Measuring Keyness

Stephanie Evert
Computational Corpus Linguistics Group
FAU Erlangen-Nürnberg, Germany

18 April 2022

1 Keywords in corpus linguistics and DH

In corpus linguistics, the notion of **keywords** refers to words (and sometimes also multiword units, semantic categories or lexico-grammatical constructions) that "occur with unusual frequency in a given text" (Scott 1997: 236) or a text collection, i.e. a corpus. Keywords are deemed to represent the characteristic vocabulary of the target text or corpus and thus have many applications in corpus linguistics, digital humanities and computational social science. They can capture the aboutness of a text (Scott 1997), the terminology of a text genre or technical domain (Paquot and Bestgen 2009), important aspects of literary style (Culpeper 2009), linguistic and cultural differences (Oakes and Farrow 2006), etc.; they give insight into historical perspectives (Fidler and Cvrček 2015) and provide a basis for measuring the similarity of text collections (Rayson and Garside 2000). Keywords are also an important starting point for corpus-based discourse analysis (Baker 2006), where manually formed clusters of keywords represent central topics, actors, metaphors, and framings (e.g. McEnery et al. 2015). Since this process is guided from the outset by human understanding, it provides a more interpretable alternative to topic models in hermeneutic text analysis.

Keywords are usually operationalised in terms of a statistical frequency comparison between the **target corpus** and a **reference corpus**. Different research questions can be addressed depending on the particular constellation of target T and reference R, e.g. (i) T = a single text vs. R = a text collection ($\rightarrow$ aboutness), (ii) T and R = collections of articles on the same topic in left-leaning and right-leaning newspapers ($\rightarrow$ contrastive framings), or (iii) T = texts from a given domain or genre vs. R = a large general-language reference corpus ($\rightarrow$ terminology).

Although keyword analysis is a well-established approach and has been implemented in many standard corpus-linguistic software tools such as WordSmith[1], AntConc[2], SketchEngine[3], and CQPweb (Hardie 2012), it is still unclear what the "right" way of measuring keyness is (see overview in Hardie 2014). In this paper, I propose (i) a mathematically well-founded **best-practice technique** and (ii) introduce a **visual approach** for exploring the empirical properties of different keyness measures.

2 Keyness measures

Keyword analysis is operationalised as a comparison of relative frequencies: For each **candidate** word, its frequency f_1 in a target corpus T of n_1 tokens is compared to its frequency f_2 in a reference corpus R of n_2 tokens. The candidate set of m items typically includes words that only occur in the target corpus ($f_2 = 0$).

[1] `https://www.lexically.net/wordsmith/`
[2] `https://www.laurenceanthony.net/software/antconc/`
[3] `https://www.sketchengine.eu/`

A candidate is considered a ("positive") keyword if its relative frequency $p_1 = f_1/n_1$ in T is substantially higher than its relative frequency $p_2 = f_2/n_2$ in R. A large number of **keyness measures** have been proposed to quantify the comparison and thus provide a basis for a ranking of the candidates and/or cut-off thresholds. Three main groups of measures can be distinguished:

1. Measures based on **hypothesis tests** put the focus on establishing a statistically significant difference between p_1 and p_2. The most widely-used measures are chi-squared X^2 and log-likelihood G^2 (Dunning 1993). These measures are biased towards high-frequency keywords, often including function words and other non-specific words.
2. **Effect size** measures instead focus on how many times more frequent a candidate is in T than in R. The most intuitive measure is relative risk $r = p_1/p_2$, also known as LogRatio $= \log_2 r$ (Hardie 2014). Some other effect-size measures are equivalent (%DIFF, Gabrielatos and Marchi 2012) or closely related (odds ratio, Pojanapunya and Watson Todd 2018) to LogRatio. These measures are biased towards very low-frequency keywords and are often combined with an additional significance filter (typically based on G^2).
3. Various **heuristic** measures lack any statistical foundation. They are often particularly easy to compute such as SketchEngine's SimpleMaths (Kilgarriff 2009), which also offers a user parameter to adjust its bias towards high-frequency or low-frequency keywords.

3 Mathematical discussion and visualisation

Hypothesis-test measures are subject to the criticism raised more generally against p-value testing in corpus linguistics and other fields (e.g. Gries 2005). In particular, they are biased towards high-frequency keywords irrespective of effect size, selecting candidates that are not very salient for the target corpus. When they are applied more reasonably as a significance filter, the problem of multiple testing is often ignored: a single analysis may carry out frequency comparisons for hundreds of thousands of candidates, resulting in large numbers of false positives at customary significance levels such as $p < .001$ (Gries 2005; Hardie 2014).

By contrast, effect-size measures such as LogRatio are biased towards low-frequency keywords because they completely ignore the statistical significance of the observed difference in relative frequency. Moreover, many of these measures are undefined for $f_2 = 0$ and need special heuristics for this case; e.g. Hardie (2014) simply substitutes $f_2 = \frac{1}{2}$ without mathematical justification.

Traditionally, keyness measures are computed from cumulative token frequency counts for T and R. However, two recent studies have independently concluded that keywords based on document counts are more robust (Evert et al. 2018; Egbert and Biber 2019).

Keyness measures can also be understood from a more intuitive angle by visualising them as **topographic maps**, which show the scores assigned to all possible combinations of frequencies f_1 in T and f_2 in R on a logarithmic scale (similar to the visualisation of collocations in Evert 2004: Sec. 3.3). The examples in Fig. 1 show the respective frequency biases of G^2 and LogRatio – which is hardly mitigated by an additional significance filter – in the top row (dark red colours indicate frequency profiles of highly-ranked keywords).

4 Best-practice recommendation

Conservative estimates based on statistical confidence intervals combine the advantages of hypothesis tests and effect-size measures into a single score. I therefore propose **LRC**, a conservative estimate of LogRatio, as a best-practice keyness measure. LRC uses an exact conditional Poisson test (Fay 2010: 55) to obtain reliable confidence intervals corrected for multiple testing. The full procedure for computing LRC scores is as follows:

1. Collect the frequency data f_1, f_2 for each candidate and the sample sizes n_1, n_2 of T and R. Wherever suitable, document frequencies should be preferred.

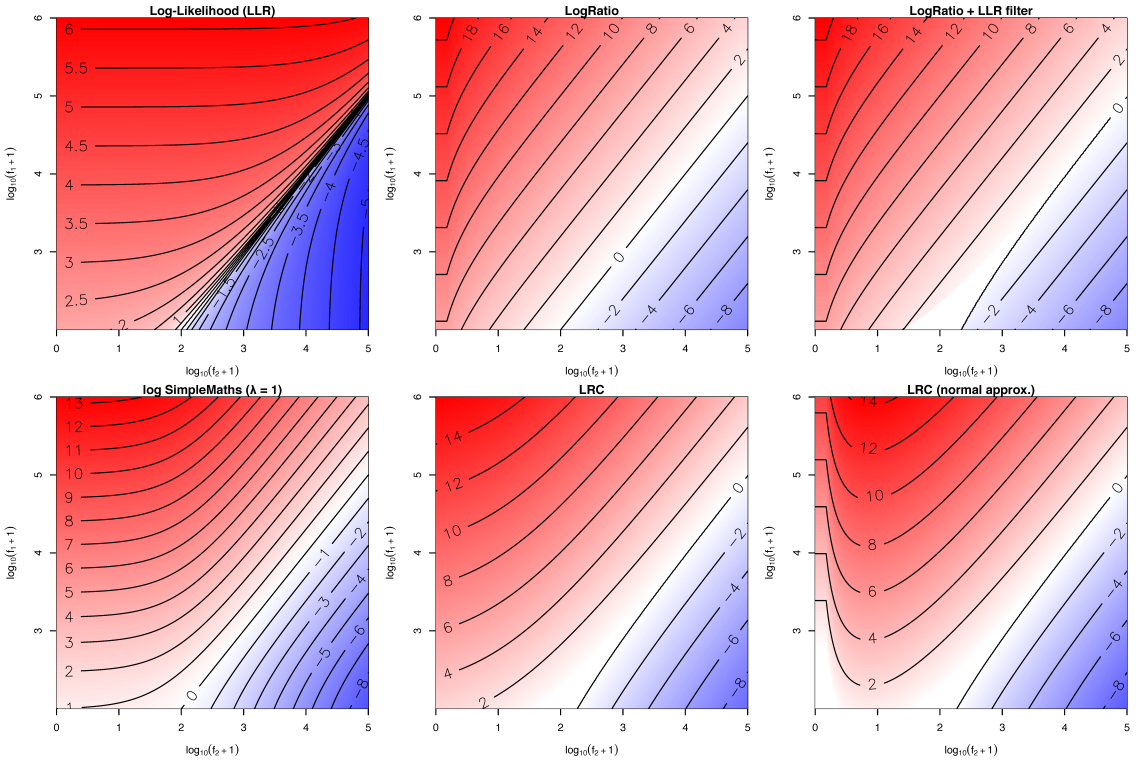

Figure 1: Visualisation of keyness measures as topographic maps for $n_1 = n_2 =$ 100M words. The bottom right panel highlights problems of an earlier version of LRC currently used by CQPweb.

2. Compute a two-sided Pearson-Clopper binomial confidence interval $[\pi_-, \pi_+]$ for f_1 successes out of $f_1 + f_2$ trials, with Bonferroni-adjusted significance level $\alpha = 0.05/m$.
3. Convert the binomial proportions to $[\text{LRC}_-, \text{LRC}_+] = [\log_2(\frac{n_2}{n_1} \cdot \frac{\pi_-}{1-\pi_-}), \log_2(\frac{n_2}{n_1} \cdot \frac{\pi_+}{1-\pi_+})]$.
4. If the test is not significant ($\text{LRC}_- \leq 0 \leq \text{LRC}_+$), set LRC $= 0$. Otherwise, set LRC $= \text{LRC}_-$ if $p_1 > p_2$ and LRC $= \text{LRC}_+$ if $p_1 < p_2$.

LRC has several **advantages** over other keyness measures: (i) it balances out the high-frequency bias of hypothesis tests and the low-frequency bias of effect-size measures (cf. right panel of Fig. 2); (ii) unlike heuristics such as SimpleMaths it does this in a mathematically well-justified way; (iii) it can be applied to candidates with $f_2 = 0$ without special precautions; (iv) it detects both positive ($p_1 \gg p_2$) and negative ($p_1 \ll p_2$) keywords; (v) it includes a reliable significance filter (LRC $= 0$) and does not require arbitrary frequency thresholds; (vi) robust and efficient implementations of the underlying binomial confidence intervals are available in standard statistical software packages, so very large candidate sets can easily be processed. The left panel of Fig. 2 shows that LRC overlaps well with established keyness measures, again indicating that it provides an excellent compromise.

A reference implementation of LRC is available at `https://osf.io/cy6mw/` together with a more detailed analysis. It is also included in version 0.6 of the `corpora` package for R.[4]

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum Books, London.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics*, 14(1):29–59.

[4] `https://cran.r-project.org/web/packages/corpora/`

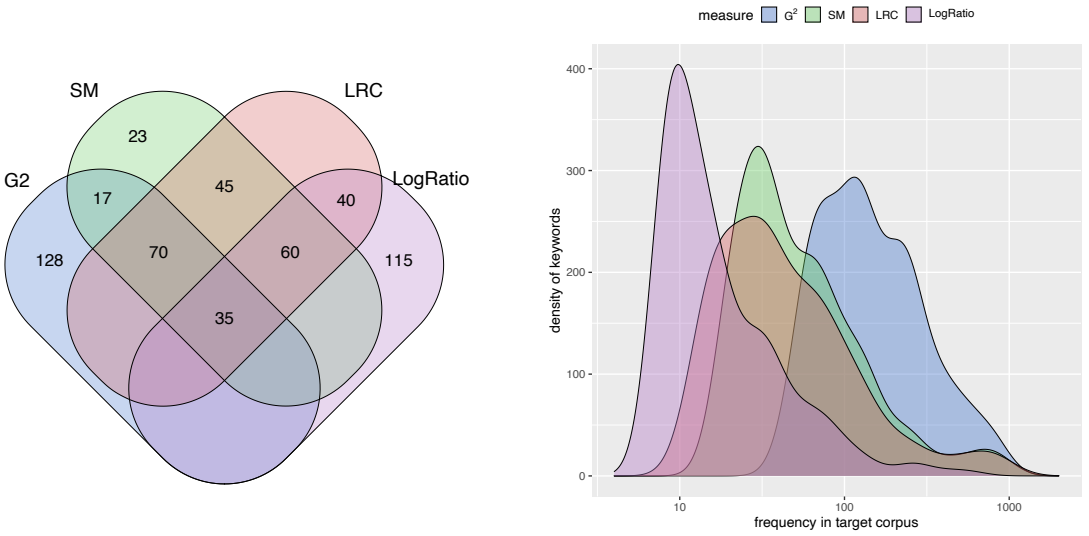

Figure 2: Quantitative analysis of top-250 keyword lists for the data of Evert et al. (2018): overlap between four measures (left panel) and frequency distribution in the target corpus (right panel).

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Egbert, J. and Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from `http://www.collocations.de/phd.html`.

Evert, S., Dykes, N., and Peters, J. (2018). A quantitative evaluation of keyword measures for corpus-based discourse analysis. Presentation at the Corpora & Discourse International Conference (CAD 2018), Lancaster, UK.

Fay, M. P. (2010). Two-sided exact tests and matching confidence intervals for discrete data. *The R Journal*, 2(1):53–58.

Fidler, M. and Cvrček, V. (2015). A data-driven analysis of reader viewpoints: reconstructing the historical reader using keyword analysis. *Journal of Slavic Linguistics*, 23(3):197–239.

Gabrielatos, C. and Marchi, A. (2012). Keyness: Appropriate metrics and practical issues. Presentation at the Corpora and Discourse Studies Conference (CADS 2012), Bologna, Italy. Available from `https://www.researchgate.net/publication/261708842_Keyness_Appropriate_metrics_and_practical_issues`.

Gries, S. T. (2005). Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory*, 1(2):277–294.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Hardie, A. (2014). A single statistical technique for keywords, lockwords, and collocations. Internal CASS working paper no. 1, unpublished.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics 2009 Conference*, Liverpool, UK.

McEnery, T., McGlashan, M., and Love, R. (2015). Press and social media reaction to ideologically inspired murder: the case of Lee Rigby. *Discourse and Communication*, 9(2):1–23.

Oakes, M. P. and Farrow, M. (2006). Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing*, 22(1):85–99.

Paquot, M. and Bestgen, Y. (2009). Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction. In Jucker, A., Schreier, D., and Hundt, M., editors, *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora*, pages 247–269. Rodopi, Amsterdam.

Pojanapunya, P. and Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1):133–167.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the ACL Workshop on Comparing Corpora*, pages 1–6, Hong Kong.

Scott, M. (1997). PC analysis of key words – and key key words. *System*, 25(2):233–245.