Corpus Linguistics in the South 15
University of Cambridge, 28 October 2017

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

⇨ The larger the item frequencies and/or corpora, the smaller the differences that will be significant.

Statistical Significance: Nature (2)

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).

Statistical Significance: Metrics & Thresholds (1)

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=0.0000000001 (or even higher) (Pojanapunya, & Watson Todd, 2016).

Statistical Significance: Metrics & Thresholds (2)

Sensitivity of stat.sig. 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: **BIC ≈ LL – log(N)** (Wilson, 2013: 6).
- BIC can be seen as a way to normalise *p*-values ⇨ BIC values can be compared across different keyness analyses.

Interpreting BIC Values

BIC	Degree of evidence against H_0
<0	No evidence – favours H_0
0-2	Not worth more than a bare mention
2-6	Positive evidence against H_0
6-10	Strong evidence against H_0
>10	Very strong evidence against H_0

Selecting KIs for manual analysis

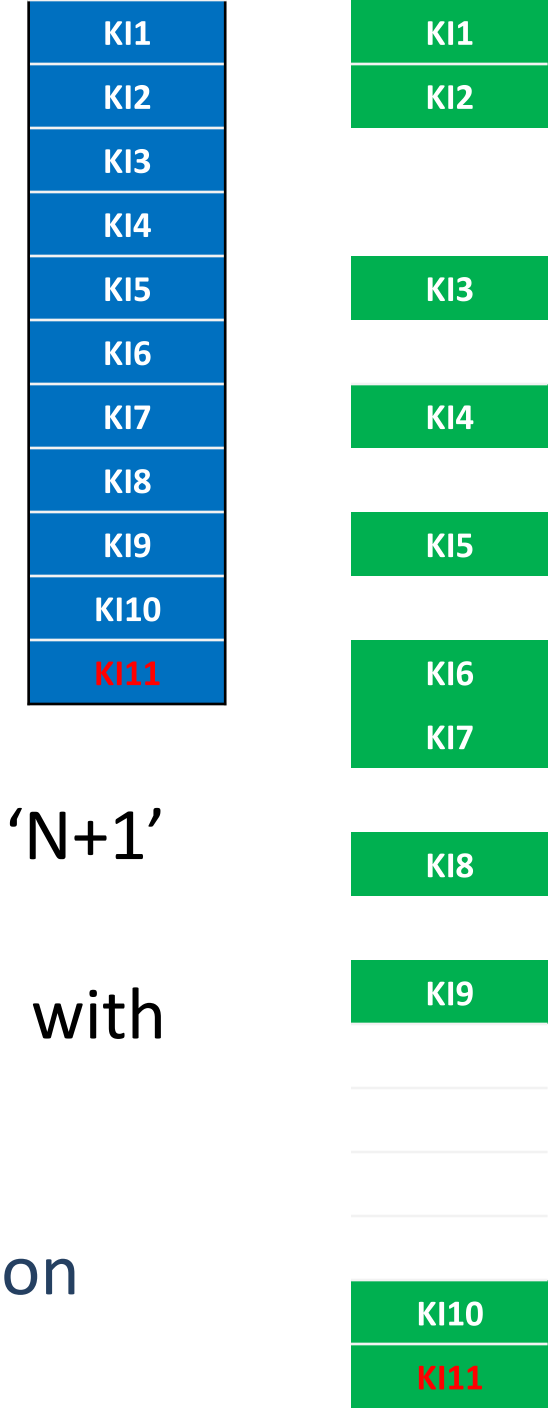
Current approaches

- 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.

$$\boxed{\text{No of clusters} = \frac{\text{Total N of CKIs}}{\text{N of CKIs to be examined}}}$$

- 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.

Clustering: approach and settings (3)

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.
  - CM2017: 29,954 words; LM2017: 23,691 words.
- 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≤0.01 (G^2≥6.63), 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

In CM, but not in LM	**In LM**, but not in CM	
United	Labour’s	banks
Kingdom	equality	renters
universities	unions	women’s
shall	LGBT	failure
shale	reinstate	enforce
	scrap	extending
	privatisation	centres

Differences: BIC≥2

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

The effect of changing the stat.sig. threshold

- Let’s compare the CKIs returned when BIC≥2 (p≤0.001) with those returned with the less strict statistical significance threshold of p≤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≤0.001) are marked in yellow.

Cluster	Difference: CKIs in CM2017 (p<0.01)
1	1:UNITED
2	2:KINGDOM
3	3:UNIVERSITIES
4	4:SHALL
5	5:SHALE
6	6:YOUNGER; 7:AHEAD; 8:YOUR
7	9:EASIER; 10:MERITOCRACY
8	11:DESIGN; 12:MIGHT ; 13:ELDERLY; 14:COMPETITIVE; 15:DEEP; 16:ACTIVE; 17:ATTRACT; 18:PUPILS
9	19:EXCEPTIONAL; 20:THINGS; 21:LEADERS; 22:WRONG; 23:GLOBE; 24:EDINBURGH; 25:REGULATORS; 26:EXPLORE; 27:COMBAT; 28:WORRY; 29:GOVERN
10	30:STABLE; 31:DATA; 32:PROSPEROUS; 33:DIFFICULT; 34:FRAMEWORK; 35:BELIEVE; 36:MUCH; 37:GENERATIONS; 38:GO; 39:INFORMATION; 40:ONLINE; 41:IF; 42:INSTITUTIONS; 43:LEADERSHIP; 44:TECHNICAL; 45:OPPORTUNITY; 46:TECHNOLOGY; 47:OLD; 48:SIGNIFICANT; 49:POOR; 50:DIGITAL; 51:GREAT; 52:REMAIN; 53:WORLD'S; 54:STRONG; 55:PARTNERSHIP; 56:THERESA; 57:BETTER; 58:WANT; 59:MARKETS; 60:STRONGER; 61:HELP; 62:INTERESTS; 63:PROSPERITY; 64:NATION; 65:UNION; 66:GREATER; 67:NOW; 68:WORLD; 69:DO; 70:TOGETHER; 71:LEAVE; 72:SCHOOL; 73:CONTINUE; 74:BEST; 75:EUROPEAN; 76:RIGHT; 77:SHOULD; 78:ABOUT; 79:USE; 80:AROUND; 81:TAKE; 82:BRITISH; 83:SO; 84:THOSE; 85:CAN; 86:MAKE; 87:WE; 88:THIS; 89:IT; 90:BRITAIN; 91:PEOPLE; 92:IN

Clusters	Difference: CKIs LM2017 (p<0.01)
1	1:LABOUR'S
2	2:EQUALITY
3	3:UNIONS
4	4:LGBT
5	5:REINSTATE
6	6:SCRAP
7	7:PRIVATISATION; 8:BANKS
8	9:RENTERS; 10:WOMEN'S; 11:FAILURE; 12:ENFORCE; 13:EXTENDING; 14:CENTRES; 15:NEGOTIATING; 16:PROBATION; 17:ADULT
9	18:PROCUREMENT; 19:INSECURE; 20:WAGES; 21:HIV; 22:TOURISM; 23:PRIORITISE; 24:REINTRODUCE; 25:PROFIT; 26:YOUTH; 27:TRANSITION; 28:REVERSE; 29:RESOLUTION; 30:NEGLECT; 31:ABOLISH; 32:PROFITS; 33:MATERNITY; 34:OPERATIVE; 35:UNLIKE; 36:LIBRARIES; 37:RECOGNITION; 38:LATE; 39:CONTROLS; 40:HANDS; 41:BALANCE; 42:MUSIC; 43:DELIVERS; 44:JUDICIAL; 45:OPTIONS; 46:FARES
10	47:LABOUR; 48:CUTS; 49:OFFICERS; 50:UN; 51:FAILED; 52:OWNERSHIP; 53:EQUAL; 54:ECONOMIES; 55:CRISIS; 56:WAR; 57:FORMS; 58:PEACE; 59:ALLOWANCE; 60:TARGETS; 61:FEES; 62:GUARANTEE; 63:REGIONAL; 64:LEGISLATION; 65:TRADING; 66:ARRANGEMENTS; 67:VITAL; 68:STAFF; 69:LED; 70:RANGE; 71:PLANS; 72:RIGHTS; 73:HOURS; 74:TOWARDS; 75:WOULD; 76:FULLY; 77:OWNED; 78:WORKERS; 79:DISABILITIES; 80:STANDARDS; 81:DISCRIMINATION; 82:FOOD; 83:UNDER; 84:BACK; 85:CLIMATE; 86:CONSULT; 87:CUT; 88:CONSERVATIVES; 89:PRIVATE; 90:JOBS; 91:ENVIRONMENTAL; 92:TRANSPORT; 93:INVEST; 94:WOMEN; 95:EMPLOYMENT; 96:SECTOR; 97:HOMES; 98:END; 99:MANY; 100:ALL; 101:FUNDING; 102:PROTECT; 103:REVIEW; 104:BEEN; 105:COMMUNITIES; 106:INTO; 107:ON

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)

- Gabrielatos, C. (2018) Keyness analysis: nature, metrics and techniques. In Taylor, C. & Marchi, A. (eds.) *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge.
- Gabrielatos, C. & Marchi, A. (2011) Keyness: matching metrics to definitions. *Corpus Linguistics in the South* 1, University of Portsmouth, 5 November 2011. [http://eprints.lancs.ac.uk/51449].
- Gabrielatos, C. & Marchi, A. (2012) Keyness: appropriate metrics and practical issues. *CADS International Conference*, Bologna, Italy, 13-15 September 2012. [https://repository.edgehill.ac.uk/4196].
- Gries, S.Th. (2010). Useful statistics for corpus linguistics. In: Sánchez, A. & Almela, M. (eds.) *A Mosaic of Corpus Linguistics: selected approaches*. Frankfurt am Main: Peter Lang, 269-291.
- Kilgarriff, A. (2001) Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133.

References (2)

- Partington, A. (2014) Mind the gaps: the role of corpus linguistics in researching absences. *International Journal of Corpus Linguistics*, 19(1), 118–146.
- Partington, A. & Duguid, A. (2018) Using corpus linguistics to investigate absence/s. In Taylor, C. & Marchi, A. (eds) *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge.
- Pojanapunya, P. & Watson Todd, R. (2016) Log-likelihood and odds ratio: keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory,* DOI: 10.1515/cllt-2015-0030.
- Rayson P., Berridge D. & Francis B. (2004) Extending the Cochran rule for the comparison of word frequencies between corpora. In Purnelle G., Fairon C. & Dister A. (eds.) *Le poids des mots: proceedings of the 7th International Conference on Statistical analysis of textual data JADT 2004., Vol. 2*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain, 926-936.

References (3)

- Rayson, P. 2009. Wmatrix: A web-based corpus processing environment, Computing Department, Lancaster University. Retrieved from http://ucrel.lancs.ac.uk/wmatrix.
- Scott, M. 2016. *WordSmith Tools version 7*. Stroud: Lexical Analysis Software.
- Taylor, C. (2013) Searching for similarity using corpus-assisted discourse studies. *Corpora*, 8(1), 81-113.
- Taylor, C. (2018) Similarity. In Taylor, C. & Marchi, A. (eds) *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge.
- Wilson, A. (2013) Embracing Bayes factors for key item analysis in corpus linguistics. In Bieswanger, M. & Koll-Stobbe, A. (eds.) *New Approaches to the Study of Linguistic Variability*. Vol. 4. Frankfurt: Peter Lang, 3-11.