FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

HS Corpus Linguistics / Korpuslinguistik

3. Statistical analysis of frequency data

Prof. Dr. Stephanie Evert

Chair of Computational Corpus Linguistics

FAU Erlangen-Nürnberg

Outline

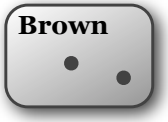
1. Quantitative analysis of corpus data:
   Why do we need statistics?

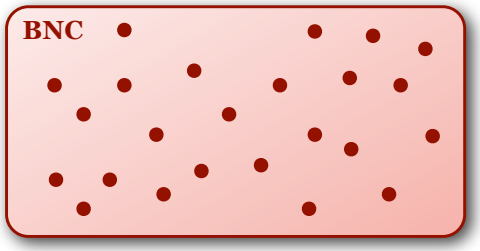
2. Keywords
   - an application of frequency comparison methods
   - the need for empirical evaluation studies

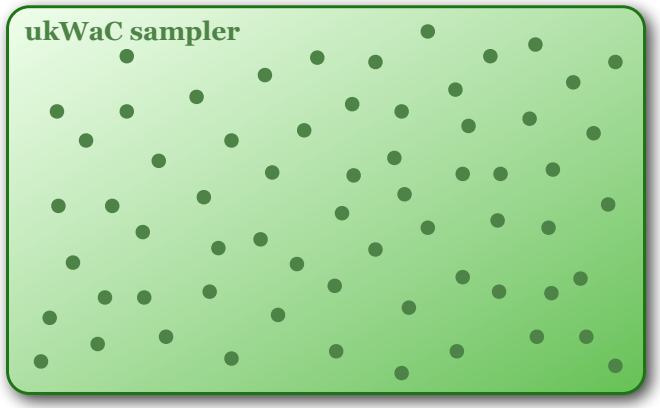
3. Collocations
   - bonus: lexicographic evaluation of corpus-based collocation identification

Interpreting corpus frequencies

- Interpret as relative frequency (e.g. per million words)
  - Are there 20,000 passives?
    - Brown (1M words)
  - Or 1 million?
    - BNC (90M words)
  - Or 5.1 million?
    - ukWaC sampler (450M words)
- But need absolute frequency counts to assess statistical significance!







Example: déjà vu

- Spellings of *déjà vu* in the BNC
  - 38x with accents *déjà vu*
  - 23x without *deja vu*
  - 3x confused *déjá vu* etc. (we'll ignore these here)
- Can we say that the correct spelling is most common in BrE (i.e. has relative frequency π > 50%)?
- Direct estimate:

$$\pi \approx p = \frac{f}{n} = \frac{38}{38 + 23} = 62.3\%$$

NB: not frequency *per million words* in this case!

Sampling variation

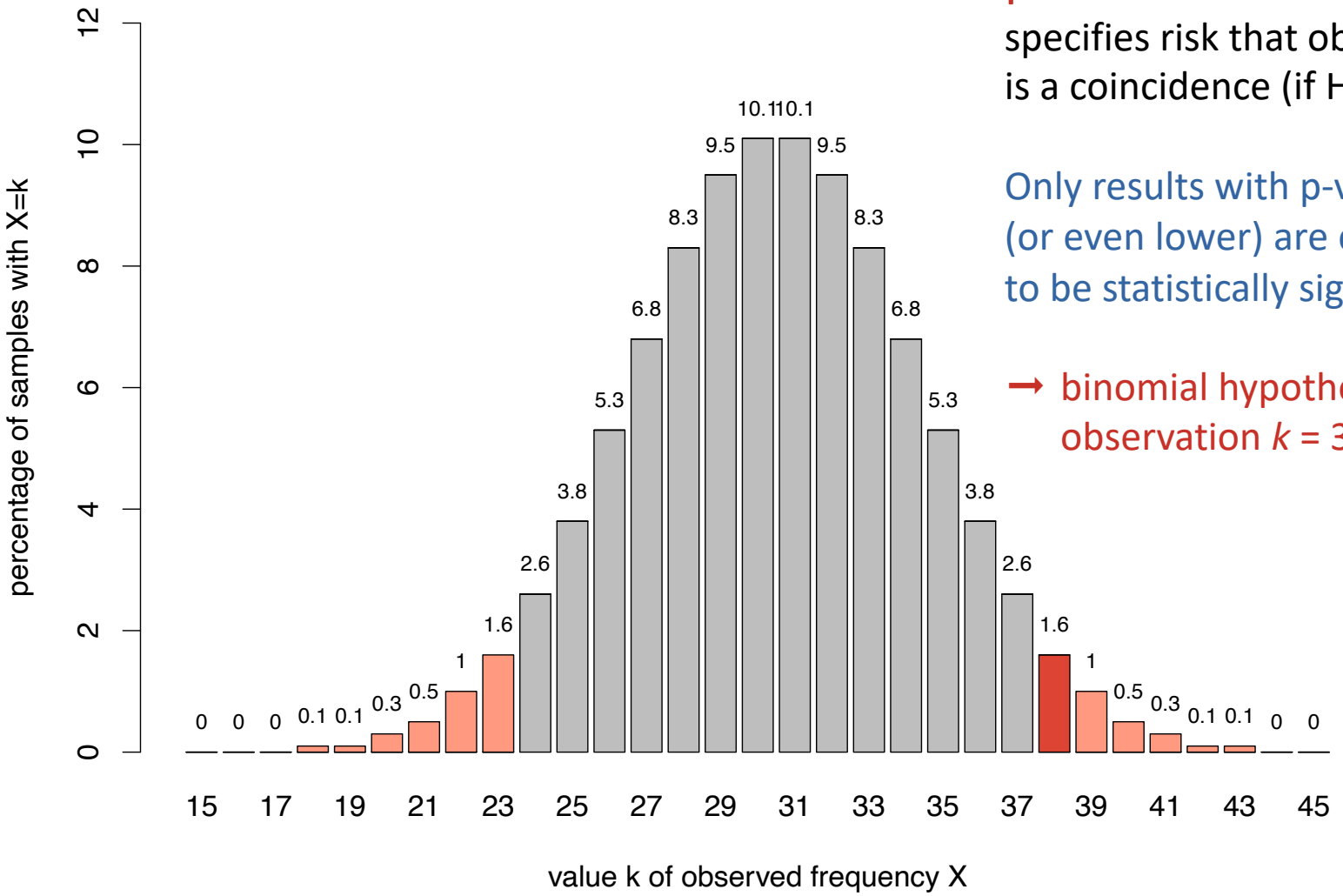
- But what if the compilers of the BNC had chosen a slightly different set of texts? Would we still find the same proportion of *déjà vu*?
- Random sample ➞ sampling variation
- Statistical hypothesis tests can be used to assess how much sampling variation has to be expected
- A "what if" game starting from null hypothesis H_0:

$$H_0: \pi = 50\%$$

☞ Simulate many random samples of same size $n = 61$

  – drawn from a population in which H_0 is true

Binomial sampling variation



p-value = 7.2%

specifies risk that observation is a coincidence (if H_0 is true).

Only results with p-value < 5% (or even lower) are considered to be statistically significant.

→ binomial hypothesis test for observation $k = 38$ / $n = 61$

Confidence intervals & effect size

- What about *–ize* vs. *–ise*?
  - 1398x *characterise*
  - 1278x *characterize*
- Binomial test for 1398 / 2676: p-value = 0.02* (= 2%)
  - input: "successes" *k* / sample size *n*
  - try it for yourself: http://sigil.collocations.de/wizard.html
  - significance is achieved because of large sample size
- Confidence interval for π = all H_0 that cannot be rejected (with $p < .05$ or other significance level)
- Indicates small effect size:

$$50.3\% \leq \pi \leq 54.2\%$$

Frequency estimates

- Confidence intervals should always be computed for corpus frequencies!
- E.g. *kick the bucket* in BNC
  - 16 hits for CEQL query `{kick/V} <<3>> {bucket/N}`
  - corresponds to p = 0.16 pmw
  - confidence interval for 16 / 98313429:
    0.085 pmw ≤ π ≤ 0.252 pmw
  - footnote: is this really the usage frequency of the idiom?

sample size = number of tokens

- Not supported by standard corpus tools yet
  - but we're working on it
  - various statistical problems (e.g. non-randomness)

Hypothesis tests in practice

SIGIL: Corpus Frequency Test Wizard

back to main page

This site provides some online utilities for the project **Statistical Inference: A Gentle Introduction for Linguists (SIGIL)** by Marco Baroni and Stefan Evert. The main SIGIL homepage can be found at purl.org/stefan.evert/SIGIL.

One sample: frequency estimate (confidence interval)

back to top

Frequency count	Sample size
19	100

Clear fields

Calculate

95% confidence interval

in automatic format

with 4 significant digits

extrapolate to items

Two samples: frequency comparison

back to top

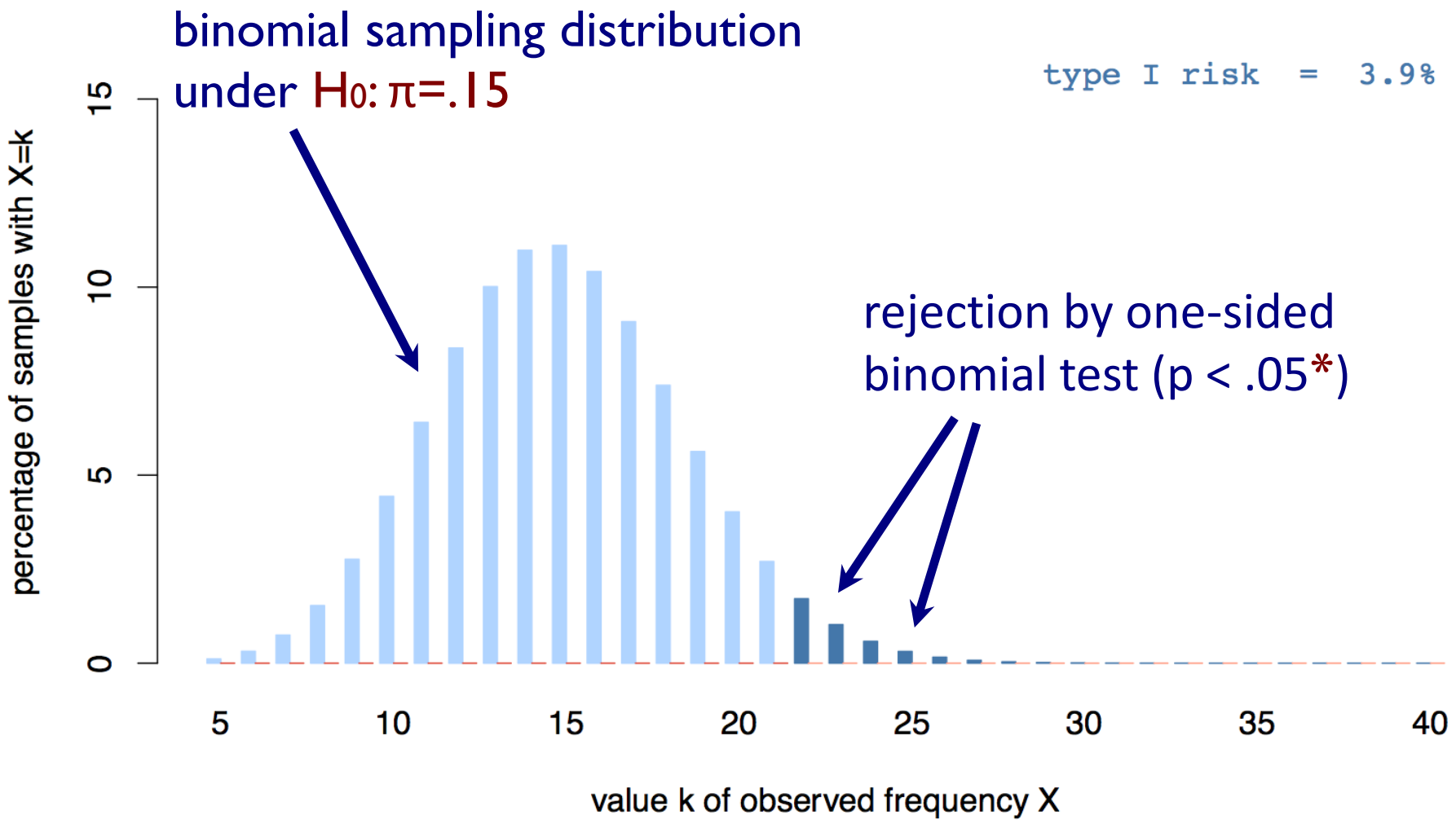
	Frequency count	Sample size
Sample 1	19	100
Sample 2	25	200

Clear fields

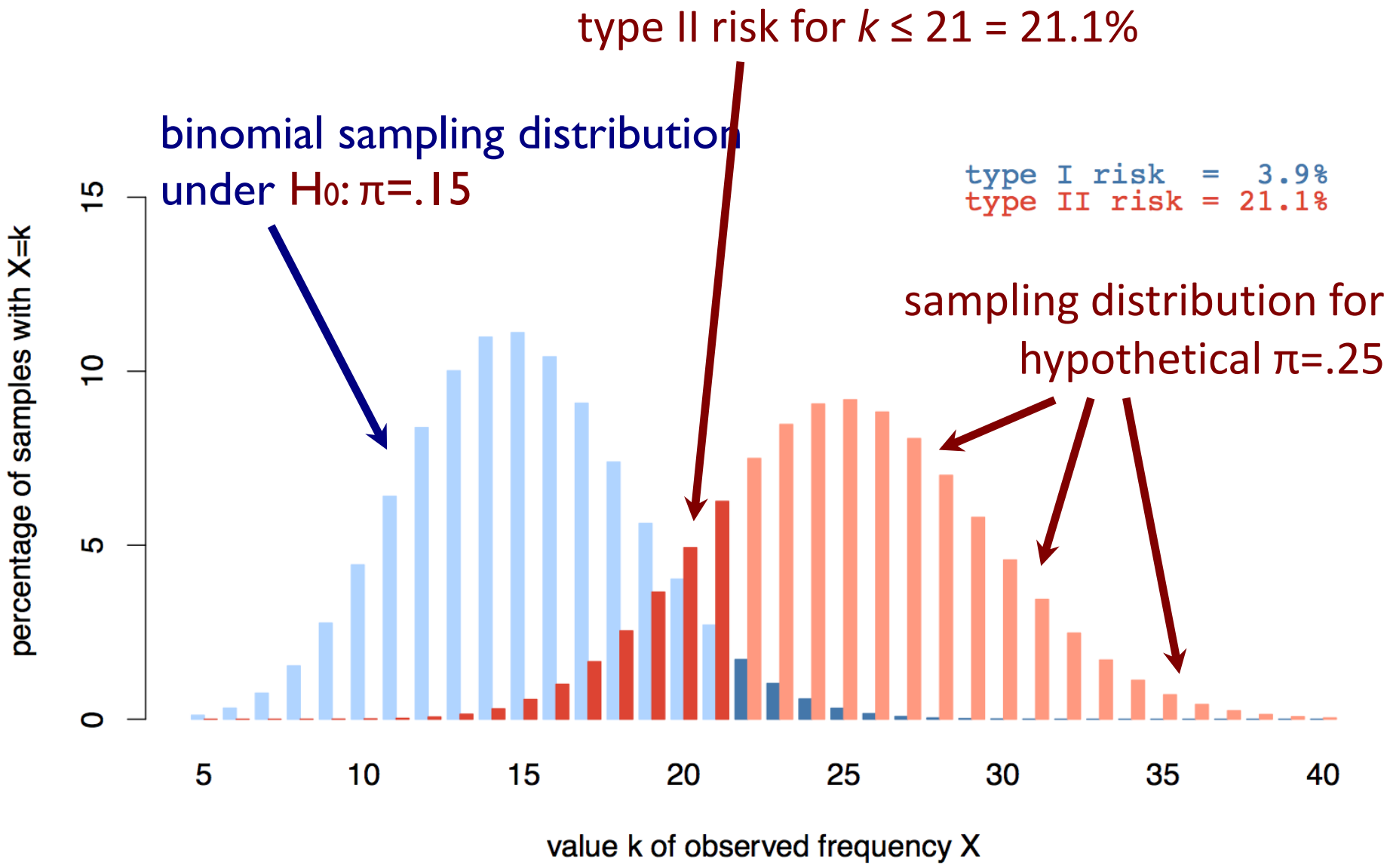
Calculate

- http://sigil.collocations.de/wizard.html
- http://corpora.lancs.ac.uk/sigtest/
- http://vassarstats.net/
- SPSS, SAS, Excel, …
- best to use dedicated software:

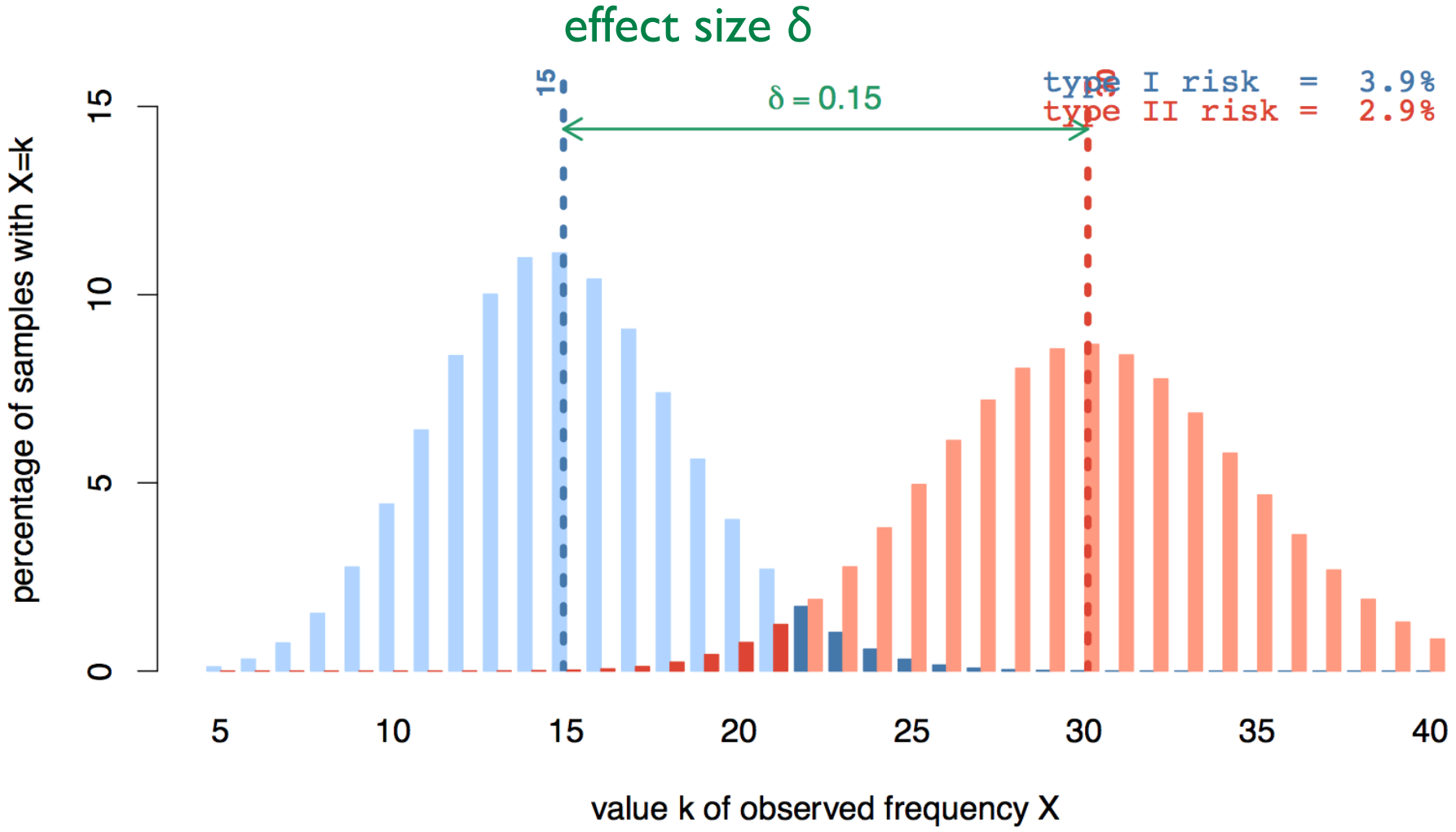
Type I errors



Type II errors

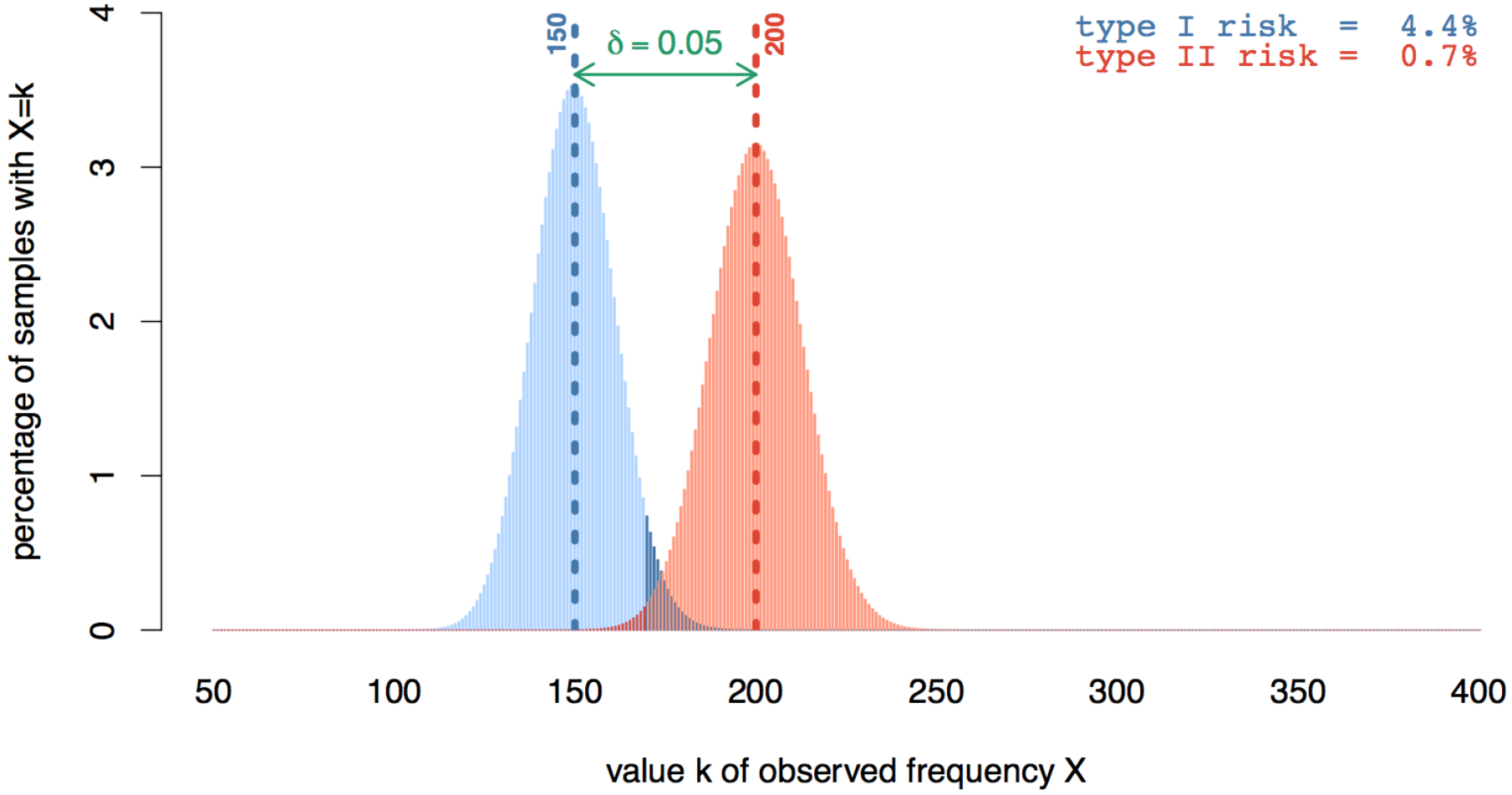


Type II errors & effect size



Type II errors & sample size

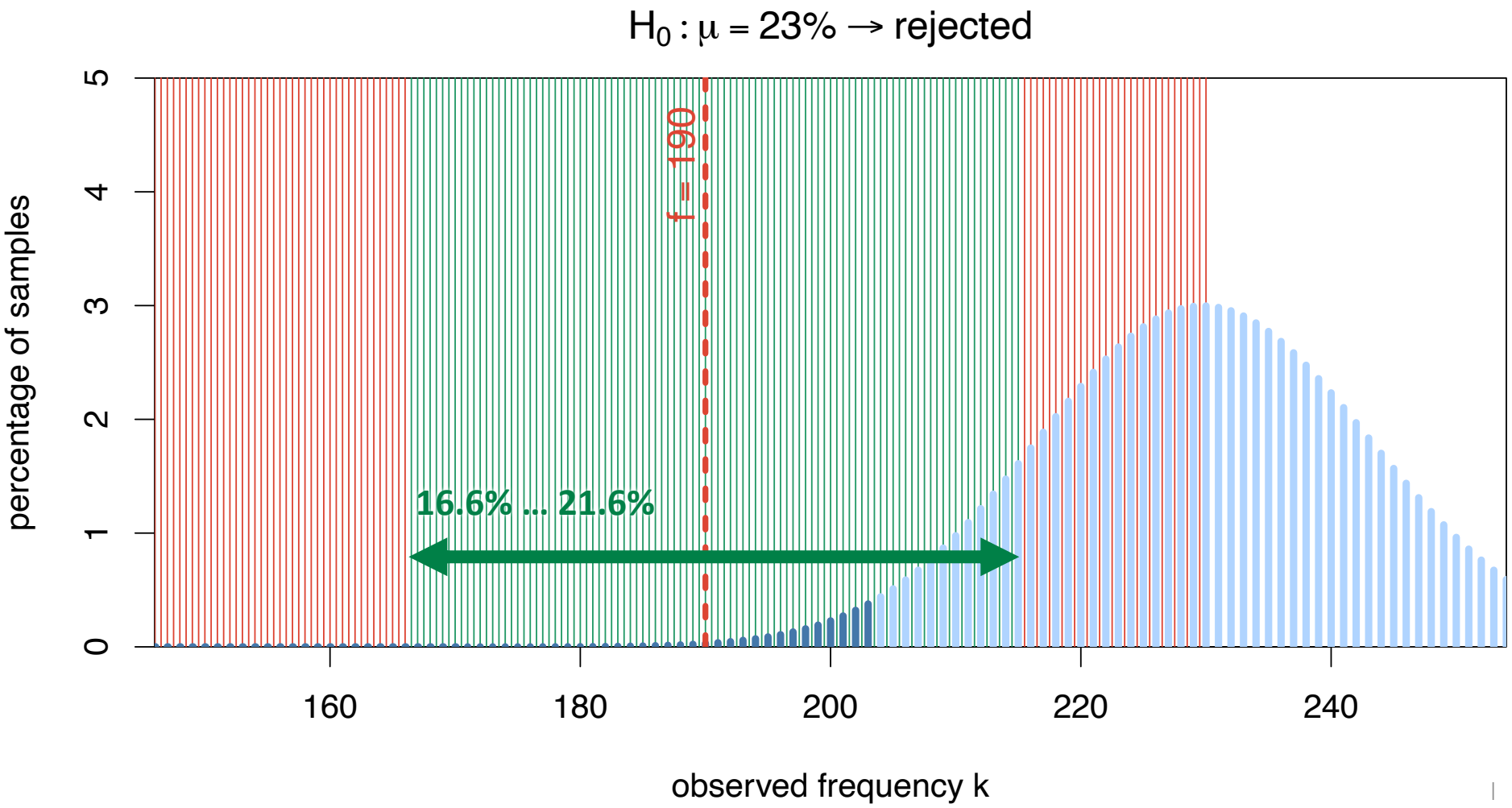
n = 1000



Confidence interval

observed data:
k = 190 / *n* = 1000

95% confidence
two-sided $p < .05 = \alpha$



Confidence intervals

- Confidence interval = range of plausible values for true population proportion
  - H_0 rejected by test iff π_0 is outside confidence interval
- Size of confidence interval depends on power of the test (i.e. sample size and significance level)

	$n = 100$ $k = 19$	$n = 1{,}000$ $k = 190$	$n = 10{,}000$ $k = 1{,}900$
$\alpha = .05$	11.8% . . . 28.1%	16.6% . . . 21.6%	18.2% . . . 19.8%
$\alpha = .01$	10.1% . . . 31.0%	15.9% . . . 22.4%	18.0% . . . 20.0%
$\alpha = .001$	8.3% . . . 34.5%	15.1% . . . 23.4%	17.7% . . . 20.3%

Frequency comparison

- Similar principles can be used to compare frequencies in different corpora / sub-corpora
- Application to lexical frequencies: keyword analysis

- Data for frequency comparison test:
  - f_1 = frequency in corpus 1
    n_1 = sample size of corpus 1
  - f_2 = frequency in corpus 2
    n_2 = sample size of corpus 2
- Collected in contingency table
  = cross-classification

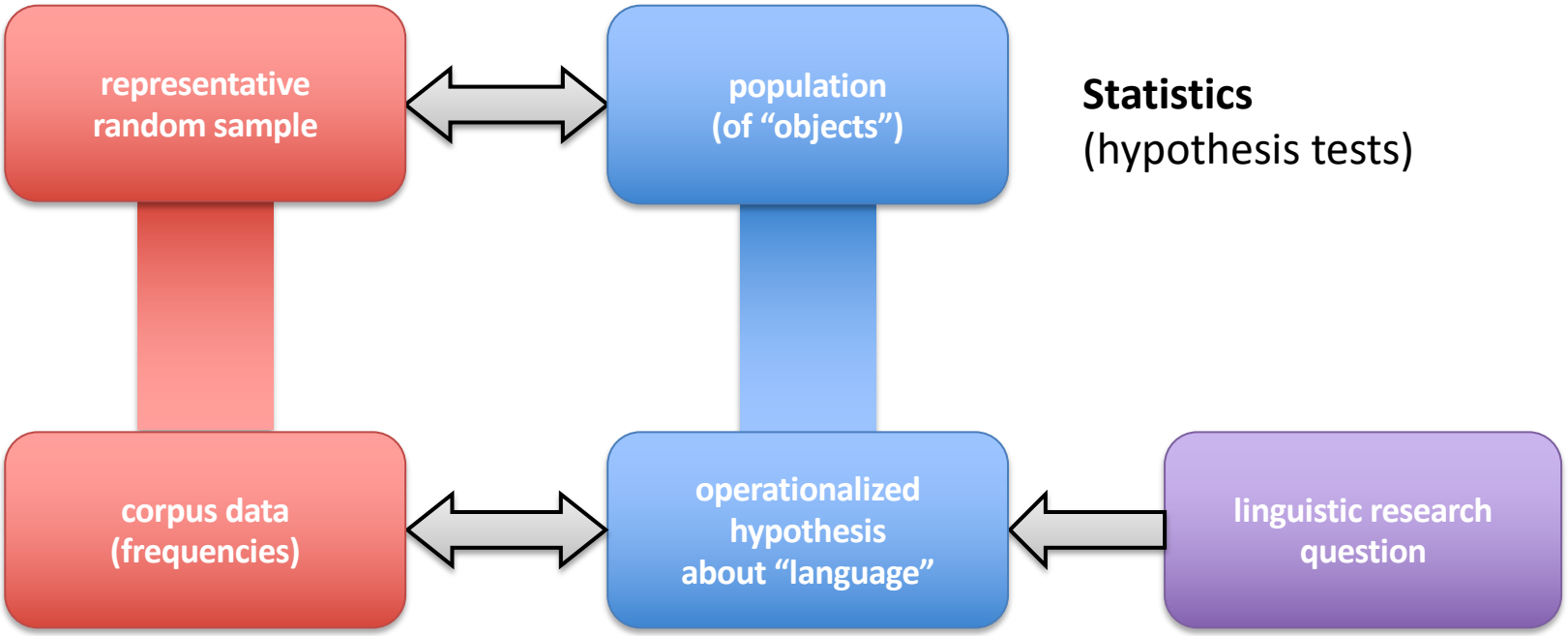
corpus 1	corpus 2
f_1	f_2
$n_1 - f_1$	$n_2 - f_2$

Frequency comparison

- Null hypothesis H_0: $\pi_1 = \pi_2$
  - precise value of common relative frequency not specified
- Different significance testing procedures:
  - Pearson's chi-squared test → X^2 statistic (↑)
  - likelihood ratio test ("log-likelihood") → G^2 statistic (↑)
  - Fisher's exact test → p-value (↓)

- Different ways of measuring effect size
  - difference of proportions: $\delta = \pi_1 - \pi_2$
  - relative risk = proportion ratio: $r = \pi_1 / \pi_2$
  - odds ratio (nice for maths): $\theta = \pi_1(1 - \pi_2) / \pi_2(1 - \pi_1)$

The broader picture on statistics in corpus linguistics



Statistics
(hypothesis tests)

Further reading

- Ch. 5 of Hoffmann, Sebastian *et al.* (2008). *Corpus Linguistics with BNCweb – a Practical Guide*. Peter Lang, Frankfurt.
- Baroni, Marco and Evert, Stefan (2008). Statistical methods for corpus exploitation. In *HSK 29.2 Corpus Linguistics*, Ch. 36.
- Evert, Stefan (2013). Tools for the acquisition of lexical combinatorics. In *HSK 5.4 Computational Lexicography*, Ch. 104.
- Evert, Stefan (2008). Corpora and collocations. In *HSK 29.2 Corpus Linguistics*, Ch. 58.
- Hardie, Andrew (unpubl.). A single statistical technique for keywords, lockwords, and collocations. Technical report.

… and many recent textbooks on statistical methods for (corpus) linguistics by Stefan Gries, Guillaume Desagulier, Bodo Winter, Vaclav Brezina, Sean Wallis, Gerold Schneider & Max Lauber.

https://dlf.uzh.ch/openbooks/statisticsforlinguists/

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

An evaluation study

KEYWORDS

Keyword extraction techniques

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

- 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: $\mathbf{LR_{cons}}$ conservative LR (Evert 2022)
  - lower bound of confidence interval (Hardie's formula)
  - with Bonferroni correction

Keyword extraction techniques

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

- 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)

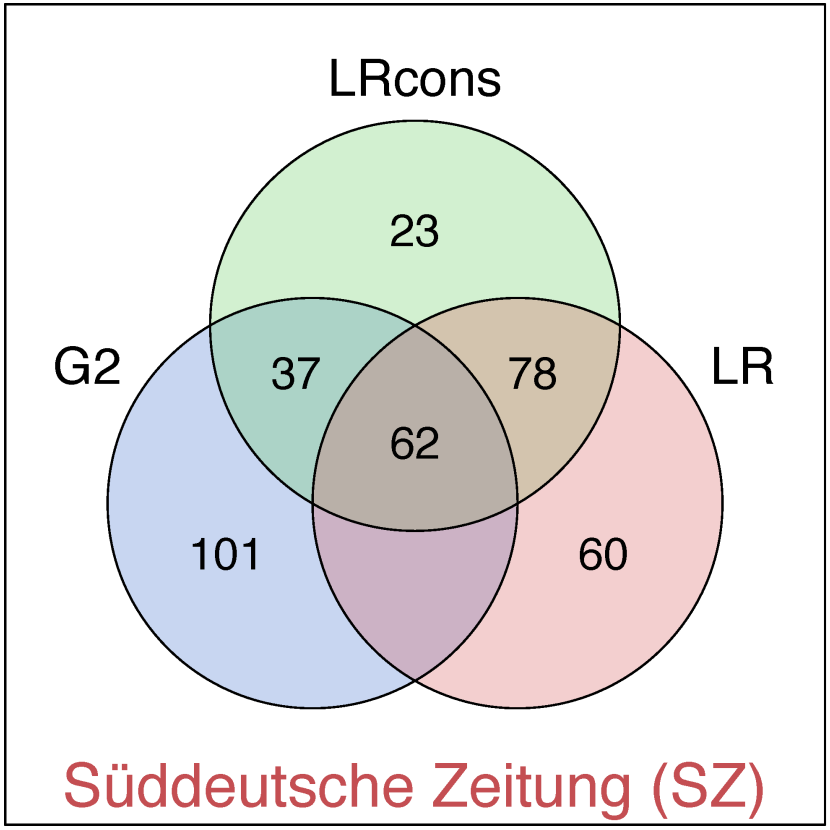
Corpus

- 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
- Annotated manually with metadata!
- Target subcorpus: 1.3M tokens (1,177 texts) of mass media texts and reader comments
  - Actor – author: media
  - Actor – reader: general public
  - Topic: MRO

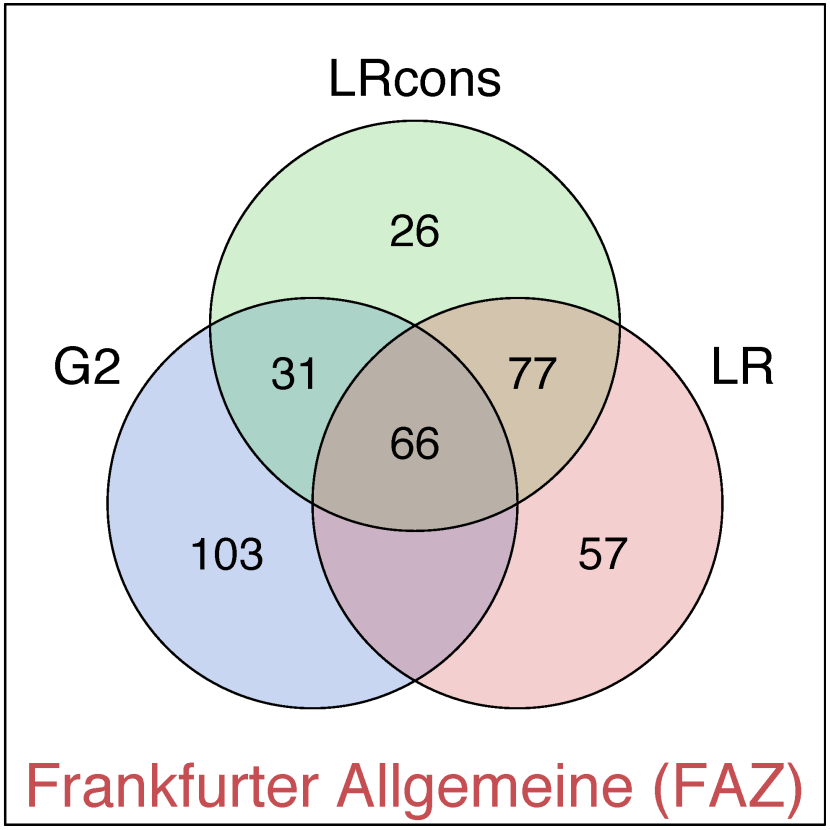
Experiments

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

Overlap between techniques



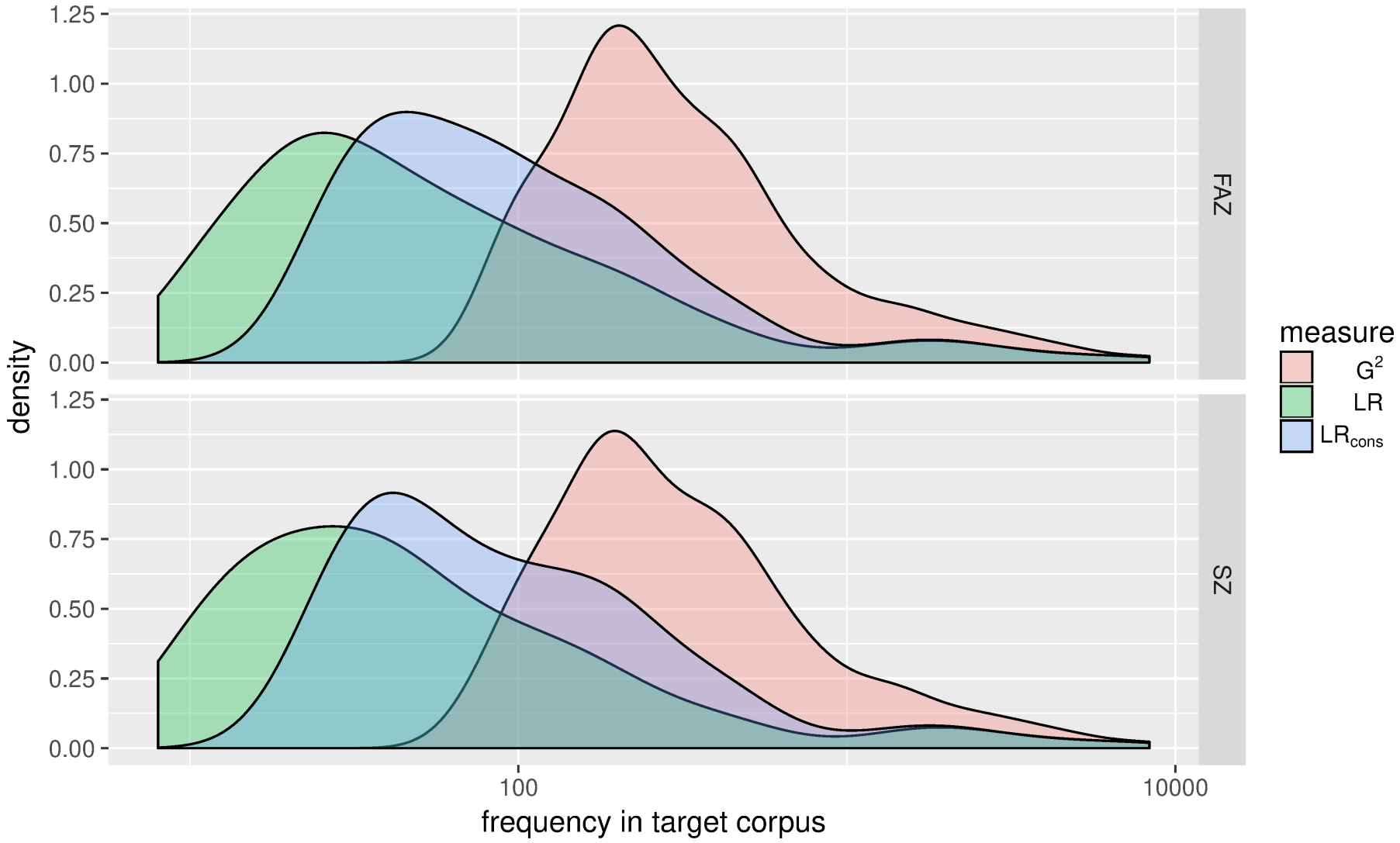
Süddeutsche Zeitung (SZ)



Frankfurter Allgemeine (FAZ)

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

Frequency bias



Manual annotation

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)

Manual annotation

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

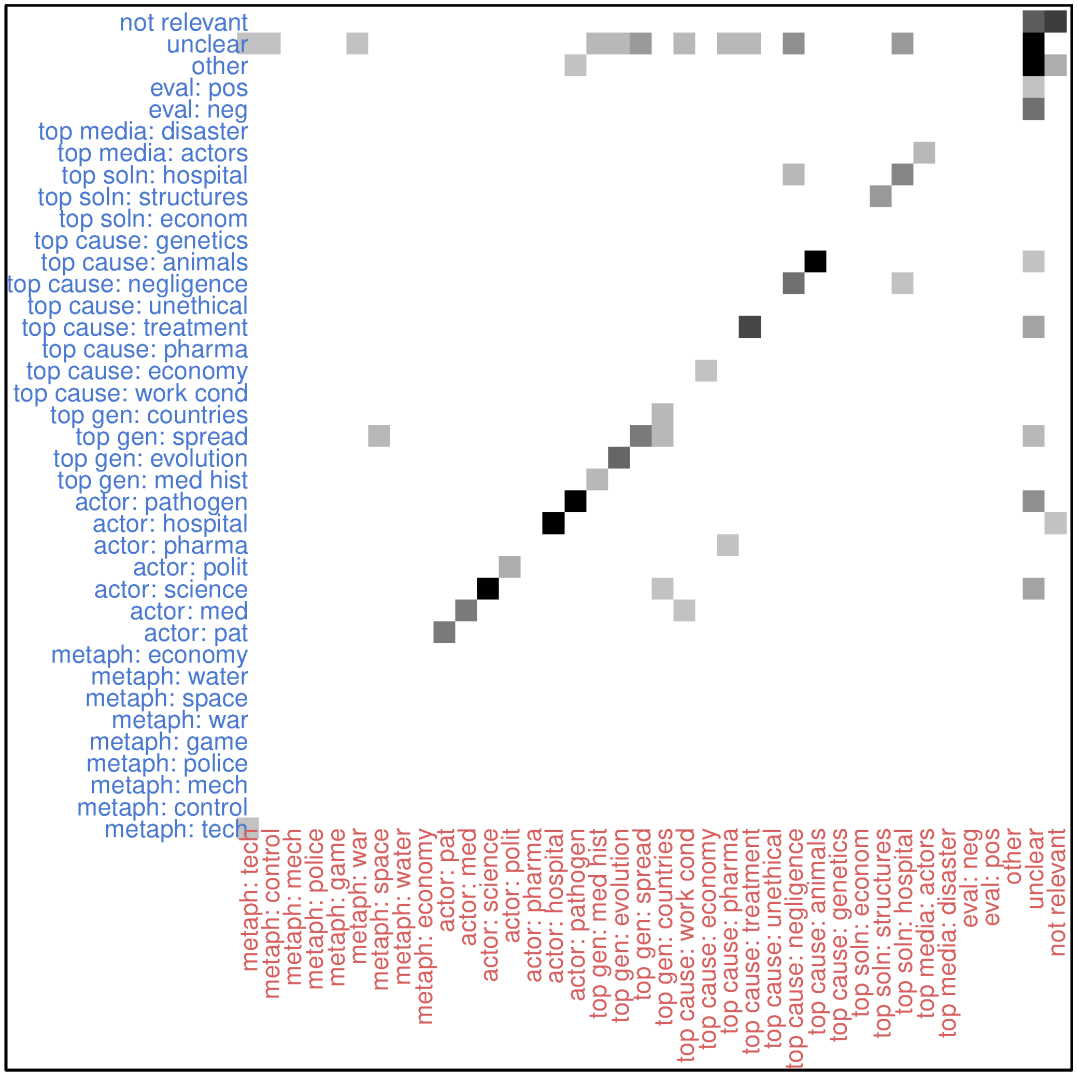
Agreement

- 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

IAA

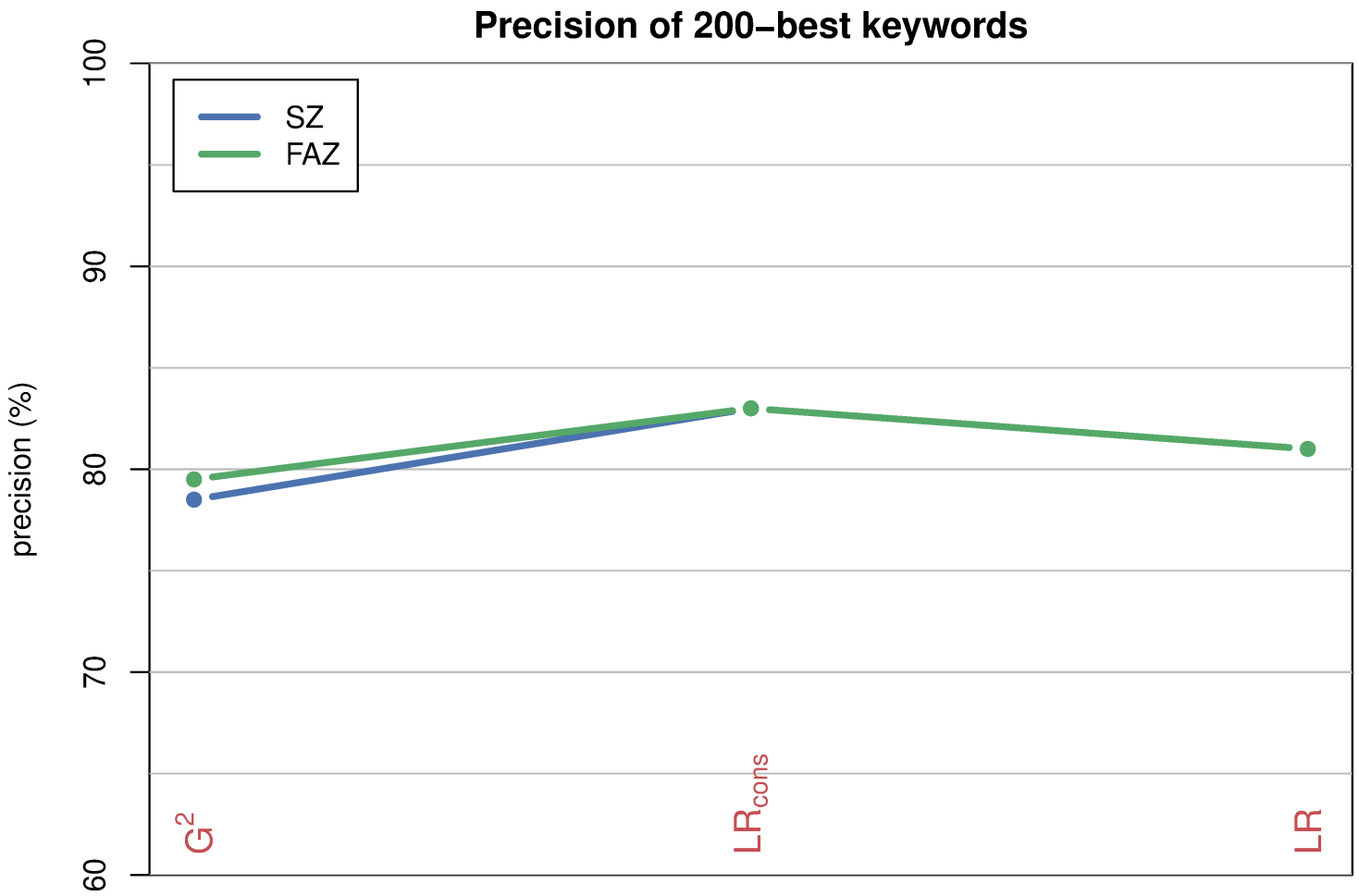
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Confusion matrix (primary category)



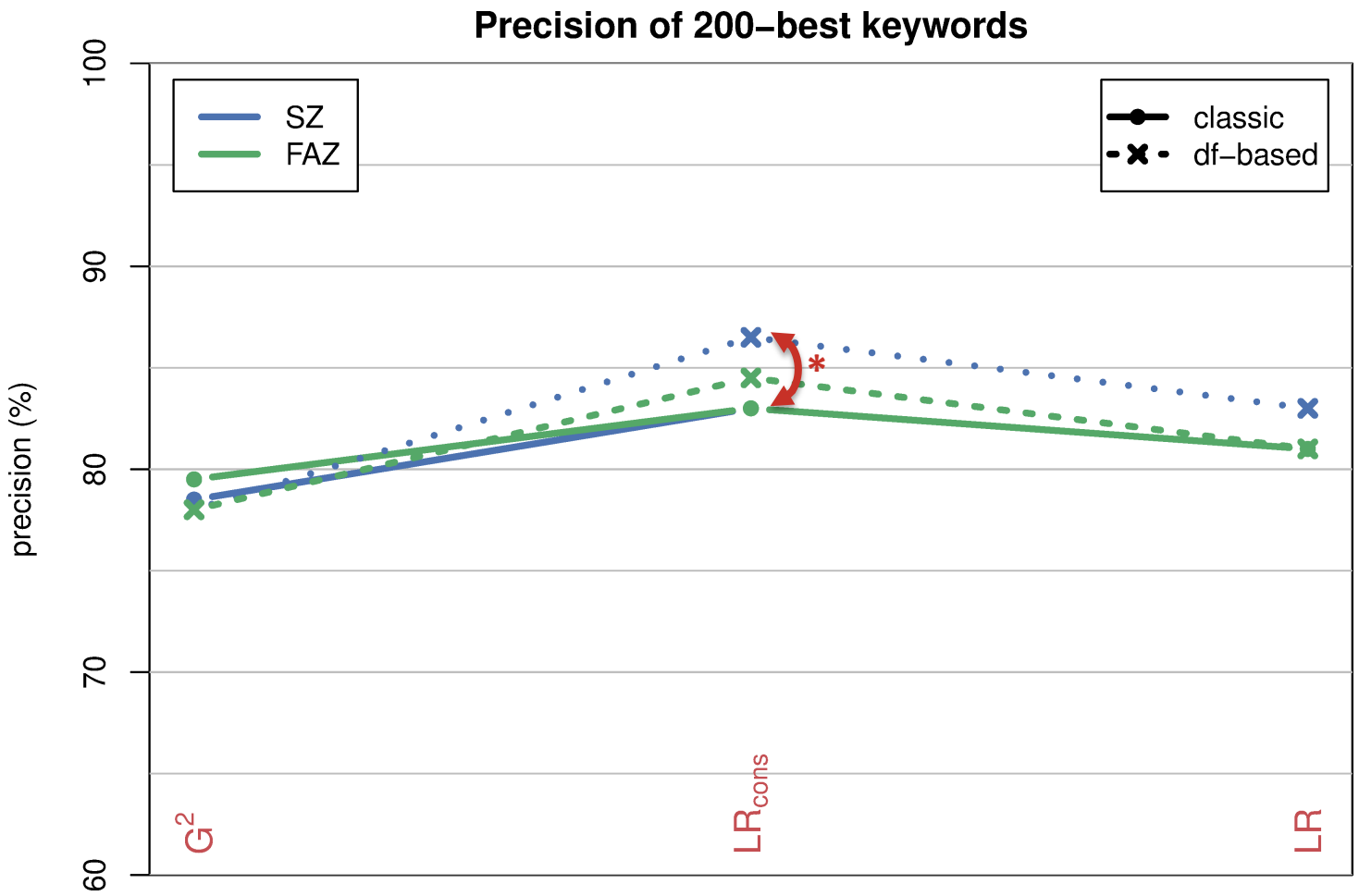
Precision = #TP / 200 cand.

TP = assigned to category and/or evaluative



Precision = #TP / 200 cand.

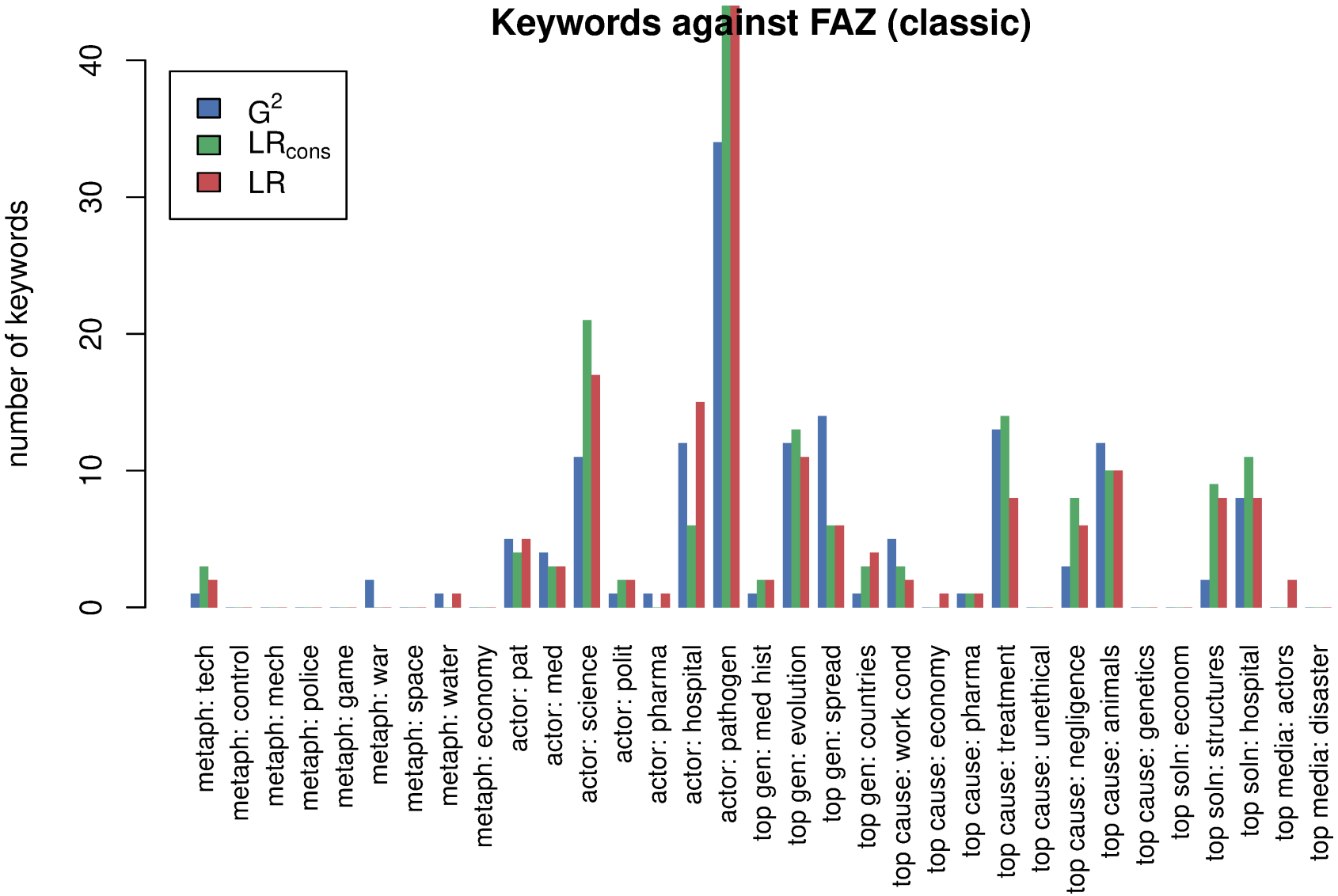
TP = assigned to category and/or evaluative



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

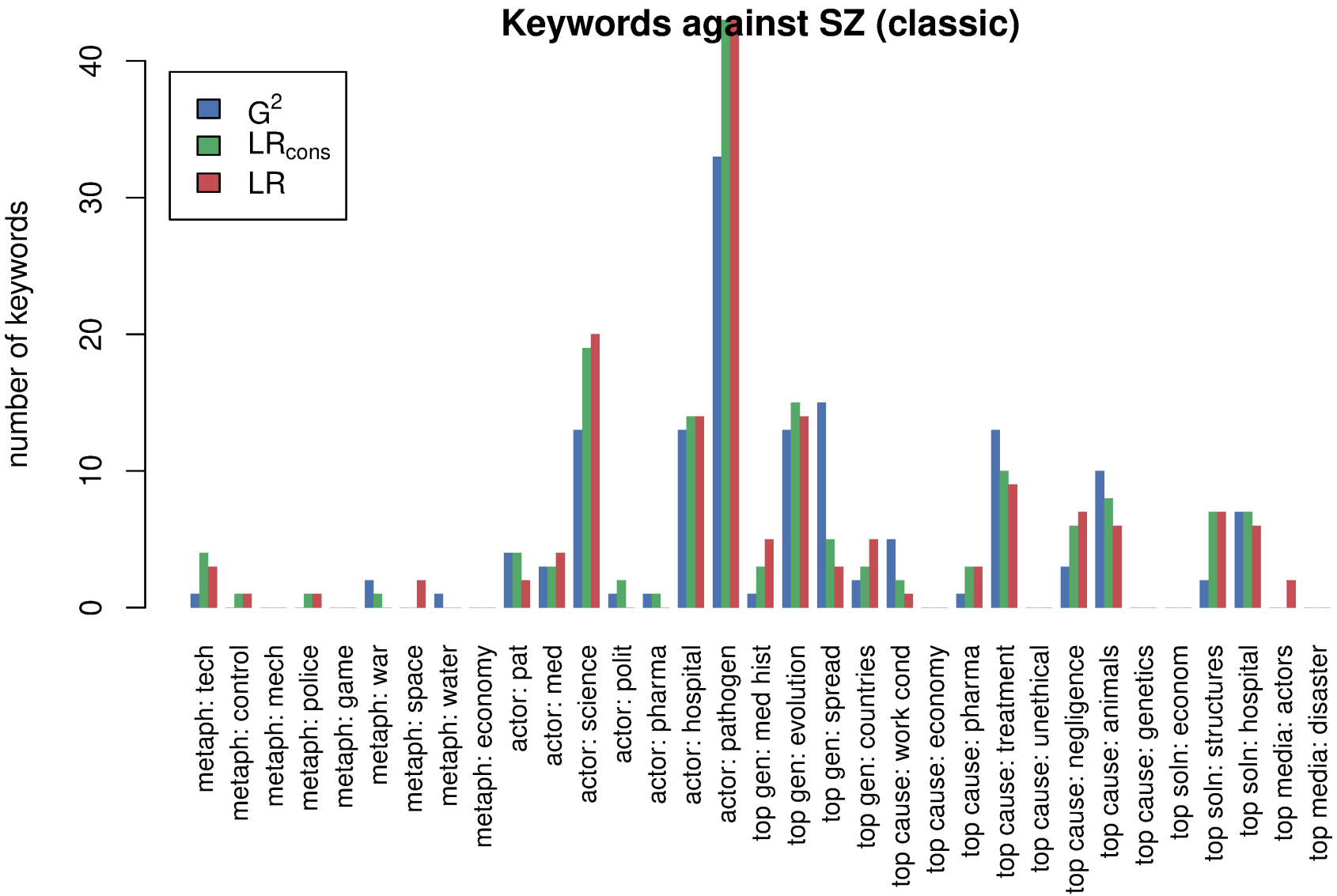
Recall = #kw for each category



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

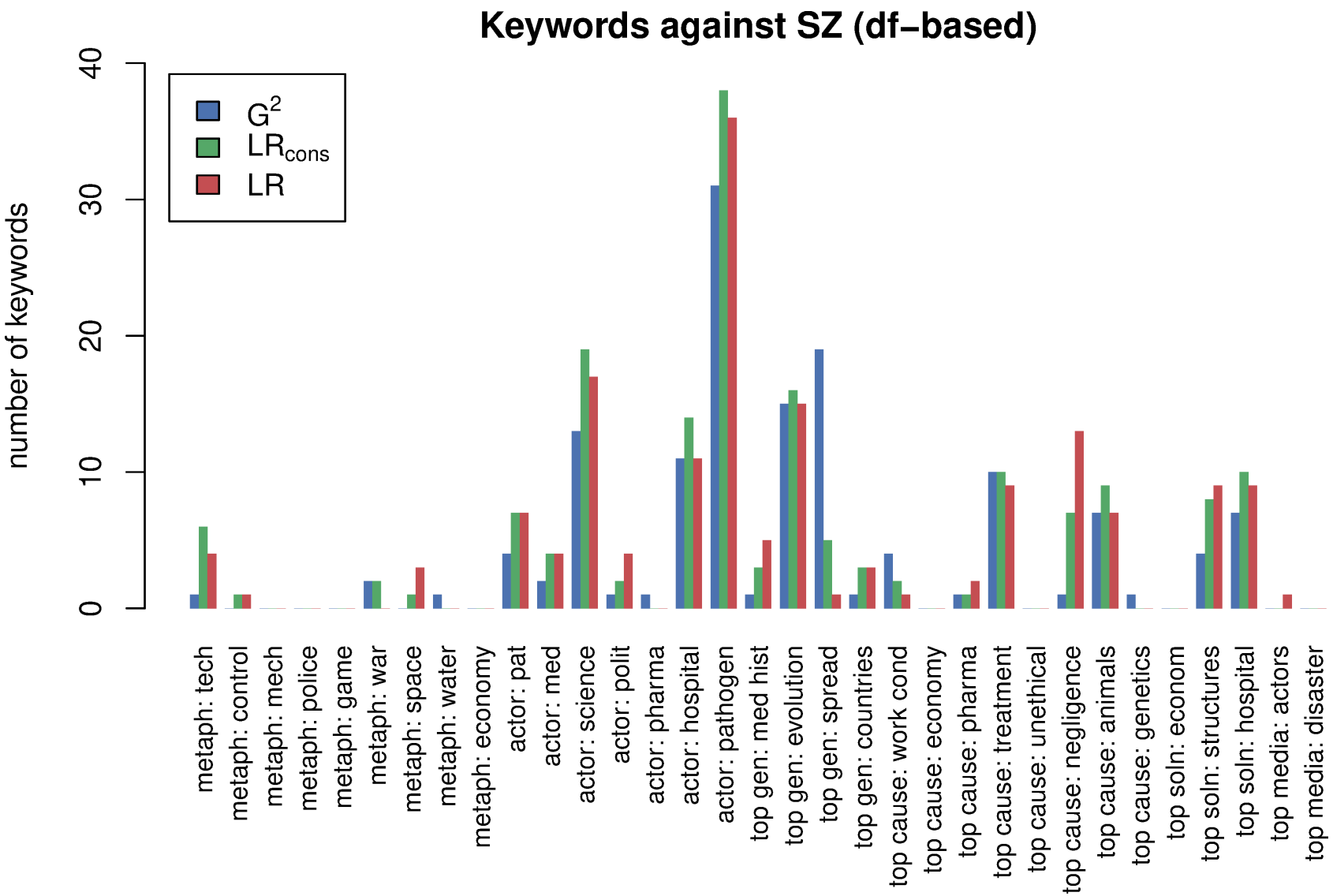
Recall = #kw for each category



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Recall = #kw for each category



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

An evaluation study

COLLOCATIONS

Collocations

- Corpus linguistics: collocations as (statistically) salient co-occurrence patterns of words
  - most common perspective: single-word collocates of a "node" word or phrase, ranked by statistical salience
  - evidence for lexicalised multiword expressions (MWE)
  - typical usage patterns (➞ discourse analysis, CALL, …)
  - indicative of word meaning and different word senses (in NLP: "one sense per collocation")
  - widely used in computational lexicography
- Different types of co-occurrence: surface span, textual unit, syntactic relation (Evert 2008)

Word sketch = syntactic co-occurrence

cat British National Corpus freq = 5381

object_of 964 2.0		and/or 1056 1.7	
skin	9 7.91	dog	208 8.49
diddle	7 7.85	cat	68 8.01
stroke	10 7.09	kitten	13 8.01
torture	5 6.57	fiddle	9 7.71
feed	22 6.34	mouse	29 7.68
rain	4 6.3	monkey	15 7.55
chase	9 6.27	budgie	4 6.74
rescue	7 6.15	rabbit	12 6.48

pp_obj_like-p 106 28.9	
grin	11 7.63
fight	9 4.62
smile	4 4.24
look	11 2.04

pp_among-p 17 14.8	
pigeon	15 8.66

possessor 91 1.9	
Schrödinger	8 10.87
witch	4 6.82
gardener	4 6.0
Henry	8 4.91
neighbour	5 4.28

possession 232 4.7	
cradle	24 9.91
whisker	9 8.92
paw	5 7.44
fur	9 7.14
tray	4 5.34
tail	5 4.91
tongue	5 4.89
ear	5 4.0

subject_of 842 3.3	
purr	7 7.76
miaow	5 7.57
mew	4 7.18
jump	20 6.95
scratch	8 6.84
leap	10 6.78
stalk	4 6.56
react	4 5.33

adj_subject_of 142 2.6	
asleep	4 6.09
alive	4 5.06
concerned	4 2.94
black	4 2.36
likely	4 1.96

pp_obj_of-p 324 1.3	
moral	4 7.06
breed	6 5.77
signal	4 3.89
sight	4 3.77
species	5 3.36
game	9 3.14
picture	6 2.99
death	7 2.71

modifier 1622 1.2		
pussy	76	10.42
Cheshire	45	8.9
stray	25	8.7
siamese	17	8.35
tabby	17	8.35
wild	53	7.94
pet	31	7.92
tom	12	7.8

modifies 610 0.5	
flap	16 8.39
litter	15 8.15
phobia	5 7.64
burglar	8 7.55
faeces	6 7.47
assay	10 7.38
Hastings	7 6.91
scan	4 6.59

adjective-noun
(pre-nominal modifier)

Syntactic co-occurrence

In an *open barouche* [. . .] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots; two *young ladies* in scarfs and feathers; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers; a lady of *doubtful age*, probably the aunt of the aforesaid; and [. . .]

f(young, gentleman)

	•|gent.	•|¬gent
young|•		
¬young|•		

Syntactic co-occurrence

In an *open barouche* [. . .] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots; two *young ladies* in scarfs and feathers; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers; a lady of *doubtful age*, probably the aunt of the aforesaid; and [. . .]

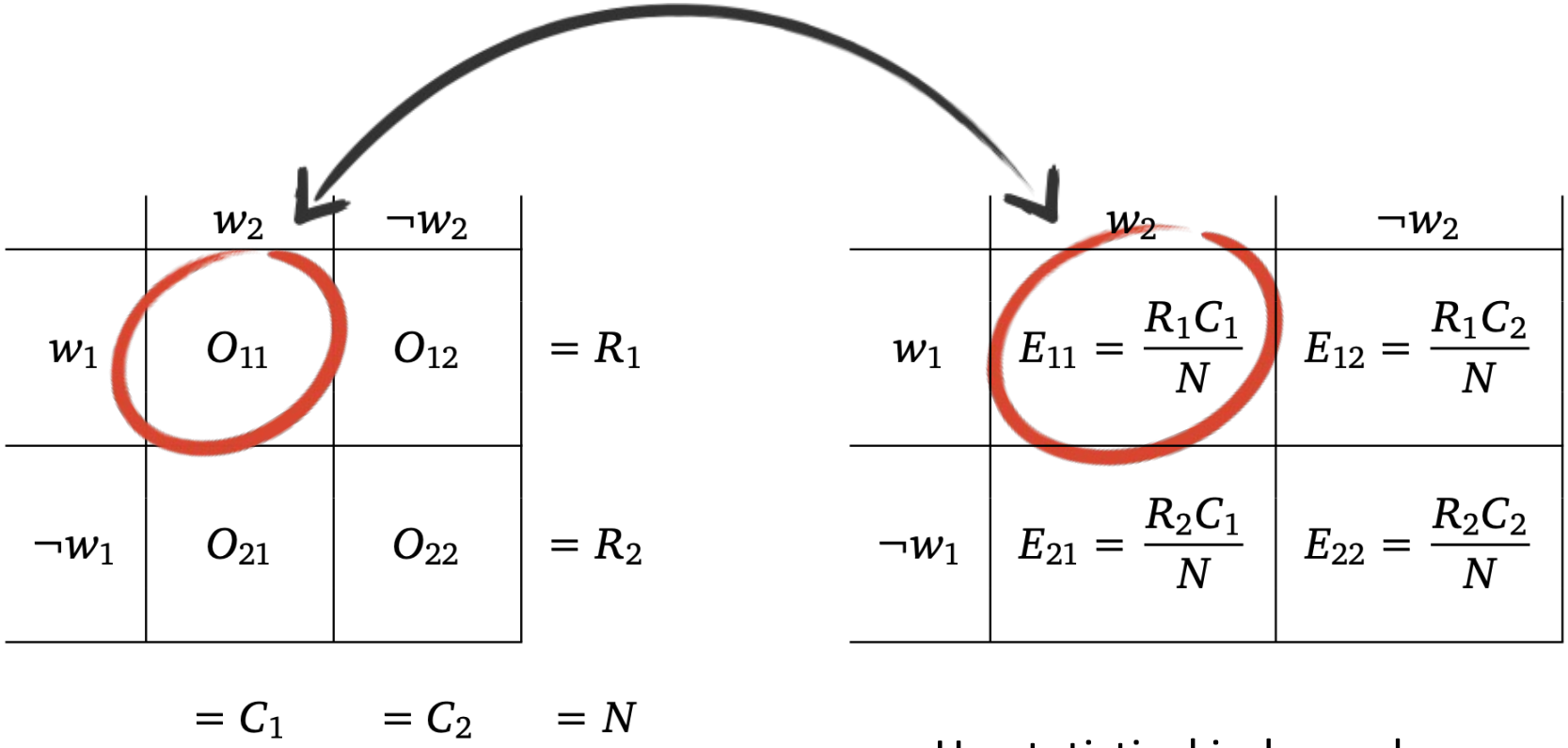
→

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

f(young, gentleman) = 1
sample size N = 9

	•|gent.	•|¬gent	
young|•	1	2	3
¬young|•	2	4	6
	3	6	9

Statistical association measures (AM)



	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	$= R_1$
$\neg w_1$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

	w_2	$\neg w_2$
w_1	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$\neg w_1$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

H_0: statistical independence

$\pi_{cooc} = \pi_1 \times \pi_2$

observed

expected

Statistical association measures (AM)

recommended measures (Evert 2008)

$$\text{MI} = \log_2 \frac{O}{E}$$

$$\text{MI}^k = \log_2 \frac{O^k}{E}$$

$$\text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}}$$

$$\text{t-score} = \frac{O - E}{\sqrt{O}}$$

$$\text{simple-ll} = 2 \left(O \cdot \log \frac{O}{E} - (O - E) \right)$$

$$\text{chi-squared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\text{chi-squared}_{\text{corr}} = \frac{N\left(|O_{11}O_{22} - O_{12}O_{21}| - N/2\right)^2}{R_1 R_2 C_1 C_2}$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

$$\text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}$$

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1}$$

$$\text{odds-ratio} = \log \frac{\left(O_{11} + \frac{1}{2}\right)\left(O_{22} + \frac{1}{2}\right)}{\left(O_{12} + \frac{1}{2}\right)\left(O_{21} + \frac{1}{2}\right)}$$

Sketch Engine™

Evaluating collocation identification (Evert et al. 2017)

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

- Many parameters for collocation extraction, but which settings are most suitable?
- Here: broad-scale comparative evaluation
  - Which AM correlates best with collocativity?
  - What is an appropriate co-occurrence context?
  - Which source corpora provide the best results?
  - Does size matter? Or representativeness?
  - Are there interactions between these parameters?
  - Are crawled Web corpora and n-gram databases a viable substitute for expensive reference corpora?

Gold standard

- BBI: The BBI Combinatory Dictionary of English (Benson, Benson & Ilson 1986)
  - based on lexicographic native-speaker intuitions
  - pre-corpus era ➞ no bias towards specific method/corpus

- OCD2: Oxford Collocations Dictionary for students of English, 2nd ed. (McIntosh, Francis & Poole 2009)
  - corpus-based, much more comprehensive
  - clearer notion of collocation (≈ our subtypes of MWE)

The Bartsch224 gold standard

- Set of 203 node words selected by Sabine Bartsch
  - original set contained approx. 224 node words
  - some obscure nodes with few collocates omitted
- Manually extracted all lexical words (nouns, verbs, adjectives, adverbs) from corresponding BBI entries
  - set of 2,845 node-collocate pairs
  - lemmatized, reduced to two-word collocations
- Automatic extraction from XML version of OCD2
  - also from other entries (our node word listed as collocate)
  - set of 18,545 node-collocate pairs

Gold standard example: BBI

Node: measure (noun or verb)

👉 cubic, dry, liquid, metric, tape, certain, good, make, take,

measure I *n.* 1. a cubic; dry; liquid; metric ~ 2. a tape ~ 3. in a certain ~ (in large ~) 4. (misc.) for good ~ ('as smt. extra'); made to ~ ('custom-made'); to take smb.'s ~ ('to evaluate smb.') (see also **measures**)

measure II *v.* 1. (d; tr.) to ~ against (to ~ one's accomplishments against smb. else's) 2. (P; intr.) the room ~s twenty feet by ten

Gold standard example: BBI

Node: measure (noun or verb)

👉 cubic, dry, liquid, metric, tape, certain, good, make, take, carry, coercive, compulsory, draconian, drastic, harsh, stern, stringent, emergency, extreme, radical, preventive, prophylactic, safety, security, stopgap, temporary

measures *n.* 1. to carry out, take ~ 2. coercive; compulsory; draconian; drastic, harsh, stern, stringent; emergency; extreme, radical; preventive, prophylactic; safety, security; stopgap, temporary ~ 3. ~ to + inf. (we took ~ to insure their safety) 4. ~ against (to take ~ against smuggling)

Gold standard example: OCD

Node: measure (noun or verb)

- ~~cubic~~, ~~dry~~, ~~liquid~~, metric, tape, ~~certain~~, ~~good~~, ~~make~~, take, ~~carry~~, ~~coercive~~, ~~compulsory~~, draconian, drastic, harsh, ~~stern~~, ~~stringent~~, emergency, extreme, radical, preventive, ~~prophylactic~~, safety, security, stopgap, temporary
- ability, able, accurate, accurately, achievement, activity, additional, adopt, aim, angle, appropriate, approve, austerity, autonomy, ballot, brandy, broad, calculate, carefully, change, circumference, composition, conservation, considerable, control, corrective, cost-cutting, crude, cup, defeat, defensive, density, derive, and 158 more

Parameters: association measure

- Mutual Information (MI, Church & Hanks 1990)
- t-score (t, Church et al. 1991)
- MI^2, MI^3, MI^4 (Daille 1994) + MI_{conf} (Johnson 2001)
- chi-squared (X^2) and z-score (z) with Yates correction
- Dice (SketchEngine), Jaccard coefficient
- minimum sensitivity (MS, Pedersen & Bruce 1996)
- odds ratio (log θ, log θ_{disc}), relative risk (log *r*)
- log-likelihood (G^2, Dunning 1993)
- ΔP (Gries 2013) in 4 variants (fwd, bwd, min, max)
- co-occurrence frequency (f)

Parameters: co-occurrence context

- syntactic co-occurrence: dependency relations
  - direct dependency (all types, both directions)
- surface co-occurrence: L1 / R1
- surface co-occurrence: L2 / R2
- surface co-occurrence: L3 / R3
- surface co-occurrence : L5 / R5
- surface co-occurrence : L10 / R10
- textual co-occurrence: sentence

Parameters: corpus & annotation

corpus	annotation	size
British National Corpus (BNC, Aston & Burnard 1998)	C&C, Stanford	0.1 G
Darmstadt English Movie Subtitle Corpus (DESC)	C&C, Stanford	0.1 G
Gigaword newspaper corpus (2nd ed.)	C&C, Stanford	2.0 G
English wikipedia of 2009 (Wackypedia)	C&C, Malt, Stfd	1.0 G
Subcorpus WP500 (500 words per article)	C&C, Malt, Stfd	0.2 G
Web corpus ukWaC (Baroni et al. 2009)	C&C, Malt	2.0 G
Web corpus WebBase (Han et al. 2013)	C&C	3.0 G
Web corpus UKCOW 2012 (Schäfer et al. 2012)	C&C	4.0 G
Web corpus ENCOW 2014	C&C, Malt	10.0 G
All Web corpora + Wackypedia (JOINT)	C&C	16.0 G

Parameters: corpus & annotation

corpus	annotation	size
Google Web 1T 5-Grams (Web1T5, Brants & Franz 2006)		1000 G
Google Books N-Grams 2012 (BooksEN, Lin et al. 2012)*	parsed	500G
Google Books N-Grams 2012 GB (BooksGB)*	parsed	50 G

* Google Books data sets only include n-gram counts from contemporary books published in 1980 and later (evaluation on full 20th century yields very similar results)

Evaluation methodology

node	collocate	G^2	BBI?
measure	introduce	762.60	—
measure	take	600.13	++
measure	measure	510.33	—
measure	preventive	496.79	++
measure	success	475.13	—
measure	use	450.28	—
measure	concentration	428.35	—
measure	safety	413.83	++
measure	adopt	397.25	—
measure	distance	337.12	—

most strongly associated co-occurrences for node measure (n = 10)

P = 3 TP / 10 cand. = 30%

R = 3 / 26 TP = 11.5%

TP = true positive
(according to BBI dictionary)

Evaluation methodology

- Precision / recall of n-best lists for each node
  - strategy used by Uhrig & Proisl (2012)
  - task: determine most salient collocates for given node
  - results averaged over all 203 nodes

- Precision vs. recall for list of all candidate pairs
  - strategy used by Bartsch & Evert (2014), following Evert & Krenn (2001, 2005)
  - task: determine most salient collocational pairs

    ☞ we present results for this approach

Evaluation: global ranking

node	collocate	G^2	BBI?
minister	prime	111653.34	++
prime	minister	103587.58	—
authority	local	64395.65	++
take	place	43787.49	—
place	take	42551.75	++
set	up	37871.00	—
state	secretary	37588.70	—
door	open	37287.35	++
open	door	37193.15	—
head	shake	32301.90	++

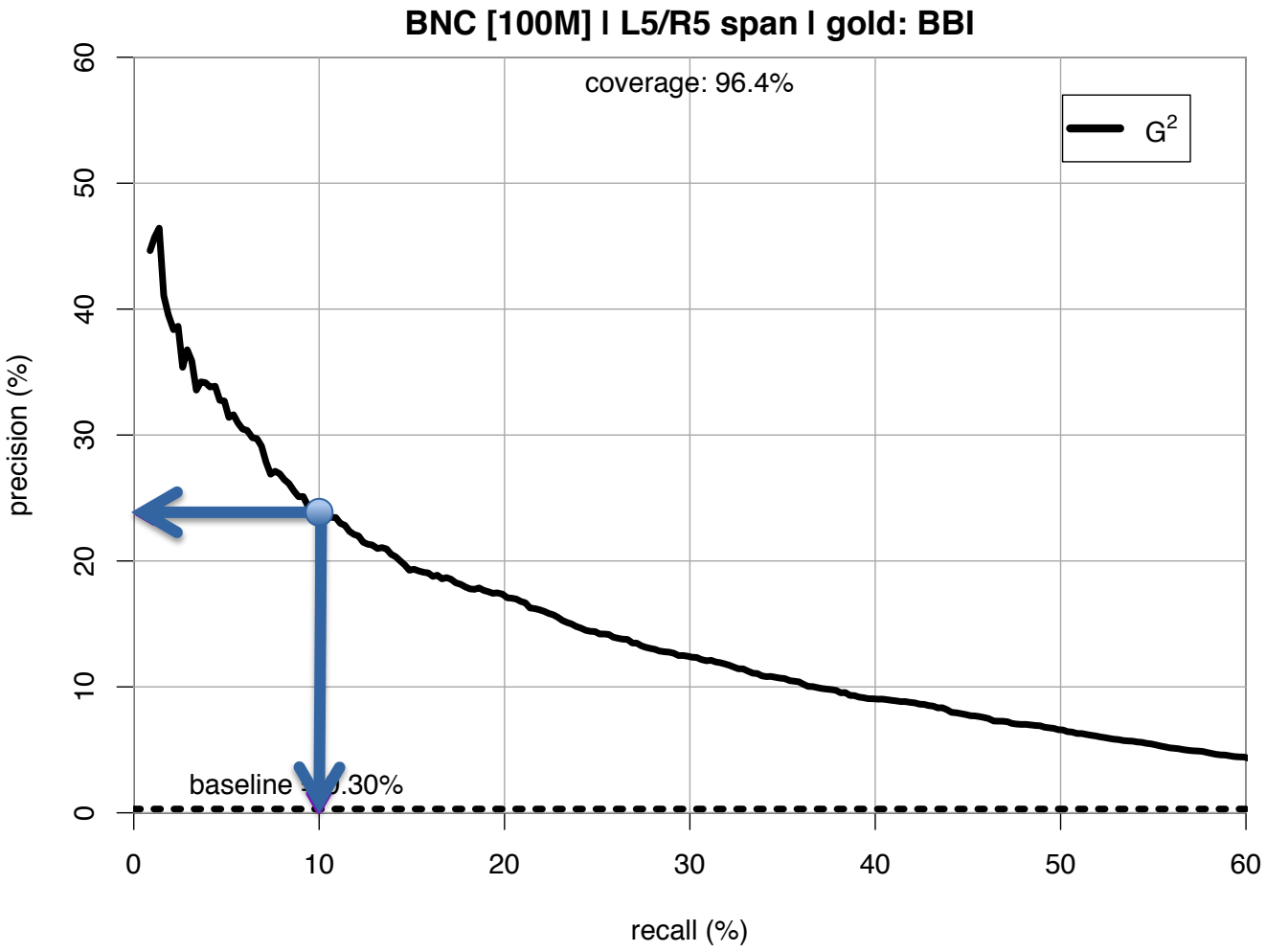
most strongly associated node-collocate pairs (n = 10)

P = 5 TP / 10 cand. = 50%

R = 5 / 2845 TP = 0.2%

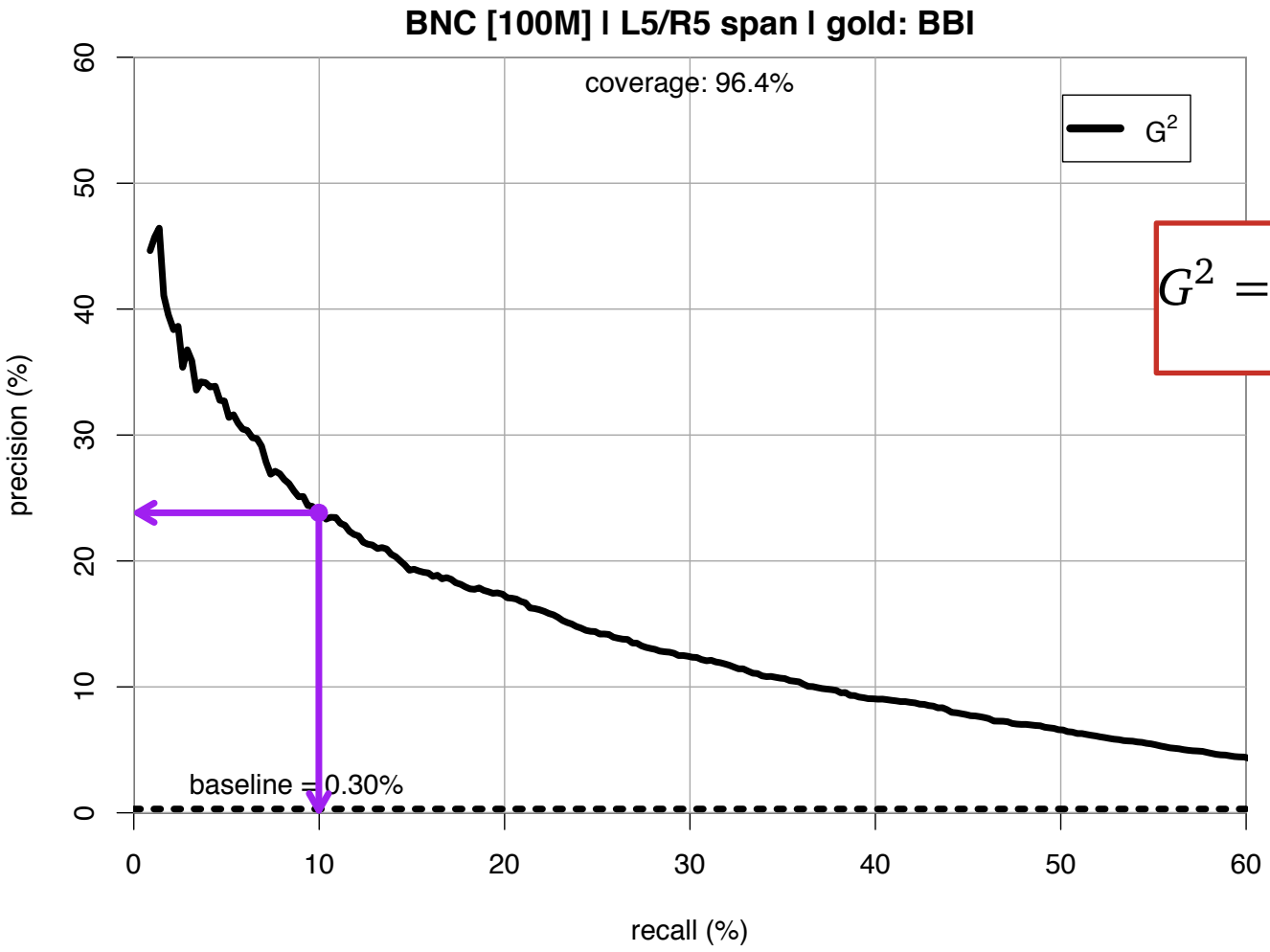
TP = true positive (according to BBI dictionary)

Evaluation: precision vs. recall | BBI



n = 1193
R = 10.0%
P = 23.9%

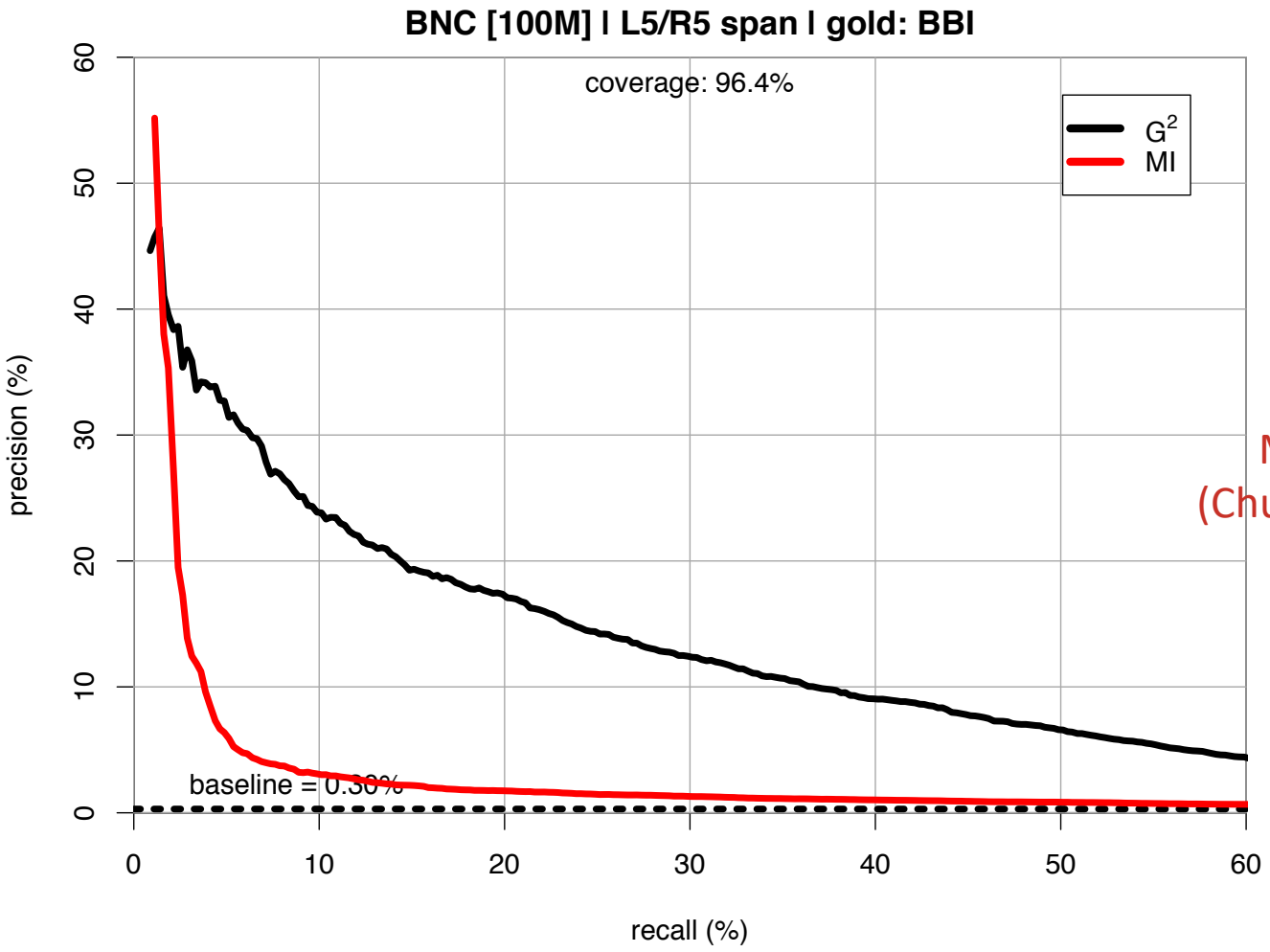
Evaluation: precision vs. recall | BBI



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

log-likelihood
(Dunning 1993)

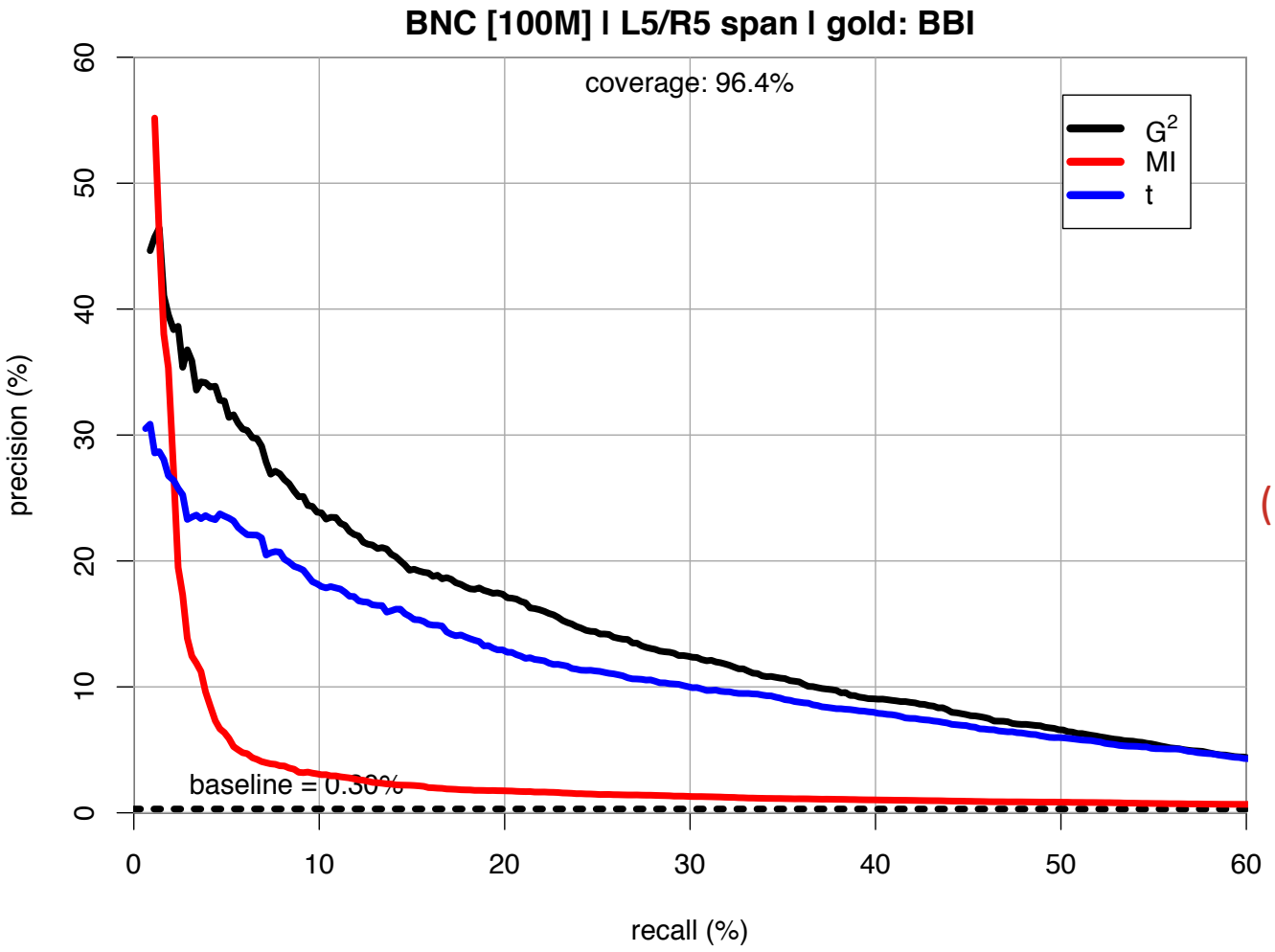
Evaluation: precision vs. recall | BBI



$$\mathrm{MI} = \log_2 \frac{O}{E}$$

Mutual Information
(Church & Hanks 1990)

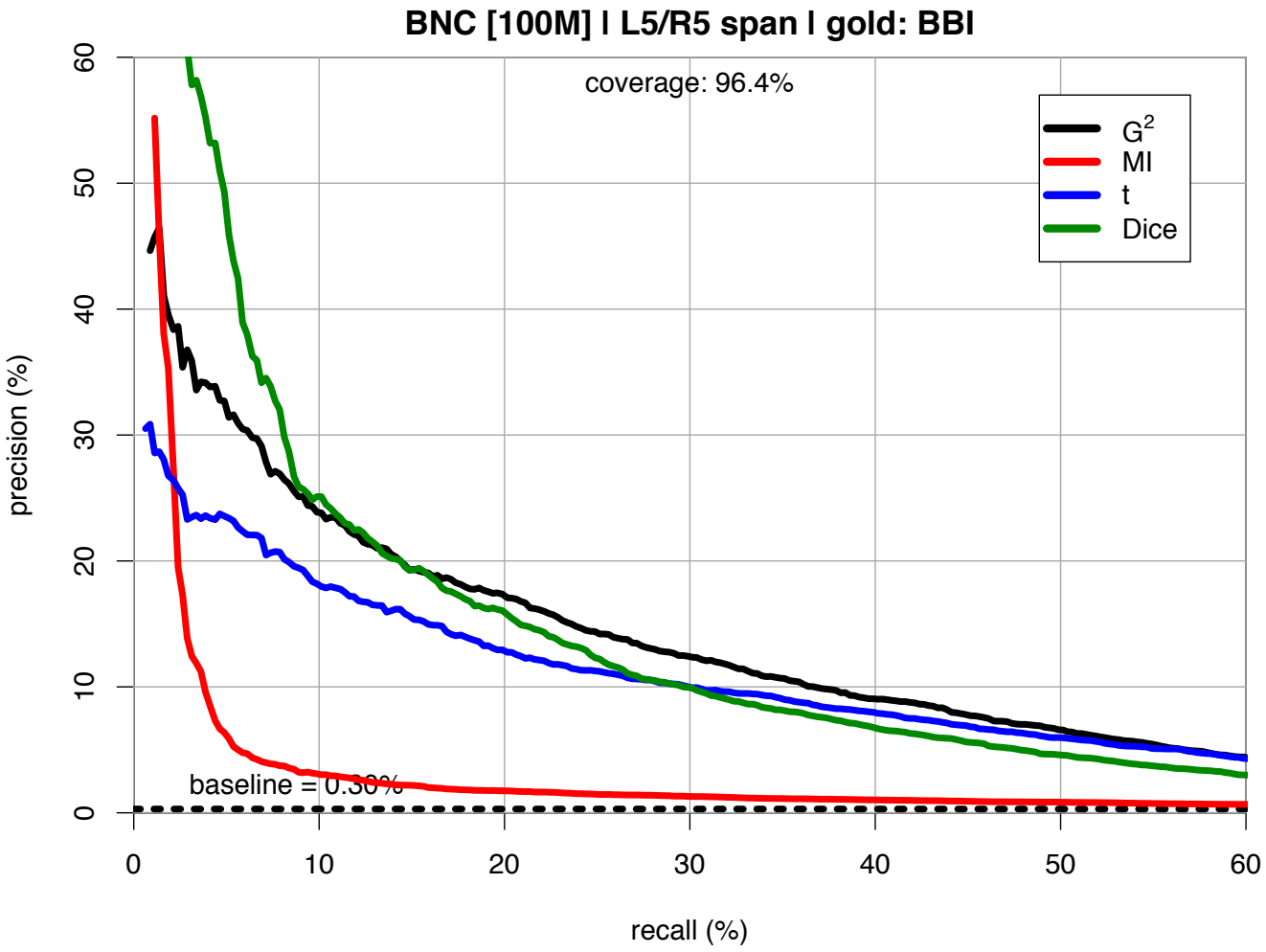
Evaluation: precision vs. recall | BBI



$$t = \frac{O - E}{\sqrt{O}}$$

t-score
(Church et al. 1991)

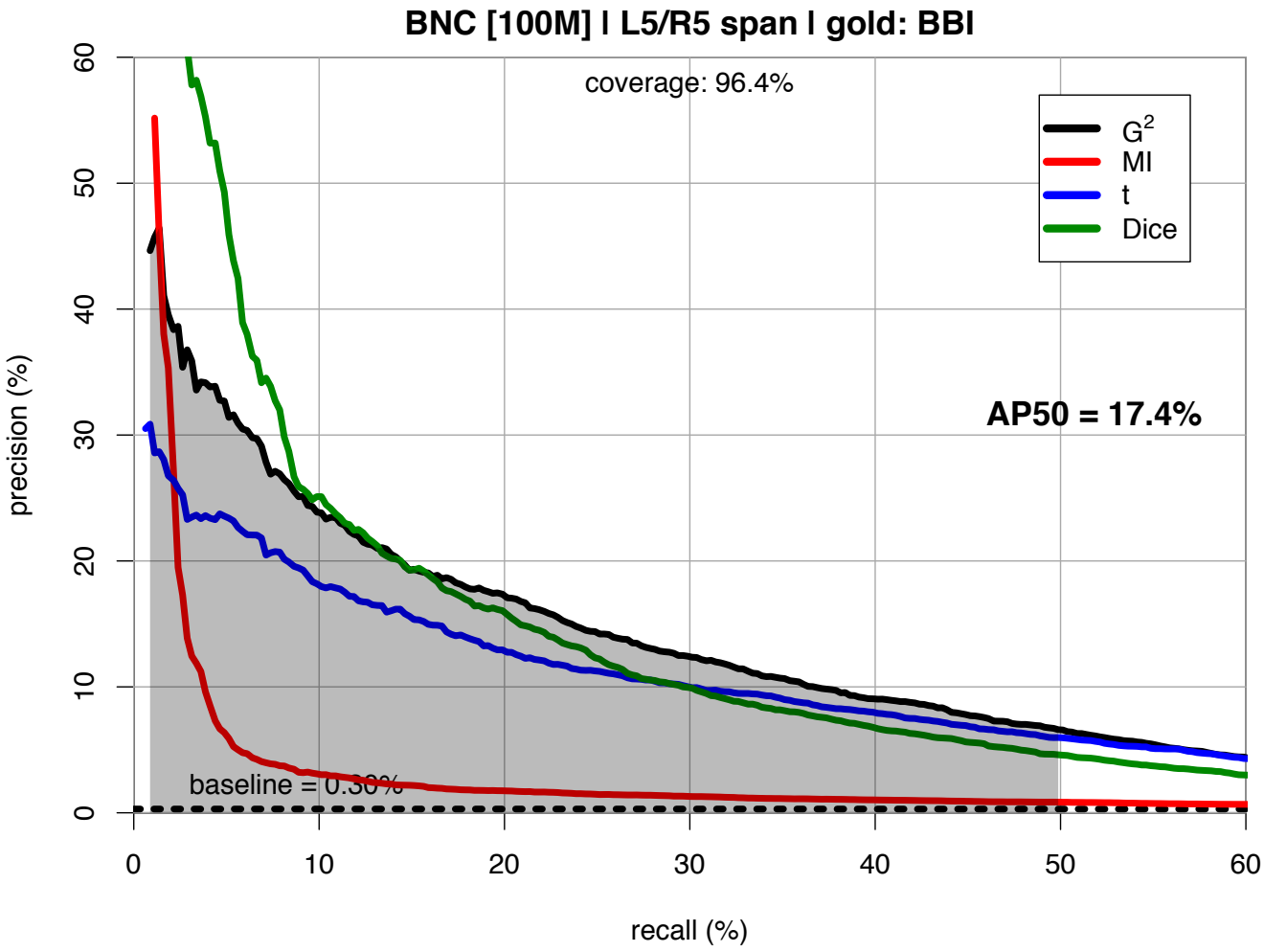
Evaluation: precision vs. recall | BBI



$$\text{Dice} = \frac{2O}{R_1 + C_1}$$

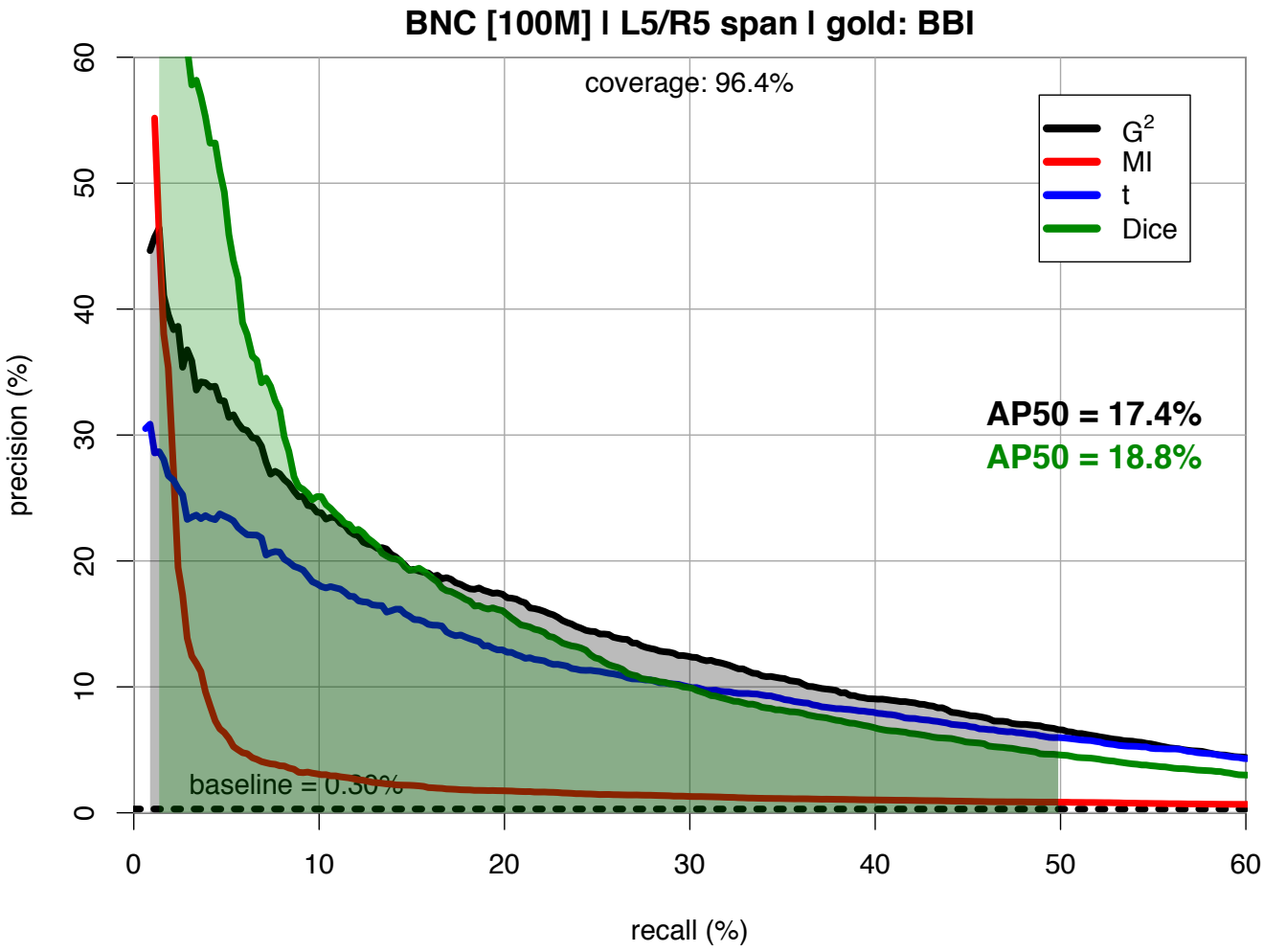
Dice coefficient
(Sketch Engine)

Evaluation: precision vs. recall | BBI



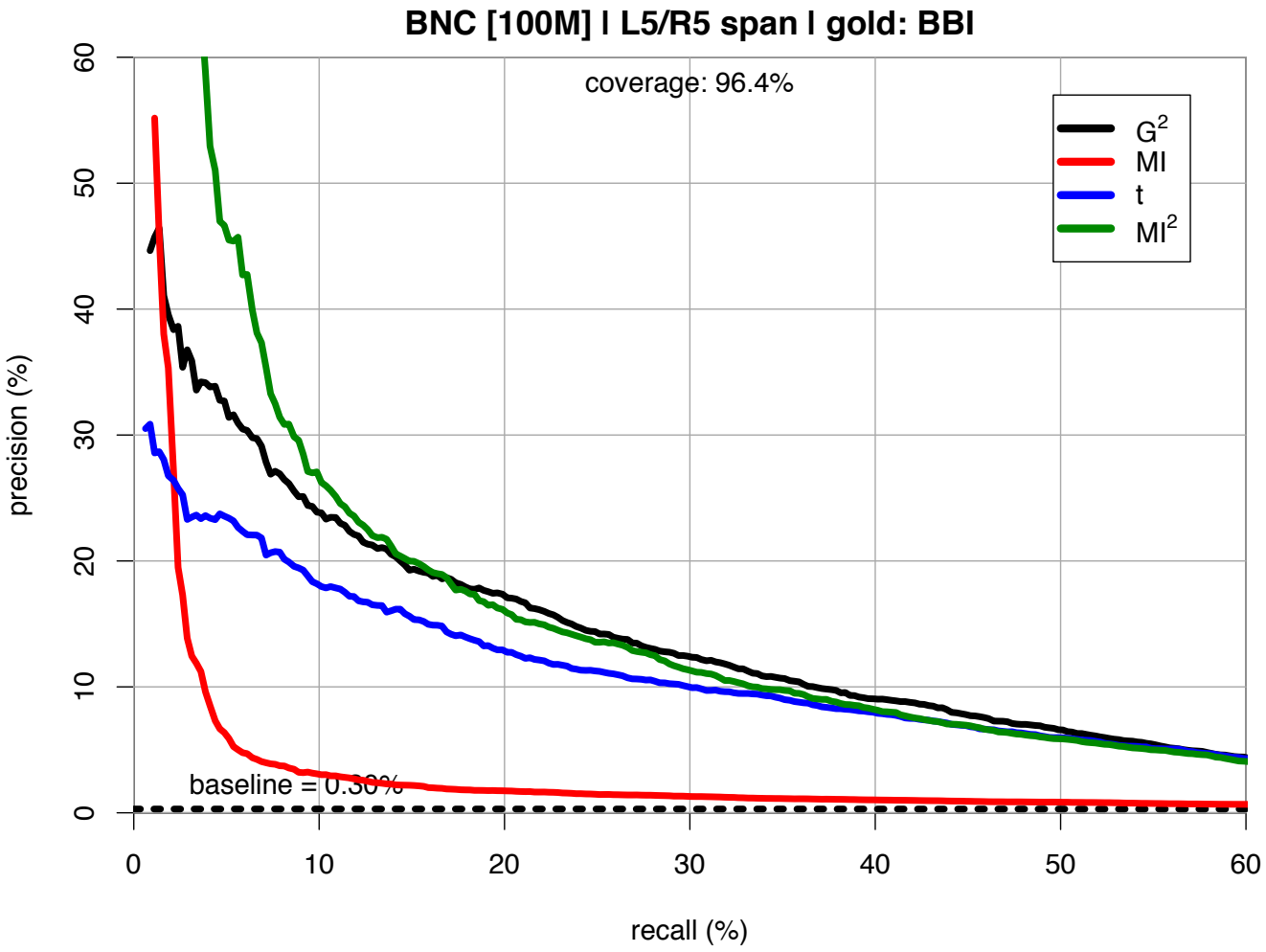
Average Precision
up to 50% recall

Evaluation: precision vs. recall | BBI



Average Precision
up to 50% recall

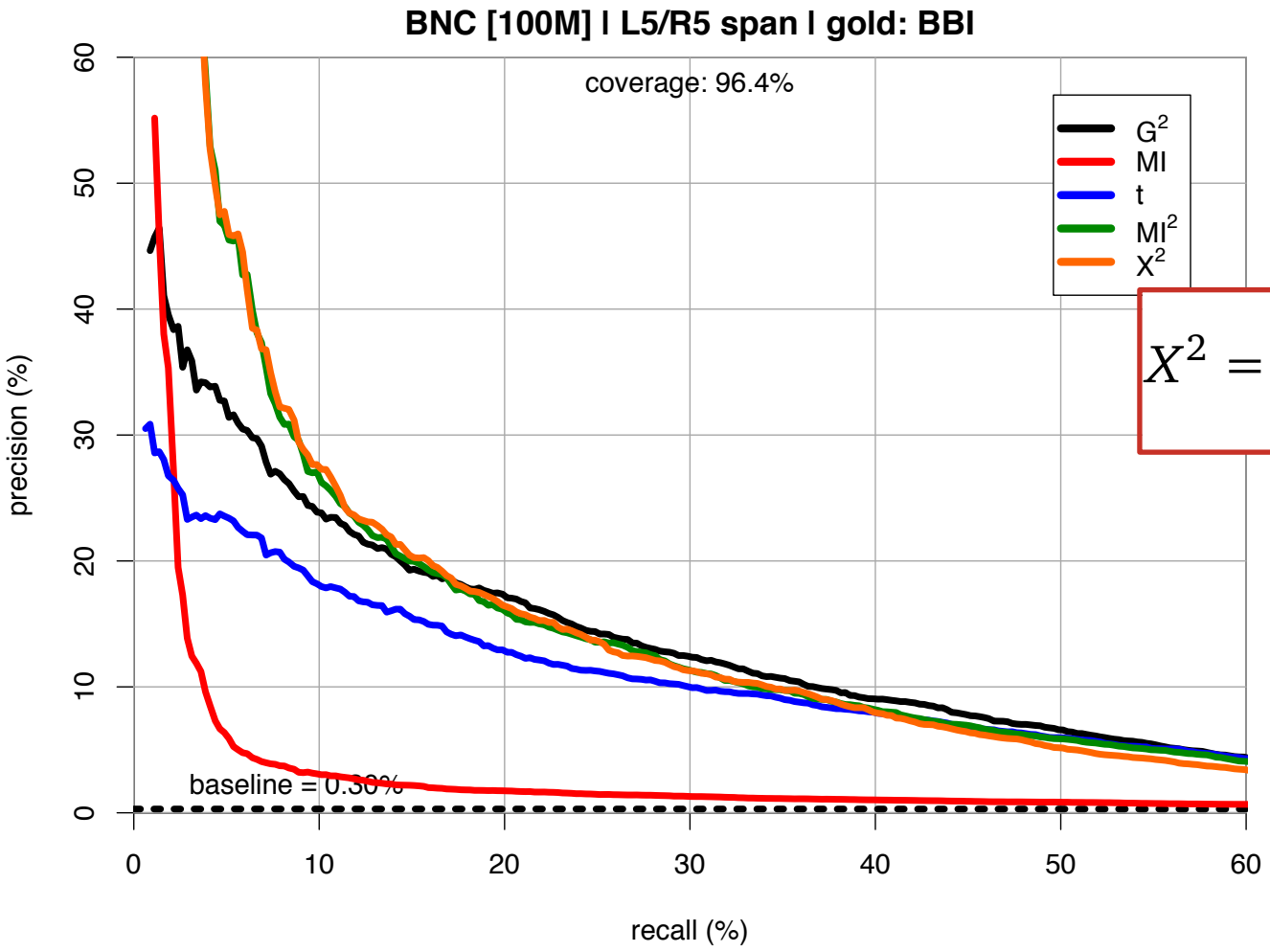
Evaluation: precision vs. recall | BBI



$$\mathrm{MI}^2 = \log_2 \frac{O^2}{E}$$

heuristic measure
(Daille 1994)

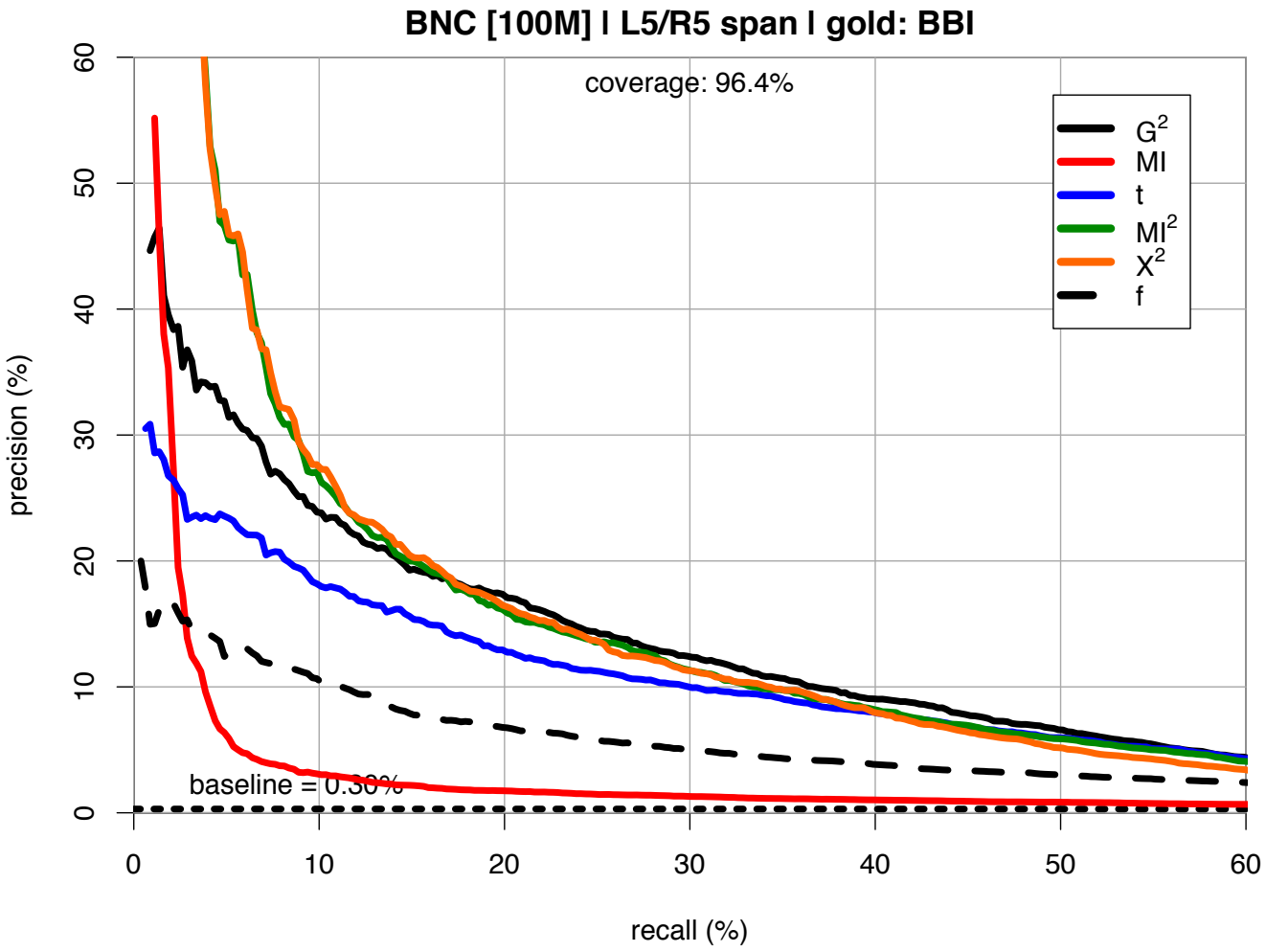
Evaluation: precision vs. recall | BBI



$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

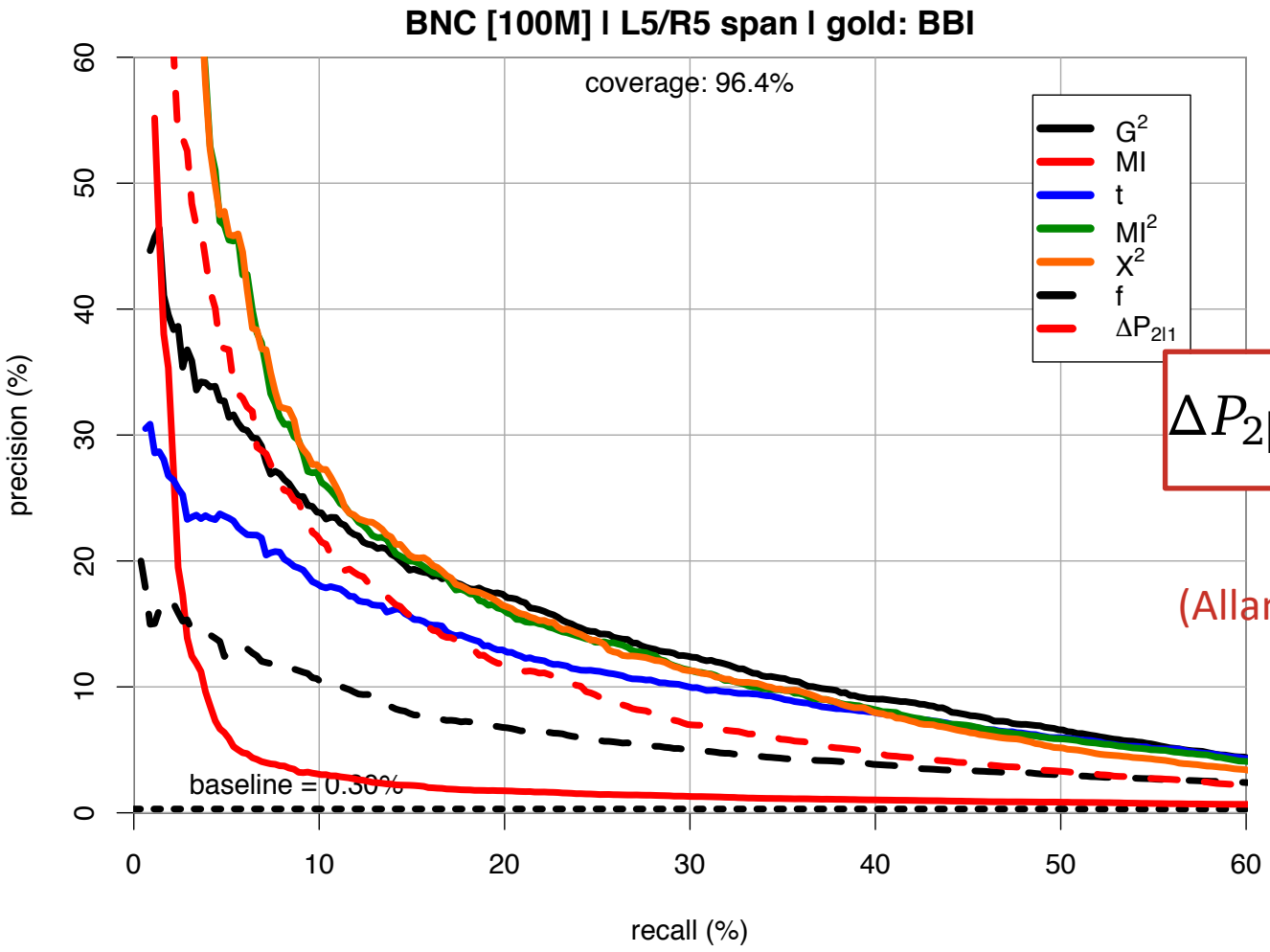
standard
chi-squared test

Evaluation: precision vs. recall | BBI



co-occurrence
frequency
for comparison

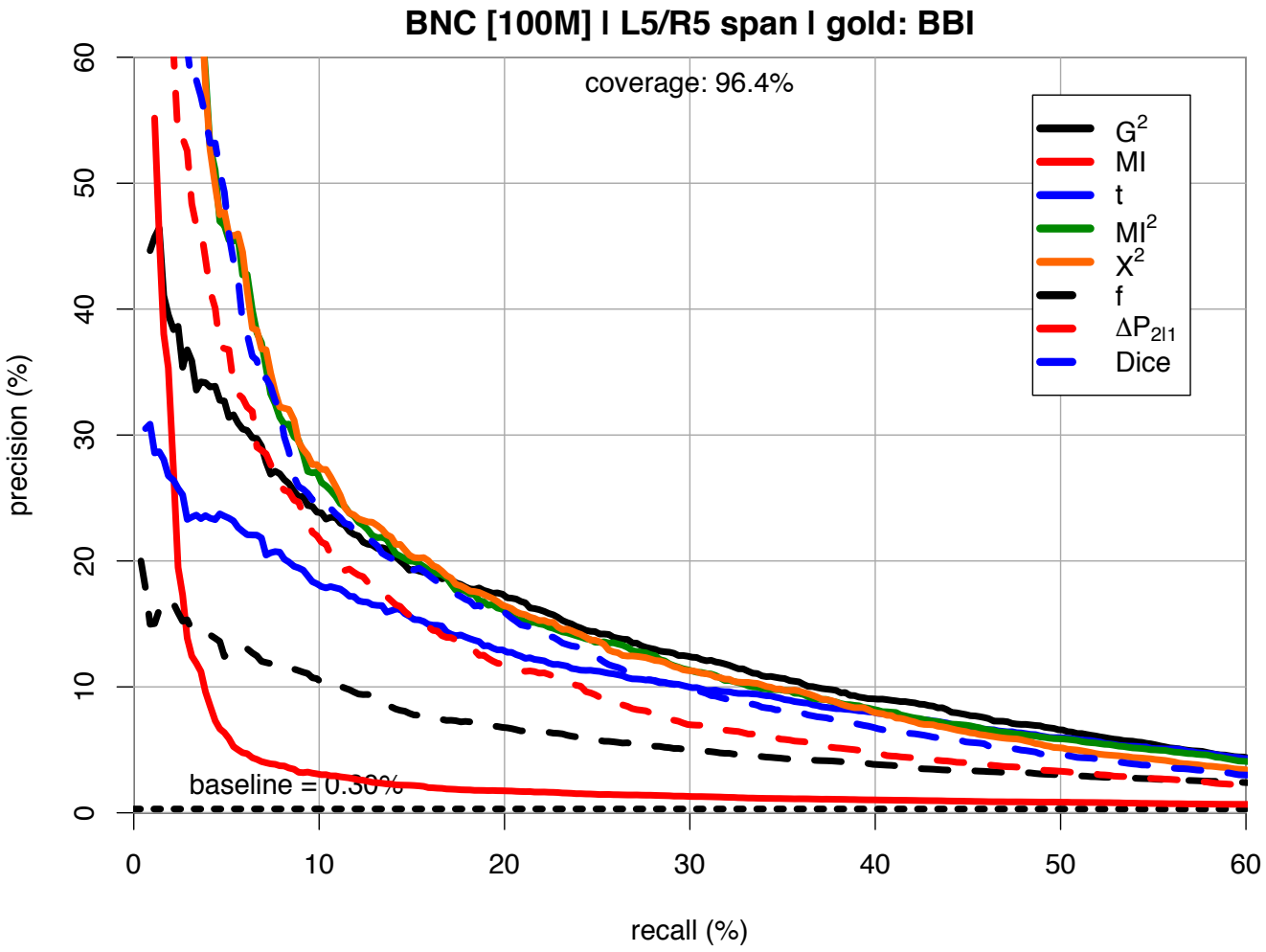
Evaluation: precision vs. recall | BBI



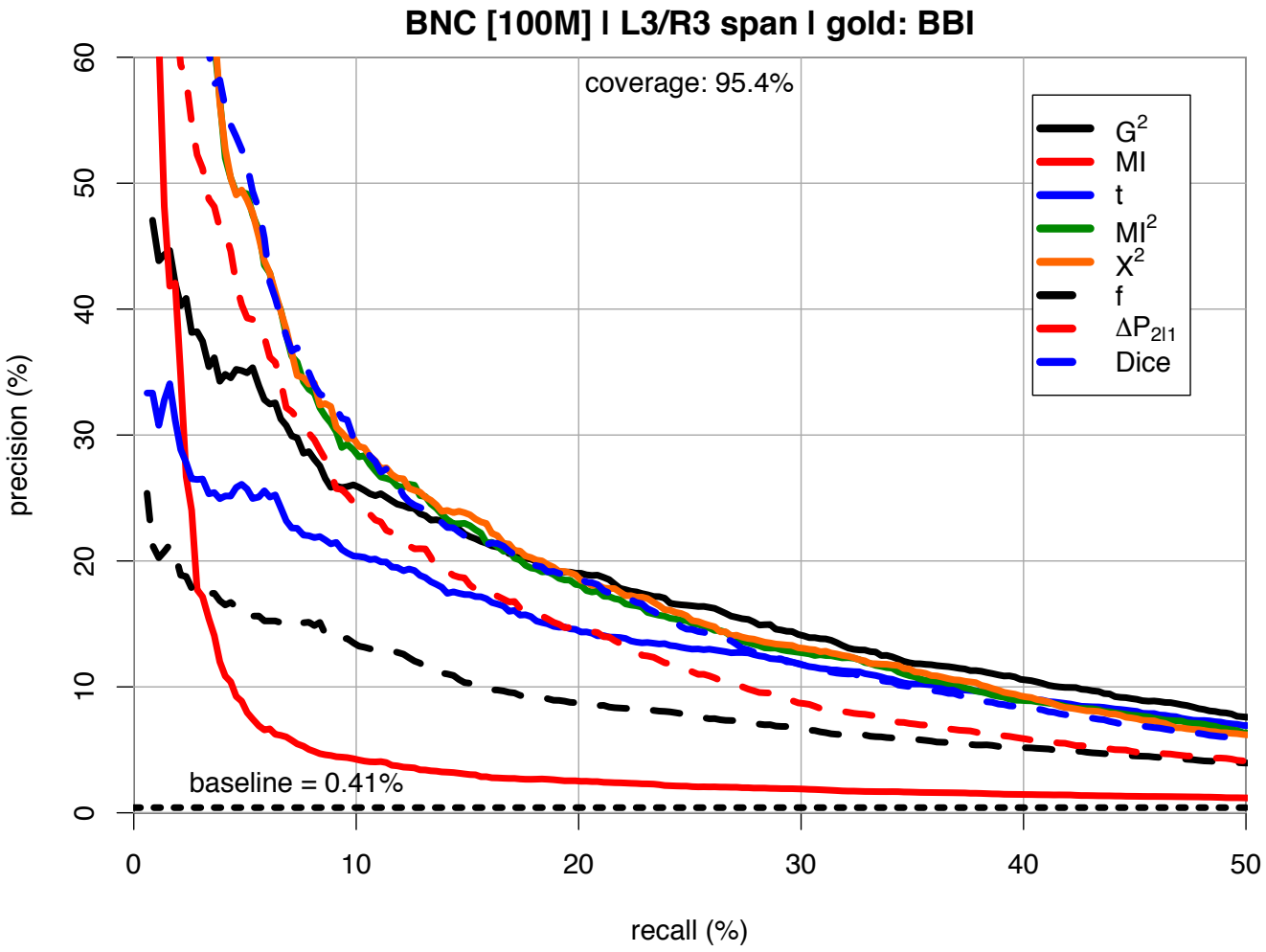
$$\Delta P_{2|1} = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$$

asymmetric ΔP
(Allan 1980; Gries 2013)

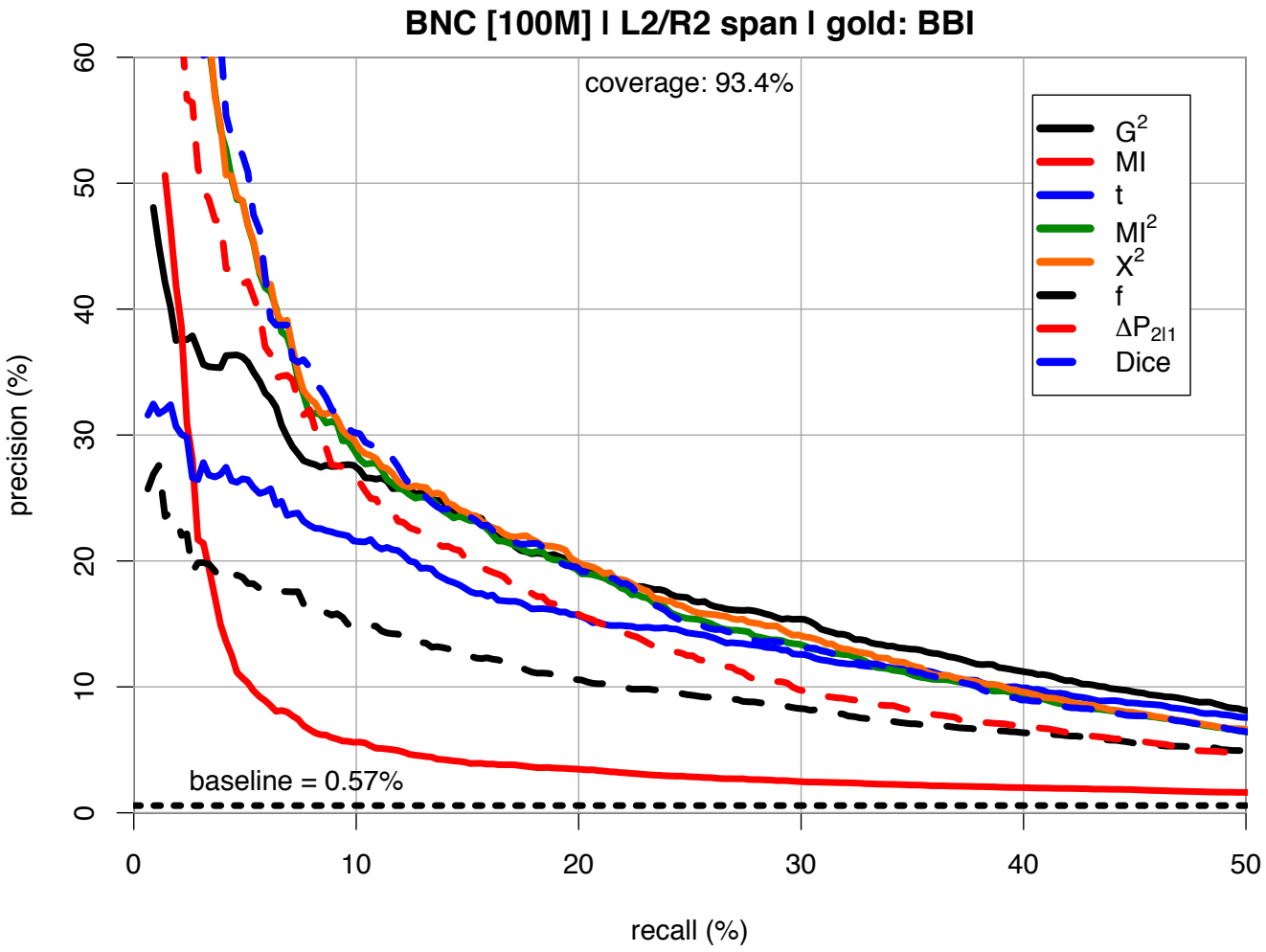
Evaluation: precision vs. recall | BBI



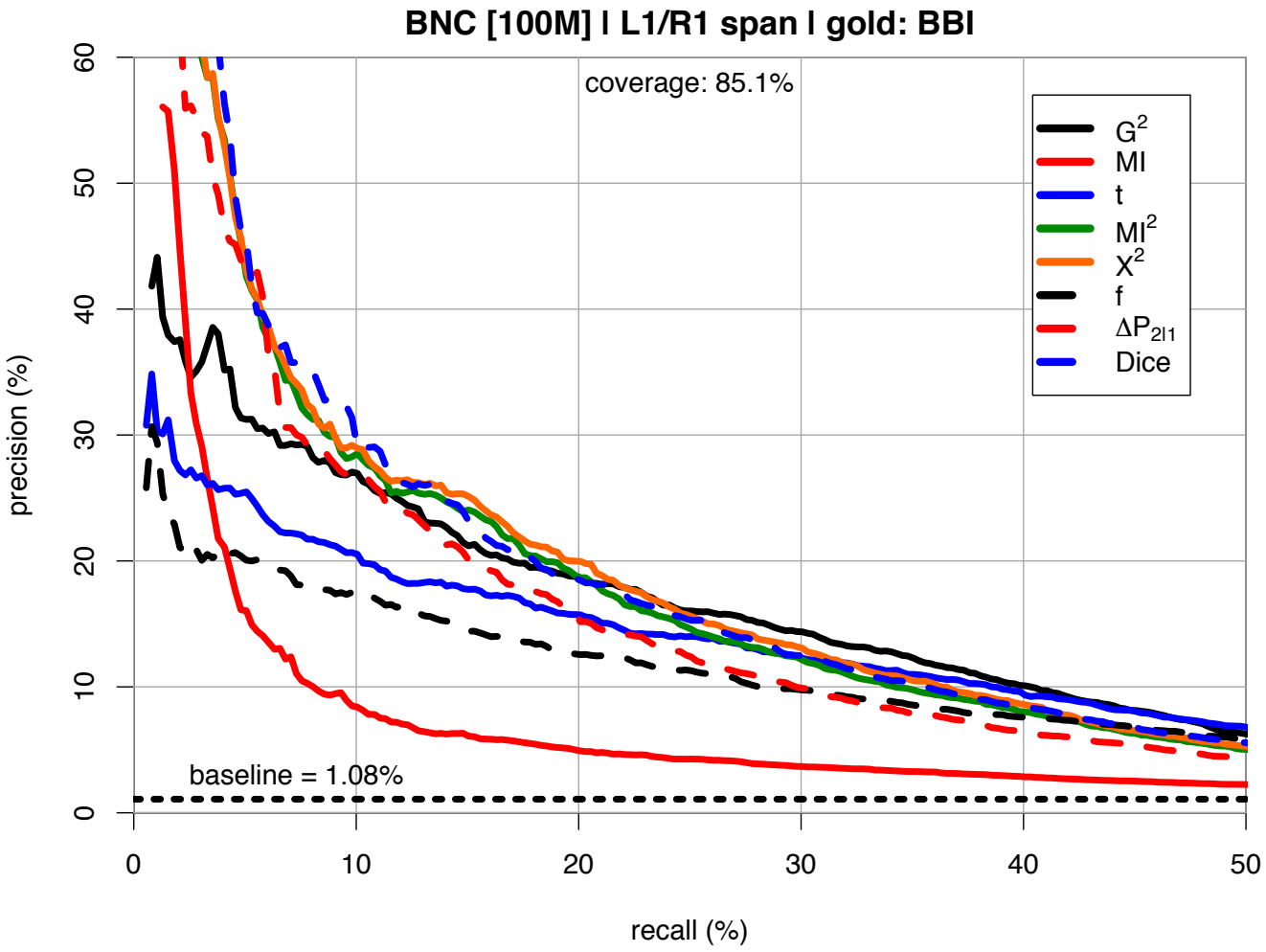
Factor: context size | BBI



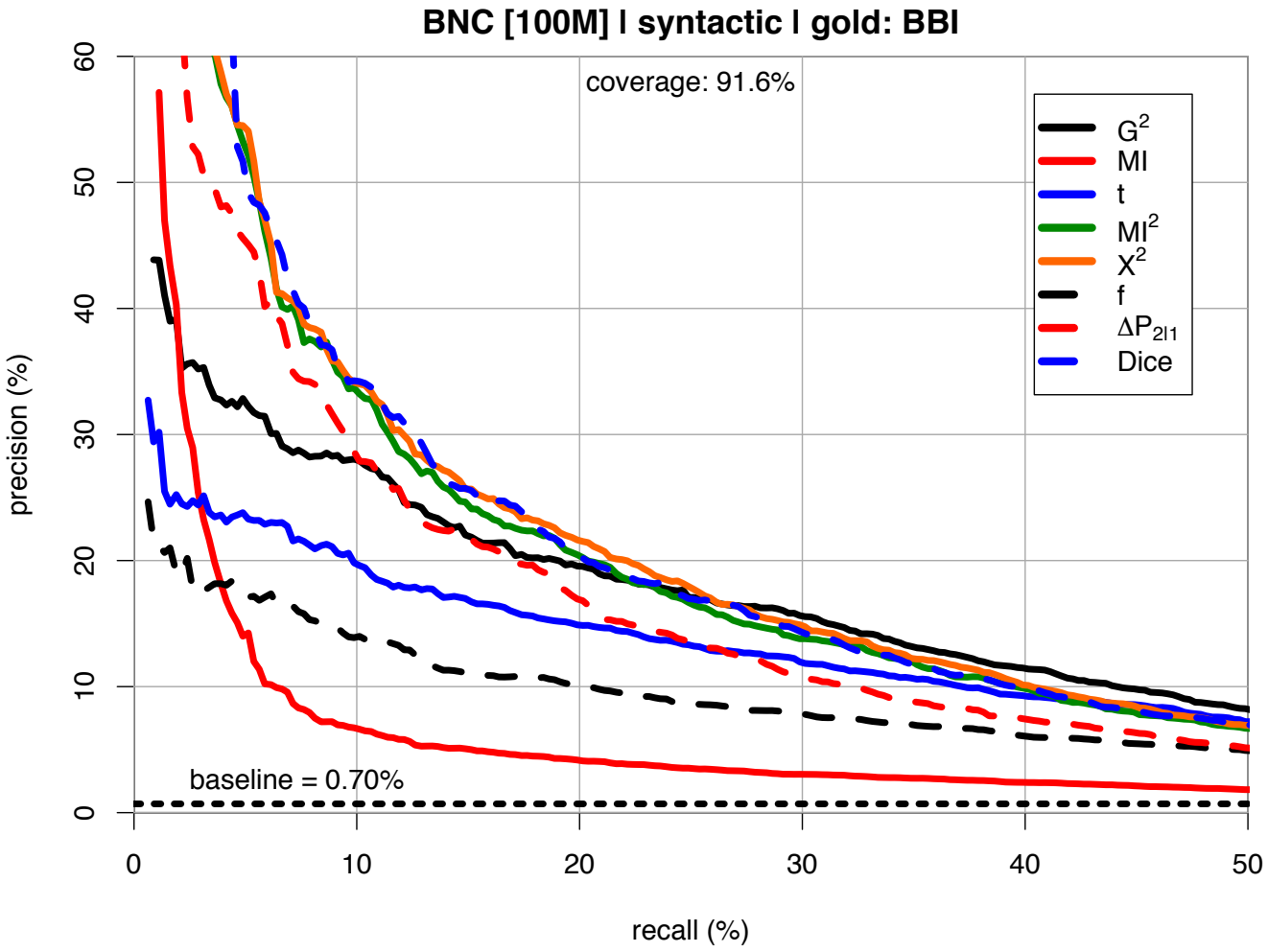
Factor: context size | BBI



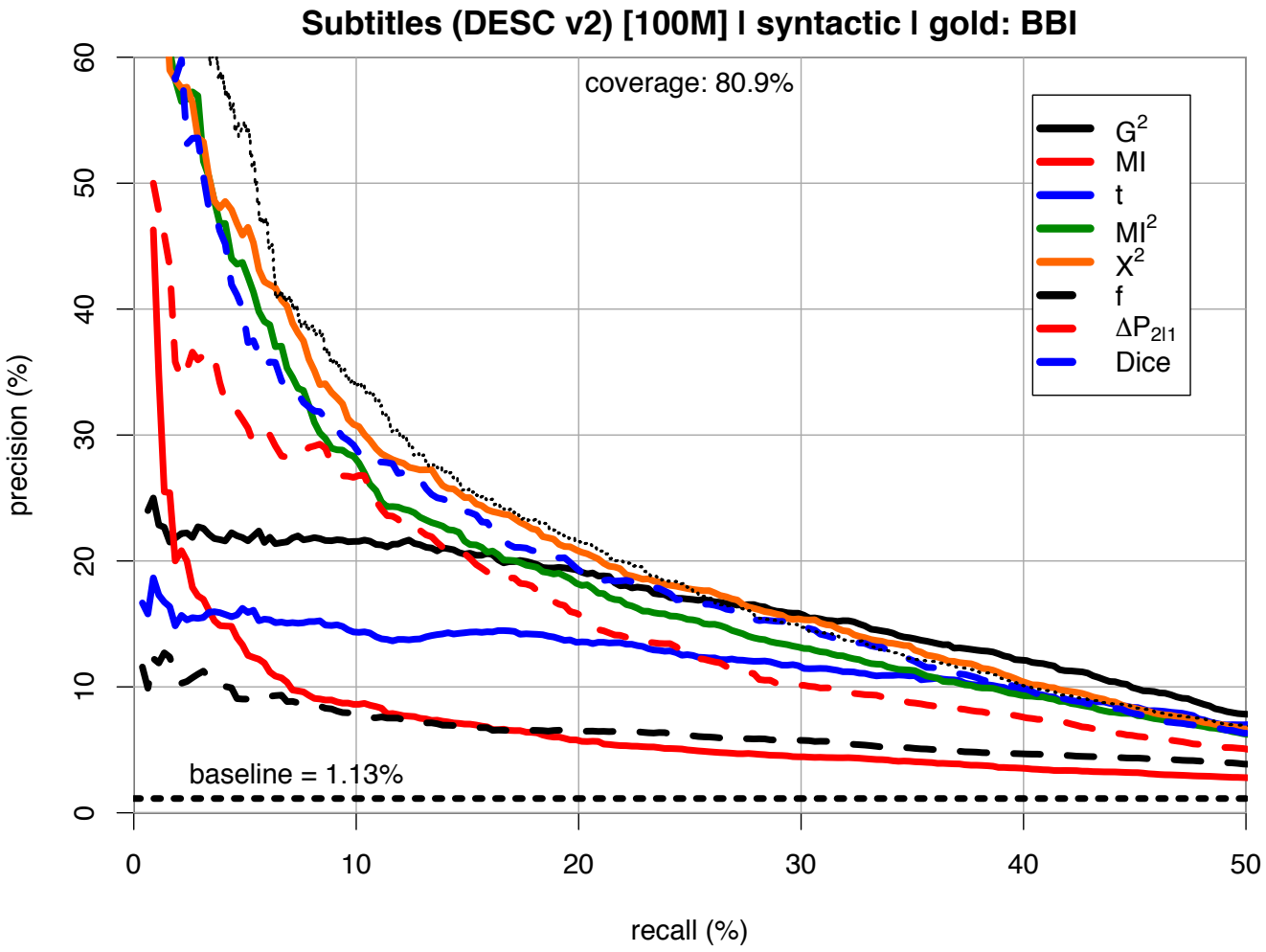
Factor: context size | BBI



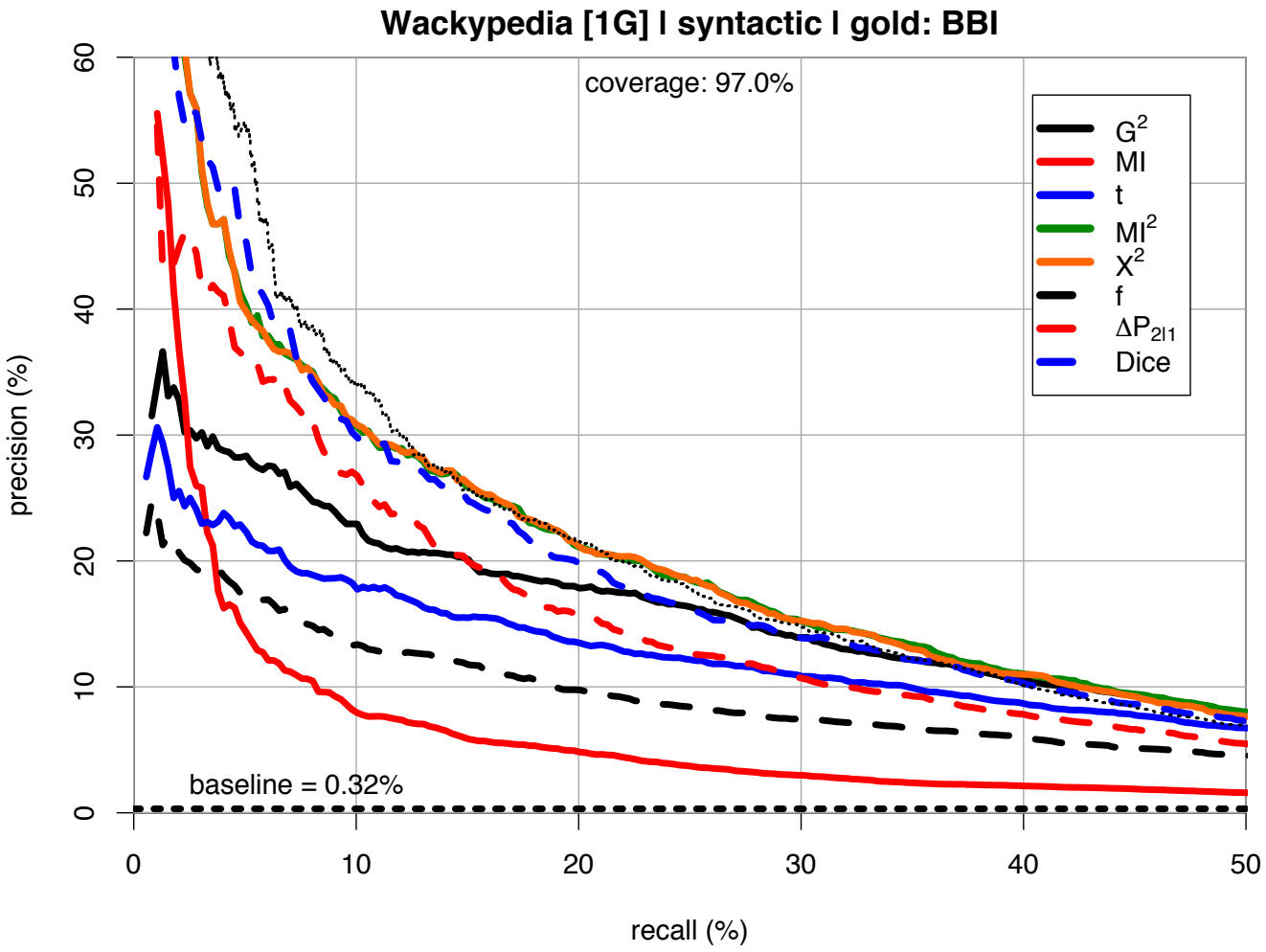
Factor: context size | BBI



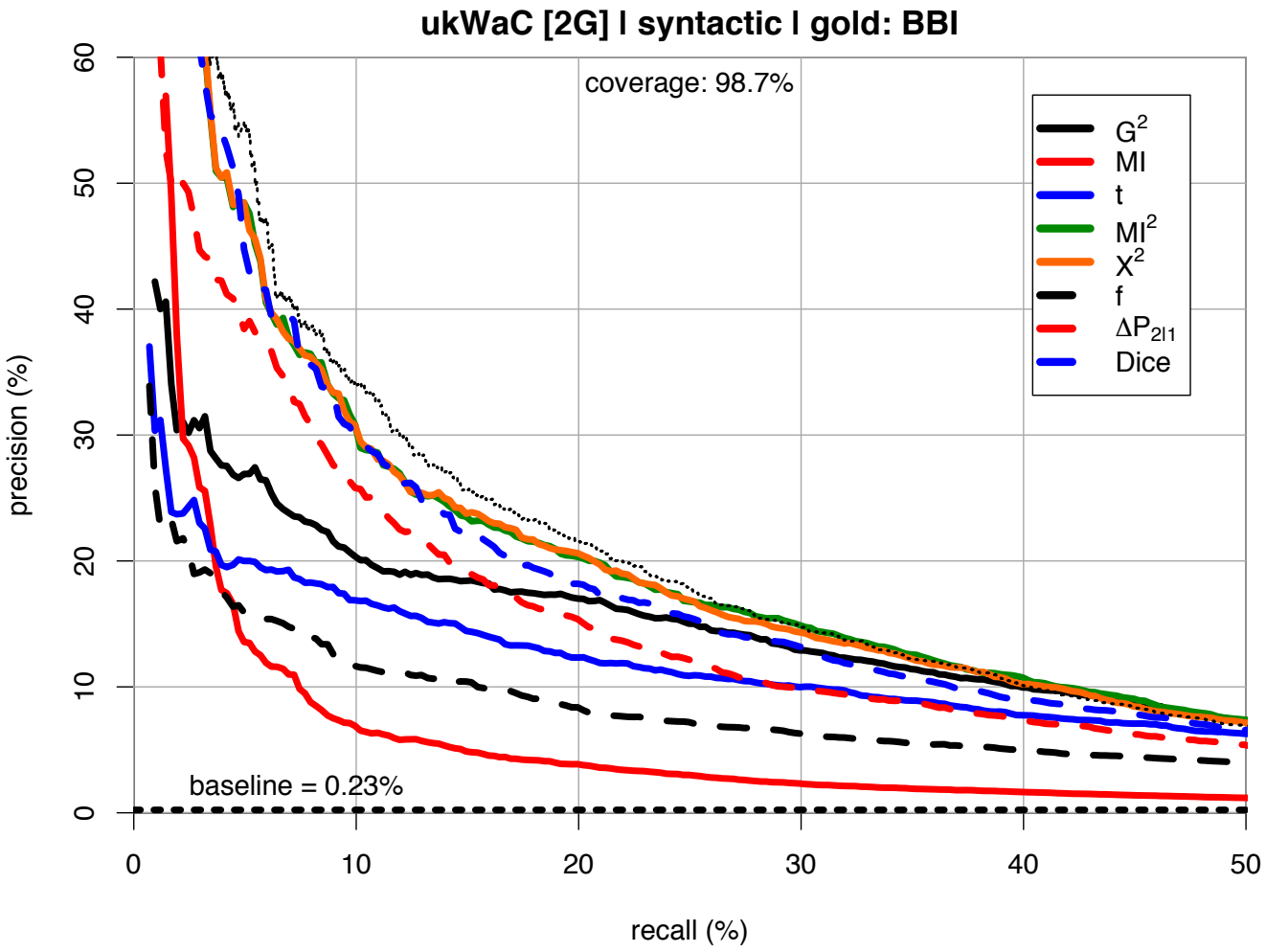
Factor: corpus | syntactic| BBI



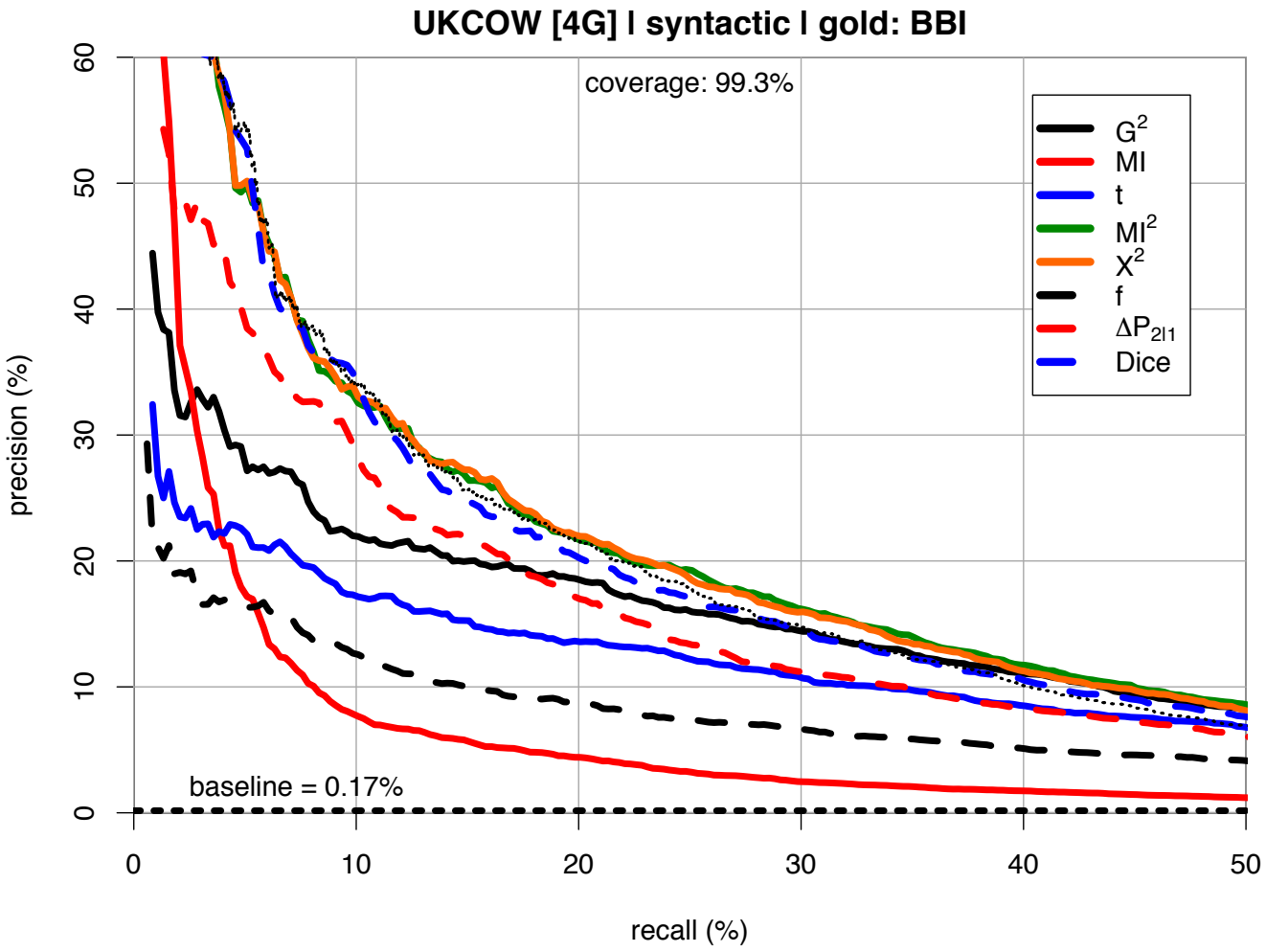
Factor: corpus | syntactic| BBI



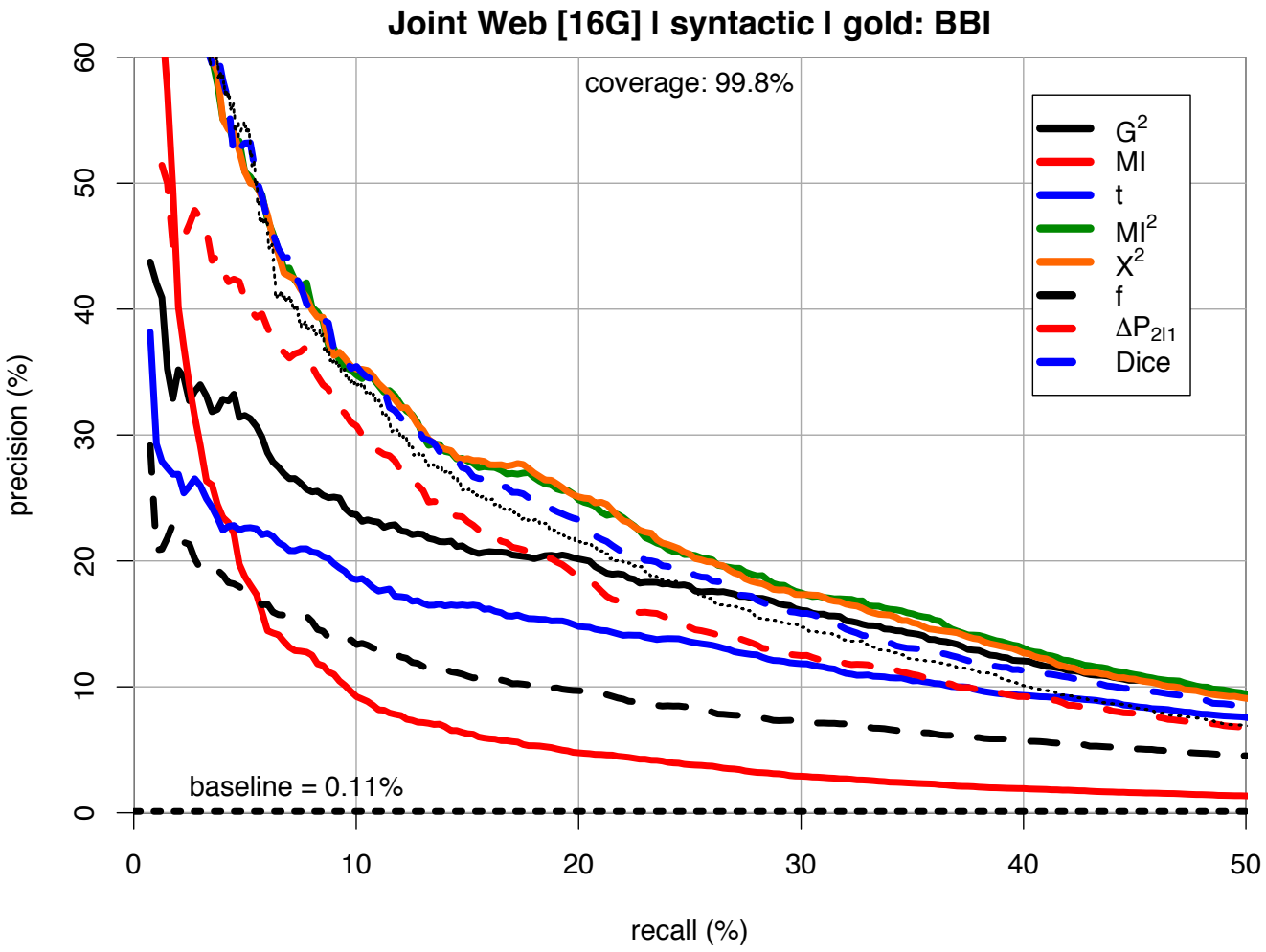
Factor: corpus | syntactic| BBI



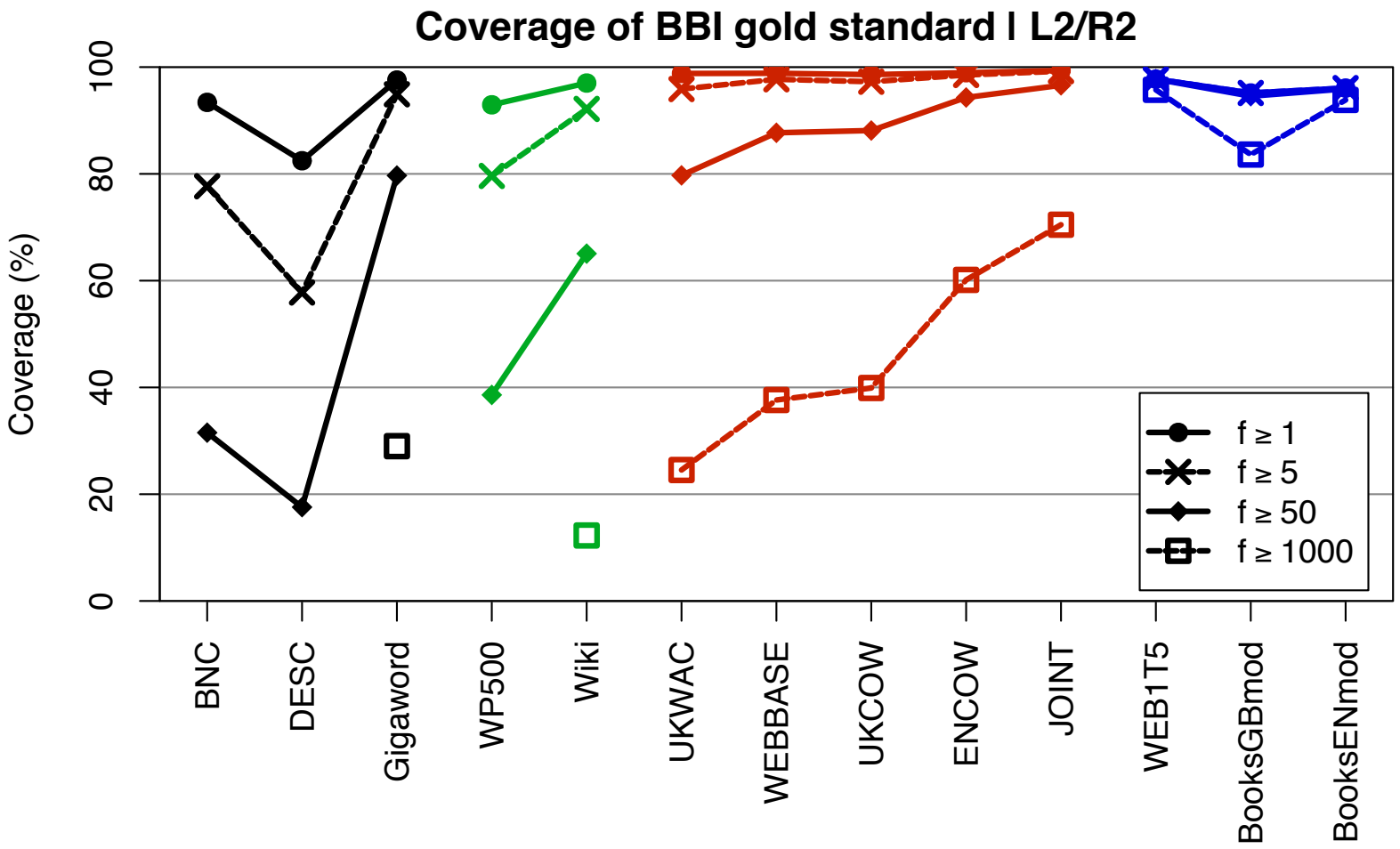
Factor: corpus | syntactic| BBI



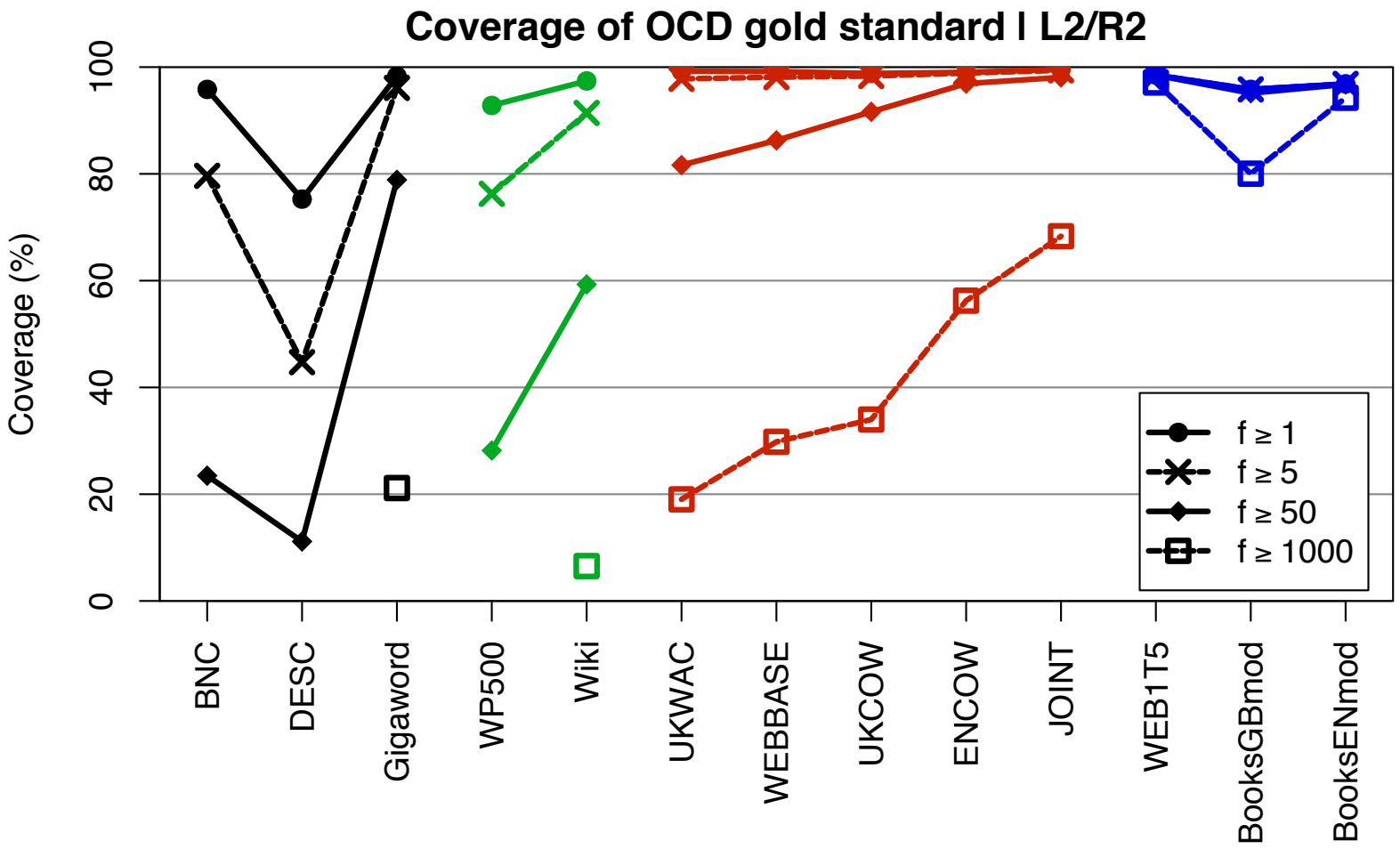
Factor: corpus | syntactic| BBI



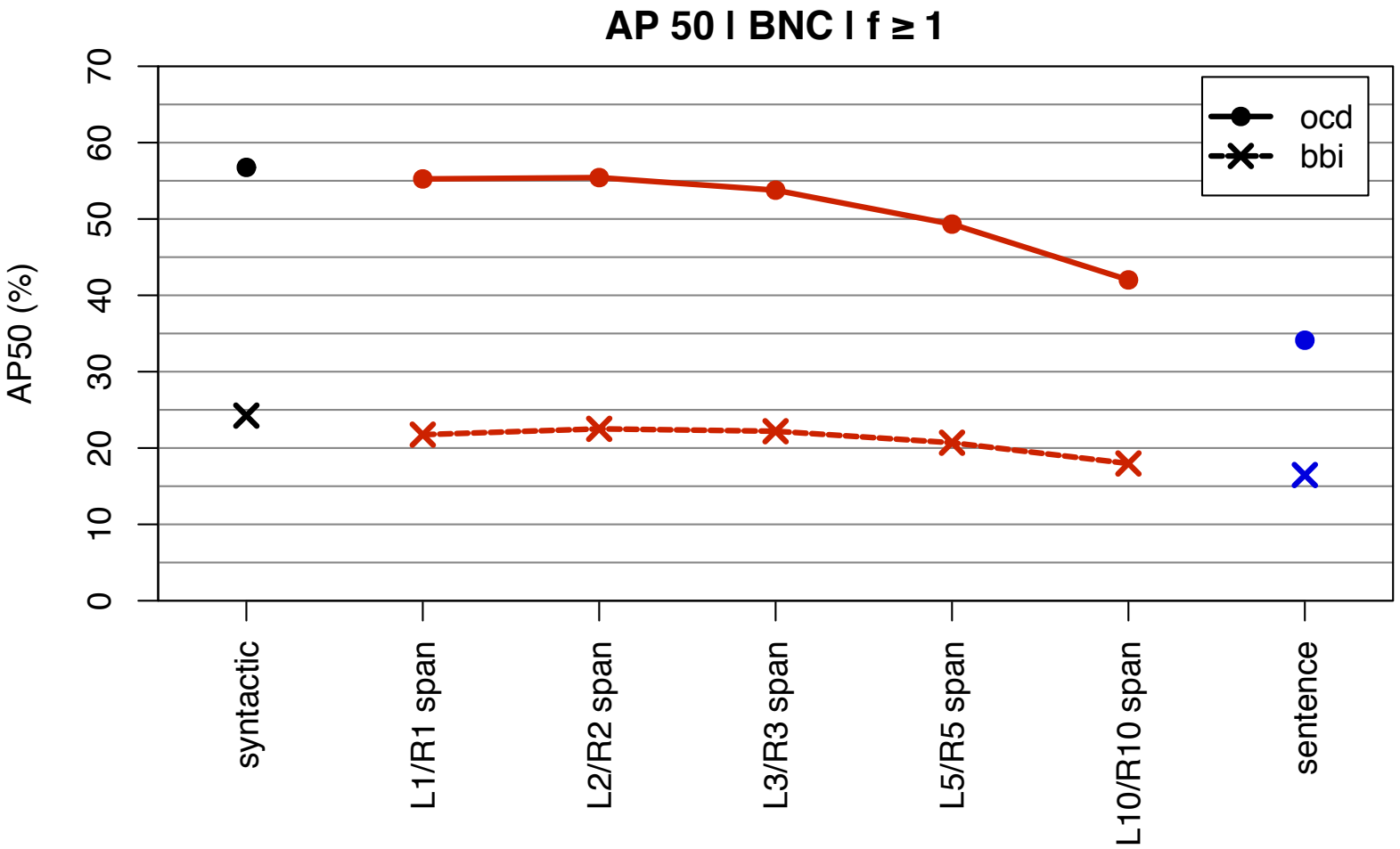
Results: coverage



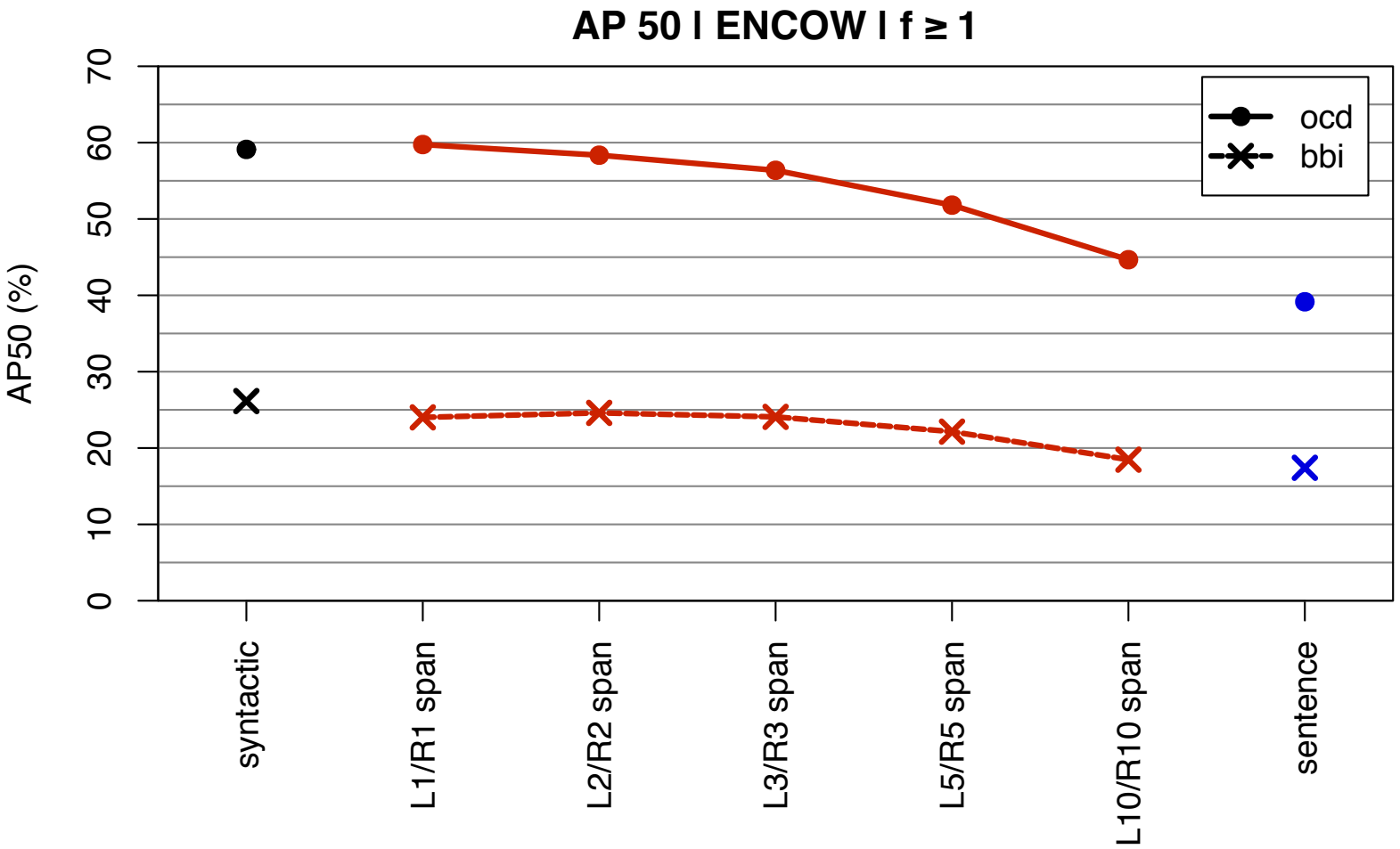
Results: coverage



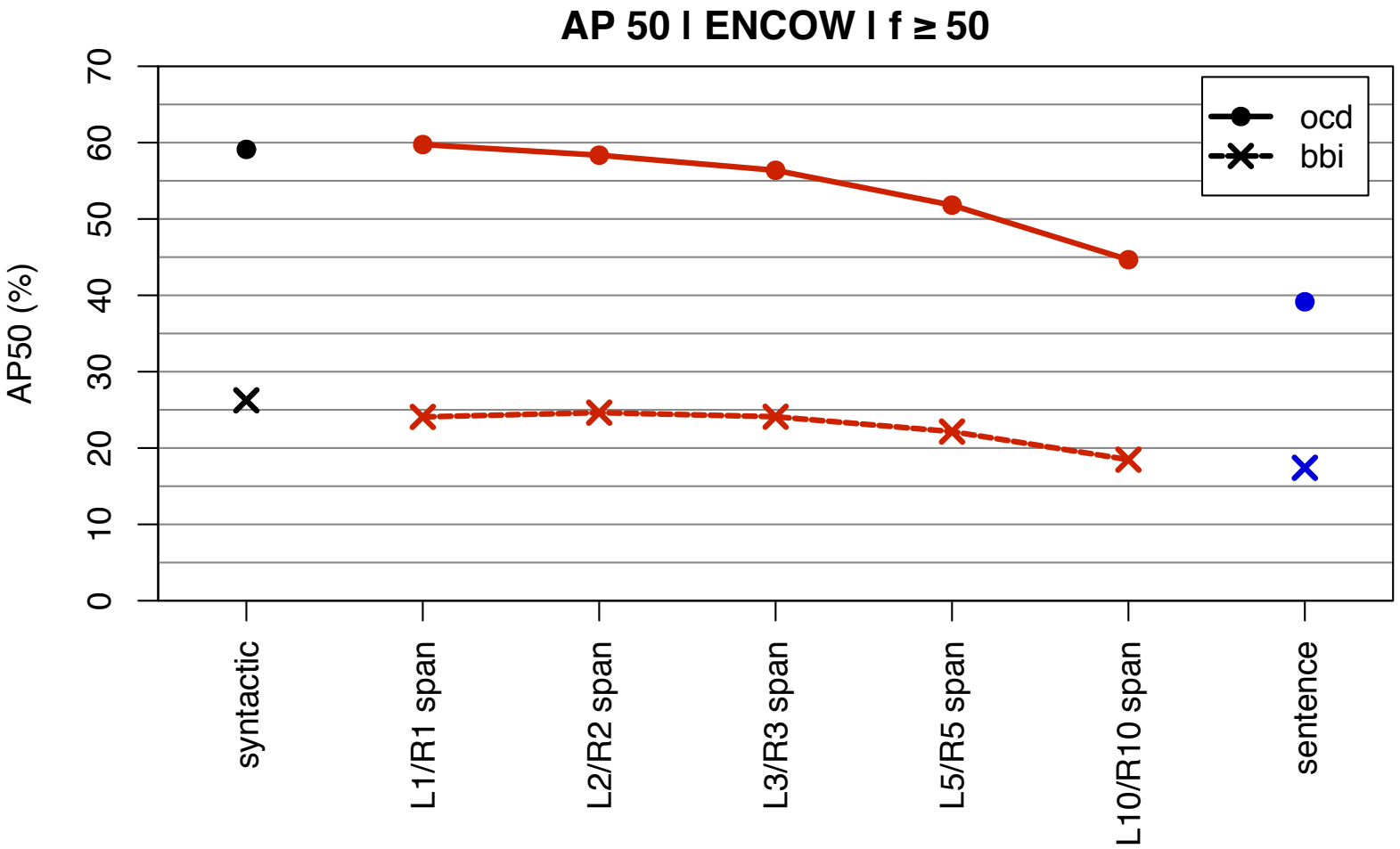
Results: context size



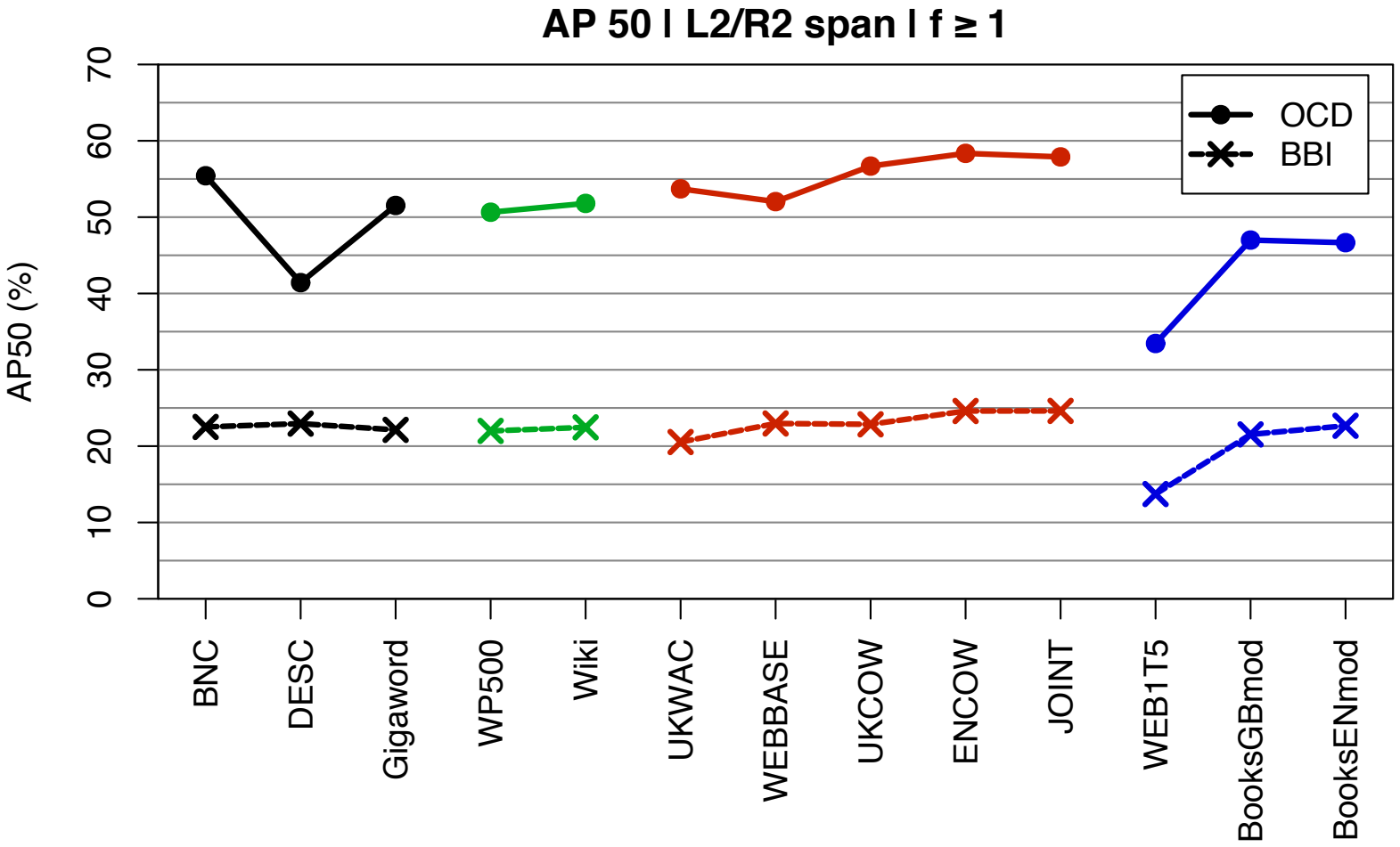
Results: context size



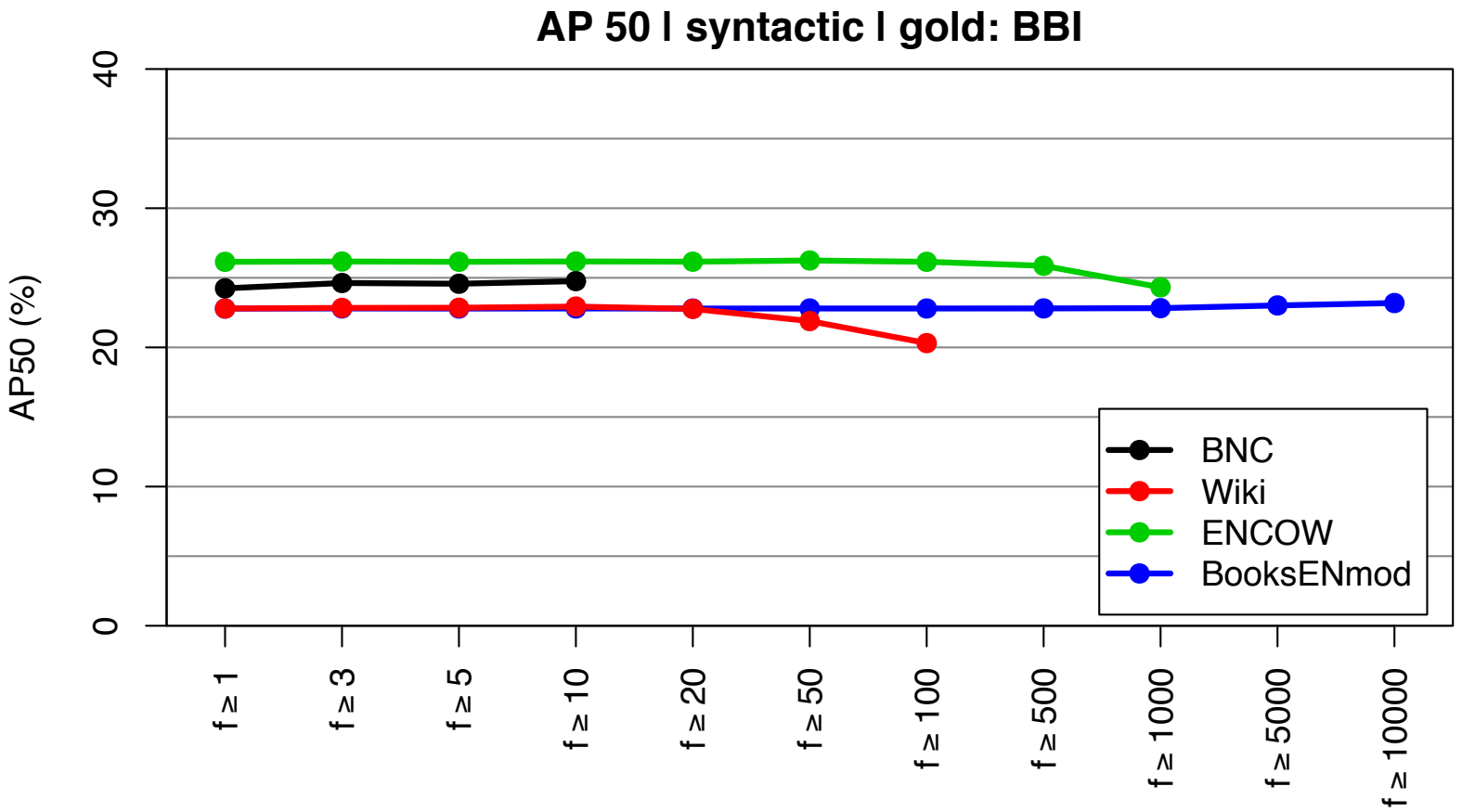
Results: context size



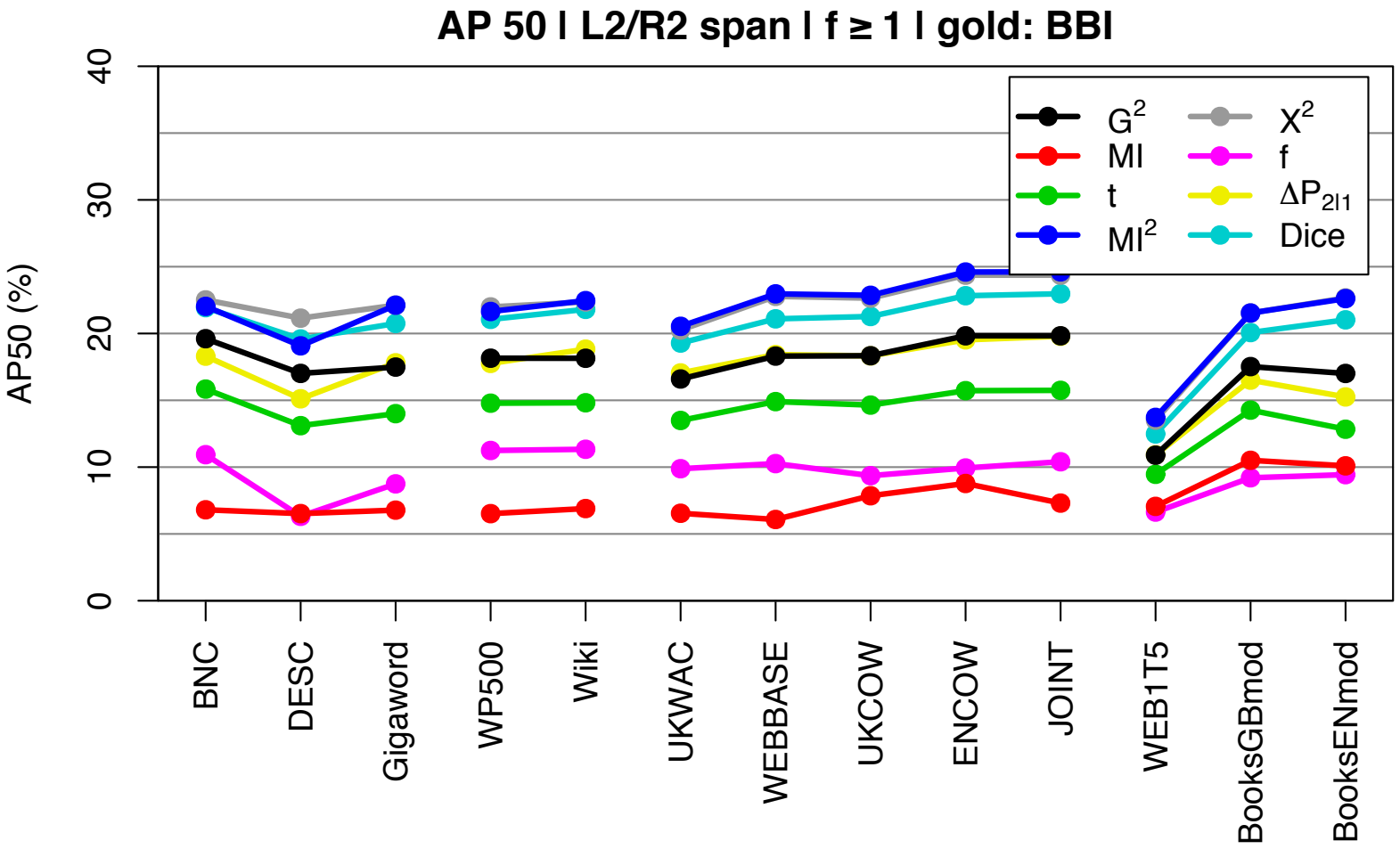
Results: corpus | L2/R2 span



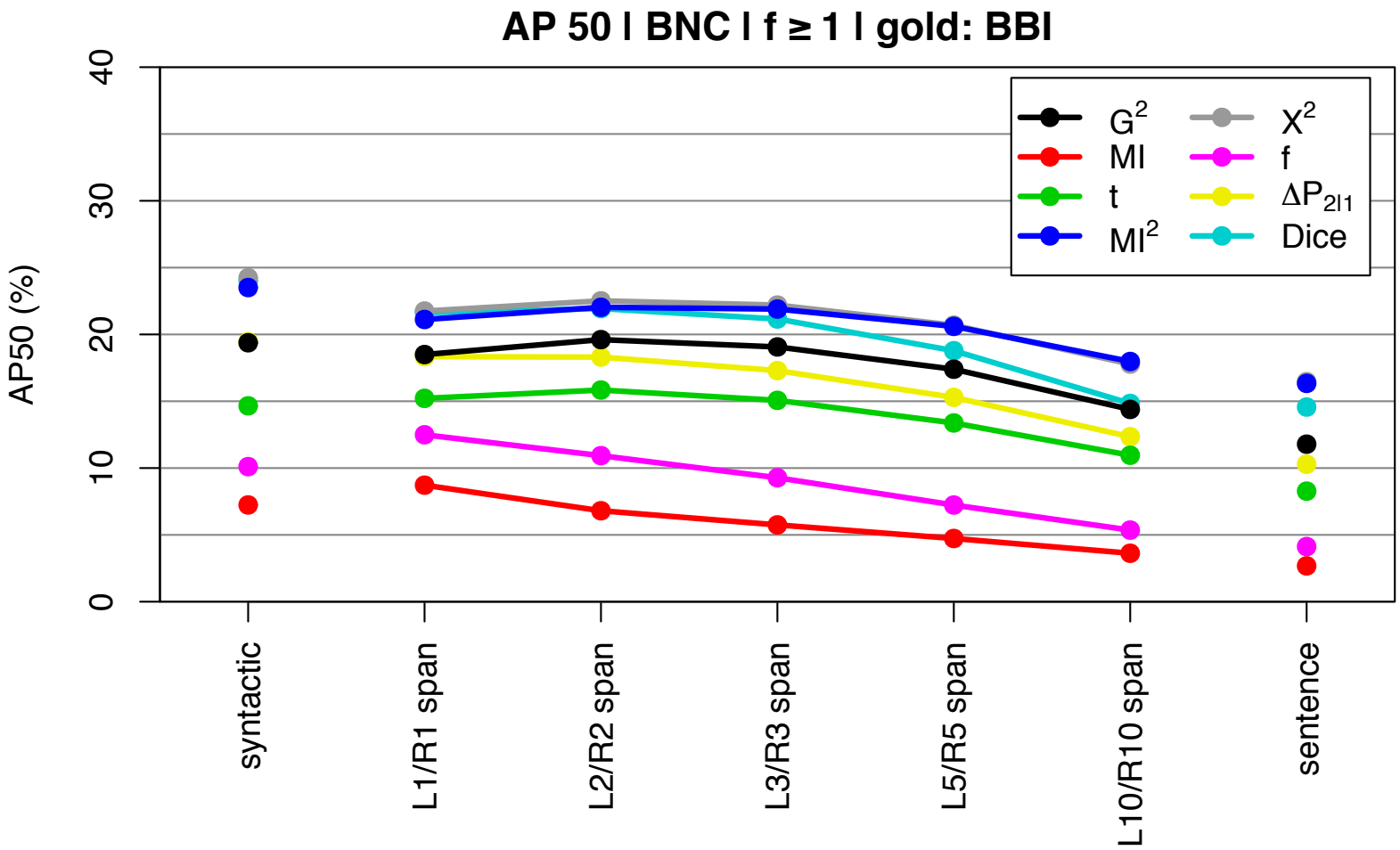
Results: frequency threshold



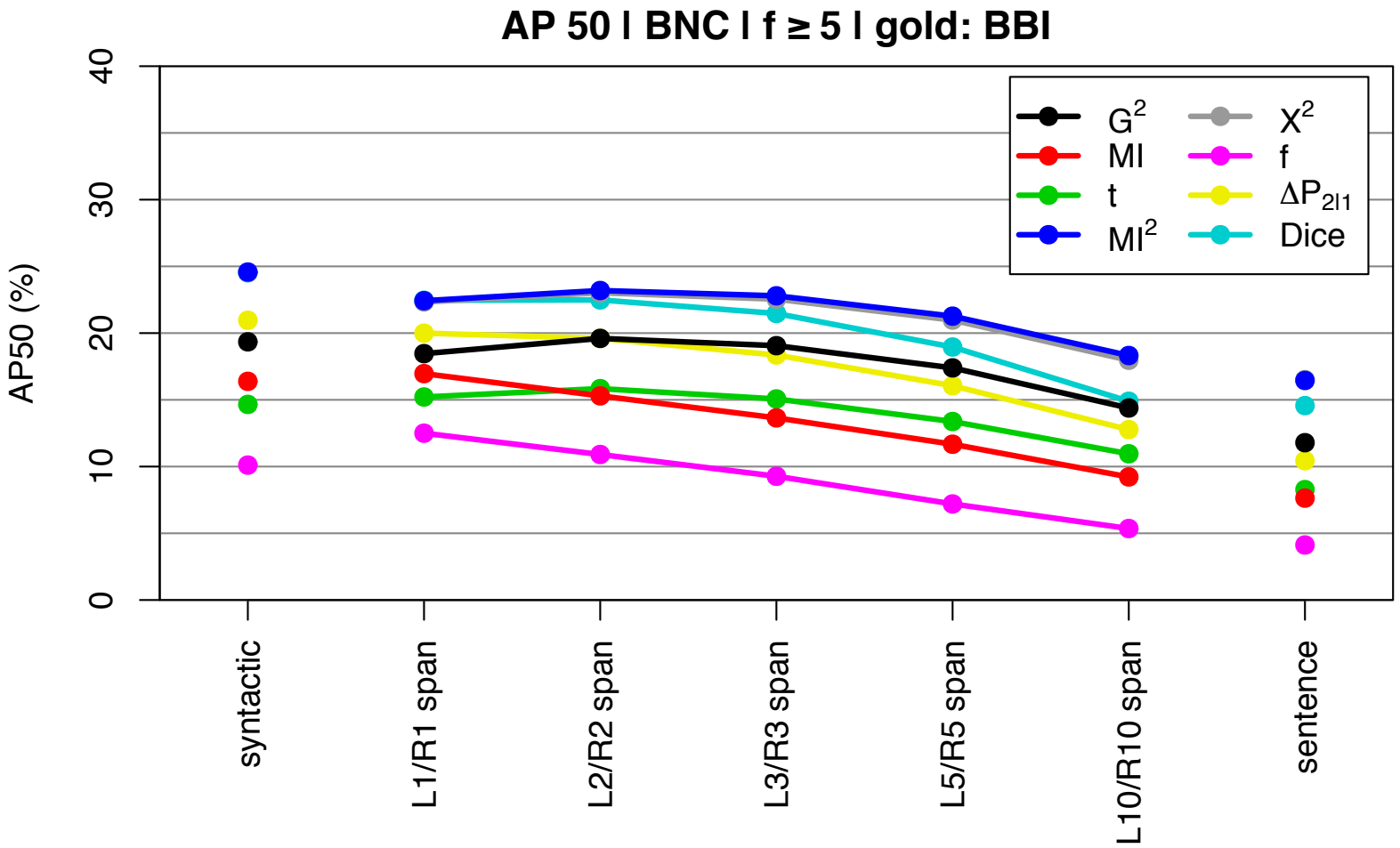
Results: interactions



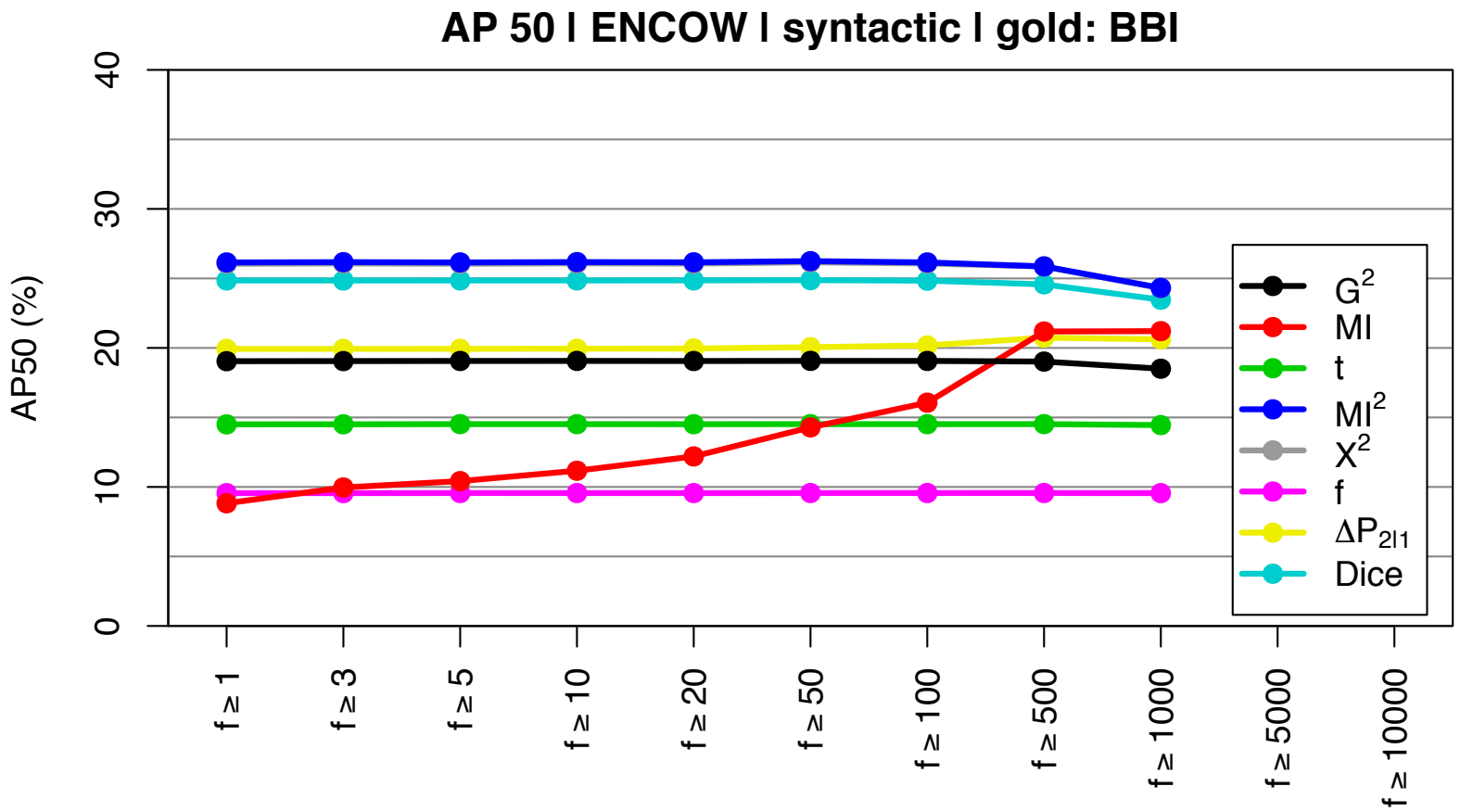
Results: interactions



Results: interactions



Results: interactions



E-VIEW-alation

- Interactive Web-based viewer for P/R plots
- Gives user full control over evaluation parameters
  - make your own animations like those in the presentation

http://www.collocations.de/eviewalation/

👉 to be released as open-source software

Conclusions

- Small co-occurrence contexts are better
- Size matters, but also corpus quality
  - very large Web corpora outperform BNC
- Frequency threshold does not improve results
  - possibly due to focus on small number of nodes
- Virtually no interactions between parameters
  - corpus vs. context size vs. AM
  - most findings hold across both gold standards (except AM)
- Share all results with E-VIEW-alation!