FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

# (Association) measure for measure: Evaluating the corpus-based identification of lexical collocations

**Stefan Evert**
FAU Erlangen-Nürnberg

— joint work with Sabine Bartsch, Thomas Proisl, Peter Uhrig —

# Lexical collocations

- Lexical collocation = "habitual" recurrent word combination (Firth 1957, Sinclair 1966)
  - complex phenomenon without clear definition
  - but supported by strong native-speaker intuitions
- Some examples
  - cow–milk, day–night (Firth 1957)
  - blonde hair, catchy tune, commit a crime, spill the beans
  - make/thwart an attempt, feeble/successful attempt, …
- Important applications in lexicography, language teaching, digital humanities & social sciences, NLP, …

# Related phenomena

- Lexicalized multiword expressions (MWE)
  - kick the bucket, eat humble pie, spill the beans, …
  - hand in, carry on, turn down, give a speech, …
  - bucket seat, black hole, New York City, …
- Psychological association
  - apple → red, orange, fruit, pie, pear, worm, …
  - dolphin → fish, sea, fin, mammal, whale, blue, …
- Co-occurrence in text corpora

👉 Overlap with lexical collocations & each other

# Collocation identification

- Automatic identification of collocations and MWE based on large electronic corpora
- Main cue: co-occurrence frequency, quantified by various statistical association measures (AM)
- 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)
- AM are most appropriate cue for lexical collocations

# Collocation identification

- Best techniques & parameters for collocation/MWE identification are still unclear
  - e.g. choice of more than 57 AMs (Pecina 2005)
- Need for empirical comparative evaluation
- Prior work: (subtype of) MWE with relatively clear definition → manual validation of MWE candidates

- Today: results of large-scale evaluation study on the automatic identification of English lexical collocations
  - special focus: linguistic usefulness of Web corpora

# Research questions

- Which AM correlates best with lexical collocations?
- What is an appropriate window of co-occurrence?
- Which source corpora provide the best results?
- Does size matter? Or representativeness?
- Is automatic annotation with NLP tools useful?
- Are there interactions between these parameters?
- Are crawled Web corpora and n-gram databases a viable substitute for expensive reference corpora?
- long-term goal: better understanding of collocations

# Gold standard

- Standard evaluation methodology (Evert & Krenn 2001, 2005)
  - obtain ranked list of candidates from automatic system
  - manual validation by expert → TPs, FPs
  - compute & compare quantitative evaluation scores such as precision (P), recall (R) and F-score (F) for n-best lists
- Difficult to apply in this study
  - not suitable for comprehensive large-scale evaluation
  - lack of clear annotation criteria → pro-corpus bias

# Gold standard

- Alternative: use existing gold standard of "known" lexical collocations → collocation dictionary
  - but see critical remarks by Evert (2004) 😉
- 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

# 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 (node word listed as collocate)
  - set of 18,545 node-collocate pairs

# Gold standard example

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

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)

# Parameters: association measure

- log-likelihood ($G^2$, Dunning 1993)
- Mutual Information (MI, Church & Hanks 1990)
- t-score (t, Church et al. 1991)
- $MI^2$ (Daille 1994)
- chi-squared ($X^2$) with Yates's correction
- co-occurrence frequency (f)
- $\Delta P$ (Gries 2013, a directional measure)
- Dice coefficient (Dice, SketchEngine)

(selection based on recommendations of Evert 2008)

# 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 1996) | 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 | $G^2$ | TP? |
|---|---|---|---|
| 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 (2013), following Evert & Krenn (2001, 2005)
  - task: determine most salient collocational pairs
    - ☞ we present results for this approach

# Evaluation: global ranking

| node | collocate | $G^2$ | TP? |
|---|---|---|---|
| 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 | ++ |

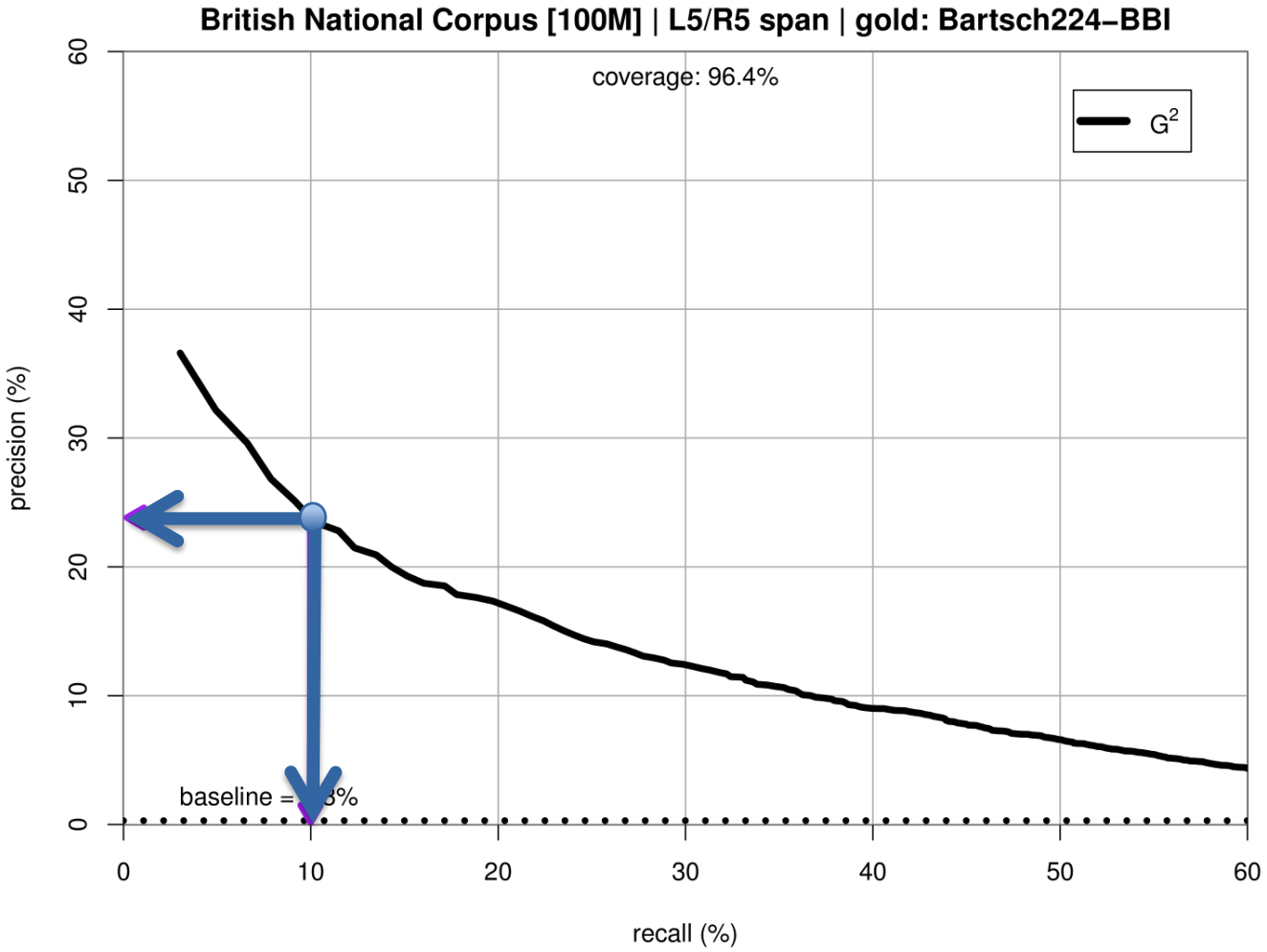
most strongly associated node-collocate pairs (n = 10)

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

*R* = 5 / 2845 TP = 0.2%

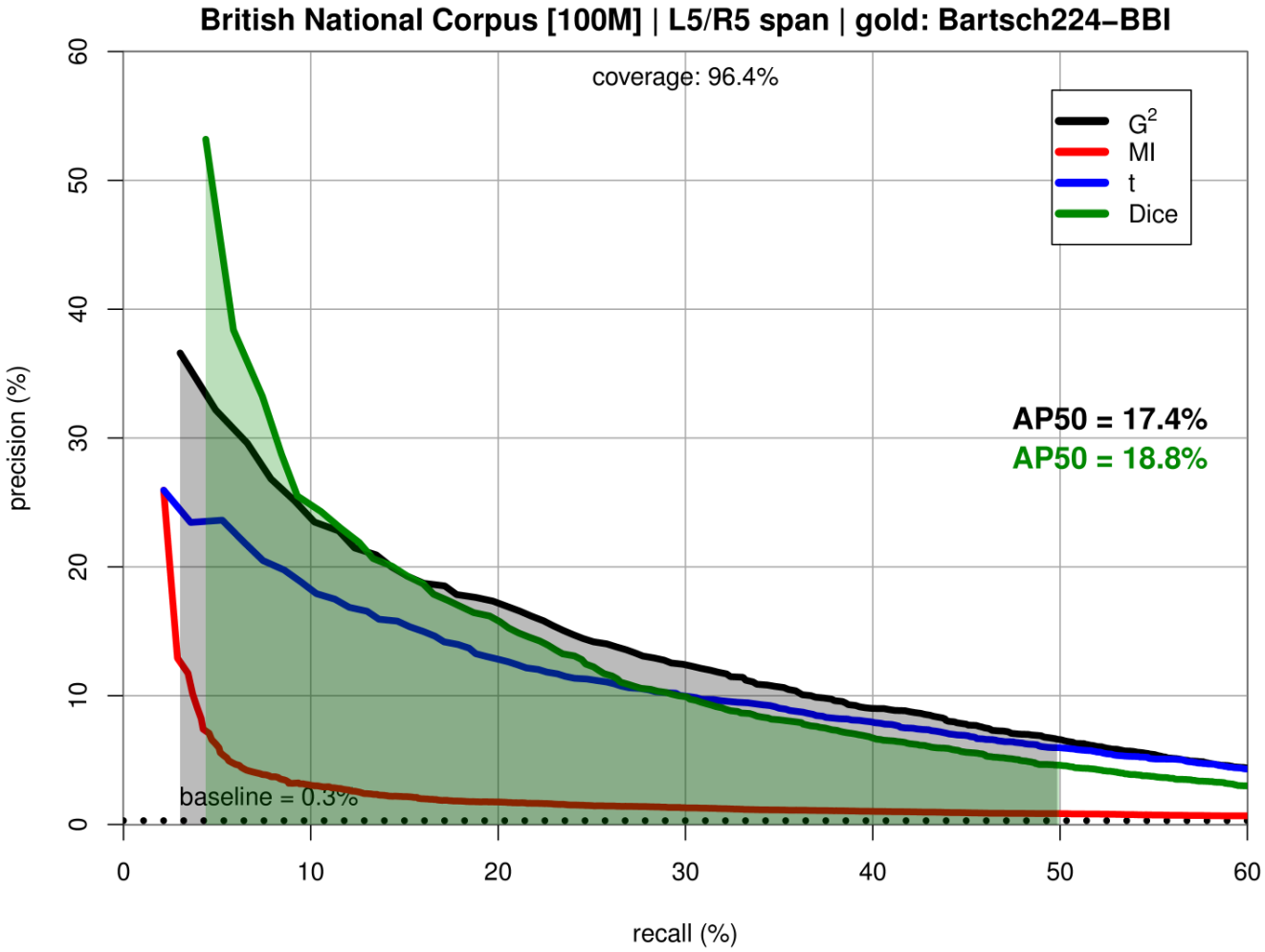
TP = true positive (according to BBI dictionary)

# Evaluation: precision vs. recall | BBI



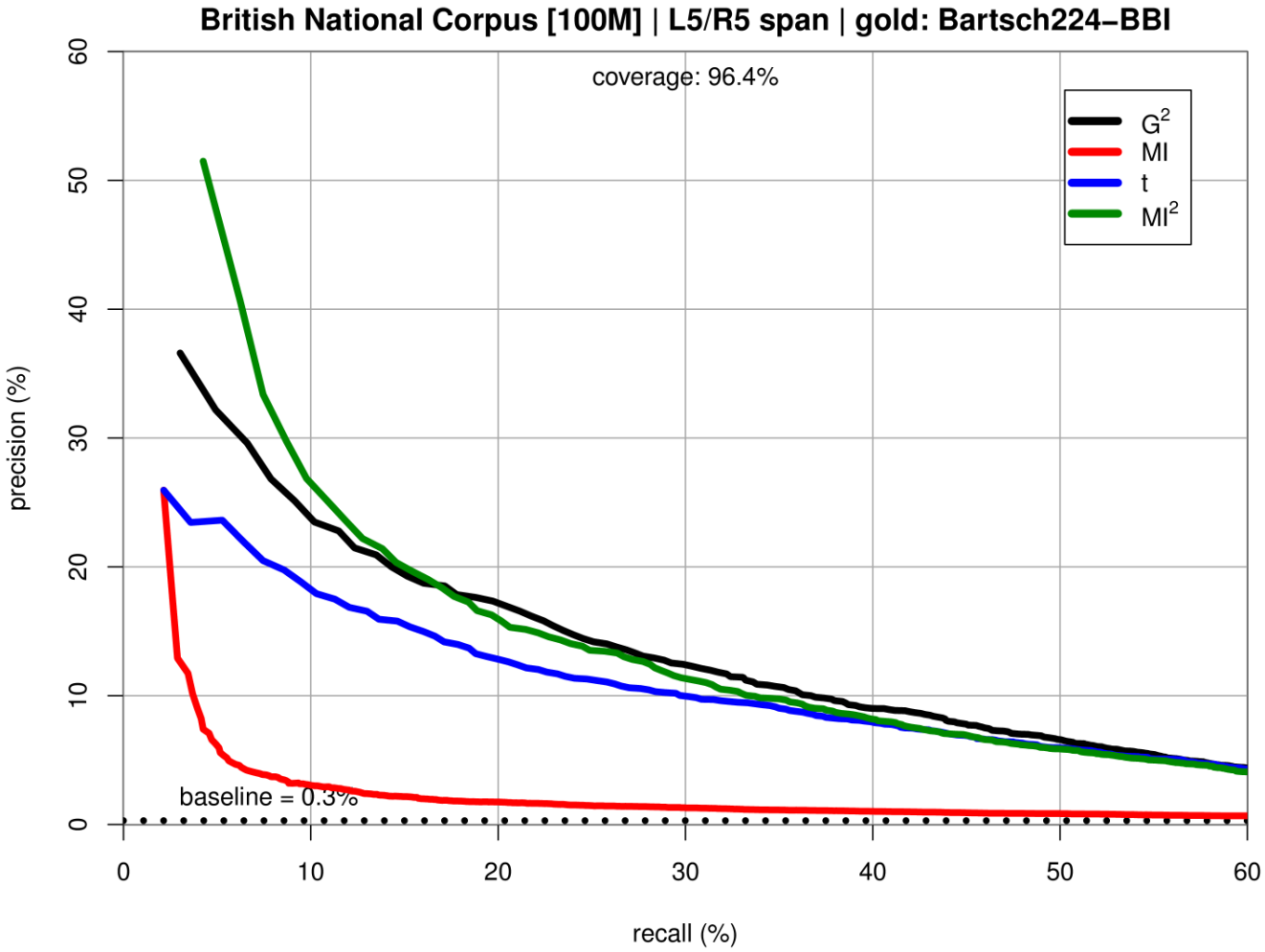

n = 1193
*R* = 10.0%
*P* = 23.9%

# Evaluation: precision vs. recall | BBI




Average Precision
up to 50% recall

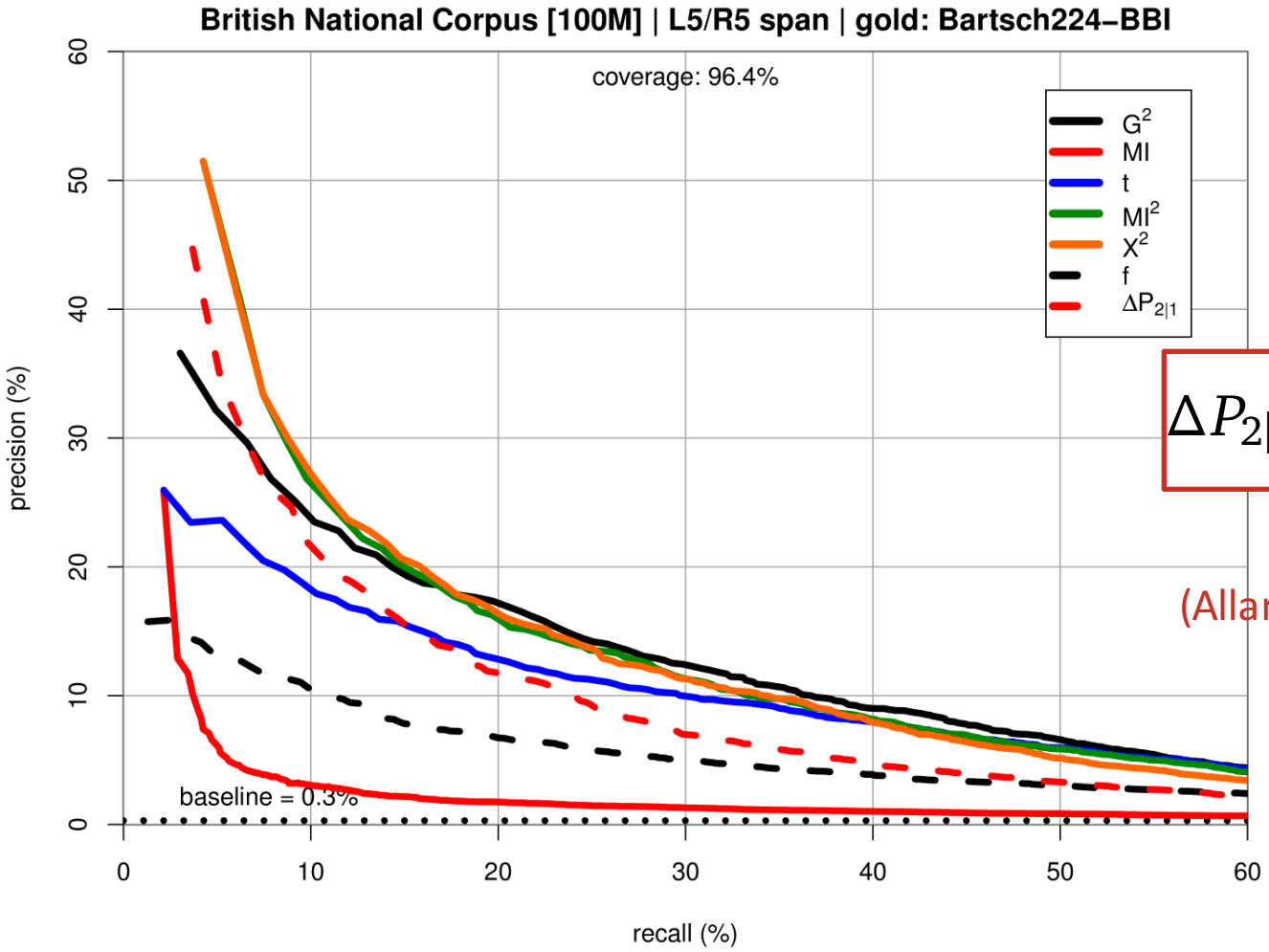
# Evaluation: precision vs. recall | BBI




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

heuristic measure
(Daille 1994)

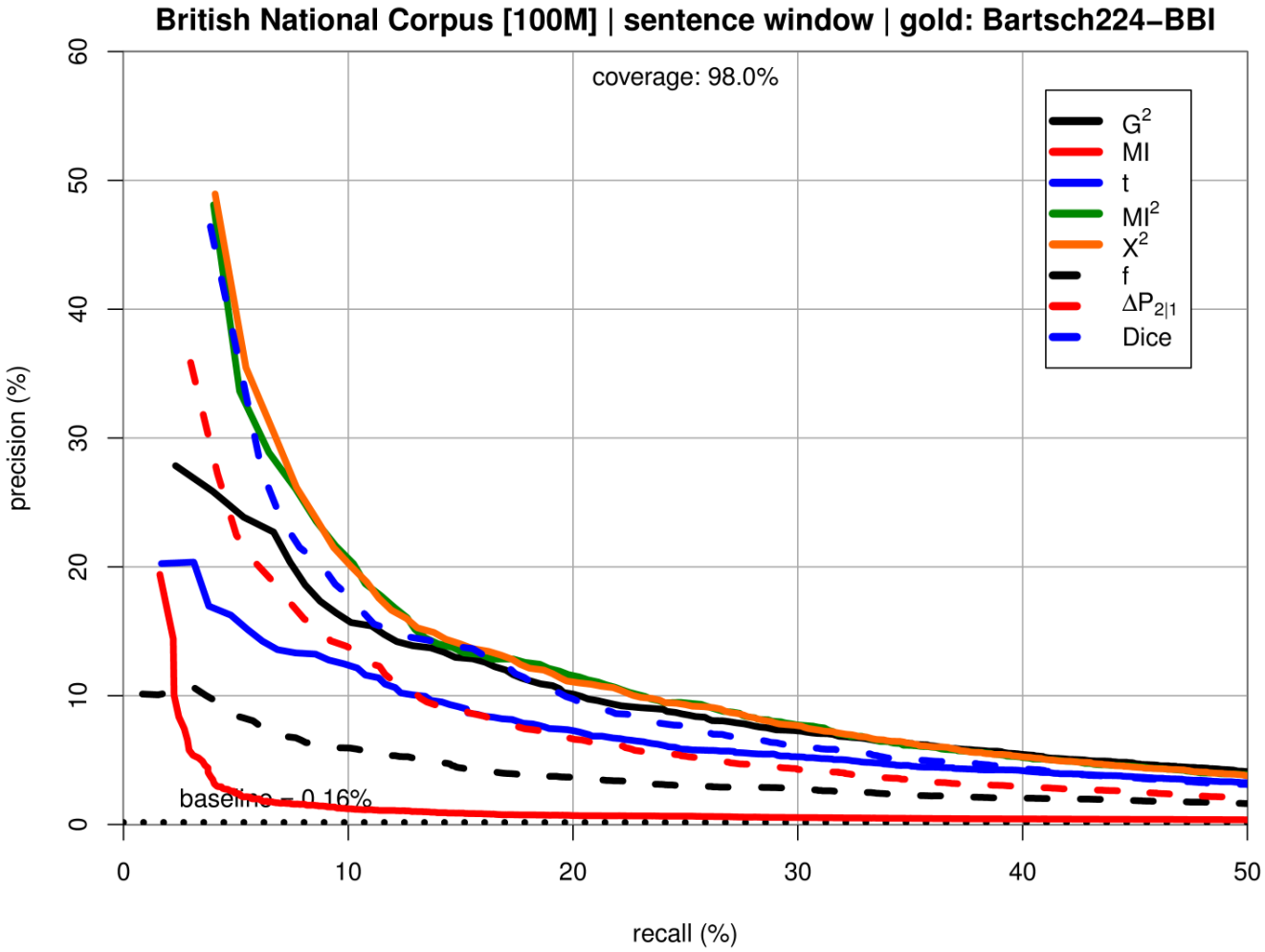
# Evaluation: precision vs. recall | BBI



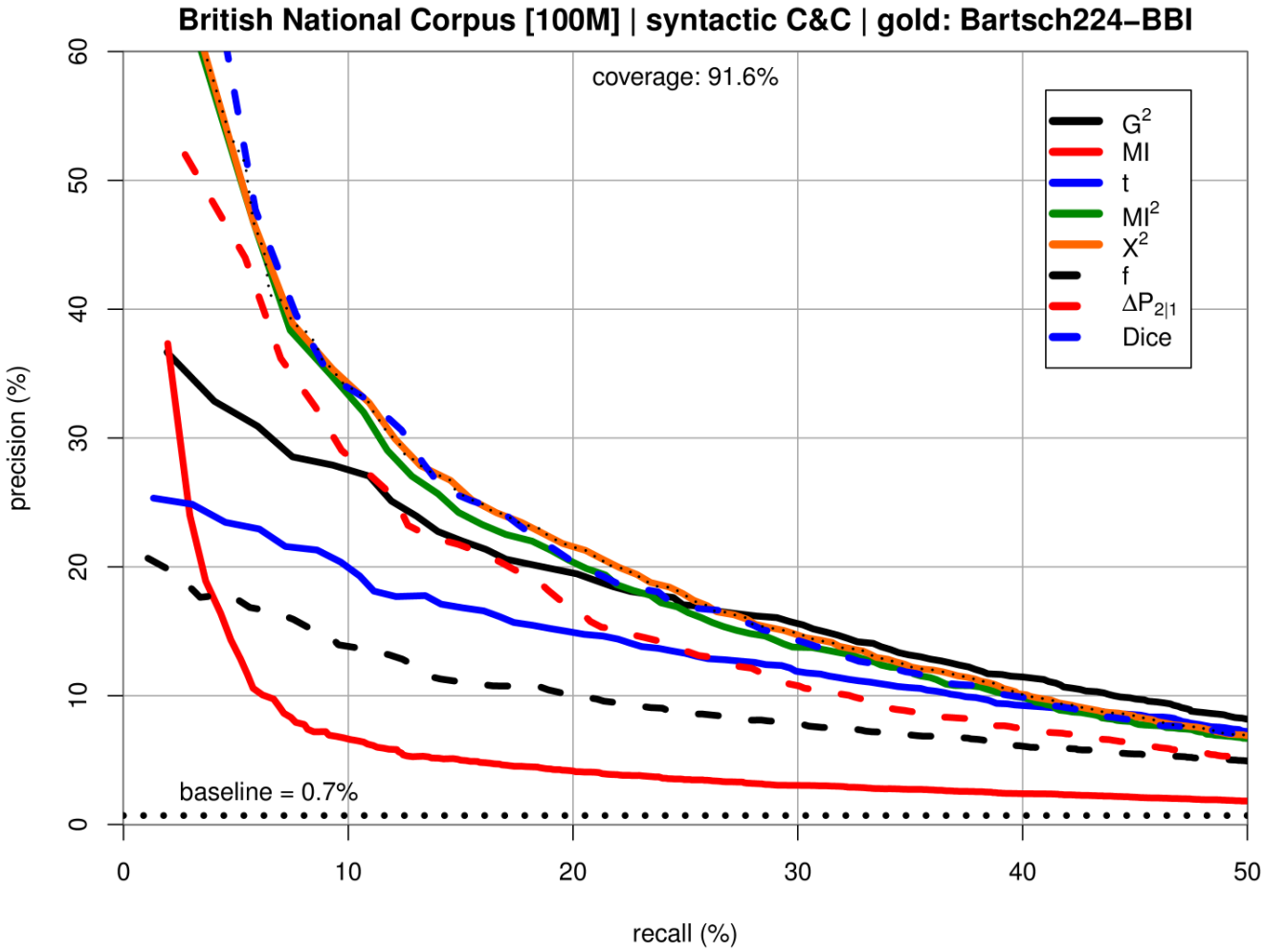

$$\Delta P_{2|1} = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$$

asymmetric ΔP
(Allan 1980; Gries 2013)

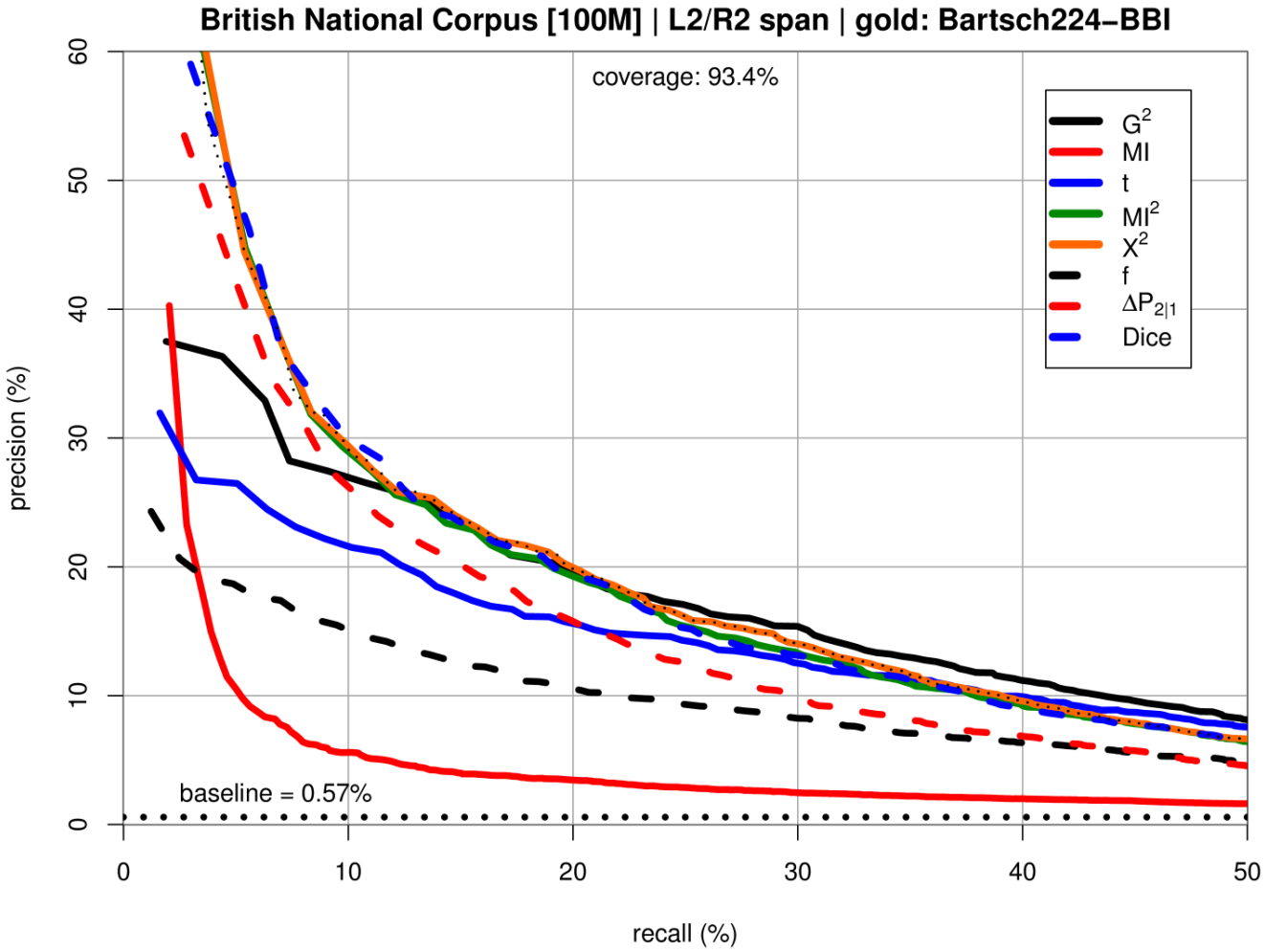
# Factor: context size | BBI



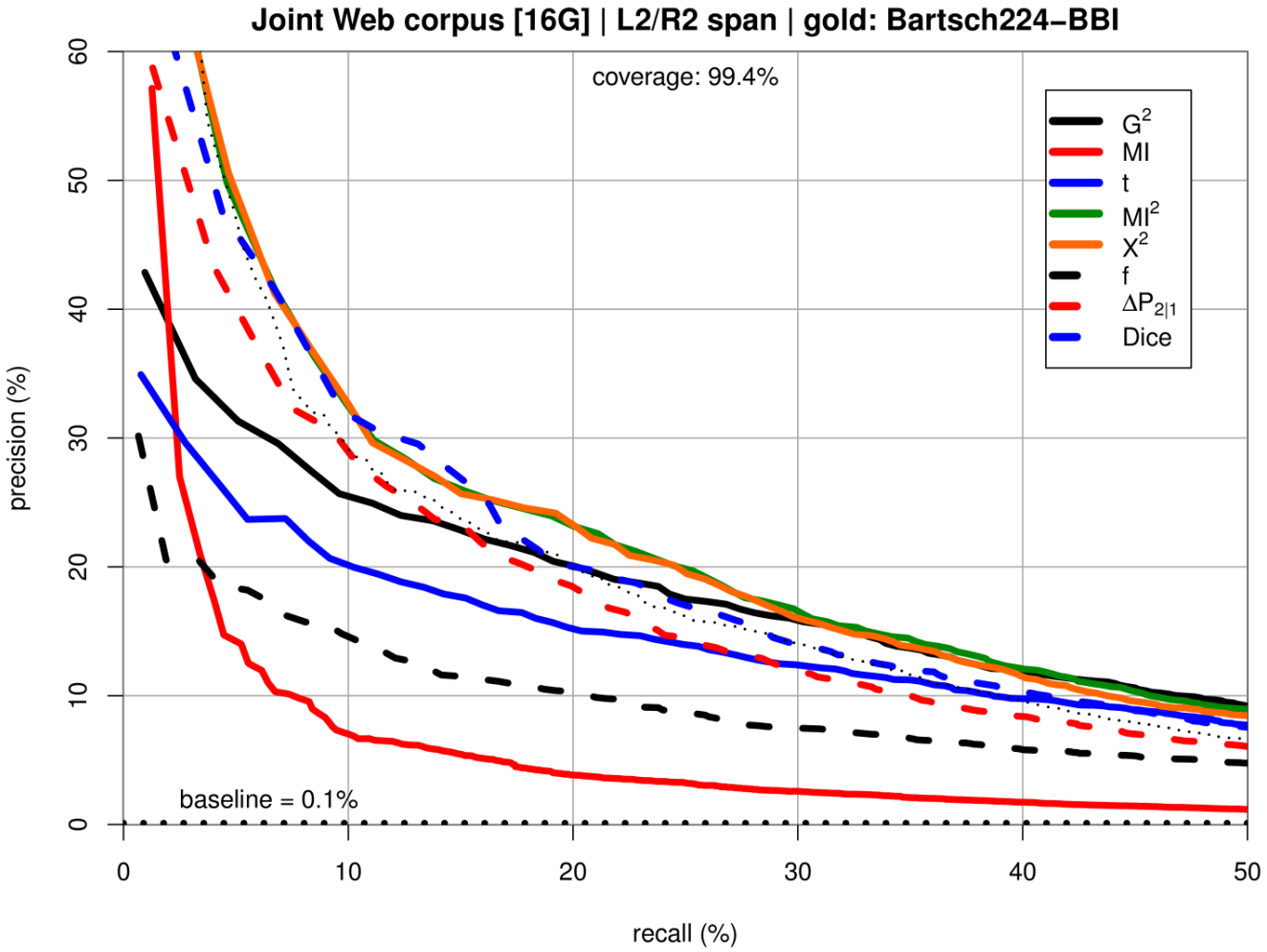
# Factor: context size | BBI



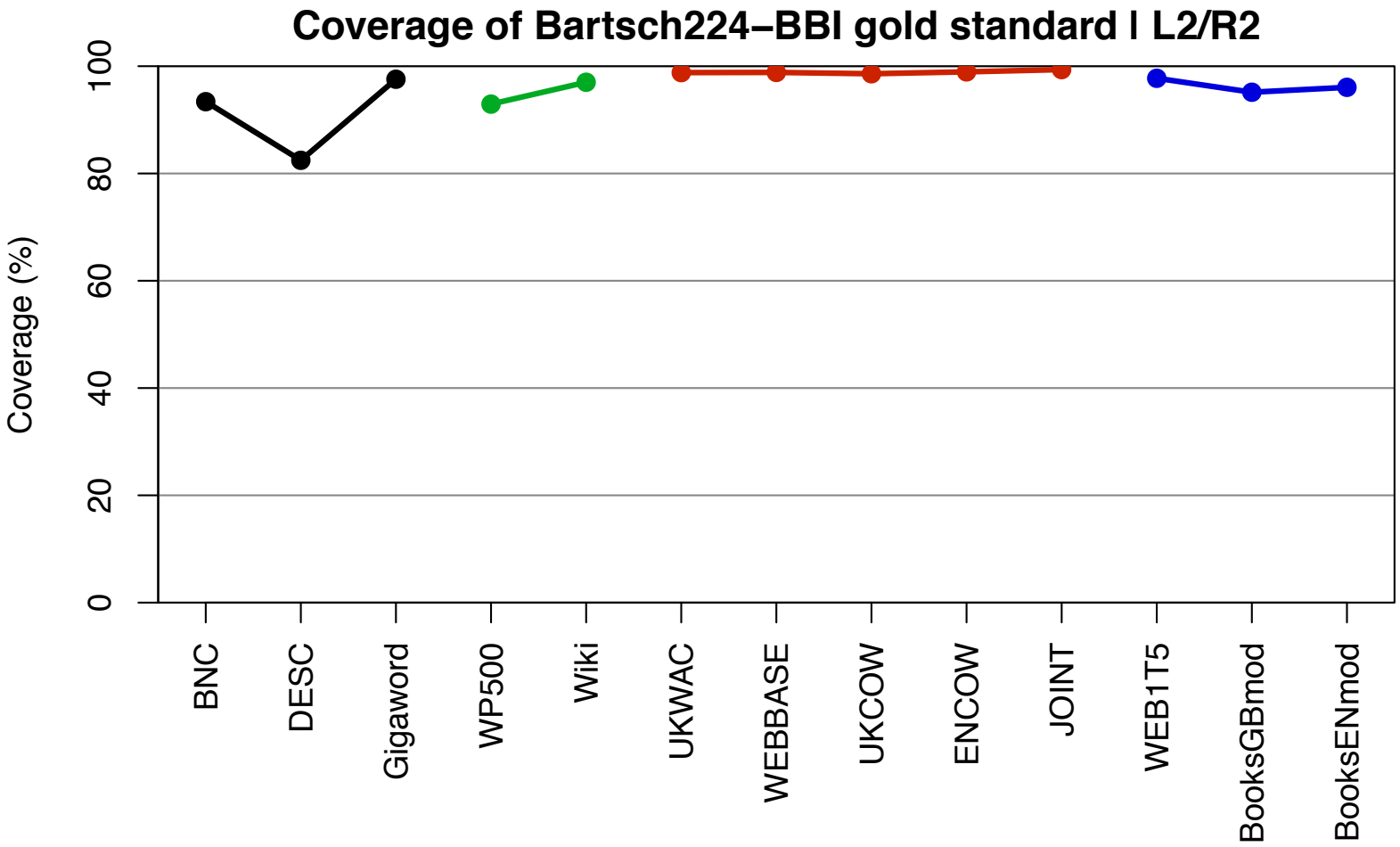
# Factor: different corpora | L2 / R2 | BBI



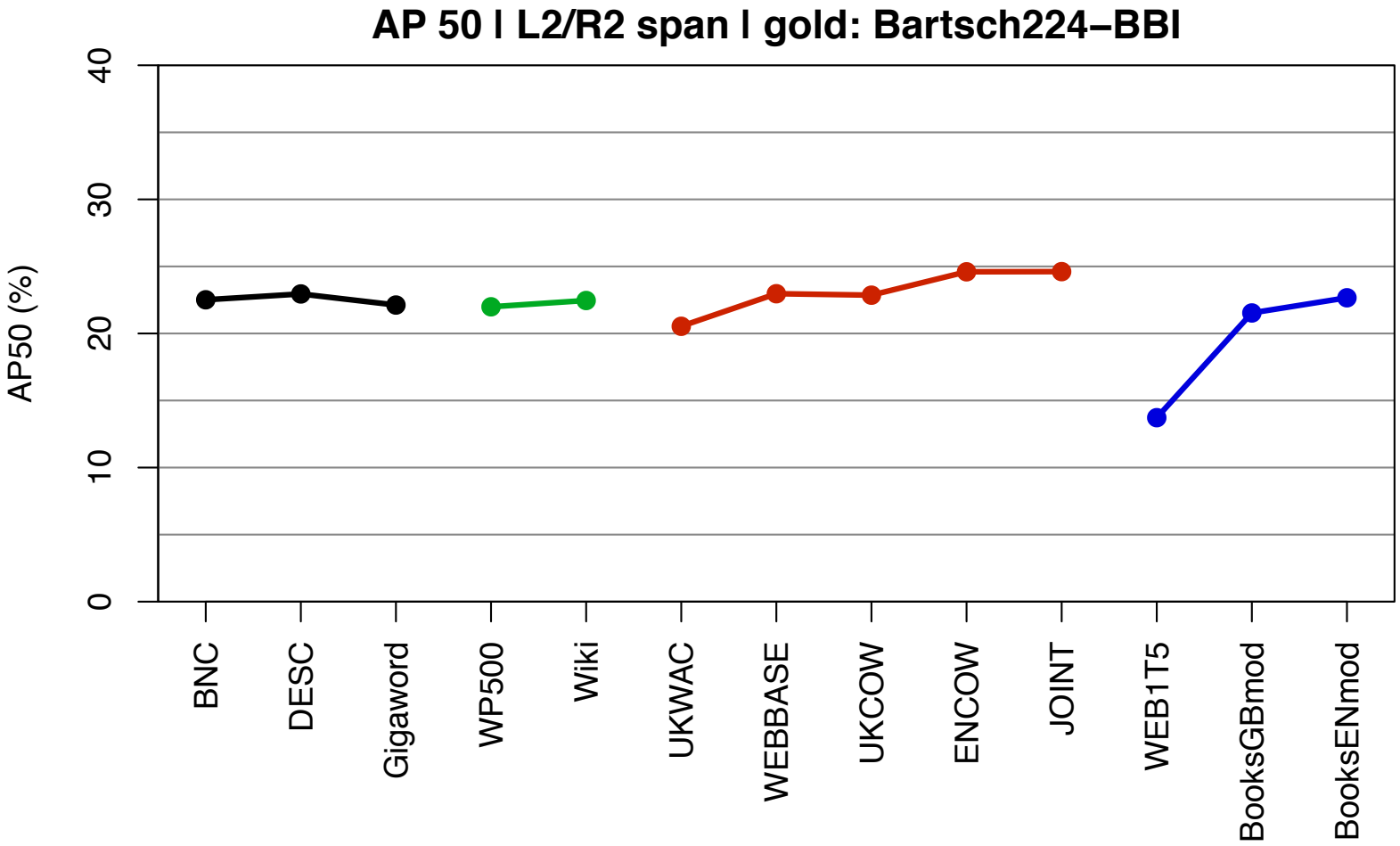
# Factor: different corpora | L2 / R2 | BBI



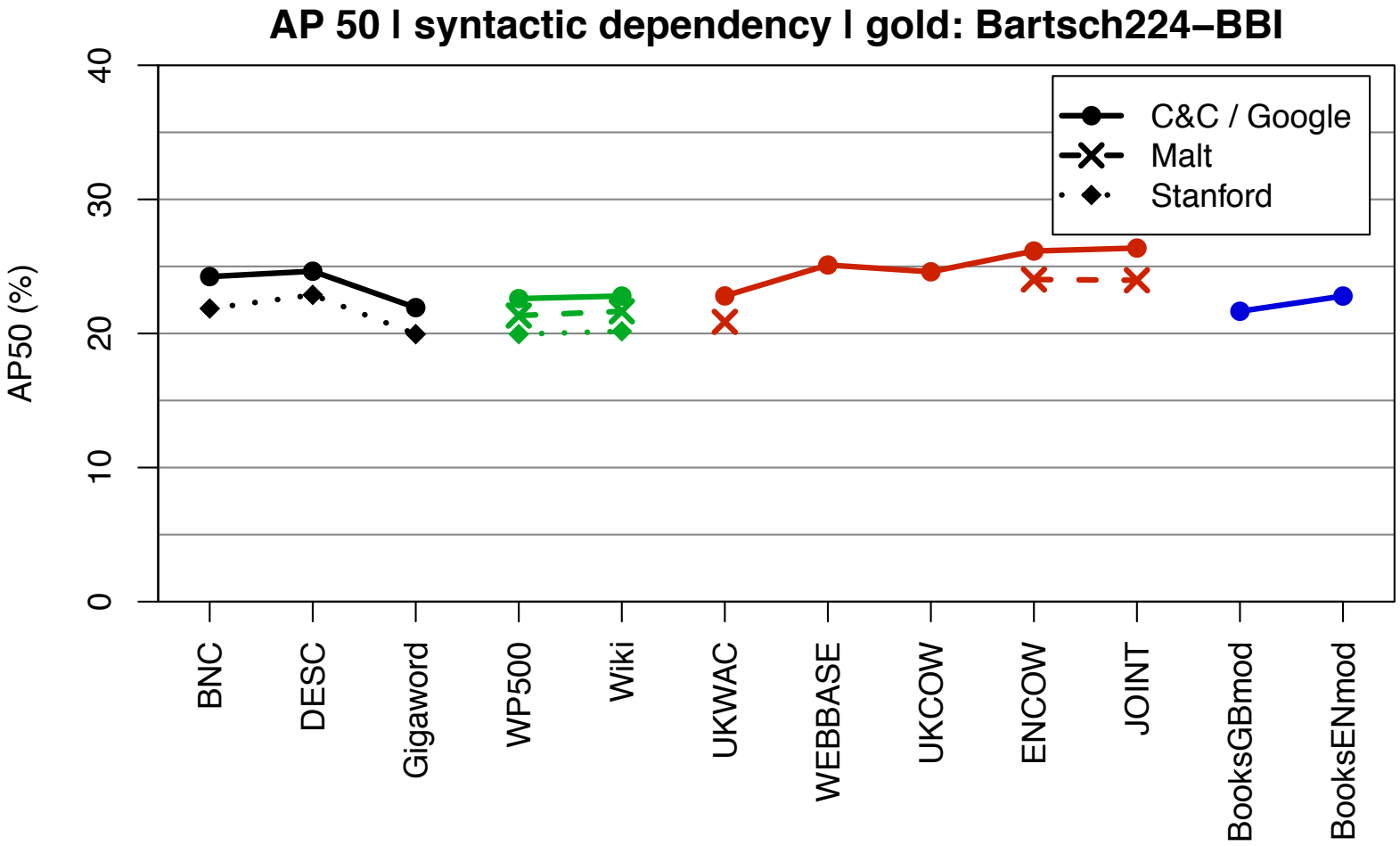
# Result overview | BBI



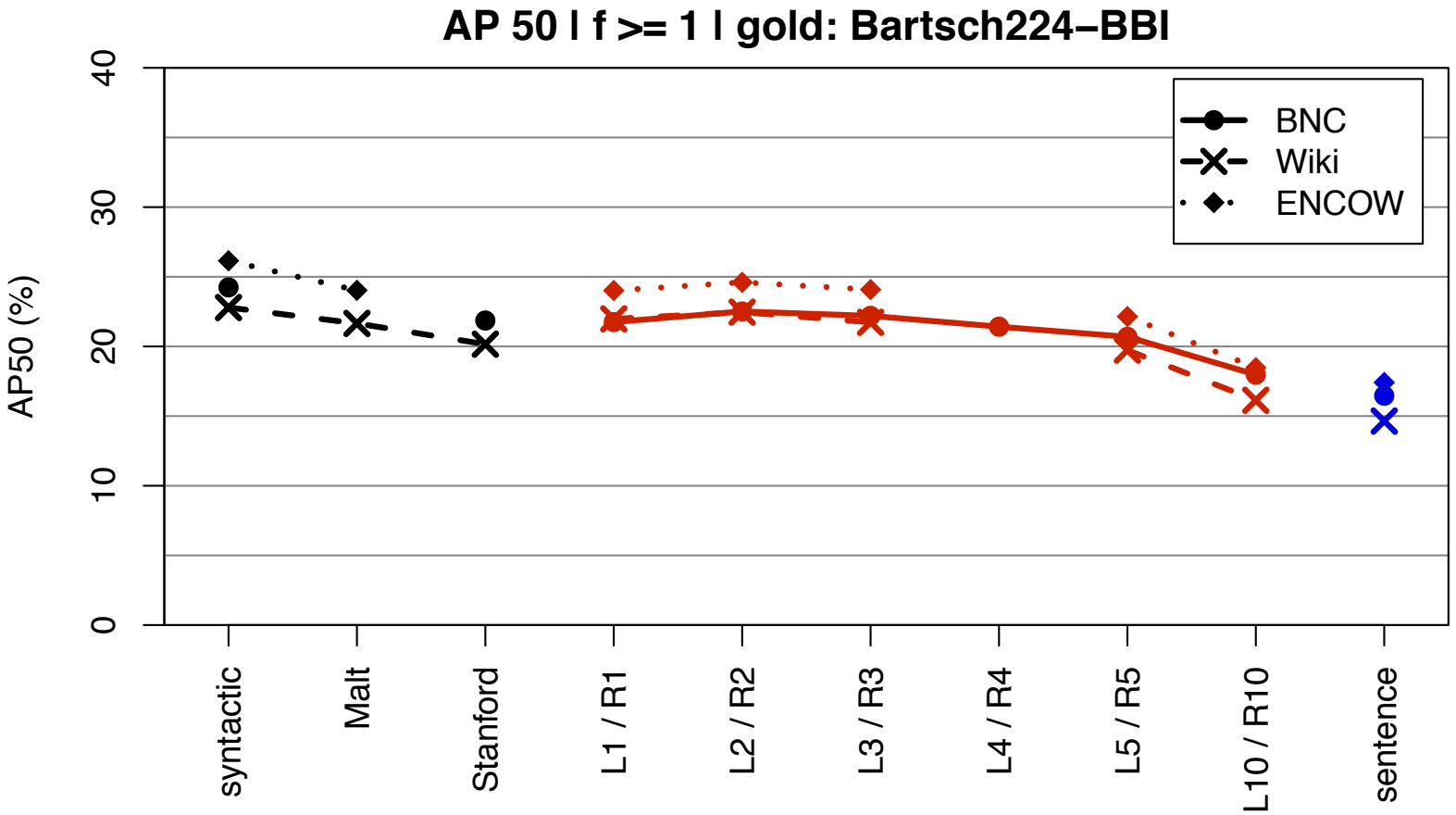
# Result overview | BBI



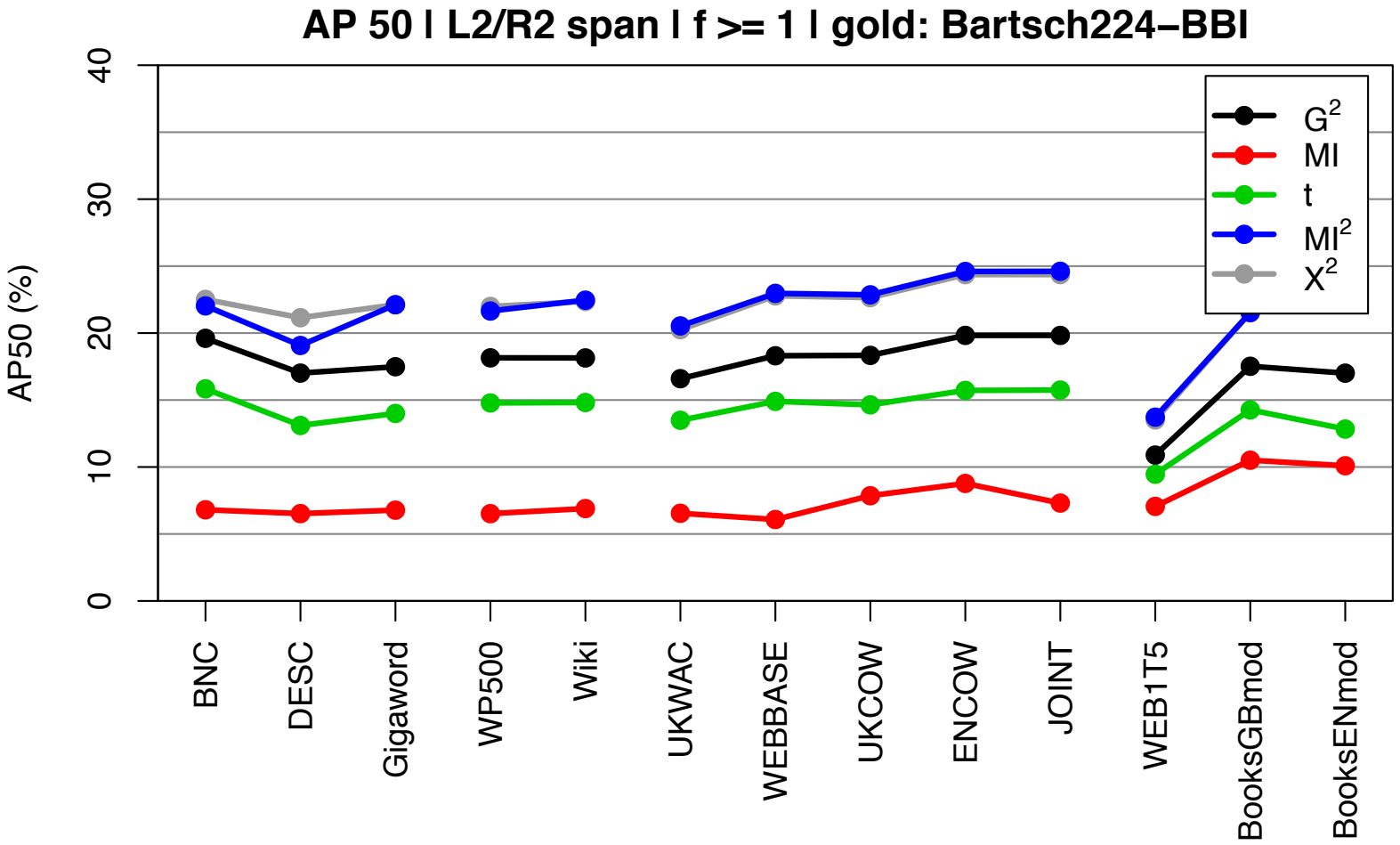
# Result overview | BBI



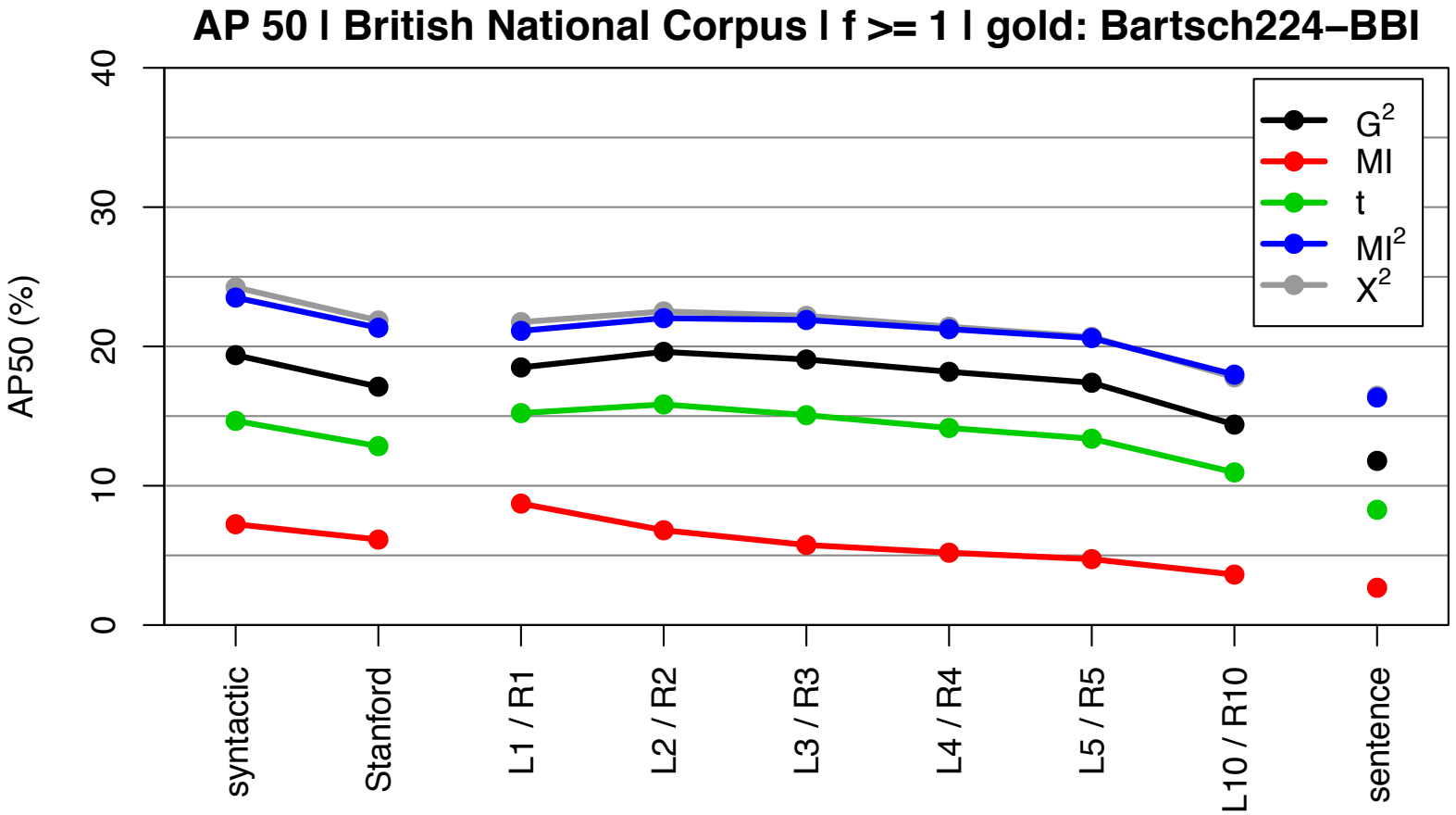
# Result overview | BBI



# Result overview | BBI



# Result overview | BBI



# Conclusions I

- Chi-squared / MI2 seem to be best-suited AMs for identification of (BBI) lexical collocations
- Web corpora are a good replacement for BNC, but they need to be much larger & reasonably clean
- Corpus size matters, but very slow "learning curve"
- Small context windows & syntactic context optimal
- NLP annotation helps, but only with the right parser
- Few interactions: always same ranking of AMs
- But how good and reliable is our "gold standard"?

# BBI vs. OCD2

**rose** *n.* 1. a long-stemmed; rambling; wild ~ 2. a bouquet of ~s

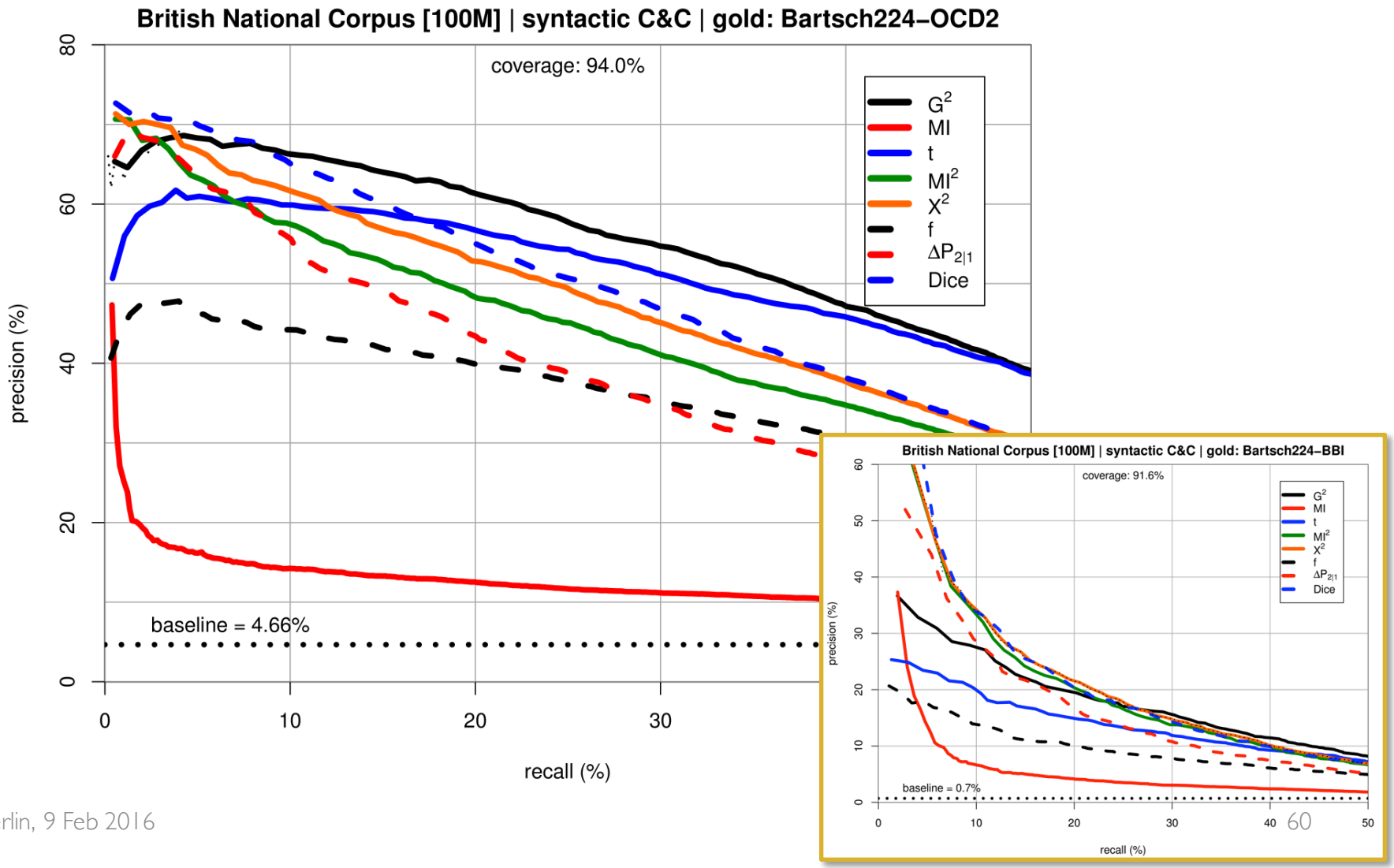
**rose** *noun*

ADJ. **deep pink, red, yellow, etc. | early, late | climbing, rambler, rambling, shrub | long-stemmed | scented | fresh** *She put fresh roses in the vases.* **| wild**

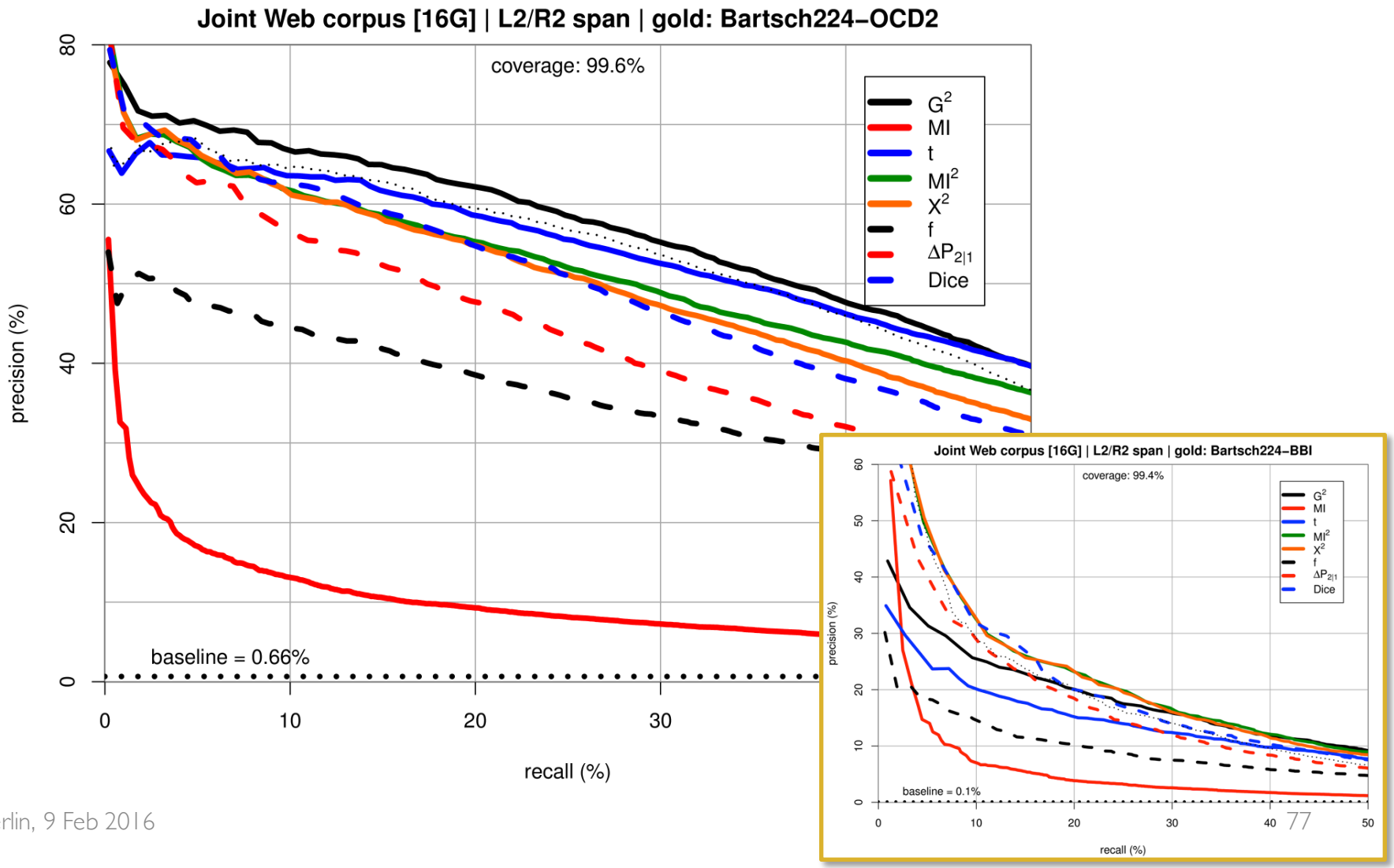
ROSE + NOUN **garden | bed | bush | petals | bowl**

PHRASES **a bed of roses, the scent/smell of a rose** > Note at FLOWER(for verbs)

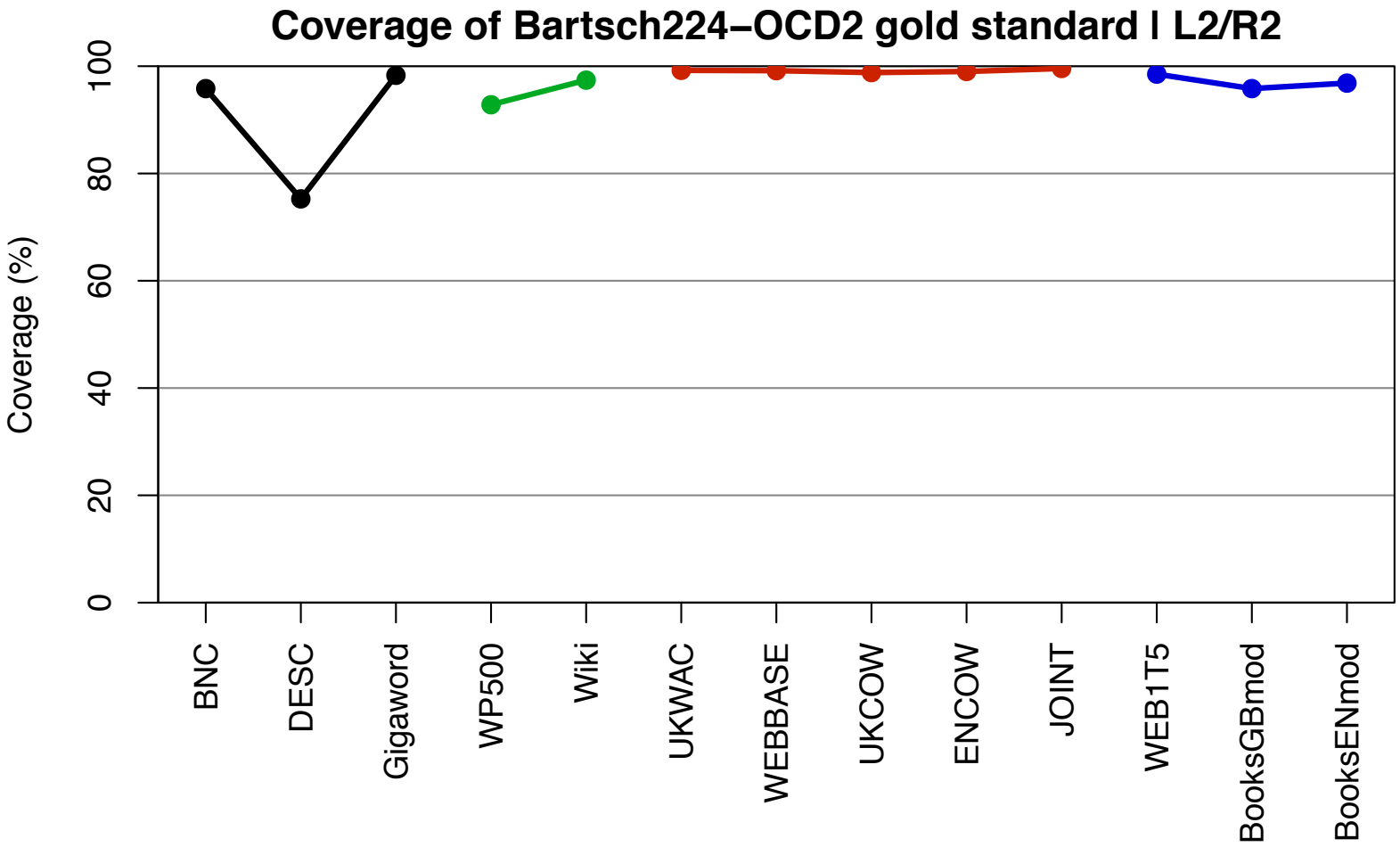
# Replication on OCD2 gold standard



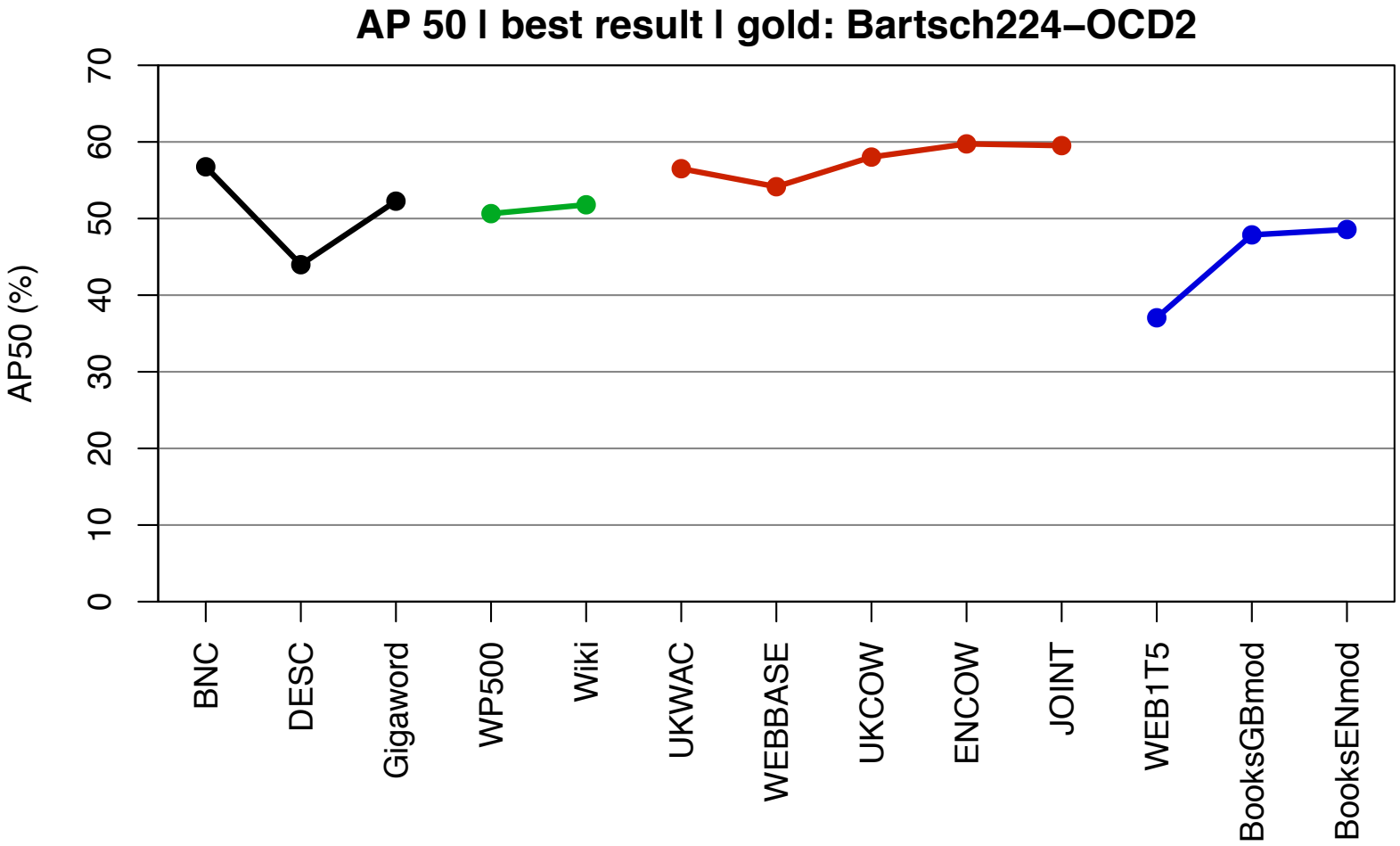
# Replication on OCD2 gold standard



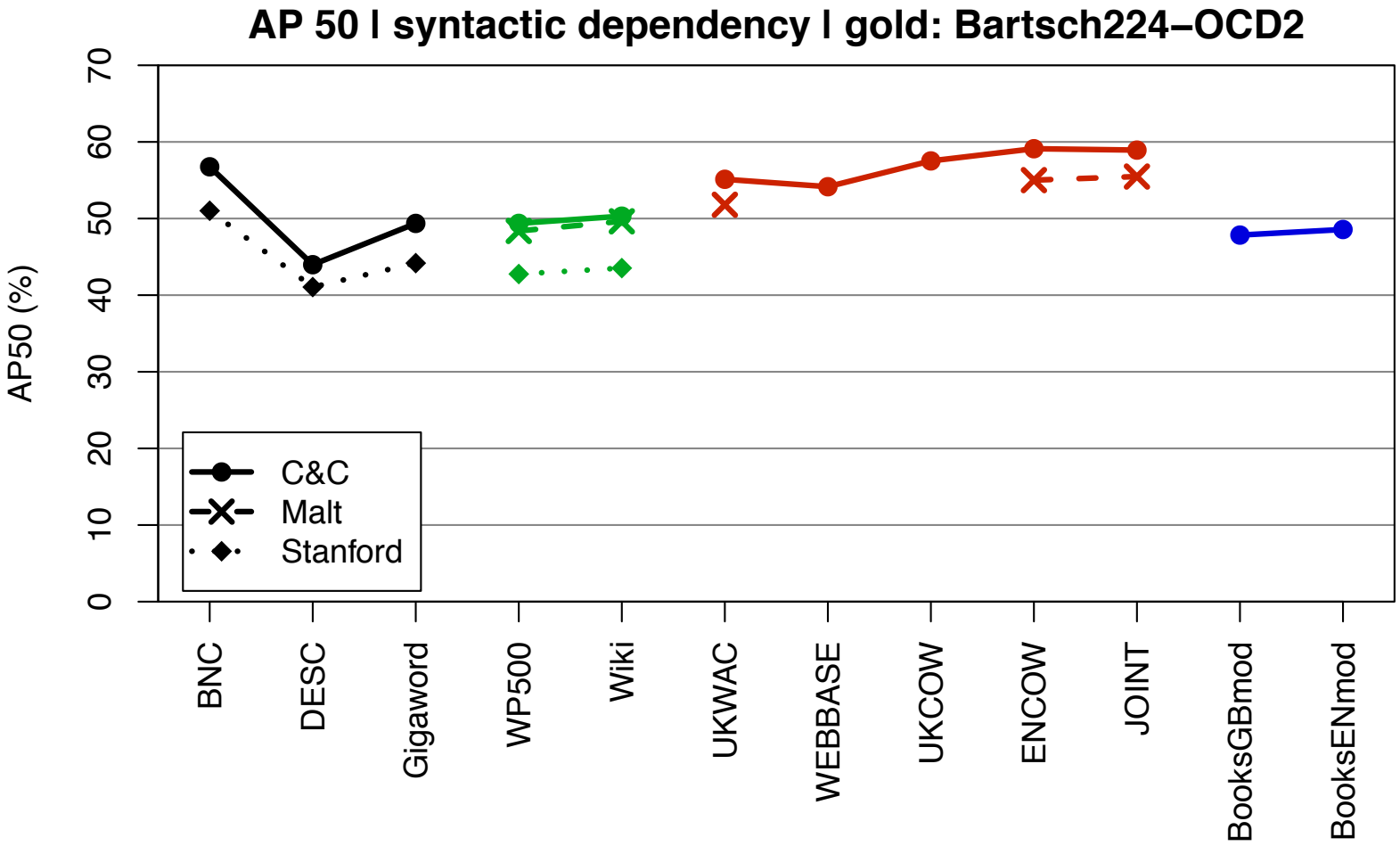
# Result overview | OCD



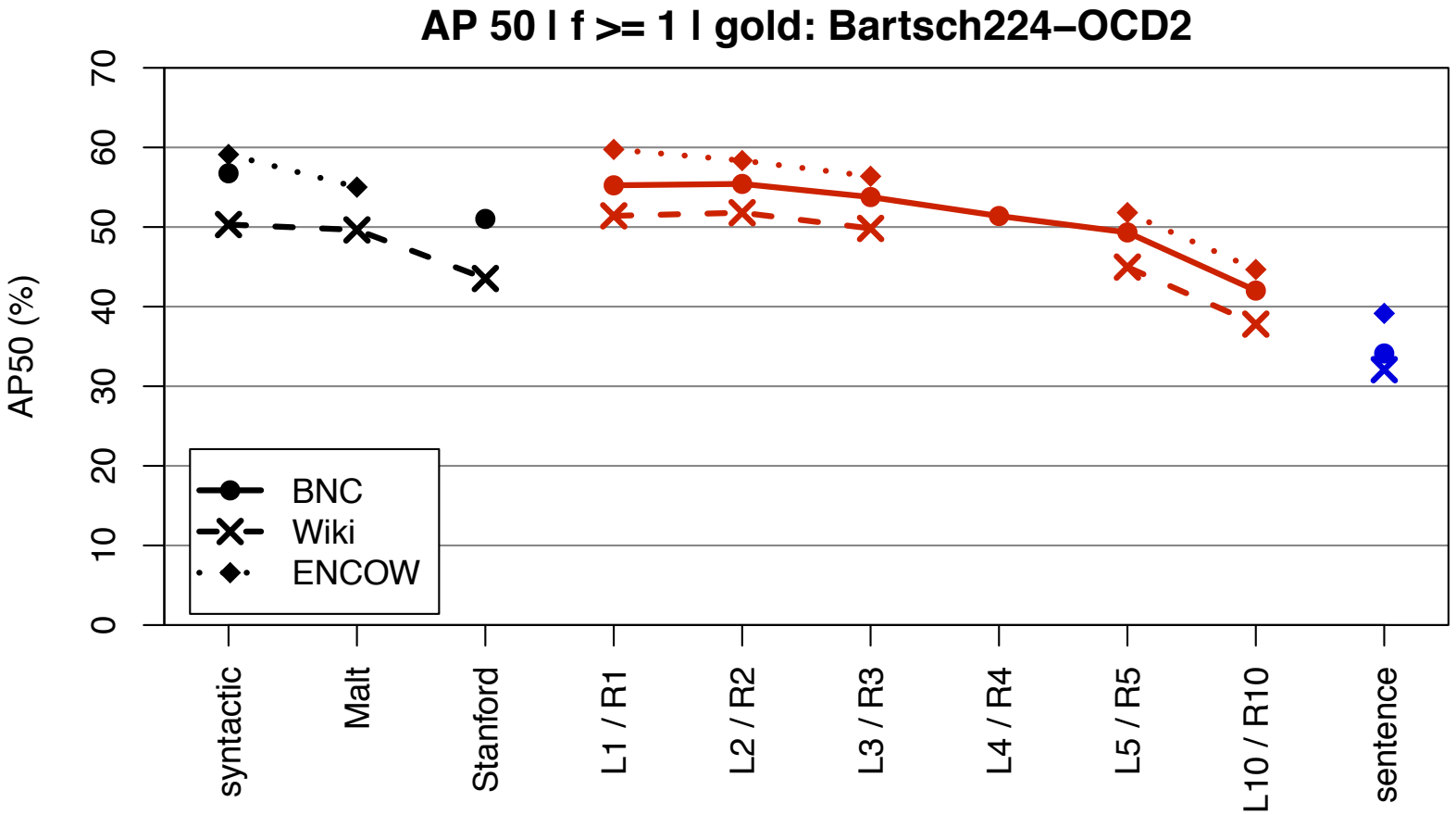
# Result overview | OCD



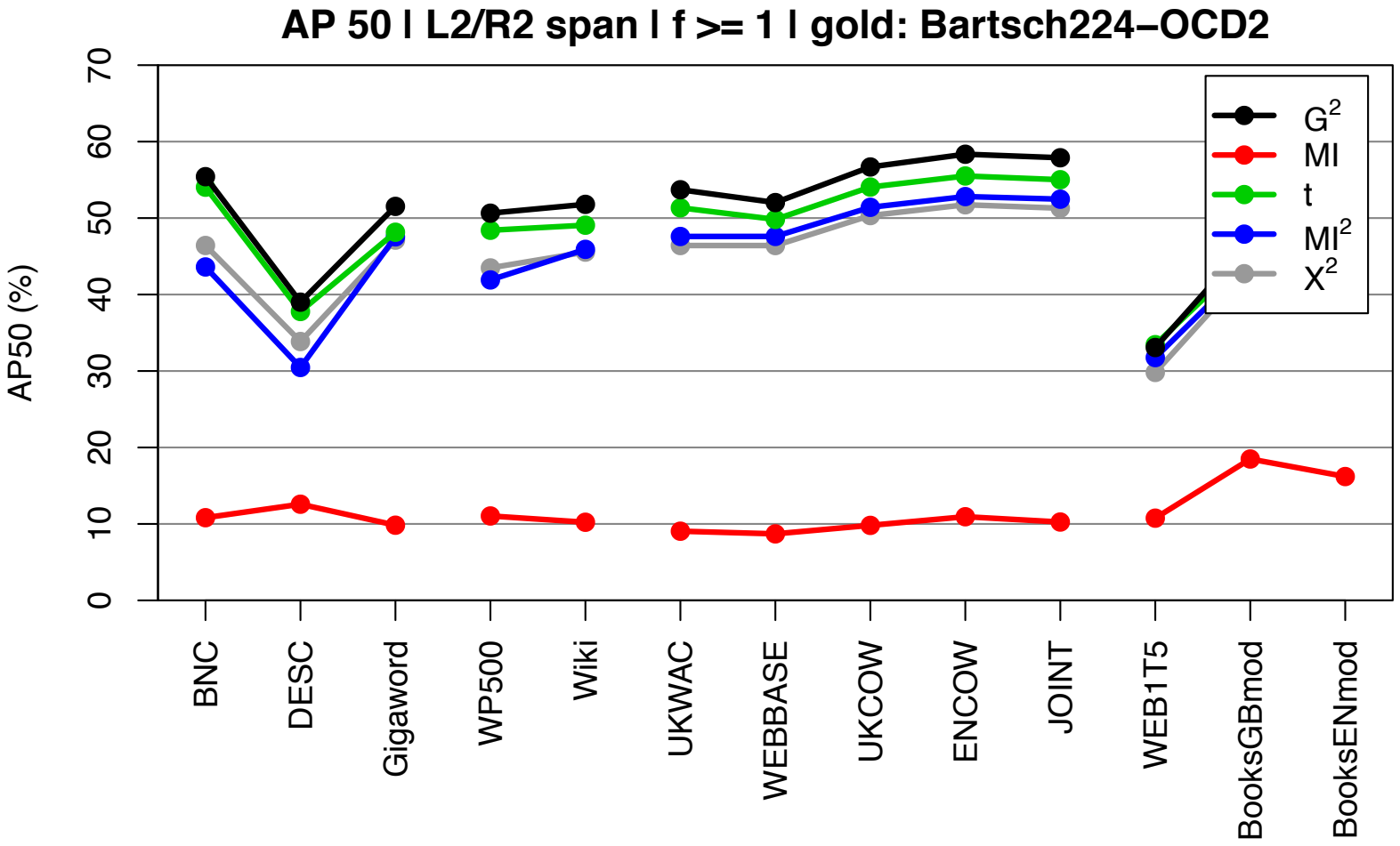
# Result overview | OCD



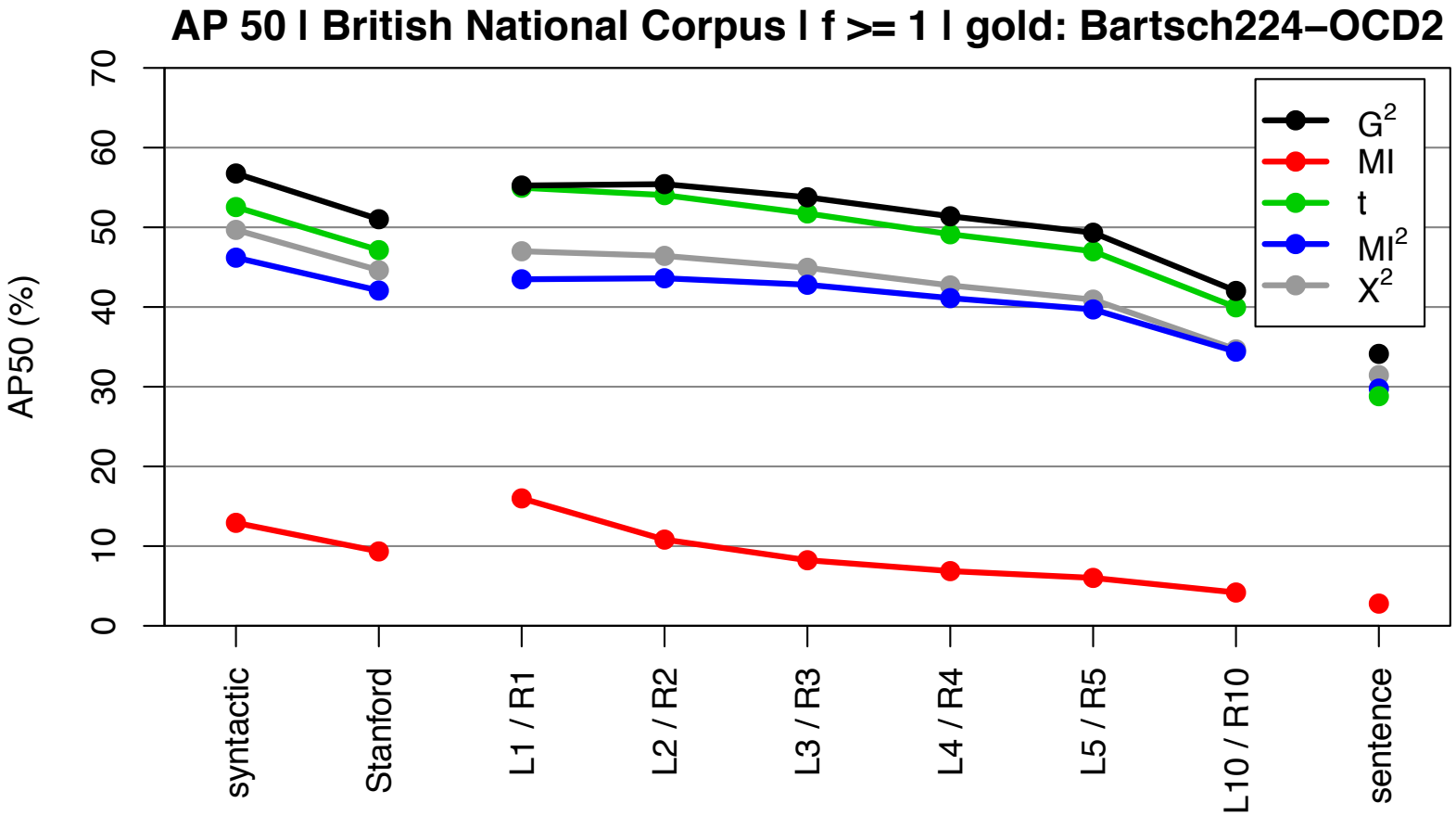
# Result overview | OCD



# Result overview | OCD



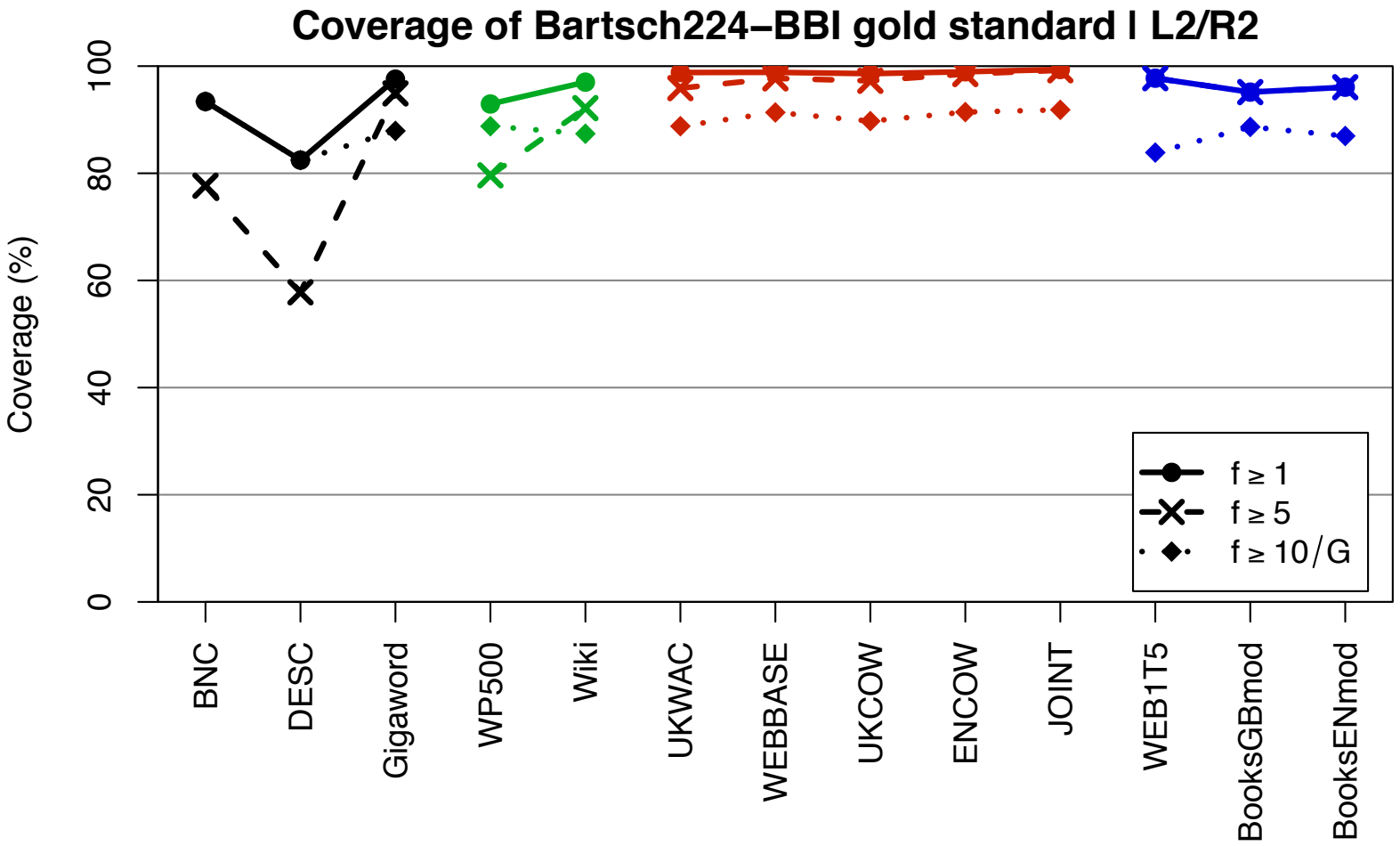
# Result overview | OCD



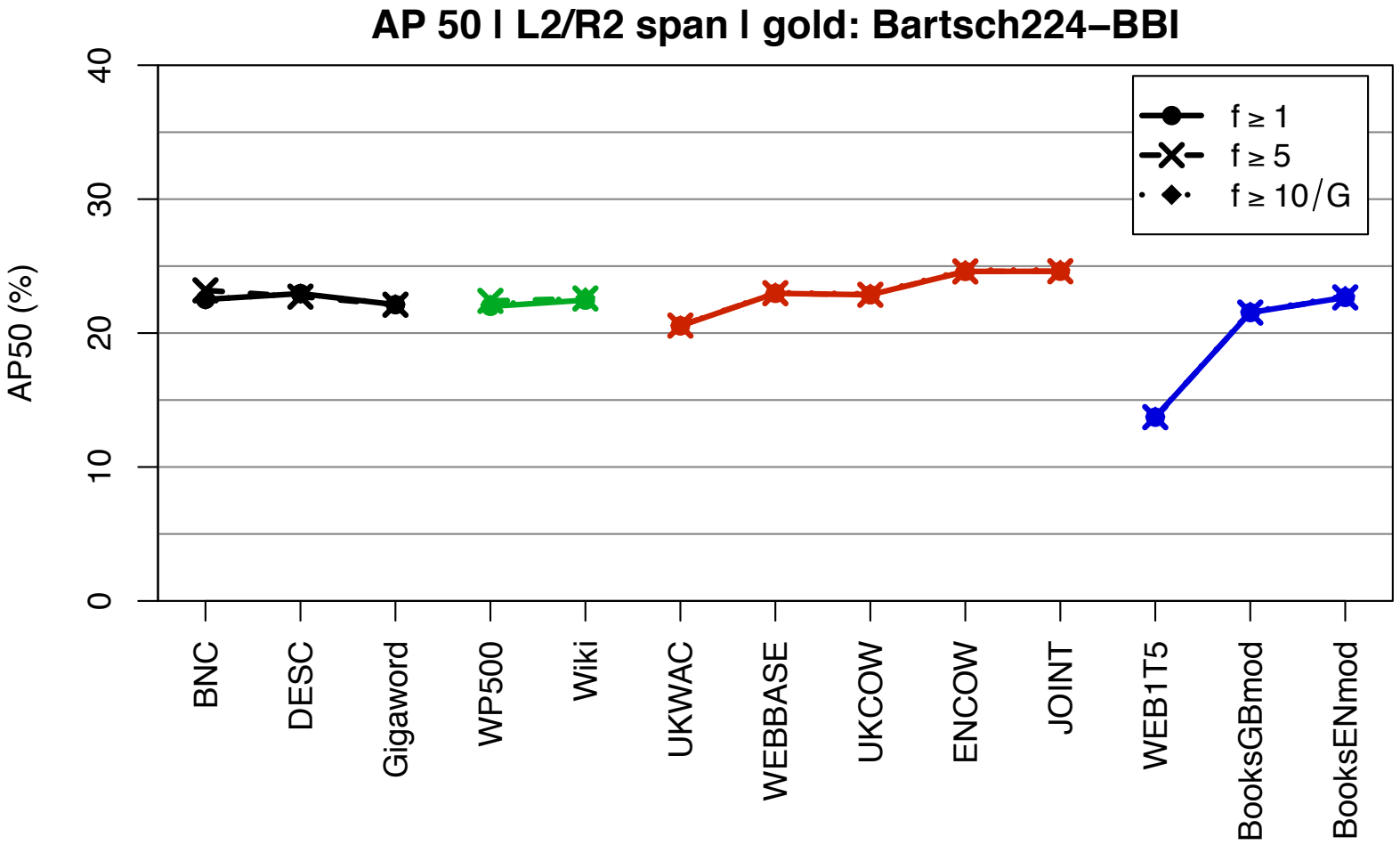
# But what about frequency thresholds?

- Evaluation experiments based on *all* co-occurrences found in corpora, even if not recurrent (f = 1)
- It is common in collocation research to apply a frequency threshold (esp. with MI measure)
  - prior work often reports substantially better results than without the frequency threshold
- Approach 1: fixed threshold, e.g. f ≥ 5
- Approach 2: relative threshold, e.g. f ≥ 10 / G words

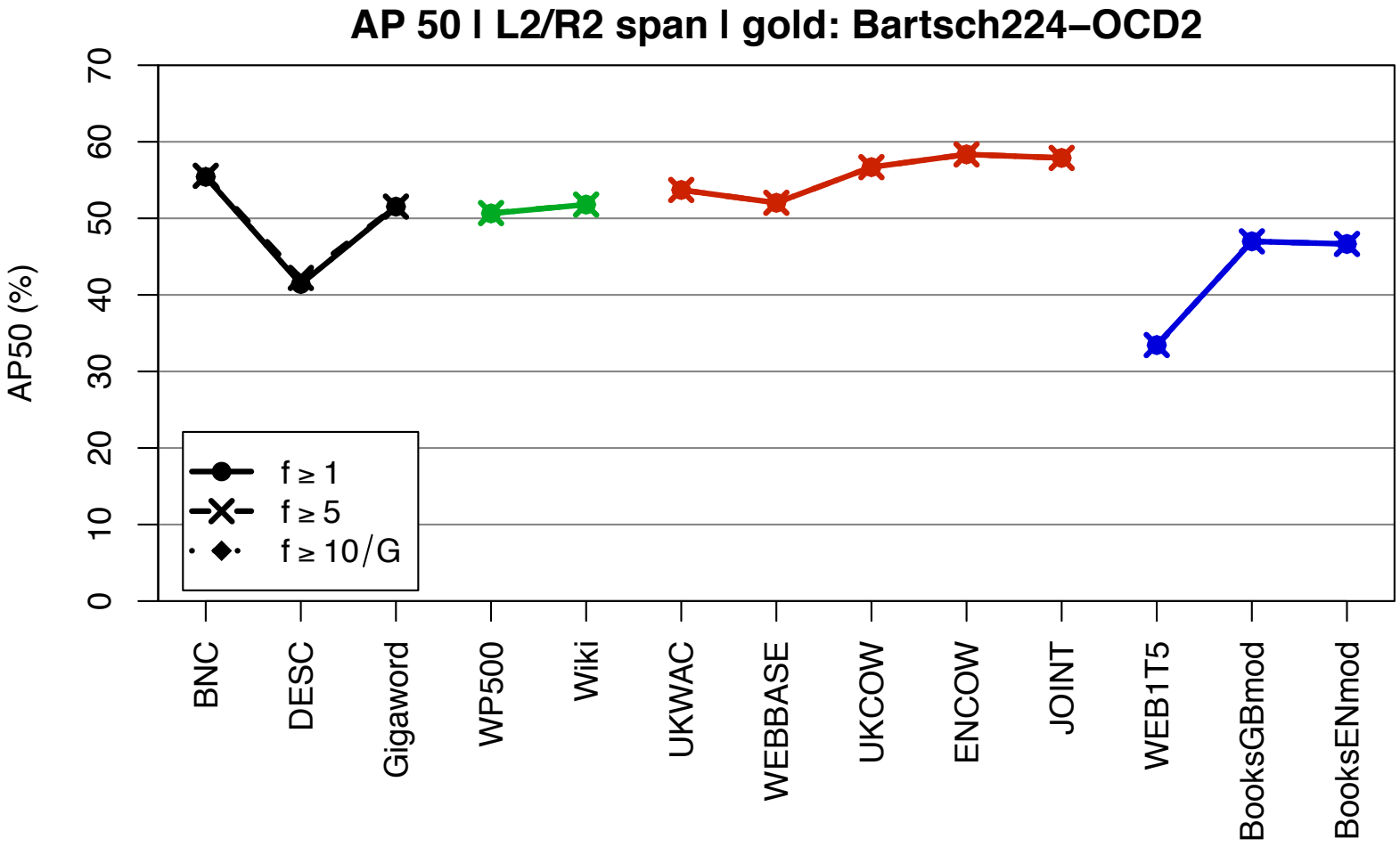
# Frequency thresholds



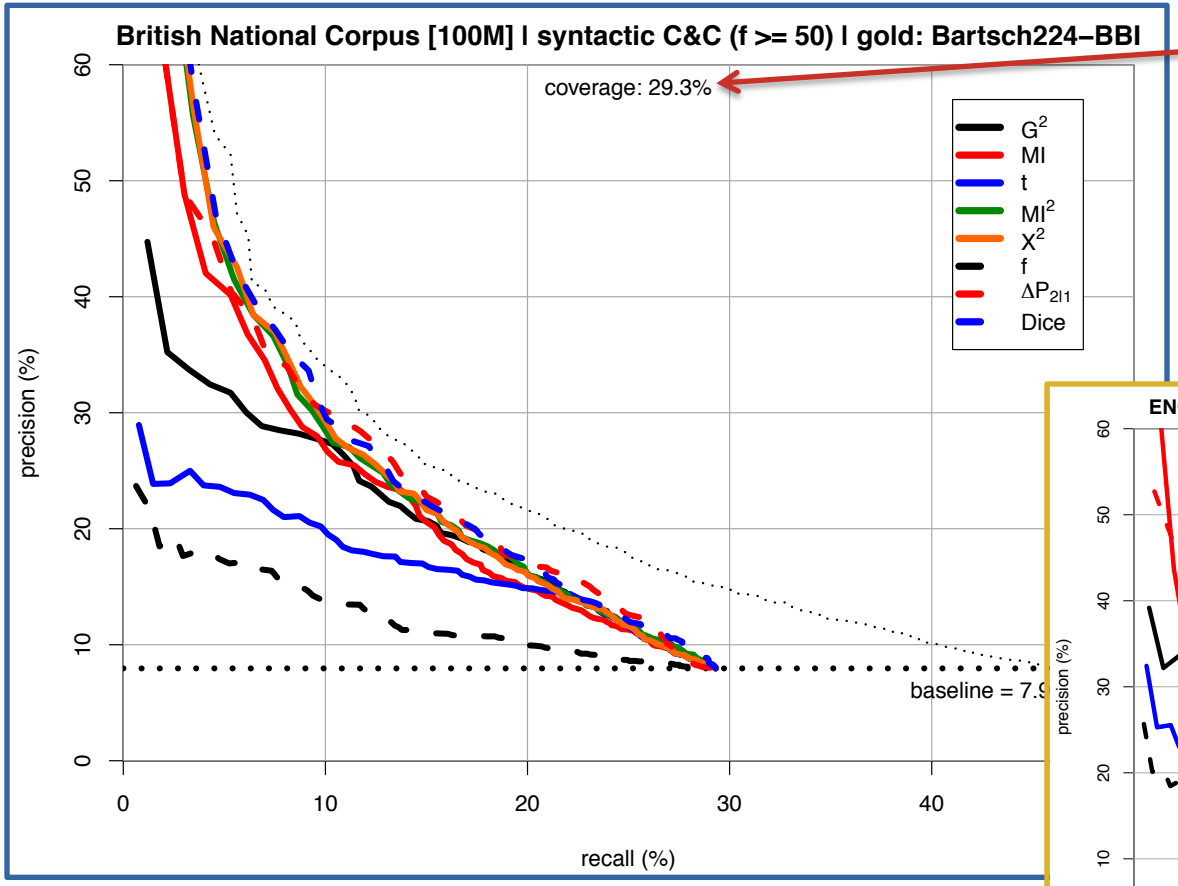
# Frequency thresholds



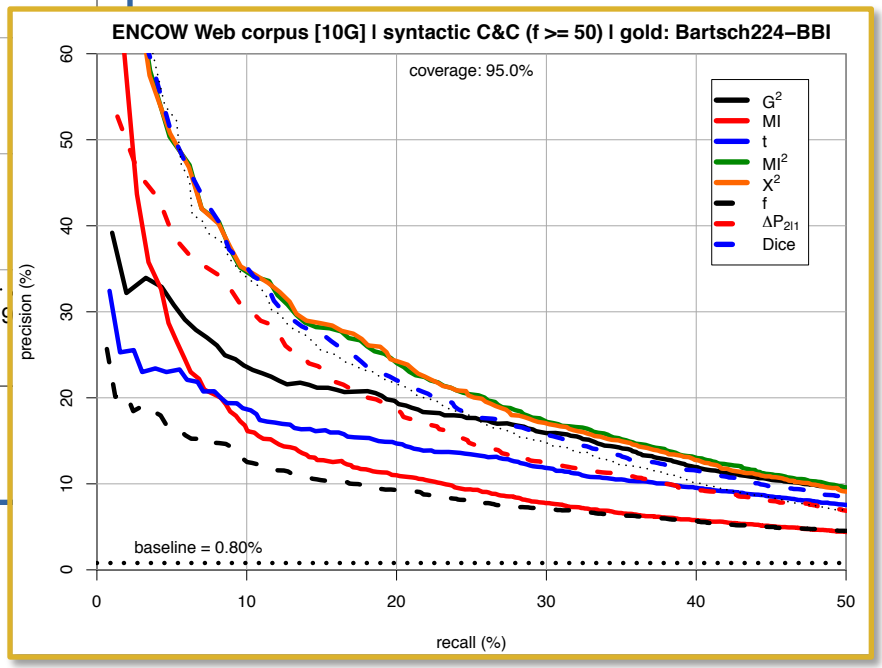
# Frequency thresholds



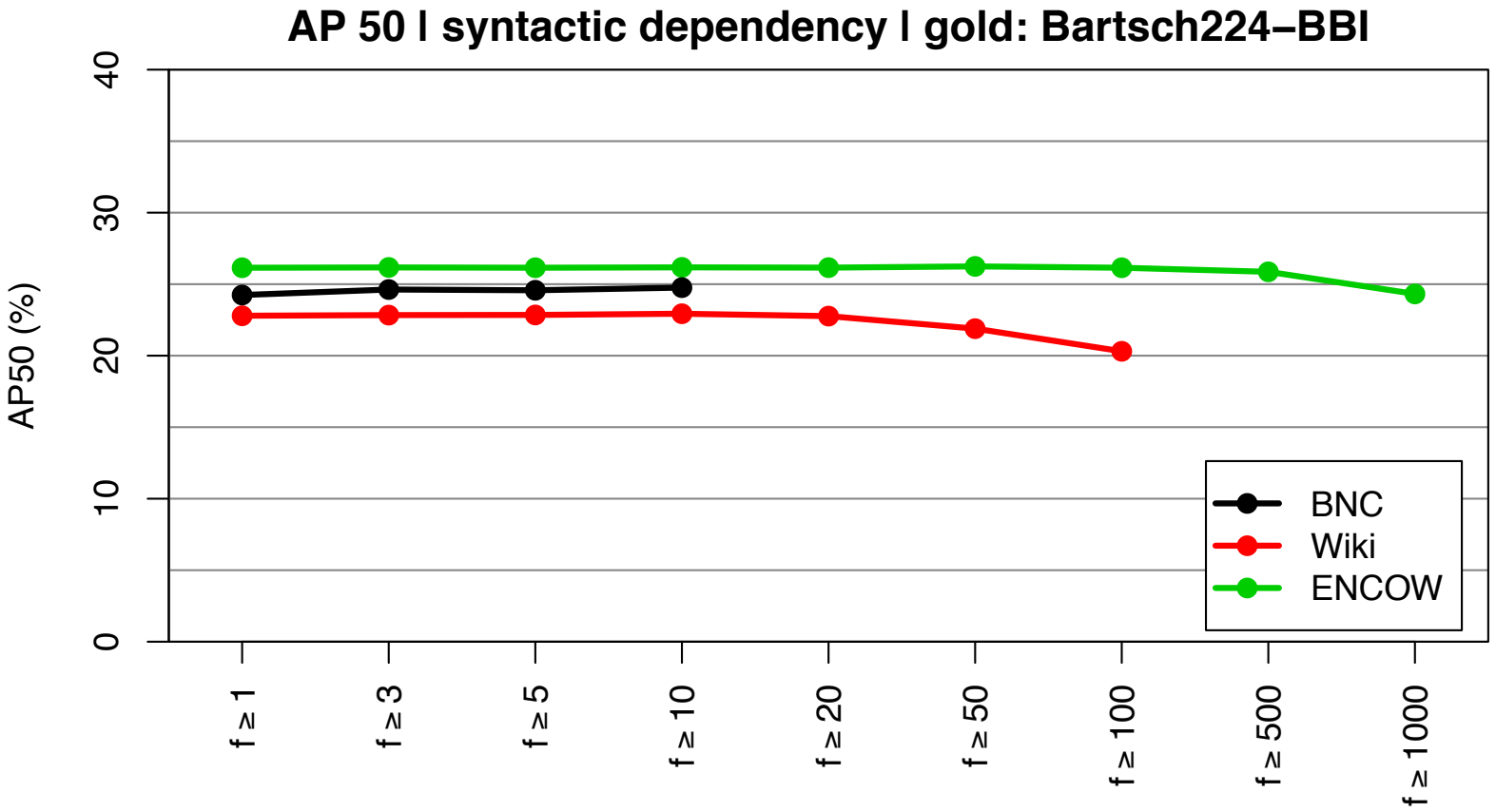
# Frequency effect?



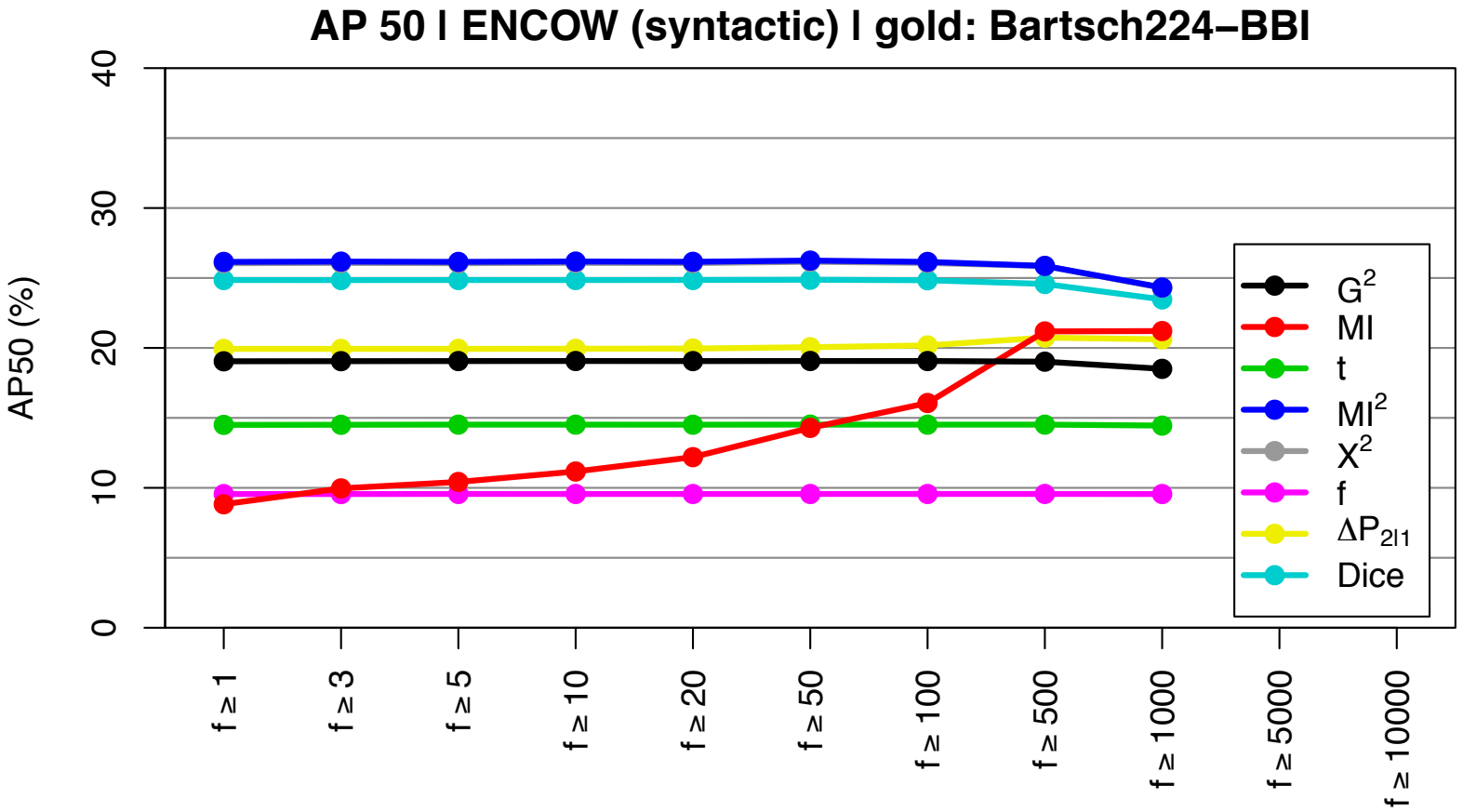

for high thresholds,
coverage falls below 50%



# Frequency effect?



# Frequency effect?



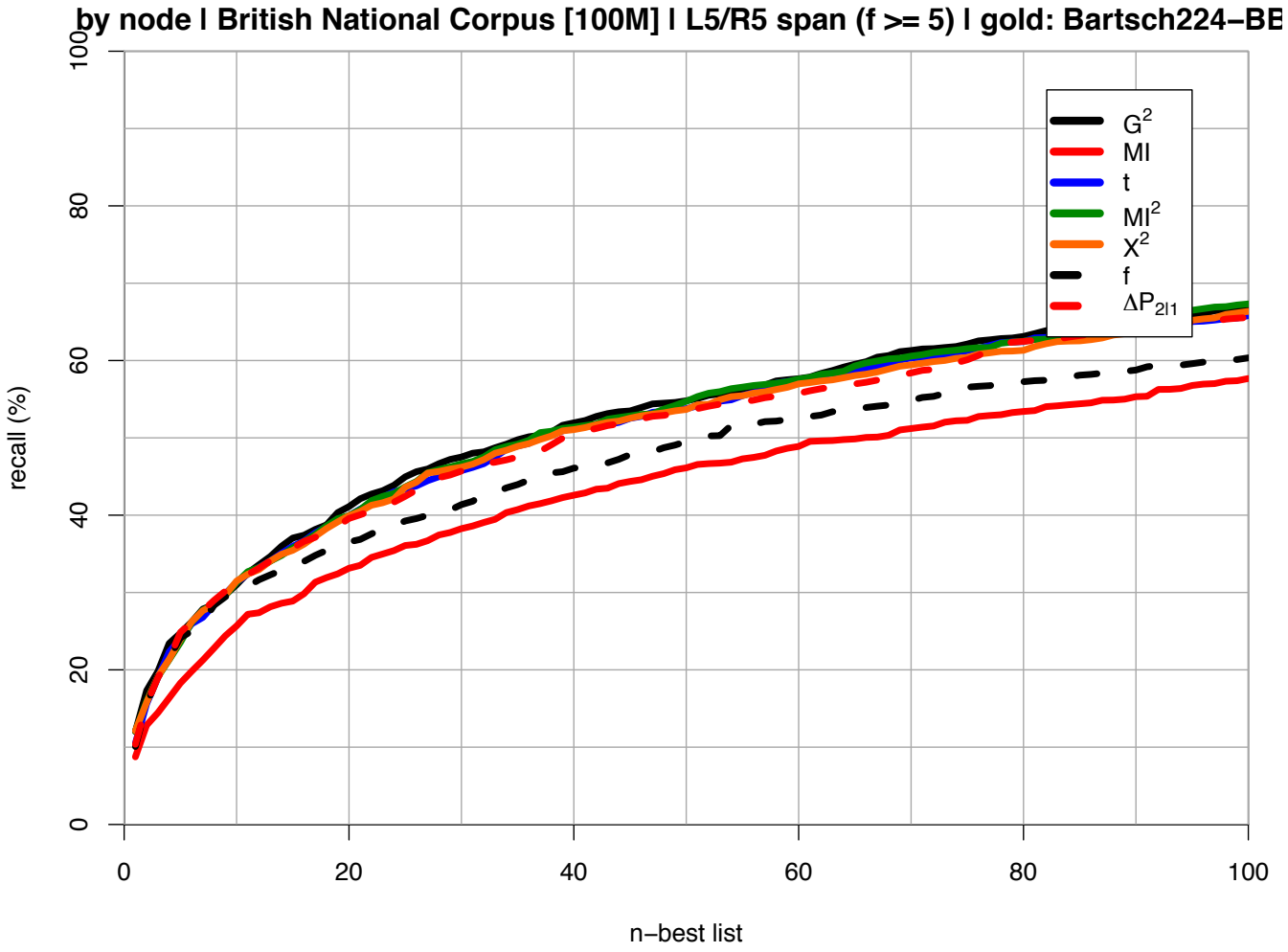
# Conclusions 2

- Suitable choice of AM depends on gold standard, i.e. the precise type of collocations to be extracted
  - chi-squared or $MI^2$ for BBI collocations
  - log-likelihood (or t-score) for OCD2 collocations
- Plausible bias of OCD2 towards BNC & certain AMs
- Good news: no interaction with other parameters
  - best AMs are the same for (nearly) all parameter settings
  - consistent effect of window size, source corpus & parser
- Frequency thresholds have virtually no effect
  - AMs successful at weeding out low-frequency noise

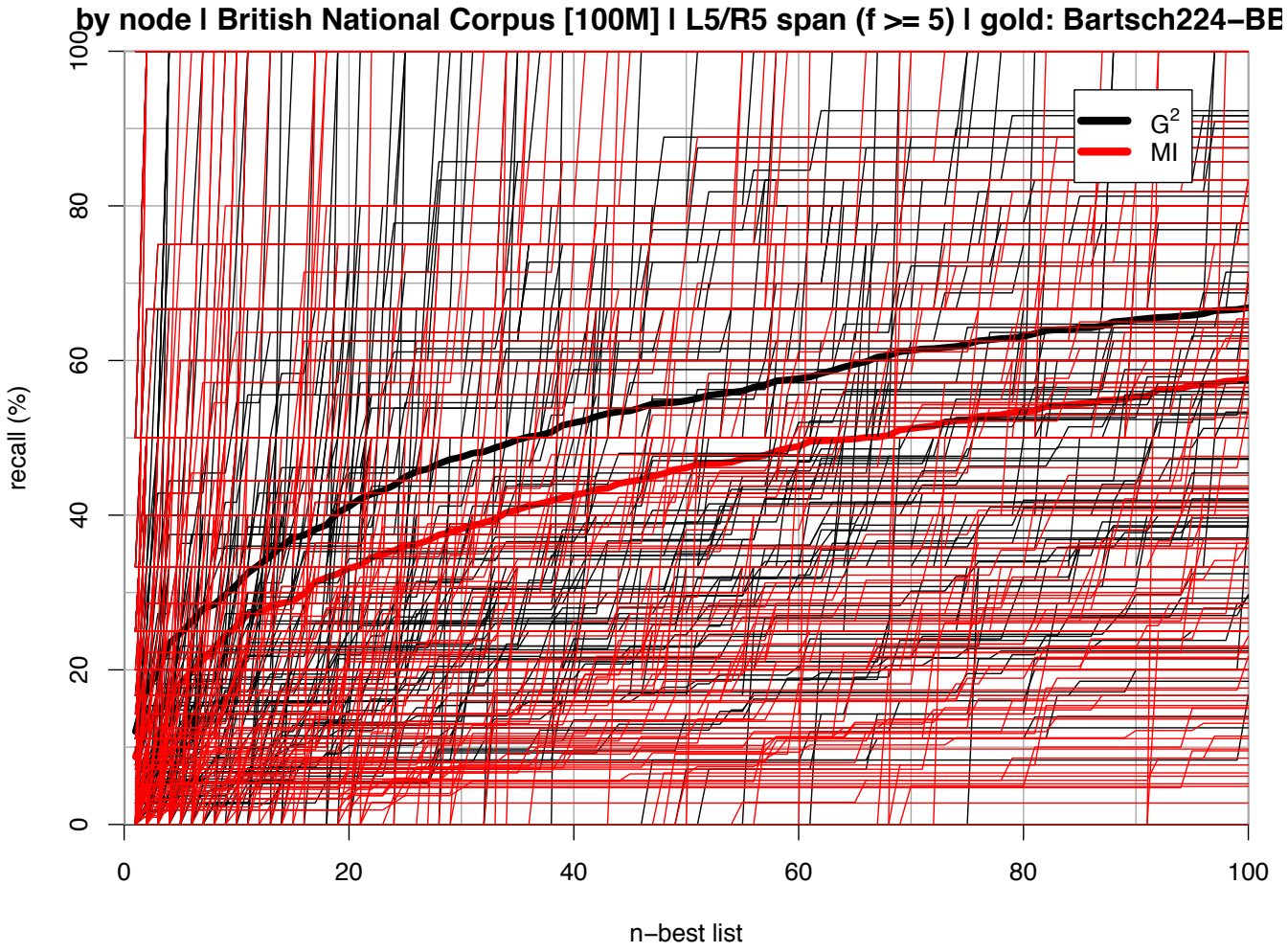
Questions?

# THANK YOU!

# Appendix: Per-node evaluation



# Appendix: Per-node evaluation



# 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


- Allan, Lorraine G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, **15**(3), 147–149.
- 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.
- 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.

# References


- 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 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.

# References


- Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford. Reprinted in Palmer (1968), pages 168–205.
- 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.
- Hausmann, Franz Josef (1979). Un dictionnaire de collocations, est-il possible? *Travaux de linguistique et de littérature*, **17**, 187–195.
- Hausmann, Franz Josef (1984). Wortschatzlernen ist Kollokationslernen. *Praxis des neusprachlichen Unterrichts*, **31**, 395–406.
- Hausmann, Franz Josef (1985). Kollokationen im deutschen Wörterbuch. ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz and J. Mugdan (eds.), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch*, number 3 in Lexicographica. Series Maior, pages 118–129.

# References


- Palmer, F. R. (ed.) (1968). *Selected Papers of J. R. Firth 1952–59*. Longmans, London.
- 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.
- Sinclair, John McH. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.
- 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.