A quantitative evaluation of keyword measures for corpus-based discourse analysis

Stefan Evert, Natalie Dykes, Joachim Peters
University of Erlangen-Nuremberg, Germany
stefan.evert@fau.de

Keywords: keyness, association measures, quantitative evaluation

We evaluated the automatic identification of keywords in our target corpus of German press texts on multi-resistant pathogens; a recurring and highly controversial topic in public discourse (1.3M tokens). In corpus-based discourse analysis, keywords have been described as "a quick and simple 'way in'" to corpus comparison (Baker et al. 2013), while the details on how to best 1) calculate and 2) categorise them lacks consensus. There have been multiple attempts to answer question 1) (cf. Kilgarriff 2001, Paquot & Bestgen 2009, Lijffijt et al. 2016). However, these studies focus mostly on either the number of keywords or the mathematical adequacy, but do not evaluate against an authentic use-case specifically tailored to discourse analysis. Our approach draws on a set of previously determined qualitative linguistic categories and evaluates statistically generated keyword lists against them. The keywords were calculated using three measures:

- log-likelihood (G^2) (Dunning 1993)
- log ratio (LR), an intuitive measure implemented in the CQPweb corpus analysis software (Hardie 2012)
- LR_{conf}, a conservative estimate for the log ratio coefficient using the lower bound of a 99% confidence interval with Bonferroni correction

and two different German national broadsheet newspapers as reference corpora (RC):

- years 2011–2014 of *Süddeutsche Zeitung* (SZ), a left-leaning daily newspaper (290M tokens)
- years 2011–2014 of *Frankfurter Allgemeine Zeitung* (FAZ), a right-leaning daily newspaper (150M tokens)

All corpora were POS-tagged and lemmatized using the state-of-the-art morphological analyzer SMOR (Schmid et al. 2004).

Applying a frequency threshold of $f \geq 5$ in the RC, we obtained the 200 top-scoring keywords for each combination of keyness measure and RC. The various measures show substantial differences (Figure 1). The results suggest that LR_{conf} is intermediate between the other two measures, but it is closer to LR than G^2. The choice of the RC has more influence than expected – according to the literature, the keywords should be relatively stable for large RC (Scott & Tribble 2006). While the top-200 lists against SZ vs. FAZ have an overlap of 89.0% for G^2, the overlap is only 78.0% for LR_{conf} and 58.5% for LR. Especially for LR, which tends to select lower-frequency keywords, the differences are partly due to the frequency threshold on the RC; but even for keywords occurring $f \geq 5$ times in both RC, the ranking differences are bigger than for G^2. A crucial difference between the keyness measures is that G^2 prefers keywords with high frequency in the target corpus, while LR is biased towards low-frequency keywords (Hardie 2014) . Figure 2 confirms this expectation, showing LR_{conf} as a compromise between the two extremes.

Our evaluation is based on a manual classification of the keywords into 33 categories previously identified in a detailed qualitative analysis (Peters 2017). A keyword is considered a true positive (TP) if it can be clearly assigned to a category or expresses positive or negative sentiment, otherwise it is considered a false positive (FP). All six lists were pooled for the manual annotation, resulting in a total of 455 distinct lemmas.

Figure 3 shows the precision achieved for each keyness measure and each of the two RC. Overall, keyword identification works fairly well with precision ranging from 60.5% to 66.0%. Differences between the measures and RC are small, with LR and LR_{conf} slightly better than G^2. None of the differences are significant according to a McNemar-style test (Evert 2004).

In corpus-based discourse analysis, precision is only of secondary concern: manually discarding many FPs may be tedious, but does not lower the quality of the final analysis if sufficient support can be found for all relevant categories. The main aspect of our evaluation is thus on the recall of top-200 keyword lists, i.e. how much support they give to each category. The plot below shows the number of keywords supporting each category. Results are shown for the larger SZ RC; those for FAZ look very similar.

A key observation is that some categories have substantial support from automatically extracted keywords, whereas others – notably metaphors – can hardly be inferred from the corpus-based analysis. There are moderate differences between the keyness measures: G^2 identifies many evaluative terms, whereas LR finds more keywords related to pathogens, hospitals and scientific background. However, a corpus-based analysis using either measure would likely come up with the same categories (and also miss the same set, especially metaphors). Surprisingly, LR_{conf} does not appear to be simply a compromise between G^2 and LR; in some categories, it is even more extreme than LR.

References

Baker, Paul, Gabrielatos, Costas, & McEnery, Tony (2013). *Discourse analysis and media attitudes: The Representation of Islam in the British Press*. Cambridge: Cambridge University Press.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, **19**(1), 61–74.

Evert, Stefan (2004). Significance tests for the evaluation of ranking methods. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004), pages 945–951, Geneva, Switzerland.

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

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

Kilgarriff, Adam (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6**(1), 97–133.

Lijffijt, Jefrey, Nevalainen, Terttu, Säily, Tanja, Papapetrou, Panagiotis, Puolamäki, Kai, & Mannila, Heikki (2016). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, **31**(2), 374–397.

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

Peters, Joachim (2017). Den Feind beschreiben. Multiresistente Erreger im deutschen Pressediskurs. Eine diskurslinguistische Untersuchung der Jahre 1994–2015 (master's thesis). Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, unpublished.

Schmid, Helmut; Fitschen, Arne; Heid, Ulrich (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004), pages 1263–1266, Lisbon, Portugal.

Scott, Mike; Tribble, Christopher (2006). *Textual patterns – Key words and corpus analysis in language education*. Studies in Corpus Linguistics: Vol. 22. Amsterdam, Philadelphia: John Benjamins.

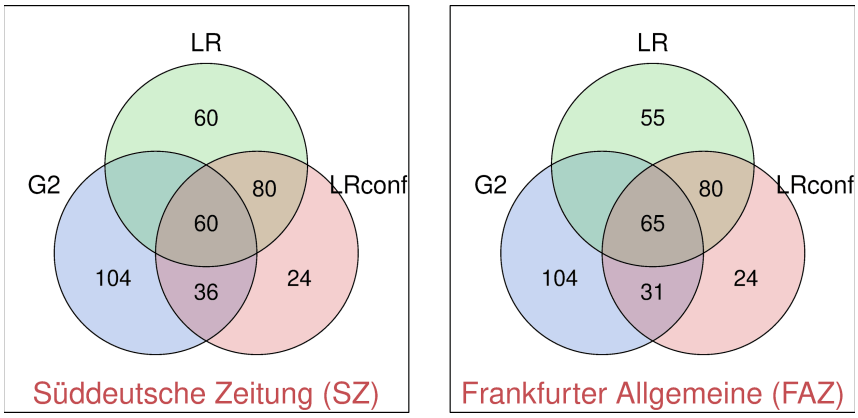

Figure 1: Overlap between top-200 keyword lists for each RC.

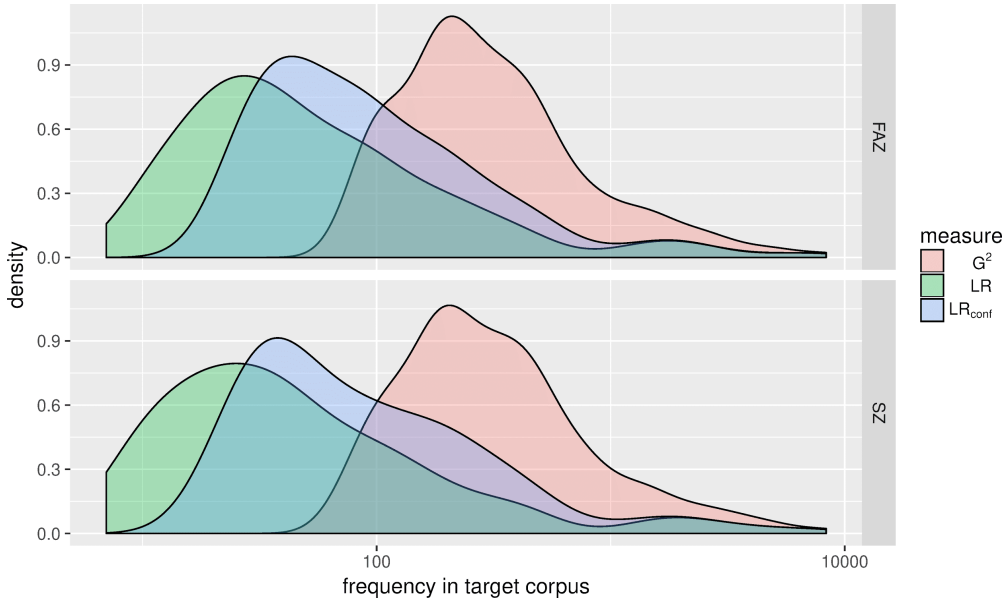

Figure 2: Distribution of target corpus frequency in top-200 lists according to different keyness measures.

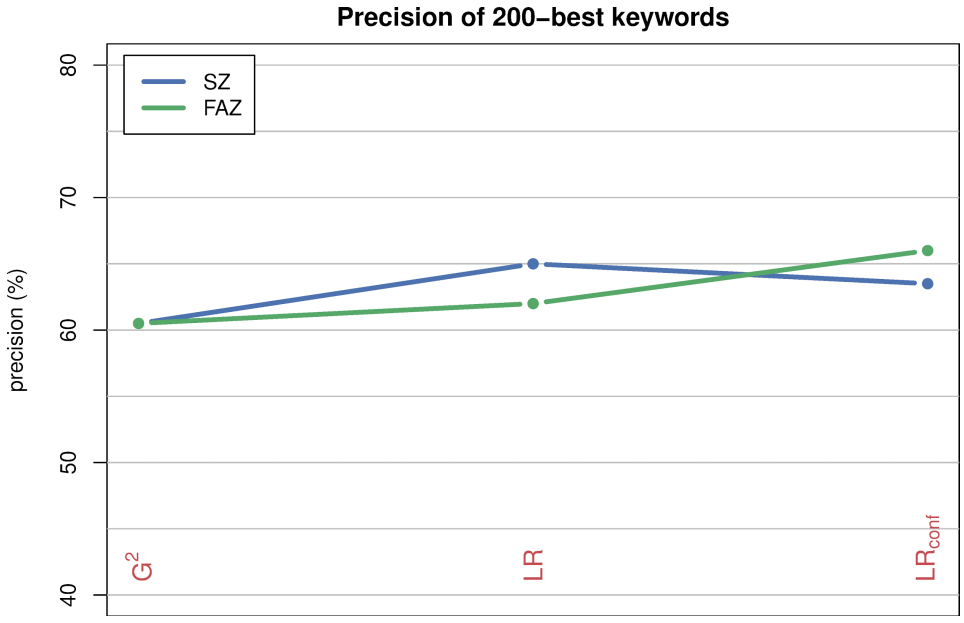

Figure 3: Precision of the top-200 keyword lists.

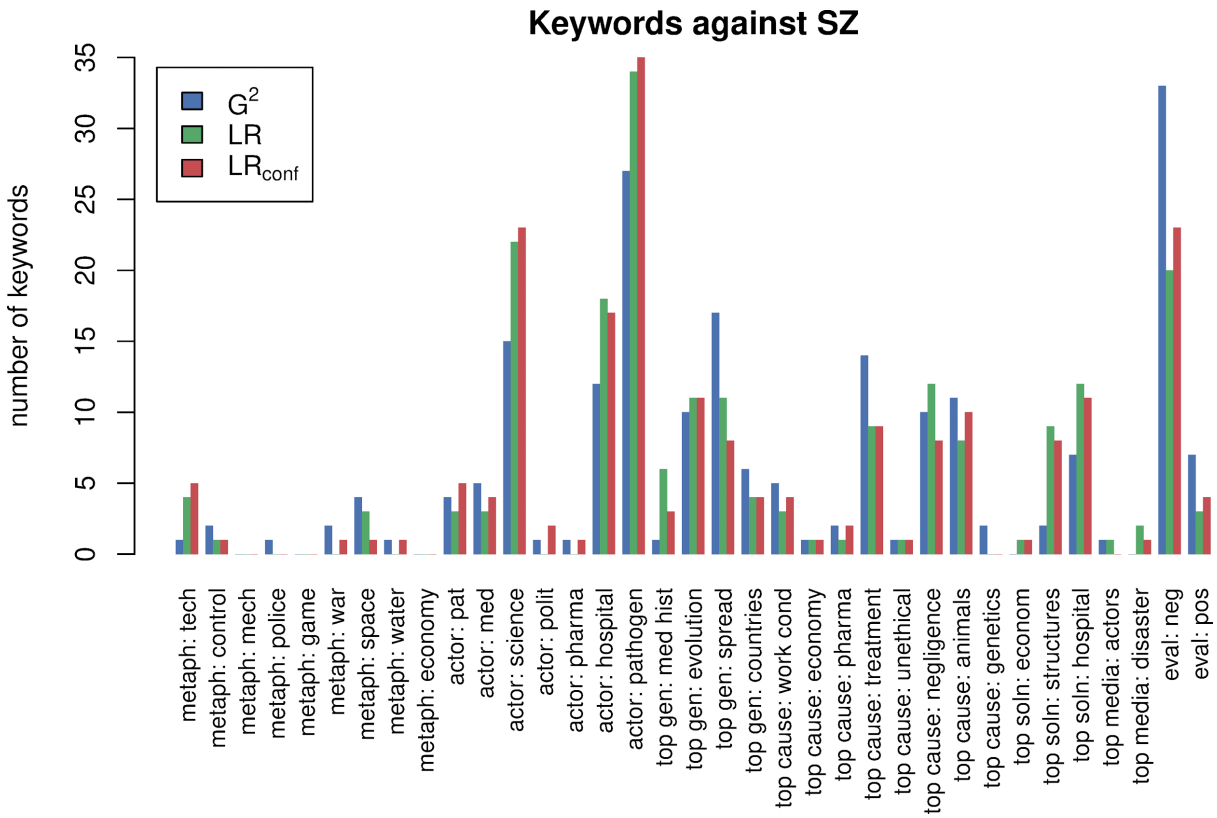

Figure 4: Number of keywords supporting each category among top-200 lists for three different keyness measures.