Unit 4: Measuring Keyness

Statistics for Linguistics with R – a SIGIL course

Prof. Dr. Stephanie Evert

Computational Corpus Linguistics

FAU Erlangen-Nürnberg

https://www.stephanie-evert.de/ | @schtepf

What are Donald Trump's favourite words?

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

https://www.thetrumparchive.com/

	Trump tweets **(target)**	other tweets **(reference)**
crooked	p = 340 pmw TTA: $f = 453$	p = 6.4 pmw
everyone	p = 404 pmw TTA: $f = 538$	p = 404 pmw

→ **keywords** "occur with unusual frequency in a given text" or text collection (Scott 1997: 236)

→ basis: frequency comparison with reference corpus

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

Keywords in corpus linguistics

- Aboutness of a text ➞ key keywords (Scott 1997)
- Technical/genre terminology (Paquot & Bestgen 2009)
- Literary style (Culpeper 2009)
- Linguistic & cultural differences (Oakes & Farrow 2006)
- Historical perspectives (Fidler & Cvrcek 2015)
- Similarity of text collections (Rayson & Garside 2000)
- Corpus-based discourse analysis (Baker 2006)
  - also know as corpus-assisted discourse studies (CADS)
  - clusters of keywords represent central topics, actors, metaphors, and framings (e.g. McEnery et al. 2015)

Keyness

- More generally, **keyness** is one of the most fundamental concepts in corpus linguistics
- Frequency comparison between corpora **A** and **B** (representative of underlying linguistic populations)
- For different kinds of lexico-grammatical items
  - word forms, lemmas, n-grams, multiword expressions
  - morphemes, grammatical constructions, n-grams of tags
- Wide range of applications depending on choice of lexico-grammatical items and of corpora **A** and **B**

Applications of keyness

Bibliographic keywords

- **A** = text, **B** = collection ➞ aboutness of text
- also: key keywords (that are key in many texts)

Target corpus vs. reference

- **A** = domain, **B** = general language ➞ terminology
- items = n-grams / MWE ➞ multiword terms (SkE)
- **A** = thematic corpus, **B** = reference ➞ discourse (CADS)

Applications of keyness

Symmetric keyword analysis

- **A**, **B** similar but “opposite” ➞ contrastive framings (e.g. liberal vs. conservative newspaper)

Collocation identification

- **A** = contexts of node word, **B** = rest of corpus ➞ collocations of node word

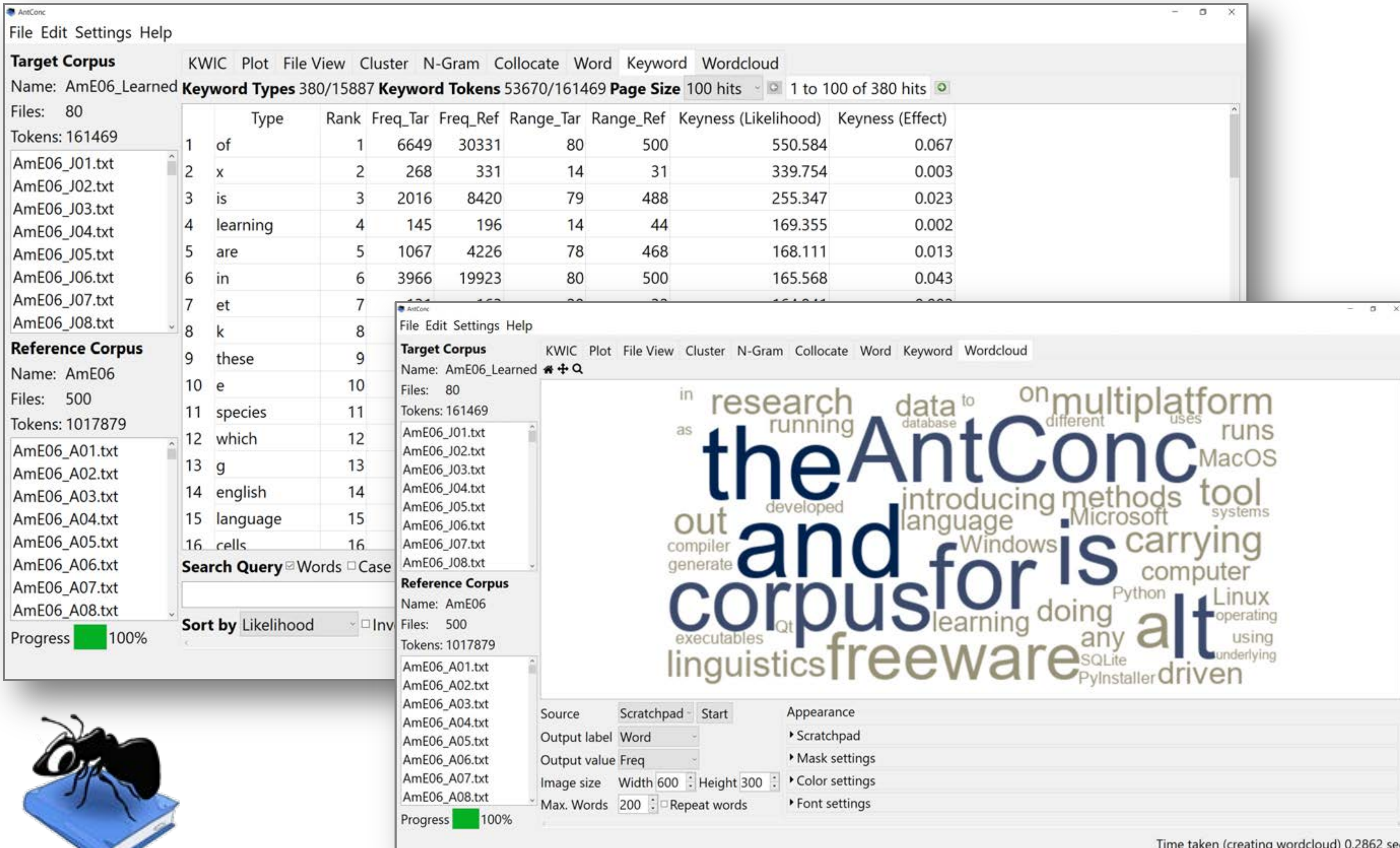
Corpus comparison

- **A**, **B =** comparable corpora, items = grammatical constructions ➞ language variation

Keywords in CQPweb

No.	Word	In whole "German COVID-19 tweets (v2)":		In corpus "German Reference Tweets (2018/2019)":		+/-	Conservative LR
		Frequency (absolute)	Frequency (per mill)	Frequency (absolute)	Frequency (per mill)		
1	**Corona**	2,114,391	9,453.42	42	0.37	+	13.14
2	**#Corona**	1,048,731	4,688.87	5	0.04	+	12.34
3	**paNdEMIe**	167,967	750.98	20	0.18	+	9.88
4	**lockDown**	143,748	642.70	25	0.22	+	9.56
5	**#Lockdown**	113,326	506.68	5	0.04	+	9.13
6	**NeuINFEKTIONEN**	70,034	313.12	5	0.04	+	8.44
7	**rki**	63,840	285.43	23	0.20	+	8.43
8	**#Pandemie**	55,078	246.25	5	0.04	+	8.09
9	**impfstoff**	82,550	369.08	69	0.61	+	8.07
10	**Quarántäne**	71,280	318.69	58	0.51	+	8
27	**@BAG_OFSP_UFSP**	23,571	105.39	34	0.30	+	6.78
28	**Infektion**	39,620	177.14	90	0.80	+	6.77
29	**fAlLzAhLeN**	25,385	113.50	46	0.41	+	6.69
30	**Biontech**	20,703	92.56	5	0.04	+	6.68
31	**#querdenker**	16,895	75.54	18	0.16	+	6.6
32	**infiziert**	58,135	259.92	192	1.70	+	6.56
33	**IntensivstationeN**	16,277	72.77	18	0.16	+	6.54

Keywords in AntConc



But what is happening behind the scenes when you use such software?

INSIDE THE BLACK BOX

Measuring keyness

- Compare frequency in **A** with frequency in **B** separately for each candidate term $w \in C$

Frequency data for *w*

- f_1 = freq. in corpus **A**
- n_1 = sample size of **A**
- f_2 = freq. in corpus **B**
- n_2 = sample size of **B**

	A	**B**
w	f_1	f_2
$\neg w$	$n_1 - f_1$	$n_2 - f_2$
	$= n_1$	$= n_2$

Measuring keyness

- Recent studies: document frequency more robust than term frequency (e.g. Egbert & Biber 2019)

Frequency data for *w*

- f_1 = df in corpus **A**
- n_1 = no. of texts in **A**
- f_2 = df in corpus **B**
- n_2 = no. of texts in **B**

	A	**B**
w	f_1	f_2
$\neg w$	$n_1 - f_1$	$n_2 - f_2$
	$= n_1$	$= n_2$

Measuring keyness

- Goal: compare frequencies π_1 and π_2 of candidate item in sublanguages represented by corpora **A** and **B**
  - statisticians speak of "populations"
- Best sample estimates (MLE)

$$\hat{\pi}_1 = \frac{f_1}{n_1}, \qquad \hat{\pi}_2 = \frac{f_2}{n_2}$$

	A	**B**
w	f_1	f_2
¬ w	$n_1 - f_1$	$n_2 - f_2$
	$= n_1$	$= n_2$

- positive keyword if $\pi_1 >> \pi_2$
- negative keyword if $\pi_1 << \pi_2$

Keyness measures: significance

- Inference about frequency in population **A** vs. **B**

$$H_0 : \pi_1 = \pi_2$$

- Observed contingency table

$O_{11} = f_1$	$O_{12} = f_2$
$O_{21} = n_1 - f_1$	$O_{22} = n_2 - f_2$

- Contingency table of expected frequencies

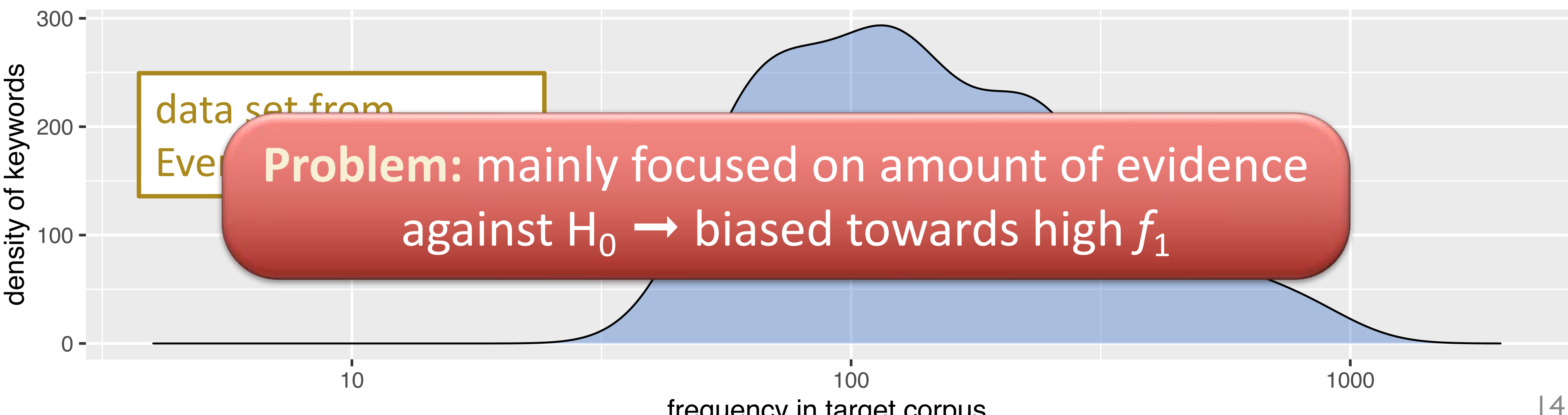
$E_{11} = n_1 \cdot \left(\frac{f_1 + f_2}{n_1 + n_2} \right)$	$E_{12} = n_2 \cdot \left(\frac{f_1 + f_2}{n_1 + n_2} \right)$
$E_{21} = n_1 - E_{11}$	$E_{22} = n_2 - E_{22}$

Keyness measures: significance

Statistical hypothesis tests for H_0 in contingency table:

- **log-likelihood** G^2 (Rayson & Garside 2000)
- chi-squared test X^2 (Scott 1997)
- Fisher's exact test (Lafon 1980)

$$G^2 = 2 \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$



Problem: mainly focused on amount of evidence against H_0 → biased towards high f_1

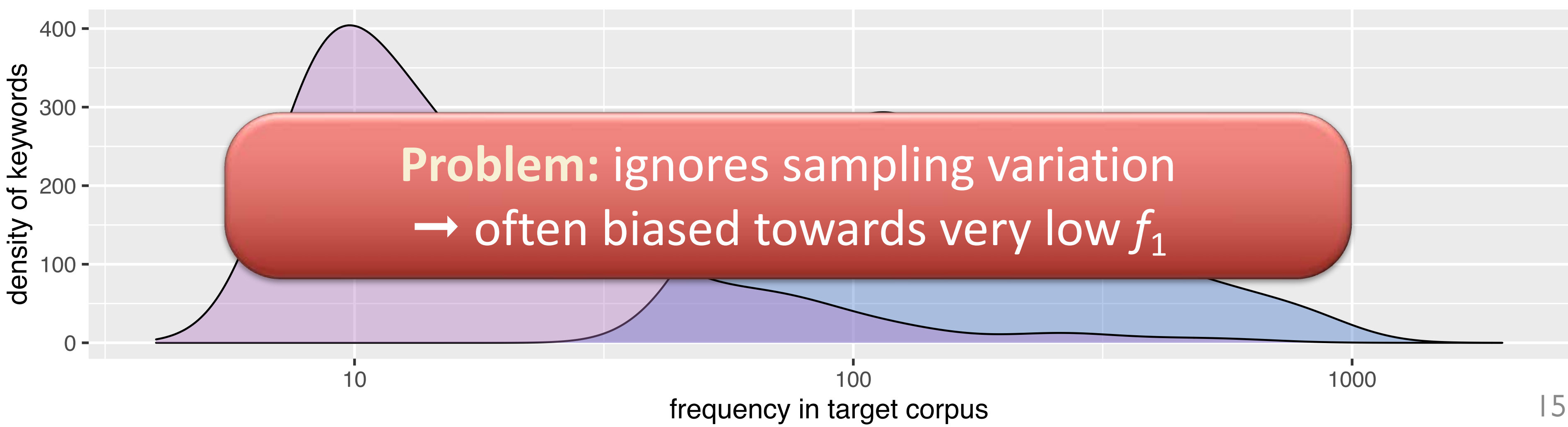
Keyness measures: effect size

Focus on magnitude of difference between π_1 and π_2:

- **LogRatio** (Hardie 2014) = log relative risk r
  - a better version (Walter 1975)

$$\mathrm{LR} = \log_2 \frac{f_1 + \frac{1}{2}}{n_1 + \frac{1}{2}} - \log_2 \frac{f_2 + \frac{1}{2}}{n_2 + \frac{1}{2}}$$

- closely related measures: %DIFF (Gabrielatos & Marchi 2012), RRF, odds ratio, ΔP



Problem: ignores sampling variation
→ often biased towards very low f_1

Keyness measures: significance filter

- Effect-size measures combined with **significance filter**: set score = 0 if not significant according to G^2
- Hardie (2014): control family-wise error rate (FWER) in data set by using **adjusted significance level**

$$\alpha' = 1 - (1 - \alpha)^{\frac{1}{m}} \qquad \text{or} \qquad \alpha' = \frac{\alpha}{m}$$

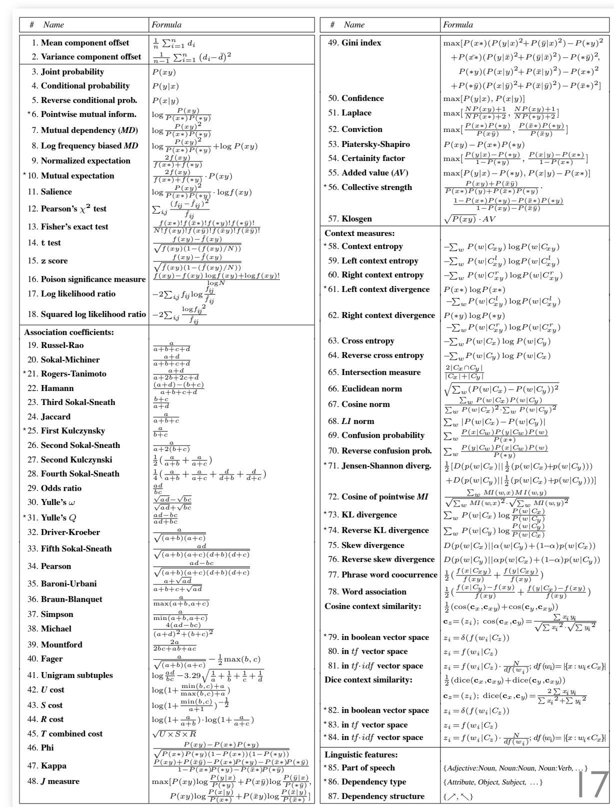
- Heuristic alternative: frequency **threshold**
  - typically $f_1 \geq 5, 10, 100, \ldots$
  - often also requirement $f_2 > 0$ in reference corpus

Keyness measures: heuristics

- Another heuristic: **SimpleMaths** (Kilgarriff 2009)

$$\mathrm{SM} = \frac{10^6 \cdot \frac{f_1}{n_1} + \lambda}{10^6 \cdot \frac{f_2}{n_2} + \lambda} \qquad (\lambda > 0)$$

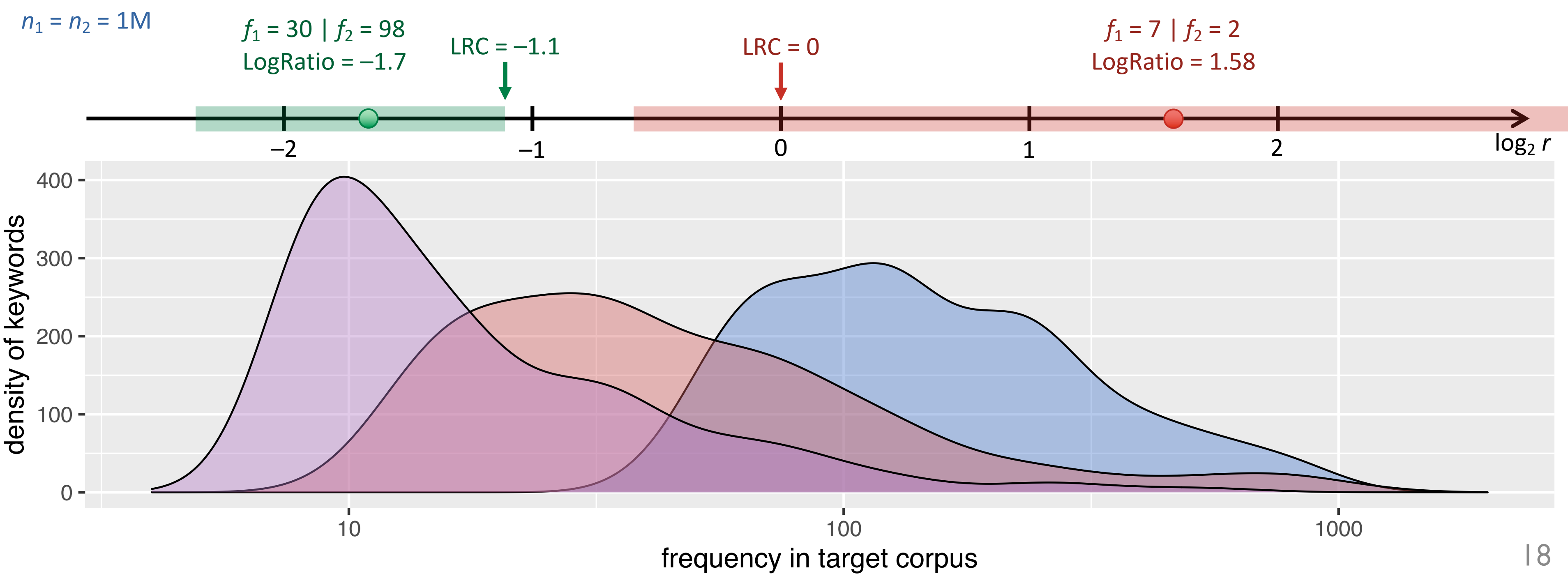
- Mathematician: no comment!
- Many other (often heuristic) **association measures** have been suggested for collocation extraction (e.g. Pecina 2005)
- Hardie (2014) includes AM in his list of keyness measures



#	Name	Formula
1.	**Mean component offset**	$\frac{1}{n}\sum_{i=1}^{n} d_i$
2.	**Variance component offset**	$\frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2$
3.	**Joint probability**	$P(xy)$
4.	**Conditional probability**	$P(y\|x)$
5.	**Reverse conditional prob.**	$P(x\|y)$
*6.	**Pointwise mutual inform.**	$\log\frac{P(xy)}{P(x*)P(*y)}$
7.	**Mutual dependency (*MD*)**	$\log\frac{P(xy)^2}{P(x*)P(*y)}$
8.	**Log frequency biased *MD***	$\log\frac{P(xy)^2}{P(x*)P(*y)} + \log P(xy)$
9.	**Normalized expectation**	$\frac{2f(xy)}{f(x*)+f(*y)}$
*10.	**Mutual expectation**	$\frac{2f(xy)}{f(x*)+f(*y)} \cdot P(xy)$
11.	**Salience**	$\log\frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$
12.	**Pearson's χ^2 test**	$\sum_{ij}\frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
13.	**Fisher's exact test**	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
14.	**t test**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
15.	**z score**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
16.	**Poison significance measure**	$\frac{\hat{f}(xy)-f(xy)\log\hat{f}(xy)+\log f(xy)!}{\log N}$
17.	**Log likelihood ratio**	$-2\sum_{ij} f_{ij}\log\frac{f_{ij}}{\hat{f}_{ij}}$
18.	**Squared log likelihood ratio**	$-2\sum_{ij}\frac{\log f_{ij}^2}{\hat{f}_{ij}}$
	Association coefficients:	
19.	**Russel-Rao**	$\frac{a}{a+b+c+d}$
20.	**Sokal-Michiner**	$\frac{a+d}{a+b+c+d}$
*21.	**Rogers-Tanimoto**	$\frac{a+d}{a+2b+2c+d}$
22.	**Hamann**	$\frac{(a+d)-(b+c)}{a+b+c+d}$
23.	**Third Sokal-Sneath**	$\frac{b+c}{a+d}$
24.	**Jaccard**	$\frac{a}{a+b+c}$
*25.	**First Kulczynsky**	$\frac{a}{b+c}$
26.	**Second Sokal-Sneath**	$\frac{a}{a+2(b+c)}$
27.	**Second Kulczynski**	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
28.	**Fourth Sokal-Sneath**	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$
29.	**Odds ratio**	$\frac{ad}{bc}$
30.	**Yulle's ω**	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
*31.	**Yulle's Q**	$\frac{ad-bc}{ad+bc}$
32.	**Driver-Kroeber**	$\frac{a}{\sqrt{(a+b)(a+c)}}$
33.	**Fifth Sokal-Sneath**	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
34.	**Pearson**	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
35.	**Baroni-Urbani**	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
36.	**Braun-Blanquet**	$\frac{a}{\max(a+b,a+c)}$
37.	**Simpson**	$\frac{a}{\min(a+b,a+c)}$
38.	**Michael**	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
39.	**Mountford**	$\frac{2a}{2bc+ab+ac}$
40.	**Fager**	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2}\max(b,c)$
41.	**Unigram subtuples**	$\log\frac{ad}{bc}-3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
42.	***U* cost**	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$
43.	***S* cost**	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$
44.	***R* cost**	$\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$
45.	***T* combined cost**	$\sqrt{U \times S \times R}$
46.	**Phi**	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
47.	**Kappa**	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
48.	***J* measure**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
49.	**Gini index**	$\max[P(x*)(P(y\|x)^2+P(\bar{y}\|x)^2)-P(*y)^2$ $+P(\bar{x}*)(P(y\|\bar{x})^2+P(\bar{y}\|\bar{x})^2)-P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2$ $+P(*\bar{y})(P(x\|\bar{y})^2+P(\bar{x}\|\bar{y})^2)-P(\bar{x}*)^2]$
50.	**Confidence**	$\max[P(y\|x), P(x\|y)]$
51.	**Laplace**	$\max[\frac{NP(xy)+1}{NP(x*)+2}, \frac{NP(xy)+1}{NP(*y)+2}]$
52.	**Conviction**	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})}, \frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
53.	**Piatersky-Shapiro**	$P(xy)-P(x*)P(*y)$
54.	**Certainity factor**	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)}, \frac{P(x\|y)-P(x*)}{1-P(x*)}]$
55.	**Added value (*AV*)**	$\max[P(y\|x)-P(*y), P(x\|y)-P(x*)]$
*56.	**Collective strength**	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)}\cdot$ $\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
57.	**Klosgen**	$\sqrt{P(xy)}\cdot AV$
	Context measures:	
*58.	**Context entropy**	$-\sum_w P(w\|C_{xy})\log P(w\|C_{xy})$
59.	**Left context entropy**	$-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
60.	**Right context entropy**	$-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
*61.	**Left context divergence**	$P(x*)\log P(x*)$ $-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
62.	**Right context divergence**	$P(*y)\log P(*y)$ $-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
63.	**Cross entropy**	$-\sum_w P(w\|C_x)\log P(w\|C_y)$
64.	**Reverse cross entropy**	$-\sum_w P(w\|C_y)\log P(w\|C_x)$
65.	**Intersection measure**	$\frac{2\|C_x \cap C_y\|}{\|C_x\|+\|C_y\|}$
66.	**Euclidean norm**	$\sqrt{\sum_w (P(w\|C_x)-P(w\|C_y))^2}$
67.	**Cosine norm**	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2\cdot\sum_w P(w\|C_y)^2}$
68.	***L1* norm**	$\sum_w \|P(w\|C_x)-P(w\|C_y)\|$
69.	**Confusion probability**	$\sum_w \frac{P(x\|C_w)P(y\|C_w)P(w)}{P(x*)}$
70.	**Reverse confusion prob.**	$\sum_w \frac{P(y\|C_w)P(x\|C_w)P(w)}{P(*y)}$
*71.	**Jensen-Shannon diverg.**	$\frac{1}{2}[D(p(w\|C_x)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))$ $+D(p(w\|C_y)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))]$
72.	**Cosine of pointwise *MI***	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2}\cdot\sqrt{\sum_w MI(w,y)^2}}$
*73.	**KL divergence**	$\sum_w P(w\|C_x)\log\frac{P(w\|C_x)}{P(w\|C_y)}$
*74.	**Reverse KL divergence**	$\sum_w P(w\|C_y)\log\frac{P(w\|C_y)}{P(w\|C_x)}$
75.	**Skew divergence**	$D(p(w\|C_x)\|\|\alpha(w\|C_y)+(1-\alpha)p(w\|C_x))$
76.	**Reverse skew divergence**	$D(p(w\|C_y)\|\|\alpha p(w\|C_x)+(1-\alpha)p(w\|C_y))$
77.	**Phrase word coocurrence**	$\frac{1}{2}(\frac{f(x\|C_{xy})}{f(xy)}+\frac{f(y\|C_{xy})}{f(xy)})$
78.	**Word association**	$\frac{1}{2}(\frac{f(x\|C_y)-f(xy)}{f(xy)}+\frac{f(y\|C_x)-f(xy)}{f(xy)})$
	Cosine context similarity:	$\frac{1}{2}(\cos(\mathbf{c}_x,\mathbf{c}_{xy})+\cos(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \cos(\mathbf{c}_x,\mathbf{c}_y)=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\cdot\sqrt{\sum y_i^2}}$
*79.	**in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
80.	**in *tf* vector space**	$z_i=f(w_i\|C_z)$
81.	**in *tf·idf* vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x: w_i \epsilon C_x\}\|$
	Dice context similarity:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_i y_i}{\sum x_i^2+\sum y_i^2}$
*82.	**in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
*83.	**in *tf* vector space**	$z_i=f(w_i\|C_z)$
*84.	**in *tf·idf* vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x: w_i \epsilon C_x\}\|$
	Linguistic features:	
*85.	**Part of speech**	{*Adjective:Noun, Noun:Noun, Noun:Verb, …*}
*86.	**Dependency type**	{*Attribute, Object, Subject, …*}
87.	**Dependency structure**	{↗, ↖}

My measure: LRC (Evert 2022)

- Combine effect-size and significance aspects: confidence interval [$\log_2 r_{-}$, $\log_2 r_{+}$] for relative risk
- Conservative estimate **LRC** (conservative LogRatio)
  - use value closest to 0 (not significant if 0 in interval ➞ LRC = 0)



The maths behind LRC

Be careful with approximations such as the one used by CQPweb

- Exact inference for relative risk in contingency table with conditional Poisson test (Fay 2010: 55)

$$\mathbb{P}(f_1 | f_1 + f_2) = \binom{f_1 + f_2}{f_1} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{f_1} \left(1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{f_2}$$

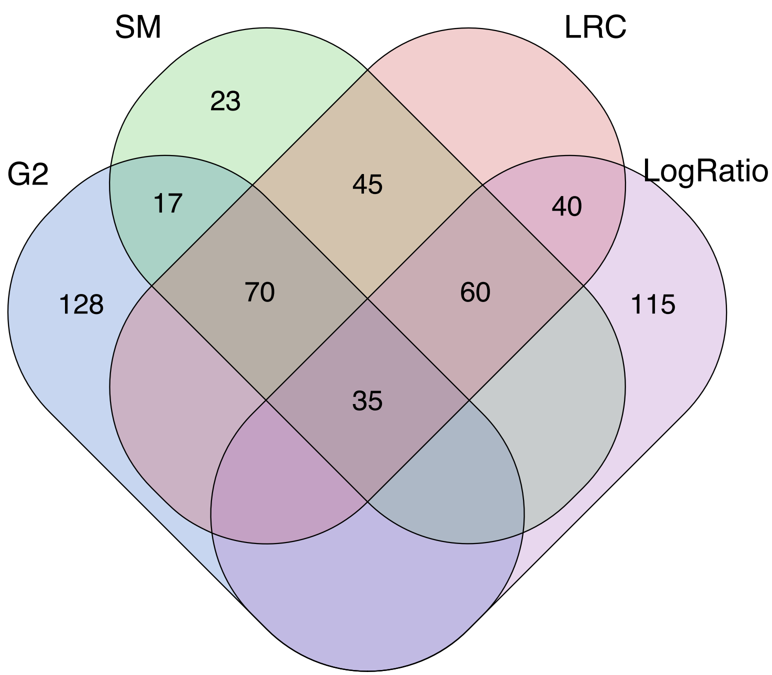
$$\lambda_1 = n_1 \pi_1, \quad \lambda_2 = n_2 \pi_2$$

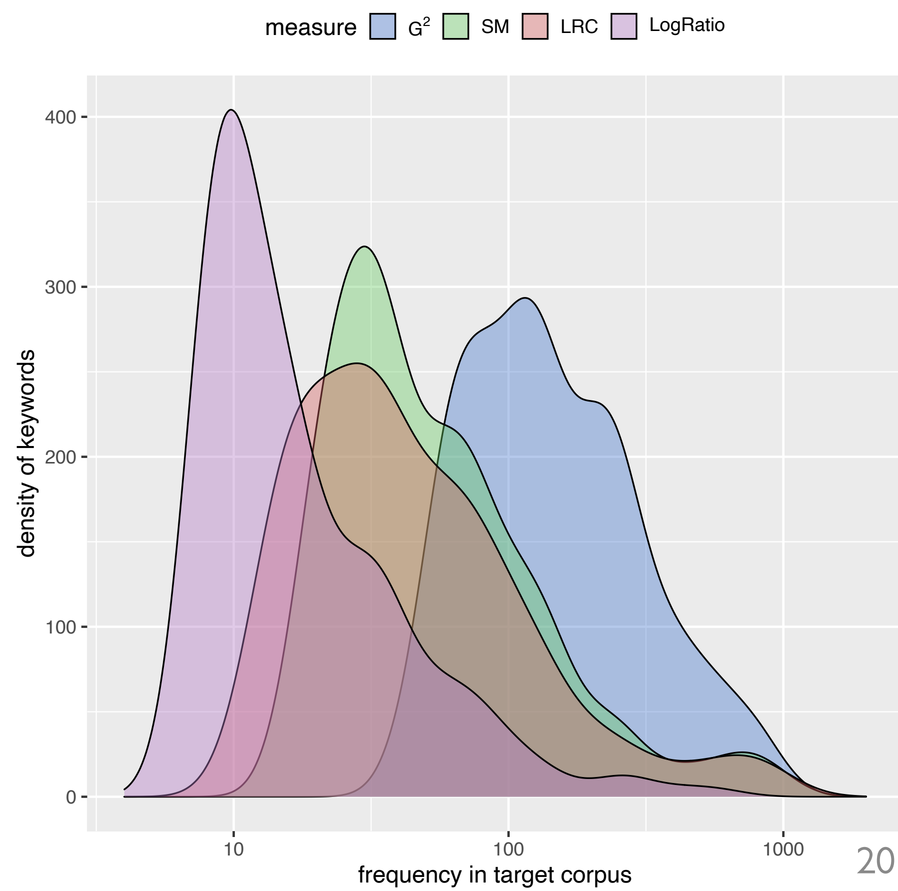
- Two-sided confidence interval
  - with Bonferroni correction
  - LRC = 0 if not significant
  - LRC > 0 ➞ significant pos. KW
  - LRC < 0 ➞ significant neg. KW

	A	B
w	f_1	f_2
$\neg w$	$n_1 - f_1$	$n_2 - f_2$
	$= n_1$	$= n_2$

Comparison

- Based on candidate data from Evert et al. (2018)
- Top-250 keywords from each measure





How well does it work in practice?

EVALUATION

Evaluating keywords

- Key challenge: many different applications of keyness
  → different requirements and evaluation goals
- Evaluation always wrt. a specific goal (e.g. CADS)
- What to evaluate? – measures, reference corpora, …
- Primarily manual validation of KW candidates
  - occasionally evaluation against gold standard possible (e.g. for identification of domain terminology)
  - special case: keyness measures for corpus comparison (Rayson & Garside 2000) can be evaluated with known similiarity corpora (Kilgarriff & Rose 1998)

Evaluation: a case study

- 14.3M token corpus on German web data about multi-resistant pathogens (MRO) collected with BootCat (Baroni & Bernardini 2004)
  - 9,750 texts of varying genres and lengths
- Target corpus: 1.3M tokens (1,177 texts) of mass media texts and reader comments from MRO corpus
- Evaluation of different keyword extraction techniques for CADS analysis of MRO discourses (Evert et al. 2018)

Evaluation: a case study

- Three keyness measures: G^2, LogRatio, LRC
- Two comparable reference corpora: *Süddeutsche* (SZ) vs. *Frankfurter Allgemeine* (FAZ)
- Keywords based on raw frequency (classic) vs. document frequency (df-based)

- Extract top-200 keywords for each technique
  - frequency threshold $f \geq 5$ in reference corpus, because we are not interested in terminology extraction
- Manual annotation of TPs (categories, evaluative)
  - pre-determined category scheme from qualitative study

Annotation procedure

MRSA: Traditional Keywords (iteration #2) [mrsa]

9 / 29 Go << >> missing LABEL2 for entry #178 set to eval: neg [undo] [export] back to main page

161	Furunkel	other	other	other	---	Symptome	Set
162	Gastmeier	actor: science	actor: science	actor: science	---		Set
163	Gatermann	actor: science	actor: science	actor: science	---		Set
164	Gebietsgrenze	top gen: spread	top gen: spread	top gen: spread	---		Set
165	Gefahr	unclear	unclear	unclear	eval: neg		Set
166	gefährlich	unclear	unclear	unclear	eval: neg		Set
167	Geflügelfleisch	top cause: animals	top cause: animals	top cause: animals	---		Set
168	Geflügelmast	top cause: animals	top cause: animals	top cause: animals	---		Set
169	gelangen	top gen: spread	top gen: spread	top gen: spread	---		Set
170	Gen	top gen: evolution	top gen: evolution	top gen: evolution	---		Set
171	Geno	actor: hospital	actor: hospital	actor: hospital	---		Set
172	Gentransfer	top gen: evolution	top gen: evolution	top gen: evolution	---		Set
173	geschwächt	unclear	unclear	unclear	eval: neg		Set
174	gescreent	top soln: hospital	top soln: hospital	top soln: hospital	---		Set
175	gesund	unclear	unclear	unclear	eval: pos		Set
176	Gesundheit	unclear	unclear	unclear	eval: pos		Set
177	Gesundheitsamt	actor: polit	actor: polit	actor: polit	---		Set
178	Gesundheitskris			top gen: spread	eval: neg		Set
179	Gesundheitssenator			---	---		Set
180	Gesundheitssenatorin	actor: polit	actor: polit	actor: polit	---		Set

Sie isolierten von beiden Immunzellen (Makrophagen , **Fresszellen**) - und brachten sie mit Bakterien und Viren in Kontakt .

Afro-Fresszellen fressen rascher Das im Fachmagazin Cell veröffentlichte Ergebnis : Die **Fresszellen** der Amerikaner afrikanischen Ursprungs killten die Bakterien drei Mal so rasch wie die Fresszellen der Amerikaner europäischen Ursprungs .

Afro-Fresszellen fressen rascher Das im Fachmagazin Cell veröffentlichte Ergebnis : Die Fresszellen der Amerikaner afrikanischen Ursprungs killten die Bakterien drei Mal so rasch wie die **Fresszellen** der Amerikaner europäischen Ursprungs .

Die können angeblich für jedes Bakterium ein **Fresszelle** herstellen .

Dann gelingt es ihnen leicht , die körpereigenen **Fresszellen** , die eigentlich für die Abwehr der Eindringlinge zuständig sind , zu zerstören , um sich dann ungehindert auszubreiten .

Als Antibiotikaersatz taugen sie bisher nicht , weil sie im menschlichen Immunsystem schnell von **Fresszellen** verspeist werden .

Man geht konventionellerweise davon aus , daß die **Fresszellen** des Immunsystems die Bakterien dann beseitigen . chen-men 16. 11. 2015 24. Noch manche Krankheit wird als Bakterien-Folge erkannt werden Dazu eine hochinteressante Information .

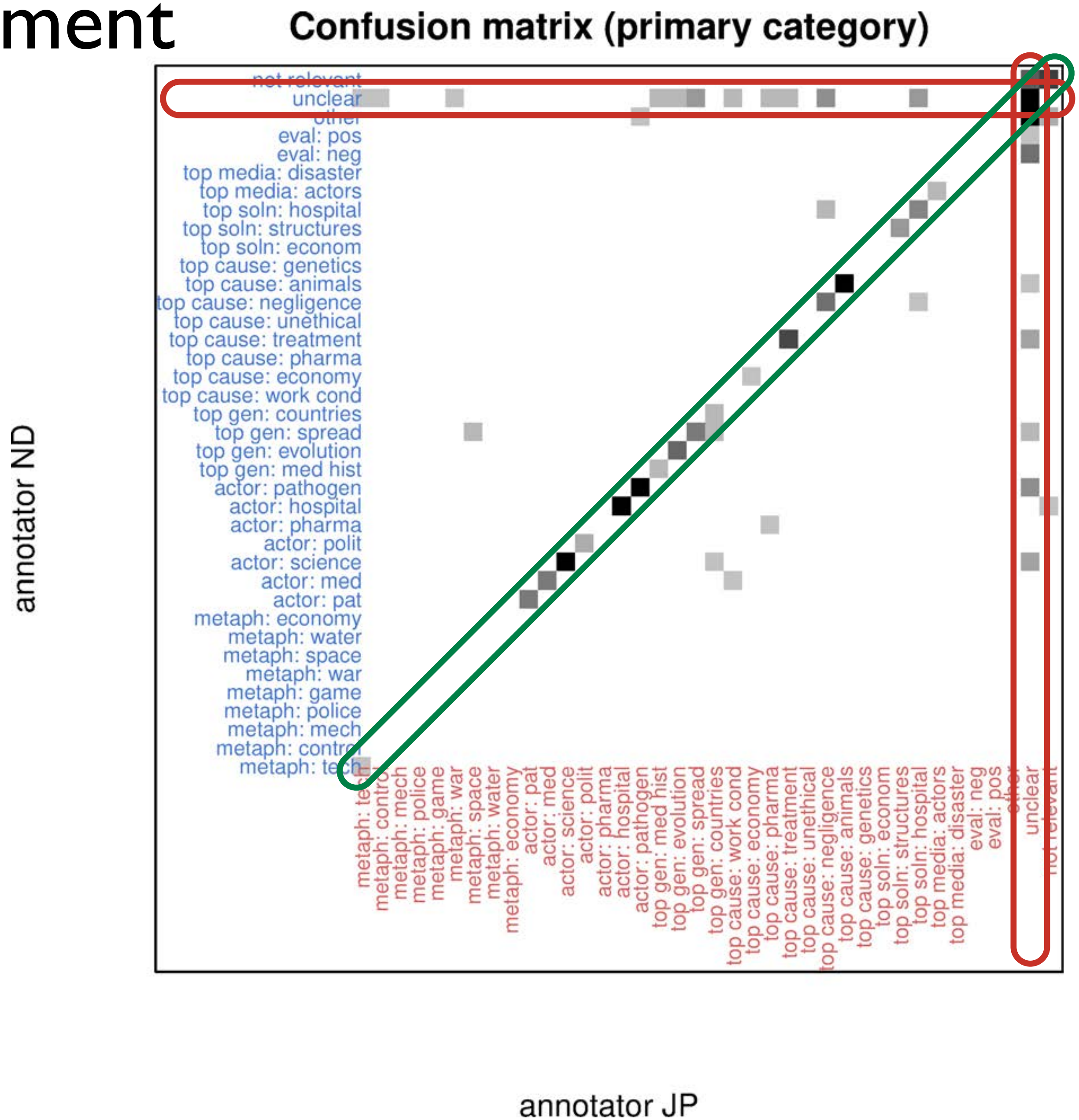
Im Übrigen sind die von Ihnen benannten " **Fresszellen** " immer Bestandteil der Immunantwort , egal ob mit Antibiotikum oder ohne .

Web-based annotation platform MiniMarker

Agreement

- Two independent annotators
- Agreement of 82.2% on distinction TP vs. FP (but Cohen κ = .566 fairly low)
- Domain-specific, highly frequent words often marked "unclear" (FP) by one annotator and TP by the other
- Disagreements between TP categories less frequent; mostly due to overlap between discourse levels
  - metaphors as part of topoi
  - intertwined argumentational levels
- Final gold standard jointly reconciled by annotators

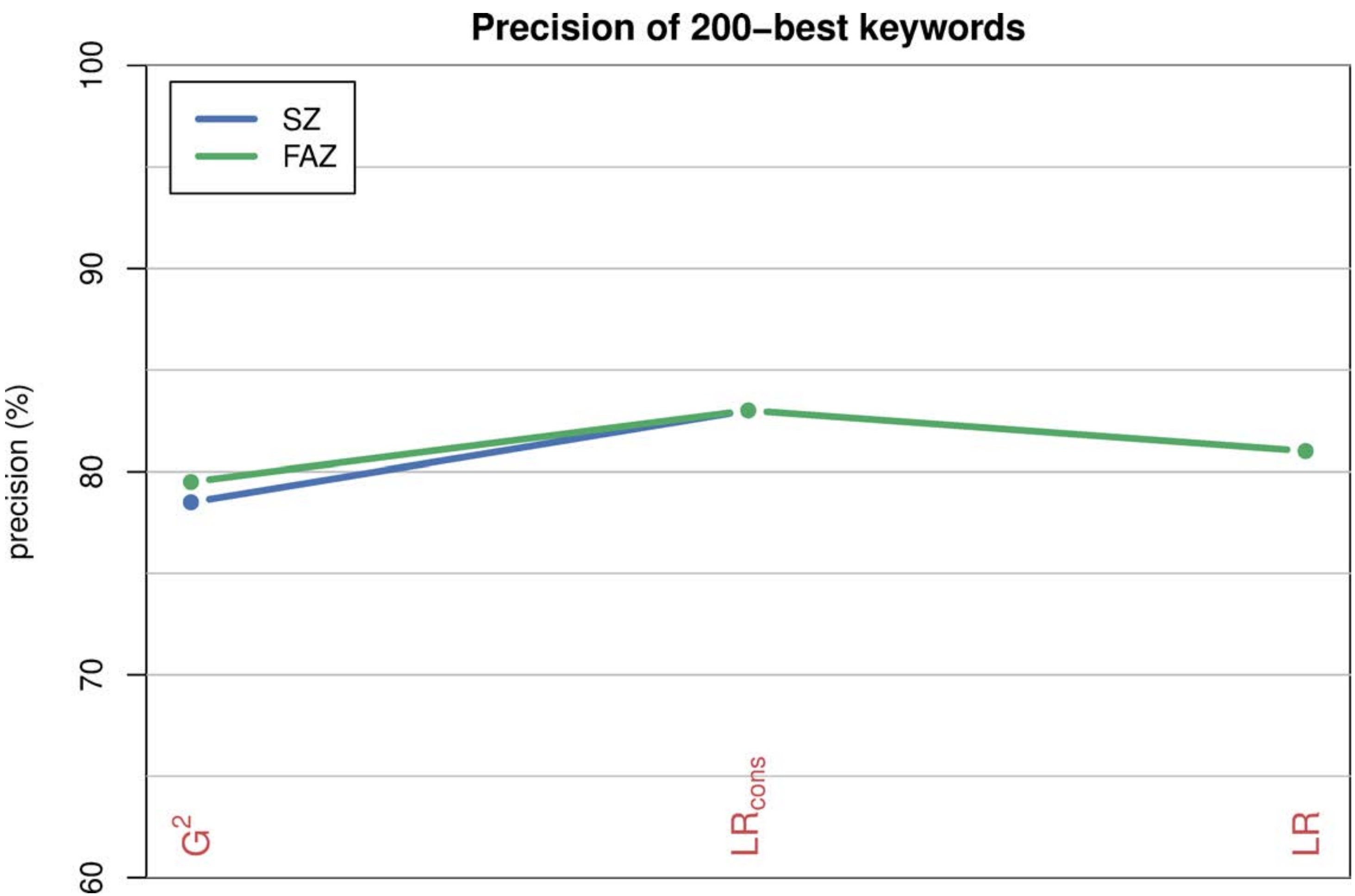
Agreement



Confusion matrix (primary category)
annotator ND
annotator JP
not relevant
unclear
other
eval: pos
eval: neg
top media: disaster
top media: actors
top soln: hospital
top soln: structures
top soln: econom
top cause: genetics
top cause: animals
top cause: negligence
top cause: unethical
top cause: treatment
top cause: pharma
top cause: economy
top cause: work cond
top gen: countries
top gen: spread
top gen: evolution
top gen: med hist
actor: pathogen
actor: hospital
actor: pharma
actor: polit
actor: science
actor: med
actor: pat
metaph: economy
metaph: water
metaph: space
metaph: war
metaph: game
metaph: police
metaph: mech
metaph: contr
metaph: tech

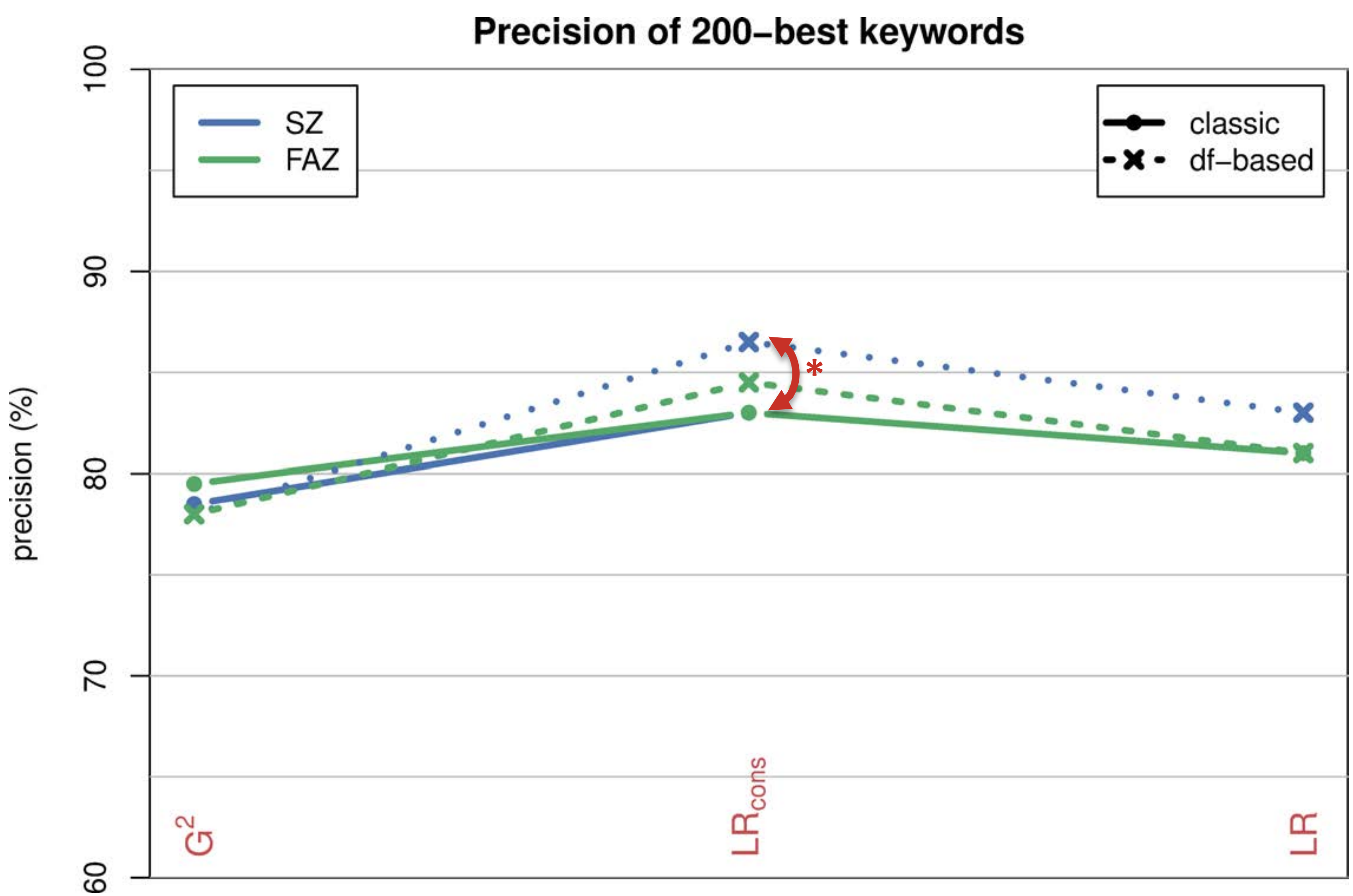
Precision = #TP / 200 candidates

TP = assigned to category and/or evaluative

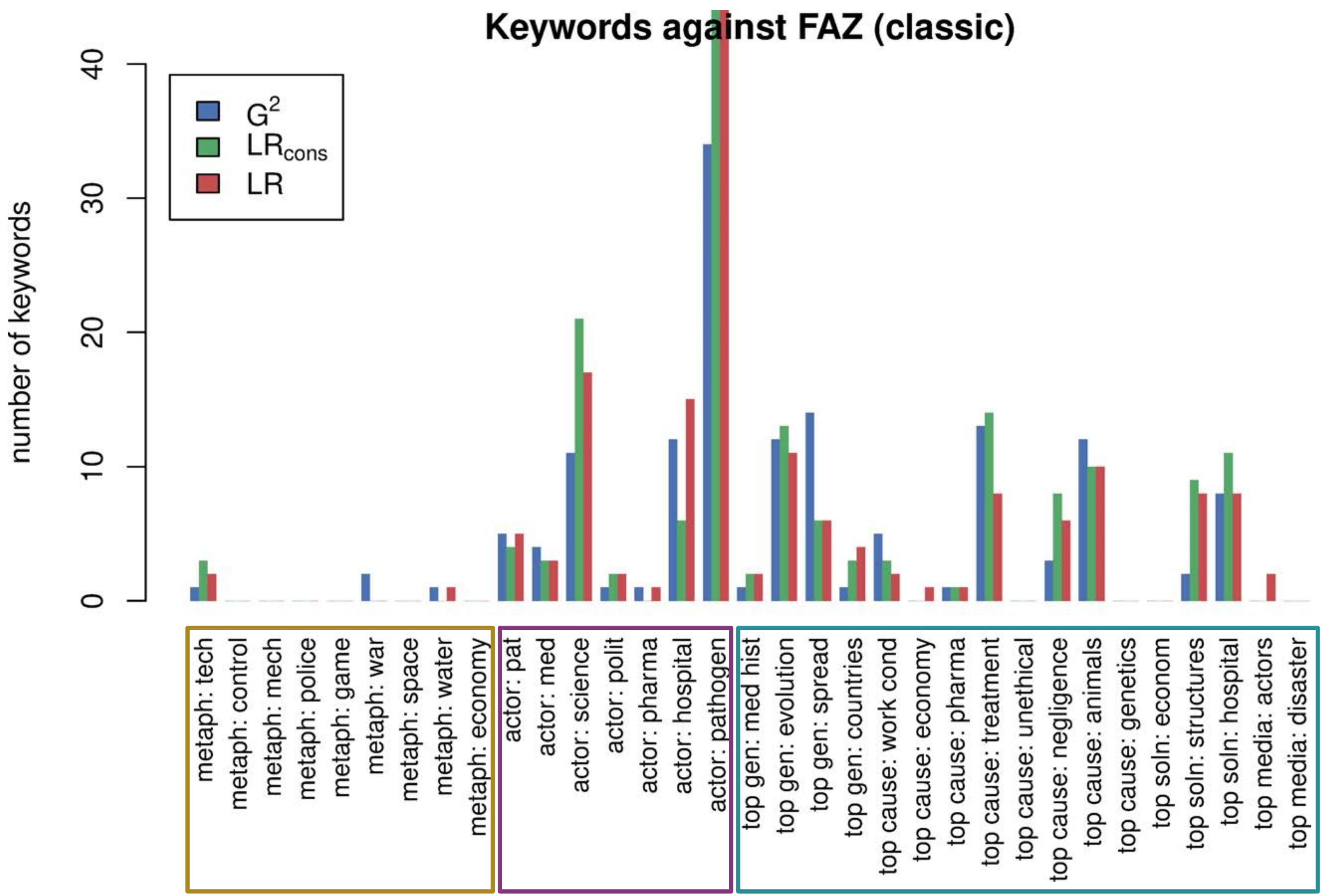


Precision = #TP / 200 candidates

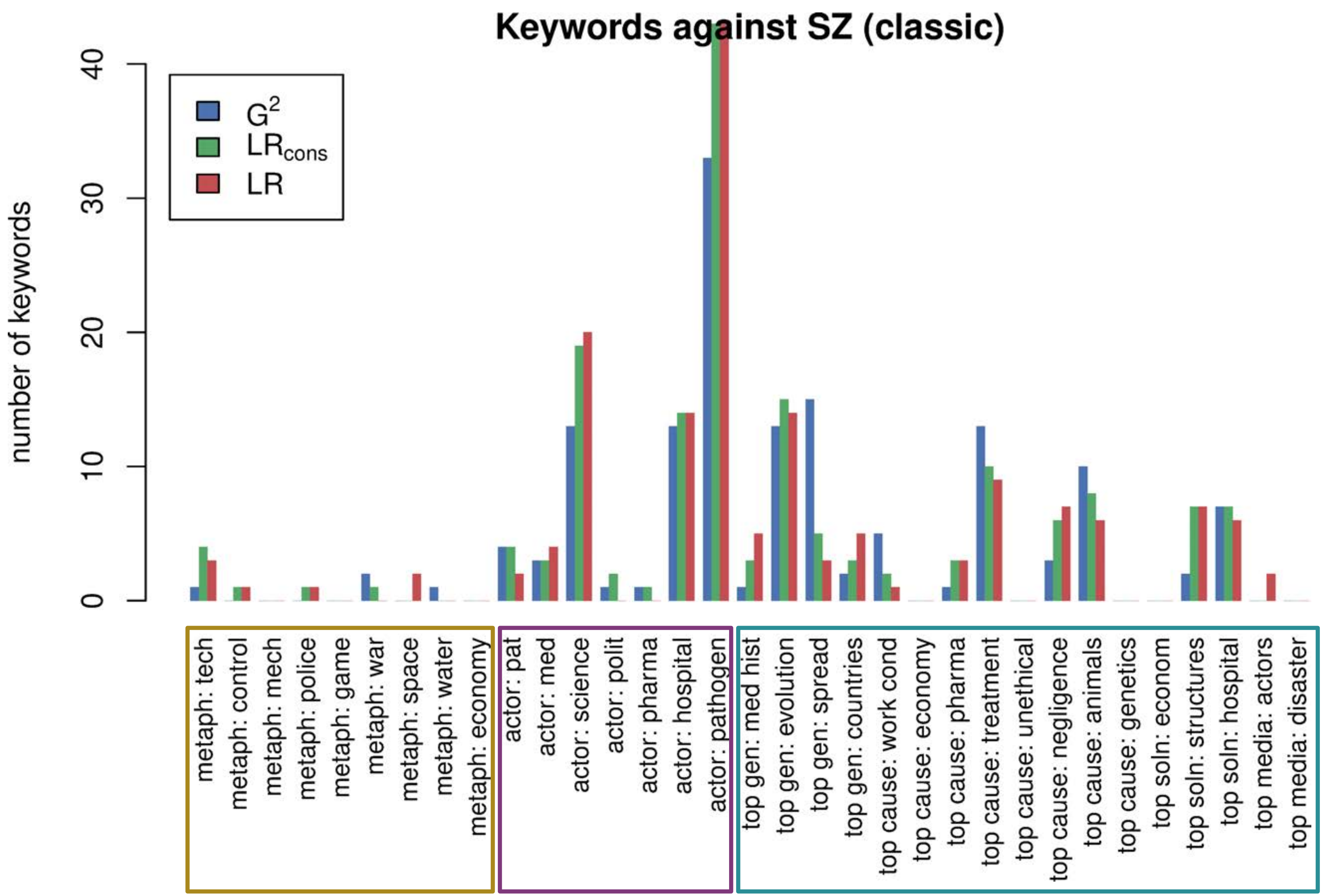
TP = assigned to category and/or evaluative



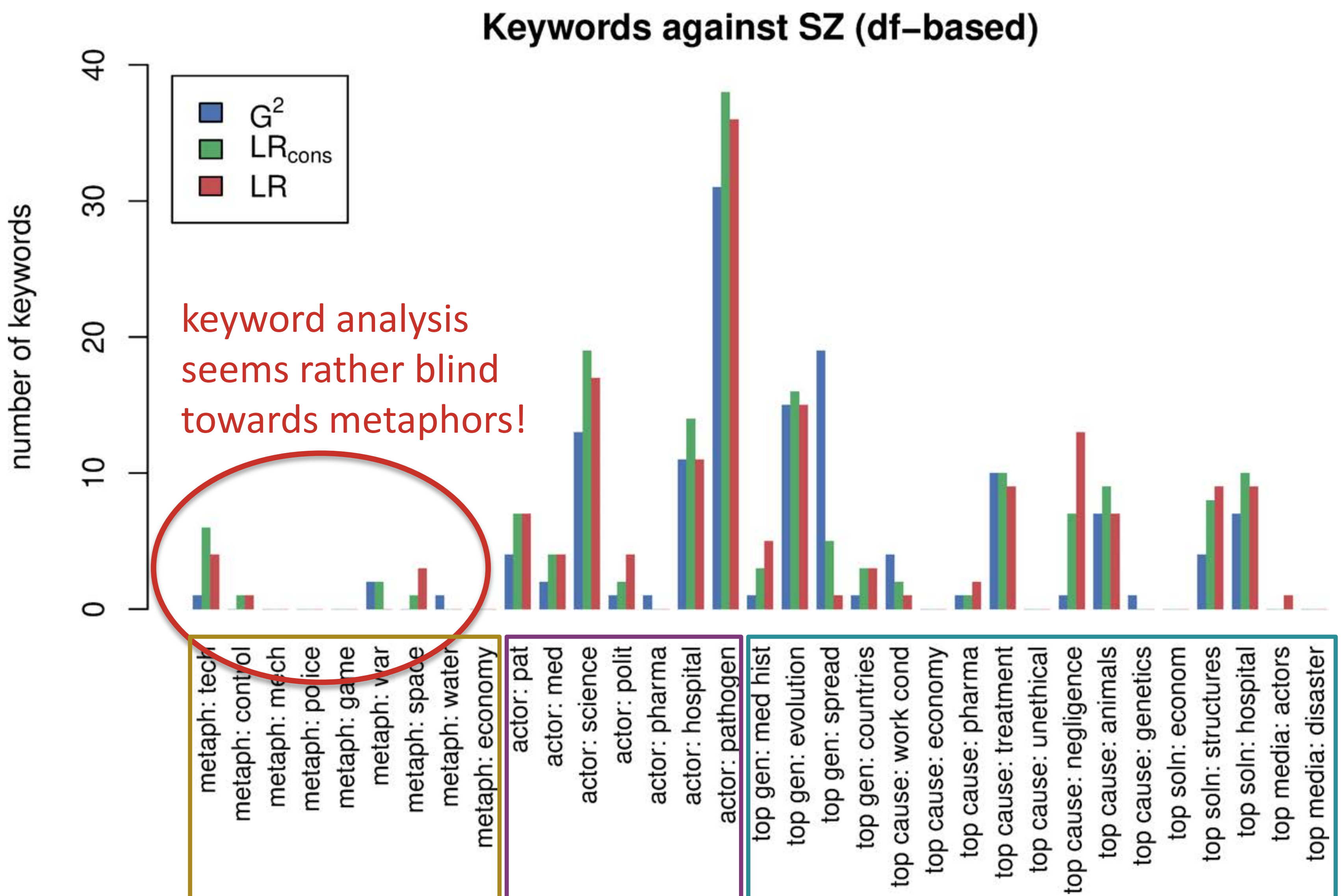
Recall = #KW for each category



Recall = #KW for each category



Recall = #KW for each category



A few quiz questions

- Which is the best keyness measure?
- What impact has the choice of reference corpus?
- How many keywords should you look at?
- Should you only consider significant keywords? Why?
- What's the best way of reading a keyword list?
  Ranked by keyness? Alphabetical? Word cloud? …
- What is "keyness" really?
- What are limitations of keyword analysis?

NB: None of these questions has a clear-cut answer!

Interactive session

COMPUTING KEYWORDS WITH R

What you will need

- R from https://cran.r-project.org
- RStudio from https://posit.co/downloads/
- R packages (install via RStudio)
  - `tidyverse` (to manipulate frequency lists)
  - `corpora` version 0.6 (or newer)
  - `Rtsne`, `ggrepel` (for a really cool visualisation)
  - `fastTextR` (to apply this visualisation to your own data)

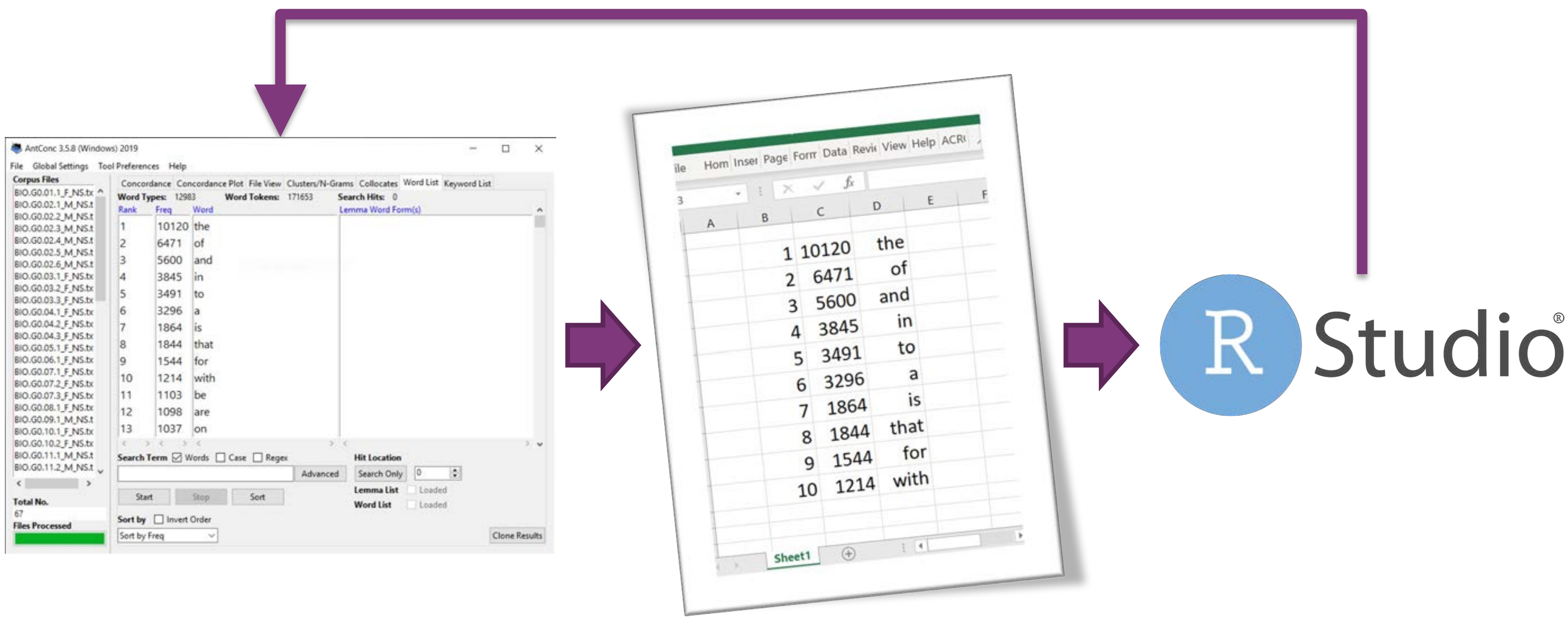
- RStudio project with data sets & worked example
  - provided as ZIP archive `04_keyness_hands_on.zip`

Interoperability

- At least three steps in a keyword analysis
  - pre-processing & linguistic annotation of corpora **A** and **B**
  - extraction of frequency data (optionally with filters, df counts, dispersion-adjusted frequencies, etc.)
  - statistical analysis ➞ keyness measures & beyond
  - optional 4th step: visualisation (scattertext, semantic map, …)
- Many end-user tools integrate all three steps (CQPweb, AntConc, WordSmith)
- … but better to use specialised state-of-the art tools for each step (in particular, R for statistical analysis)

Interoperability with tabular data

- Tabular data in MTSV format (Anthony & Evert 2019)
  - data set = collection of TAB-delimited tables
    - word frequencies, positional data (for dispersion), kwic, …
  - important: link back from statistical analysis to corpus



MTSV for keywords

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

target

freqlist

type	frequency	_reference
the	2	{"corpus": "xyz", "search": "the", "case": "0", ...}
cat	1	{"corpus": "xyz", "search": "cat", "case": "0", ...}
mat	1	{"corpus": "xyz", "search": "mat", "case": "0", ...}
on	1	{"corpus": "xyz", "search": "on", "case": "0", ...}
sat	1	{"corpus": "xyz", "search": "sat", "case": "0", ...}

meta

_semantic_model	size	types	case	kind	threshold	comments
type_frequency_list	7	6	lower	token_counts	1	Target corpus frequency list

reference

freqlist

type	frequency	_reference
the	23383	{"corpus": "xyz", "search": "the", "case": "0", ...}
cat	0	{"corpus": "xyz", "search": "cat", "case": "0", ...}
mat	282	{"corpus": "xyz", "search": "mat", "case": "0", ...}
on	2582	{"corpus": "xyz", "search": "on", "case": "0", ...}
sat	7892	{"corpus": "xyz", "search": "sat", "case": "0", ...}

meta

_semantic_model	size	types	case	kind	threshold	comments
type_frequency_list	19238145	8293	lower	token_counts	1	Reference corpus frequency list

Tabular data in practice

- Little support for MTSV yet, except for AntConc
- How to obtain MTSV word frequency lists:
  - open desired corpus as *Target Corpus*
  - create word frequency list (in *Word* tab)
  - select *Save Current Tab Database Tables* from menu
  - creates ZIP archive with several CSV tables

- But most tools can easily read/write tabular files: CQPweb, WordSmith, CWB, Python, R, Excel, …
  - we'll look at examples from AntConc, CWB and CQPweb

Tabular data in practice

- CSV = comma-separated values (RFC 4180)
  - https://datatracker.ietf.org/doc/html/rfc4180
  - comma-separated columns (usually), values double-quoted if necessary, data types of columns inferred from values
- TSV = TAB-delimited text files
  - columns delimited by TAB characters (ASCII 0x09, `"\t"`)
  - no quotes (values must not contain TABs or line breaks)

- Strategy: export frequency lists for corpora **A** and **B** from favourite corpus tool + note down sample sizes
  - some corpus tools create "tidier" tabular data than others

And finally …

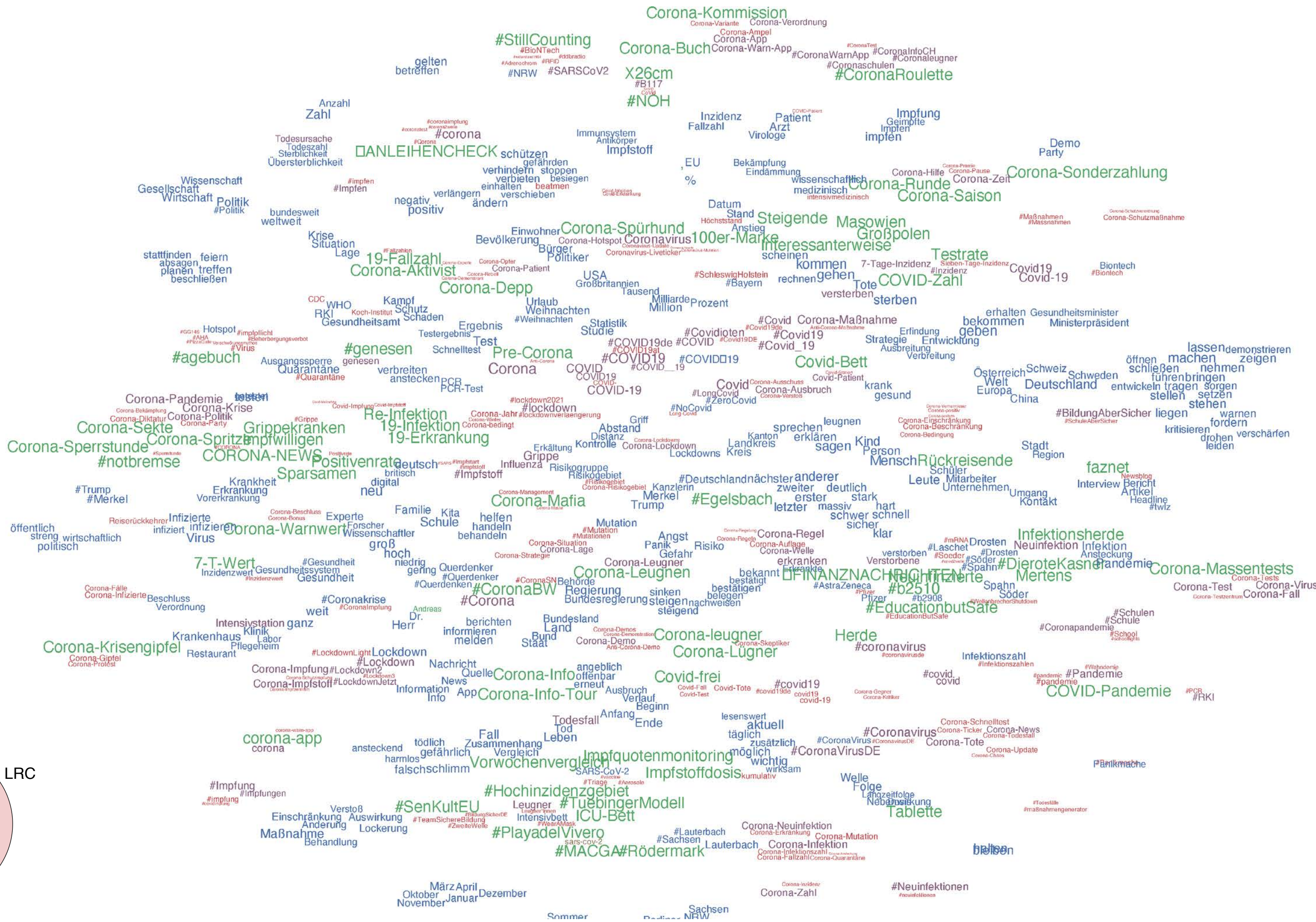
Hands on!

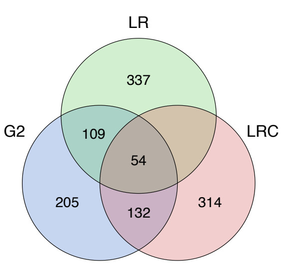
- LRC reference implementation, example data and mathematical details at https://osf.io/cy6mw/
- Implementation for end users:
  `keyness()` function in `corpora` package v0.6
- Unpack ZIP archive `keyness_hands_on.zip` then double-click the `.Rproj` file to open RStudio

Interactive session

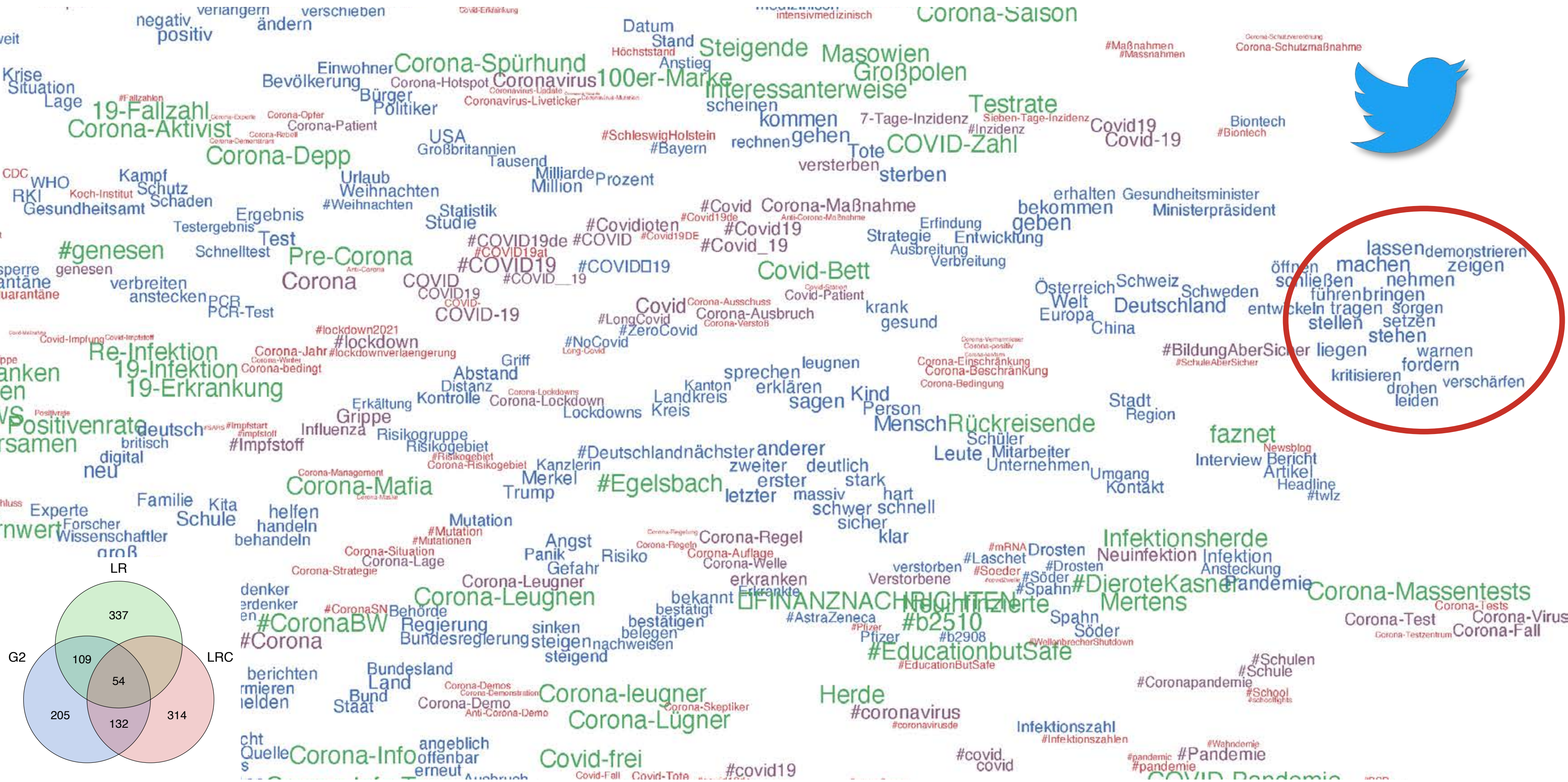
VISUALISING KEYWORDS

Visualisation as semantic map

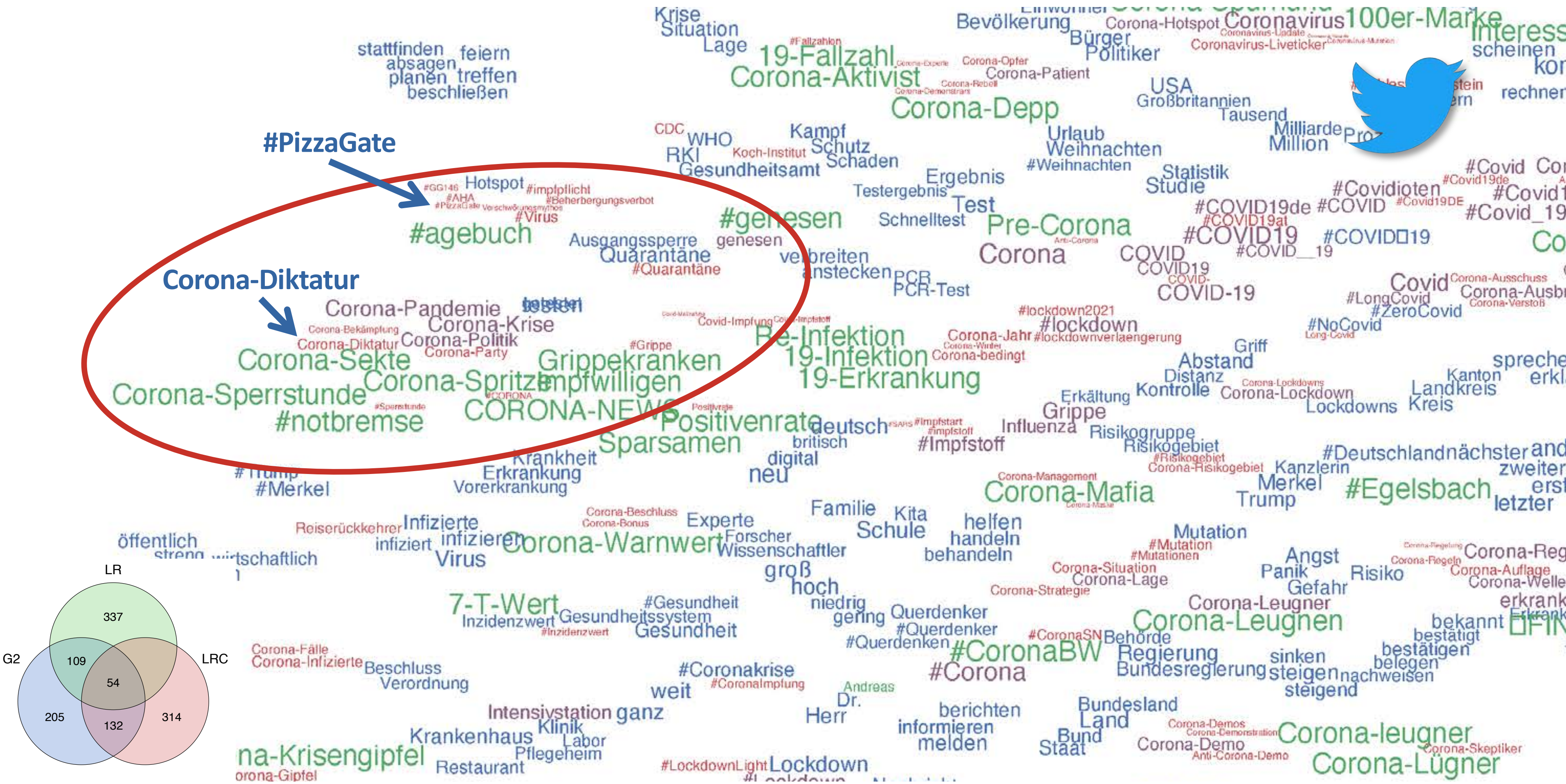




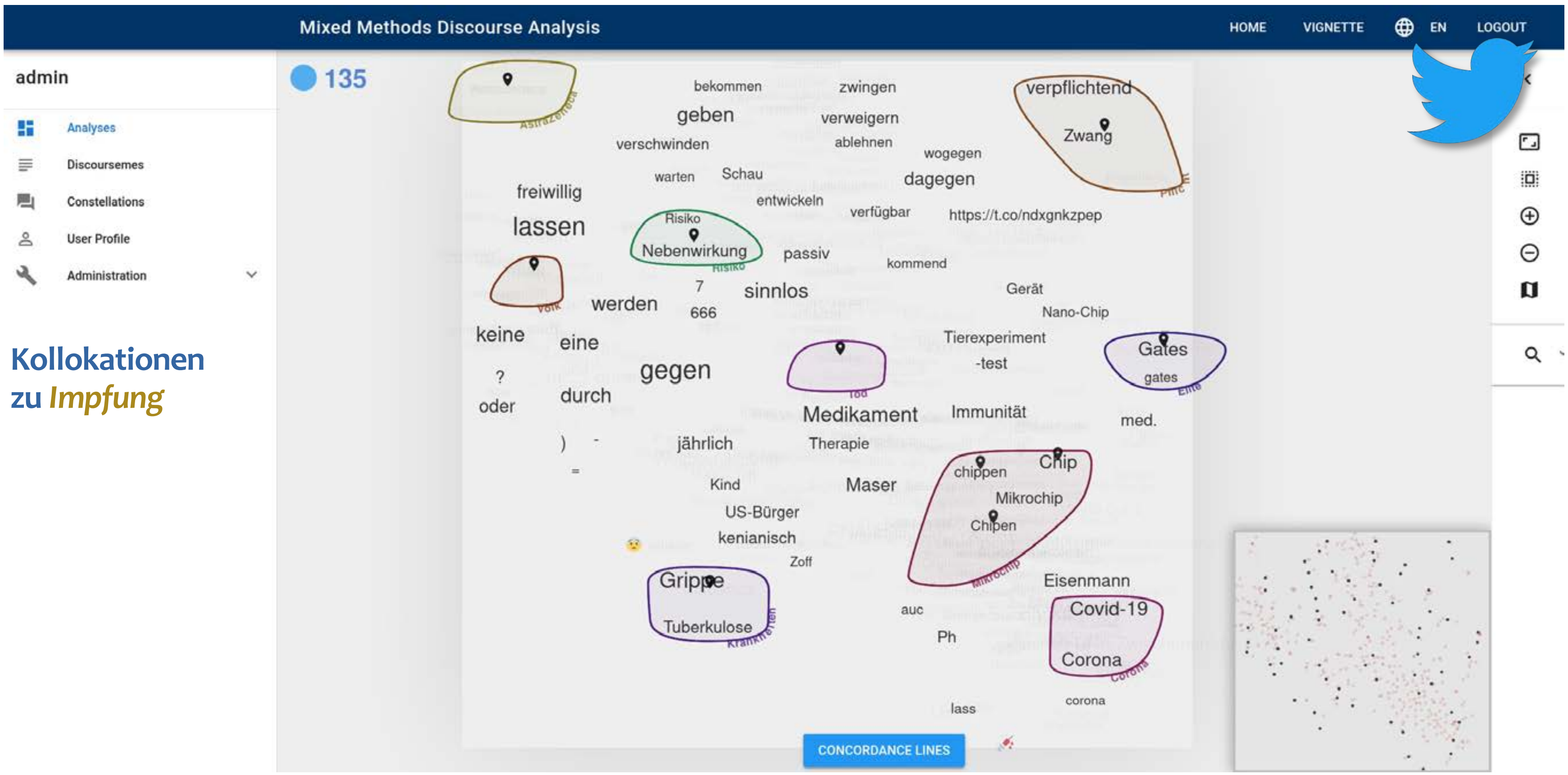
Visualisation as semantic map



Visualisation as semantic map



Interactive grouping with MMDA



Kollokationen zu *Impfung*

https://www.linguistik.phil.fau.de/projects/efe/mmda-toolkit/

EMERGING FIELDS INITIATIVE

Exploring the Fukushima Effect

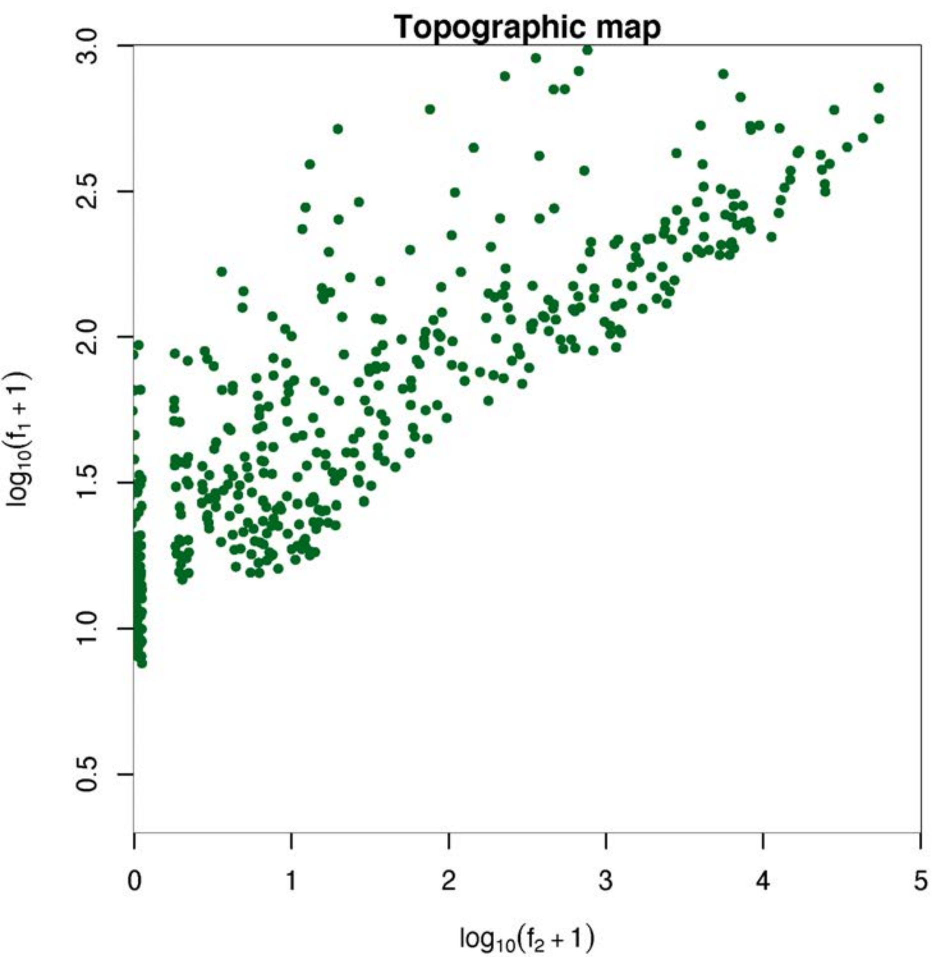
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

Interactive session

WHAT IS KEYNESS?

Understanding keyness measures



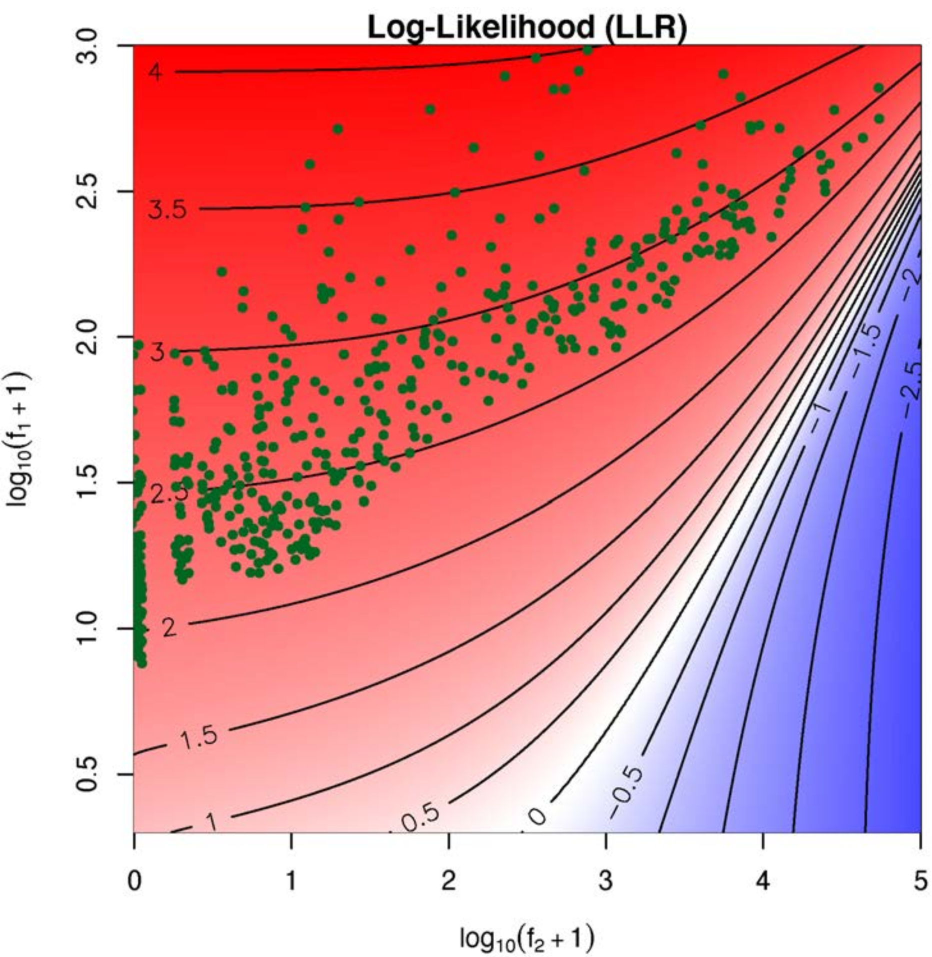
Candidates from data set of Evert et al. (2018) that are among top-250 keywords for any of several keyness measures

Topographic map visualises f_1 vs. f_2 (sufficient since n_1 and n_2 are fixed for data set) on a logarithmic scale

→ similar to ScatterText

https://spacy.io/universe/project/scattertext

Topographic maps

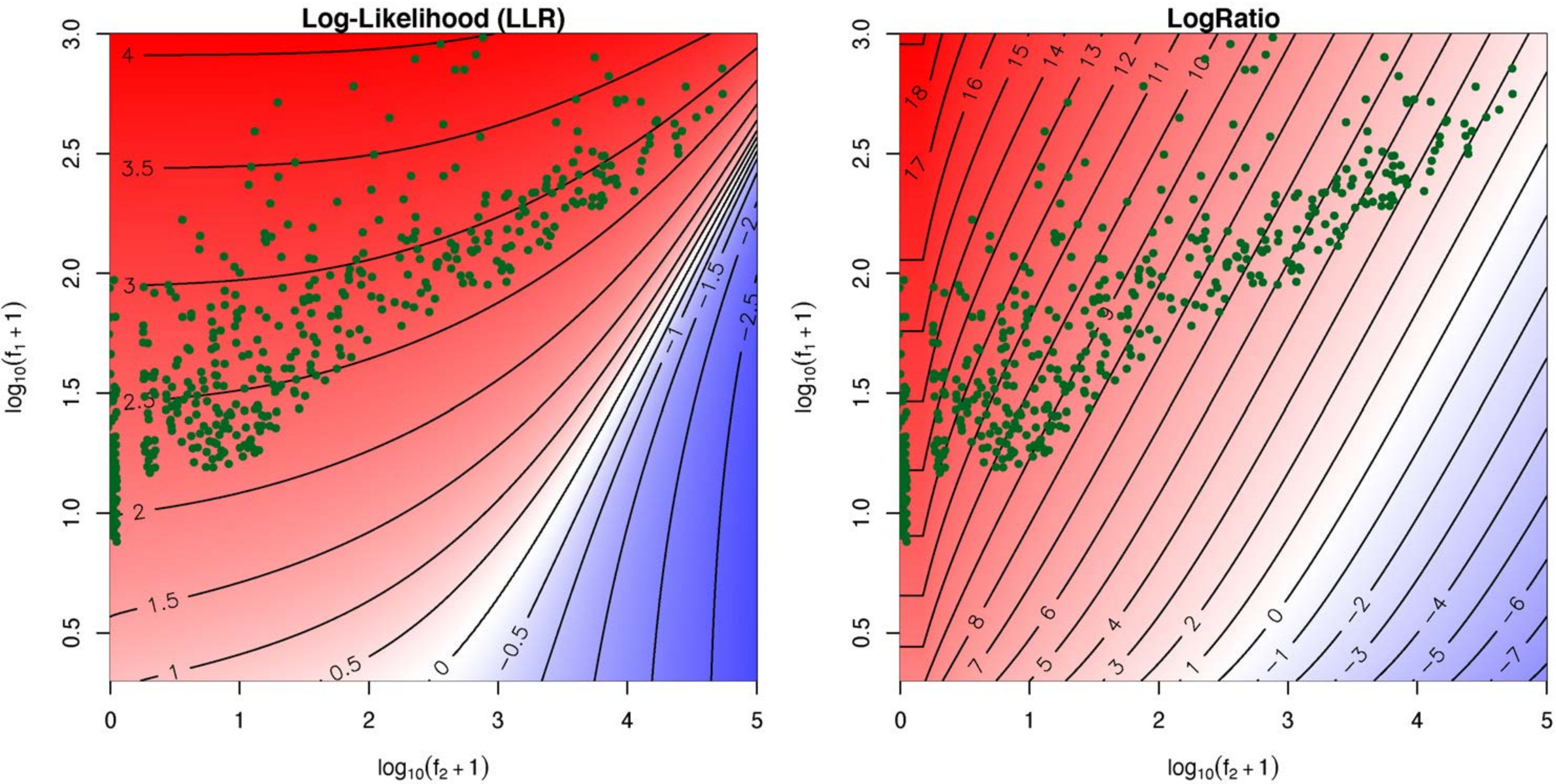


Topographic map visualises f_1 vs. f_2 (sufficient since n_1 and n_2 are fixed for data set) on a logarithmic scale

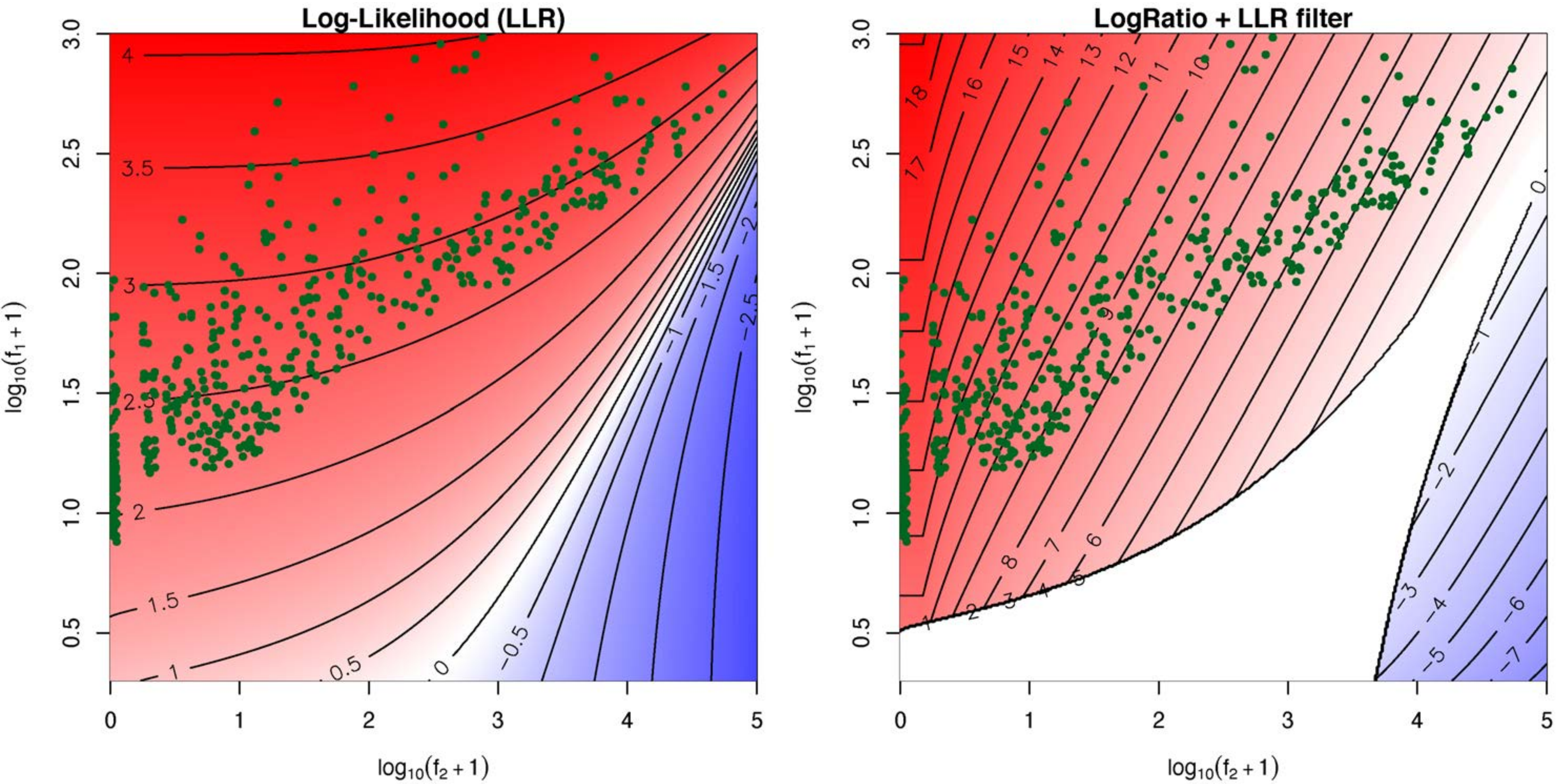
→ similar to ScatterText

https://spacy.io/universe/project/scattertext

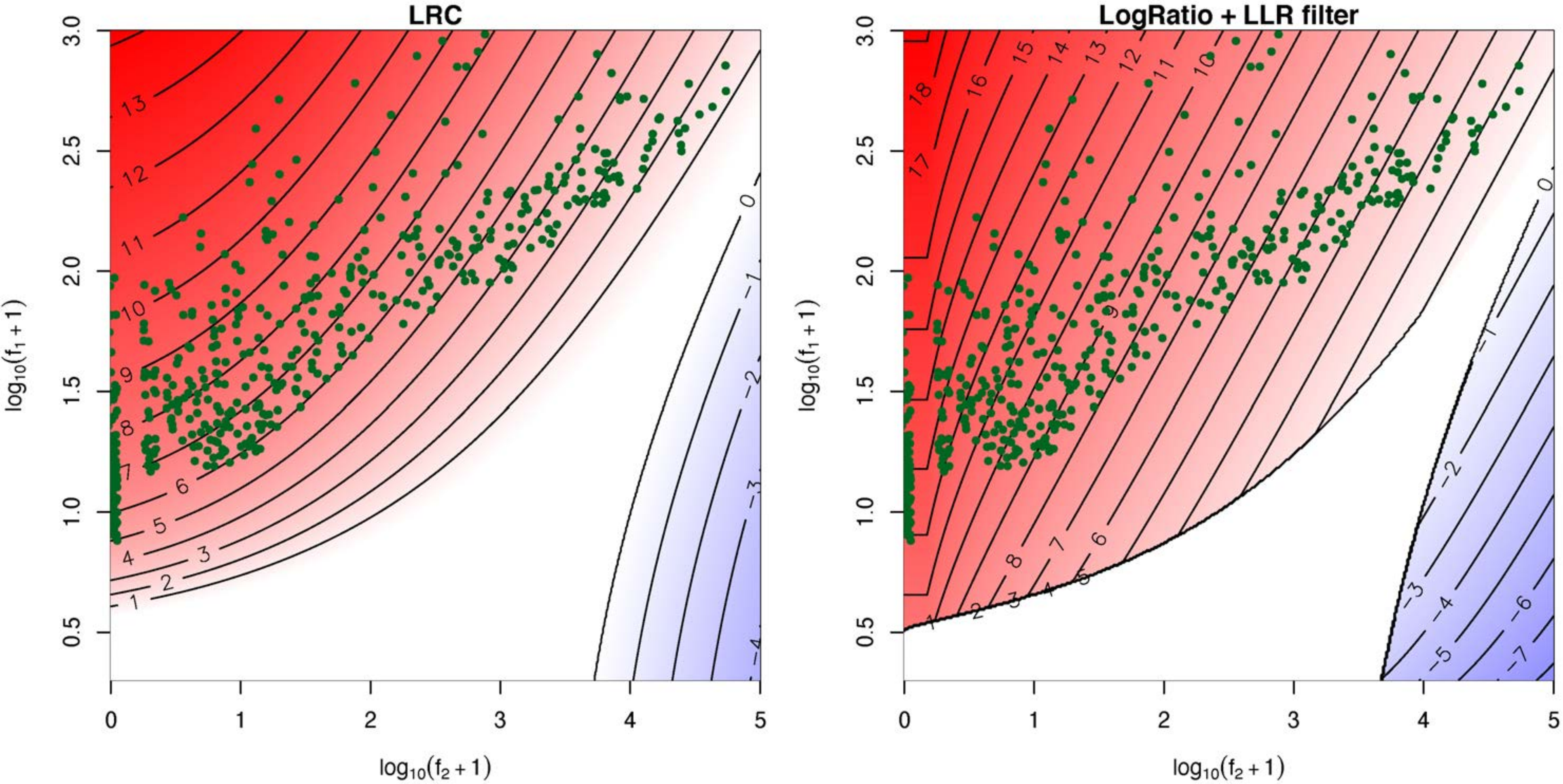
Topographic maps



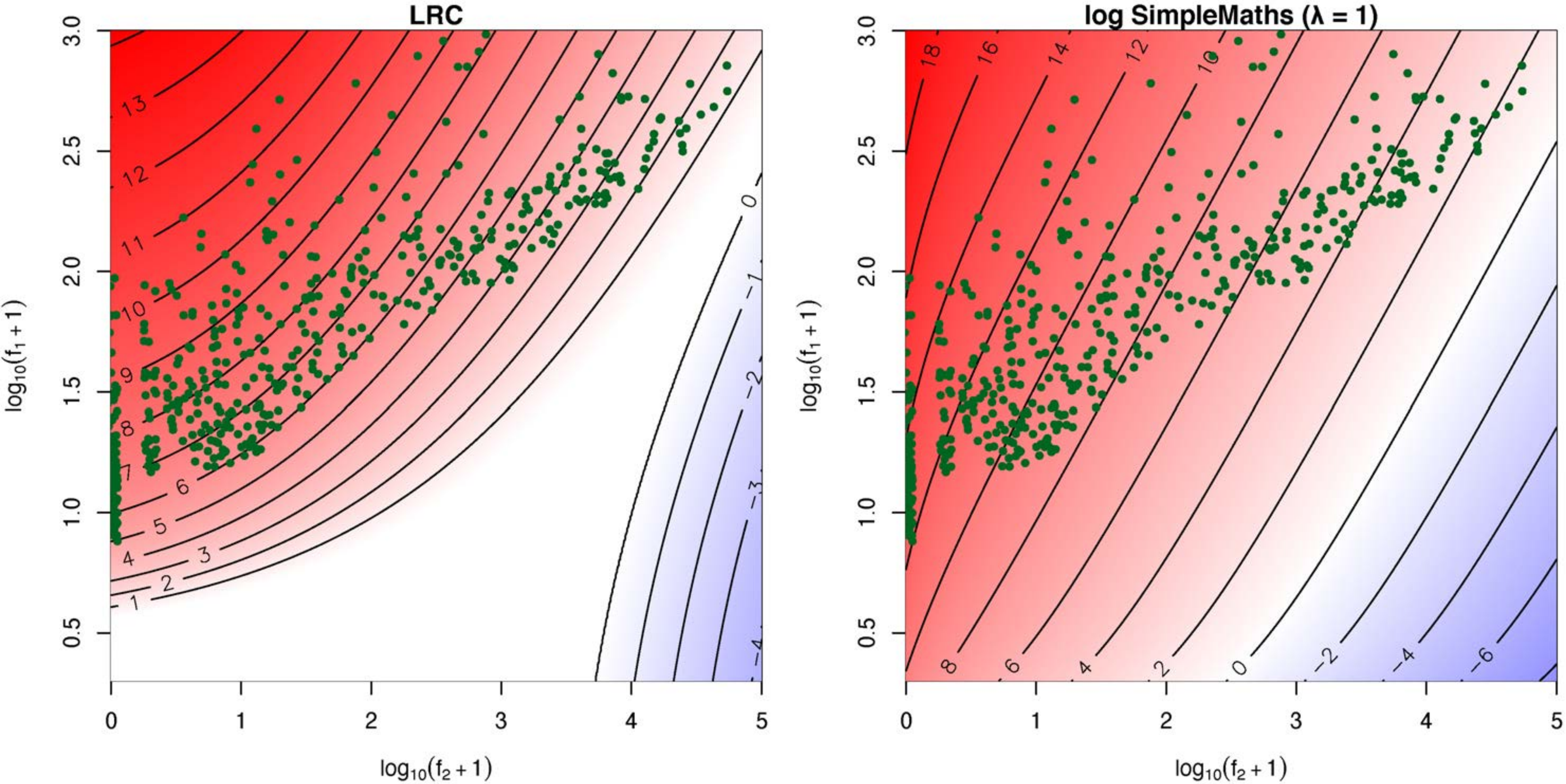
Topographic maps



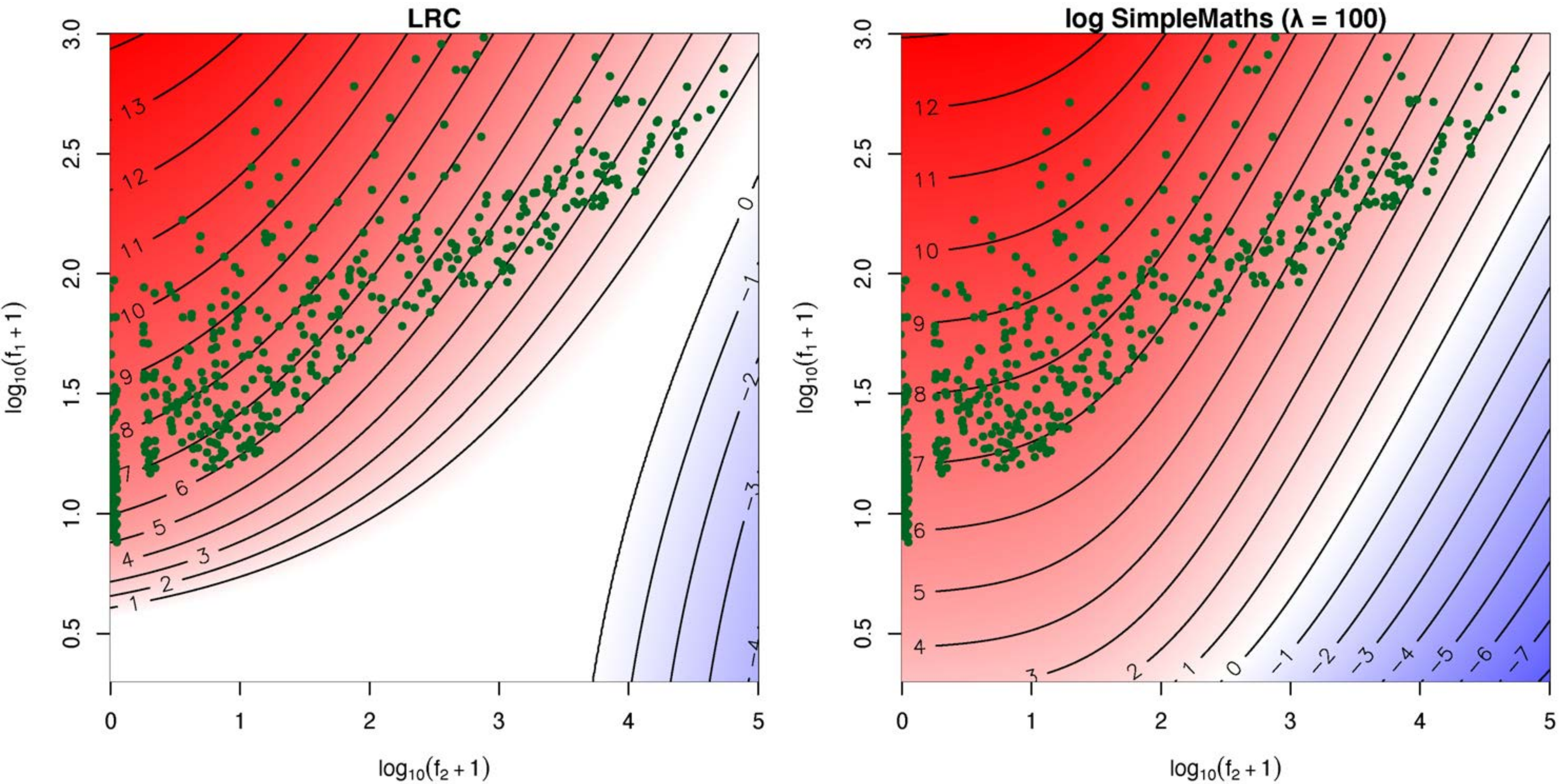
Topographic maps



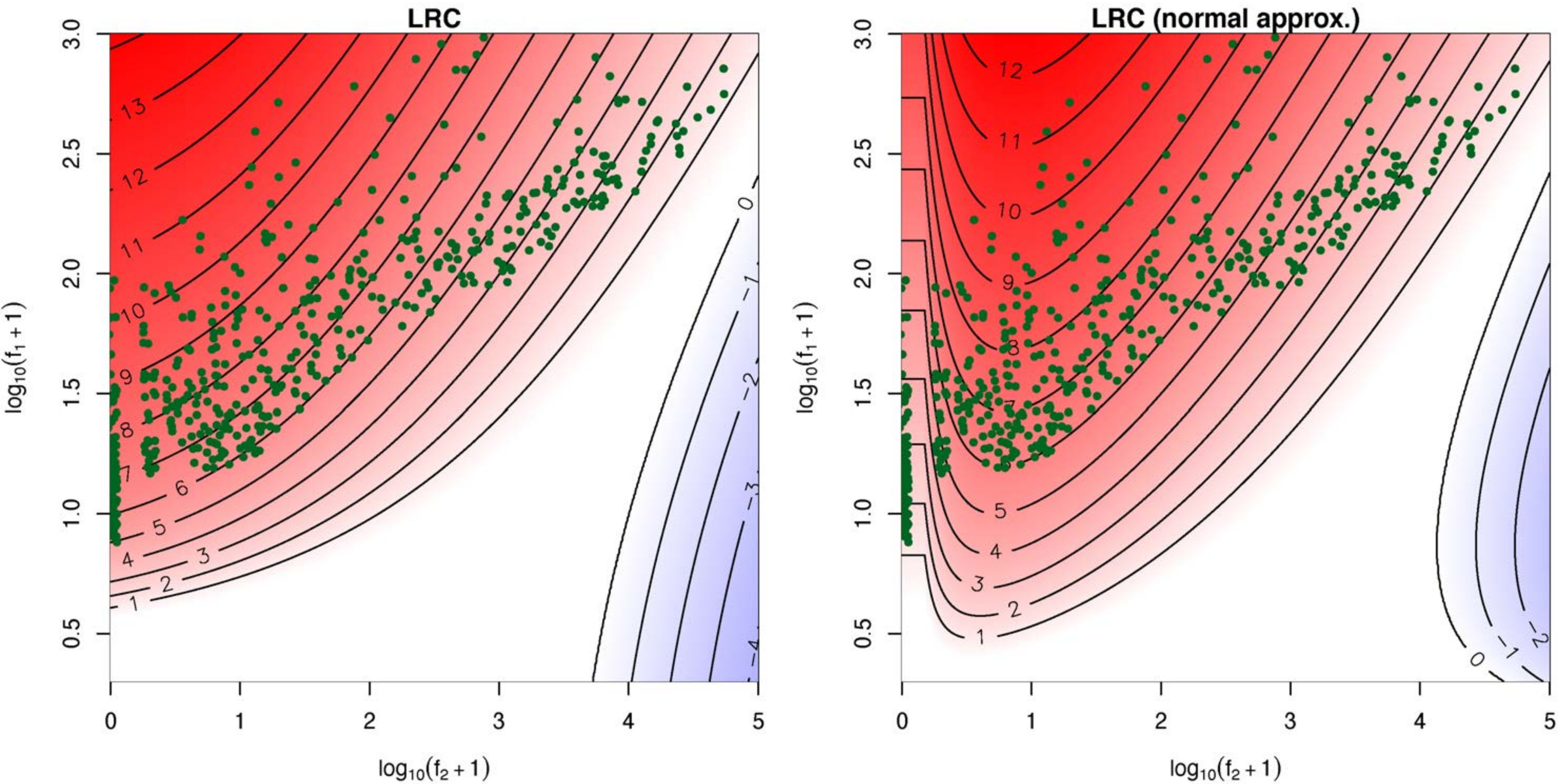
Topographic maps



Topographic maps



Topographic maps



Interactive session

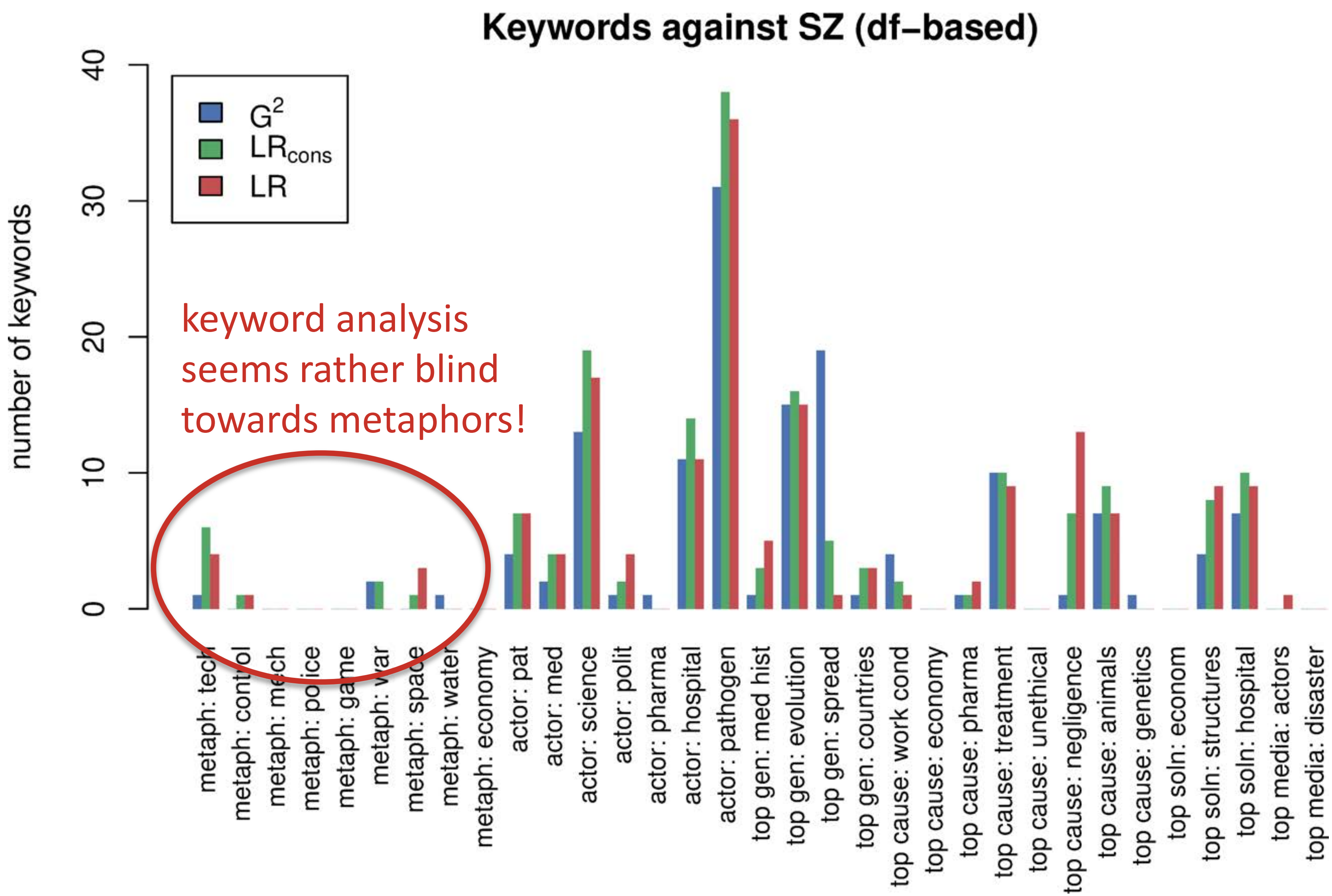
FINDING METAPHORS

Finding better keywords

- Keyness as simple frequency comparison very limited
- But more sophisticated approaches often share its limitations → still based on surface frequencies

- Certain types of keywords (terminology, topics) are easy to detect, other seem to be very challenging
- Perhaps more knowledge-rich approaches needed!
- Let's get back to case study from Evert et al. (2018)

Recall = #KW for each category



Why so few metaphor keywords?

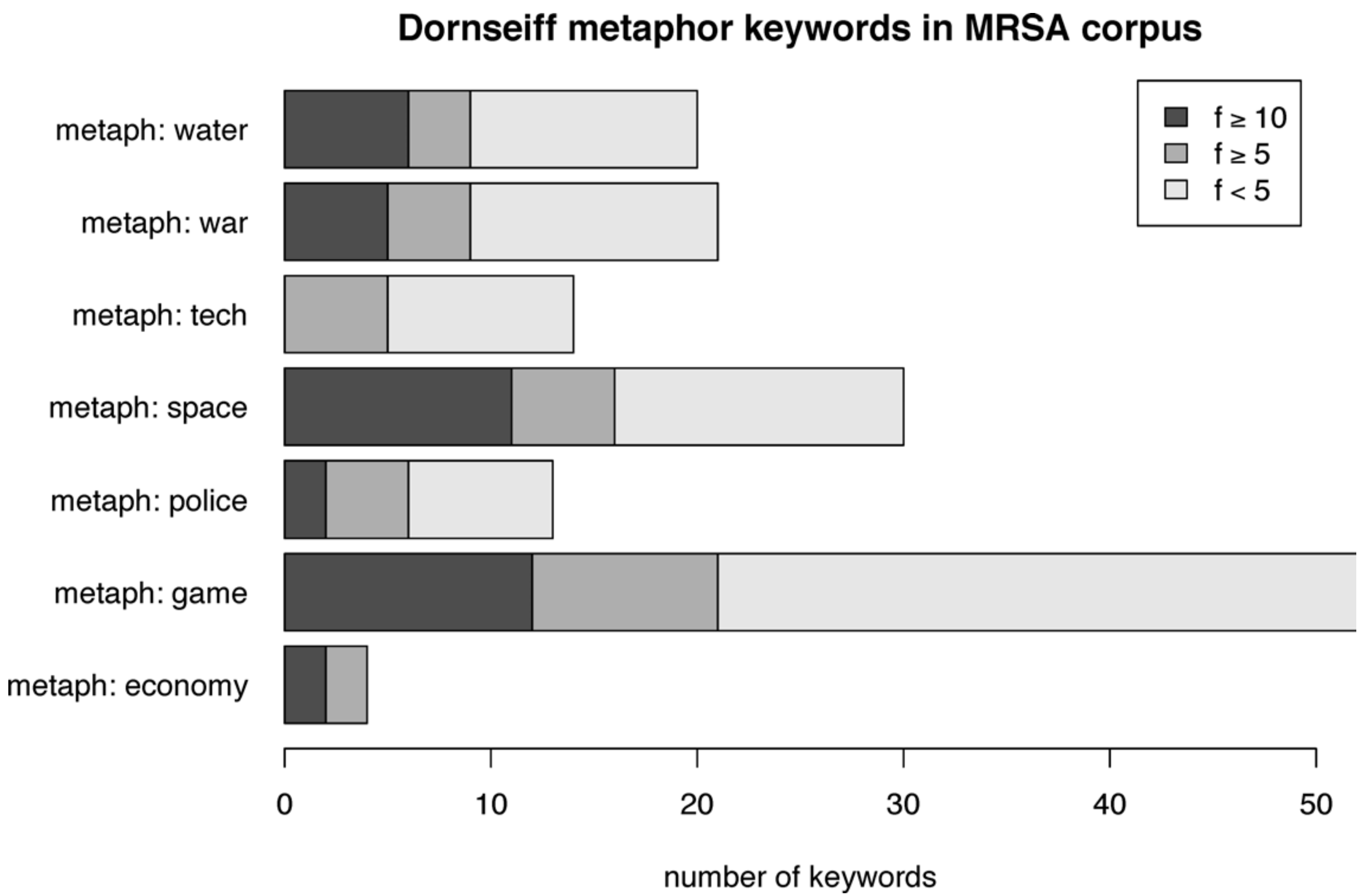
Possible causes:

- No metaphors in online media discourse (unlikely)
- Cannot be reduced to single words
- Keywords occur, but are too infrequent

A case study

- List of plausible keywords for each metaphor category from thesaurus (Dornseiff 2004)
  - e.g. POLICE: *Indiz* clue, *Killer* killer, *Mord* murder, *Täter* culprit, *fahnden* search, *heimtückisch* insidious, …
  - manually validated against concordance in target corpus
- Comparison with full set of keyword candidates
  - frequency in target corpus
  - removed because of reference corpus threshold?
  - keyness score and rank in candidate set

A case study

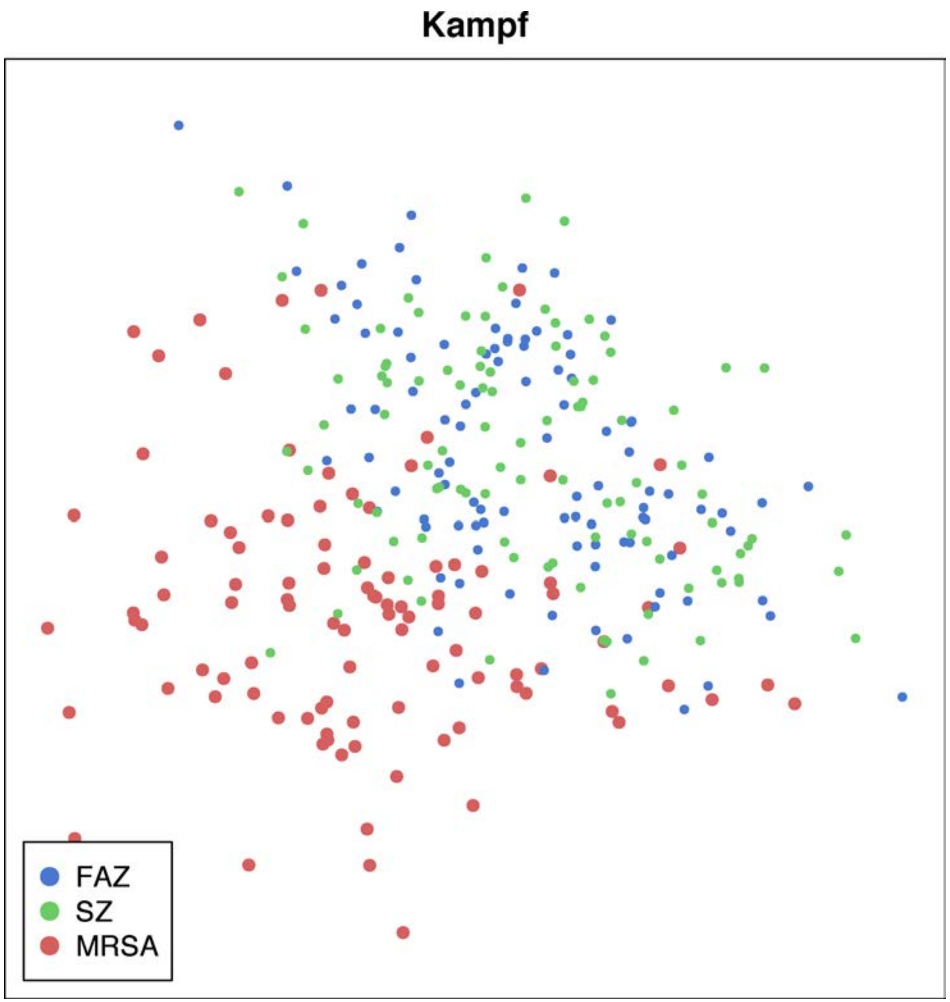


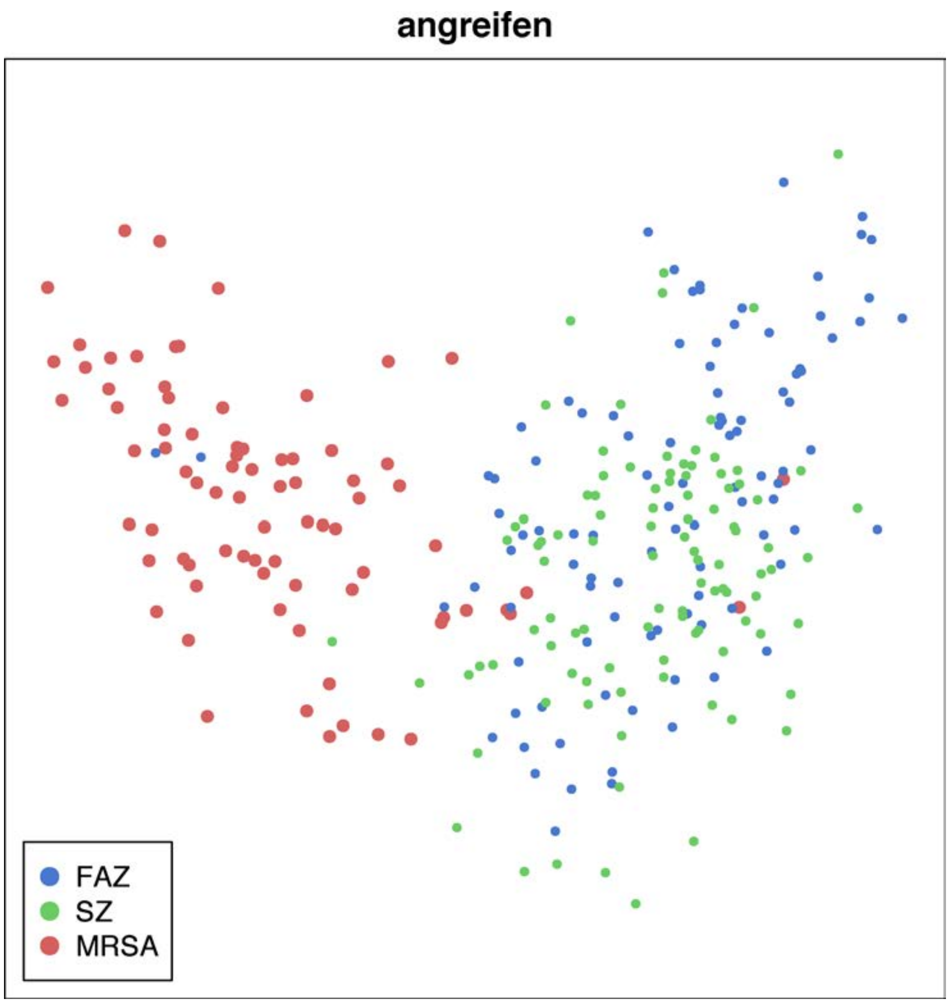
Dornseiff metaphor keywords in MRSA corpus
metaph: water
metaph: war
metaph: tech
metaph: space
metaph: police
metaph: game
metaph: economy
f ≥ 10
f ≥ 5
f < 5
0
10
20
30
40
50
number of keywords

Finding metaphor keywords

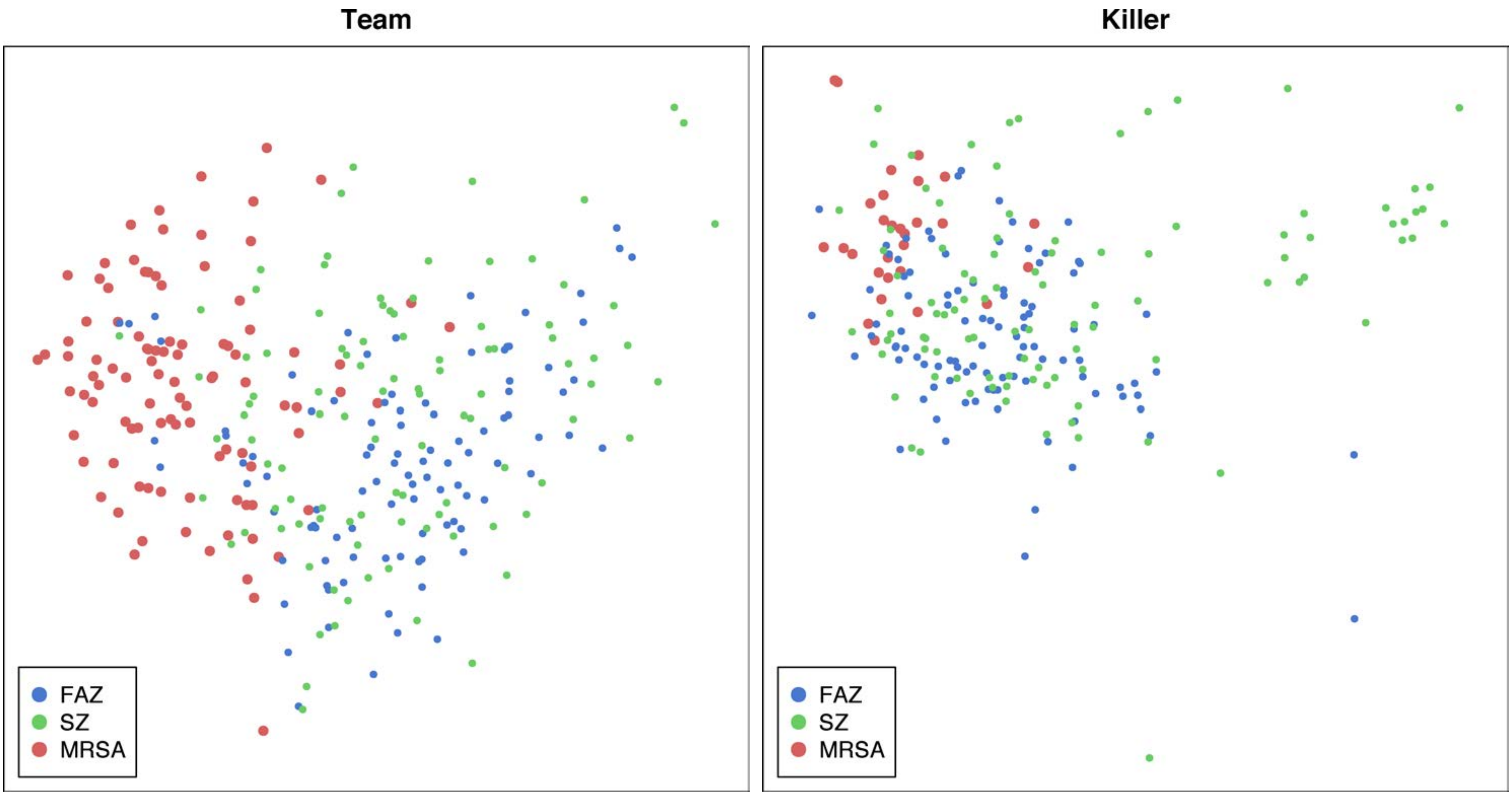
- Substantial number of plausible keywords for all metaphor categories except ECONOMY
  - frequent in target corpus & pass threshold in reference
  - but very low ranks (> 1000) from all keyness measures
- Reason: literal senses very frequent in reference
  - aggregating all keywords from category doesn't help
- Approximate semantics with distributional context vectors (Schütze 1998)
  - three-sentence context around each potential keyword
  - bag-of-words centroids of word embeddings
  - MRSA contexts clearly separated from reference contexts?

Finding metaphor keywords



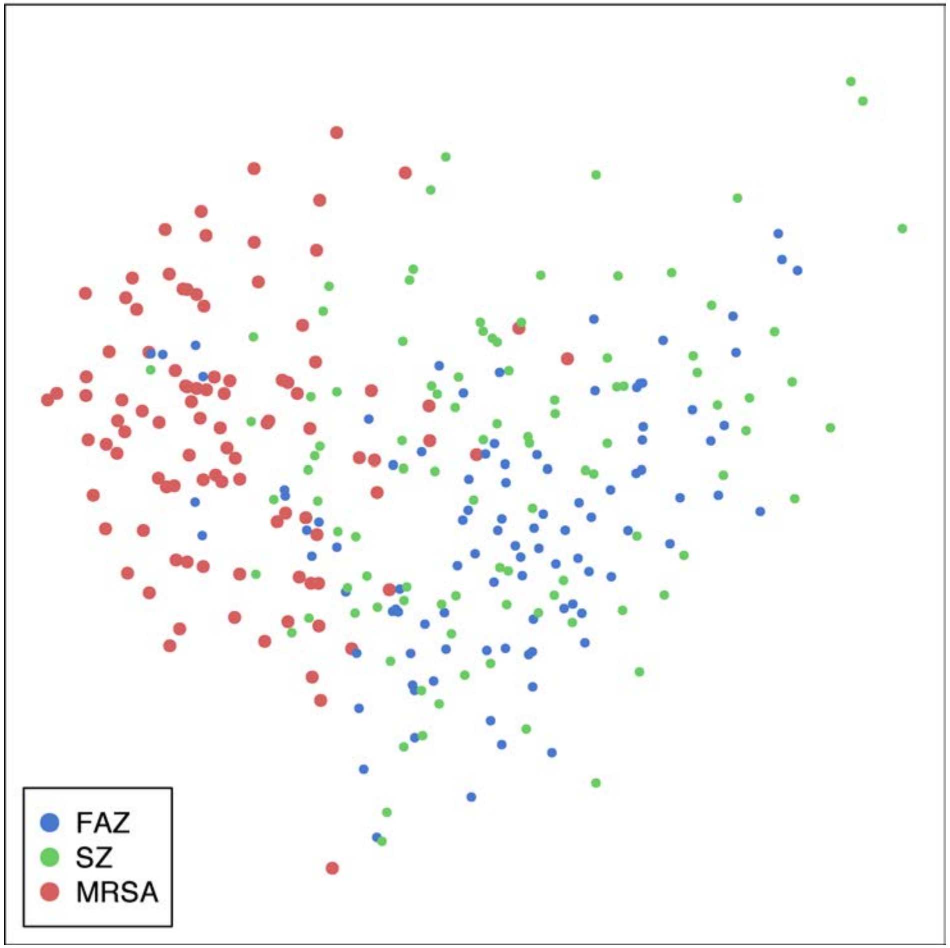


Finding metaphor keywords

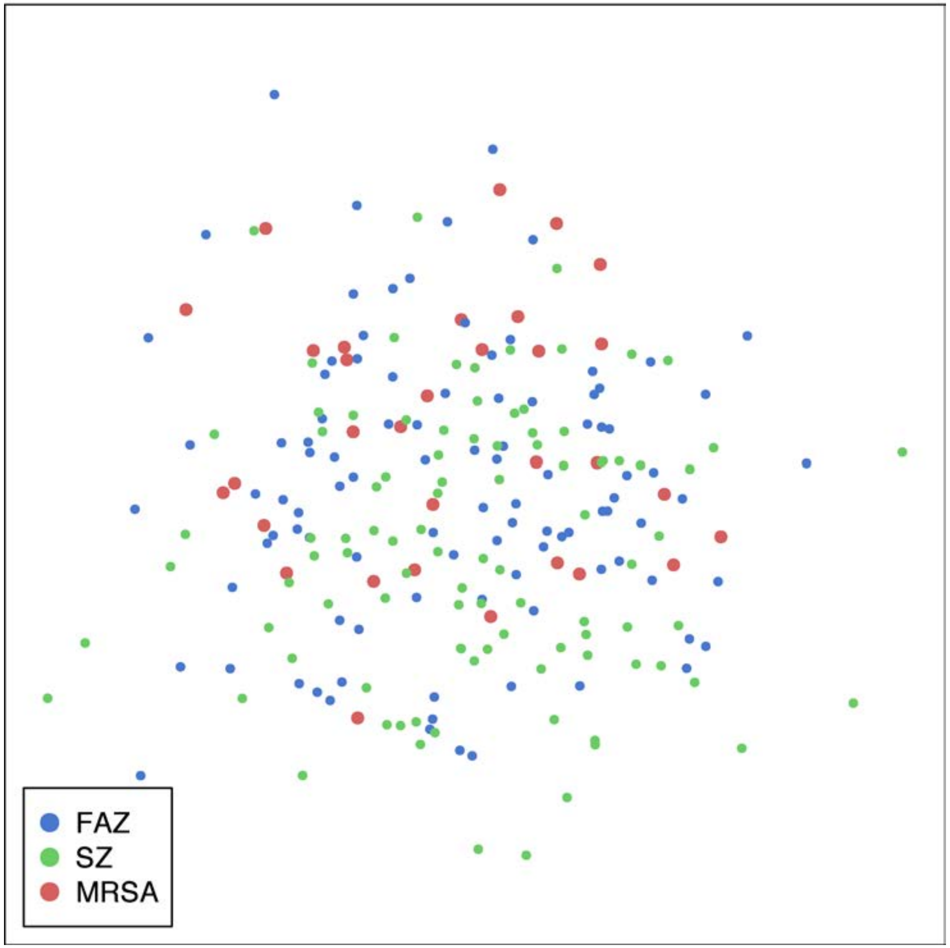


Finding metaphor keywords

Team



Forscherteam



References I

- Anthony, L. and Evert, S. (2019). Embracing the concept of data interoperability in corpus tools development. In *Proceedings of the Corpus Linguistics 2019 Conference*, Cardiff, UK.
- Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of LREC 2004*, pages 1313–1316, Lisbon, Portugal.
- 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.
- Dornseiff, F. (2004). *Der deutsche Wortschatz nach Sachgruppen*. De Gruyter, Berlin, 8th edition.
- Egbert, J. and Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104.
- 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.

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

References II

- Evert, S. (2022). Measuring keyness. In *Digital Humanities 2022: Conference Abstracts*, pages 202–205, Tokyo, Japan / online. https://osf.io/cy6mw/
- Fay, M. P. (2010). Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics*, 11(2):373–374.
- Fidler, M. and Cvrcek, 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.
- Hardie, A. (2014). A single statistical technique for keywords, lockwords, and collocations. Internal CASS working paper no. 1, unpublished.
- Kilgarriff, A. and Rose, T. (1998). Measures for corpus similarity and homogeneity. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 46–52, Granada, Spain.

References III

- Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics 2009 Conference*, Liverpool, UK.
- Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.
- 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.

References IV

- Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, MI.
- 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.
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.
- Scott, M. (1997). PC analysis of key words – and key key words. *System*, 25(2):233–245.
- Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika*, 62(2):371–374.