TECHNISCHE UNIVERSITÄT DARMSTADT

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

# E-VIEW-alation: A large-scale evaluation study of association measures for collocation identification

**Stefan Evert[1], Peter Uhrig[1], Sabine Bartsch[2], Thomas Proisl[1]**

[1]FAU Erlangen-Nürnberg [2]TU Darmstadt

# TL; DR

- Comprehensive evaluation study
  - lexicographic gold standard (collocations dictionaries)
  - focus: corpus, co-occcurence context, freq. threshold
- Novel approach to sharing results
  - interactive Web-based viewer
  - gives access to *complete* evaluation results

# Lexicalised MWE in lexicography

- idioms (*kick the bucket*, *play cat and mouse*)
- figurative expressions (*spill the beans*)
- multiword units (*in front of*, *déjà vu*, *to and fro*)
- particle verbs (*give up*, *hand in*)
- lexical collocations (*brush teeth*, *heavy smoker*)
- light verb constructions (*give talk*, *draw conclusion*)
- compounds (*skeleton key*, *penalty kick*)
- named entities (*New York City*, *Red Cross*)
- clichés (*thigh-high boots*, *bucket and spade*)

# Lexicalised MWE in lexicography

- idioms (*kick the bucket*, *play cat and mouse*)
- figurative expressions (*spill the beans*)
- multiword units (*in front of*, *déjà vu*, *to and fro*)
- particle verbs (*give up*, *hand in*)
- lexical collocations (*brush teeth*, *heavy smoker*)
- light verb constructions (*give talk*, *draw conclusion*)
- compounds (*skeleton key*, *penalty kick*)
- named entities (*New York City*, *Red Cross*)
- clichés (*thigh-high boots*, *bucket and spade*)

# MWE identification

- Main cue: co-occurrence frequency, quantified by statistical association measures (AM, Sinclair 1966)
- Other criteria derived from properties of MWE (Manning & Schütze 1999, 184)
  - non-compositionality (→ distributional semantics)
  - non-modifiability (→ syntactic flexibility, fixed ordering)
  - non-substitutability (→ substitution tests)
- Recent work focuses on feature combination by machine learning and on specific subtypes of MWE
  - AMs still play central role, esp. for collocation identification

# Evaluation

- Evaluation studies usually test a single new algorithm of focus on a specific subtype of MWE
- Our goal is a 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 ($\log \theta$, $\log \theta_{disc}$), relative risk ($\log r$)
- log-likelihood ($G^2$, Dunning 1993)
- $\Delta P$ (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

- 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 (2013), 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 | ++ |

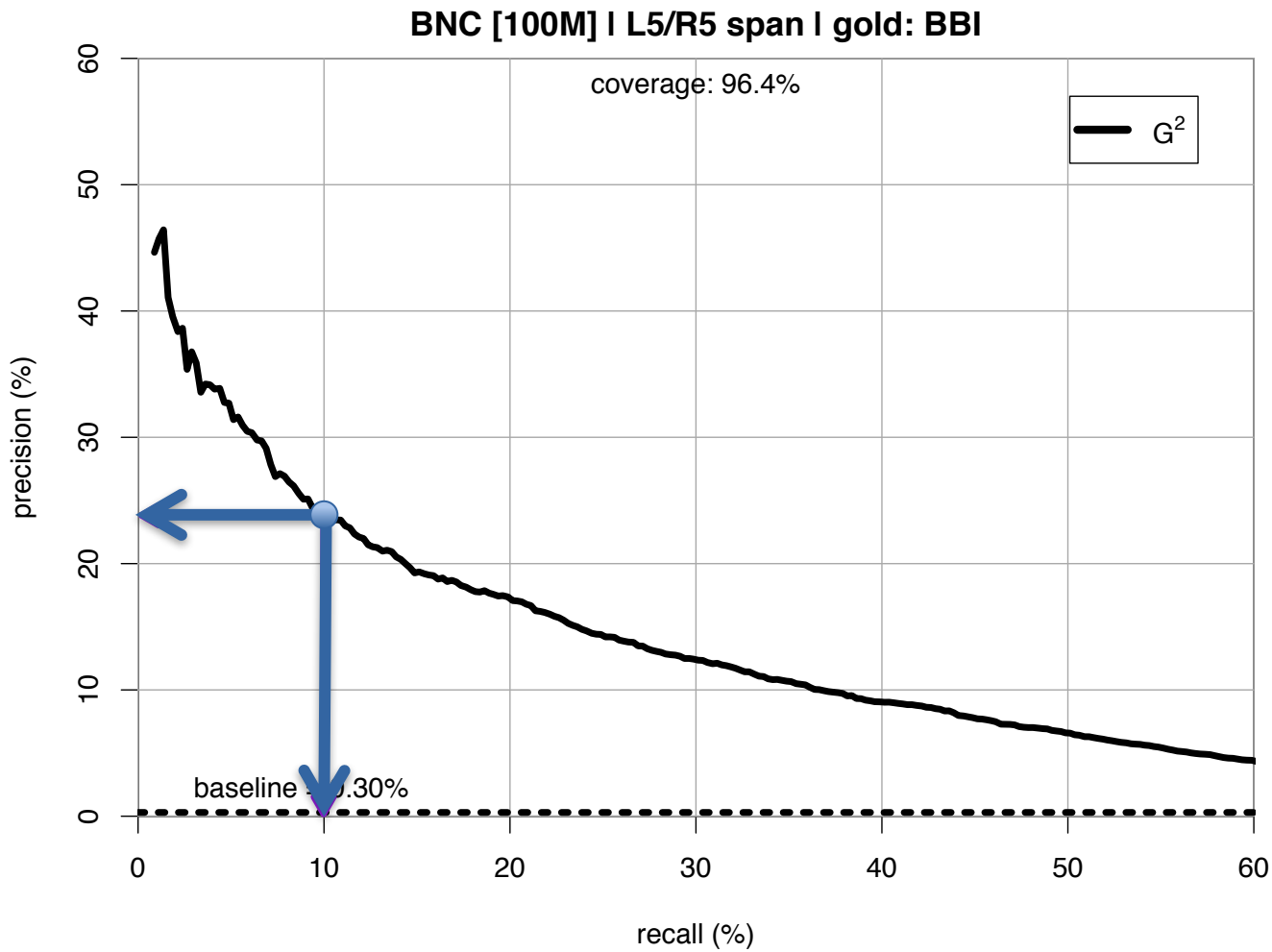
most strongly associated node-collocate pairs (n = 10)

*P* = 5 TP / 10 cand. = 50%

*R* = 5 / 2845 TP = 0.2%

TP = true positive (according to BBI dictionary)

# Evaluation: precision vs. recall | BBI

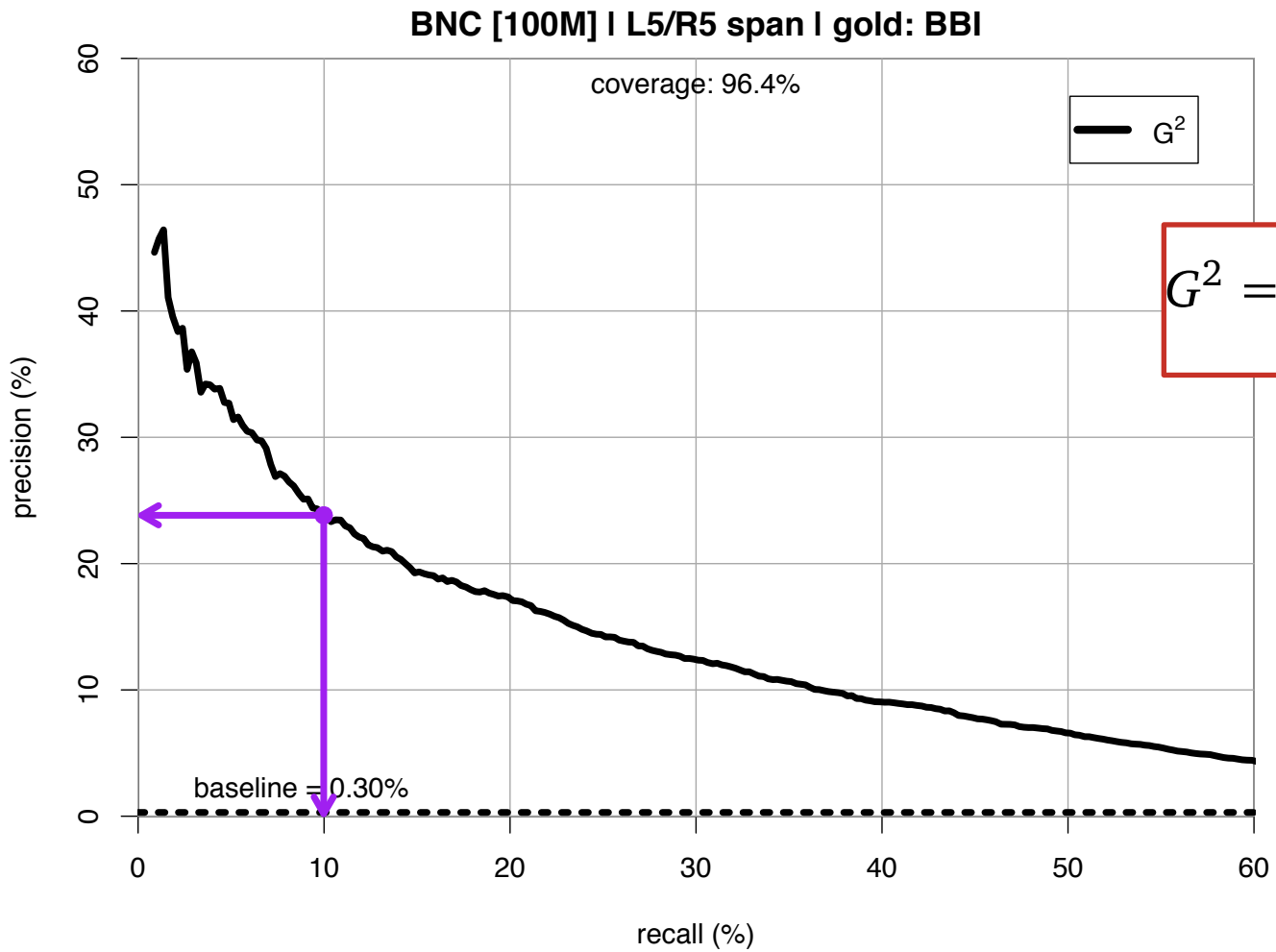



n = 1193

*R* = 10.0%

*P* = 23.9%

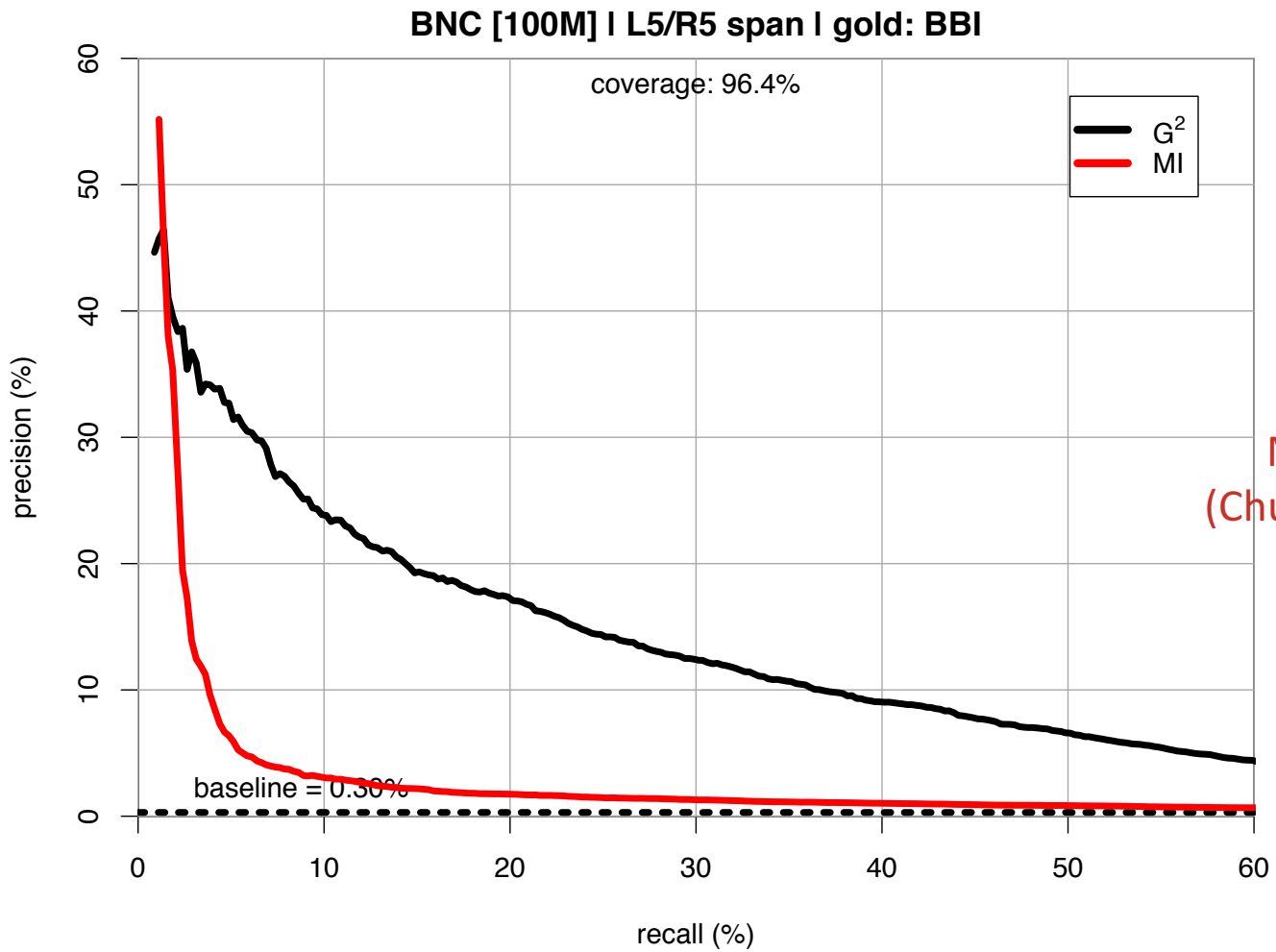
# Evaluation: precision vs. recall | BBI




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

log-likelihood
(Dunning 1993)

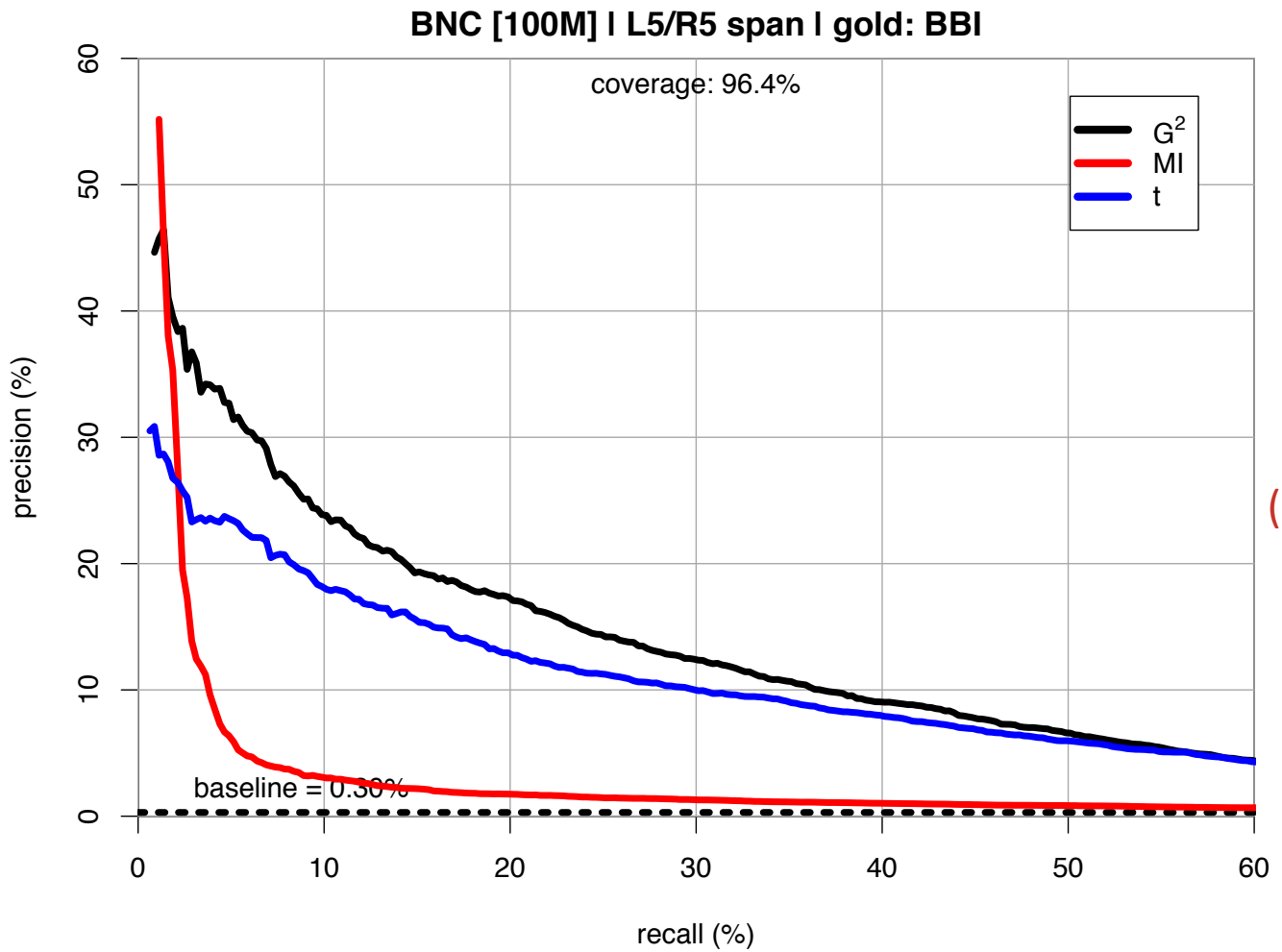
# Evaluation: precision vs. recall | BBI




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

Mutual Information
(Church & Hanks 1990)

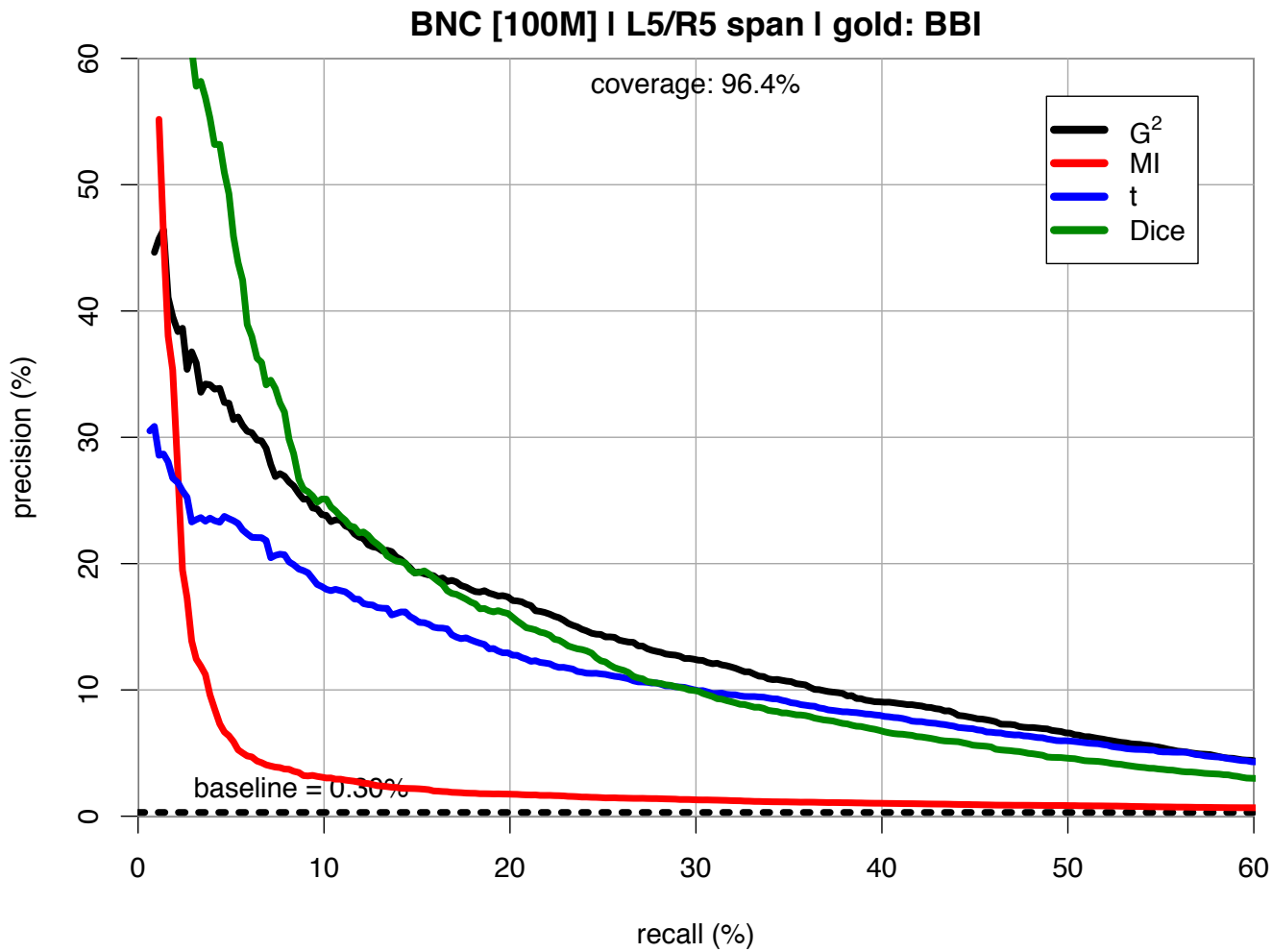
# Evaluation: precision vs. recall | BBI




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

t-score
(Church et al. 1991)

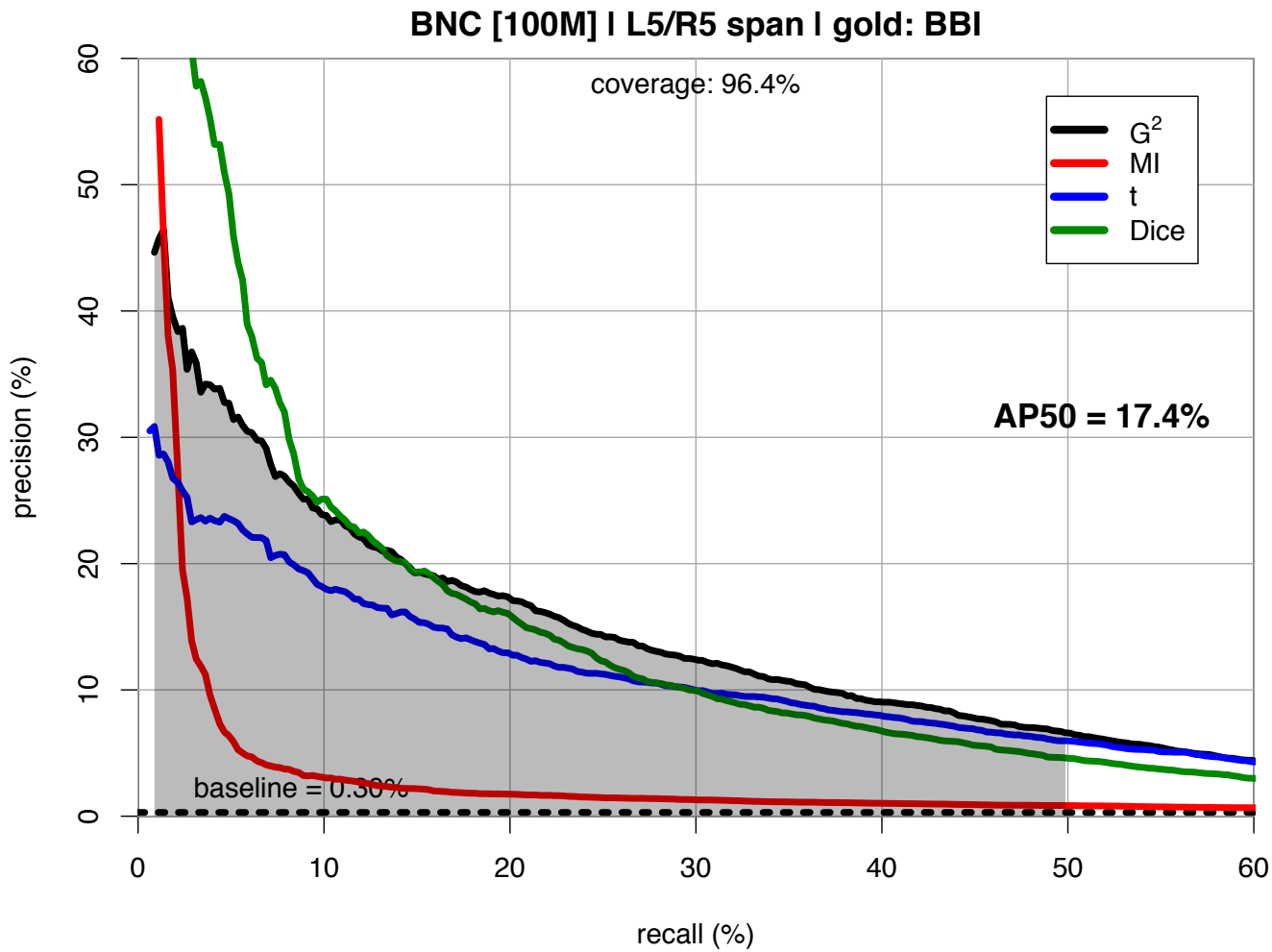
# Evaluation: precision vs. recall | BBI




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

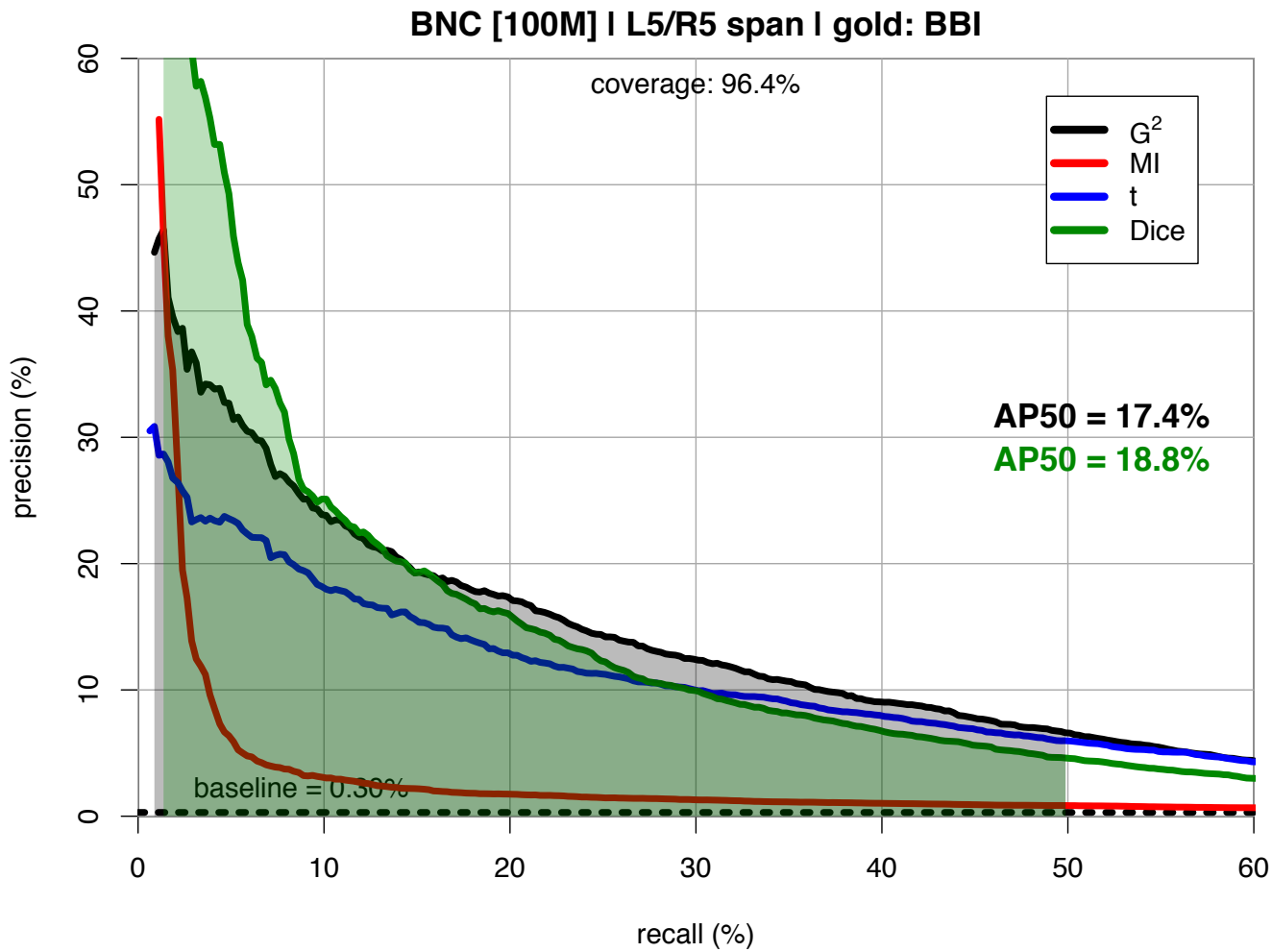
Dice coefficient
(Sketch Engine)

# Evaluation: precision vs. recall | BBI



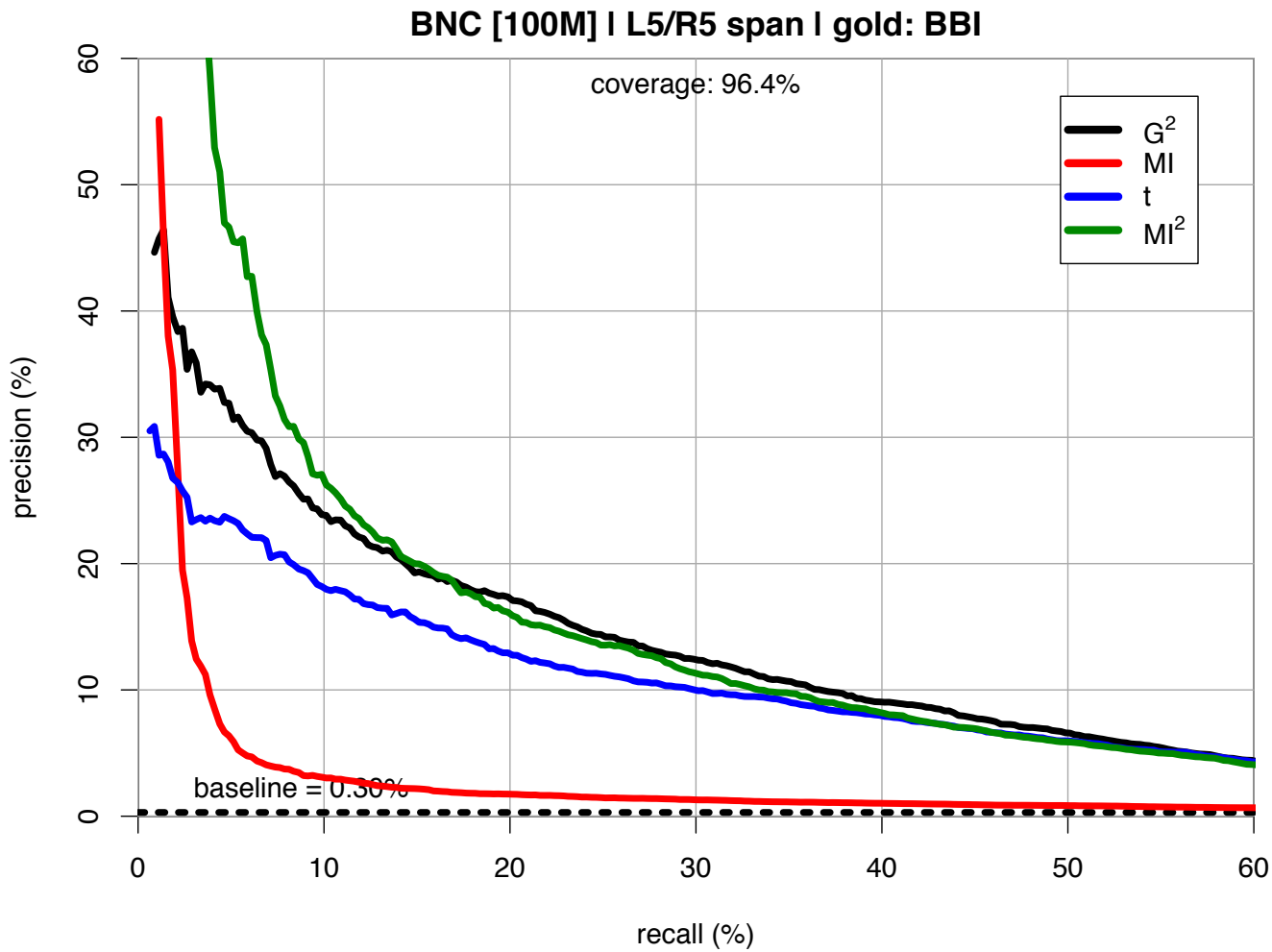

Average Precision up to 50% recall

# Evaluation: precision vs. recall | BBI




Average Precision up to 50% recall

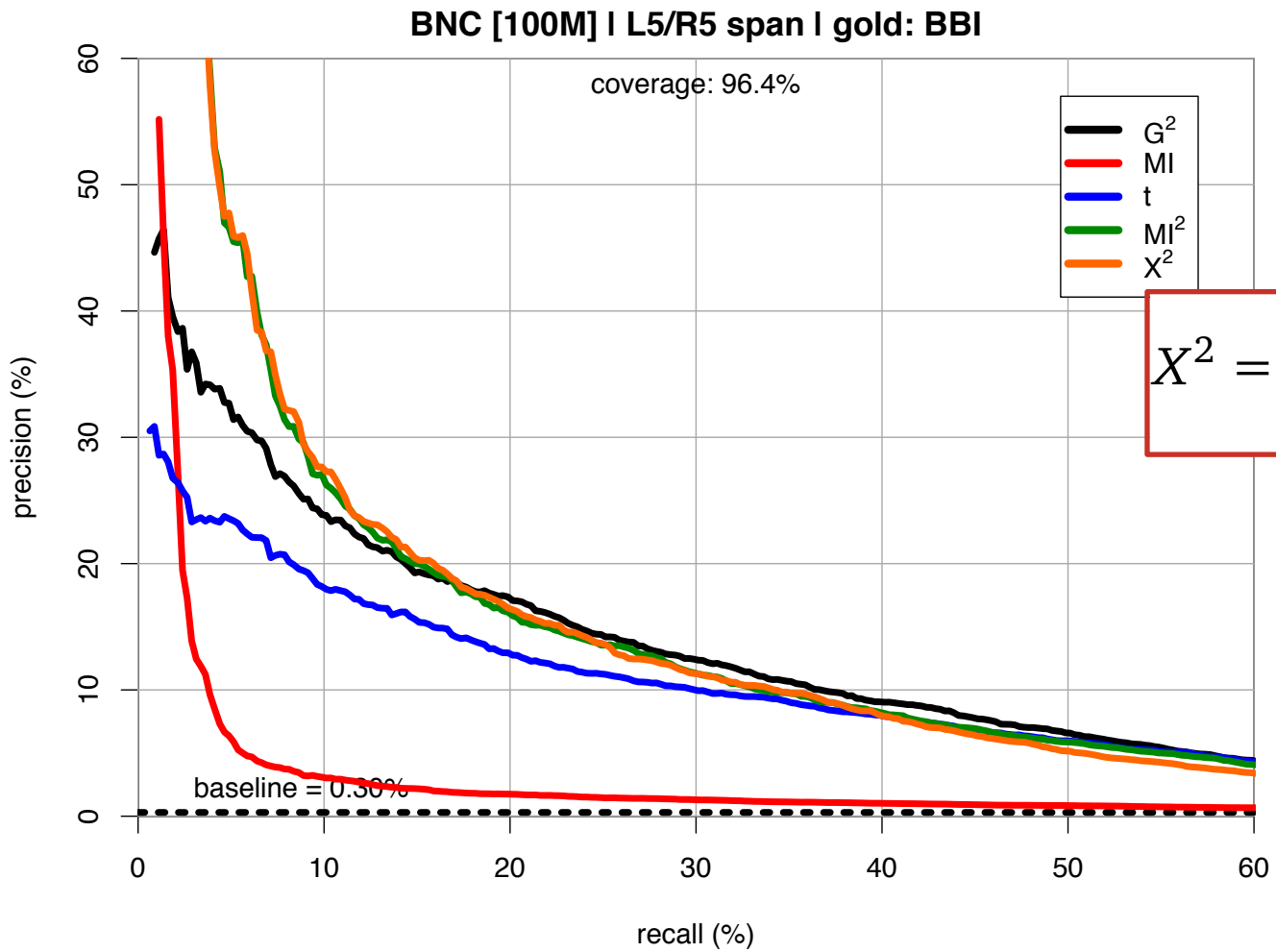
# Evaluation: precision vs. recall | BBI




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

heuristic measure
(Daille 1994)

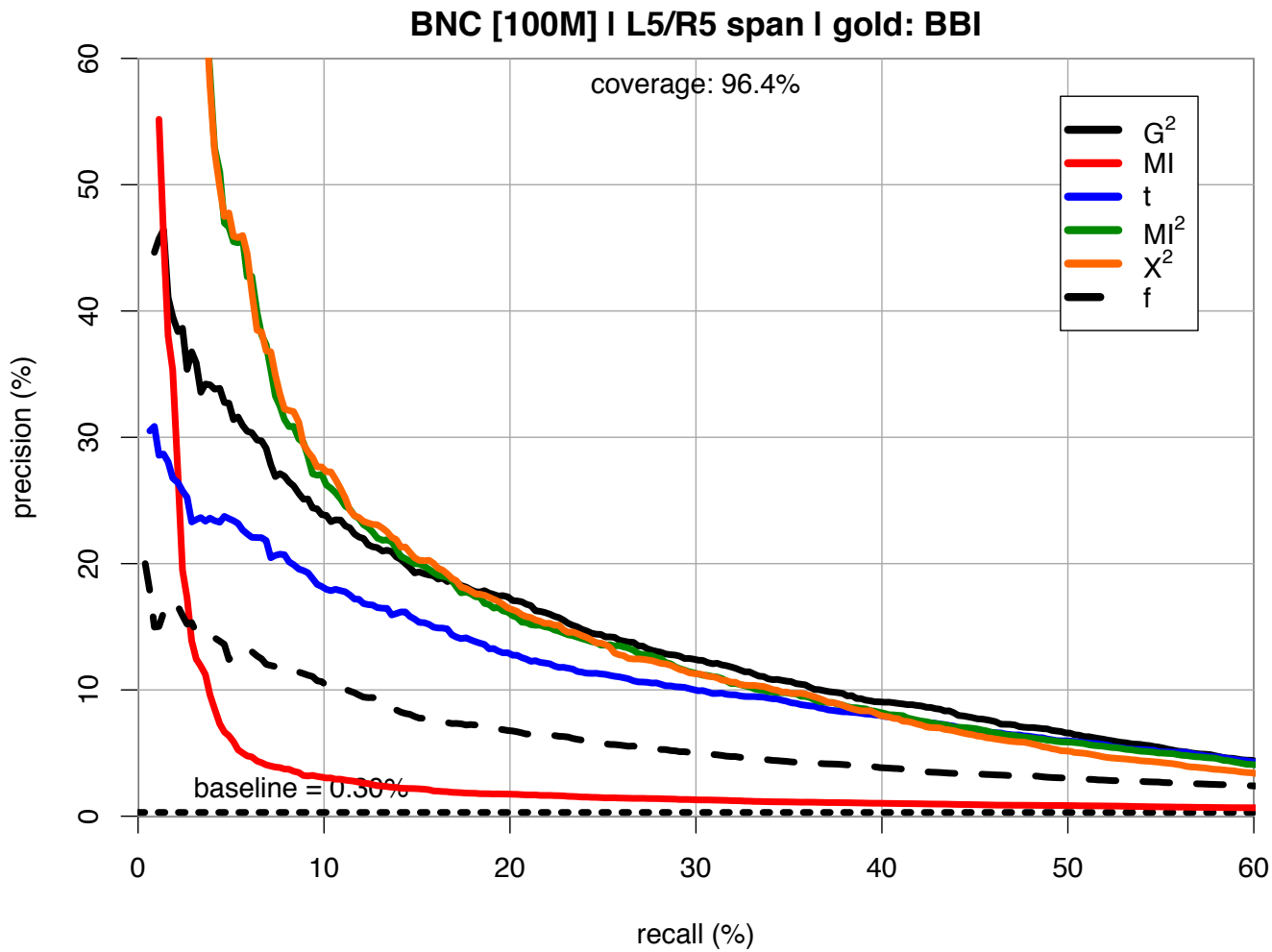
# Evaluation: precision vs. recall | BBI




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

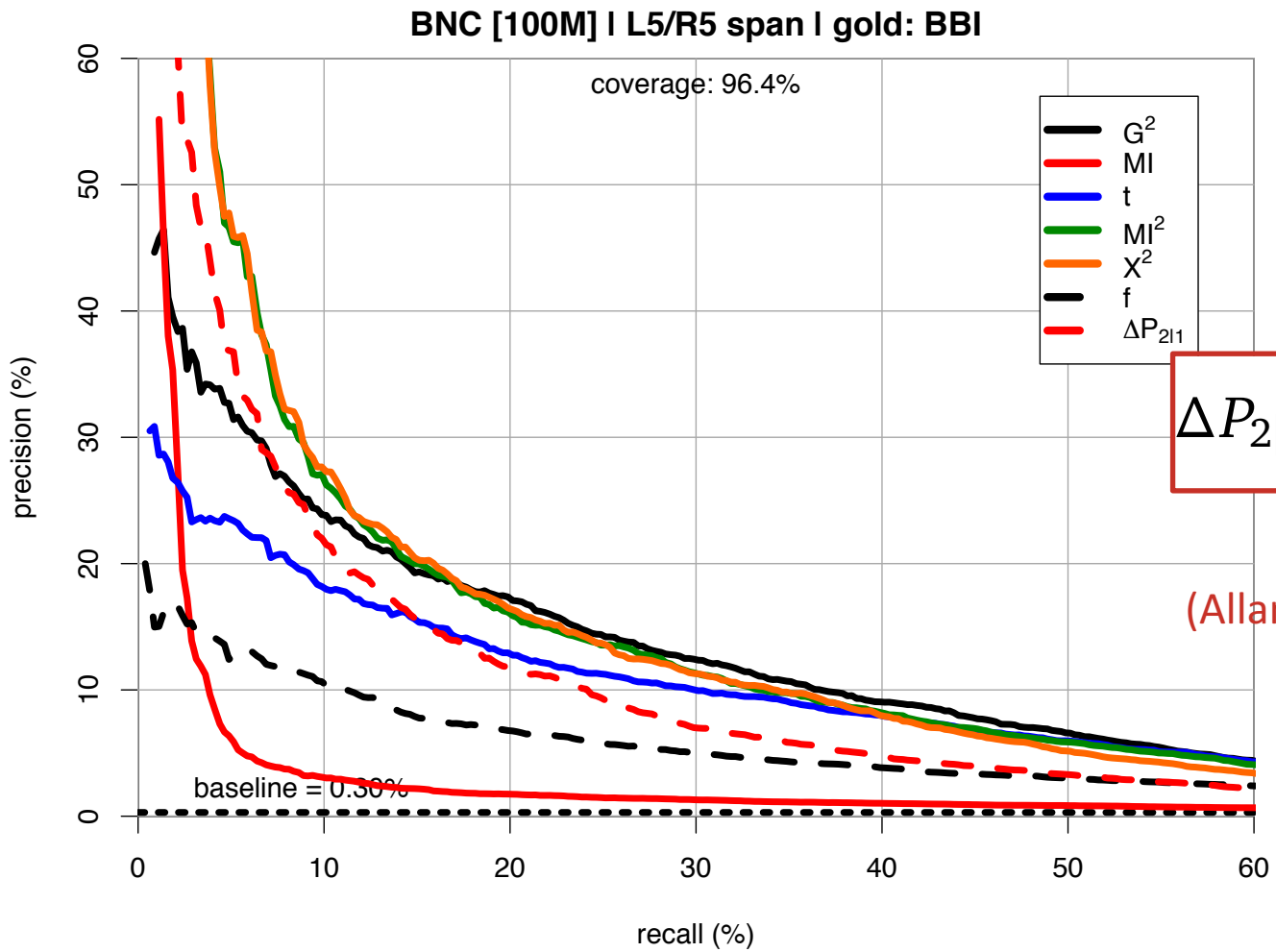
standard
chi-squared test

# Evaluation: precision vs. recall | BBI




co-occurrence frequency for comparison

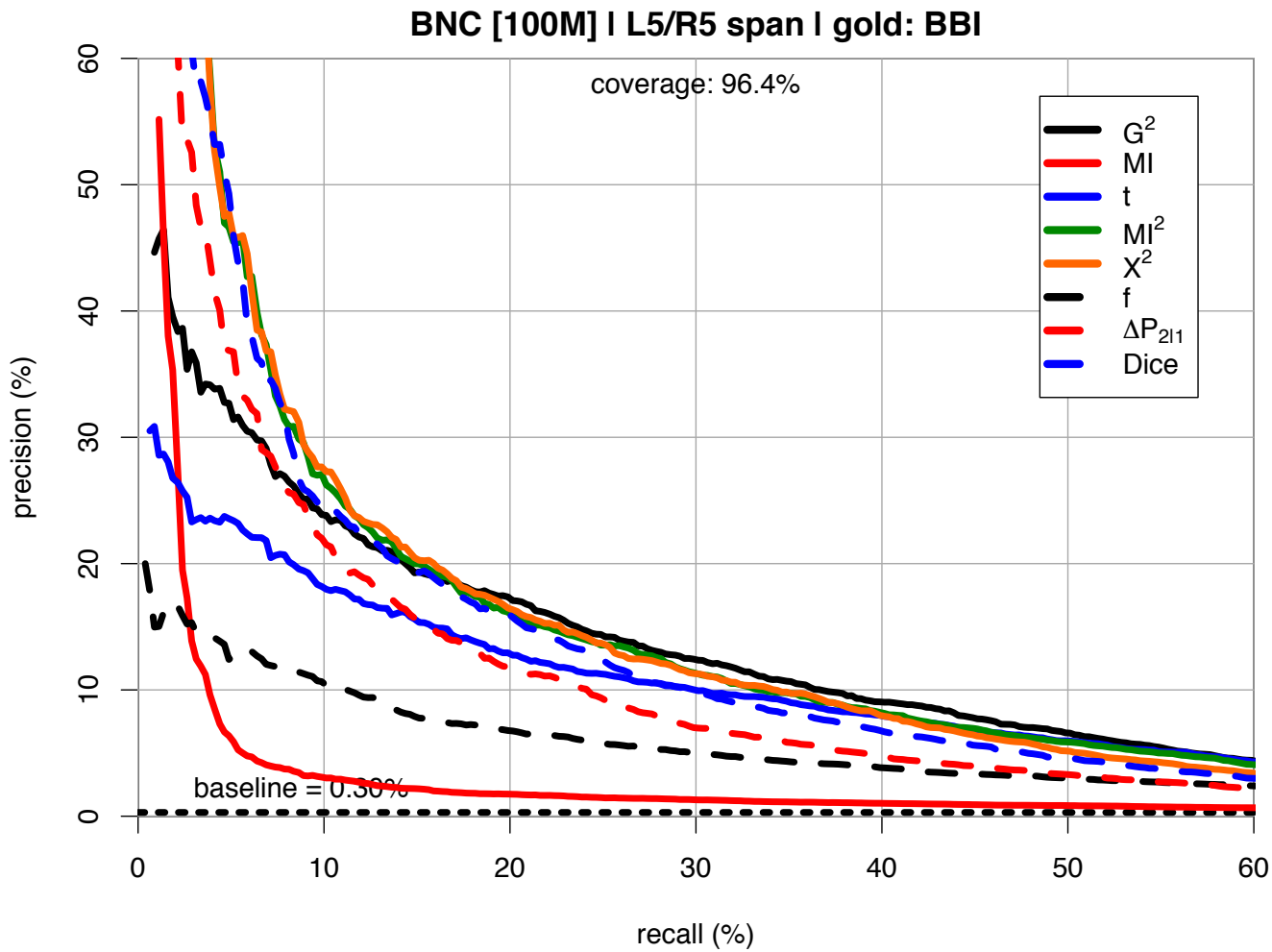
# Evaluation: precision vs. recall | BBI



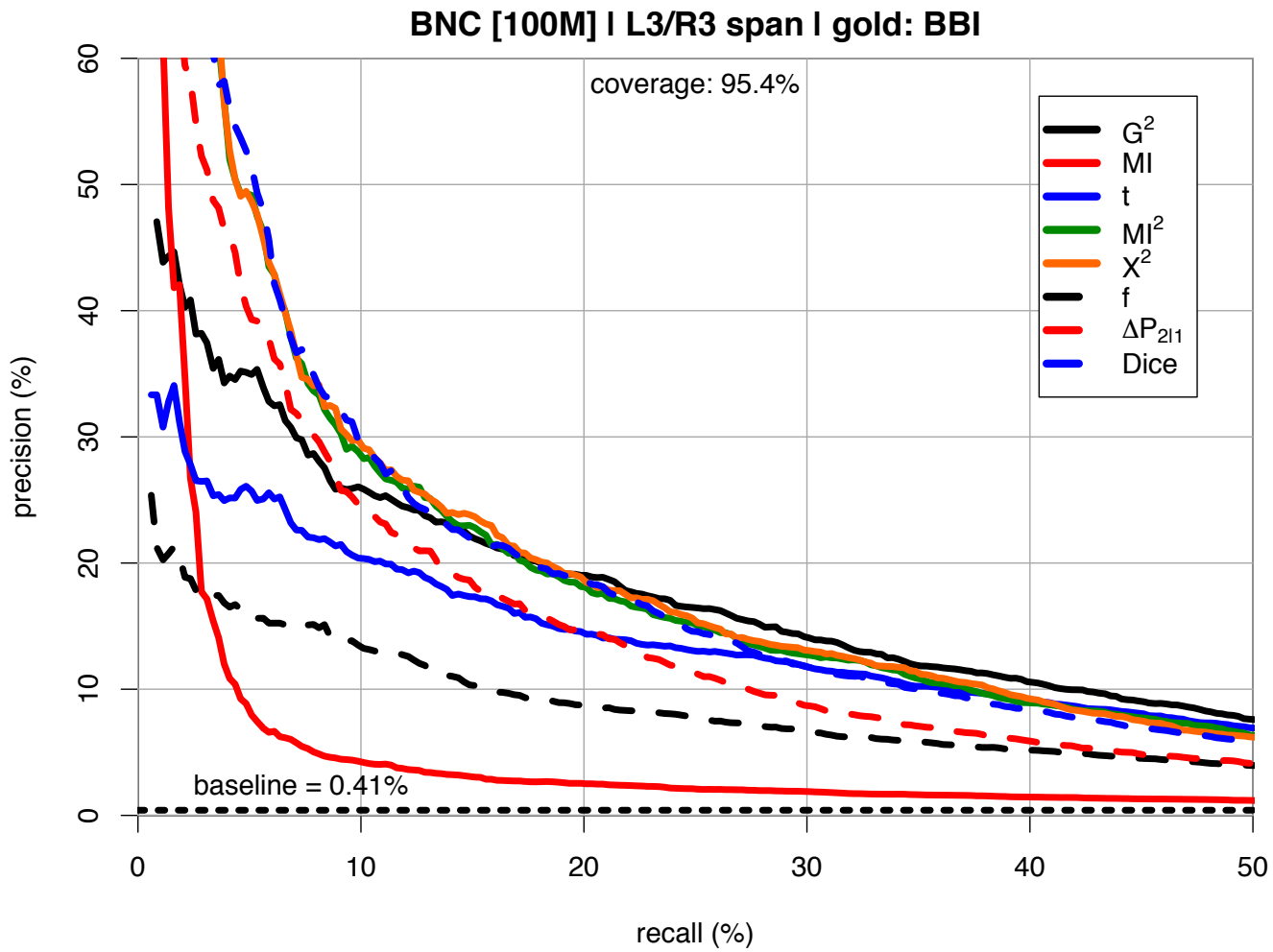

$$\Delta P_{2|1} = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$$

asymmetric ΔP
(Allan 1980; Gries 2013)

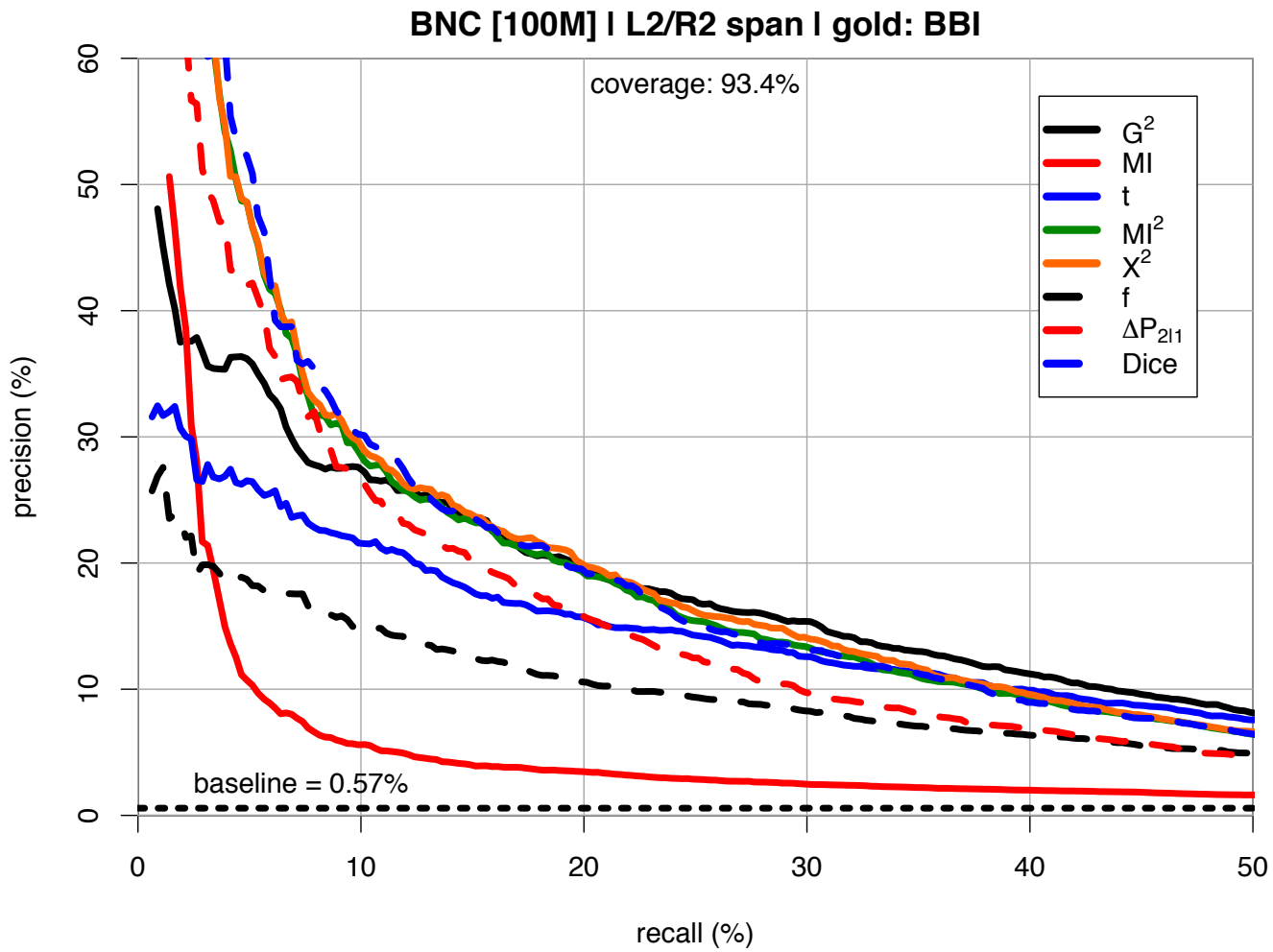
# Evaluation: precision vs. recall | BBI



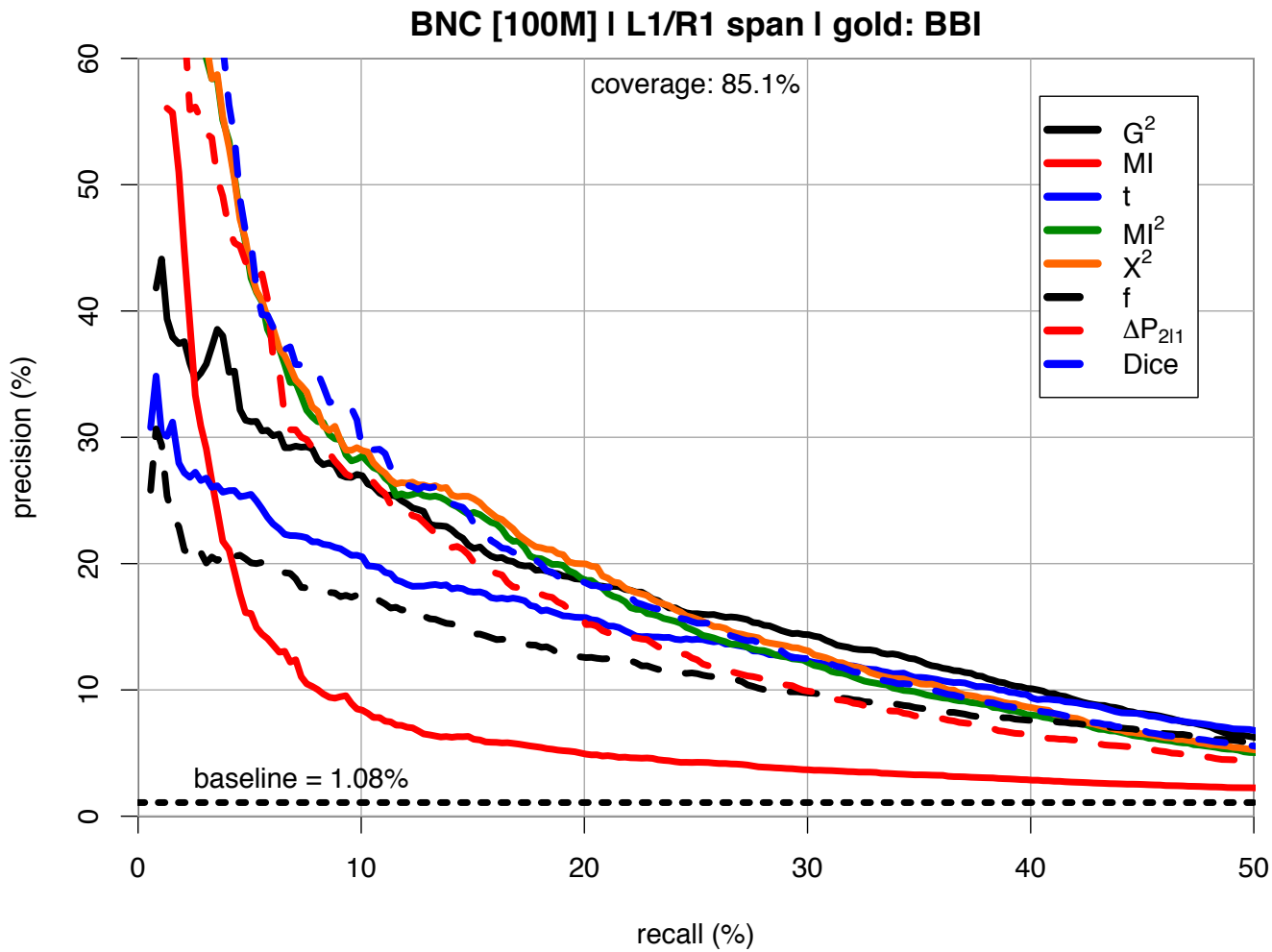
# Factor: context size | BBI



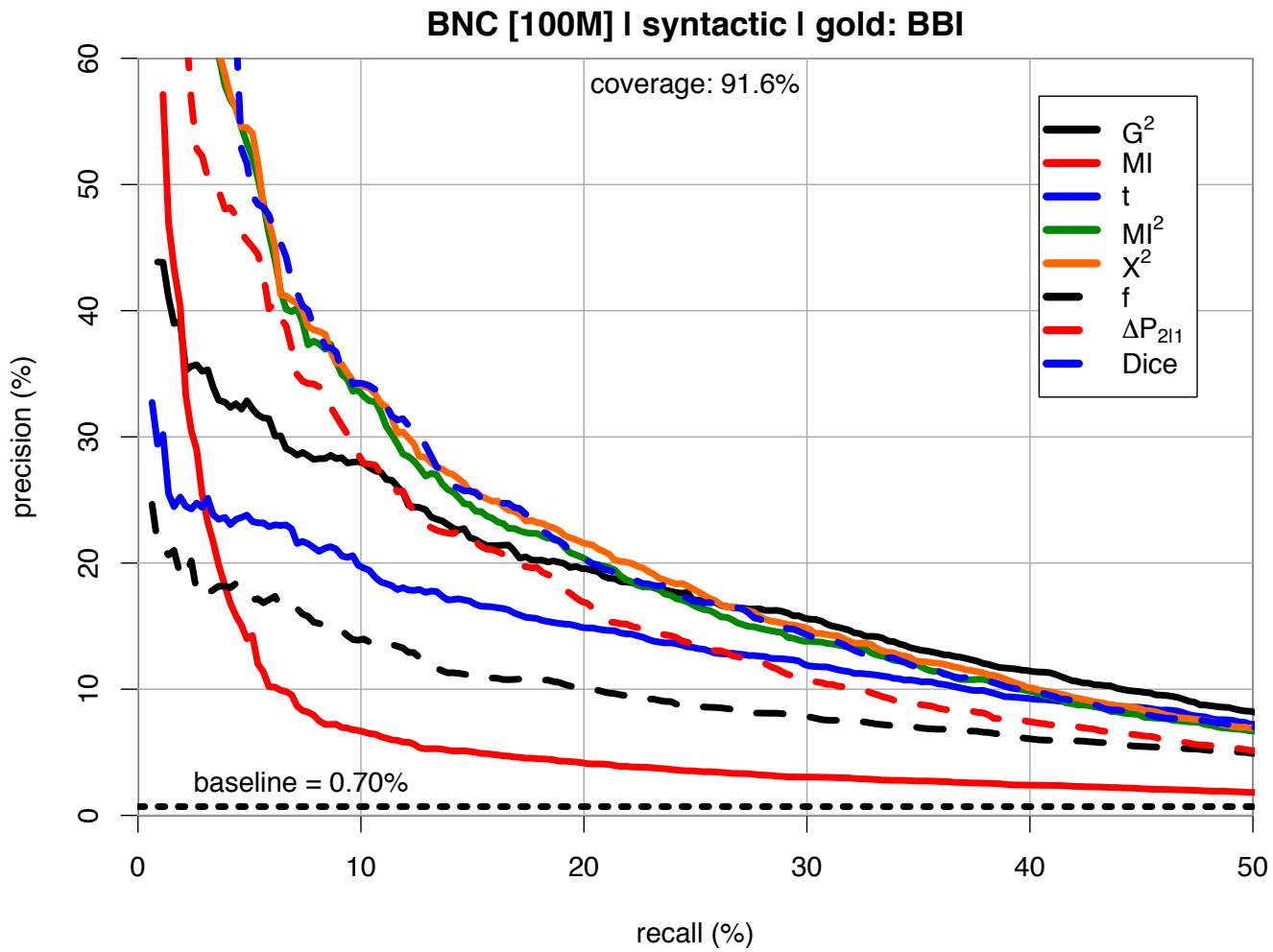
# Factor: context size | BBI



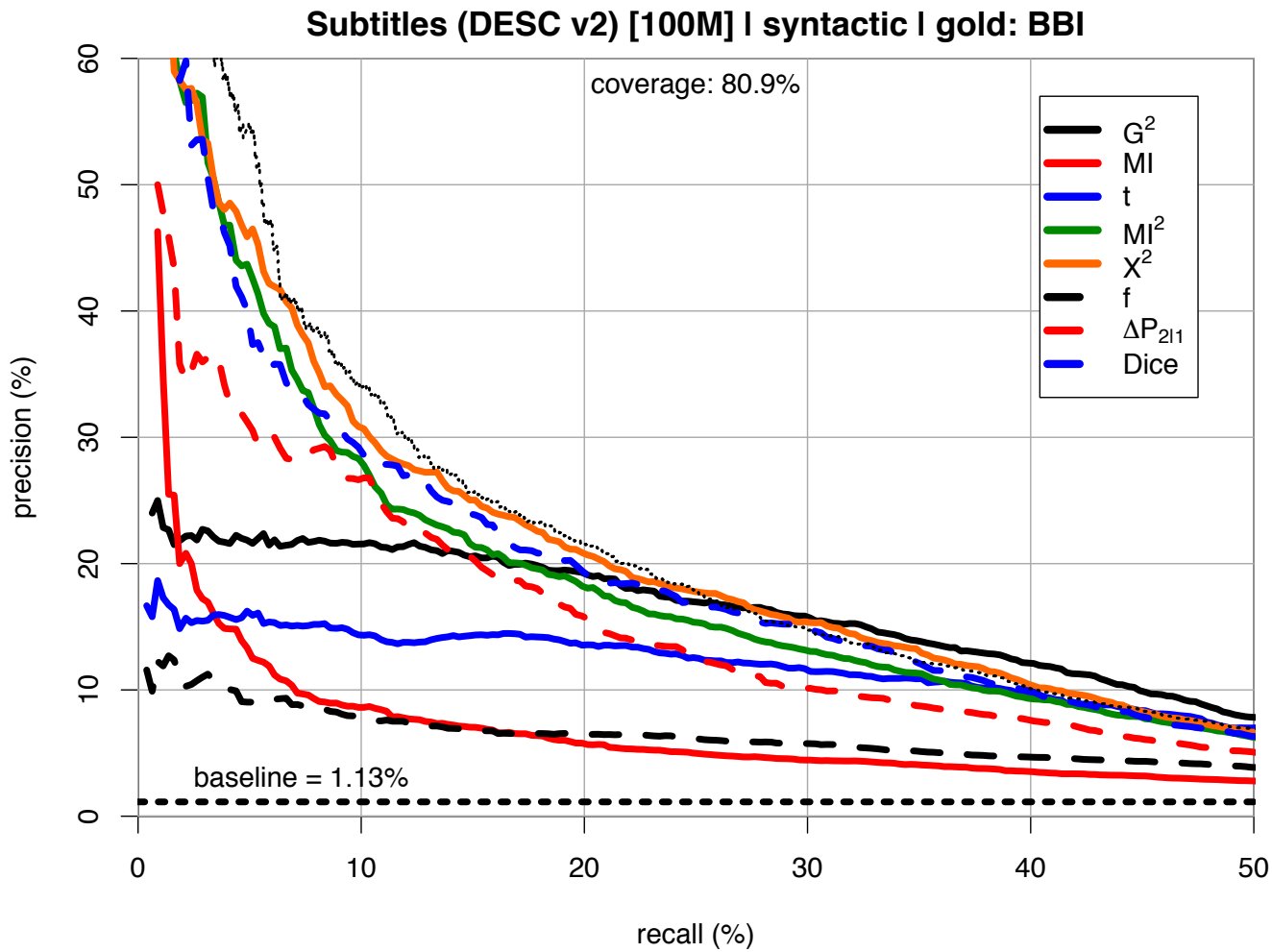
# Factor: context size | BBI



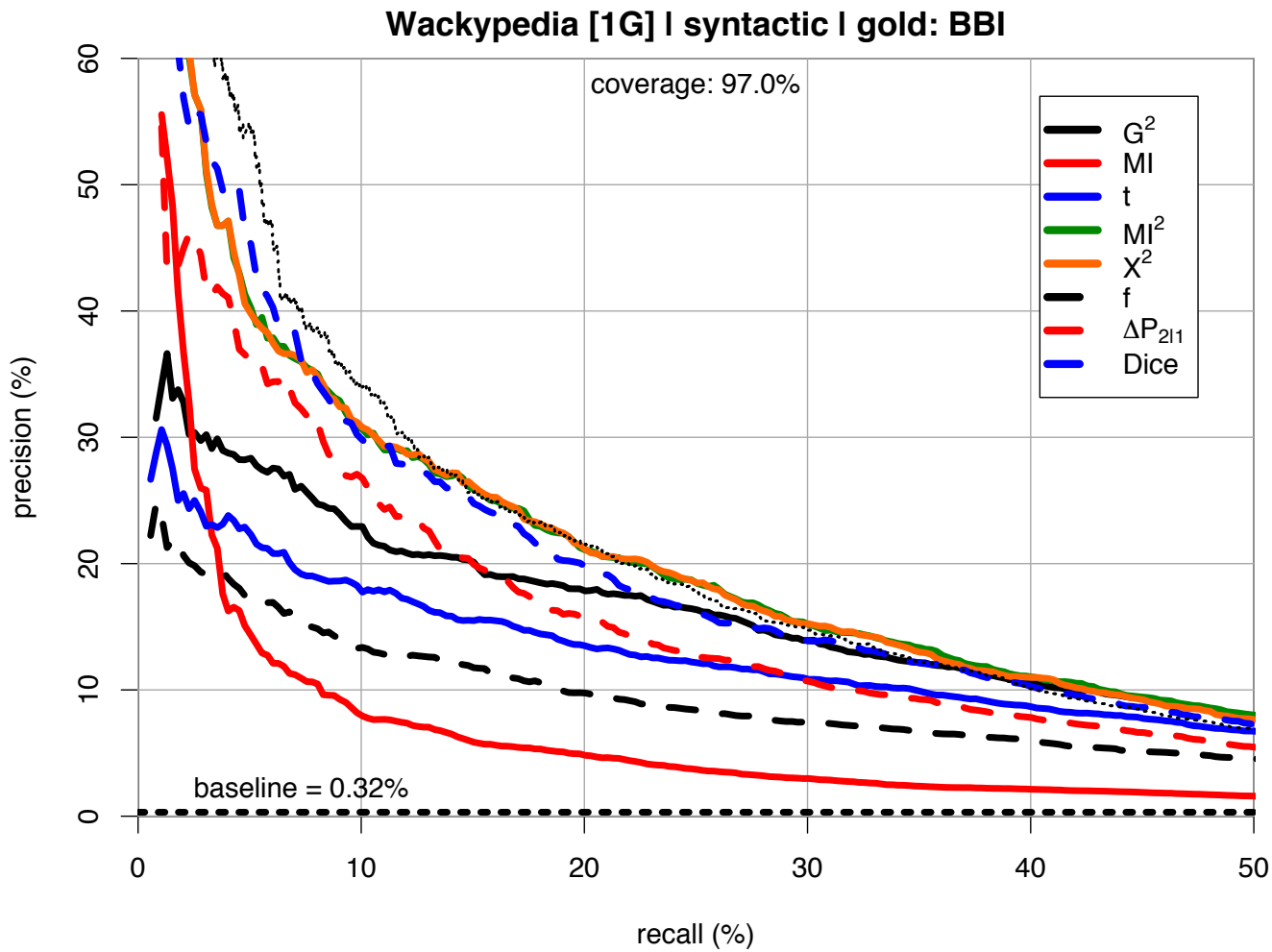
# Factor: context size | BBI



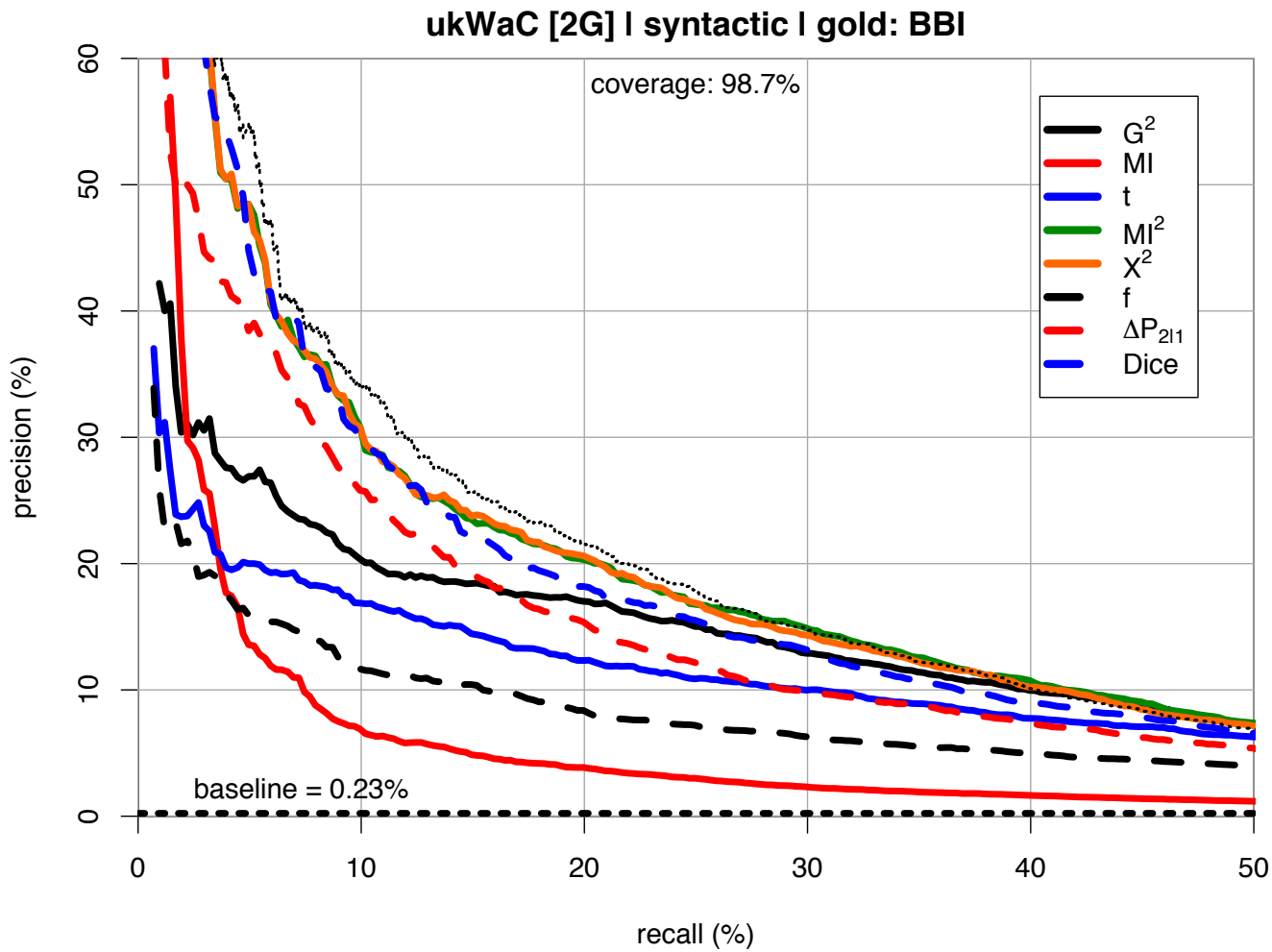
# Factor: corpus | syntactic| BBI



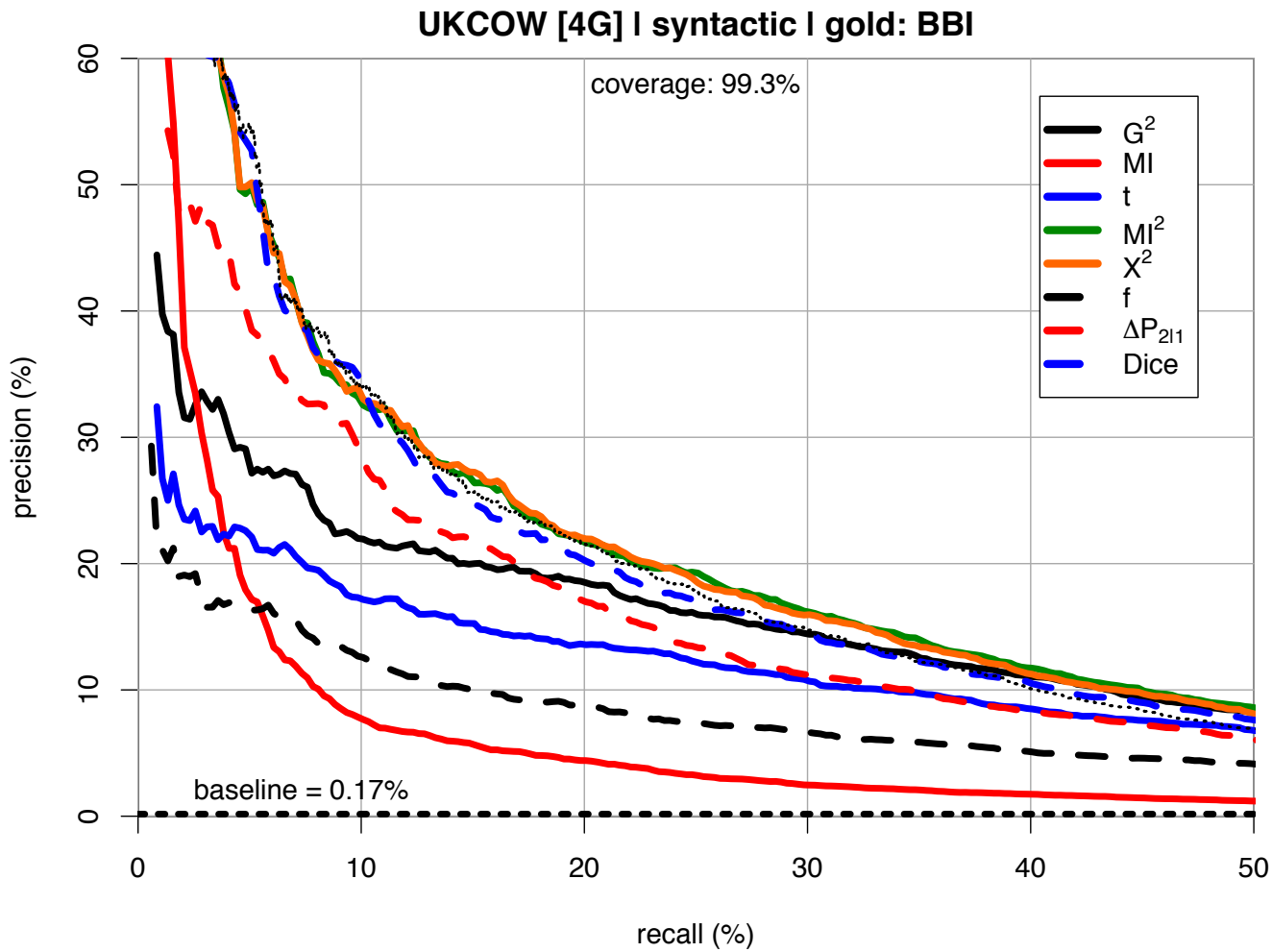
# Factor: corpus | syntactic| BBI



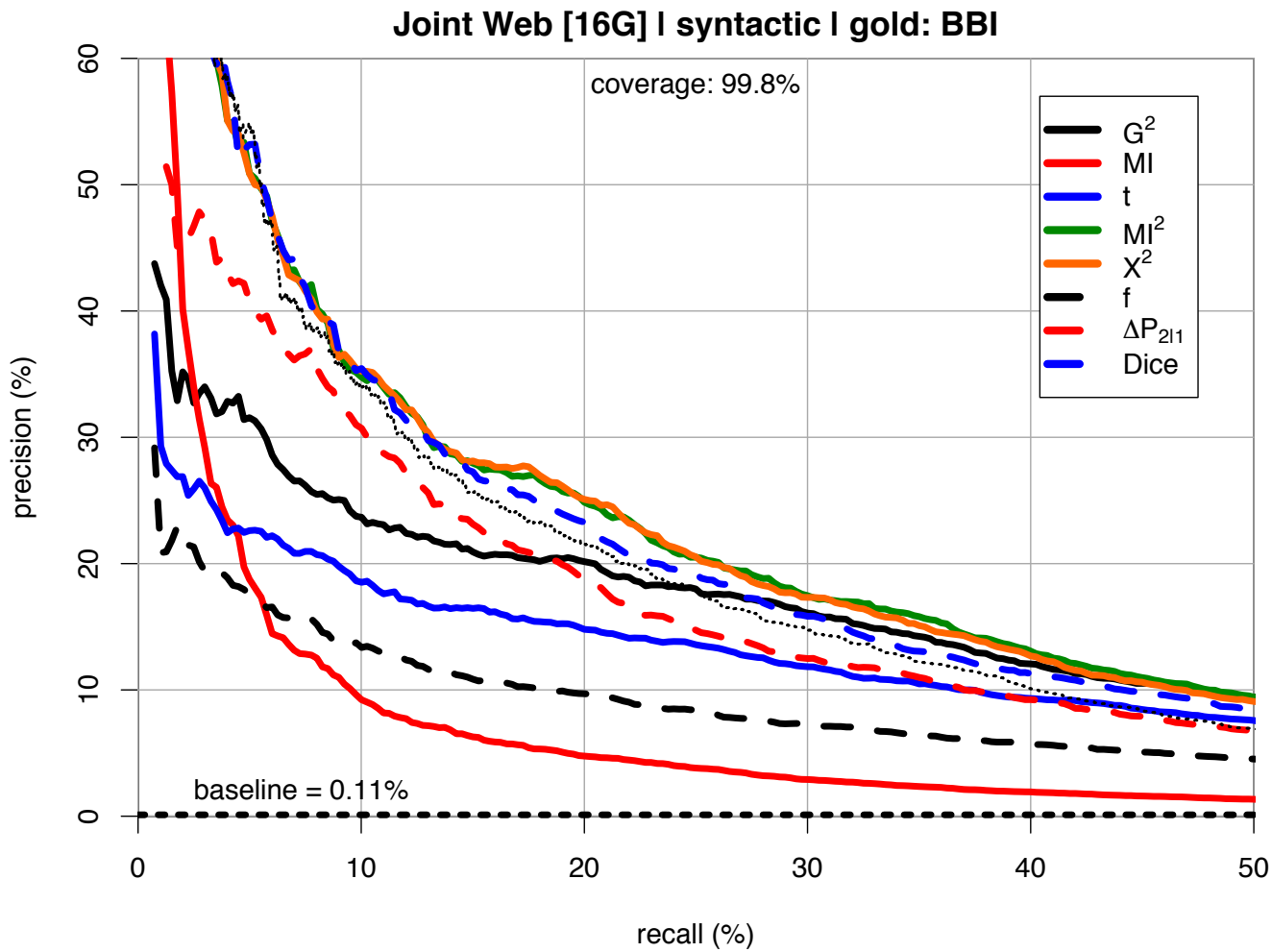
# Factor: corpus | syntactic| BBI



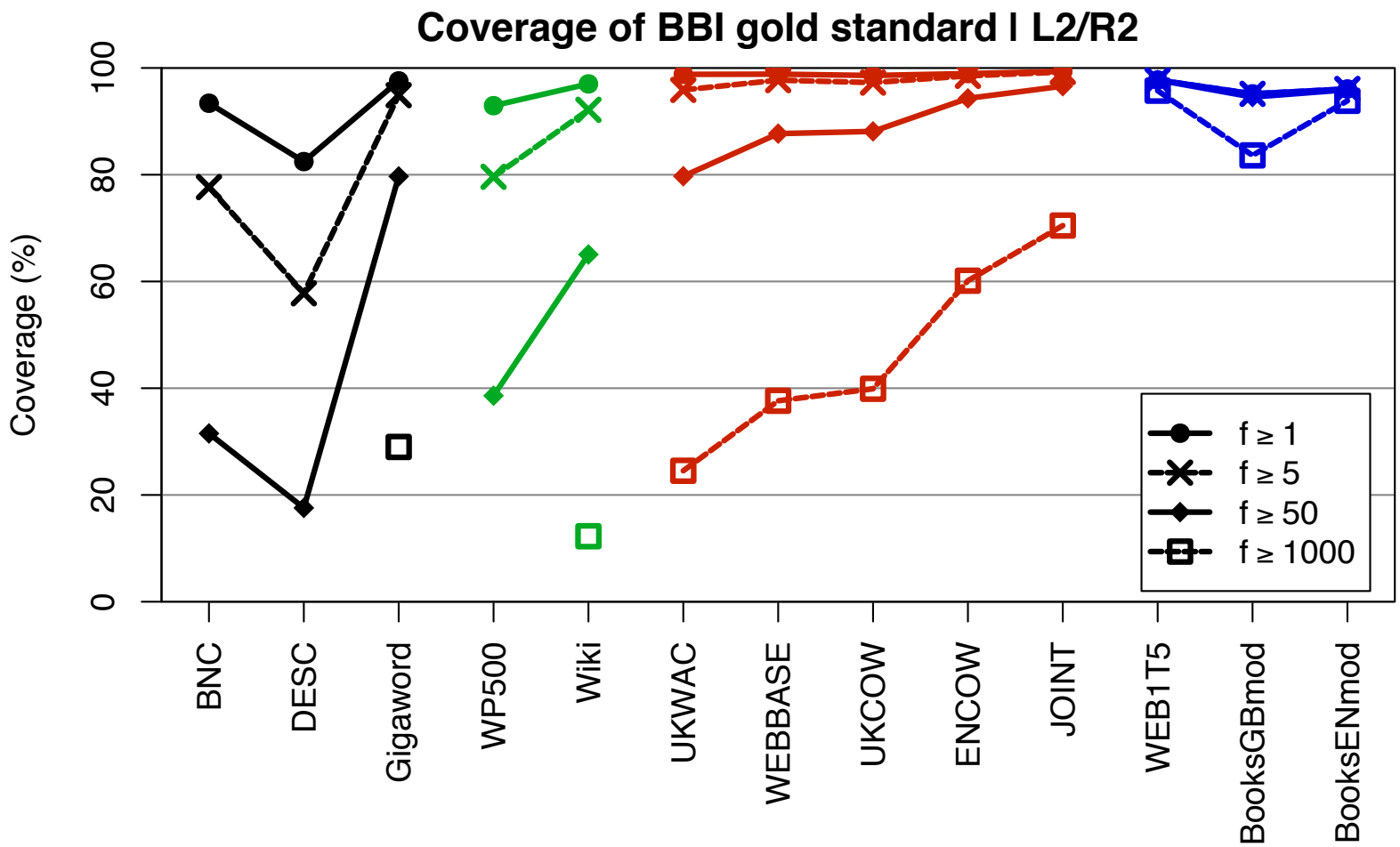
# Factor: corpus | syntactic| BBI



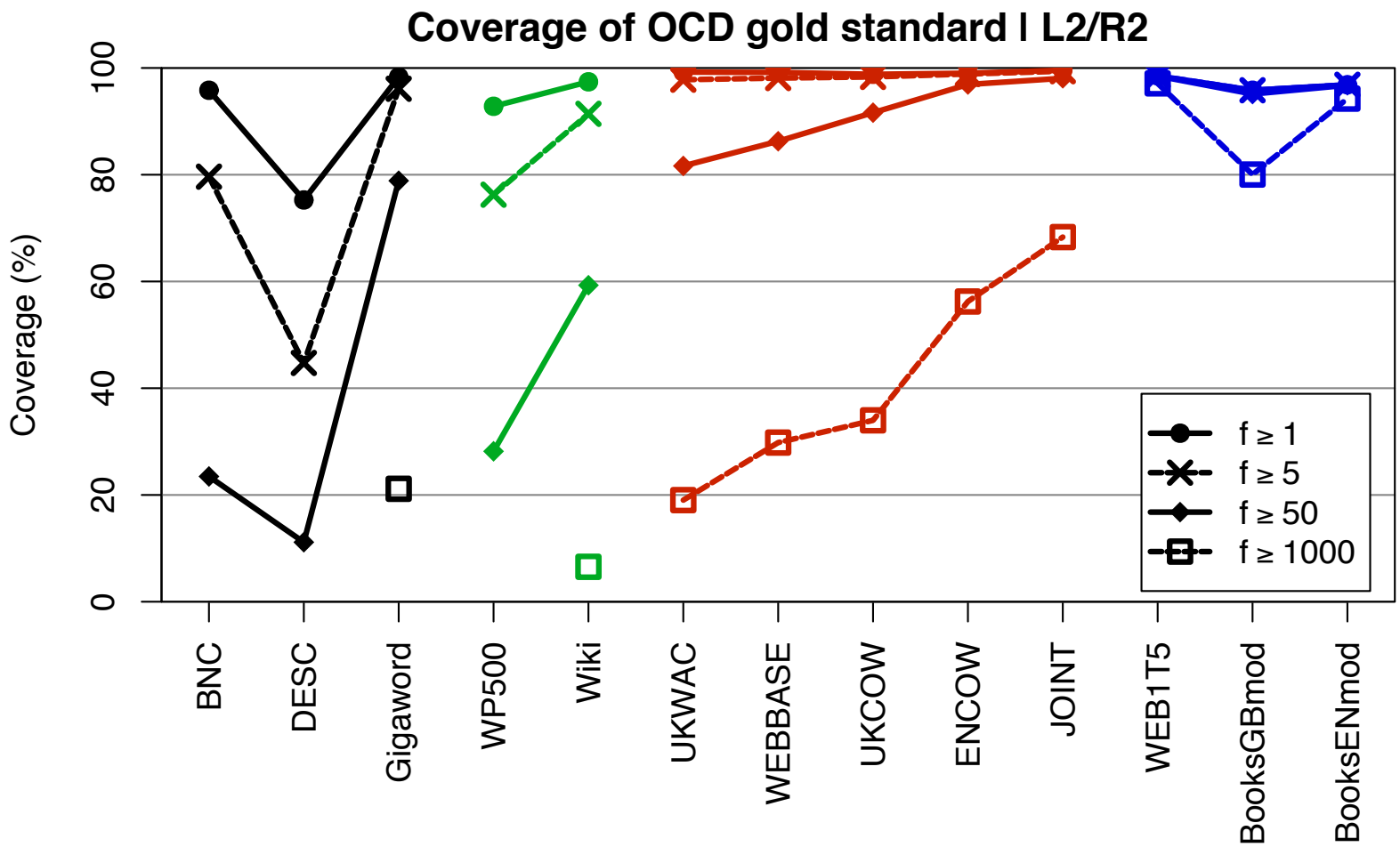
# Factor: corpus | syntactic| BBI



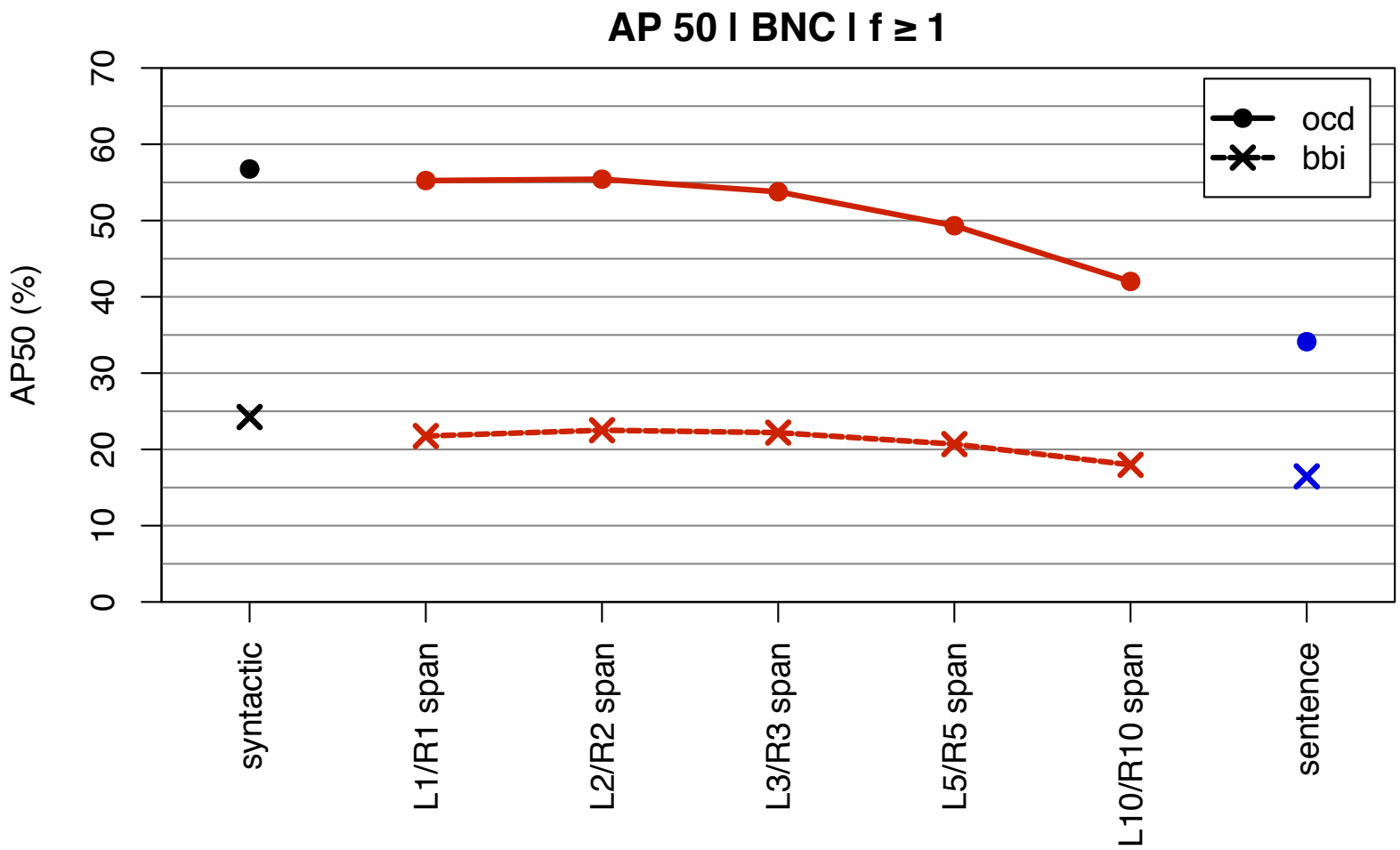
# Results: coverage



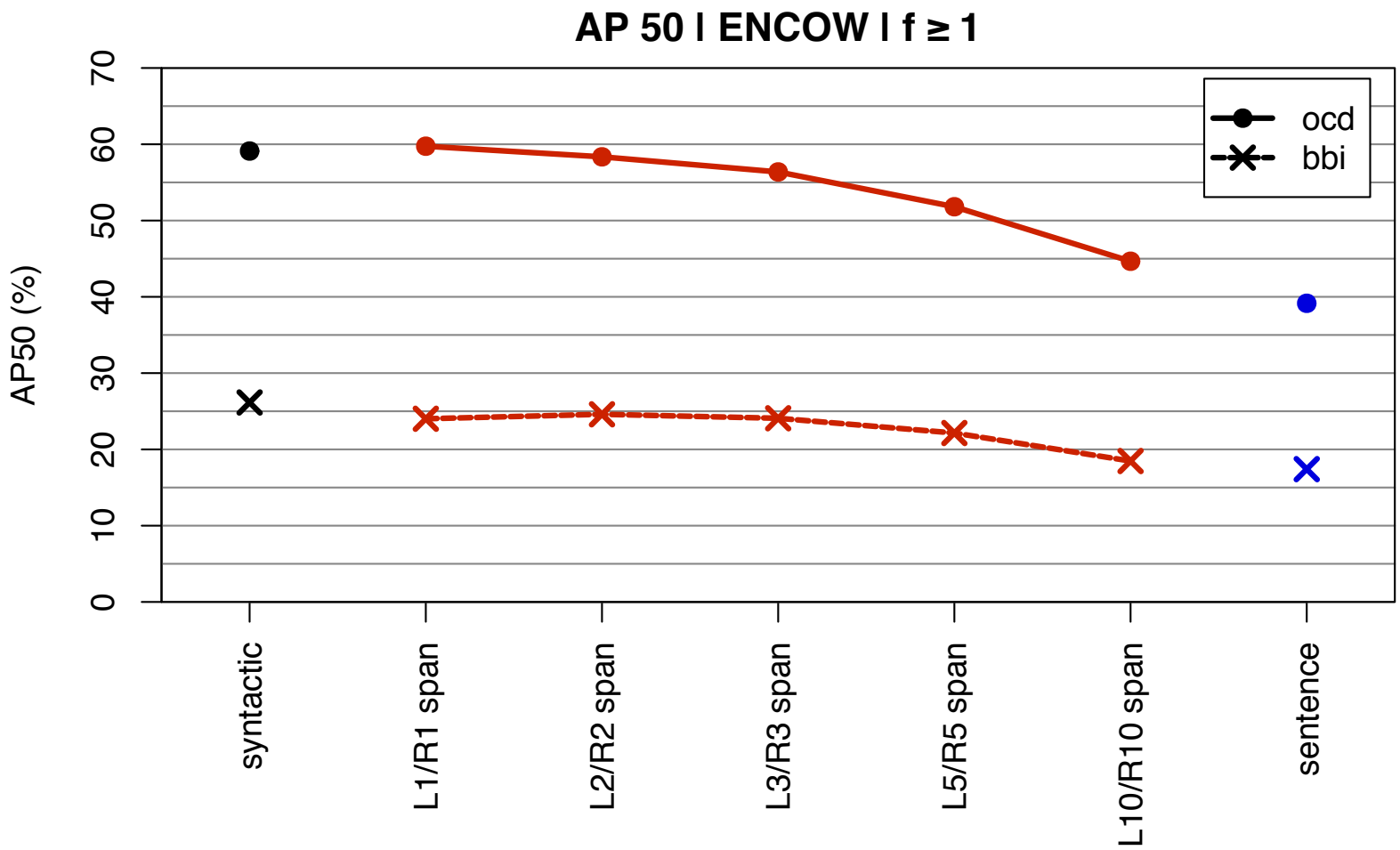
# Results: coverage



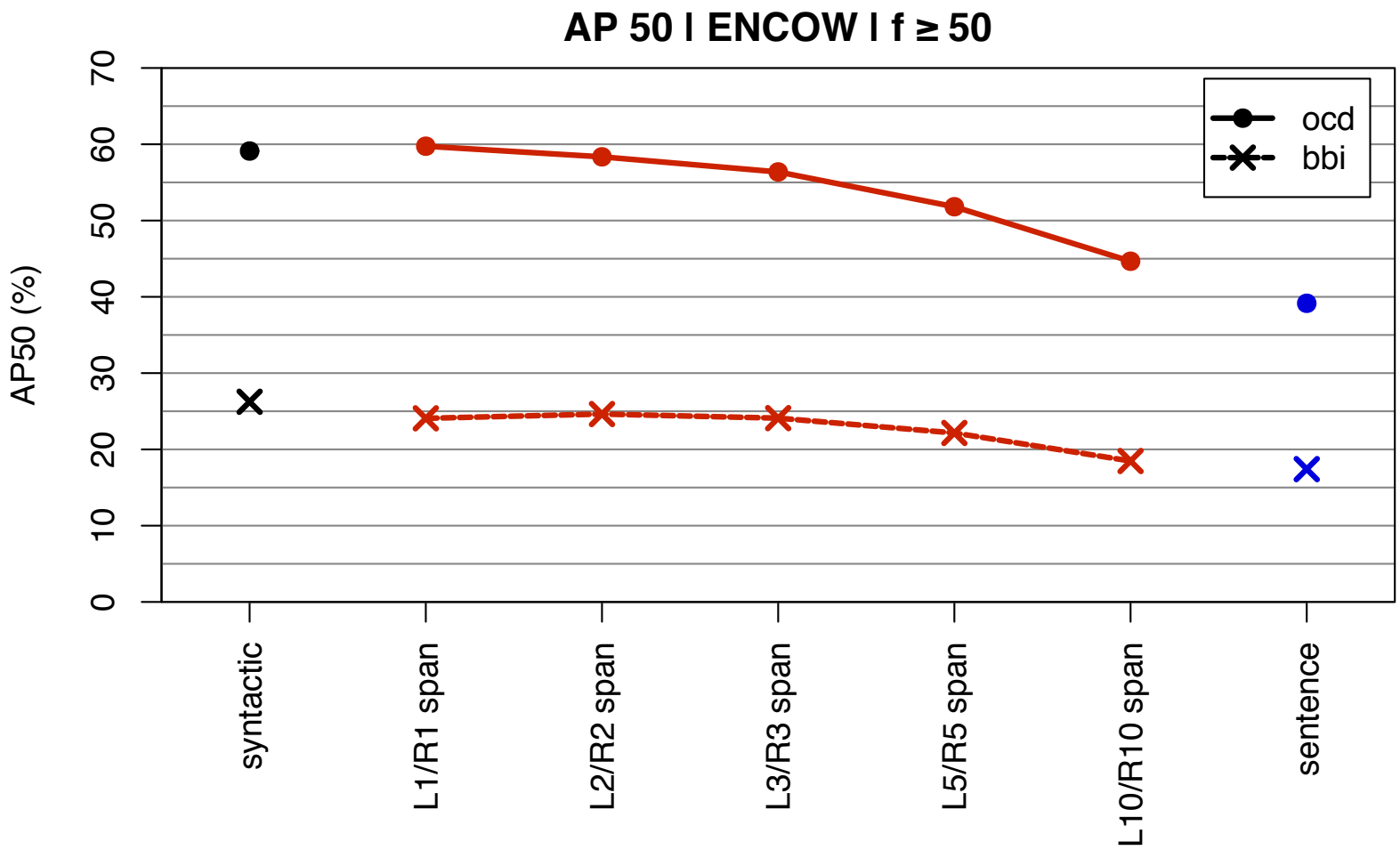
# Results: context size



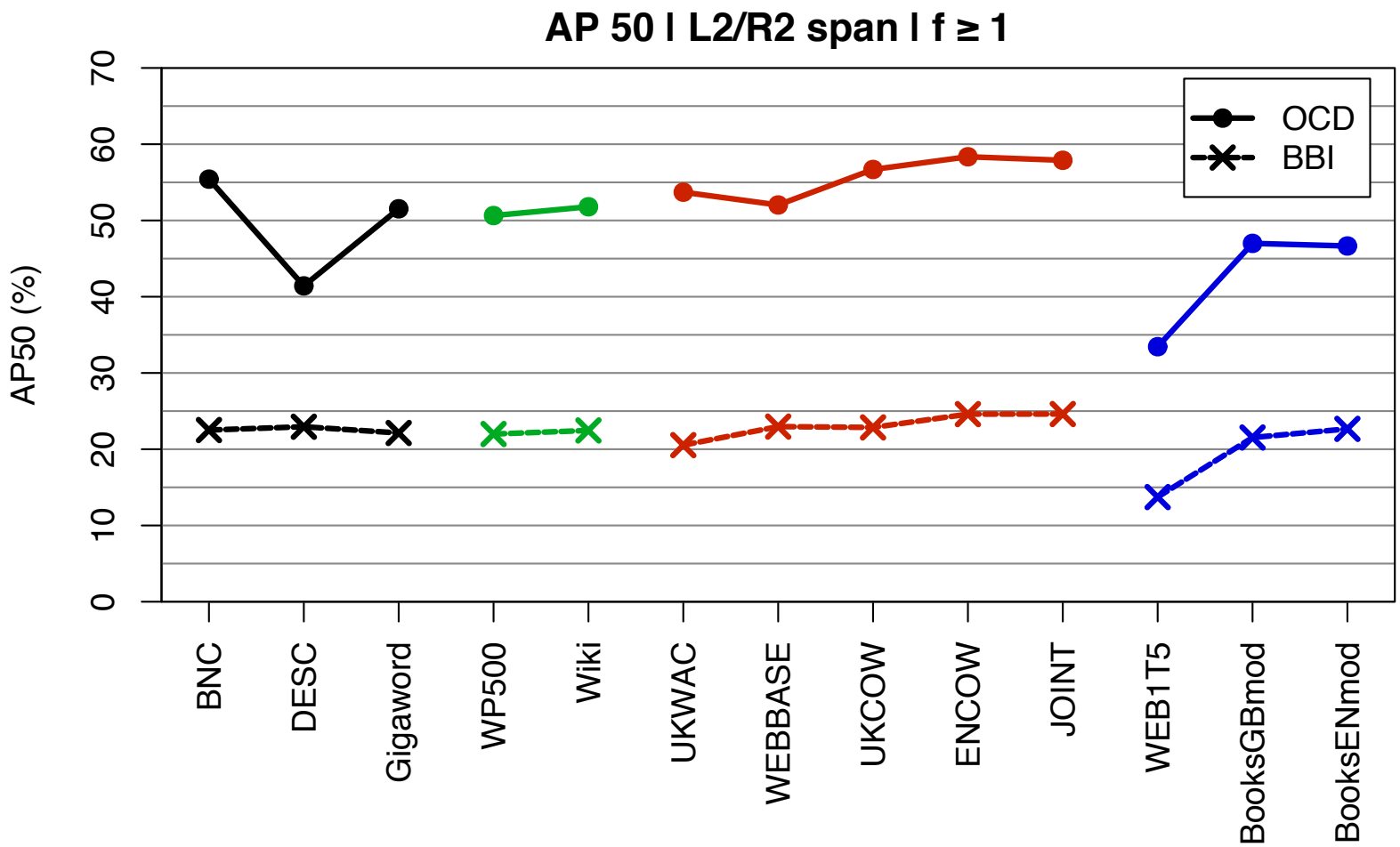
# Results: context size



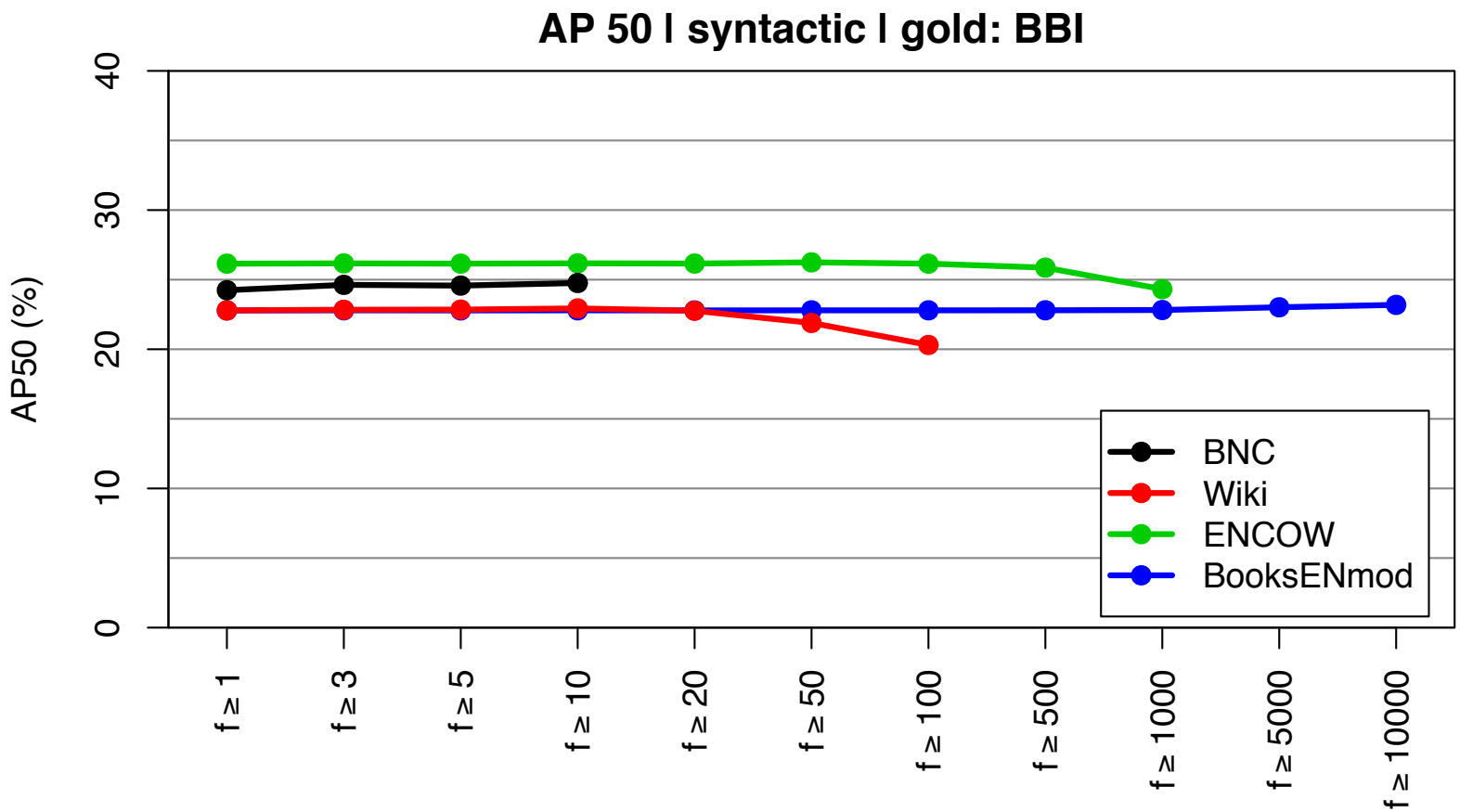
# Results: context size



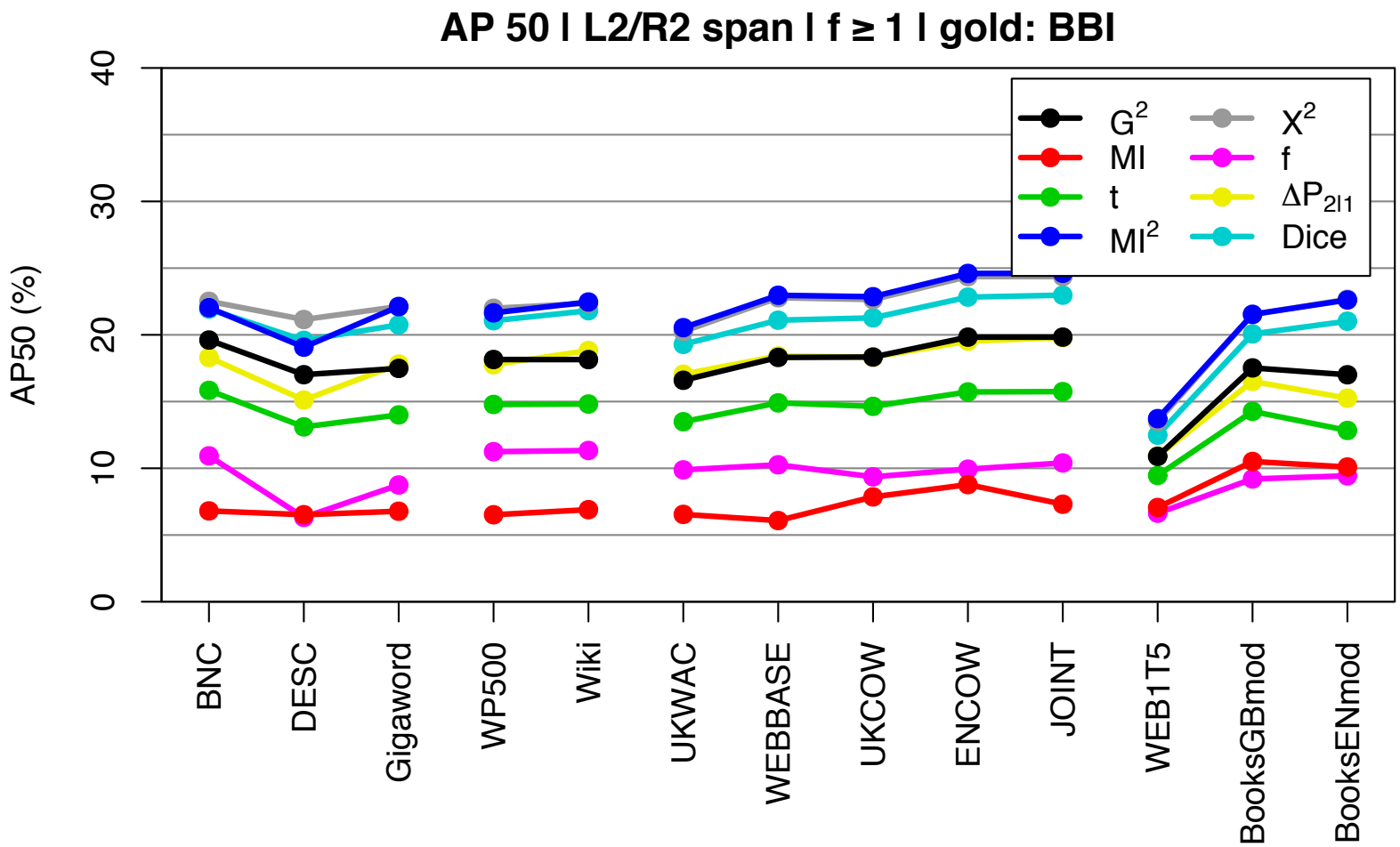
# Results: corpus | L2/R2 span



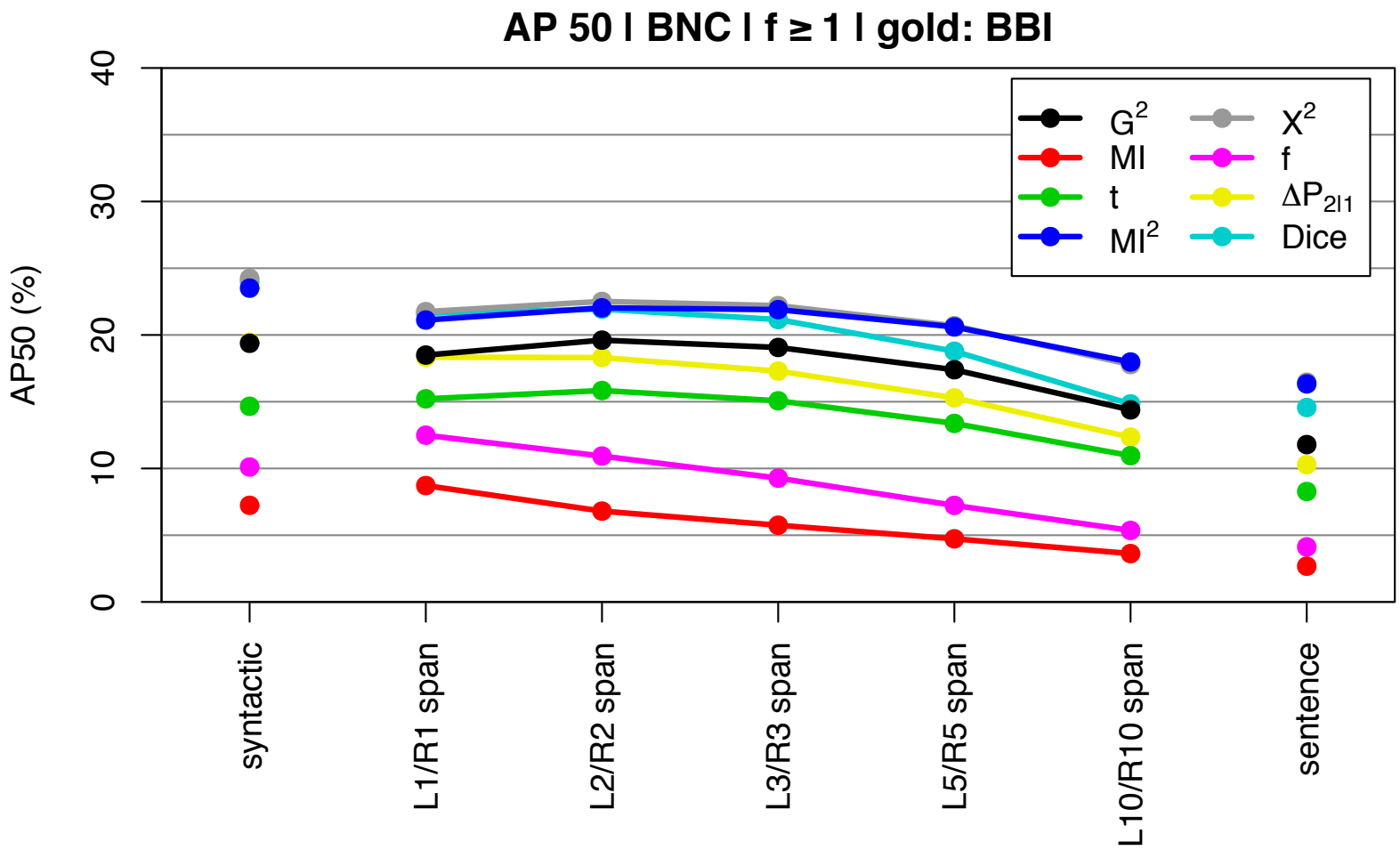
# Results: frequency threshold



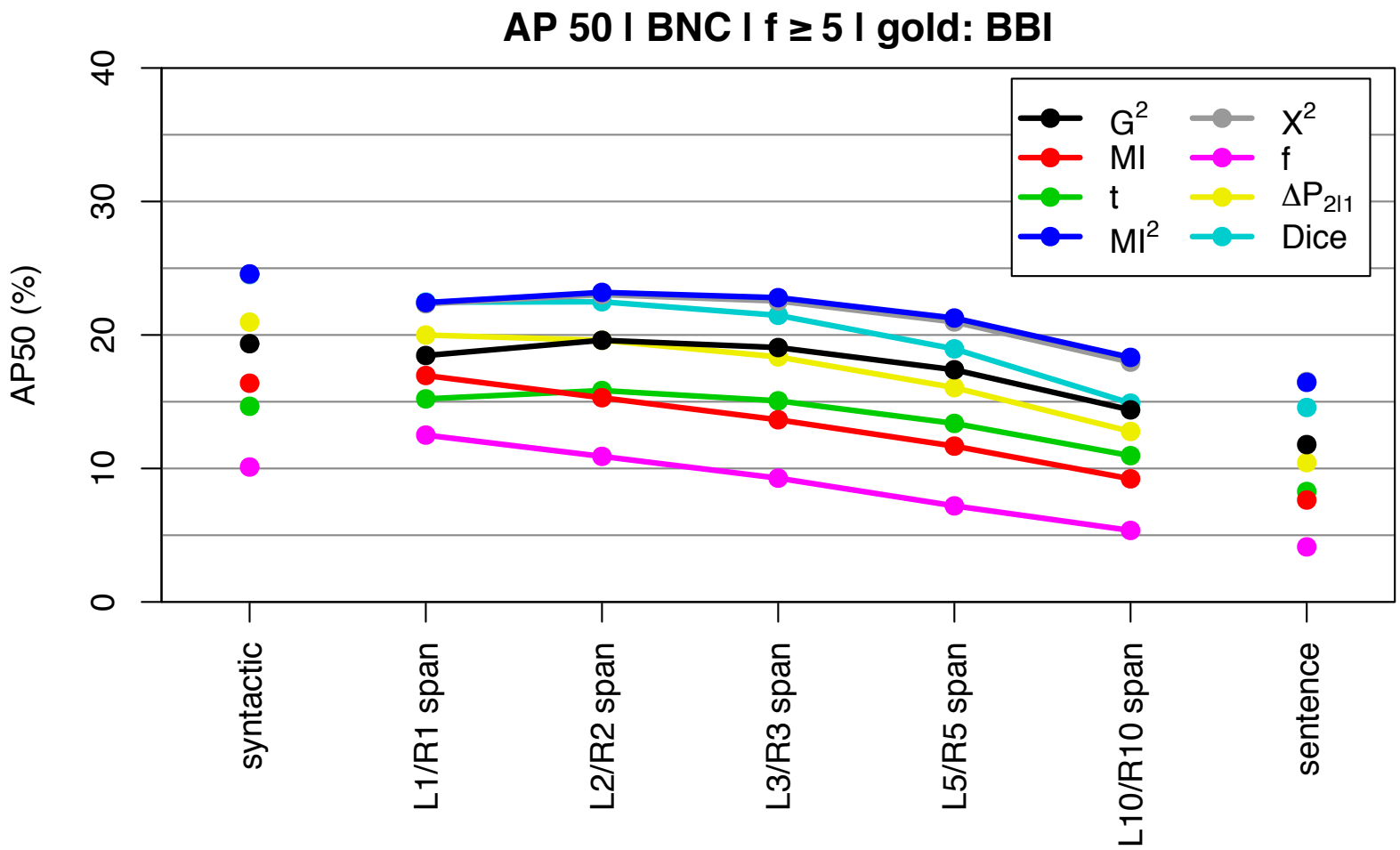
# Results: interactions



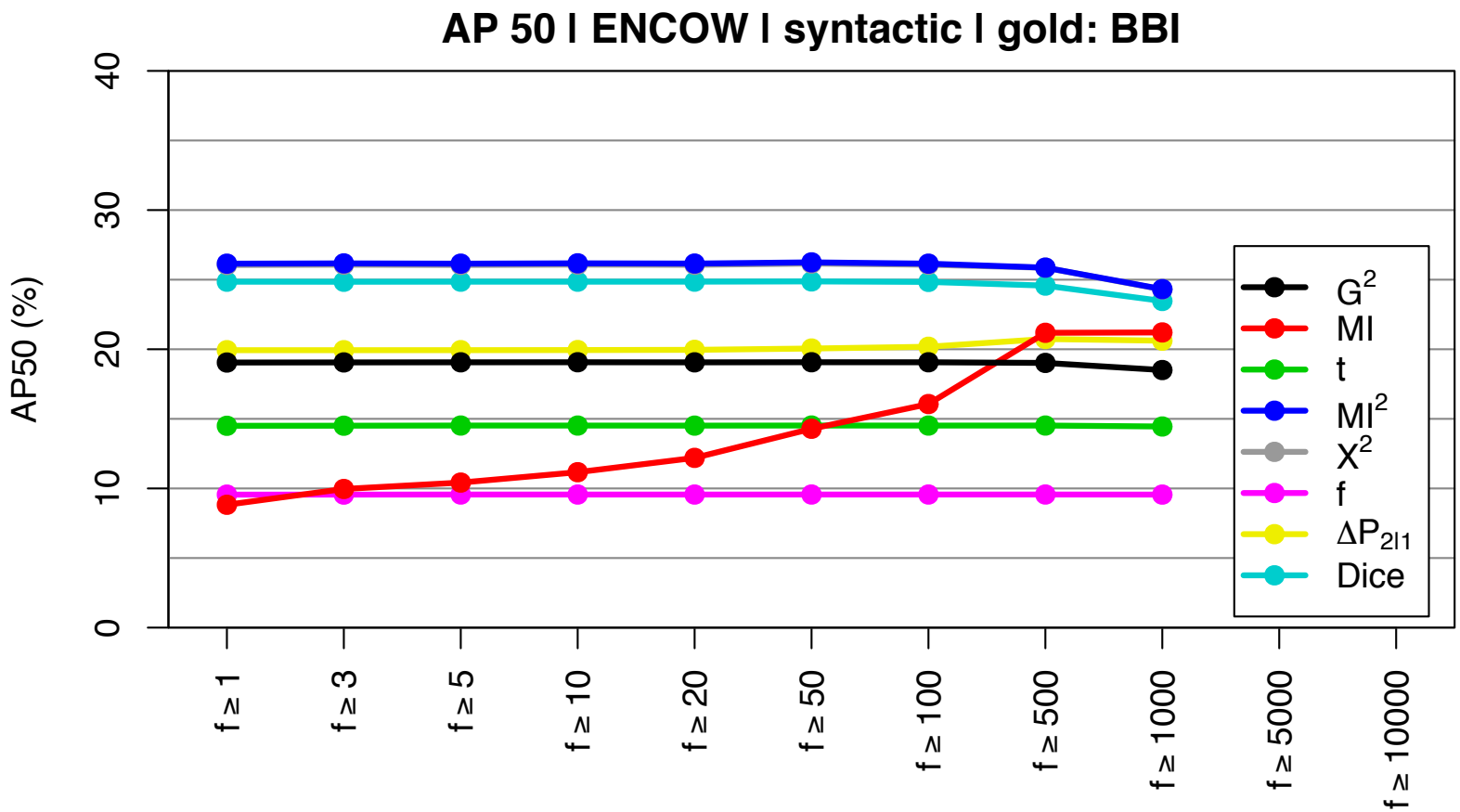
# Results: interactions



# Results: interactions



# Results: interactions



# Lost in results

- No space for complete results in talk or paper
  - you have to take our word that they look similar
- No room for the other 12 association measures
- AP50 may not be the most appropriate criterion
  - perhaps 30% recall sufficient, perhaps 80% needed
  - hides details of trade-off between precision and recall
  - do parameters affect the shape of P/R curves?
- Too much data for supplementary materials
  - 2.3 GiB of co-occurrence data (compressed)
  - gold standard cannot be redistributed

# 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!

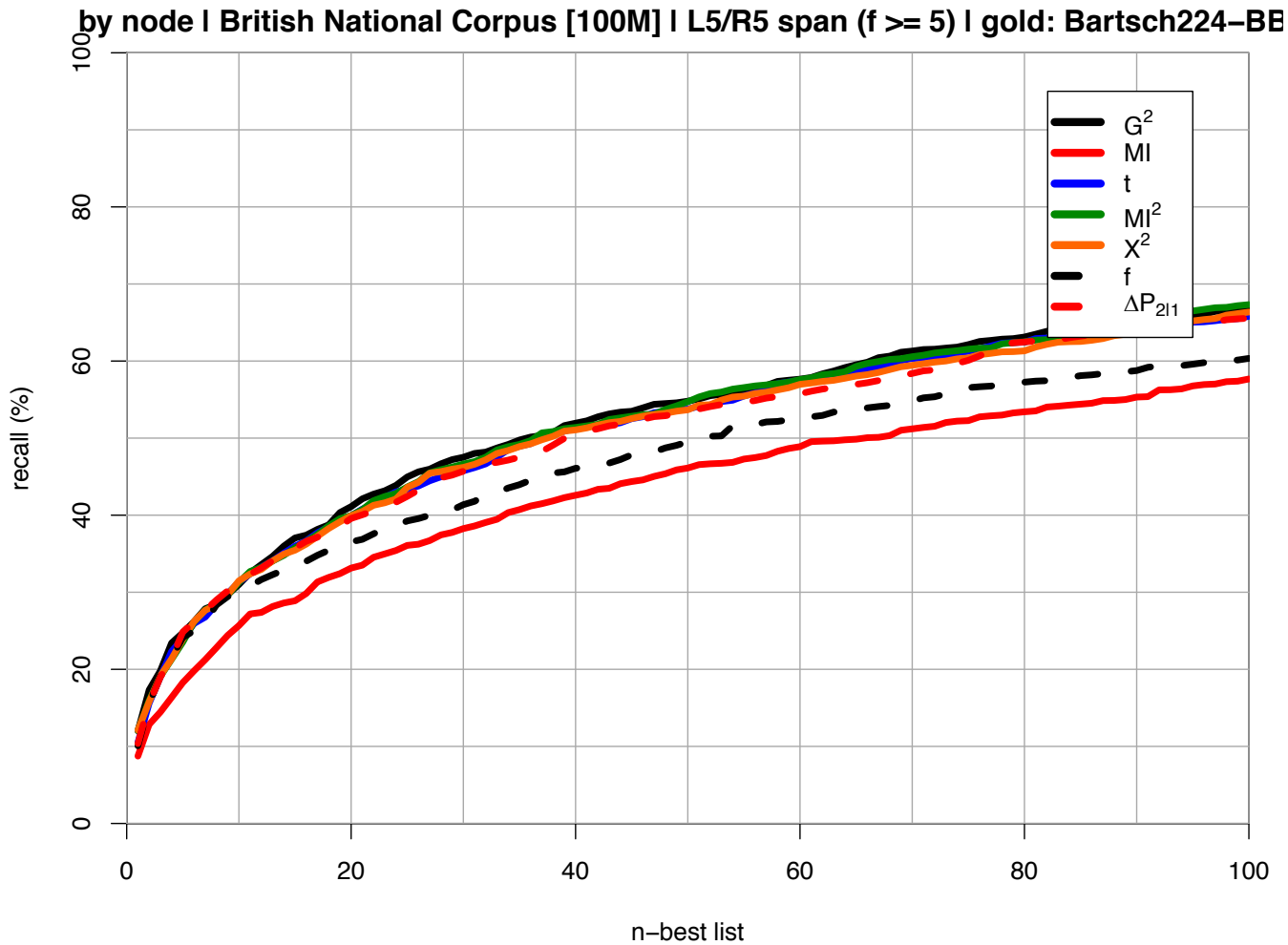
Questions?

# THANK YOU!

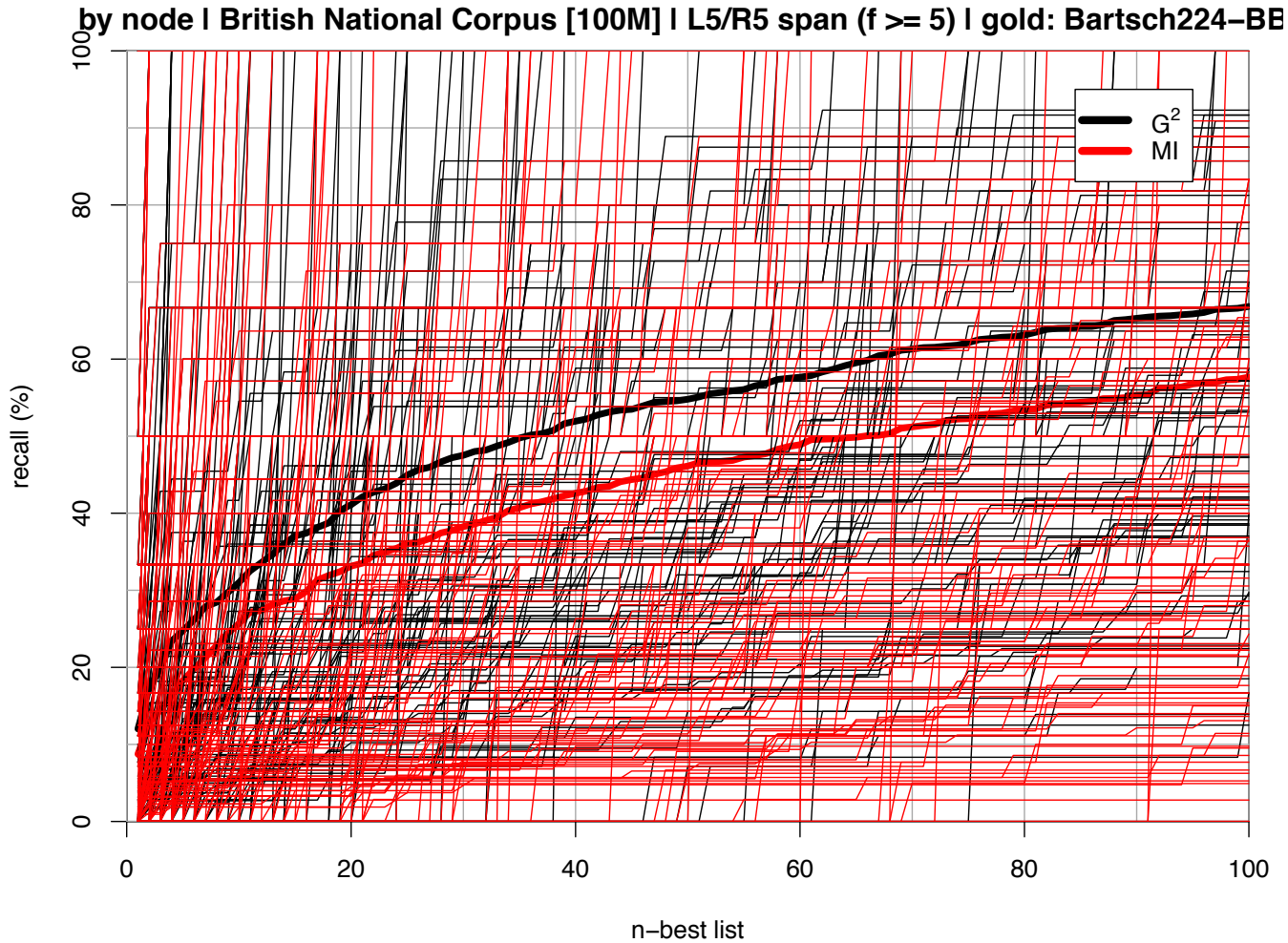
# Appendix: Technicalities

- Corpora indexed with IMS CWB (Evert & Hardie 2011)
- Data extraction with UCS toolkit (Evert 2004)
- Evaluation & plots with UCS/R
- Precision-recall data exported as JSON files
- E-VIEW-alation
  - client: Vega 2.6 + JQuery UI
  - server: Perl CGI script serves requested JSON data

# Appendix: Per-node evaluation



# Appendix: Per-node evaluation



# Appendix: Manual validation

## discrepancies between BBI / OCD2 and corpus data

**Bartsch 224, ENCOW, Malt dependencies, X2 'shake' (1000 candidates)** [bartsch]

3 / 50 Go << >> label for entry #3287924 set to FP [export] back to main page

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 11477613 | shake | furiously | 41 | 6971.790 | --- | OCD | | TP | | Set |
| 17066993 | shake | quake | 42 | 6818.999 | --- | --- | | TP | | Set |
| 15079184 | shake | milk | 43 | 6617.603 | --- | OCD | (under collocate) | TP | | Set |
| 17993610 | shake | snort | 44 | 6486.647 | --- | --- | | TP | | Set |
| 17878103 | shake | scuttle | 45 | 6073.180 | --- | --- | | TP | LSP | Set |
| 16052416 | shake | pan | 46 | 5520.637 | --- | --- | | TP | | Set |
| 7184709 | shake | cry | 47 | 5308.692 | --- | --- | | FP | | Set |
| 1753930 | shake | ass | 48 | 5103.859 | --- | --- | | TP | | Set |
| 11878568 | shake | grab | 49 | 4952.691 | --- | --- | | --- | | Set |
| 10226053 | shake | explosion | 50 | 4783.420 | --- | OCD | (under collocate) | TP | | Set |
| 17994048 | shake | swarm | 51 | 4777.921 | --- | --- | | TP | LSP | Set |
| 3287924 | shake | body | 52 | 4619.348 | --- | OCD | (under collocate) | FP | | Set |

variable ordering: A..B 68.2% / B..A 31.8% non-contiguous: 90.9% adjacent

The GTC 's chassis , the stiffest of any convertible in the world , provides a firm foundation for suspension control , minimising **scuttle shake** and contributing to the GTC 's refined handling .

They do handle well , have only driven the cab which to be honest can suffer from **scuttle shake** on poorly surfaced roads but it still handles way better than my previous Saab 9 -3 Aero ( later shape on Hirsch springs ) , god only knows what a well fettled Coupe on coilovers , new bushes and uprate ARB 's is like , suspect not much modern stuff is any better .

Ride quality is generally acceptable for this class of car but there are signs of **scuttle shake** if you hit mid-corner bumps at speed .

Only the very occasional sound of movement between the tiny rear quarter window and the side window rubber betrays any hint of convertible **scuttle shake** .

The latest round of negotiations ends in just 4 days - but outcries in each of our countries could **shake** the confidence of negotiators and **scuttle**

# Appendix: Manual validation

## discrepancies between BBI and OCD2

**Bartsch 224, ENCOW, Malt dependencies, X2 'argue' (1000 candidates)** [bartsch]

1 / 50 Go << >> [export] back to main page

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1535292 | argue | convincingly | 1 | 740479.262 | --- | OCD | | TP | V ADV | Set |
| 1536640 | argue | forcefully | 2 | 131422.442 | --- | OCD | | TP | V ADV | Set |
| 1539055 | argue | plausibly | 3 | 108643.811 | BBI | OCD | | TP | V ADV | Set |
| 1534717 | argue | case | 4 | 99395.432 | --- | OCD | (under collocate) | TP | directional: collocate entry | Set |
| 1539037 | argue | plaintiff | 5 | 94527.407 | --- | OCD | (under collocate) | TP | directional: collocate entry | Set |
| 1540747 | argue | strongly | 6 | 82931.817 | --- | OCD | | TP | | Set |
| 1535589 | argue | defendant | 7 | 79217.847 | --- | --- | | FP | argument structure constraint on semantic field of subj | Set |
| 1534170 | argue | author | 8 | 78598.802 | --- | OCD | (under collocate) | TP | directional: collocate entry | Set |
| 1538856 | argue | passionately | 9 | 72731.072 | BBI | OCD | | TP | | Set |
| 1536059 | argue | economist | 10 | 52627.136 | --- | --- | | FP | argument structure constraint on semantic field of subj | Set |
| 1540725 | argue | strenuously | 11 | 52533.752 | BBI | OCD | | TP | V ADV | Set |
| 989811 | argue | also | 12 | 50528.000 | --- | --- | | FP | | Set |

variable ordering: A..B 74.5% / B..A 25.5% non-contiguous: 79.2% adjacent

Although no one has argued that amendments other than the Tenth and Fourteenth have impliedly amended the AC , one might **plausibly argue** that each has necessarily done so .

It cannot **plausibly** be **argued** , in my opinion , that the Human Rights Act erodes the sovereignty of Parliament or amounts to a usurpation of power by the judges .

Rather than viewing them as epistemically-morally-politically pernicious forms of hasty generalization by contrast , say , with Gadamerian pre-judgements or putatively more benign practices of categorization , Fricker **argues plausibly** for a " neutral " sense of stereotype which catches their frequent reliability as part of a " hearer 's rational resources " in making credibility judgements .

But , it can be **plausibly argued** that we are getting there .

Given any statement , we can **argue plausibly** that it is about Maine .

# References


- Aston, Guy and Burnard, Lou (1998). *The BNC Handbook.* Edinburgh University Press, Edinburgh. [http://www.natcorp.ox.ac.uk/]
- Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano; Zanchetta, Eros (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, **43**(3), 209–226.
- Bartsch, Sabine (2004). *Structural and Functional Properties of Collocations in English*. Narr, Tübingen.
- Bartsch, Sabine and Evert, Stefan (2013). Exploring the Firthian notion of collocation. In: Abstract book of Corpus Linguistics 2013, Lancaster, UK.
- BBI = Benson, Morton; Benson, Evelyn; Ilson, Robert (1986). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam, New York.
- Brants, Thorsten and Franz, Alex (2006). Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia, PA.

# References


- Church, Kenneth; Gale, William A.; Hanks, Patrick; Hindle, Donald (1991). Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum.
- Church, Kenneth W. and Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.
- Daille, Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.
- Dunning, Ted E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.
- Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005.
- Evert, Stefan and Hardie, Andrew (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, UK.

# References


- Evert, Stefan and Krenn, Brigitte (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.
- Evert, Stefan and Krenn, Brigitte (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, **19**(4), 450–466.
- Gries, Stefan Th. (2013). 50-something years of work on collocations: What is or should be next … . *International Journal of Corpus Linguistics*, **18**(1), 137–165.
- Han, Lushan; Kashyap, Abhay L.; Finin, Tim; Mayfield, James; Weese, Johnathan (2013). UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, ACL.
- Johnson, Mark (2001). Trading recall for precision with confidence sets. Unpublished technical report.
- Lin, Yuri; Michel, Jean-Baptiste; Aiden, Erez Lieberman; Orwant, Jon; Brockman, Will; Petrov, Slav (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174.

# References


- Manning, Christopher D. and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- Pedersen, Ted and Bruce, Rebecca (1996). What to infer from a description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.
- Sinclair, John McH. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins (eds.), *In Memory of J. R. Firth*, pages 410–430. Longmans, London.
- Uhrig, Peter and Proisl, Thomas (2012). Less hay, more needles - using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates. *Lexicographica*, **28**(1), 141–180.
- OCD2 = McIntosh, Colin; Francis, Ben; Poole, Richard (eds.) (2009). *Oxford Collocations Dictionary for students of English*. Oxford University Press.
- Schäfer, Roland and Bildhauer, Felix (2012). Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC '12), pages 486–493.