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.
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

<table>
<thead>
<tr>
<th>n+against-i+N, 905, 35.4</th>
<th>n+on-i+N, 1238, 9.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>war</td>
<td>war</td>
</tr>
<tr>
<td>crusade</td>
<td>attack</td>
</tr>
<tr>
<td>blasphemy</td>
<td>expert</td>
</tr>
<tr>
<td>prejudice</td>
<td>view</td>
</tr>
<tr>
<td>conspiracy</td>
<td>book</td>
</tr>
<tr>
<td>fight</td>
<td>comment</td>
</tr>
<tr>
<td>struggle</td>
<td>lecture</td>
</tr>
<tr>
<td>hatred</td>
<td>assault</td>
</tr>
<tr>
<td>West</td>
<td>remark</td>
</tr>
<tr>
<td>offence</td>
<td>conference</td>
</tr>
<tr>
<td></td>
<td>back</td>
</tr>
</tbody>
</table>

This is the most salient sketch for Islam_N
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.
Corpus texts containing *Islam* tend to also contain *terror*.

However

*Islam* is a query term - *terror* is not!
<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>terrorist</td>
<td>45818</td>
</tr>
<tr>
<td>terror</td>
<td>44315</td>
</tr>
<tr>
<td>terrorism</td>
<td>43564</td>
</tr>
<tr>
<td>terrorists</td>
<td>34923</td>
</tr>
<tr>
<td>terrorised</td>
<td>646</td>
</tr>
<tr>
<td>terrorise</td>
<td>514</td>
</tr>
<tr>
<td>terrorising</td>
<td>426</td>
</tr>
<tr>
<td>terrors</td>
<td>218</td>
</tr>
<tr>
<td>terrorist-related</td>
<td>132</td>
</tr>
<tr>
<td>terrorism-related</td>
<td>119</td>
</tr>
<tr>
<td>terror-related</td>
<td>77</td>
</tr>
<tr>
<td>terrorises</td>
<td>32</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Reference to ...</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>... armed or violent conflict, and the attendant issue of death ... police / security</td>
<td>13</td>
</tr>
<tr>
<td>... 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.</td>
<td>29</td>
</tr>
<tr>
<td>... governance and leadership, human aspects (e.g. age and sex), time, money, and quantification.</td>
<td>42</td>
</tr>
<tr>
<td>➔ These types usually refer to the (willing or unwilling) participants in the conflict, their representatives/leaders, or the place/time of conflict.</td>
<td></td>
</tr>
</tbody>
</table>

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

Distribution in terms of types
<table>
<thead>
<tr>
<th>Sub-categories of ‘conflict’</th>
<th>Types</th>
<th>Tokens</th>
<th>Definition</th>
<th>Example types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argument</td>
<td>30</td>
<td>286171</td>
<td>Differences in opinion, goals etc.</td>
<td>argue, agreement, criticism, debate, division, resolution</td>
</tr>
<tr>
<td>Violent/armed conflict</td>
<td>145</td>
<td>1883949</td>
<td>Armed or violent conflict</td>
<td>guns, invasion, military, war</td>
</tr>
<tr>
<td>Death</td>
<td>27</td>
<td>441639</td>
<td>References to violent death</td>
<td>buried, death, died, genocide, kill</td>
</tr>
<tr>
<td>Exclusion/Inclusion</td>
<td>15</td>
<td>120118</td>
<td>Issues of racial (in)tolerance</td>
<td>ban, discrimination, diversity, racism, tolerance</td>
</tr>
<tr>
<td>Extremism</td>
<td>12</td>
<td>103780</td>
<td>References to extreme views and/or practices</td>
<td>extremists, fundamentalists, hardline, moderate, radicals</td>
</tr>
<tr>
<td>Health</td>
<td>9</td>
<td>69412</td>
<td>References to ill-health</td>
<td>disease, doctors, medical</td>
</tr>
<tr>
<td>Law/Crime</td>
<td>55</td>
<td>688748</td>
<td>References to law-breaking and punishment</td>
<td>charged, conviction, crime, criminal, detainees, drugs, (il)legal, jail, judge, legitimate, prisoner, suspects</td>
</tr>
<tr>
<td>Peace/Safety</td>
<td>15</td>
<td>245719</td>
<td>Although nominally positive, they are used in discussions of their lack and need.</td>
<td>aid, calm, friend, peace, protection, safety, welfare</td>
</tr>
<tr>
<td>Problem/Danger</td>
<td>19</td>
<td>206764</td>
<td>Reference to (potentially) problematic/dangerous actions or situations</td>
<td>chaos, corruption, crisis, dangerous, emergency, panic, problem, risk, trouble, worry</td>
</tr>
<tr>
<td>Terror</td>
<td>6</td>
<td>177512</td>
<td>Reference to terrorism or terrorists</td>
<td>hijackers, terrorists, towers</td>
</tr>
<tr>
<td>Violence/Damage</td>
<td>85</td>
<td>1055092</td>
<td>Reporting of / reference to instances of (threat of) physical/psychological violence towards humans, and damage in property.</td>
<td>attack, blast, bloody, bomb, brutal, casualties, destruction, explosion, fire, hurt, injured, rape, refugees, suffering, threatening, victims, violence</td>
</tr>
</tbody>
</table>
4k topic-indicators: tokens

- HUMAN ASPECT
- FINANCE/BUSINESS
- TRANSPORT/TRAVEL
- LEISURE/SPORT
- MATERIALS/RESOURCES
- FOOD/DRINK
- MEDIA/COMMUNICATION
- QUERY
- RELIGION/CULTURE/EDUCATION

- CONFLICT etc.

16.3%
48.9%
4k topic-indicators: types

- RELIGION/CULTURE/EDUCATION: 51.0%
- CONFLICT etc.: 48.6%
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)

<table>
<thead>
<tr>
<th>Islam-UK</th>
<th>BNCnews</th>
<th>Islam-UK %</th>
<th>BNCnews %</th>
<th>Diff.</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(142,962,543)</td>
<td>(9,897,378)</td>
<td>3.70</td>
<td>2.75</td>
<td>+34.5%</td>
<td>24,618.53</td>
</tr>
<tr>
<td>5,278,904</td>
<td>272,131</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1,894,614</td>
<td>5,626</td>
<td>4.29</td>
<td>3.22</td>
<td>+33.2%</td>
<td>517.78</td>
</tr>
</tbody>
</table>

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
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.