CLARET Workshop

Compiling topic-specific corpora from limited-access online databases

Costas Gabrielatos
Lancaster University

Lancaster University, 31 March 2008

Menu

- Motivation
- Defining ‘topic-specific corpora’
- Compiling a topic-specific corpus
- Online text databases
- Selecting query terms

Case study

Task

- Corpus for the project "Discourses of refugees and asylum seekers in the UK Press 1996-2006".

Project aims

- To explore the discourses surrounding refugees and asylum seekers, and account for the construction of the identities of these groups, in the UK press.

Methodology

- Collocational analysis
- Keyword analysis (broadsheets vs. tabloids)
- Concordance analysis

Topic-specific corpora

- 'Topic': entities, concepts, issues, relations, states, processes.
- Mainly used in critical discourse studies.
- Focus usually on groups / issues
  - representation of minority / disadvantaged groups in mainstream or political texts (e.g. refugees)
  - self-presentation of minority / disadvantaged groups
  - self-presentation of dominant groups (e.g. corporate executives)
  - moral panics (social, political, economic or health issues)

Compiling topic-specific corpora: Issues (1)

- **Precision**:
  Is the corpus free of irrelevant documents?
- **If not**, …
  - statistical results (e.g. keyness) may be skewed;
  - corpus compilation/annotation can become unduly time-consuming.

- **Recall**:
  Does the corpus contain all relevant documents existing in the database?
- **If not**, some aspects of the entities etc. in focus may be over/under-represented or even missed.

Compiling topic-specific corpora: Issues (2)

- **Sub-corpora are important**
  - source (e.g. per newspaper)
  - time period (e.g. per month)
- ***Why?***
- Comparisons
  - e.g. between years, between newspapers
- Diachronic aspect
  - e.g. frequency developments of terms / collocations

Downloading should facilitate sub-corpora creation

Compiling topic-specific corpora: Issues (3)

- Careful when selecting core query terms.
  - Be clear about the topic.
  - Topic under investigation vs. Expected attitudes.
  - e.g. ‘racism’

Online text databases: pros/cons (1)

- **Targeted search**: source, time span, content (using indexing or query)
- 'Blank query': **all texts** in terms of source, time span, content.
- Restricted number of texts returned per query
  - e.g. Lexis Nexis
  - 1-2 weeks from a single UK national newspaper
  - Less than a day (= nothing) from all UK national newspapers
- Restricted number of texts per download
- Indexing not always helpful

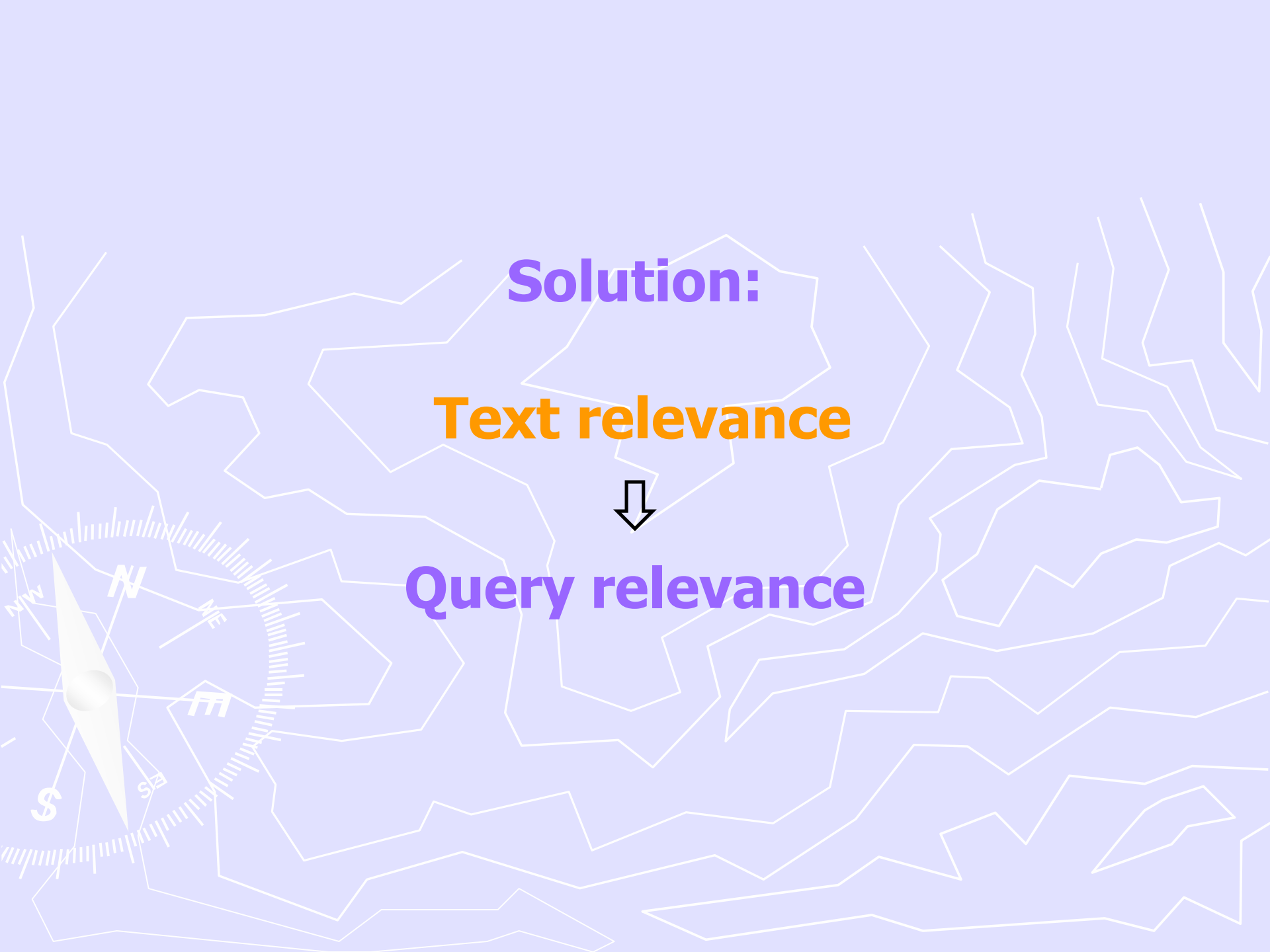
⇨ Use of a query

⇨ Source and time span adjustments

⇨ Repeated downloads

Online text databases: pros/cons (2)

- **Calculation of precision/recall problematic**
- Calculation requires:
- Number of relevant database documents
  - ⇨ unknown
- Number of relevant retrieved documents.
- Relevance can be established through …
  - human judgement
    - ⇨ too time consuming
  - indexing (absolute or weighted)
    - ⇨ may exclude metaphorical uses
    - ⇨ documents containing one relevant term merit inclusion as much as those containing two or more


Solution:
Text relevance
⇩
Query relevance

Selecting query terms

- "Discourses of refugees and asylum seekers in the UK Press 1996-2006".
- Obvious starting point: *refugee** OR *asylum seeker**
- *Core query terms (CQTs)*

Why not stop here?

Query expansion (1): Content

- Representations of groups in newspapers may "include or exclude social actors to suit their interests and purposes" (van Leeuwen, 1996: 38).
- Some terms may "share a common ground" (Baker & McEnery, 2005: 201).
  - ⇨Groups (and issues, concepts etc.) may be referred to using '**alternative' terms**
  - ⇨Terms may be used **interchangeably**
  - ⇨e.g. *refugees* - *immigrants*

Query expansion (2): Methodology

- If a term is frequently found in documents containing CQTs, then it may be related to them.
- It may be useful to examine the use of these terms within documents which do not contain CQTs.

- The inclusion of such terms allows the examination of …
  - collocate overlap between focus terms and related terms - or terms used as being related (e.g. *refugees / asylum seekers -- immigrants / migrants*).
  - intercollocations with related terms.
- (Baker et al., 2007, 2008, in press; Gabrielatos & Baker, 2006a, 2006b, 2008)

The analysis will be more thorough if such terms are added to the query.

Why not come up with more terms ourselves (introspectively)?

Query expansion (3): Problems

- Investment in time = money.
  - e.g., addition of a single term, *terrorism*:
    - corpus size would increase six-fold
    - data collection time would increase 50-100%
- Introspective additions may skew quantitative analysis:
  - keyword comparisons (particularly with reference corpus).
  - collocation strength / statistical significance

Needed: more objective measure of the utility of additional query terms.

Existing techniques (1)

Information retrieval

(e.g. Baeza-Yates & Ribeiro-Neto, 1999; Chowdhury, 2004)

- Large number of processes and algorithms, but all require knowledge of…
  - number of relevant database documents
    - ⇨ unknown
  - number of relevant retrieved documents
    - ⇨ time consuming

Existing techniques (2)

BootCat

(Baroni & Bernardini, 2003, 2004; Baroni & Sharoff, 2005; Baroni, et al., 2006; Ghani, et al., 2001)

- Uses search engine queries.
- Selection of 'seeds' → Compilation of interim corpus from top *n* retrieved pages → Successive keyword comparisons and compilation of interim corpora → Query terms

⇨ Requires open access to database.

⇨ Theoretically possible with restricted access database, **but** prohibitively time consuming (multiple downloads).

⇨ Problems with keyword analysis.

Problems with keywords

- **Available reference corpora may cover a different time span from corpora to be constructed. In this case ...**
  - A large number of keywords will be seasonal.
  - Other KWs may be related to topic, but also related to a large number of other issues.

- **KW analysis treats the corpus as one document:**
  - can hide high frequency in small number of documents.
  - some KWs may be not representative of the majority of corpus documents.

Why not use Key KW analysis?

⇨ preparation of corpus would be prohibitively time consuming.

⇨ would not address problem of different time spans.

Utility of keywords

- A KW analysis can be used to suggest candidate terms.
- ***How?***
- Construction of **sample corpus** using the core query (*refugee** OR *asylum seeker**).
  - the sample corpus should contain texts spanning the target period
  - e.g. UK6: October 1996, December 1998, February 2000, April 2002, June 2004, August 2005 (2.6 mil. words)
- KW comparison with relevant general corpus.

Top 40 Keywords

UK6 * BNC Sampler

ISRAELI	2,620.0
PALESTINIAN	2,060.9
ISRAEL	1,637.5
POUNDS	1,306.7
JENIN	1,100.7
CAMP	1,081.6
PALESTINIANS	977.5
IMMIGRATION	954.7
HOME	909.6
BRITAIN	831.3
WHO	780.6
PEOPLE	741.6
BLAIR	731.7
SHARON	728.4
POLICE	660.3
ARAFAT	641.6
SAYS	639.0
SUICIDE	608.0
HE	591.1
WAR	571.1

ISRAELIS	546.0
ISRAEL'S	497.5
SECRETARY	496.2
SOLDIERS	490.6
UN	481.4
KILLED	478.9
IMMIGRANTS	478.7
EU	465.2
LAST	420.3
SAID	414.7
ARMY	406.4
CIVILIANS	397.0
THEY	387.3
HAS	386.7
GAZA	380.9
ATTACKS	378.8
AFGHANISTAN	374.4
BLUNKETT	371.6
POWELL	368.3
IRAQ	365.1

Query term relevance (QTR)

QTR: Purpose

- To select additional query terms which can be expected to return a sufficient number of relevant documents not containing the CQTs, without creating undue noise.

QTR: Nature

- Checks the extent to which a candidate term is found in texts containing at least one CQT.
- Looks for co-occurrence of a candidate term and the CQTs in every text.
  - ⇨Akin to *collocation* - span is the whole article (e.g. Kim & Choi, 1999).
  - ⇨Akin to *key KW* analysis.
- Is independent of reference corpora.

QTR: Calculation

- Use of **exploratory queries** on the same sources and time spans used for the sample corpus.
  - ⇨ To derive document frequencies containing each query.
- These sample corpora are **temporary**:
  - ⇨ Only accessible through database interface by use of a query.
- Use of **simple formula** to derive score suggesting degree of relevance for each candidate term.

QTR: Specifics

- If hits are above the database limit, …
  - time spans need to be broken down (e.g. weeks rather than months);
  - number of hits for each sub-query have to be tabulated and tallied.

Yes, the procedure is quite labour-intensive.

QTR: Formula

$$QTR = \frac{\text{No. of texts returned by: core query AND candidate term}}{\text{No. of texts returned by: candidate term}}$$

$$QTR = \frac{\text{No. of texts returned by: [}refugee*\text{ OR }asylum\ seeker*\text{] AND }migrant*}{\text{No. of texts returned by: }migrant*}$$

- QTR score range: *0-1*
- ***0*** = candidate term found in **no** texts containing core query
- ***1*** = candidate term found in **all** texts containing core query

OK, now what do we do with the scores?

QTR: The baseline score (B)

- QTR scores mean nothing if not compared to a score acting as a threshold for inclusion: the baseline score (B).
- B is the QTR of the lowest scoring core query term, when the other is used as the core query.

CQ AND T (*refugee** AND *asylum seeker**)	T (*asylum seeker**)	*QTR*
593	1403	*0.423*

CQ AND T (*asylum seeker** AND *refugee**)	T (*refugee**)	*QTR*
593	2596	*0.228*

B is *0.228*

Terms with QTR > 0.228 are added to the query

A note on *B*

- Does not need to be lowest QTR - it can be higher or lower according to how rich you want the corpus to be.
  - ⇨ A 'richer' corpus is expected to contain more noise.

QTR may not be enough

- Useful in establishing the baseline score (B).
- Corpus-sensitive: not helpful for inter-corpus comparisons.

Why compare QTR scores across corpora?

- Double checking:
  - ⇨ using two sample corpora from same database.
- Comparing use of same candidate terms in different sources (e.g. UK vs. US newspapers).

Corpus-sensitivity

	CQ AND T (*refugee** AND *asylum seeker**)	T (*asylum seeker**)	*QTR*
UK1	39	125	*0.312*
UK6	593	1403	*0.423*

	CQ AND T (*asylum seeker** AND *refugee**)	T (*refugee**)	*QTR*
UK1	39	349	*0.112*
UK6	593	2596	*0.228*

Relative QTR (RQTR)

RQTR

Measures relative distance of QTR from B.

$$RQTR = \frac{(QTR-B) * 100}{B}$$

- Independent of corpus.
- Min. negative score always -100 (QTR = 0).
- Max. positive score varies.
  - ⇨ Positive scores need to be normalised.

$$RQTRn = \frac{(QTR-B) * 100}{1-B}$$

Interpreting RQTR scores

RQTR	Interpretation
+100	**Full relevance**: the candidate term is always found in database texts containing one or more of the core query terms.
0	**Baseline relevance**: the candidate term has the same level of relevance as that set as the minimum for inclusion to the final query.
-100	**No relevance**: the candidate term is never found in database texts containing any of the core query terms.

If compiling more than one corpus,
the same query should be used for all corpora

RQTR: Steps

- Create **sample corpus** / corpora
- Perform **KW analyses** to identify candidate terms
- Supplement with **introspective candidates**
- Calculate **QTR** to establish **B** (can be used flexibly)
- Use QTR and B to calculate **RQTR**
  - If QTR>B use RQTR formula
  - If QTR<B use RQTRn formula

RQTR: Overview

- Not a precise measure.
- More reliable than keyness alone.
- Better than introspection.
- Allows consideration of introspectively relevant terms.
- Independent of reference corpora.
- Required minimum of two core query terms easily achieved.
- Sample corpus/corpora fairly quick to compile.
- Calculation is accessible.
- Time for establishing RQTR depends on number of candidate terms and documents returned per query.
- Ideally, additional terms should …
  - have non-negative RQTR
  - be key
  - be introspectively relevant

Details: Gabrielatos (2007)

References (1)

- Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval. London*: Addison Wesley.
- Baker, P. & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics* 4(2), 197–226.
- Baker, P., McEnery, T. & Gabrielatos, C. (2007). Using collocation analysis to reveal the construction of minority groups: The case of refugees, asylum seekers and immigrants in the UK press. Paper given at *Corpus Linguistics 2007*, University of Birmingham, UK, 27-30 July 2005. (Abstract and slides available online: http://eprints.lancs.ac.uk/602/)
- Baker, P., Gabrielatos, C. & McEnery, T. (2008). Using collocational profiling to investigate the construction of refugees, asylum seekers and immigrants in the UK press. *7th Conference of the American Association for Corpus Linguistics* (AACL 2008), Brigham Young University, Provo, Utah, 13-15 March 2008.
- Baker, P., Gabrielatos C., Khosravinik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008, in press). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3), 273-305.
- Baroni, M. & Bernardini, S. (2003). The BootCaT toolkit: Simple utilities for bootstrapping corpora and terms from the web, version 0.1.2. http://sslmit.unibo.it/~baroni/Readme.BootCaT-0.1.2.
- Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. *LREC 2004 Proceedings*, 1313–1316.
- Baroni, M. & Sharoff, S. (2005). Creating specialized and general corpora using automated search engine queries. Paper presented at *Corpus Linguistics 2005*, Birmingham University, 14–17 July 2005. (Available online: http://sslmit.unibo.it/~baroni/wac/serge_marco_wac_talk.slides.pdf.
- Baroni, M., Kilgarriff, A., Pomikálek, J. & Rychlý, P. (2006). WebBootCaT: Instant domain-specific corpora to support human translators. *Proceedings of EAMT 2006*, 247–252. (Available online: http://corpora.fi.muni.cz/bootcat/publications/webbootcat_eamt2006.pdf)

References (2)

- Chowdhury, G.G. (2004, 2nd ed.) *Introduction to Modern Information Retrieval.* London: Facet Publishing.
- Gabrielatos, C. (2007). Selecting query terms to build a specialised corpus from a restricted-access database. *ICAME Journal* 31, 5-43. (Also online: http://icame.uib.no/ij31/ij31-page5-44.pdf)
- Gabrielatos, C. & Baker, P. (2006a). Representation of refugees and asylum seekers in UK newspapers: Towards a corpus-based comparison of the stance of tabloids and broadsheets. *Critical Approaches to Discourse Analysis Across Disciplines* (CADAAD 2006), University of East Anglia, Norwich, UK, 29-30 June 2006. (Abstract and slides available online: http://eprints.lancs.ac.uk/250)
- Gabrielatos, C. & Baker, P. (2006b). Representation of refugees and asylum seekers in UK newspapers: Towards a corpus-based analysis. *Joint Annual Meeting of the British Association for Applied Linguistics and the Irish Association for Applied Linguistics* (BAAL/IRAAL 2006), 7-9 September 2006, University College, Cork, Ireland. (Abstract and slides available online: http://eprints.lancs.ac.uk/265/)
- Gabrielatos, C. & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics* 36(1), 5-38.
- Ghani, R., Jones, R. & Mladeni, D. (2001). Mining the web to create minority language corpora. *CIKM 2001*, 279–286.
- Kim, M-C. & Choi, K-S. (1999). A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management* 35(1), 19-30.
- van Leeuven, T. (1996). The representation of social actors. In C-R. CaldasCoulthard and M. Coulthard (eds.). *Texts and Practices. Readings in Critical Discourse Analysis*, 32–70. London: Routledge.