Corpus Linguistics in the South

Theoretical-methodological challenges in corpus approaches to discourse studies – and some ways of addressing them

Keyness

Matching metrics to definitions

Costas Gabrielatos & Anna Marchi

Lancaster University

University of Portsmouth, 5 November 2011

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

a. the consistency of definitions given in different sources;
b. the metrics used to calculate the level of keyness;
c. 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 100 keywords as ranked by the metric used. This renders the issue of the appropriate metric central to any study using keyword analysis.

In this pilot 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 two sets of comparisons between corpora of different 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).
- Examination of frequency differences of particular sets of words (e.g. central modals) has shown discrepancies between ranking by frequency difference and ranking by LL (Gabrielatos 2007; Gabrielatos & McEnery, 2005)

→ The ranking criterion becomes very important.

- Usually the criterion is *keyness.*

→ *What is a keyword?*
→ *What is keyness?*
→ *How is it measured?*

→ **Examination of definitions of the terms *keyword* and *keyness*.**

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.

→ ***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 two ...
  - … large corpora of unequal sizes.
  - … Small/medium-sized corpora of 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 / different the two metrics are.
  - High overlap → the two metrics are almost identical.
  - Low overlap → one metric is inappropriate.

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

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. Hard news (6 mil. words)

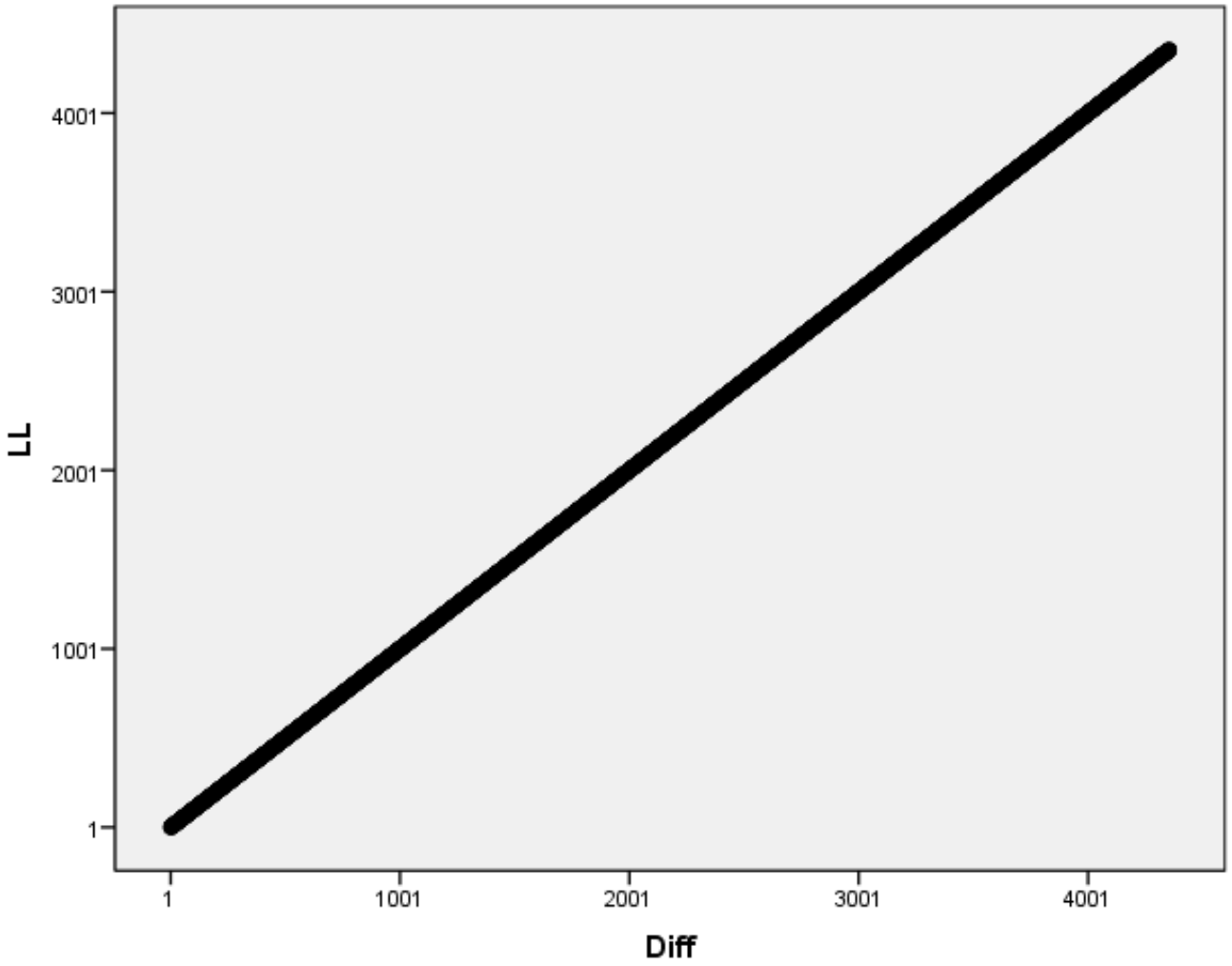
% DIFF: Calculation

$$\frac{(\text{NormFreq in SC} - \text{NormFreq in RC}) \times 100}{\text{NormFreq in RC}}$$

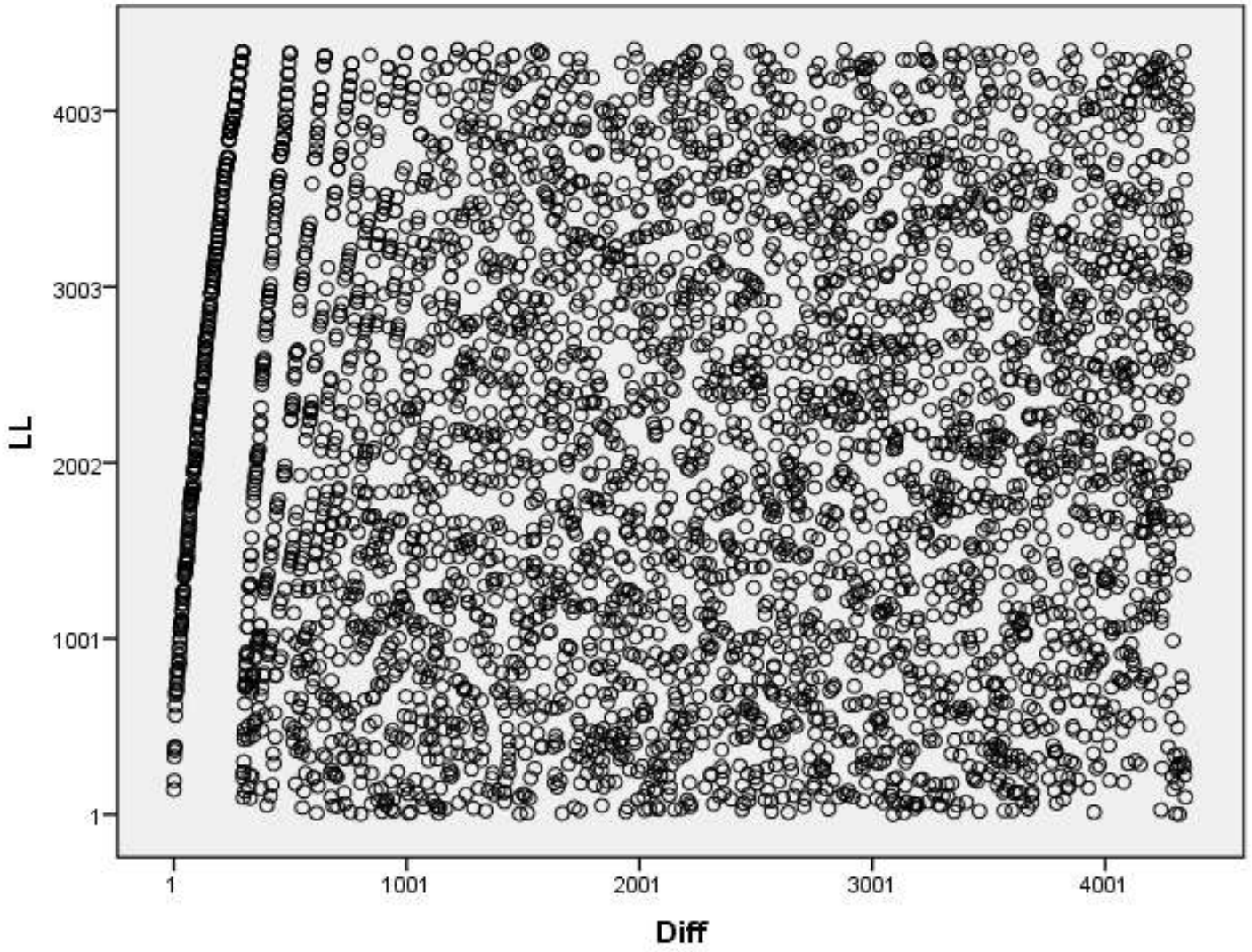
NormFreq	=	normalised frequency
SC	=	study corpus
RC	=	reference corpus

Full overlap

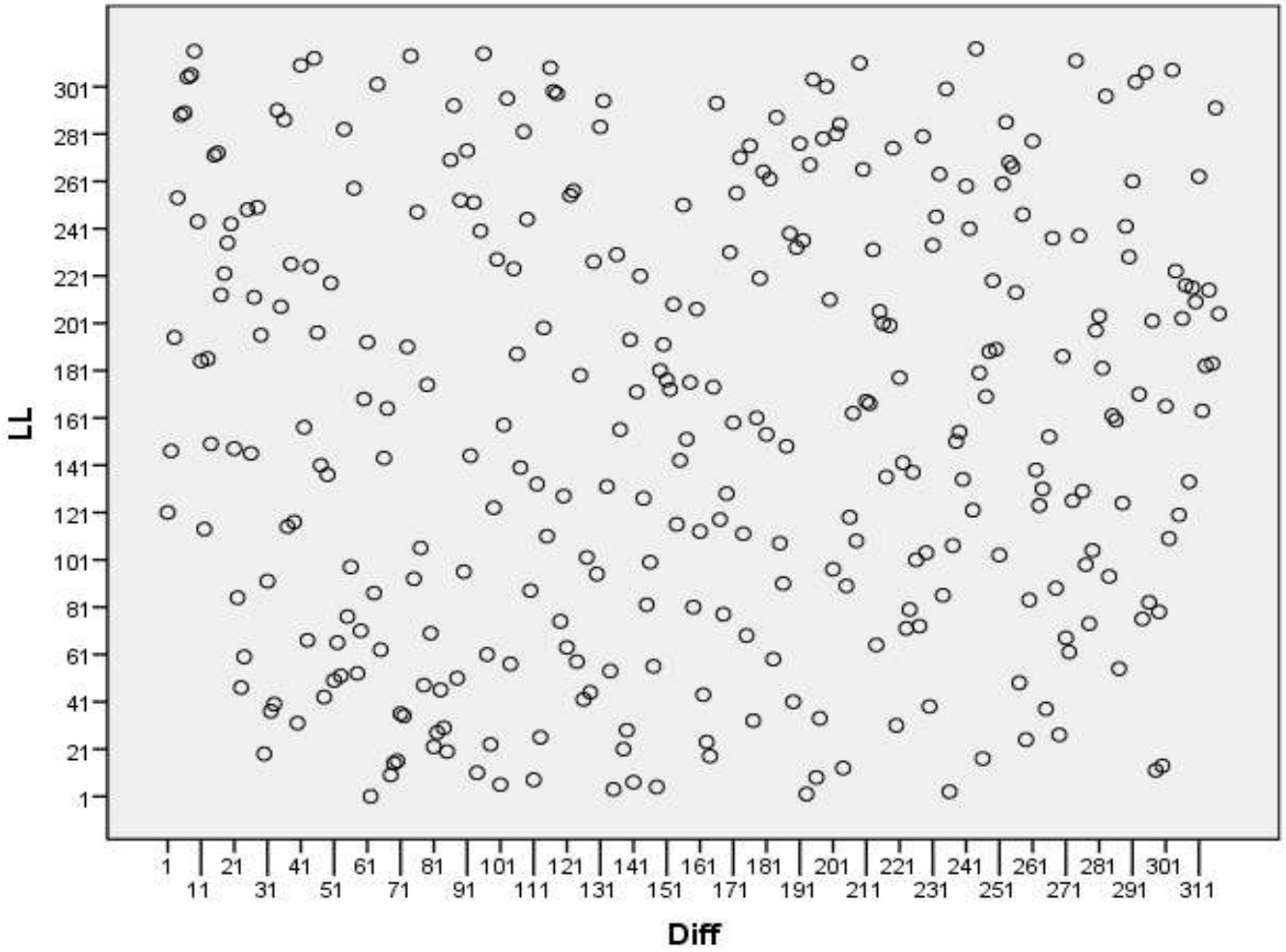
Scatterplot showing a 100% overlap between LL and %DIFF rankings.



Actual overlap: All KWs
96 mil. vs. 156 mil. (4356 KWs)



Actual overlap: All KWs
1 mil. vs. 6 mil. (317 KWs)

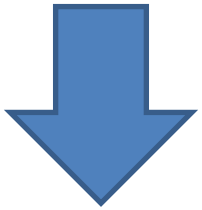


However, this very low overlap may be misleading:

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 top 100

Top 100: Overlap of ranking by LL and %DIFF

96 mil. vs. 156 mil.: ***3*** *shared KWs*

1 mil. vs. 6 mil.: ***38*** *shared KWs*

Different KWs may have **markedly different LL but similar %DIFF**

- DELORS (100th, LL=3,192.68), PAPANDREOU (761st, LL=677.85)
- But %DIFF is very similar: DELORS 5,386%, PAPANDREOU 5,340.5%

Different KWs may have **similar LL but very different %DIFF**

- SERB (33rd, LL =6,966.10), BRITISH (34th, LL=6,732.14)
- But %DIFF for SERB is **high** (1496.5%), while for BRITISH it is **low** (46.6%)

LL order (comparison 1)

N	Key word	STUDYraw	STUDYrel	REFraw	REFrel	LL	%DiFF
1	MR	206,523	0.21520862	178,174	0.114041694	38,592.73	88.7
2	THE	6,001,857	6.254273415	8,908,778	5.702134609	32,366.01	9.7
3	EC	15,204	0.015843425	668	0.000427559	24,482.87	3605.6
4	CLINTON	19,793	0.020625422	3,552	0.002273486	21,743.92	807.2
5	OF	2,782,374	2.899390697	4,051,884	2.593440771	20,935.02	11.8
6	BOSNIA	13,488	0.014055257	972	0.000622136	19,871.95	2159.2
7	1991	18,233	0.018999815	4,286	0.002743289	17,421.60	592.6
8	RECESSION	12,484	0.013009032	1,209	0.00077383	17,107.37	1581.1
9	YELTSIN	9,829	0.010242373	234	0.000149774	16,996.41	6738.6
10	CORRESPONDENT	14,743	0.015363038	2,762	0.00176784	15,873.56	769.0
11	MILLION	84,491	0.08804439	74,489	0.047677282	14,863.76	84.7
12	MAJOR	41,747	0.043502729	26,649	0.017056907	14,745.52	155.0
13	MAASTRICHT	8,669	0.009033587	325	0.000208019	14,268.10	4242.7
14	WHICH	316,733	0.330053657	388,096	0.248403952	13,946.18	32.9
15	BOSNIAN	9,159	0.009544195	677	0.000433319	13,418.82	2102.6
16	1992	16,593	0.017290842	5,476	0.003504958	12,582.16	393.3
17	CENT	89,755	0.093529768	87,509	0.056010835	11,556.27	67.0
18	SERBS	7,289	0.007595549	408	0.000261144	11,286.20	2808.6
19	LETTER	27,558	0.028716991	16,566	0.010603201	10,721.60	170.8
20	GOVERNMENT	95,247	0.099252746	96,797	0.061955694	10,559.28	60.2

% DIFF order (comparison 1)

N	Key word	STUDYraw	STUDYrel	REFraw	REFrel	LL	%DiFF
142	VANCE-OWEN	1,254	0.001306739	0	1E-19	2,423.36	1306738704442970000.0
193	KHASBULATOV	1,008	0.001050393	0	1E-19	1,947.96	1050392864271990000.0
329	BT3	658	0.000685673	0	1E-19	1,271.59	685673090629279000.0
363	RUTSKOI	615	0.000640865	0	1E-19	1,188.49	640864658635109000.0
379	BRAER	592	0.000616897	0	1E-19	1,144.04	616897363215684000.0
389	1ST-HALF	576	0.000600224	0	1E-19	1,113.12	600224477238953000.0
565	BEREGOVOY	438	0.000456421	0	1E-19	846.43	456420704722404000.0
620	HVO	410	0.000427243	0	1E-19	792.32	427243125159293000.0
681	FERRUZZI	381	0.000397023	0	1E-19	736.28	397023482946678000.0
700	PRESTRIDGE	374	0.000389729	0	1E-19	722.75	389729102607816000.0
707	NARBROUGH	371	0.000386603	0	1E-19	716.96	386602914659306000.0
713	RUTSKOY	367	0.000382435	0	1E-19	709.23	382434693165123000.0
742	OFGAS	356	0.000370972	0	1E-19	687.97	370972091332077000.0
752	WAIGEL	353	0.000367846	0	1E-19	682.17	367845903383567000.0
786	GOODA	344	0.000358467	0	1E-19	664.78	358467397745698000.0
811	HOFBRAU	333	0.000347005	0	1E-19	643.52	347004766808822000.0
818	DOULL	328	0.000341795	0	1E-19	633.86	341794511768966000.0
826	ADLEY	327	0.000340752	0	1E-19	631.93	340752449119463000.0
827	KRAVCHUK	327	0.000340752	0	1E-19	631.93	340752449119463000.0
861	ADOLLARS	317	0.000330332	0	1E-19	612.6	330331880832090000.0

LL vs. %DIFF

The same KW may have **very high LL but very low %DIFF**

- THE : LL= 32,366.01 (2nd) but % DIFF = 9.7%.
- OF : LL = 20,935.02 (5th) but % DIFF = 11.8%

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 **very high %DIFF but (relatively) low LL**

- ADVENTISTS: %DIFF = 2086.3% but LL= 137.49
- EX-COMMUNIST: %DIFF = 679.1% but LL= 136.61

However high, the %DIFF also needs to be statistically significant.

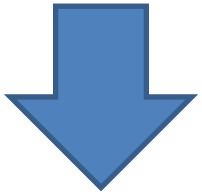
Conclusions (1)

- 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

Proposal

- %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

Further considerations and research

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

- *%DIFF is straightforward and easily computed.*
  → *Possibility of more sophisticated metric for effect size?*

- *Does absolute corpus size matter?*
- *Do relative corpus sizes matter?*
- *Does the corpus type matter (e.g. general, specialised)?*

Watch this space

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 to create a column for %DIFF in Excel (1)

In the ‘relative frequencies’ column of the REF corpus , substitute all zero frequencies with **1E-19** (*0.00000000000000000001*)

→ Why? Because you can’t divide by zero.

→ Very small number (1 divided by ten quadrillion) …

… 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 (2)

1. Add a column with header **% DIFF** .
2. In the cell below the header, write this 'function':

$$= (X2 - Y2) / Y2 * 100$$

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

- Andrew, D.P.S., Pedersen, P.M. & McEvoy, C.D. (2011). *Research Methods and Design in Sport Management*. Human Kinetics.
- Biber, D., Connor, U. & Upton, A. with Anthony, M. & Gladkov, K. (2007). Rhetorical appeals in fundraising. In D. Biber, U. Connor & A. Upton. *Discourse on the Move: Using corpus analysis to describe discourse structure* (pp. 121-151). Amsterdam: John Benjamin.
- Gabrielatos, C. (2007). *If*-conditionals as modal colligations: A corpus-based investigation. In M. Davies, P. Rayson, S. Hunston & P. Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference: Corpus Linguistics 2007*. Birmingham: University of Birmingham.
- Gabrielatos, C. & McEnery, T. (2005). Epistemic modality in MA dissertations. In P.A. Fuertes Olivera (ed.), *Lengua y Sociedad: Investigaciones recientes en lingüística aplicada*. Lingüística y Filología no. 61. (pp. 311-331). Valladolid: Universidad de Valladolid.
- Mujis, D. (2010). *Doing Quantitative Research in Education with SPSS*. Sage.
- Ridge, E. & Kudenko, D. (2010). Tuning an algorithm using design of experiments. In T. Batz-Beiselstein, M. Chiarandini, L. Paquette & M. Preuss (Eds.), *Experimental Methods for the Analysis of Optimization Algorithms* (265-286). Springer.
- Rosenfeld, B. & Penrod, S.D. (2011). *Research Methods in Forensic Psychology*. John Wiley and Sons.
- Scott, M. (1996). *WordSmith Tools Manual*. Oxford: Oxford University Press.
- Scott, M. (1997). PC analysis of key words - and key key words. *System,* 25(2), 233-45.
- Scott, M. (2011). *WordSmith Tools Manual, Version 6*. Liverpool: Lexical Analysis Software Ltd.
- Taylor, C. (2011). Searching for similarity: The representation of boy/s and girl/s in the UK press in 1993, 2005, 2010. Paper given at *Corpus Linguistics 2011*, University of Birmingham, 20-22 July 2011.