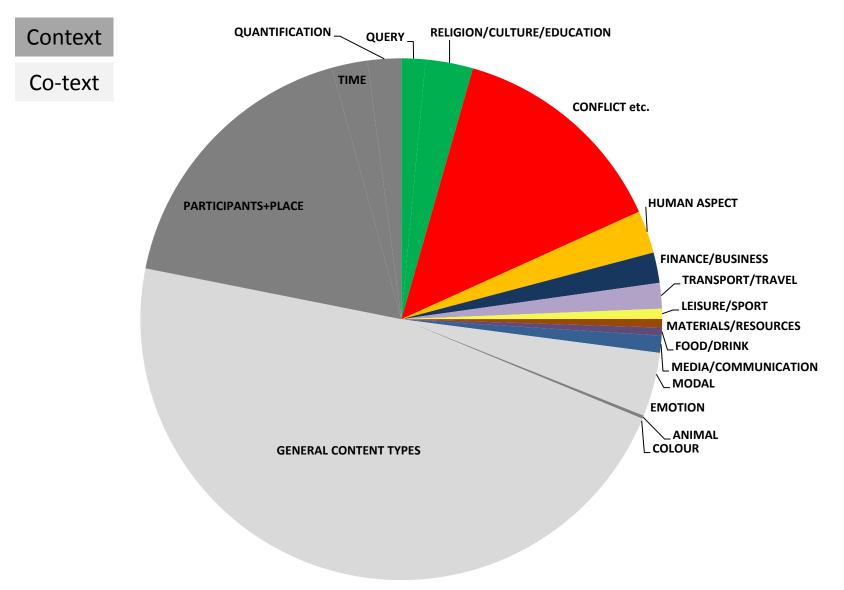
4k tokens

Context

Co-text

QUANTIFICATION

QUERY

RELIGION/CULTURE/EDU

CONFLICT etc.

TIME

HUMAN ASPECT

FINANCE/BUSINESS

TRANSPORT/TRAVEL

SPORT-LEISURE

PARTICIPANTS+PLACE

MATERIALS/RESOURCES

FOOD/DRINK

MEDIA/COMM.

MODAL

EMOTION

ANIMAL

COLOUR

GENERAL CONTENT TOKENS

Could the distribution in terms of tokens be the result of some very frequent types?

Distribution in terms of types

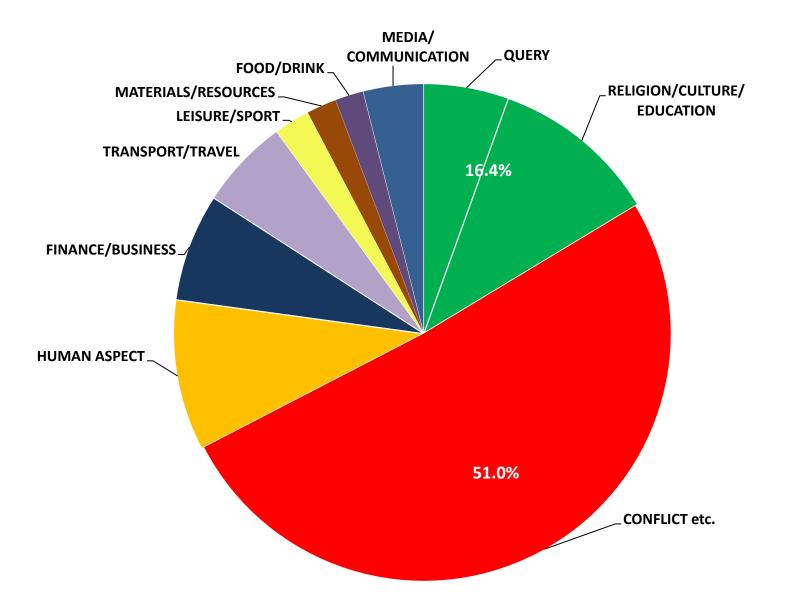# 4k types

Sub-categories of 'conflict'	Types	Tokens	Definition	Example types
Argument	30	286171	Differences in opinion, goals etc.	*argue, agreement, criticism, debate, division, resolution*
Violent/armed conflict	145	1883949	Armed or violent conflict	*guns, invasion, military, war*
Death	27	441639	References to violent death	*buried, death, died, genocide, kill*
Exclusion/ Inclusion	15	120118	Issues of racial (in)tolerance	*ban, discrimination, diversity, racism, tolerance*
Extremism	12	103780	References to extreme views and/or practices	*extremists, fundamentalists, hardline, moderate, radicals*
Health	9	69412	References to ill-health	*disease, doctors, medical*
Law/ Crime	55	688748	References to law-breaking and punishment	*charged, conviction, crime, criminal, detainees, drugs, (il)legal, jail, judge, legitimate, prisoner, suspects*
Peace/ Safety	15	245719	Although nominally positive, they are used in discussions of their lack and need.	*aid, calm, friend, peace, protection, safety, welfare*
Problem/ Danger	19	206764	Reference to (potentially) problematic/dangerous actions or situations	*chaos, corruption, crisis, dangerous, emergency, panic, problem, risk, trouble, worry*
Terror	6	177512	Reference to terrorism or terrorists	*hijackers, terrorists, towers*
Violence/ Damage	85	1055092	Reporting of / reference to instances of (threat of) physical/psychological violence towards humans, and damage in property.	*attack, blast, bloody, bomb, brutal, casualties, destruction, explosion, fire, hurt, injured, rape, refugees, suffering, threatening, victims, violence*

4k topic-indicators: tokens

FOOD/DRINK

MEDIA/COMMUNICATION

MATERIALS/RESOURCES

LEISURE/SPORT

QUERY

RELIGION/CULTURE/
EDUCATION

TRANSPORT/TRAVEL

16.3%

FINANCE/BUSINESS

HUMAN ASPECT

48.9%

CONFLICT etc.

4k topic-indicators: types

The topic or the press?

Is this focus on conflict characteristic of the corpus, or of general newspaper reporting practices?

Manual keyword analysis

Manual keyword analysis (1)

Frequency comparison of the group of 'conflict' tokens in Islam-UK and the news subcorpus of BNC (BNCnews)

Islam-UK (142,962,543)	BNCnews (9,897,378)	*Islam-UK %*	*BNCnews %*	Diff.	LL
5,278,904	272,131	*3.70*	*2.75*	**+34.5%**	24,618.53

- More than two-thirds (69.1%) of 'conflict' tokens are key in Islam-UK

However, BNC articles are from the late 1990s

Manual keyword analysis (2)

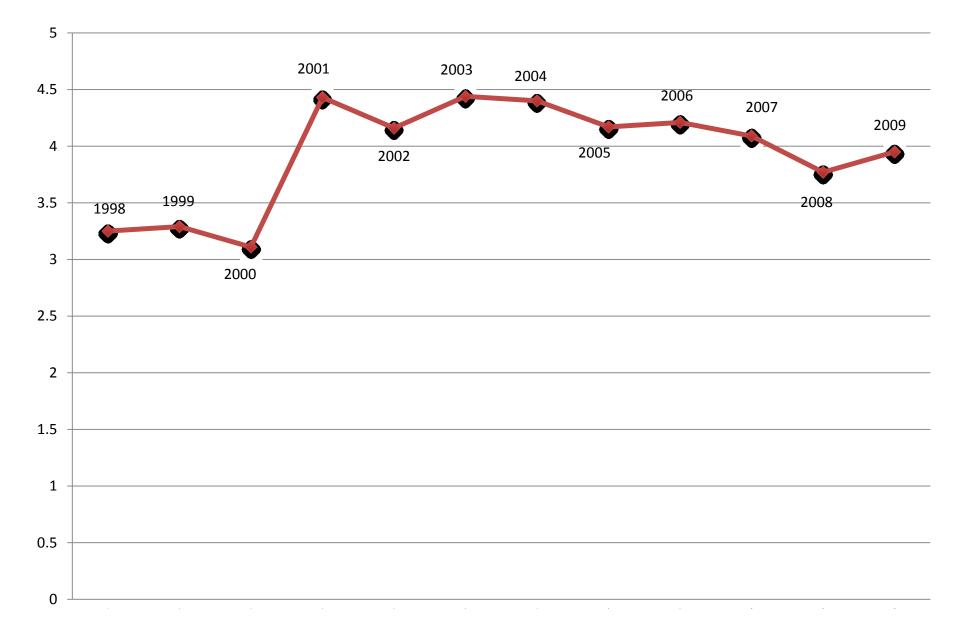Frequency comparison of the group of 'conflict' tokens in:
- IslamUK_2005-2007
- BE2006news (Baker, 2009) - 82% of texts between 2005-2006.

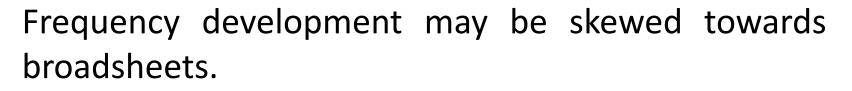| Islam-UK 2005-2007 (44,120,772)	BE2006 news (174,963)	*Islam-UK 2005-2007 %*	*BE2006 news %*	Diff.	LL
1,894,614	5,626	*4.29*	*3.22*	**+33.2%**	517.78

Breakdown of all keywords:
- Origin: 27.8%
- Conflict: 24.9%
- Religion: 13.6%
- Gov.+Org.:   7.7%

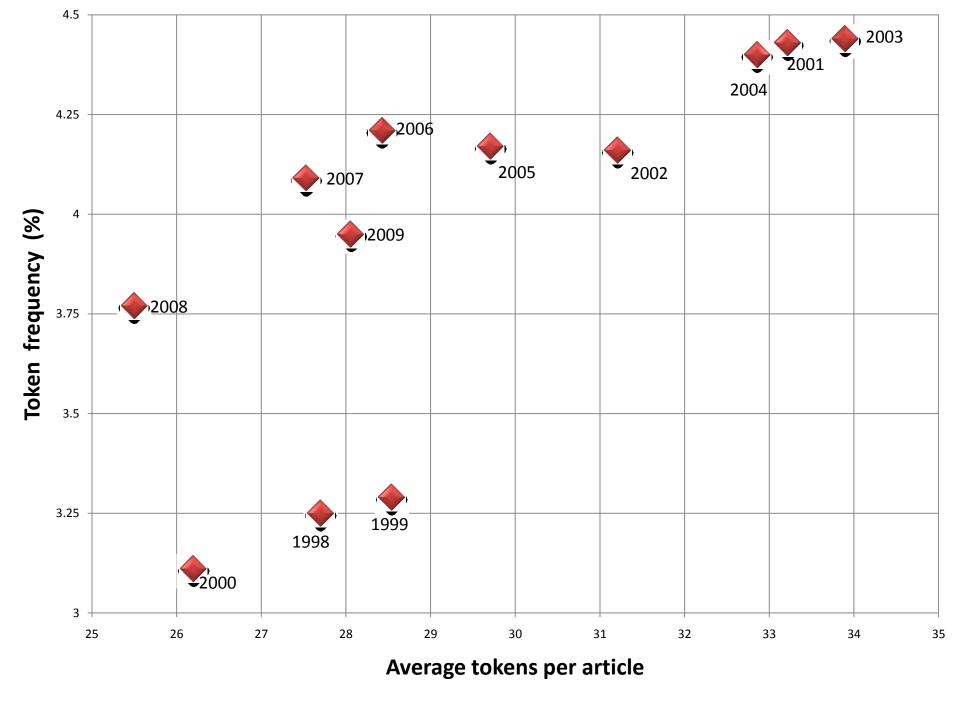# Diachronic frequency development of 'conflict' tokens

In the corpus, broadsheet articles are, on average, about 80% longer than tabloid articles.
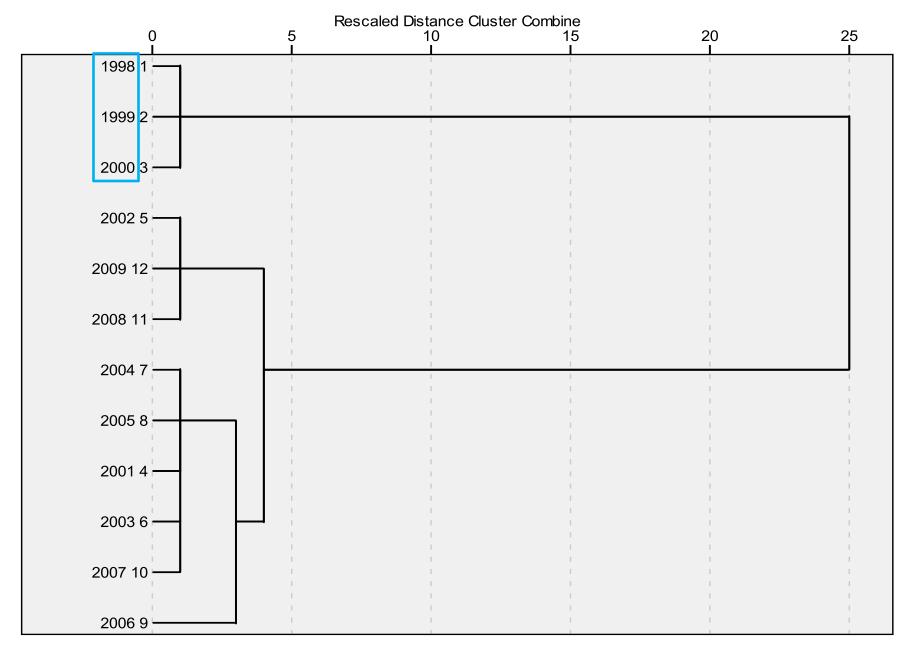
Frequency development may be skewed towards broadsheets.

Account of average tokens per article

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine

Conclusions

- The dominant discourses on Islam and Muslims in the UK national press seem to revolve around issues of conflict - and its participants, logistics and repercussions.

- The link between Islam/Muslims and conflict in UK newspapers seems to have intensified after 9/11.

- High-frequency content words can reliably reveal central topics in newspaper articles.