Keyness Analysis
A critical overview

Costas Gabrielatos
Edge Hill University
gabrielc@edgehill.ac.uk
Definitions (1)

**Keyness Analysis**

- The comparison of normalised frequencies of linguistic items between two corpora.

**Linguistic Item**

- Word-form, word-form+POS, lemma, multi-word unit, POS, grammatical construction, semantic meaning, pragmatic meaning (see also Wilson, 2013)

**Normalised Frequency**

- Frequency per N number of words (e.g. per million words)
Frequency

- **Raw** frequency
  - The number of times an item is found in a corpus

- **Normalised** (relative) frequency
  - Per 100 words (%), 1000 words, 1 mil. words etc.

- Why normalised frequencies?
  - They enable frequency comparisons between corpora.
The need for normalised frequencies

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus 1</td>
<td>100</td>
</tr>
<tr>
<td>Corpus 2</td>
<td>100</td>
</tr>
</tbody>
</table>

Q: Does this item have the same frequency in the two corpora?

A: We can’t tell. We need to know the size of the two corpora.
The need for normalised frequencies

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Raw</th>
<th>Corpus size</th>
<th>Normalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus 1</td>
<td>100</td>
<td>1,000,000</td>
<td>100 pmw</td>
</tr>
<tr>
<td>Corpus 2</td>
<td>100</td>
<td>5,000,000</td>
<td>20 pmw</td>
</tr>
</tbody>
</table>

Q: Does this item have the same frequency in the two corpora?

A: No, it’s five times more frequent in Corpus 1

[For other normalisation options, see Gabrielatos (2013)]
Definitions (2)

**Study Corpus**
- The corpus we want to investigate.

**Reference Corpus**
- The corpus we compare the study corpus to.

⇒ SC and RC are interchangeable.
Keyword Analysis: Focus

Usually the focus is *difference*.

• However, the focus can also be ... 

**Absence** (Duguid & Partington, 2018; Partington, 2014).


Both can be approached in relation to difference (Gabrielatos, 2018):

– *absence*: extreme case of difference
– *similarity*: lack of difference

⇒ Two types of keyness: *Keyness-D, Keyness-S* (Gabrielatos, 2018)
Keyword Analysis: Utility

Frequency differences/similarities

Differences/similarities in topic / attitude / ideology
Two types of Keyness Analysis

**Exploratory**

• We compare the frequency of all items in two corpora, and then examine concordance lines of the key items to establish patterns of difference, absence, and similarity.

• This is how keyness analysis is overwhelmingly used in corpus-based discourse studies.

**Focused**

• We compare the normalised frequency of particular items in two corpora to address particular research questions or hypotheses.

(Gabrielatos, 2018; Gries, 2010: 285; Partington, 2009: 286)
Keyness: Metrics

**Effect size**

- The size of the frequency difference.
- Five main metrics: *difference coefficient, ratio, %DIFF, log-ratio, odds ratio*.

**Confidence**

- The statistical significance score.
- Metric (in CL):
  - log likelihood \((G^2 \text{ or } LL)\) (Rayson et al., 2004)
  - Bayes Information Criterion (BIC) (Wilson, 2013)

⇒ When the focus is *similarity*, statistical significance testing is irrelevant -- and useless.
Ratio = \frac{\text{NFC1}}{\text{NFC2}} \quad \text{(Kilgarriff, 2009)}

OR = \frac{\frac{\text{RFC1}}{\text{C1} - \text{RFC1}}}{\frac{\text{RFC2}}{\text{C2} - \text{RFC2}}} \quad \text{(Pojanapunya & Watson Todd, 2016: 15)}

\text{Log Ratio} = \log \frac{\text{NFC1}}{\text{NFC2}} \quad \text{(Hardie, 2014)}

\% \text{DIFF} = \frac{(\text{NFC1} - \text{NFC2}) \times 100}{\text{NFC2}} \quad \text{(Gabrielatos & Marchi, 2011)}

\text{Diff Coefficient} = \frac{\text{NFC1} - \text{NFC2}}{\text{NFC1} + \text{NFC2}} \quad \text{(Hofland & Johansson, 1982)}
Keyness-D at a glance

Sizeable difference

Key

Statistically significant difference

(Gabrielatos, 2014)
Confidence is not enough – and it may mislead
Selecting Corpora: Considerations

The comparison must be meaningful.

Any frequency differences/similarities must lend themselves to useful interpretations.
Selecting Corpora: Considerations

Ideally, the two corpora will differ in only one respect

• *Medium* (speech/writing)
• *Topic*
• *Time*
• *Language Variety*
• *Nativeness* (L1/L2)
• *Expertise/Level* (e.g. professional vs. novice writers)
• *Text Type*
• *Source* (e.g. newspaper A vs. newspaper B)
Selecting Corpora: Considerations

Possible comparisons

Variety A vs. Variety B
L1/L2 spoken vs. L1/L2 written
L1 spoken/written vs. L2 spoken/written
Expert vs. Novice/Learner
Topic-specific vs. General
Topic A vs. Topic B
Time-period 1 vs. Time-period 2
Selecting Corpora: Considerations

Make-up of the two corpora

- Keyness analysis using whole corpora makes the implicit assumption that they are homogeneous – they may not be.
- If they are not homogeneous, then keyness analysis must be carried out between sub-corpora.
  - e.g. newspaper corpus: tabloid v. broadsheet, but also left-leaning v. right-leaning.
Selecting Corpora: Addressing Misconceptions

• The reference corpus need not be a general corpus.

• The study corpus need not be smaller than the reference corpus.

• Both SC and RC can be sub-corpora – i.e. sections of the same corpus.
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

Blue
• Implicit assumption/expectation

Green
• Usual reality

(Gabrielatos, 2017)
‘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).
Setting a high item frequency 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 frequency differences between mid-freq. items

- May include small(er) frequency differences just because of extremely high statistical significance.
  - e.g. very small frequency differences between extremely high frequency items.
Selecting KIs: Proposed approach

Cluster Analysis

• Candidate KIs are grouped, according their frequency difference, in a predetermined number of clusters.

• KIs are selected in whole clusters, starting from the top one.
  – If an item in a cluster is selected for analysis, all other items in that cluster must also be selected.
  – Avoids cherry-picking.

(For details, see Gabrielatos, 2018)
Features of proposed approach

• Primarily takes into account effect-size.
  – Clustering according to 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)
But this is not the end

Identifying key items is not enough.
• Simply looking at a list of linguistic items does not help us identify aspects of content (e.g. topics, attitudes) ...
• ... although it may help us formulate initial hypotheses.

We need to examine instances of each key item in enough co-text (a.k.a concordance analysis)
• Annotation of each instance – keeping in mind the relevant context.
• Quantification of results.
Corpora

• 2017 UK election Conservative and Labour manifestos.

• Texts from Paul Rayson’s Wmatrix webpage (Rayson, 2009): [http://ucrel.lancs.ac.uk/wmatrix/ukmanifestos2017](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 and quotation marks
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)
## 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>
by Britain, we believe that the future of our country is a bright one. We are already the fifth largest economy in the world, the biggest recipient of foreign investment in Europe and the fastest growing economy in the G7. We have three of the world’s top ten universities and, despite forming less than one percent of the global population, we boast more Nobel Laureates than any country other than America. We have the finest intelligence services and hugely respected armed forces that can project power around the number of scientists working in the UK and enable leading scientists from around the world to work here. We will work hard to ensure we have a regulatory environment that encourages innovation. University investment funds Our world-beating universities will lead the expansion of our R&D capacity. We must help them make a success of their discoveries - while they have a number of growing investment funds specialising in spin-outs, we have more to do to replicate the success of similar university of their discoveries - while they have a number of growing investment funds specialising in spin-outs, we have more to do to replicate the success of similar university funds in the United States. To fix that, we will work to build up the investment funds of our universities across the UK. We want larger, aggregated funds to increase significantly the amounts invested in and by universities. We want universities to enjoy the commercial fruits of their research, through funds that are large enough to list, thereby giving their success of similar university funds in the United States. To fix that, we will work to build up the investment funds of our universities across the UK. We want larger, aggregated funds to increase significantly the amounts invested in and by universities. We want universities to enjoy the commercial fruits of their research, through funds that are large enough to list, thereby giving British investors a chance to share in their success. National Productivity Investment Fund If our modern industrial to do so in a way that encourages the development of new clusters of public services, private businesses and, where appropriate, universities. We will proceed on the same basis with the arts and cultural organisations that give our United Kingdom such strength. Channel 4 will remain publicly owned and will be relocated out of London, and we will work with the nation’s most eminent mobility - which is why we are looking at ways to make sure civil service recruitment is as diverse as possible, not only from the perspective of gender and race but social class too. We will continue to fund schemes to get graduates from Britain’s leading universities to serve in schools, police forces, prisons, and social care and mental health organisations. These programmes are now some of the UK’s largest graduate employers, taking the brightest and best from our universities and using their talents to go further. We will provide seed funding for similar schemes to recruit older professionals from other sectors, including those returning to the workplace having cared for that tackle drug and alcohol abuse. We will introduce dedicated provision for women offenders. THE WORLD’S GREAT MERITOCRACY Theresa May’s Conservatives will deliver more good school places, ending the ban on selective schools and asking universities and independent schools to help run state schools. World class technical education, underpinned by prestigious new institutes of technology with the freedoms that make our universities great. A government unafraid to confront the burning injustic
scrap in LM

rates - including switching from RPI to CPI indexation. Scrap new investment in plant and machinery from valuations, and ensuring that businesses have access to a proper appeals process - while reviewing the entire business rates system in the longer run. Scrap quarterly reporting for businesses with a turnover of under £85,000. Declare war on late payments by: Using government procurement to ensure that anyone bidding for a government contract pays its own suppliers within 30 days. Developing a version of the
standards, provide certainty to EU nationals and give a meaningful role to Parliament throughout negotiations. We will end Theresa May's reckless approach to Brexit, and seek to unite the country around a Brexit deal that works for every community in Britain. We will scrap the Conservatives' Brexit White Paper and replace it with fresh negotiating priorities that have a strong emphasis on retaining the benefits of the Single Market and the Customs Union - which are essential for maintaining industries, jobs and businesses in Britain.

on short hours contracts (some only guaranteed a few hours per week), but who regularly work far more. We will strengthen the law so that those who work regular hours for more than 12 weeks will have a right to a regular contract, reflecting those hours. We will also scrap the changes brought in by the Conservatives in 2014 to TUPE, which weakened the protections for workers transferring between contractors, and we will abolish the loophole to the agency workers regulations known as the Swedish derogation. We will consult with

have slashed social security over the last seven years, leaving more people in poverty, subject to a punitive sanctions regime, and reliant on food banks. Labour will act immediately to end the worst excesses of the Conservative government's changes. We will scrap the punitive sanctions regime, scrap the Bedroom Tax, restate Housing Benefit for under-21s, scrap cuts to Bereavement Support Payment. The cuts to work allowances in Universal Credit (UC), and the decision to limit tax credit and UC payments to the first two

over the last seven years, leaving more people in poverty, subject to a punitive sanctions regime, and reliant on food banks. Labour will act immediately to end the worst excesses of the Conservative government's changes. We will: scrap the punitive sanctions regime Scrap the Bedroom Tax, restate Housing Benefit for under-21s, scrap cuts to Bereavement Support Payment. The cuts to work allowances in Universal Credit (UC), and the decision to limit tax credit and UC payments to the first two children in a family, are an

to a punitive sanctions regime, and reliant on food banks. Labour will act immediately to end the worst excesses of the Conservative government's changes. We will: Scrap the punitive sanctions regime Scrap the Bedroom Tax, restate Housing Benefit for under-21s, scrap cuts to Bereavement Support Payment. The cuts to work allowances in Universal Credit (UC), and the decision to limit tax credit and UC payments to the first two children in a family, are an attack on low-income families and will increase child poverty. Labour will

for work element. Increase Carer's Allowance by £11 to the level of Jobseekers' Allowance. Implement the court decision on Personal Independence Payment (PIP) so that there is real parity of esteem between those with physical and mental-health conditions. Scrap the Work Capability and Personal Independence Payment assessments and replace them with a personalised, holistic assessment process that provides each individual with a tailored plan, building on their strengths and addressing barriers. Labour will end

the biggest council building programme for at least 30 years...we will ditch the Conservatives' ban on long-term council tenancies to give council tenants security in their homes. We want more people to have a secure tenure in a home built to high standards. Labour will scrap the punitive bedroom tax, which has caused many people to be evicted from their home and their community. More council homes have been sold off under the Conservatives and only one in five have been replaced, despite long housing waiting lists. Labour will

to guarantee the best possible services for patients, Labour will invest in our health and care workforce. A Labour government will step in with a long-term workforce plan for our health service that gives staff the support they need to do the best for their patients. Labour will scrap the NHS pay cap, put pay decisions back into the hands of the independent pay review body and give our NHS workers the pay they deserve. Labour will protect patients and legislate to ensure safe staffing levels in the NHS. Labour's long-term ambition is for our

Strategy. We will continue to upgrade our highways and improve roadworks at known bottlenecks. The A1 North, the Severn Bridge and the A30 provide essential connections and require our urgent consideration. We will work with the Welsh Government to scrap the tolls on the Severn Bridge. After seven years of stalled progress, Labour will reset the UK's road safety vision and ambitiously strive for a transport network with zero deaths, reintroducing road-safety targets, setting out bold measures that will continuously improve
Largest Differences

**CKIs in CM2017**
Smallest %DIFF: 79.14

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

**CKIs in LM2017**
Smallest %DIFF: 61.81

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
Conclusions

• Keyness refers to the size of the frequency difference – not the level of confidence.
• Keyness can refer to difference or similarity.
• Keyness analysis can be exploratory or focused.
• Corpus selection should serve the study aims.
• A keyness analysis (or any quantitative analysis) does not necessarily entail objectivity.
  – Decisions regarding thresholds are largely subjective -- and determine which and how many items are deemed ‘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.
• For replicability, studies must report and justify:
  ...any thresholds or pre-filtering;
  ...the inclusion/exclusion of particular (types of) CKIs;
  ...the proportion of CKIs selected for analysis.
Keyness Analysis: Steps

1. Formulate hypotheses or research questions.
2. Select or compile appropriate corpora.
3. Carry out frequency comparisons to identify key items.
4. Carry out concordance analysis (i.e. explicit annotation).
5. Quantify the results.
6. Interpret results.
References (1)


References (2)

- Gabrielatos, C. & Marchi, A. (2012) Keywords: appropriate metrics and practical issues. *CADS International Conference*, Bologna, Italy, 13-15 September 2012. [https://repository.edgehill.ac.uk/4196](https://repository.edgehill.ac.uk/4196).
References (3)


References (4)


