Keyness

Appropriate metrics and practical issues

Costas Gabrielatos & Anna Marchi
(Edge Hill University, Lancaster University)

University of Bologna, 14 September 2012
Abstract

In this paper we examine the definitions of two widely-used interrelated constructs in corpus linguistics, *keyness* and *keywords*, as presented in the literature and corpus software manuals. In particular, we focus on

• the consistency of definitions given in different sources;
• the metrics used to calculate the level of keyness;
• the compatibility between definitions and metrics.

Our survey of studies employing keyword analysis has indicated that the vast majority of studies examine a subset of *keywords* – almost always the top X number of keywords as ranked by the metric used. This renders the issue of the appropriate metric central to any study using keyword analysis.

In this study, we first argue that an appropriate, and therefore useful, metric for *keyness* needs to be fully consistent with the definition of *keyword*. We then use four sets of comparisons between corpora of different types and sizes, in order to test whether and to what extent the use of different metrics affects the ranking of keywords. More precisely, we look at the extent of overlap in the keyword rankings resulting from the adoption of different metrics, and we discuss the implications of ranking-based analysis adopting one metric or another. Finally, we propose a new metric for *keyness*, and demonstrate a simple way to calculate the metric, which supplements the keyword extraction in existing corpus software.
Motivation

• Keyword analysis is one of the most widely used techniques in corpus studies.
• The vast majority of studies do not examine all keywords, but the top X (usually the top 100).
  → The ranking criterion becomes very important.
  → Usually the criterion is keyness.

• Manual examination of frequency differences of particular sets of words has shown large discrepancies between ranking by frequency difference (%DIFF) and ranking by statistical significance (LL) (Gabrielatos 2007; Gabrielatos & McEnery, 2005)
• The frequency ratio has been proposed as another metric for keyness (e.g. Kilgariff, 2001, 2012)
Questions and Aims

• What is a keyword?
• What is keyness?
• How is it measured?

→ Examination of definitions of the terms keyword and keyness.
→ Comparison of metrics.
Definitions: *Keywords*

• “Key words are those whose frequency is unusually high in comparison with some norm” (Scott, 1996: 53).

• “A key word may be defined as a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind” (Scott, 1997: 236).

Keywords are defined in relation to **frequency difference**.

The metric of **keyness** would be expected to represent the **extent of the frequency difference**.

*However ...*
Definitions: Keyness

• “The keyness of a keyword represents the value of log-likelihood or Chi-square statistics; in other words it provides an indicator of a keyword’s importance as a content descriptor for the appeal. The significance (p value) represents the probability that this keyness is accidental” (Biber et al., 2007: 138).

• “A word is said to be "key" if [...] its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p value specified by the user” (Scott, 2011).
Keyword vs. Keyness: Contradictions

• “Key words are those whose frequency is unusually high in comparison with some norm” (Scott, 2011: 165).

• “A word is said to be "key" if [...] its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p value specified by the user” (Scott, 2011: 174).

→ The current literature/practice treats the statistical significance of a frequency difference as a metric for that difference.

→ Is this appropriate? Is this good practice?

→ Some help from statistics
Effect size vs. Statistical significance

**What do they measure?**

- Effect size “indicates the magnitude of an observed finding” (Rosenfeld & Penrod, 2011: 342).
- Effect size “is a measure of the practical significance of a result, preventing us claiming a statistical significant result that has little consequence” (Ridge & Kudenko, 2010: 272).
- “Just because a particular test is statistically significant does not mean that the effect it measures is meaningful or important” (Andrew et al., 2011: 60).
- “A very significant result may just mean that you have a large sample. [...] The effect size will be able to tell us whether the difference or relationship we have found is strong or weak.” (Mujis, 2010: 70).

Frequency difference and statistical significance are not the same
Effect size vs. Statistical significance

The influence of corpus size

“Tests of statistical significance are dependent on the sample size used to calculate them. [...] With very large sample sizes, even very weak relationships can be significant. Conversely, with very small sample sizes, there may not be a significant relationship between the variables even when the actual relationship between the variables in the population is quite strong. Therefore, different conclusions may be drawn in different studies because of the size of the samples, if conclusions were drawn based only on statistical significance testing. Unlike tests of significance, effect size estimates are not dependent on sample size. Therefore, another advantage of using effect size estimates is that they provide information that permits comparisons of these relationships across studies” (Rosenfeld & Penrod, 2011: 84).
Keyness: 
*Effect size or statistical significance?*

- **Effect size:** The % difference of the frequency of a word in the study corpus when compared to that in the reference corpus.

- **Statistical significance:** The $p$ value of the frequency difference, as measured by a statistical test – usually log-likelihood or Chi-square.

$\rightarrow$ *Does the choice of metric make a difference …*

... *when all the KWs are examined?*

... *when only the top X keywords are examined?*
Methodology

• Comparisons between corpora of different types and/or unequal sizes.

• Examination of the proportion of overlap between the ranking derived through the two metrics when examining ...
  ... all KWs
  ... the top 100 KWs

• The extent of overlap will indicate how similar or different the two metrics are.
  – High overlap $\rightarrow$ the two metrics are almost identical.
  – Low overlap $\rightarrow$ one metric is inappropriate.

• In all comparisons, the cut-off point for statistical significance is $p<0.01$ (LL=6.63).
% DIFF: Calculation

\[
\text{NF in SC} - \text{NF in RC} \times 100
\]

NF in RC

NF = normalised frequency
SC = study corpus
RC = reference corpus
LL and %DIFF ranking: Visualisation of full overlap
Data

Comparison 1: large corpus vs. large corpus

• Corpora of three British broadsheets in 1993 and 2005
  • SiBol 1993 (96 mil. words) vs. SiBol 2005 (156 mil. words)

Comparison 2: small corpus vs. medium-sized corpus

• Corpora of individual sections from the Guardian in 2005
  • Media section (1 mil. words) vs. Home news (6 mil. words)
SiBol 1993 (96 mil.) vs. SiBol 2005 (156 mil.)
(4356 KWs)
Guardian 2005: Media (1 mil.) vs. Home (6 mil.) (317 KWs)
Absolute and relative size of compared corpora doesn’t seem to make a difference.

But all four corpora compared so far contain newspaper articles.

What about comparisons between specialised and general corpora?
Comparison 3: specialised vs. large general corpus
• Hutton Inquiry corpus (1 mil. words) vs. BNC (100 mil. words)

Comparison 4: specialised vs. small general corpus
• Hutton (1 mil. words) vs. FLOB (1 mil. words)

Comparison 5: small general corpus vs. large general corpus
• FLOB (1 mil. words) vs. BNC (100 mil. Words)
Hutton (1 mil.) vs. BNC (100 mil.)
(11853 KWs)
Hutton (1 mil). vs. FLOB (1 mil.)
(10631 KWs)
FLOB (1 mil.) vs. BNC (100 mil.)
(6971 KWs)
However, this very low overlap may be misleading: differences in the ranking of KWs may be very small.

e.g.
A word may be at position 25 in one ranking and 27 in the other.

Examination of overlap in the top 100 KWs by LL and %DIFF.
## Top 100: Overlap of ranking by LL and %DIFF

<table>
<thead>
<tr>
<th>Compared corpora</th>
<th>Shared top-100 KWs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiBol 1993 vs. SiBol 2005</td>
<td>3</td>
</tr>
<tr>
<td>Guardian 2005: Media vs. Home</td>
<td>0</td>
</tr>
<tr>
<td>Hutton vs. BNC</td>
<td>2</td>
</tr>
<tr>
<td>Hutton vs. FLOB</td>
<td>8</td>
</tr>
<tr>
<td>FLOB vs. BNC</td>
<td>22</td>
</tr>
</tbody>
</table>
The same KW may have very high LL but very low %DIFF

- **THE:**  LL = 32,366.01 (2\textsuperscript{nd})  \textbf{but}  \% DIFF = 9.7% (4302\textsuperscript{nd})
- **OF:**  LL = 20,935.02 (5\textsuperscript{th})  \textbf{but}  \% DIFF = 11.8% (4304\textsuperscript{th})

What the high LL values indicate here is that we can be highly confident that there is a very small frequency difference.

The same KW may have \textbf{very high %DIFF} but (relatively) low LL

- **ADVENTISTS:**  \%DIFF = 2086.3% (1020\textsuperscript{th})  \textbf{but}  LL= 137.49 (4019\textsuperscript{th})
- **EX-COMMUNIST:**  \%DIFF = 679.1% (1584\textsuperscript{th})  \textbf{but}  LL= 136.61 (4048\textsuperscript{th})

The LL value does not accurately reflect the size of a difference.
Different KWs may have markedly different LL but similar %DIFF
- LL: DELORS (100th, LL=3192.68), PAPANDREOU (761st, LL=677.85)
- %DIFF: DELORS (676th, 5386%), PAPANDREOU (677th, 5340.5%)

Using the LL values as a measure of keyness would exclude a true keyword from the analysis.

Different KWs may have similar LL but very different %DIFF
- LL: SERB (33rd, LL=6,966.10), BRITISH (34th, LL=6,732.14)
- %DIFF: SERB (1167th, 1496.5%), BRITISH (3670th, 46.6%)

Using the LL values as a measure of keyness would result in treating a low-level keyword as a high-level one.
Conclusions (1)

• High LL does not necessarily correlate with high %DIFF.

• LL and %DIFF result in different rankings.

Why?
Conclusions (2)

• The metric of **keyness** needs to measure **effect size** (i.e. frequency difference) – **not** statistical significance.

• LL measures **statistical significance**, **not** frequency difference.

• LL is **sensitive** to word frequencies and **corpus sizes**

LL is not an appropriate metric for **keyness**
Proposition

- %DIFF is fully consistent with the definition of *keyword*.
- %DIFF measures *effect size*.
- %DIFF reveals not only *differences* but also *similarities* (e.g. Taylor, 2011).

We propose %DIFF as an appropriate metric for keyness.

Only statistically significant %DIFF should be considered.
Practical issues

- Can/Should there be a threshold for %DIFF?
  (Stat. Sig. has a widely accepted threshold in CL: $p \leq 0.01$)

- How do we get, and sort by, %DIFF values?
  (Current tools don’t accommodate that)

- How do we handle zero occurrences in the RC?
  (We can’t divide by zero)
A threshold for %DIFF?

• Reminder 1: the current threshold for statistical significance \((p \leq 0.01, \text{ LL} \geq 6.63)\) is arbitrary.

• Threshold has to be relative to the resulting range of %DIFF values.
  – E.g., a 50% DIFF is relatively ...
  – ... small, if most values are larger than 200%.
  – ... large, if most values are smaller than 20%

• Reminder 2: if you focus on the top X, make sure that you include all KWs with the same %DIFF as the Xth one.
How to prepare WordSmith KW output for Excel

1. Wordsmith: change visualization settings
   
   `view > layout > RC% > decimals`
   
   → Increase number of decimal points until non-zero digits show

2. Copy list and paste it on an Excel file
How do we handle zero frequencies in the RC?

In the ‘relative frequencies’ column of the RC, substitute all zero frequencies with an extremely small number (e.g. $0.000000000000000001 = \text{quadrillion}$)

Why?

→ This small number ...

... is a very good approximation of zero for calculation purposes ...

... while allowing for divisions by it.
How to create a column for %DIFF in Excel

1. Add a column with header %DIFF.
2. In the cell below the header, write this ‘function’:
   \[=(X2 - Y2) / Y2 \times 100\]

   \(X\) = column with normalised frequencies in study corpus
   \(Y\) = column with normalised frequencies in reference corpus

   • Why row 2 (X2, Y2)?
     → Usually the first row is reserved for the column header.
References (1)


References (2)