Techniques and Metrics for Corpus-based Approaches to Humanities and Social Sciences

Costas Gabrielatos

Edge Hill University

CONSISTENT COLLOCATION ANALYSIS (3)

Application / Utility

- Diachronic corpus analysis focusing on stability (rather than change) over time.
- Also pinpoints *seasonal collocates*: strong collocates in whole corpus because of large frequency in small number of sub-corpora.

Prerequisites

- Particular lexical items (nodes) must have been selected for analysis.
- Time-specific sub-corpora.

Definition of c-collocate

- Collocate in at least two-thirds of sub-corpora.
- Gap between appearance no larger than 10% of number of sub-corpora (e.g. if 10 annual sub-corpora, gap no larger than one year).

Calculation of c-collocates

- Collocation analysis of individual sub-corpora.
- Collocates derived through combination of two metrics, showing *effect size* (strength of attraction) and *statistical significance*, respectively (e.g. Mutual Information + Log Likelihood).
- Table with collocates per sub-corpus → Pivot tables in Excel.

Further Analysis

- Proportion of consistent collocates shared by particular nodes.
- Manual examination of multi-sorted concordances of consistent collocations → Establishment of patterns (e.g. semantic preference, discourse prosody, topics).

Table 1. Topics indexed by c-collocates in the RASIM corpus (10 years of British newspapers)

	refugees	***asylum seekers***	***immigrants***	***migrants***
refugees		ENTRY NUMBER ECON. BURDEN RETURN	ENTRY RESIDENCE	ENTRY
asylum seekers	ENTRY PLIGHT NUMBER RETURN		ENTRY LEGALITY PDT RESIDENCE	ENTRY
immigrants	ENTRY RESIDENCE PLIGHT NUMBER	ENTRY PLIGHT RESIDENCE LEGALITY		ENTRY ECON. THREAT
migrants	ENTRY RESIDENCE PLIGHT PDT	PLIGHT	PDT ENTRY RESIDENCE ECON. THREAT LEGALITY	

QUERY TERM RELEVANCE (2)

Application / Utility

- Objective establishment of query terms for the compilation of *topic-specific corpora*.
  → Corpora containing texts related to particular entities, concepts, issues, relations, actions etc.
- Particularly useful when corpus texts derived from limited access databases (e.g. LexisNexis)

Prerequisites

- Existence of at least two clearly relevant terms: *core query terms* (CQT)

Nature / Characteristics

- 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 (CT) and the CQTs in every text.
  → Akin to collocation: span is the whole article.
- Independent of reference corpora.

Procedures and Metrics

- Use of exploratory queries on the same sources to be used for the sample corpus to derive document frequencies containing each query.
- Use of simple formula to derive score suggesting degree of relevance for each candidate term.

$$QTR = \frac{\text{No. of texts returned by Core Query Terms AND Candidate Term}}{\text{No. of texts returned by Candidate Term}}$$

- QTR value range: 0-1
  0 = candidate term found in no texts containing the CQTs.
  1 = candidate term found in all texts containing the CQTs.
- **Baseline** (B): the QTR of the lowest scoring CQT, when the other(s) is/are used as the core query.
- However, QTR is corpus-sensitive: not useful for inter-corpus comparisons.
- **Relative Query Term Relevance** (**RQTR**): Measures relative (%) distance of QTR from B.

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

- Min. score always *-100* (as QTR would be 0). However, max. score varies according to B → Must be normalised (**RQTRn**): % distance from max.RQTR: RQTRn = RQTR*100 / max.RQTR).

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

- **RQTR values:** *-100* (no relevance) — *0* (baseline relevance) — *100* (full relevance)

WAVE PEAK & TROUGH (WPT) METHOD (4)

Application / Utility

- Diachronic analysis of topic-specific corpora (e.g. newspaper articles on Islam/Muslims).
- Aids a "corpus-based contextual analysis" (1): frequency peaks of articles / terms related to particular entities/processes lead to identification of significant events.
- Awareness of relevant context can assist the interpretation of corpus findings (e.g. through collocational networks);
- The WPT method can objectively pinpoint time periods within which texts can be selected (through downsampling) for qualitative CDA.
- The WPT method can also be employed in the diachronic frequency analysis of lexical items.

Prerequisites

- Topic-specific corpus
- Time-specific sub-corpora.

Description

- Identifies statistically significant peaks and troughs in the diachronic frequency development of topic-specific texts or lexical items.
- Also identifies diachronic frequency trends.

Establishing statistically significant peaks/troughs

- Calculation of relative change between two consecutive time points, using the logarithm of the frequency difference → Plot of relative change over time.
- Statistical significance was derived through a non-parametric regression analysis.
- Essentially, what is calculated is the extent to which the frequency difference between two time points is both sizable and statistically significant.

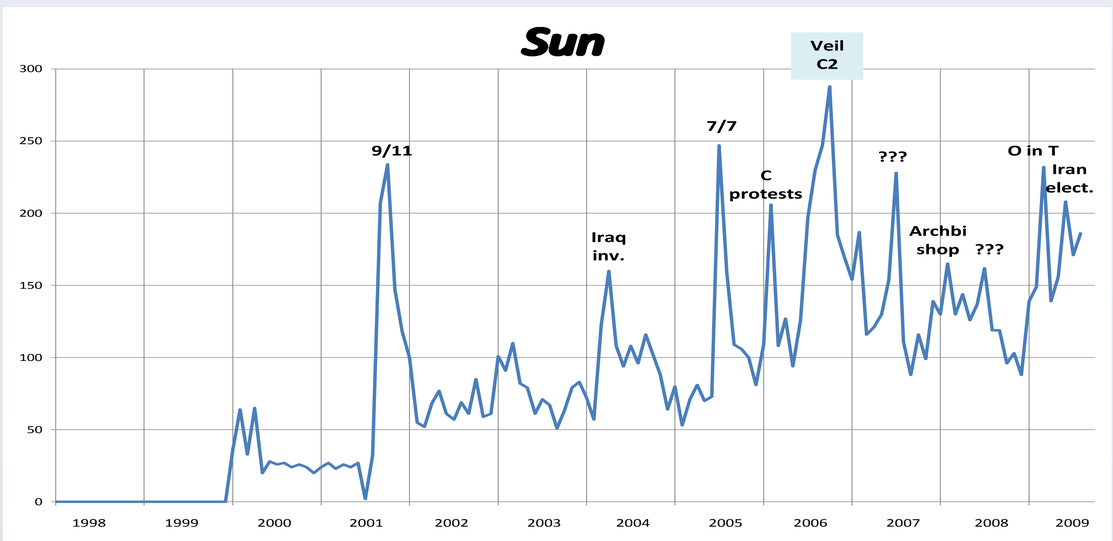

Figure 1. The Sun: Development of the raw frequency of corpus articles

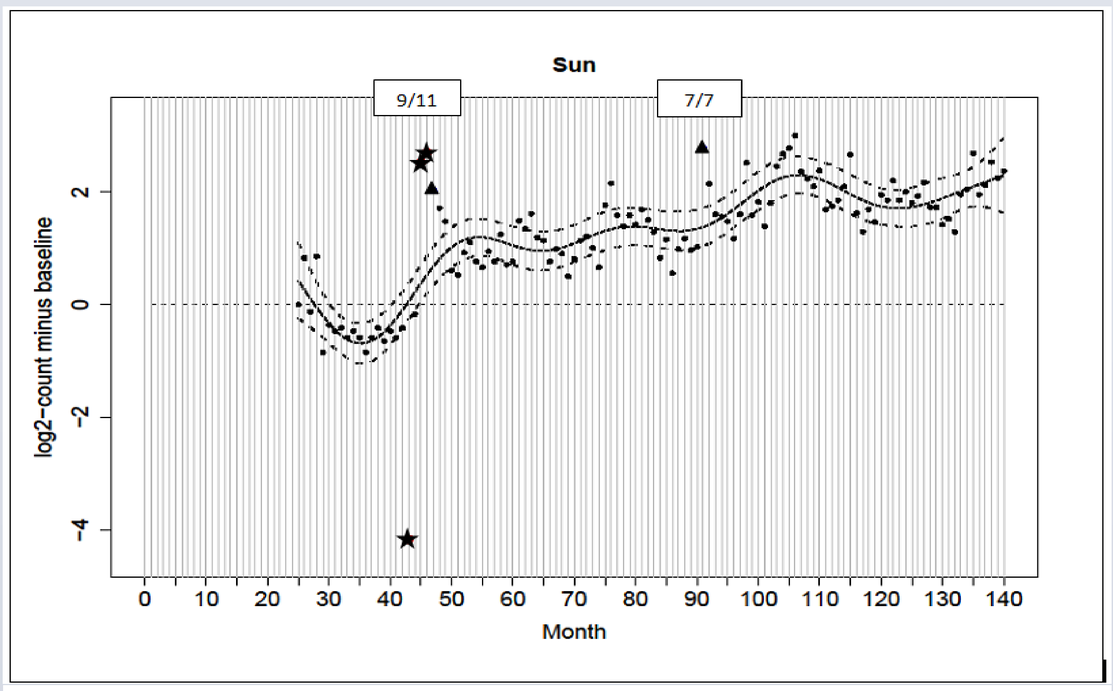

Figure 2. *The* Sun: Development of point-by-point relative differences in corpus article frequencies

Contextual Analysis

- Establishment of significant peaks of reporting → Identification of candidate trigger events → Indications of relevant contextual background.
- Trigger events are established through …
  - reading a sample of corpus articles published during the peak time period;
  - entering the corpus query in Google News, using the 'custom range' function, and examining the results for frequent news stories.

MAIN REFERENCES

1. Baker, P., Gabrielatos C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008). 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.
2. Gabrielatos, C. (2007). Selecting query terms to build a specialised corpus from a restricted-access database. *ICAME Journal* 31, 5-43.
3. 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.
4. Gabrielatos, C., McEnery, T., Diggle, P. & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics* 37(2), 151-175.

Related papers: www.gabrielatos.com/CL_methodology