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.

Prequisites

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

Definition of c-collocate

- Collocate is 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

- Proporation of consistent collocates shared by particular nodes.
- Establishment of patterns (e.g. semantic preference, discourse prosody, topics).

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

<table>
<thead>
<tr>
<th>refugees</th>
<th>asylum seekers</th>
<th>immigtrants</th>
<th>migrants</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTRY</td>
<td>ENTRY NUMBER</td>
<td>ENTRY RESIDENCE</td>
<td>ENTRY</td>
</tr>
<tr>
<td>ECON, BURDEN, RETURN</td>
<td>ECON, BURDEN, RETURN</td>
<td>ECON, BURDEN, RETURN</td>
<td>ECON, BURDEN, RETURN</td>
</tr>
<tr>
<td>asylum seekers</td>
<td>ENTRY NUMBER</td>
<td>ENTRY RESIDENCE</td>
<td>ENTRY</td>
</tr>
<tr>
<td>PLIGHT, RETURN</td>
<td>ECON, BURDEN, RETURN</td>
<td>ECON, BURDEN, RETURN</td>
<td>ECON, BURDEN, RETURN</td>
</tr>
<tr>
<td>immigtrants</td>
<td>ENTRY NUMBER</td>
<td>ENTRY RESIDENCE</td>
<td>ENTRY</td>
</tr>
<tr>
<td>PLIGHT, LEGALITY</td>
<td>PLIGHT, LEGALITY</td>
<td>PLIGHT, LEGALITY</td>
<td>PLIGHT, LEGALITY</td>
</tr>
<tr>
<td>migrants</td>
<td>ENTRY NUMBER</td>
<td>ENTRY RESIDENCE</td>
<td>ENTRY</td>
</tr>
<tr>
<td>PLIGHT</td>
<td>ECON, THREAT, LEGALITY</td>
<td>ECON, BURDEN, RETURN</td>
<td>ECON, BURDEN, RETURN</td>
</tr>
</tbody>
</table>

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)

Prequisites

- 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.

\[ QTR = \frac{\text{No. of texts retrieved by Core Query Terms AND Candidate Term}}{\text{No. of texts retrieved by Candidate Term}} \]

- 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.

\[ \text{RQTR} = \frac{(\text{QTR} - \text{B})}{(1 - \text{B})} \]

- Min. score always ≤ 100 (as QTR would be B). However, max. score varies according to B → Must be normalised (RQTRn). 5% distance from max. RQTR: RQTRn = RQTR*100 / max.RQTR.

- RQTR values:
  - 0 ≤ RQTR ≤ 100: no relevance
  - 100 ≥ RQTR: full relevance

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.

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.

Prequisites

- 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 stable 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

MAIN REFERENCES


Related papers: www.gabrielatos.co.uk/CL_methodology