FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

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

Stefan Evert, Natalie Dykes, Joachim Peters

FAU Erlangen-Nürnberg

www.linguistik.fau.de

Aim

- Keywords as “a quick and simple ‘way in’” to corpus comparison (Baker et al. 2013)
- Previous approaches to KW calculation focus on mathematical adequacy and/or number of generated items (cf. Kilgarriff 2001, Paquot & Bestgen 2009, Lijffijt et al. 2016)

Our approach:

- Previously determined qualitative linguistic categories
- Evaluate statistically generated keyword lists against them
- Procedure specifically tailored to discourse analysis

Corpus

- 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
- Overall corpus metadata (manual)
  - Actor: author
  - Actor: intended reader
  - Topic
    - MRO
    - related topic (clinical hygiene; other infections…)

Corpus

Extraction of relevant subcorpus via metadata

- Actor – author: media
- Actor – reader: general public
- Topic: MRO

1,3M tokens (1,177 texts) of mass media texts and reader comments taken from the MRO corpus

Reference corpora

- 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: POS-tagged with TreeTagger and lemmatised with SMOR (Schmid et al. 2004)

Annotation categories

Annotation of top 200 lexical KW for different techniques following gold standard based on previous analysis of a different MRO press corpus (Peters 2017)

Adaption of selected aspects of the DIMEAN model (Spitzmüller/Warnke 2011)

- Actor
- Topos
- Metaphor
- False positives (unclear/other/irrelevant)
- Additional category: evaluative lexis (positive/negative stance)

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 FP (“unclear”) 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

Keyword extraction techniques

f_1	f_2
$f_1 - n_1$	$f_2 - n_2$

- f_1 = freq. in target corpus
- n_1 = sample size of target
- f_2 = freq. in reference corpus
- n_2 = sample size of reference

- Textbook approach: G^2 log-likelihood significance test (Dunning 1993)
- Effect size measure: **LR** log ratio $f_1/n_1 : f_2/n_2$ (Hardie unpulished)
  - combined with Bonferroni-corrected significance filter
- Statistician's choice: **LR$_{cons}$** conservative LR (Evert p.c.)
  - lower bound of confidence interval (Hardie's formula)
  - with Bonferroni correction

Keyword extraction techniques

f_1	f_2
$f_1 - n_1$	$f_2 - n_2$

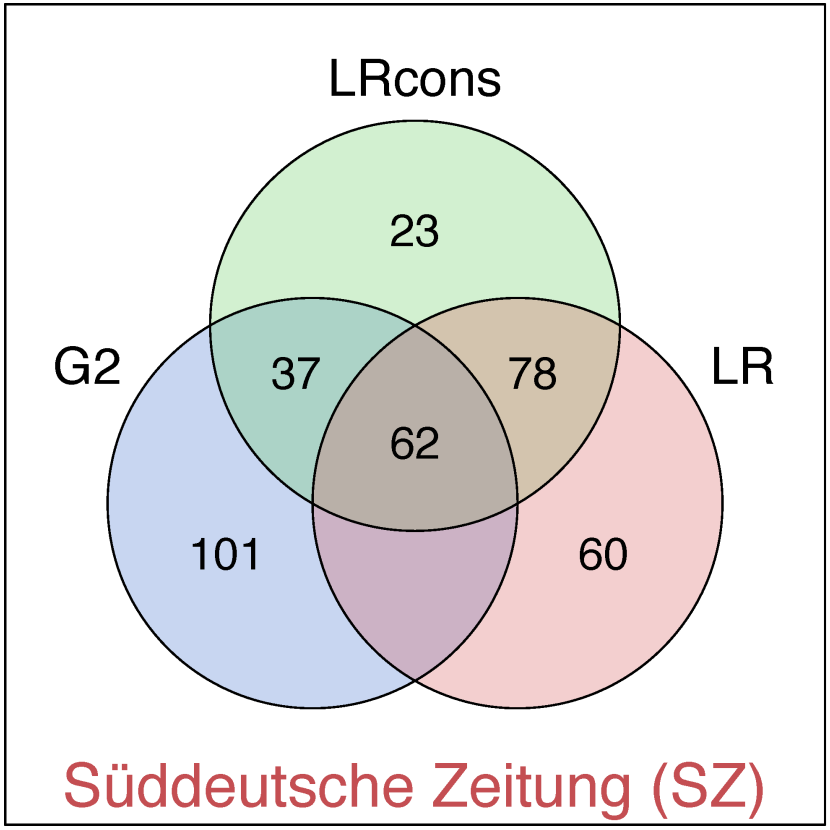
- f_1 = df in target corpus
- n_1 = #texts in target corpus
- f_2 = df in reference corpus
- n_2 = #texts in reference

- Methodological discussion: non-randomness / term clustering as key issue
- Simple correction: use document frequency (df) instead of raw frequency
- Mathematical justification as statistical inference for **α** parameter of Katz (1996)

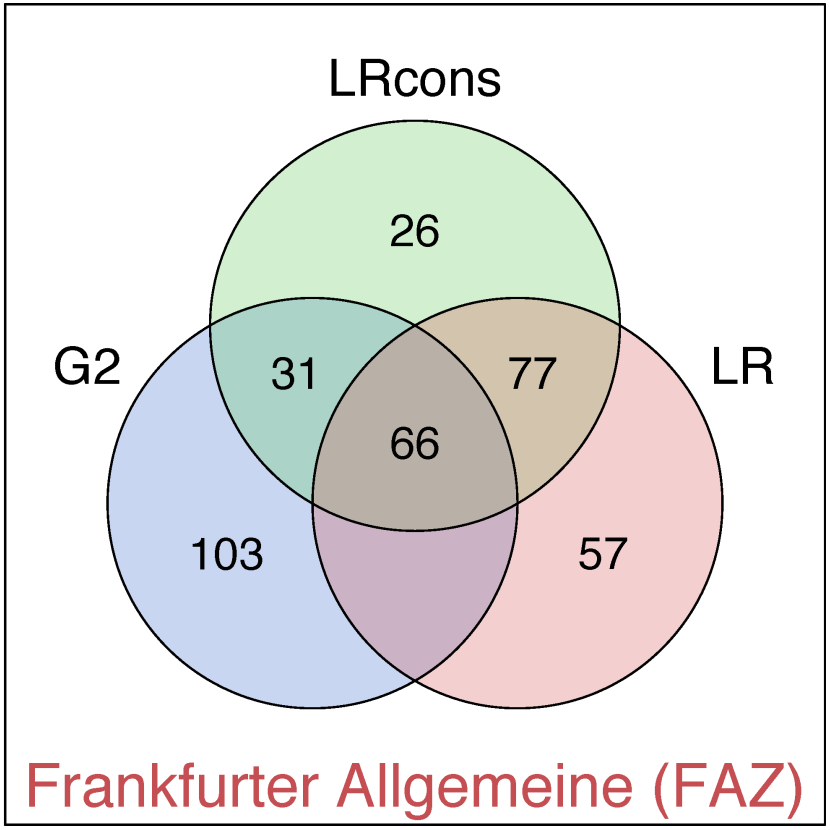
Experiments

- 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)
- Two comparable reference corpora: *Süddeutsche* (SZ) vs. *Frankfurter Allgemeine* (FAZ)
- Keywords based on raw frequency (classic) vs. document frequency (df-based)

Overlap between techniques

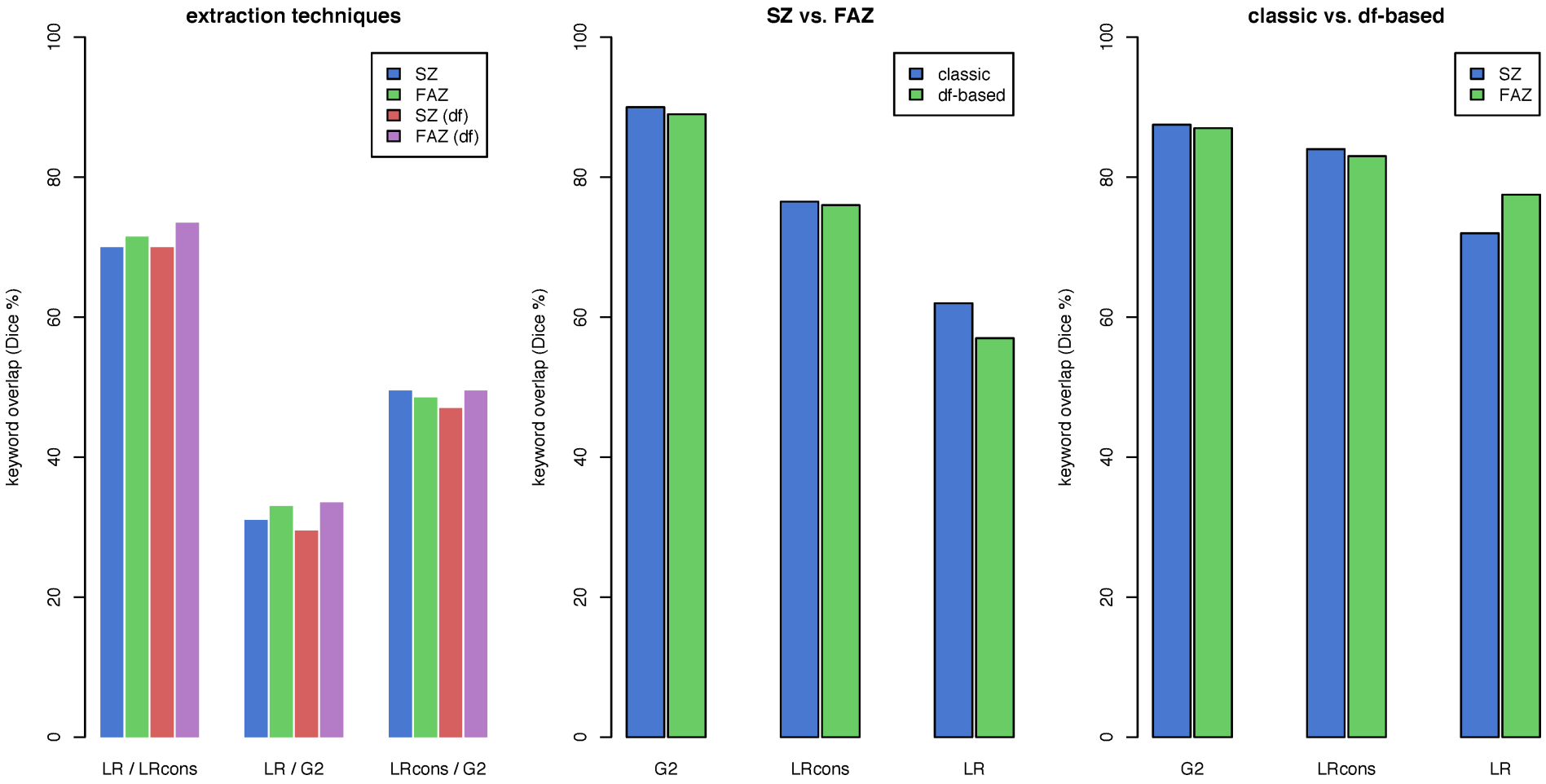


Süddeutsche Zeitung (SZ)

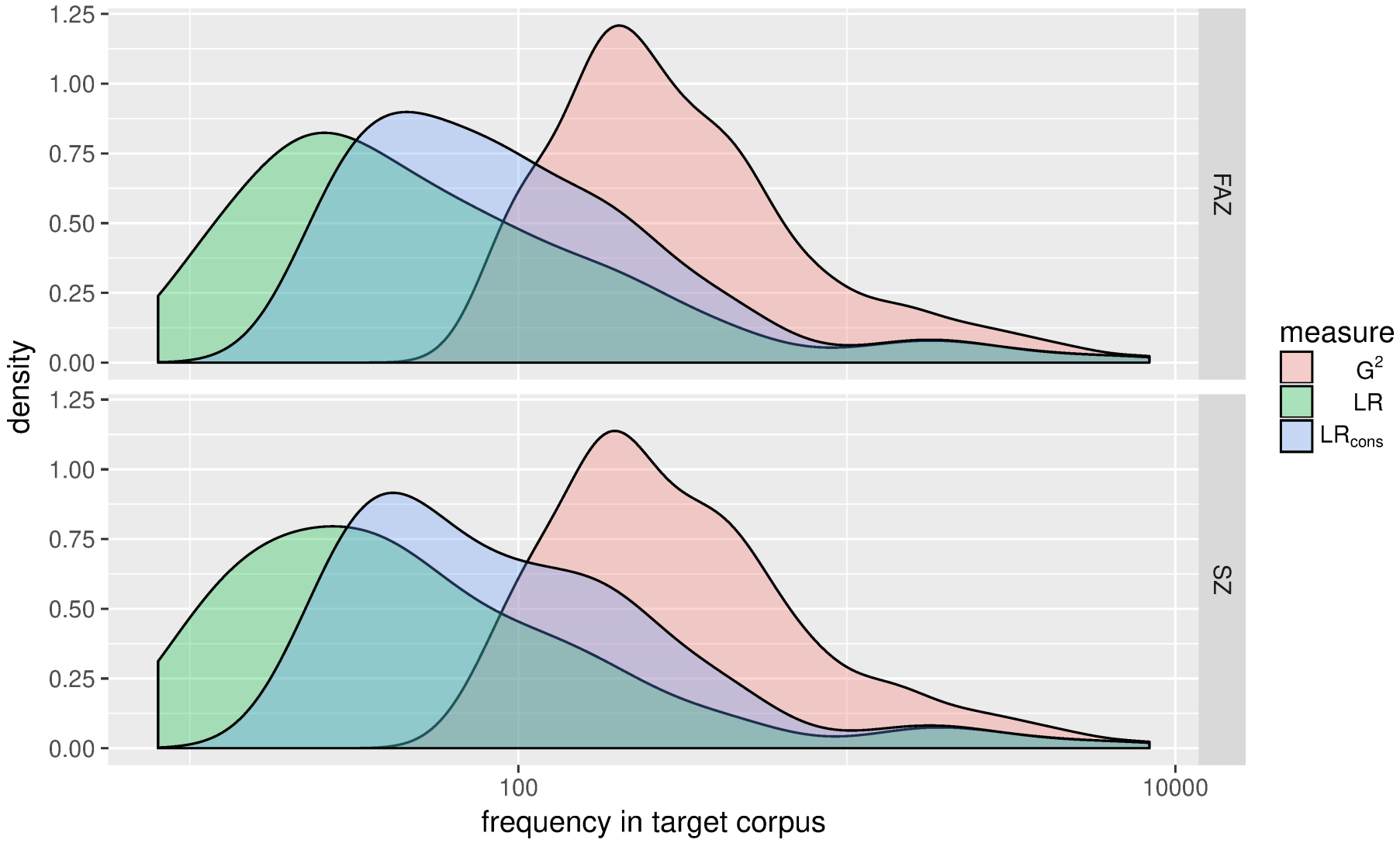


Frankfurter Allgemeine (FAZ)

Overlap between techniques

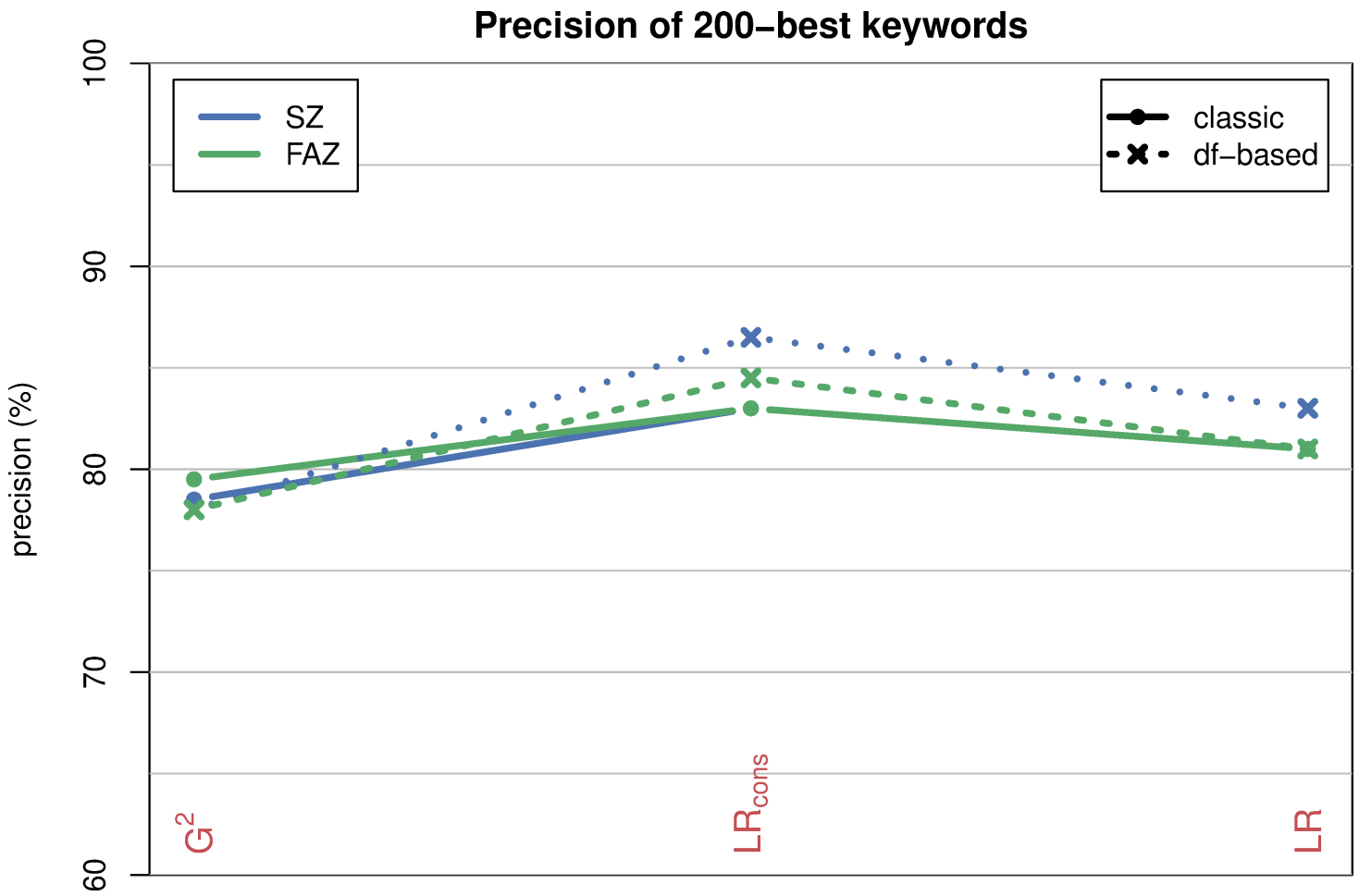


Frequency bias



Precision = #TP / 200 cand.

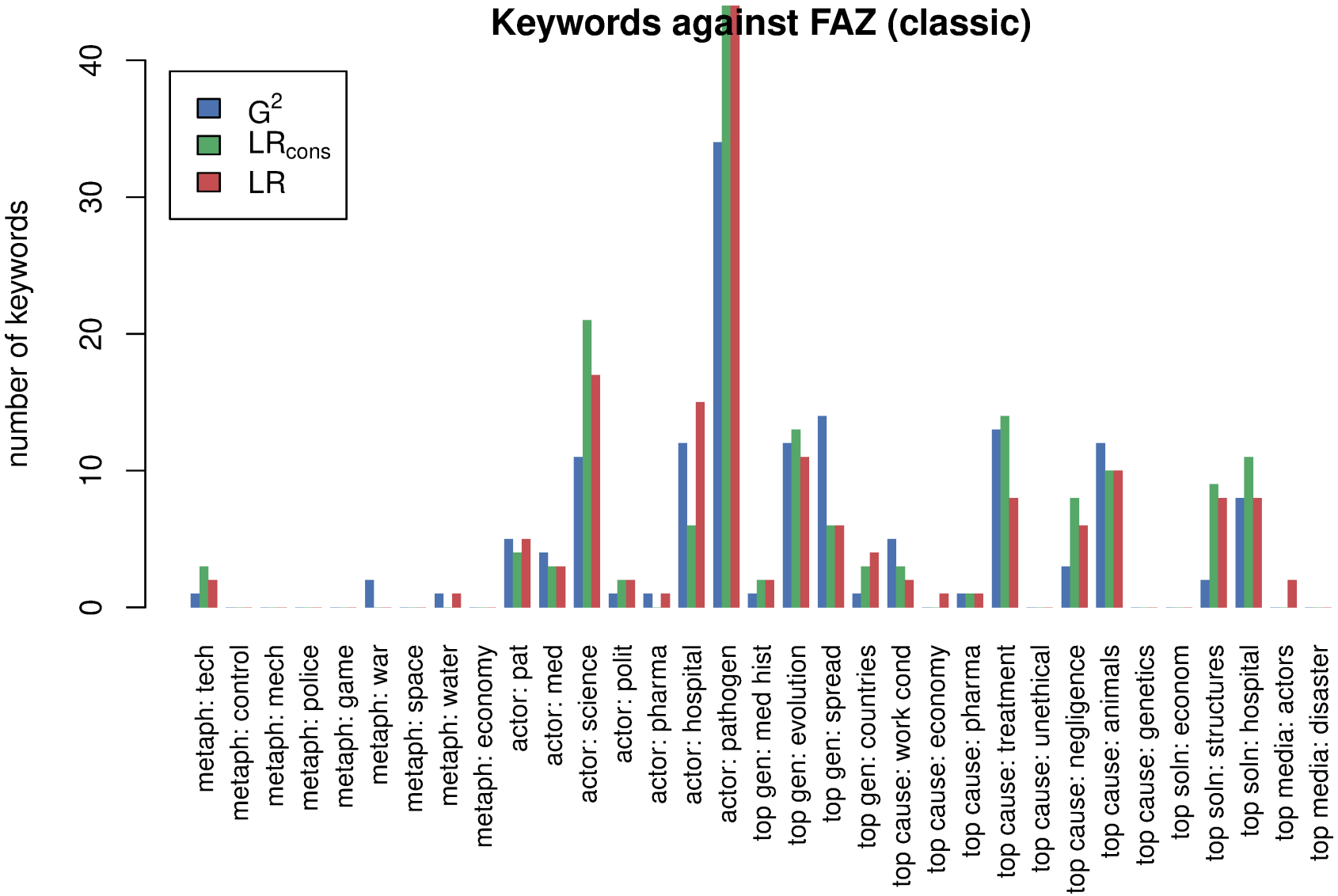
TP = assigned to category and/or evaluative



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

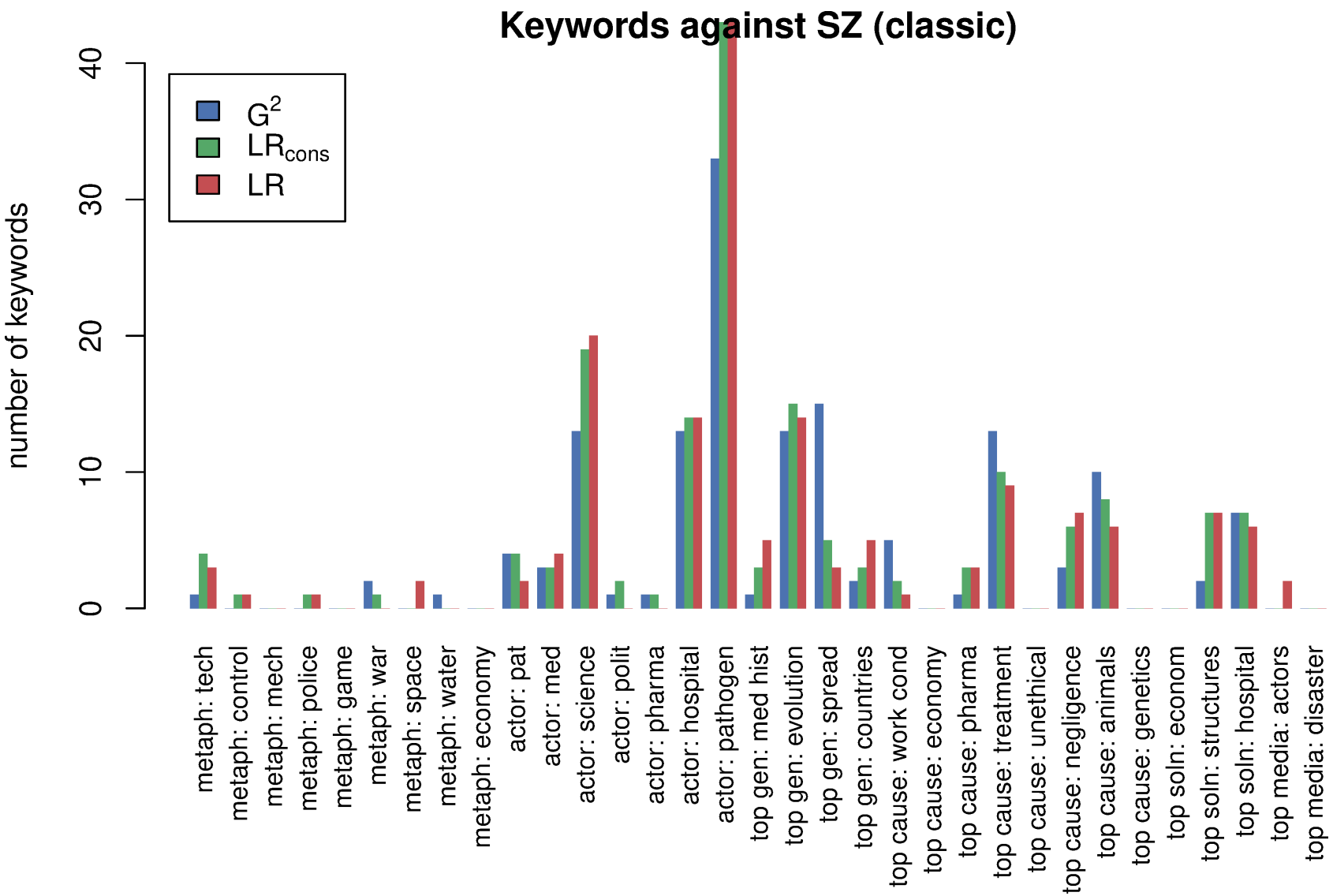
Recall = #kw for each category



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

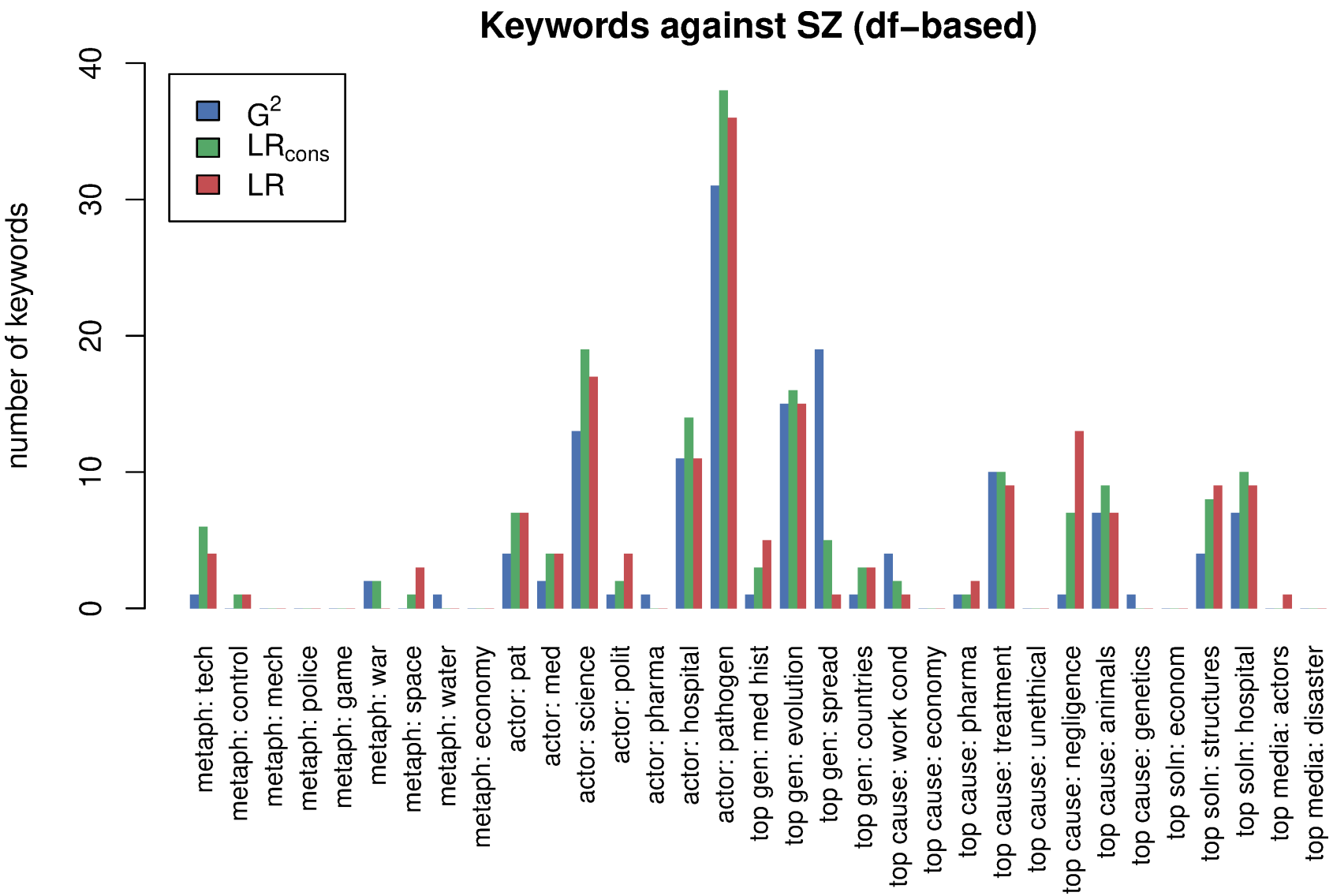
Recall = #kw for each category



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Recall = #kw for each category



Why so few metaphor keywords?

Possible causes:

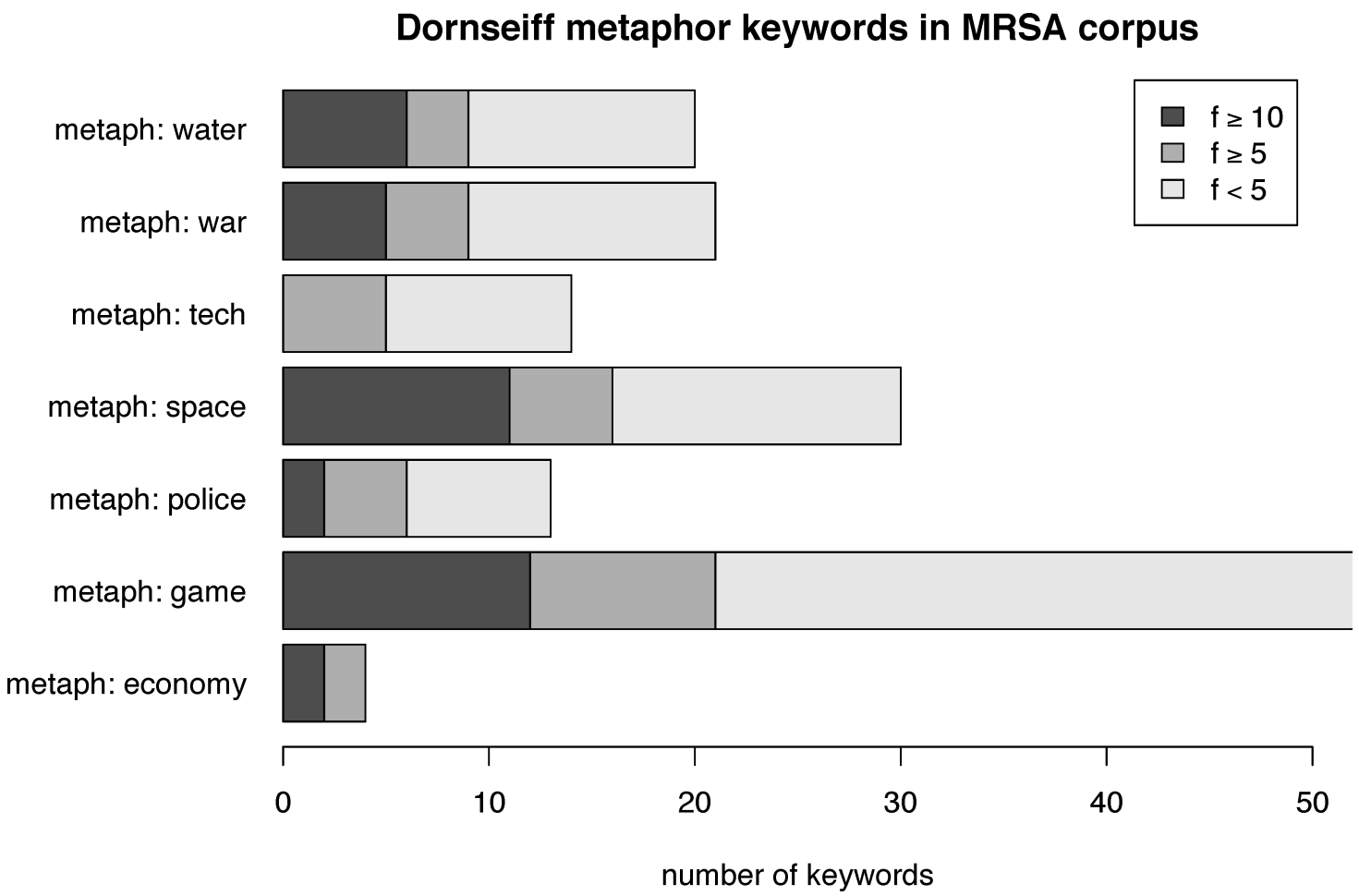
- No metaphors in online media discourse (unlikely)
- Cannot be reduced to single words
- Keywords occur, but are too infrequent

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

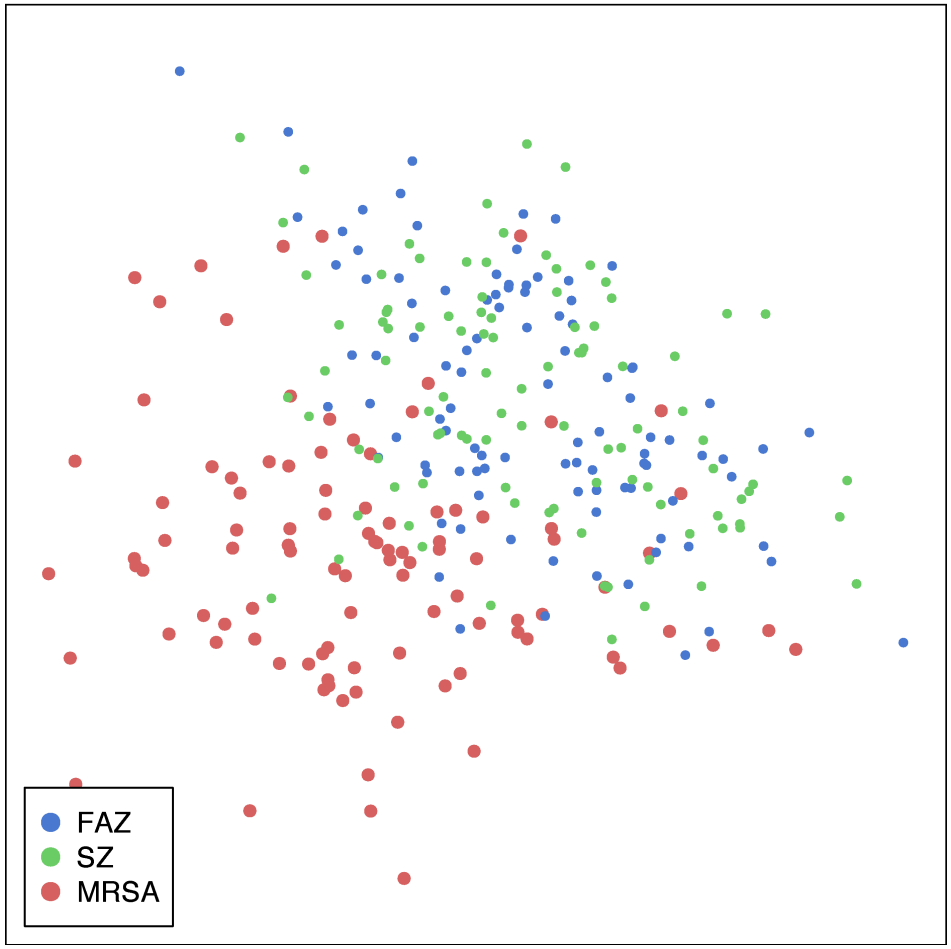

Dornseiff metaphor keywords in MRSA corpus
f ≥ 10
f ≥ 5
f < 5
metaph: water
metaph: war
metaph: tech
metaph: space
metaph: police
metaph: game
metaph: economy
0
10
20
30
40
50
number of keywords

Finding metaphor keywords

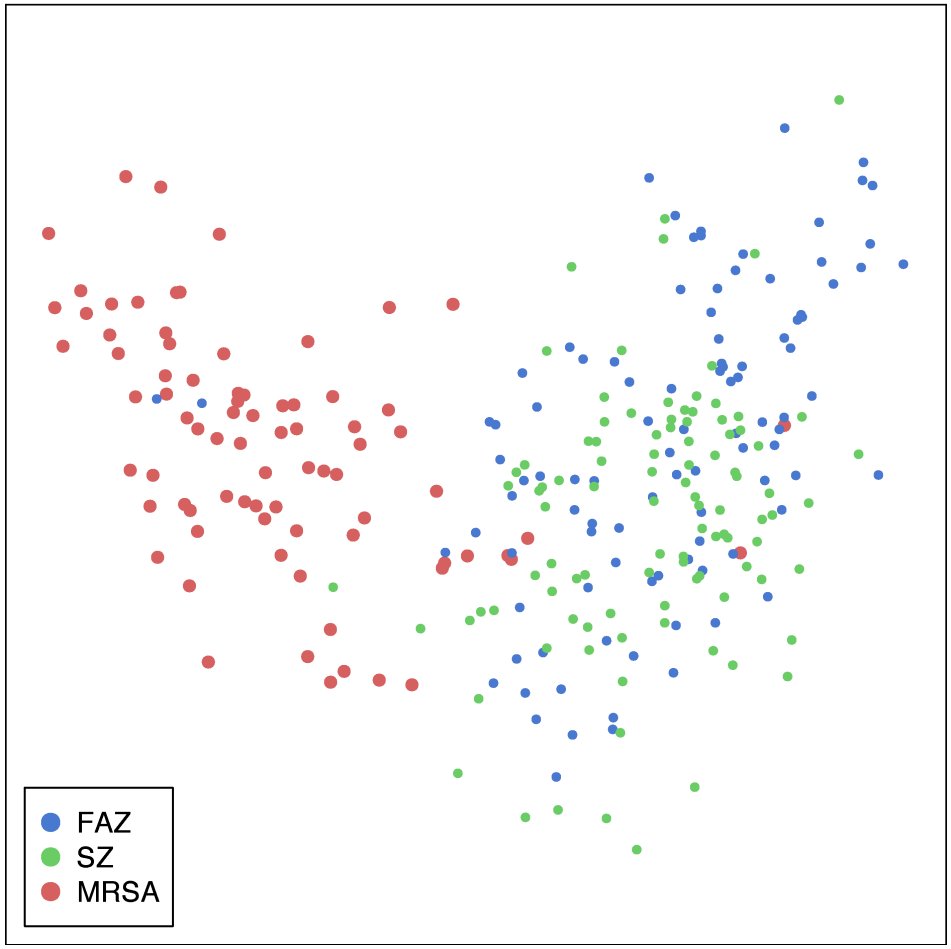
- Substantial number of plausible keywords for all metaphor categories except ECONOMY
  - frequent in target corpus & pass threshold in reference
  - but very low ranks (> 1000) from all keyness measures
- Reason: literal senses very frequent in reference
  - aggregating all keywords from category doesn't 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

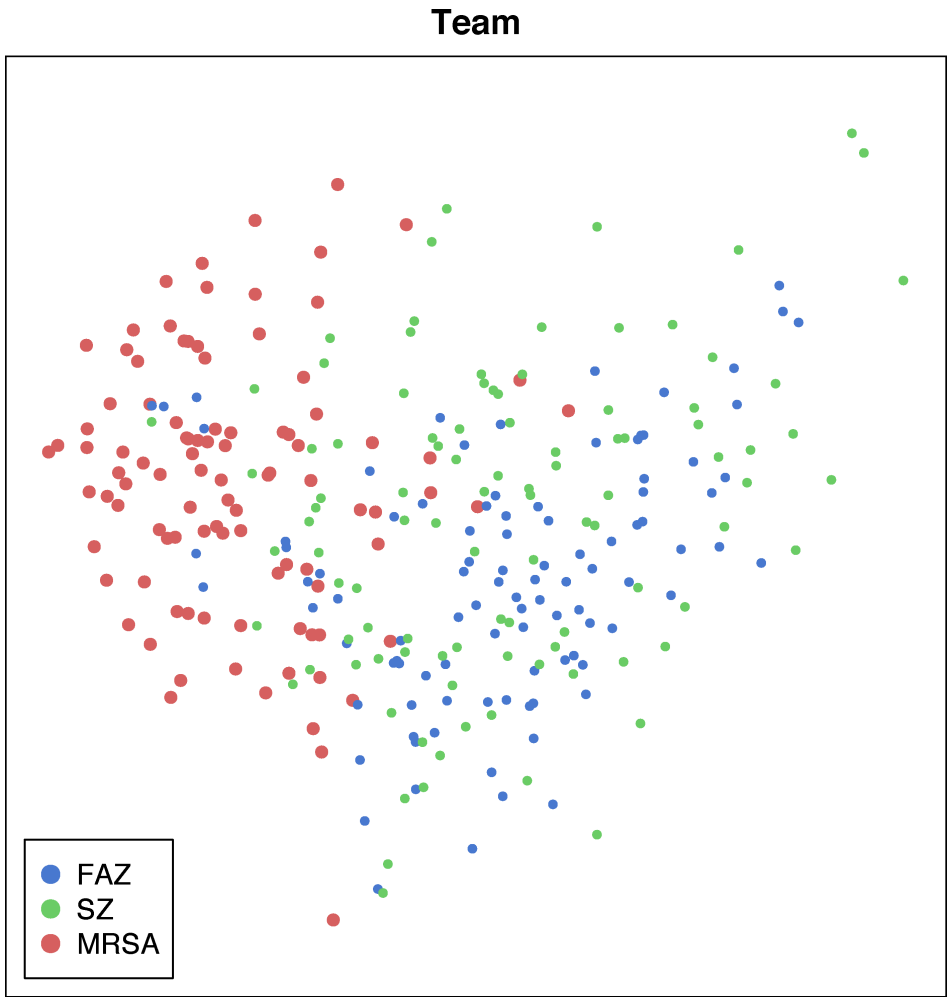
Kampf

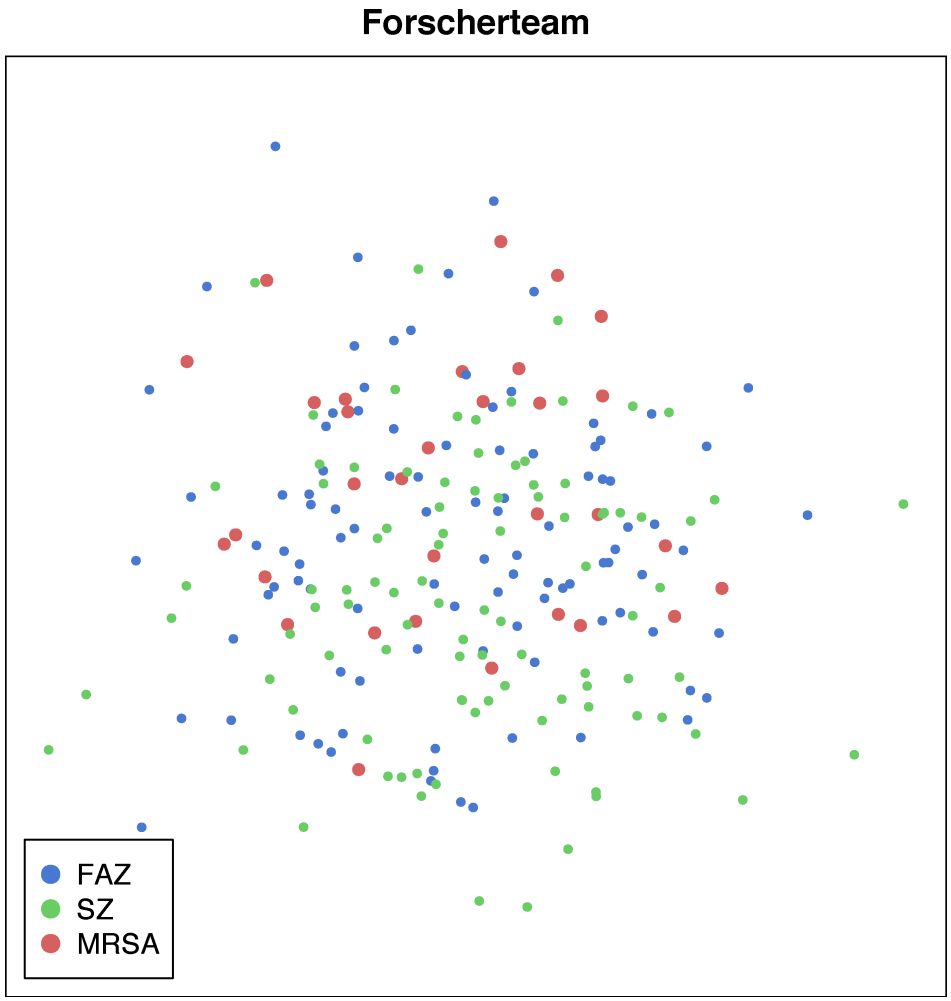


angreifen



Finding metaphor keywords





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

Conclusion

- Quantitative evaluation of keyword techniques & parameters for corpus-based discourse analysis
- Small overlap between G^2 and LR keywords
  - but choice of reference corpus makes little difference
- All techniques achieve high precision > 80%
- Recommendation: LR_{cons} on document frequency
- Good recall for some categories, poor for metaphors
- Suitable keywords *are* available → new techniques

And Thank You for your attention!

Don't forget Natalie & Joachim's talk at 14:20 (same room).

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.

Baroni,Marco & Bernardini, Silvia (2004). „BootCaT: Bootstrapping corpora and terms from the web“. In Lino, Maria et al. *Proceedings of the Ivth International Conference on Language Resources and Evaluation (LREC).* Paris: ELRA, S. 1313–1316. URL: http://clic.

cimec.unitn.it/marco/publications/lrec2004/bootcat_lrec_2004.pdf (accessed 05/06/2017).

Dornseiff, Franz (2004). *Der deutsche Wortschatz nach Sachgruppen.* Berlin: De Gruyter.

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.

References

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.

Katz, Slava M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, **2**(2), 15–59.

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.

References

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.

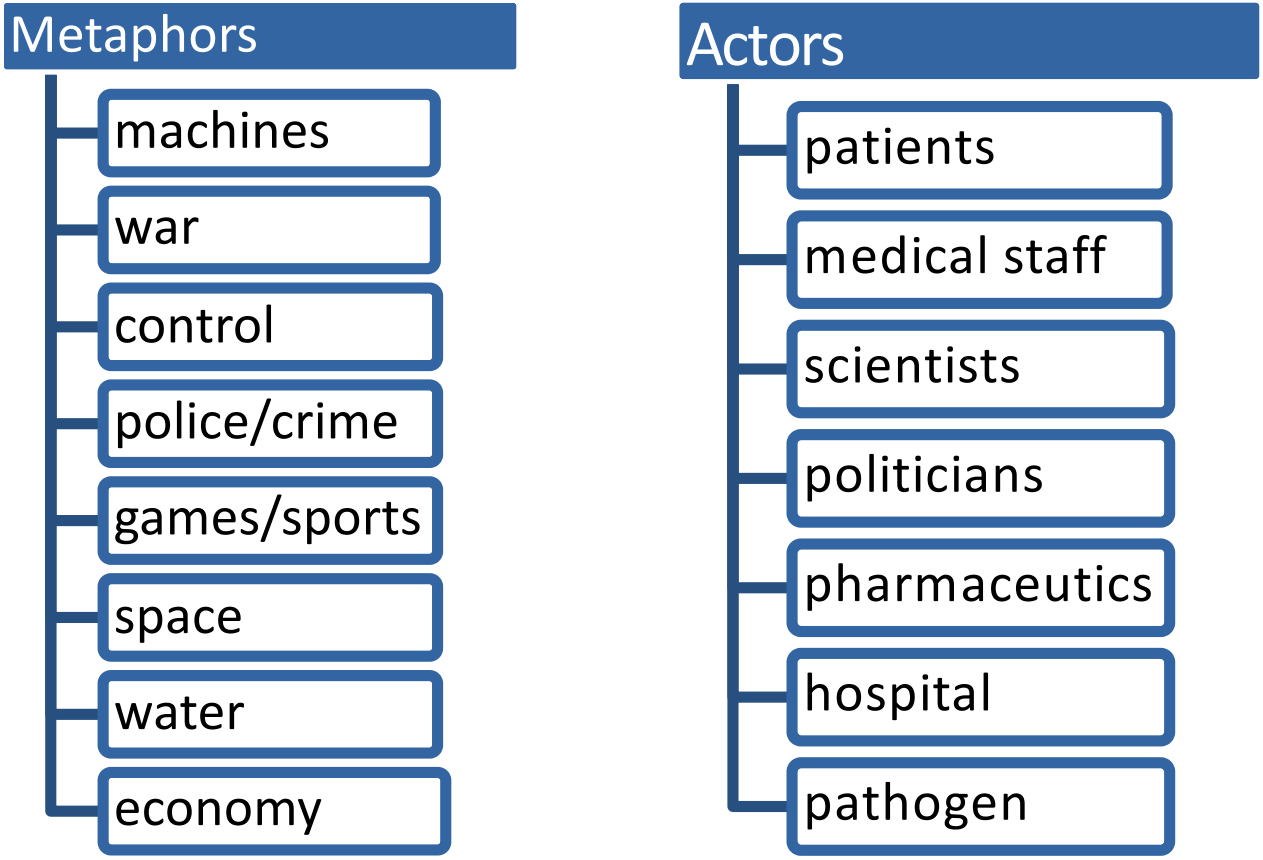
Scott, Mike; Tribble, Christopher (2006). *Textual patterns – Key words and corpus analysis in language education*. Studies in Corpus Linguistics: Vol. 22. Amsterdam, Philadelphia: John Benjamins.

Schütze, Hinrich (1998). Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–123.

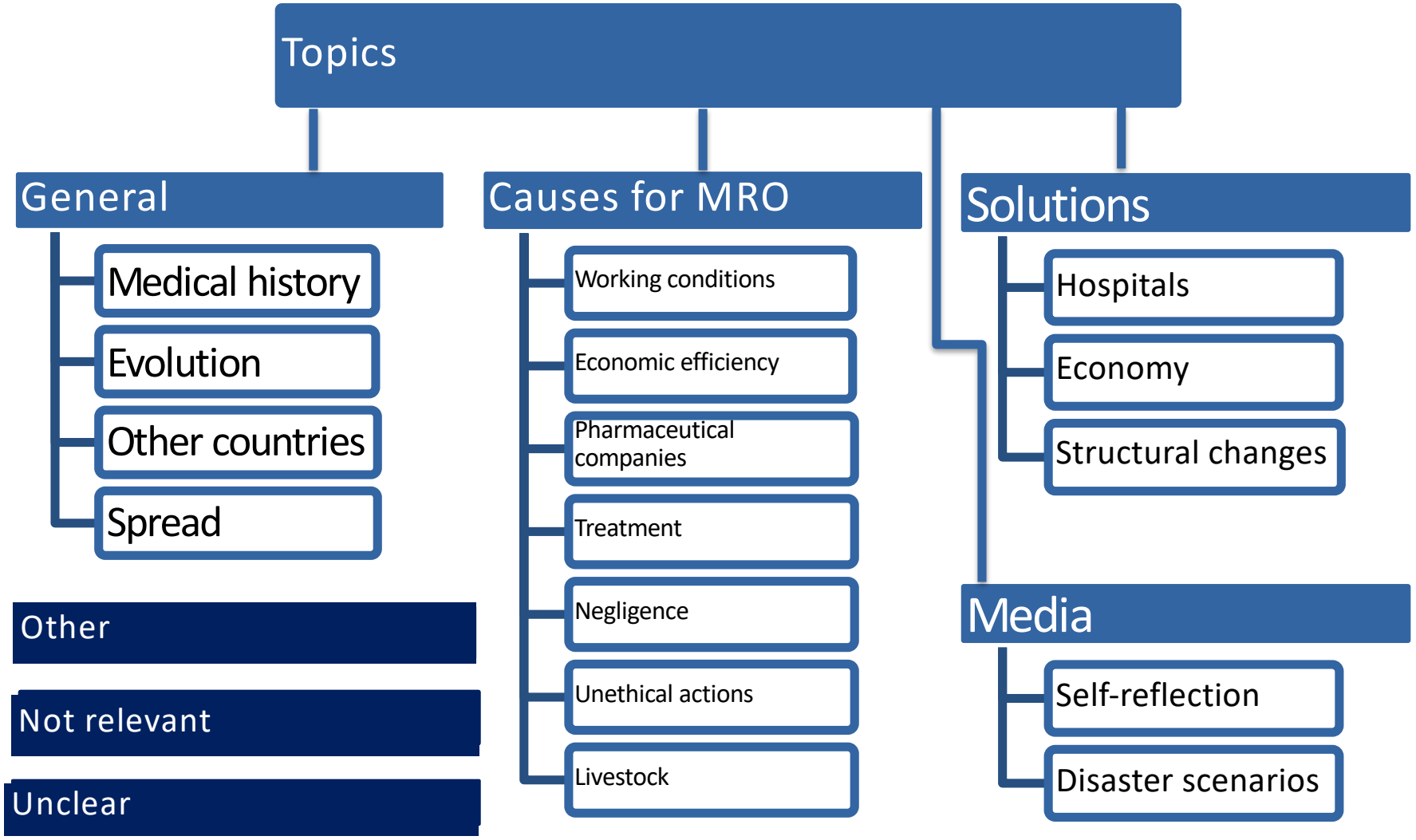
Spitzmüller, Jürgen &Warnke, Ingo (2011): Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. Berlin/New York: De Gruyter.

Annotation scheme

Categories from previous manual study on smaller corpus (Peters 2017)



Annotation scheme



IAA

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Confusion matrix (primary category)