Clusters of Keyness
A principled approach to selecting key items

Costas Gabrielatos
Edge Hill University
@congabonga
Focus

Quick overview of background to methodological decisions.
• Nature of keyness
• Foci of a keyness analysis
• Appropriate metrics

Principled selection of key items
• Decisions & Rationale
• Procedures & Techniques
Keyness: Nature & Foci

Distinctiveness in terms of frequency ...  
... identified via frequency comparisons.

So far, focus almost always on differences.

However, it is equally useful to look for ...  
• absence (Partington, 2014; Partington & Duguid, 2018)  
• similarity (Taylor, 2013, 2018)

Both can be approached in relation to difference  
(Gabrielatos, 2018):  
• absence: extreme case of difference  
• similarity: lack of difference
Keyness: Unit of analysis

Usually word-forms
• Hence, ‘keyword analysis’

However, focus can also be on:
• MWUs
• Lexicogrammatical units/patterns
• Semantic attributes
• Pragmatic functions

⇒ A better term: key item (Wilson, 2013)
Keyness: Appropriate Metrics

• The main metric should reflect the size of the frequency difference (Gabrielatos, 2018; Gabrielatos & Marchi, 2011, 2012; Gries, 2010; Kilgarriff, 2001).
  ⇒ Only purely effect-size metrics are appropriate.

• When the focus is difference or absence, it is useful to check if the observed difference is reliable.
  ⇒ Statistical significance testing.

• When the focus is similarity, statistical significance testing is irrelevant -- and useless.
Statistical Significance: Nature (1)

Stat.sig. tests examine the null hypothesis ($H_0$)
• In keyness analysis, $H_0$ is that there is no frequency difference.
• If $p$-value higher than threshold, $H_0$ is rejected.

Statistical significance scores are sensitive to ...
• the item frequency in the two corpora
• the sizes of the two corpora

دليل: The larger the item frequencies and/or corpora, the smaller the differences that will be significant.
In a keyness analysis, stat.sig. scores show the extent to which the compared corpora are large enough and/or the item is frequent enough for an observed sizeable frequency difference to be reliable.

⇒ It makes no sense to test for stat.sig. when the focus is similarity (i.e. when frequency differences are very small).
As corpus data cannot be expected to have normal distribution, the chi-squared test ($X^2$) is not appropriate, with the log-likelihood test ($G^2$) used instead (Rayson et al., 2004).

The $p$-value taken as the threshold for stat.sig. varies from study to study, ranging from $p=0.01$ to $p=0.000000000001$ (or even higher) (Pojanapunya, & Watson Todd, 2016).
Sensitivity of stat.sigan. values to item frequencies and corpus sizes

- The same $p$-value may represent different degrees of reliability in different comparisons (Wilson, 2013: 7-8).

**Solution**

- Using the Bayesian Information Criterion: $\text{BIC} \approx \text{LL} - \log(N)$ (Wilson, 2013: 6).
- BIC can be seen as a way to normalise $p$-values $\Rightarrow$ BIC values can be compared across different keyness analyses.
## Interpreting BIC Values

<table>
<thead>
<tr>
<th>BIC</th>
<th>Degree of evidence against $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;0$</td>
<td>No evidence – favours $H_0$</td>
</tr>
<tr>
<td>0-2</td>
<td>Not worth more than a bare mention</td>
</tr>
<tr>
<td>2-6</td>
<td>Positive evidence against $H_0$</td>
</tr>
<tr>
<td>6-10</td>
<td>Strong evidence against $H_0$</td>
</tr>
<tr>
<td>$&gt;10$</td>
<td>Very strong evidence against $H_0$</td>
</tr>
</tbody>
</table>
Selecting KIs for manual analysis

Current approaches

• Selection of top-N items.
• Setting a high item frequency threshold.
• Setting an extremely high stat.sig. threshold.

(Pojanapunya & Watson Todd, 2016: 3-10)
‘Top-N’ technique: problems

• Does not take account of the distance between the effect-size of consecutive items.

• In fact, it implicitly assumes equal distance.

• Item ‘N’ may have ...
  – a very similar effect-size to item ‘N+1’ (excluded from analysis),
  – while having a large difference with item ‘N-1’ (included in analysis).

• Blue = Implicit assumption/expectation
• Green = Usual reality
Setting a high item threshold: Problems

• May exclude important absences.
• May exclude very large frequency differences by excluding items with very low frequency in corpus 1 but very high frequency in corpus 2.

“It is a brave, or rather foolish, analyst who assumes that, in any given data set, the words are so unlikely to be key that they can be safely ignored from the very start.”

(McEnery, 2006: 148)
Setting an extremely high statistical significance threshold: Problems

- May exclude very large frequency differences simply because they don’t have extremely high statistical significance.
  - e.g. large freq.diffs between mid-freq. items

- May include small(er) frequency differences just because of extremely high statistical significance.
  - e.g. very small freq.diffs between extremely high-freq. items.
Features of proposed approach

• Primarily takes into account effect-size.
• Caters for focus on both difference (*keyness-D*) and similarity (*keyness-S*).
• When focus is on difference, stat.sig. is added as a secondary consideration.
• Avoids pre-filtering (no frequency or statistical significance thresholds).
  – Initially, all items are regarded as *candidate key items* (CKIs)
• Clusters CKIs according to their effect-size.
Clustering: approach and settings (1)

• Hierarchical cluster analysis: agglomerative method.
  – Bottom-up: initially, each CKI treated as a cluster.

• (Dis)similarity measured via Euclidian distance
  – Square root of sum of squares of pairwise differences between effect size scores.

• Distance between clusters measured via average group linkage
  – Average of distances between all effect-size scores in each cluster.
Clustering: approach and settings (2)

• Predetermined number of clusters
  – Accommodates the usual restriction in the number of KIs that can be manually examined.

\[
\text{No of clusters} = \frac{\text{Total N of CKIs}}{\text{N of CKIs to be examined}}
\]

• No frequency cut-off.
  – Caters for focus on absence.
  – Caters for large frequency differences due to very low item frequency in one of the corpora.

• Initially, no statistical significance cut-off.
  – Caters for focus on similarity.
Whatever the focus
• Clustering according to effect-size (%DIFF).
• Items in the same cluster are equally key
  ⇒ Selection of whole clusters.

Keyness-D
• Before clustering, frequency differences below set statistical significance threshold are removed.
• CKI clusters selected starting with cluster containing the largest frequency differences.

Keyness-S
• No statistical significance filtering.
• CKI clusters selected starting with cluster containing the from smallest frequency differences.
Corpora

• 2017 UK election Conservative and Labour manifestos.

• Texts from Paul Rayson’s WMATRIX webpage (Rayson, 2009): http://ucrel.lancs.ac.uk/wmatrix/ukmanifestos2017

• Further manual cleaning to (fully) remove
  – page numbers
  – chapter/section numbers
  – headers and footers
  – characters indicating bullet points (&bull;) and quotation marks (&bquo; and &equo;)

Tools & Metrics

Corpus tools:
• WordSmith 7 (Scott, 2016)
• Paul Rayson’s effect-size+stat.sig. Excel spreadsheet (Rayson, 2009)

Cluster analysis: SPSS 22

Effect-size metric: %DIFF (Gabrielatos & Marchi, 2011)

Stat.sig metrics: $G^2$ and BIC (threshold value: 2)
Methodological findings and comments

- BIC=2 corresponded to $p<0.001$
  - Much lower than the usual stat.sig. cut-offs.

- Filtering by BIC≥2 left a very small proportion (1.3%) of CKIs, but still a good number of KIs to examine manually: 31 (CM vs. LM) and 34 (LM vs. CM).
  - A good proportion of KIs index important absences: 16% (CM vs. LM) and 41% (LM vs. CM).

- Filtering by $p\leq0.01$ ($G^2\geq6.63$), returned about three times the number of CKIs: 92 (CM vs. LM) and 107 (LM vs. CM)

- Similarities are directional:
  - Different lists according to which corpus is treated as the study corpus ⇒ must be combined.

- CKIs are unequally distributed in clusters.
  ⇒ ‘Top-N’ technique is indeed problematic.
## Absences

<table>
<thead>
<tr>
<th>In <strong>CM</strong>, but not in <strong>LM</strong></th>
<th>In <strong>LM</strong>, but not in <strong>CM</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>United</td>
<td>Labour’s</td>
</tr>
<tr>
<td>Kingdom</td>
<td>equality</td>
</tr>
<tr>
<td>universities</td>
<td>unions</td>
</tr>
<tr>
<td>shall</td>
<td>LGBT</td>
</tr>
<tr>
<td>shale</td>
<td>reinstate</td>
</tr>
<tr>
<td></td>
<td>scrap</td>
</tr>
<tr>
<td></td>
<td>privatisation</td>
</tr>
<tr>
<td></td>
<td>banks</td>
</tr>
<tr>
<td></td>
<td>renters</td>
</tr>
<tr>
<td></td>
<td>women’s failure</td>
</tr>
<tr>
<td></td>
<td>enforce</td>
</tr>
<tr>
<td></td>
<td>extending</td>
</tr>
<tr>
<td></td>
<td>centres</td>
</tr>
</tbody>
</table>
Differences: \( \text{BIC} \geq 2 \)

<table>
<thead>
<tr>
<th>CKIs in CM2017</th>
<th>Smallest %DIFF: 79.14</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNITED, KINGDOM, UNIVERSITIES, SHALL, SHALE, STABLE, DATA BELIEVE, GENERATIONS, GO, ONLINE, IF, INSTITUTIONS, LEADERSHIP, TECHNICAL, OPPORTUNITY, TECHNOLOGY, DIGITAL, GREAT, STRONG, BETTER, WANT, HELP, UNION, WORLD, DO, CONTINUE, BEST, SO, CAN, WE</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CKIs in LM2017</th>
<th>Smallest %DIFF: 61.81</th>
</tr>
</thead>
<tbody>
<tr>
<td>LABOUR'S, EQUALITY, UNIONS, LGBT, REINSTATE, SCRAP, PRIVATISATION, BANKS, RENTERS, WOMEN'S, FAILURE, ENFORCE, EXTENDING, CENTRES, LABOUR, CUTS, OFFICERS, OWNERSHIP, CRISIS, GUARANTEE, REGIONAL, ARRANGEMENTS, VITAL, STAFF, RIGHTS, WOULD, WORKERS, STANDARDS, UNDER, BACK, CONSERVATIVES, JOBS, ALL, ON</td>
<td></td>
</tr>
</tbody>
</table>
The effect of changing the stat.sig. threshold

- Let’s compare the CKIs returned when BIC≥2 ($p \leq 0.001$) with those returned with the less strict statistical significance threshold of $p \leq 0.01$.
- *Do we simply get more CKIs, added after the ones derived with the stricter stat.sig. threshold?*
- CKIs returned when BIC≥2 ($p \leq 0.001$) are marked in *yellow.*
<table>
<thead>
<tr>
<th>Cluster</th>
<th>Difference: CKIs in CM2017 (p&lt;0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1:UNITED</td>
</tr>
<tr>
<td>2</td>
<td>2:KINGDOM</td>
</tr>
<tr>
<td>3</td>
<td>3:UNIVERSITIES</td>
</tr>
<tr>
<td>4</td>
<td>4:SHALL</td>
</tr>
<tr>
<td>5</td>
<td>5:SHALE</td>
</tr>
<tr>
<td>6</td>
<td>6:YOUNGER; 7:AHEAD; 8:YOUR</td>
</tr>
<tr>
<td>7</td>
<td>9:EASIER; 10:MERITOCRACY</td>
</tr>
<tr>
<td>Clusters</td>
<td>Difference: CKIs  LM2017 (p&lt;0.01)</td>
</tr>
<tr>
<td>----------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>1:LABOUR'S</td>
</tr>
<tr>
<td>2</td>
<td>2:EQUALITY</td>
</tr>
<tr>
<td>3</td>
<td>3:UNIONS</td>
</tr>
<tr>
<td>4</td>
<td>4:LGBT</td>
</tr>
<tr>
<td>5</td>
<td>5:REINSTATE</td>
</tr>
<tr>
<td>6</td>
<td>6:SCRAP</td>
</tr>
<tr>
<td>7</td>
<td>7:PRIVATISATION; 8:BANKS</td>
</tr>
</tbody>
</table>
The effect of changing the stat.sig. threshold

- *Do we simply get more CKIs, added after the ones derived with the stricter stat.sig. threshold?*

  - **NO**
    - **The ranking changes!**
    - **The clustering changes!**

**Why?**

- The lower stat.sig. threshold adds some items with larger frequency differences than some of the items returned by the stricter threshold.
Conclusions (1)

• Current KI selection techniques lack a sound rationale.
• Clustering CKIs provides a principled, transparent and replicable approach to selecting KIs for manual analysis.
• Pre-filtering can exclude large freq.diffs. and/or include smaller ones.
• Rankings by BIC and $G^2$ values do not correspond!
• Lowering the stat.sig. threshold doesn’t just return more CKIs, but also different rankings!
Conclusions (2)

• Keyness is not a straightforward attribute.

• A keyness analysis (or any quantitative analysis) does not necessarily entail objectivity.
  – Decisions regarding thresholds (frequency, effect-size, statistical significance) are subjective ...
  … and determine which and how many items are deemed ‘key’.

• A stricter stat.sig. threshold doesn’t necessarily return items that are ‘more key’.
Recommendations

• Methodological decisions must be principled and explicitly stated.
• Pre-filtering on the basis of frequency or POS should be avoided, as it is tantamount to cherry-picking.
• BIC seems a more reliable metric of statistical significance.
• For replicability, studies must report & justify:
  ...any thresholds or pre-filtering;
  ...the inclusion/exclusion of particular (types of) CKIs;
  ...the proportion of CKIs selected for analysis.
References (1)


References (2)

References (3)