An NLP Approach to the Evaluation of Web Corpora

Stefan Evert

Corpus Linguistics Group, Department Germanistik & Komparatistik
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
stefan.evert@fau.de

Leuven, 17 February 2015

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

Collaborators

parts of this presentation are based on the following studies

- Biemann, Chris; Bildhauer, Felix; Evert, Stefan; Goldhahn, Dirk; Quasthoff, Uwe; Schäfer, Roland; Simon, Johannes; Swiezinski, Leonard; Zesch, Torsten (2013). Scalable construction of high-quality Web corpora. *Journal for Language Technology and Computational Linguistics (JLCL)*, **28**(2), 23–59.

- Bartsch, Sabine and Evert, Stefan (2014). Towards a Firthian notion of collocation. In A. Abel and L. Lemnitzer (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern. 2. Arbeitsbericht des wissenschaftlichen Netzwerks Internetlexikografie*, OPAL 02/2014. Institut für Deutsche Sprache, Mannheim, pp. 48–61.

- Lapesa, Gabriella and Evert, Stefan (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, **2**, 531–545.

- Bartsch, Sabine; Evert, Stefan; Proisl, Thomas; Uhrig, Peter (2015). (Association) measure for measure: comparing collocation dictionaries with co-occurrence data for a better understanding of the notion of collocation. Presentation at *ICAME 36 Conference*, Trier, Germany.

Why Web corpora?

Why Web corpora?

- Properties of the Web
  - Internet English, distribution of Web genres, hyperlink graph
  - Web corpus = random sample of the (public) WWW

Why Web corpora?

- Properties of the Web
  - Internet English, distribution of Web genres, hyperlink graph
  - Web corpus = random sample of the (public) WWW
- Computer-mediated communication (CMC)
  - Twitter, Facebook, chatroom logs, discussion groups, ...
  - many Web genres share aspects of interactive CMC
  - Web corpus = targeted collection of CMC genres

Why Web corpora?

- Properties of the Web
  - Internet English, distribution of Web genres, hyperlink graph
  - Web corpus = random sample of the (public) WWW
- Computer-mediated communication (CMC)
  - Twitter, Facebook, chatroom logs, discussion groups, ...
  - many Web genres share aspects of interactive CMC
  - Web corpus = targeted collection of CMC genres
- As replacement for linguistic reference corpora
  - main goal of the early WaC(ky) community
  - cheaper, larger and more up-to-date than traditional corpora
  - Web corpus should be similar to reference corpus

Why Web corpora?

because more data are better data (Church and Mercer 1993)

- Properties of the Web
  - Internet English, distribution of Web genres, hyperlink graph
  - Web corpus = random sample of the (public) WWW
- Computer-mediated communication (CMC)
  - Twitter, Facebook, chatroom logs, discussion groups, ...
  - many Web genres share aspects of interactive CMC
  - Web corpus = targeted collection of CMC genres
- As replacement for linguistic reference corpora
  - main goal of the early WaC(ky) community
  - cheaper, larger and more up-to-date than traditional corpora
  - Web corpus should be similar to reference corpus
- Scaling up NLP training data (Banko and Brill 2001)
  - 1964: 1 million words (Brown Corpus)
  - 1995: 100 million words (British National Corpus)
  - 2003: 1,000+ million words (English Gigaword, WaCky)
  - 2006: 1,000,000 million words (Google Web 1T 5-Grams)

Why Web corpora?

because more data are better data (Church and Mercer 1993)

- Properties of the Web
  - Internet English, distribution of Web genres, hyperlink graph
  - Web corpus = random sample of the (public) WWW
- Computer-mediated communication (CMC)
  - Twitter, Facebook, chatroom logs, discussion groups, ...
  - many Web genres share aspects of interactive CMC
  - Web corpus = targeted collection of CMC genres
- As replacement for linguistic reference corpora
  - main goal of the early WaC(ky) community
  - cheaper, larger and more up-to-date than traditional corpora
  - Web corpus should be similar to reference corpus
- Scaling up NLP training data (Banko and Brill 2001)
  - 1964: 1 million words (Brown Corpus)
  - 1995: 100 million words (British National Corpus)
  - 2003: 1,000+ million words (English Gigaword, WaCky)
  - 2006: 1,000,000 million words (Google Web 1T 5-Grams)

Is bigger always better?

- From small, clean and well designed ...
  - British National Corpus (BNC)
  - movie subtitles, newspapers, ...

Is bigger always better?

- From small, clean and well designed ...
  - British National Corpus (BNC)
  - movie subtitles, newspapers, ...
- ... to large and messy ...
  - WaCky, WebBase, COW, TenTen, GloWbE, Aranea, ...
  - sampling frame unclear, lack of metadata
  - boilerplate, duplicates, non-standard language

Is bigger always better?

- From small, clean and well designed ...
  - British National Corpus (BNC)
  - movie subtitles, newspapers, ...
- ... to large and messy ...
  - WaCky, WebBase, COW, TenTen, GloWbE, Aranea, ...
  - sampling frame unclear, lack of metadata
  - boilerplate, duplicates, non-standard language
- ... to huge n-gram databases
  - largest corpora only available as n-gram databases, e.g. Google's 1-trillion-word Web corpus (Web 1T 5-Grams)
  - tend to be even messier, often w/o linguistic annotation
  - lack of context, incomplete because of frequency threshold

The Google Web 1T 5-Gram database

Brants and Franz (2006)

word 1	word 2	word 3	*f*
supplement	depend	on	193
supplement	depending	on	174
supplement	depends	entirely	94
supplement	depends	on	338
supplement	derived	from	2668
supplement	des	coups	77
supplement	described	in	200

excerpt from file `3gm-0088.gz`

Web1T5 made Easy

but not for the computer (Evert 2010)

word 1	word 2	word 3	*f*
supplement	depend	on	193
supplement	depending	on	174
supplement	depends	entirely	94
supplement	depends	on	338
supplement	derived	from	2668
supplement	des	coups	77
supplement	described	in	200

- This looks very much like a relational database table
- So why not just put the data into an off-the-shelf RDBMS?
  - built-in indexing for quick access
  - powerful query language SQL

Web1T5 made Easy

but not for the computer (Evert 2010)

word	id
depend	6094
depending	3571
depends	3846
...	...
on	14
...	...
supplement	5095

id 1	id 2	id 3	f
5095	6094	14	193
5095	3571	14	174
5095	3846	4585	94
5095	3846	14	338
5095	4207	27	2668
5095	2298	62481	77
5095	1840	11	200

- Use numeric ID coding as in IR / large-corpus query engines
- More efficient to store, index and sort in RDBMS
- Frequency-sorted lexicon is beneficial for variable-length coding of integer IDs (used by SQLite)

Web1T5-Easy database encoding procedure

Pre-processing (normalisation, filtering, . . .)

⇓

Numeric ID coding & database insertion [1d 23h]

⇓

Collapse duplicate rows (from normalisation) [6d 7h]

⇓

Indexing of each n-gram position [3d 2h]

⇓

Statistical analysis for query optimisation [not useful]

⇓

Build database of co-occurrence frequencies [ca. 3d]

⇓

✌ 211 GiB

Carried out in spring 2009 on quad-core Opteron 2.6 GHz with 16 GiB RAM — should be faster on state-of-the-art server with latest version of SQLite.

Querying the database

It's easy to search the database for patterns like

$$\textit{association} \ldots \text{X}\textit{al}\ \text{Y}$$

Querying the database

It's easy to search the database for patterns like

$$\textit{association} \ldots \text{X}\textit{al}\ \text{Y}$$

with a "simple" SQL query:

```
SELECT w3, w4, SUM(f) AS freq FROM ngrams
WHERE w1 IN (SELECT id FROM vocab WHERE w='association')
AND w3 IN (SELECT id FROM vocab WHERE w LIKE '%al')
GROUP BY w3, w4 ORDER BY freq DESC;
```

Querying the database

It's easy to search the database for patterns like

association ... X*al* Y

with a "simple" SQL query:

```
SELECT w3, w4, SUM(f) AS freq FROM ngrams
WHERE w1 IN (SELECT id FROM vocab WHERE w='association')
AND w3 IN (SELECT id FROM vocab WHERE w LIKE '%al')
GROUP BY w3, w4 ORDER BY freq DESC;
```

Web1T5-Easy implements a more user-friendly query language:

```
association ?  %al *
```

Web1T5-Easy demo

http://corpora.linguistik.uni-erlangen.de/demos/cgi-bin/Web1T5/Web1T5_freq.perl

Dutch Twitter N-Grams: http://www.let.rug.nl/~gosse/Ngrams/ngrams.html

Frequency list | Associations | Collocations

The Google Web 1T 5-Gram Database — SQLite Index & Web Interface

This is the Web interface of the Web1T5-Easy package, using a GOPHER page design.

Query Form

Search pattern: association ? %al *

Search | CSV | XML

• display first 50 N-grams with frequency ≥ 100

Help | Debug | Optim.

• variable elements are listed normally , constant elements are shown

Reset Form

Results

50 matches in 11.09 seconds

87979	association .. social workers
54756	association .. computational linguistics
54119	association .. trial lawyers
49715	association .. annual meeting
45917	association .. real estate
45703	association .. criminal defense
37246	association .. mental health
26721	association .. pharmaceutical scientists
26644	association .. professional engineers
26132	association .. artificial intelligence
24770	association .. annual conference
21821	association .. neurological surgeons

Web1T5-Easy query performance

Web1T5-Easy query	cold cache	warm cache
`corpus linguistics`	0.11s	0.01s
`web as corpus`	1.29s	0.44s
`time of *`	2.71s	1.09s
`%ly good fun`	181.03s	24.37s
`[sit,sits,sat,sitting] * ? chair`	1.16s	0.31s
`* linguistics` *(association ranking)*	11.42s	0.05s
`university of *` *(association ranking)*	1.48s	0.48s

(64-bit Linux server with 2.6 GHz AMD Opteron CPUs, 16 GiB RAM and fast local hard disk; based on timing information from the public Web interface.)

Evaluating the "quality" of Web corpora

- Statistical properties
  - type-token distributions, n-gram frequencies, other markers
  - representativeness (as sample of the Web)
  - genre distribution (traditional vs. Web genres)

Evaluating the "quality" of Web corpora

- Statistical properties
  - type-token distributions, n-gram frequencies, other markers
  - representativeness (as sample of the Web)
  - genre distribution (traditional vs. Web genres)
- Corpus comparison
  - between Web corpora (➜ reliability)
  - between Web corpus and reference corpus
  - compared to within-corpus variation

Evaluating the "quality" of Web corpora

- Statistical properties
  - type-token distributions, n-gram frequencies, other markers
  - representativeness (as sample of the Web)
  - genre distribution (traditional vs. Web genres)
- Corpus comparison
  - between Web corpora (➜ reliability)
  - between Web corpus and reference corpus
  - compared to within-corpus variation
- Training data for NLP application
  - larger amount of training data is often beneficial
  - confounding factors (NLP algorithm, training regime, ...)

Evaluating the "quality" of Web corpora

- Statistical properties
  - type-token distributions, n-gram frequencies, other markers
  - representativeness (as sample of the Web)
  - genre distribution (traditional vs. Web genres)
- Corpus comparison
  - between Web corpora (➜ reliability)
  - between Web corpus and reference corpus
  - compared to within-corpus variation
- Training data for NLP application
  - larger amount of training data is often beneficial
  - confounding factors (NLP algorithm, training regime, ...)
- Linguistic evaluation of Web corpora
  - as substitute for / extension of reference corpus
  - need linguistic tasks that can be judged quantitatively and that make immediate use of corpus frequency data

Evaluating the "quality" of Web corpora

- Statistical properties
  - type-token distributions, n-gram frequencies, other markers
  - representativeness (as sample of the Web)
  - genre distribution (traditional vs. Web genres)
- Corpus comparison
  - between Web corpora (➜ reliability)
  - between Web corpus and reference corpus
  - compared to within-corpus variation
- Training data for NLP application
  - larger amount of training data is often beneficial
  - confounding factors (NLP algorithm, training regime, ...)
- Linguistic evaluation of Web corpora
  - as substitute for / extension of reference corpus
  - need linguistic tasks that can be judged quantitatively and that make immediate use of corpus frequency data

Linguistic evaluation of Web corpora

1. Frequency comparison
   - "good" Web corpora should agree with reference corpus on core phenomena ➜ correlation between frequency counts
   - e.g. Basic English vocabulary, compound nouns, ...

Linguistic evaluation of Web corpora

1. Frequency comparison
   - "good" Web corpora should agree with reference corpus on core phenomena ➜ correlation between frequency counts
   - e.g. Basic English vocabulary, compound nouns, ...
2. Identification of multiword expressions (MWE)
   - well-know NLP task based on co-occurrence statistics
   - some gold standard data sets available
   - e.g. "phrasal verbs", lexical collocations, ...

Linguistic evaluation of Web corpora

1. Frequency comparison
   - "good" Web corpora should agree with reference corpus on core phenomena ➜ correlation between frequency counts
   - e.g. Basic English vocabulary, compound nouns, ...
2. Identification of multiword expressions (MWE)
   - well-know NLP task based on co-occurrence statistics
   - some gold standard data sets available
   - e.g. "phrasal verbs", lexical collocations, ...
3. Distributional semantic models (DSM)
   - hypothesis: semantic similarity $\sim$ distributional similarity
   - distribution quantified by co-occurrences with other words
   - DSMs can be evaluated in various shared tasks

Research questions

- Are English Web corpora a substitute for the BNC?
- What are the differences between Web corpora?
- Does size matter more than content?
- How useful are n-gram databases?
  - esp. detrimental effects of frequency thresholds
- How important is (automatic) linguistic annotation?
- Do Web corpora offer better coverage?

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G
- Wackypedia subset (WP500) Malt 0.2 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G
- Wackypedia subset (WP500) Malt 0.2 G
- ukWaC (Baroni *et al.* 2009) Malt 2.0 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G
- Wackypedia subset (WP500) Malt 0.2 G
- ukWaC (Baroni *et al.* 2009) Malt 2.0 G
- WebBase (Han *et al.* 2013) 3.0 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G
- Wackypedia subset (WP500) Malt 0.2 G
- ukWaC (Baroni *et al.* 2009) Malt 2.0 G
- WebBase (Han *et al.* 2013) 3.0 G
- UKCOW 2012 (Schäfer and Bildhauer 2012) 4.0 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G
- Wackypedia subset (WP500) Malt 0.2 G
- ukWaC (Baroni *et al.* 2009) Malt 2.0 G
- WebBase (Han *et al.* 2013) 3.0 G
- UKCOW 2012 (Schäfer and Bildhauer 2012) 4.0 G
- Joint Web corpus 10.0 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G
- Wackypedia subset (WP500) Malt 0.2 G
- ukWaC (Baroni *et al.* 2009) Malt 2.0 G
- WebBase (Han *et al.* 2013) 3.0 G
- UKCOW 2012 (Schäfer and Bildhauer 2012) 4.0 G
- Joint Web corpus 10.0 G
- Web 1T 5-Grams (Brants and Franz 2006) 1000.0 G

Corpora in the evaluation

Ref: British National Corpus (Aston and Burnard 1998) C&C	0.1 G
English Movie Subtitles (DESC v2) C&C	0.1 G
Gigaword newspaper corpus (2nd edition)	2.0 G
English Wackypedia Malt	1.0 G
Wackypedia subset (WP500) Malt	0.2 G
ukWaC (Baroni *et al.* 2009) Malt	2.0 G
WebBase (Han *et al.* 2013)	3.0 G
UKCOW 2012 (Schäfer and Bildhauer 2012)	4.0 G
Joint Web corpus	10.0 G
Web 1T 5-Grams (Brants and Franz 2006)	1000.0 G
LCC n-gram database	1.0 G

Corpora in the evaluation

Ref: British National Corpus (Aston and Burnard 1998) C&C	0.1 G
English Movie Subtitles (DESC v2) C&C	0.1 G
Gigaword newspaper corpus (2nd edition)	2.0 G
English Wackypedia Malt	1.0 G
Wackypedia subset (WP500) Malt	0.2 G
ukWaC (Baroni *et al.* 2009) Malt	2.0 G
WebBase (Han *et al.* 2013)	3.0 G
UKCOW 2012 (Schäfer and Bildhauer 2012)	4.0 G
Joint Web corpus	10.0 G
Web 1T 5-Grams (Brants and Franz 2006)	1000.0 G
LCC n-gram database	1.0 G
ENCOW 2014 Malt	10.0 G

Corpora in the evaluation

Ref: British National Corpus (Aston and Burnard 1998) C&C	0.1 G
English Movie Subtitles (DESC v2) C&C	0.1 G
Gigaword newspaper corpus (2nd edition)	2.0 G
English Wackypedia Malt	1.0 G
Wackypedia subset (WP500) Malt	0.2 G
ukWaC (Baroni *et al.* 2009) Malt	2.0 G
WebBase (Han *et al.* 2013)	3.0 G
UKCOW 2012 (Schäfer and Bildhauer 2012)	4.0 G
Joint Web corpus	10.0 G
Web 1T 5-Grams (Brants and Franz 2006)	1000.0 G
LCC n-gram database	1.0 G
ENCOW 2014 Malt	10.0 G
Google Books 2012 EN (Lin *et al.* 2012) Malt	900.0 G

Corpora in the evaluation

- Ref: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- English Movie Subtitles (DESC v2) C&C 0.1 G
- Gigaword newspaper corpus (2nd edition) 2.0 G
- English Wackypedia Malt 1.0 G
- Wackypedia subset (WP500) Malt 0.2 G
- ukWaC (Baroni *et al.* 2009) Malt 2.0 G
- WebBase (Han *et al.* 2013) 3.0 G
- UKCOW 2012 (Schäfer and Bildhauer 2012) 4.0 G
- Joint Web corpus 10.0 G
- Web 1T 5-Grams (Brants and Franz 2006) 1000.0 G
- LCC n-gram database 1.0 G

- ENCOW 2014 Malt 10.0 G
- Google Books 2012 EN (Lin *et al.* 2012) Malt 900.0 G
- Google Books 2012 GB Malt 100.0 G

Frequency Comparison

Comparison of frequency counts

- Scatterplots of (log) frequencies in BNC vs. other corpora
  - Pearson correlation r from regression $f_{\text{ukWaC}} \sim \beta \cdot f_{\text{BNC}}$ etc.
  - only consider items that occur in both corpora
    (➜ low coverage is not penalized directly)

Comparison of frequency counts

- Scatterplots of (log) frequencies in BNC vs. other corpora
  - Pearson correlation r from regression $f_{\text{ukWaC}} \sim \beta \cdot f_{\text{BNC}}$ etc.
  - only consider items that occur in both corpora
    (➜ low coverage is not penalized directly)
- Test data sets
  - Basic English words (lemmatized)
  - inflected forms of Basic English words
  - binary compound nouns extracted from WordNet 3.0

Comparison of frequency counts

- Scatterplots of (log) frequencies in BNC vs. other corpora
  - Pearson correlation r from regression $f_{\text{ukWaC}} \sim \beta \cdot f_{\text{BNC}}$ etc.
  - only consider items that occur in both corpora
    (➜ low coverage is not penalized directly)
- Test data sets
  - Basic English words (lemmatized)
  - inflected forms of Basic English words
  - binary compound nouns extracted from WordNet 3.0
- Morphological query expansion for unannotated n-grams

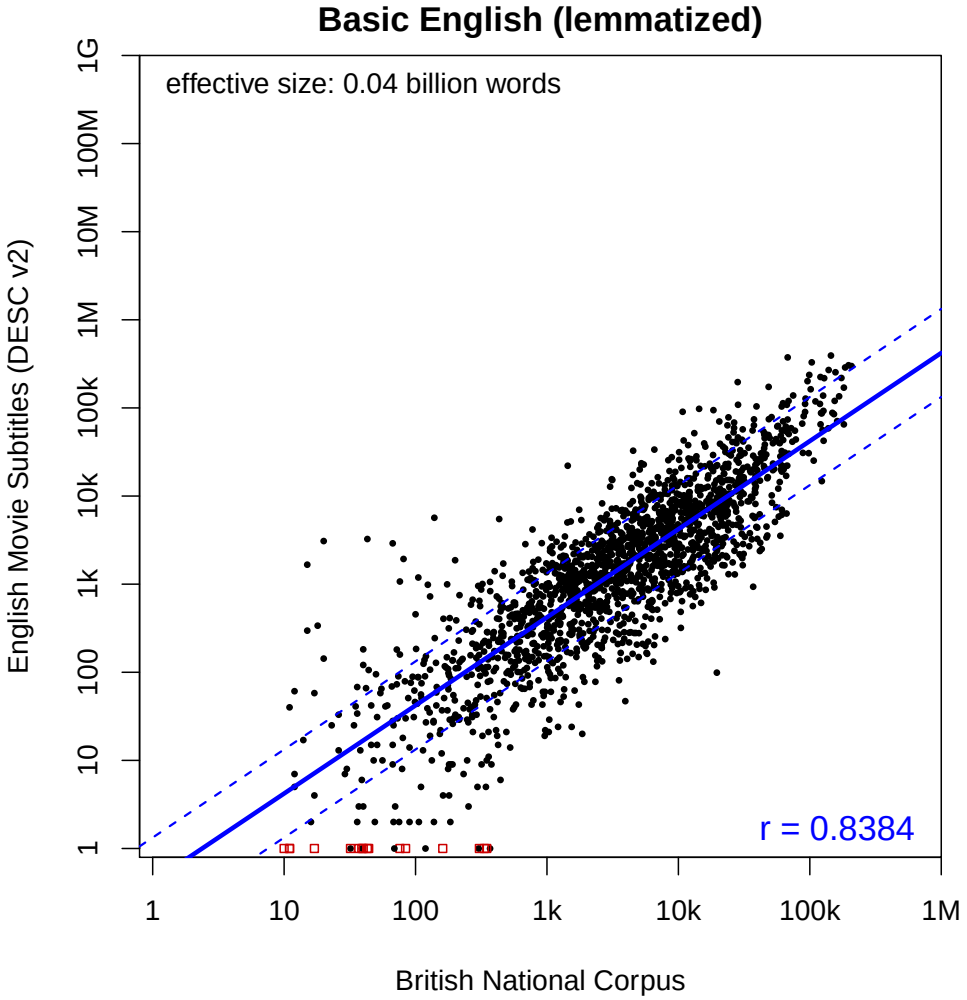
query	f
`hear sound`	36,304

Comparison of frequency counts

- Scatterplots of (log) frequencies in BNC vs. other corpora
  - Pearson correlation r from regression $f_{\text{ukWaC}} \sim \beta \cdot f_{\text{BNC}}$ etc.
  - only consider items that occur in both corpora
    (➜ low coverage is not penalized directly)
- Test data sets
  - Basic English words (lemmatized)
  - inflected forms of Basic English words
  - binary compound nouns extracted from WordNet 3.0
- Morphological query expansion for unannotated n-grams

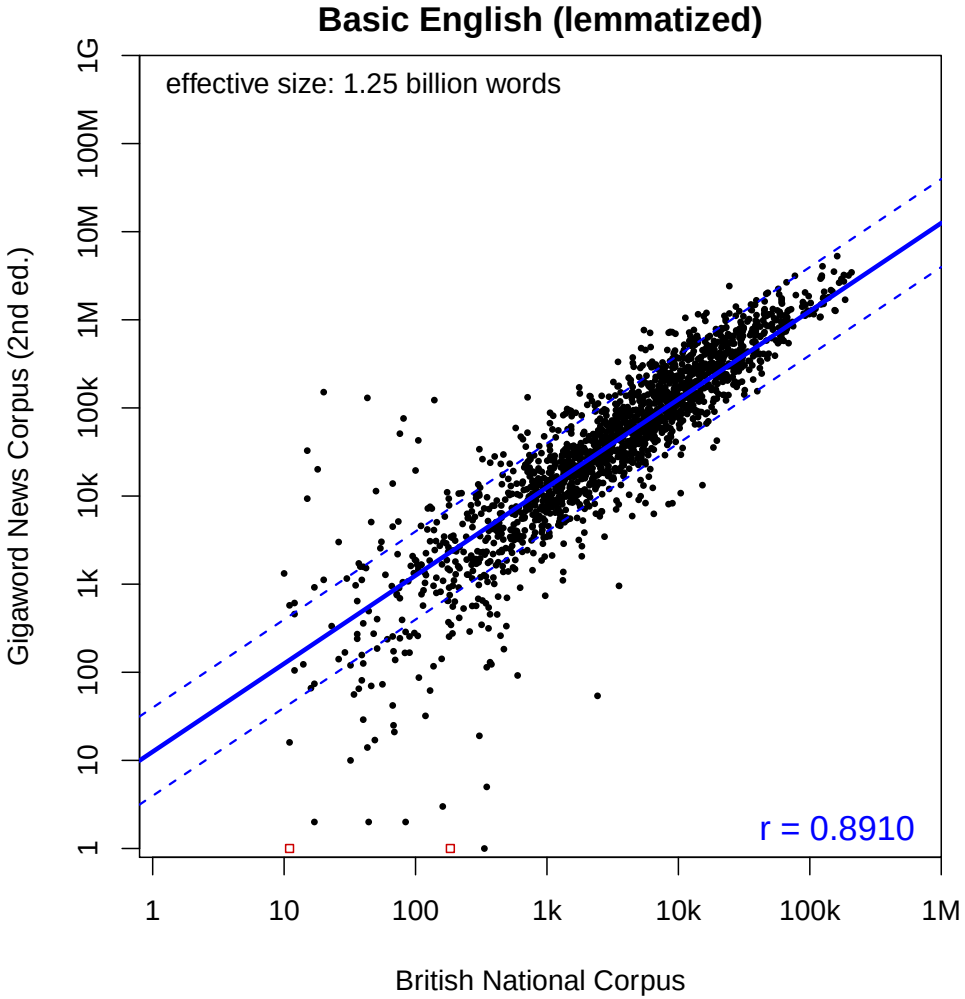
query	f
`hear sound`	36,304
`[hear,hears,heard,hearing]`	
`[sound,sounds]`	95,453

Frequency comparison: Basic English (lemmatized)



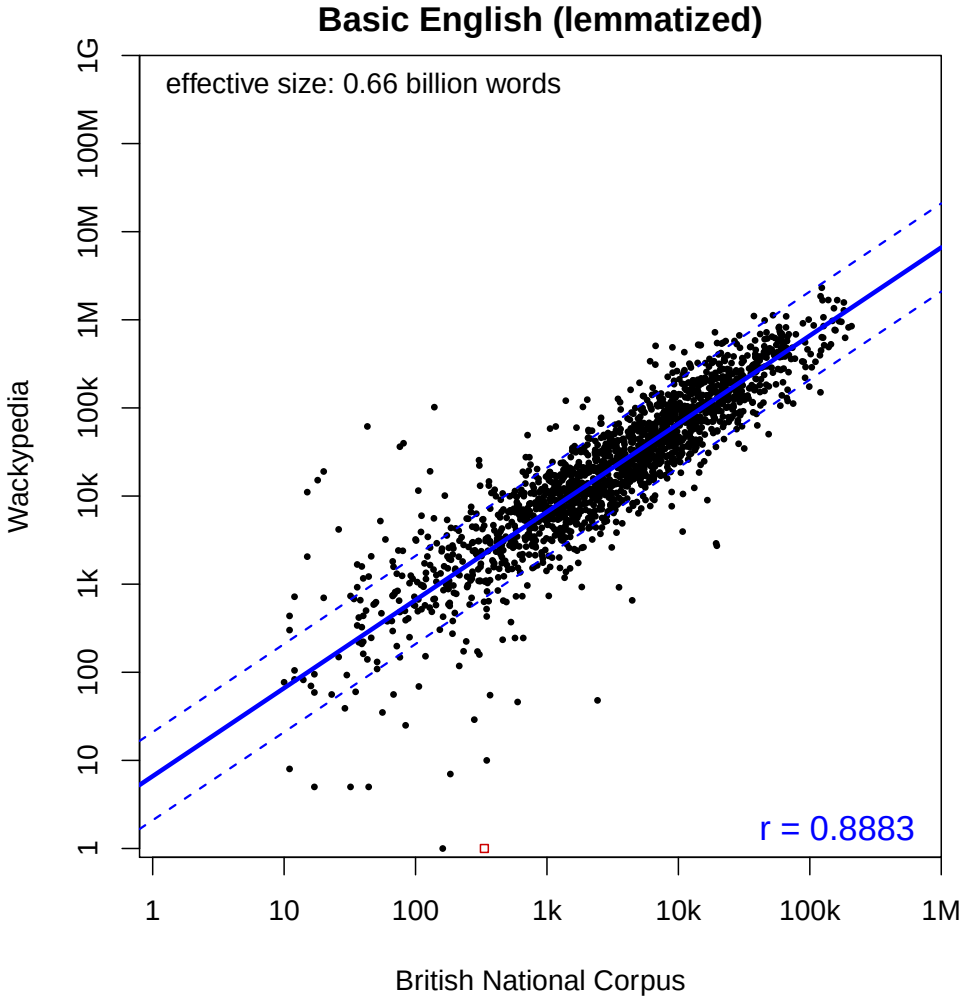
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



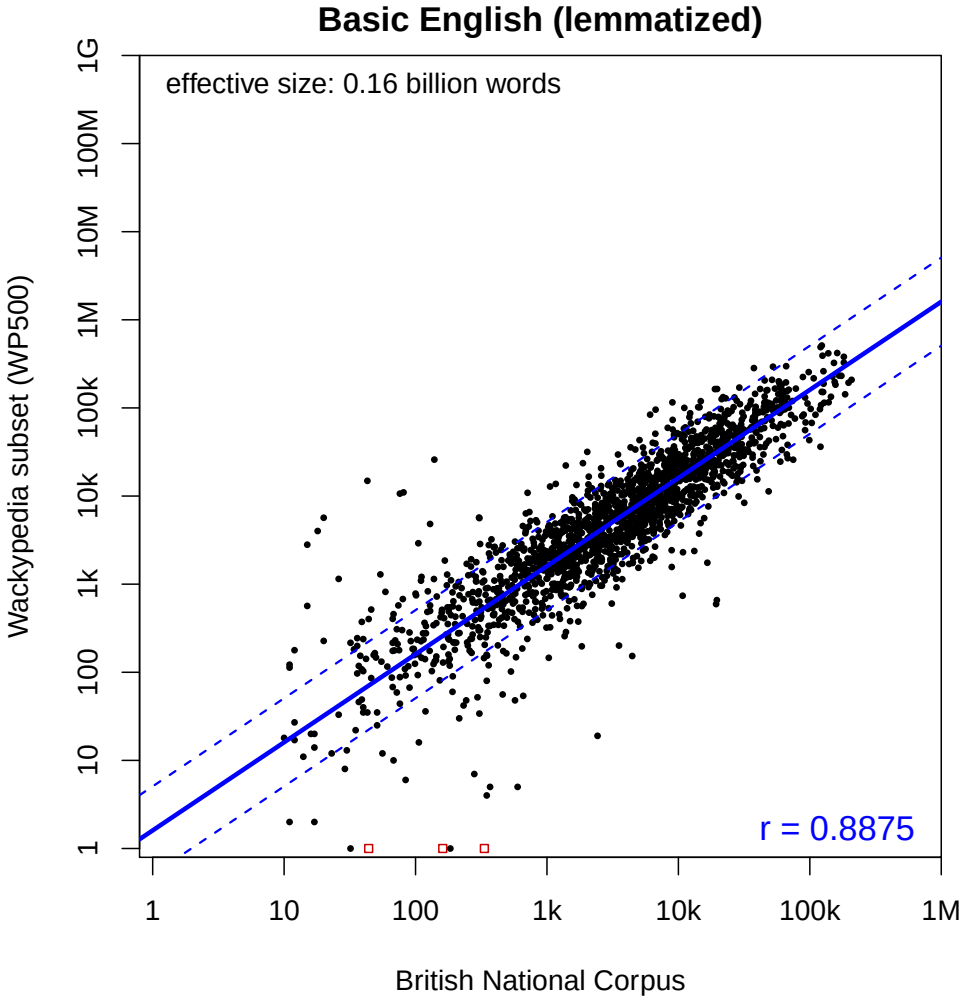
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



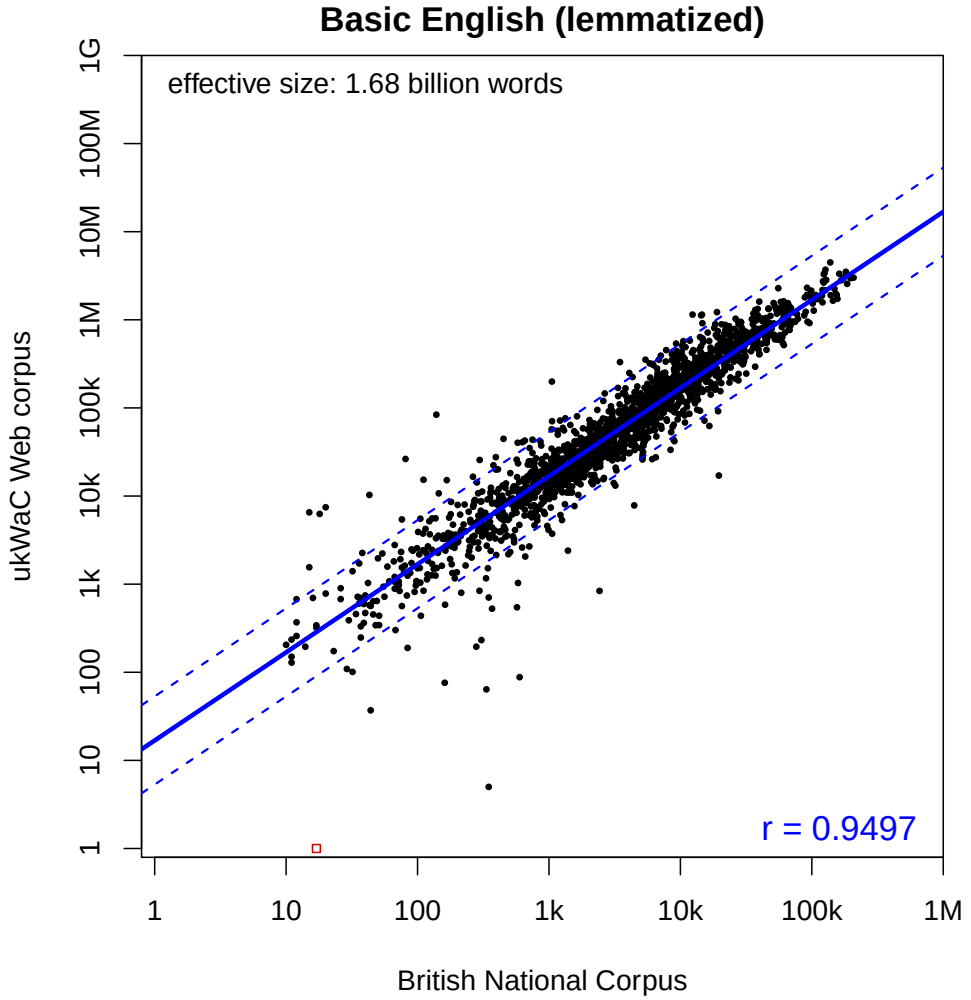
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



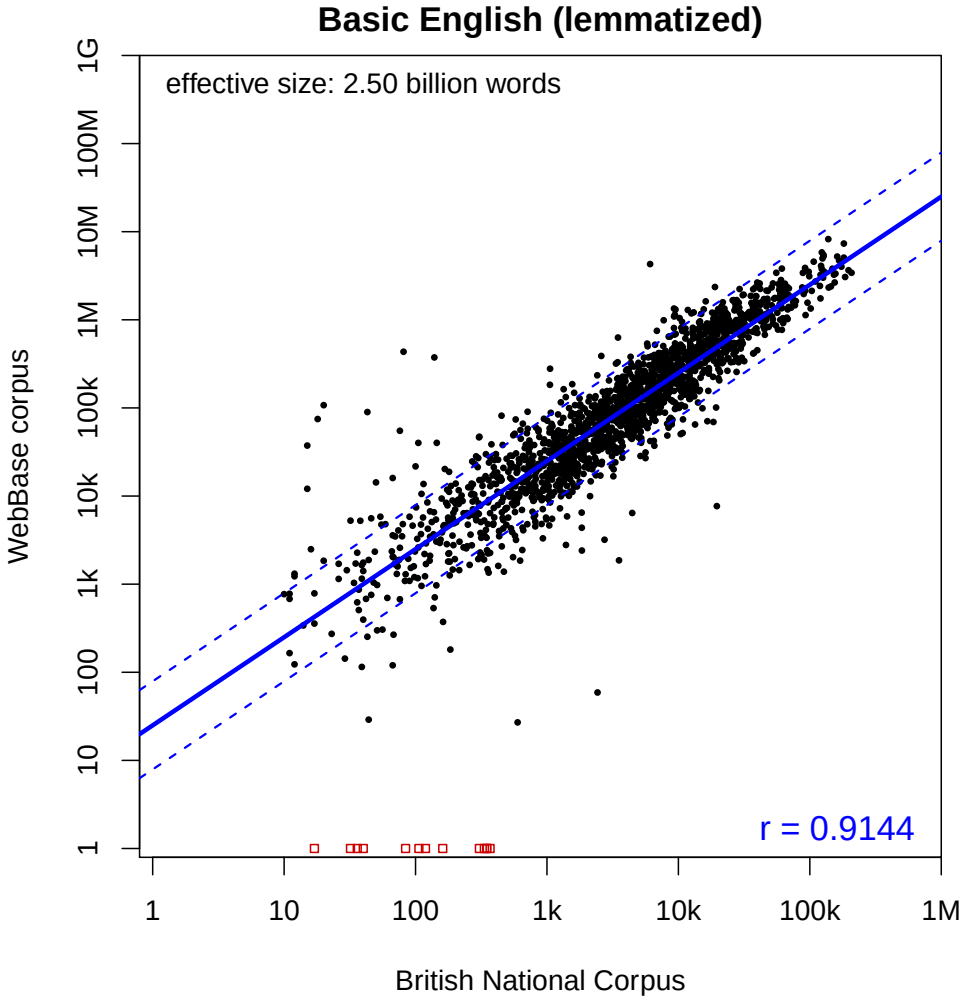
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



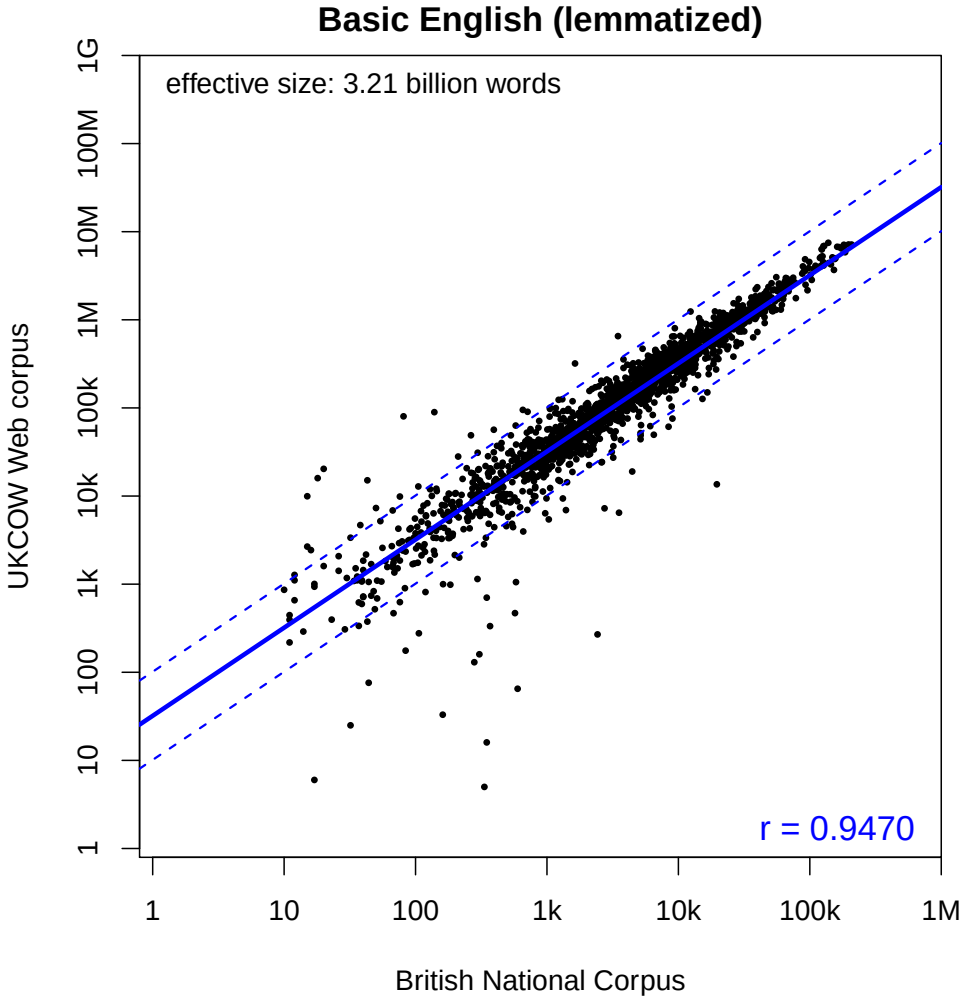
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



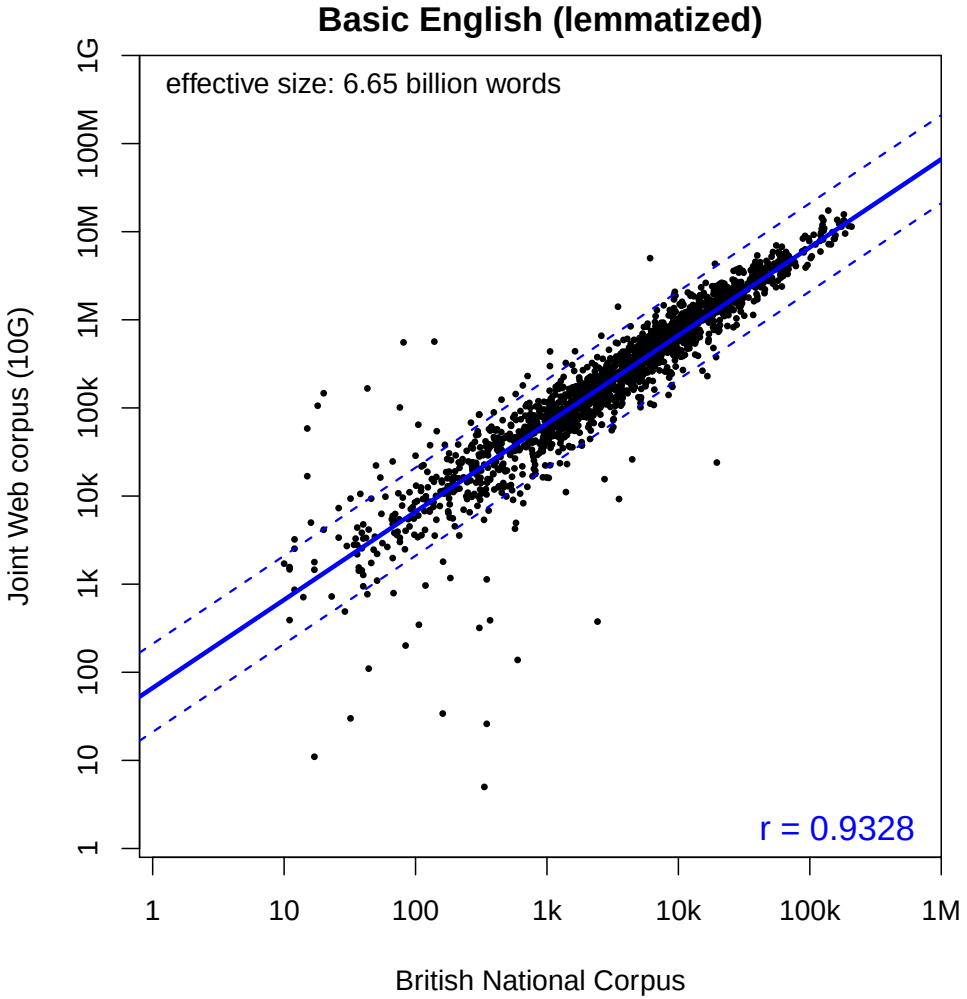
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



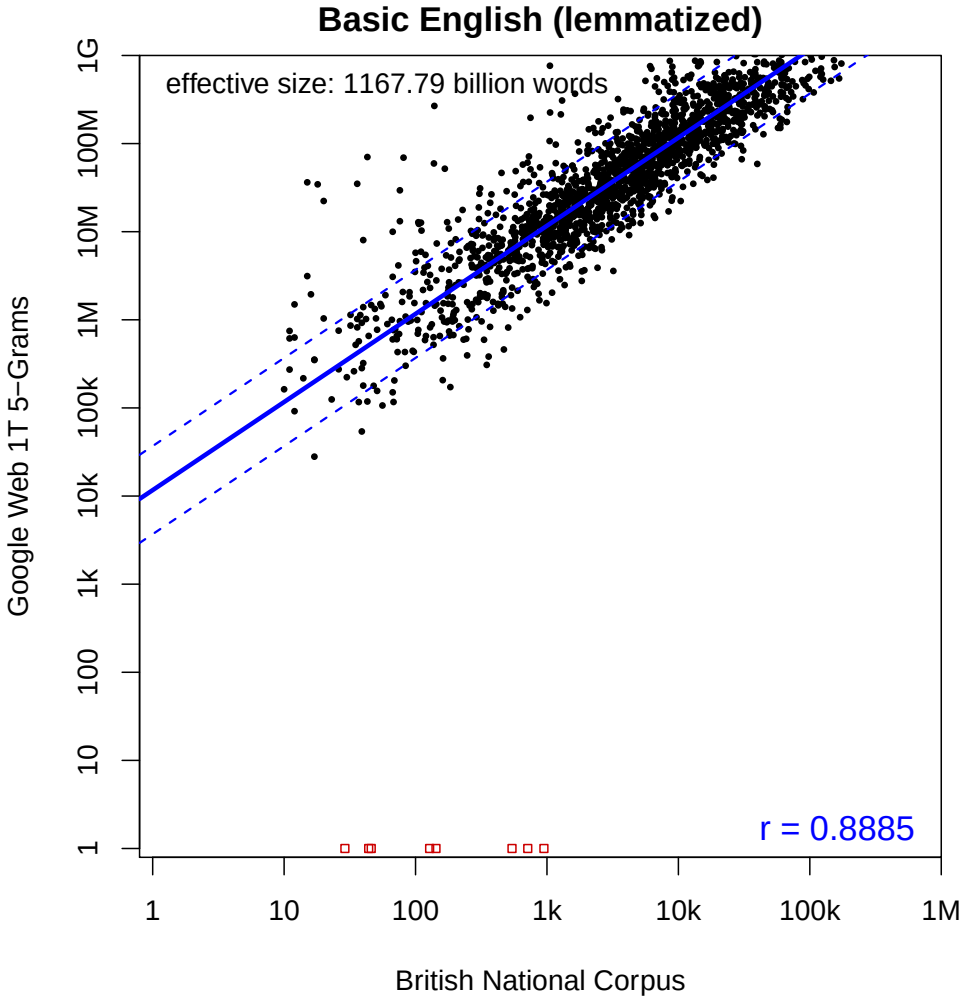
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



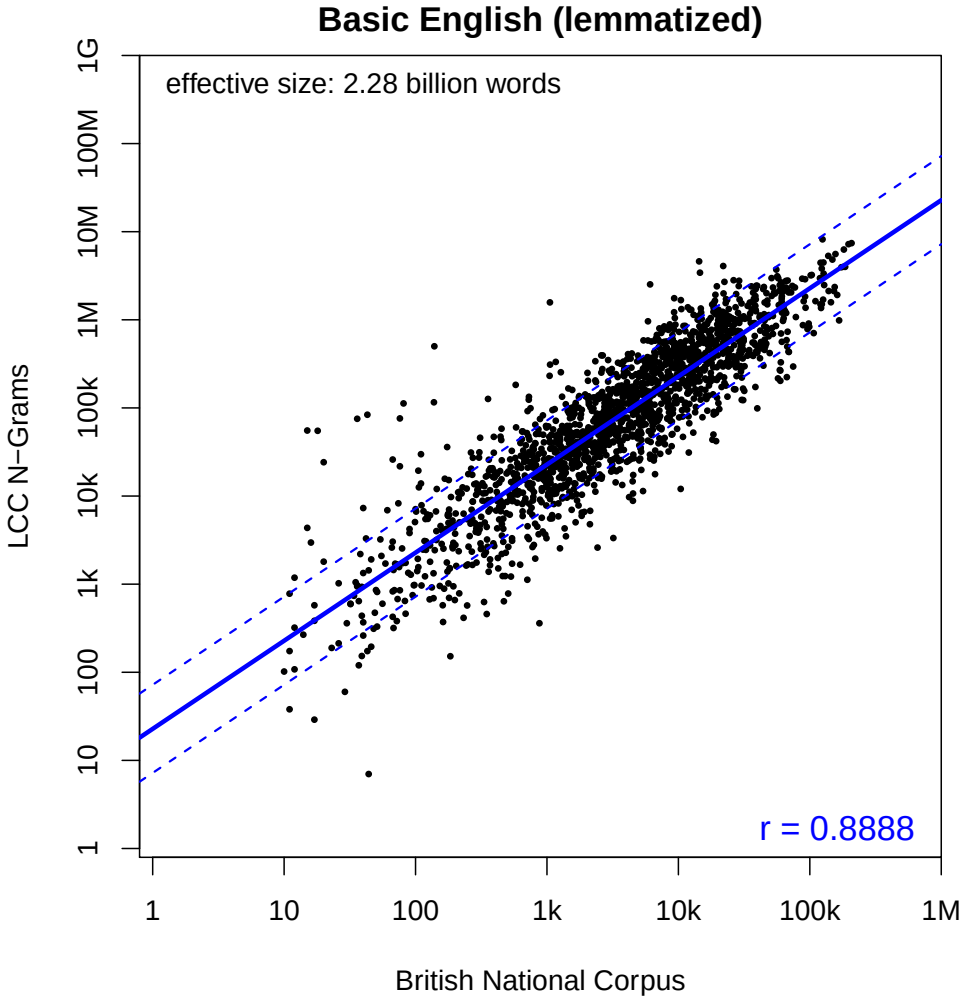
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



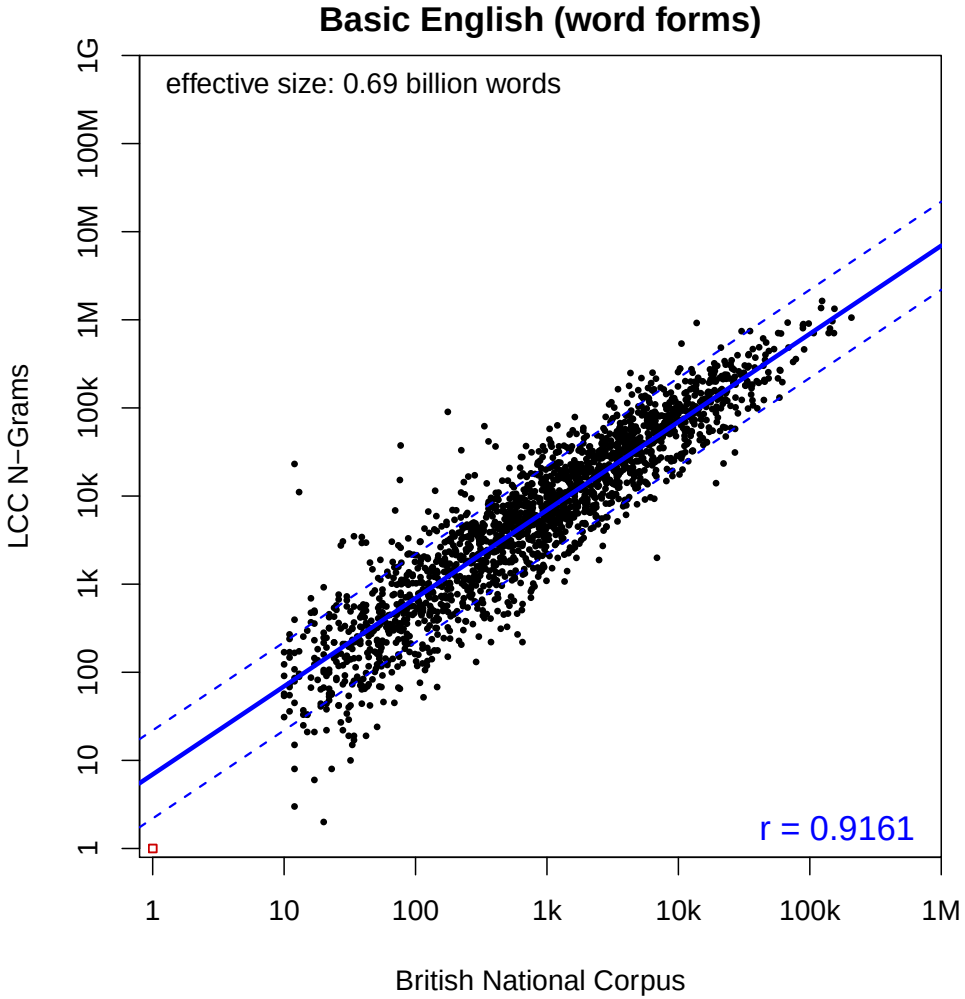
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (lemmatized)



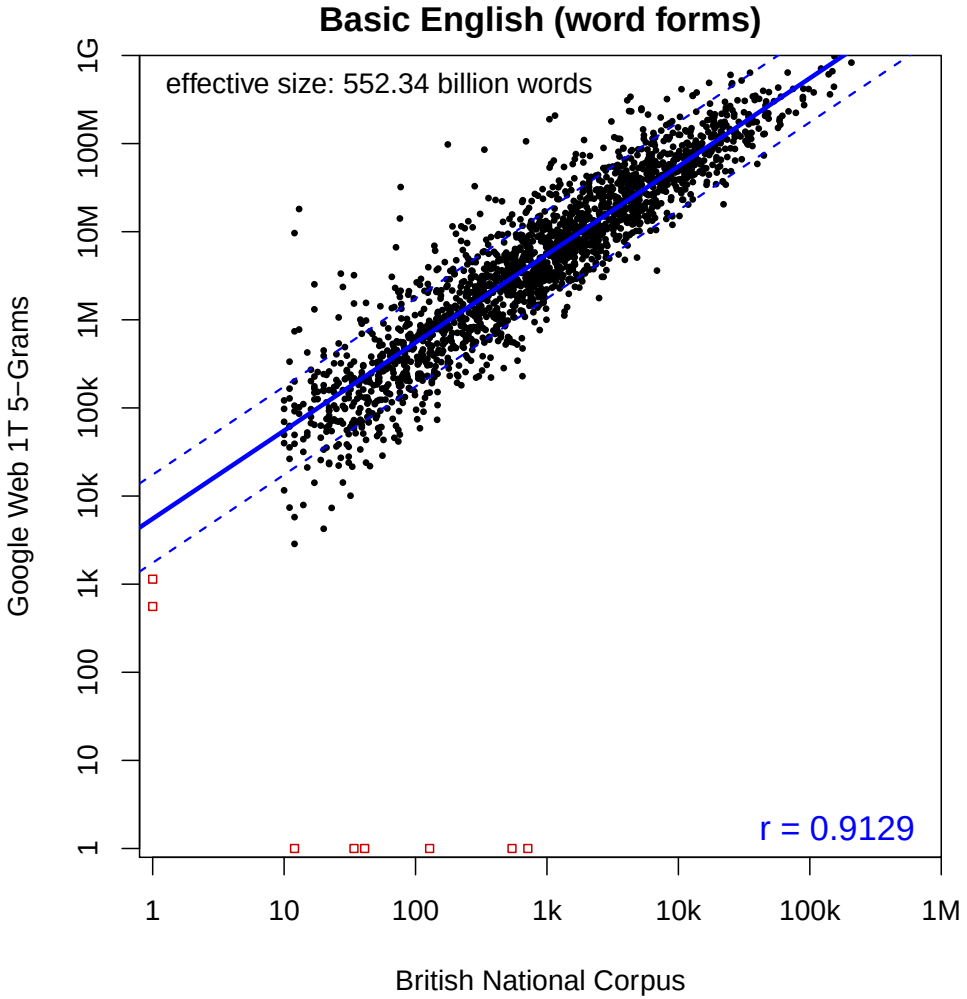
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (word forms)



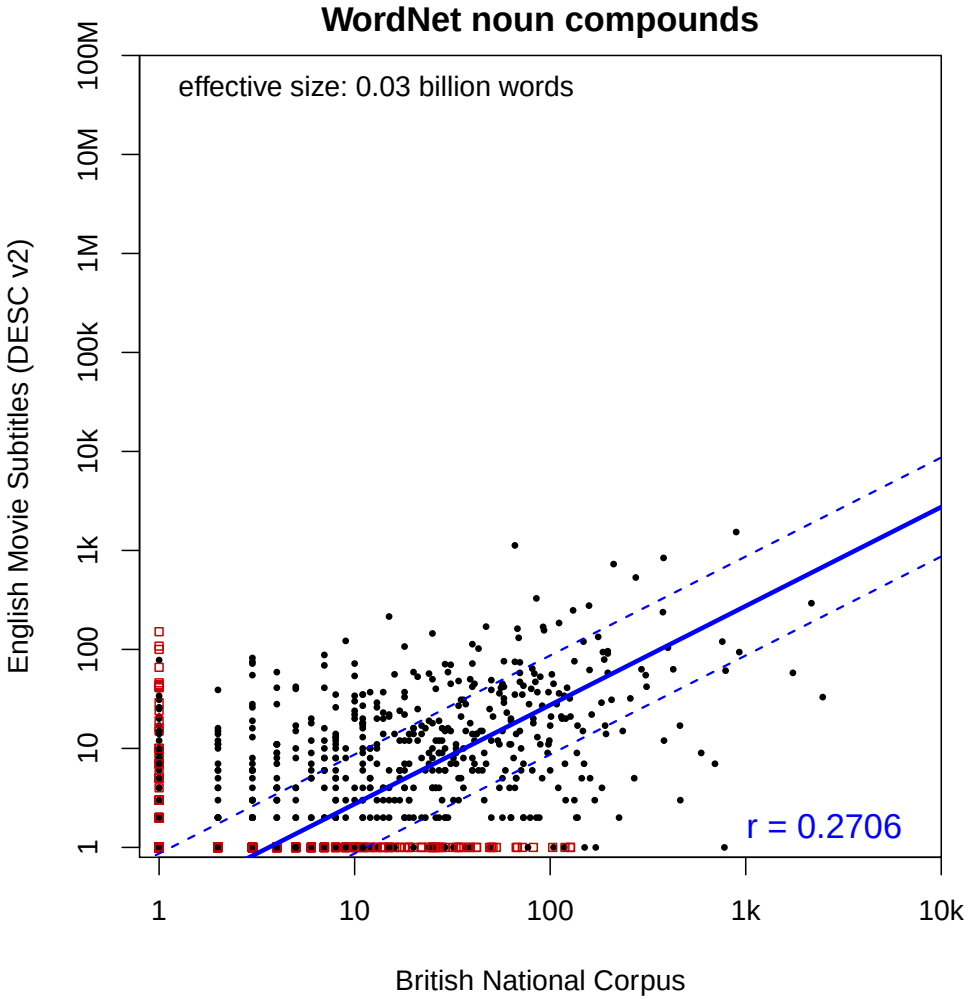
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: Basic English (word forms)



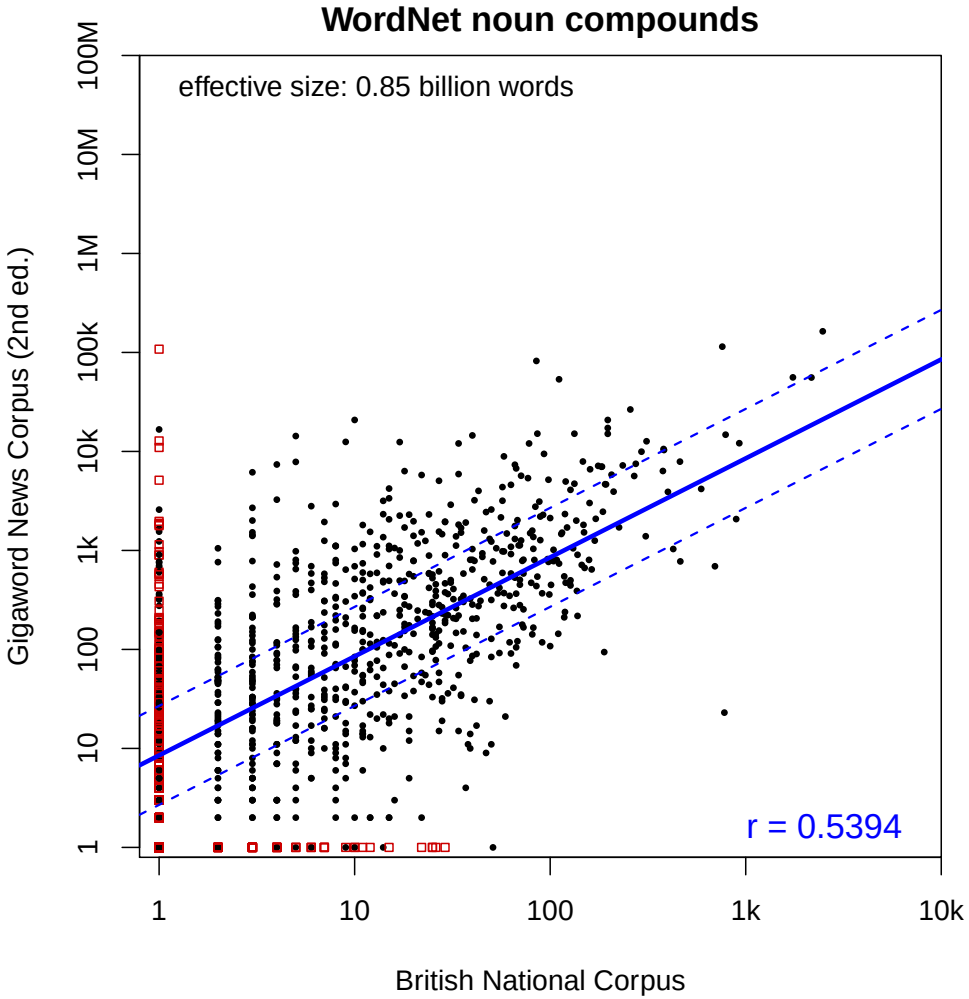
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



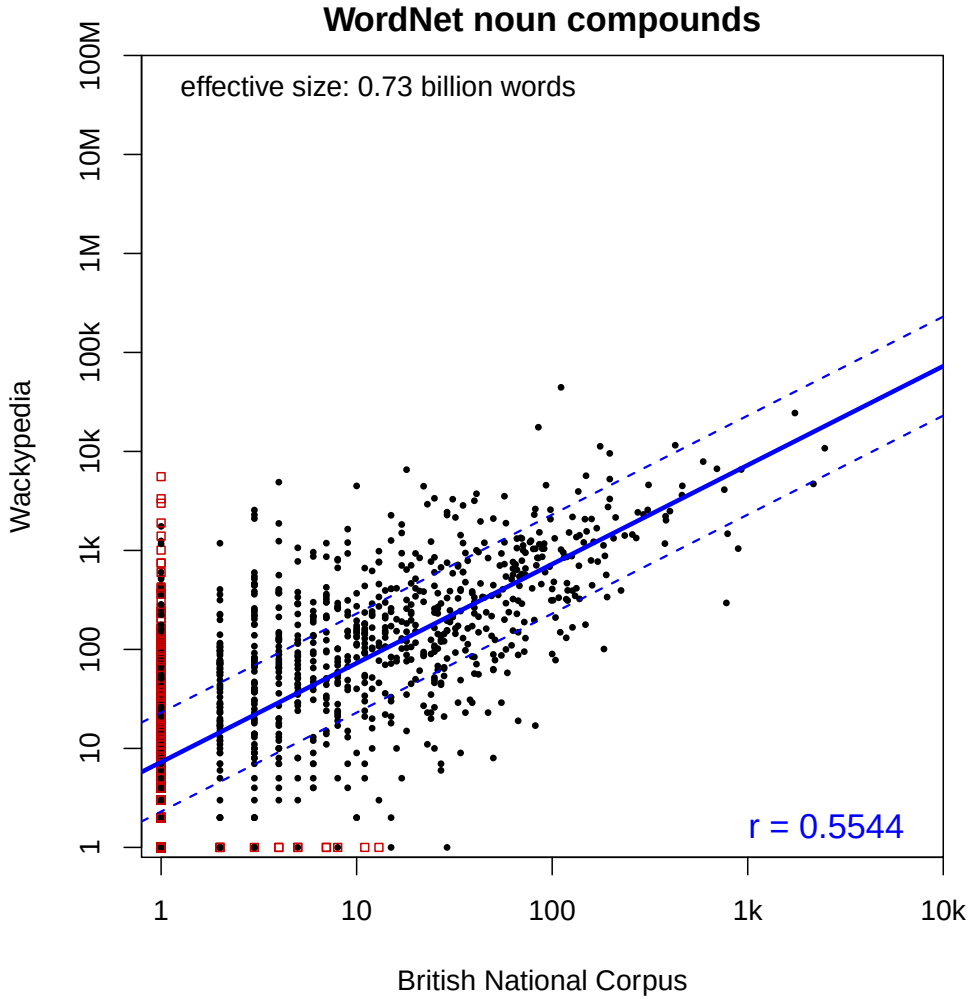
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



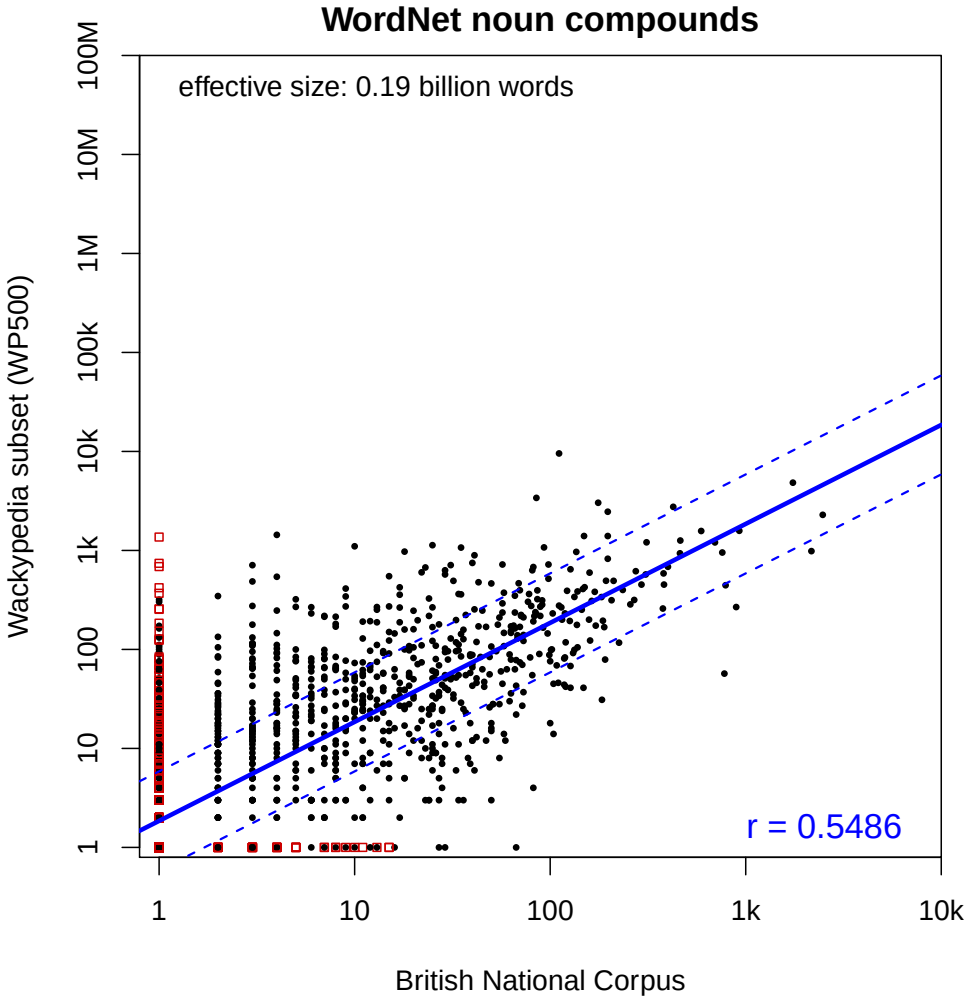
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



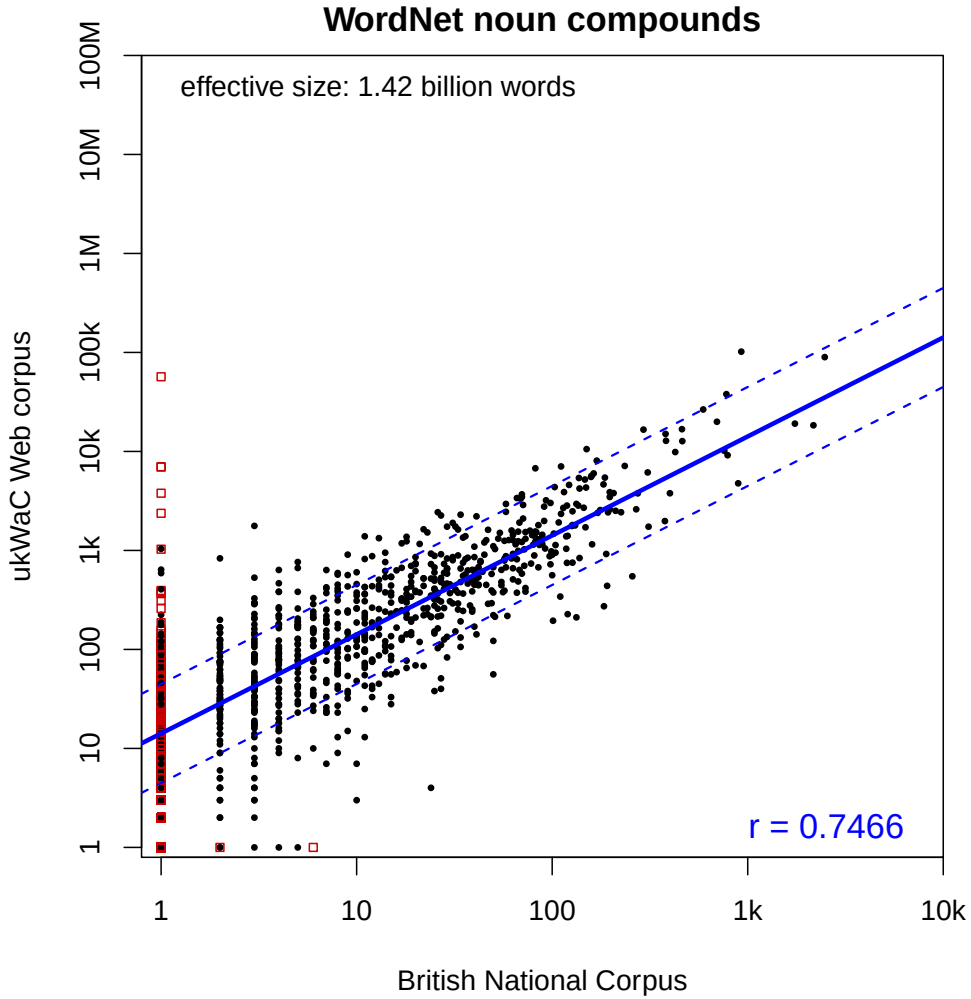
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



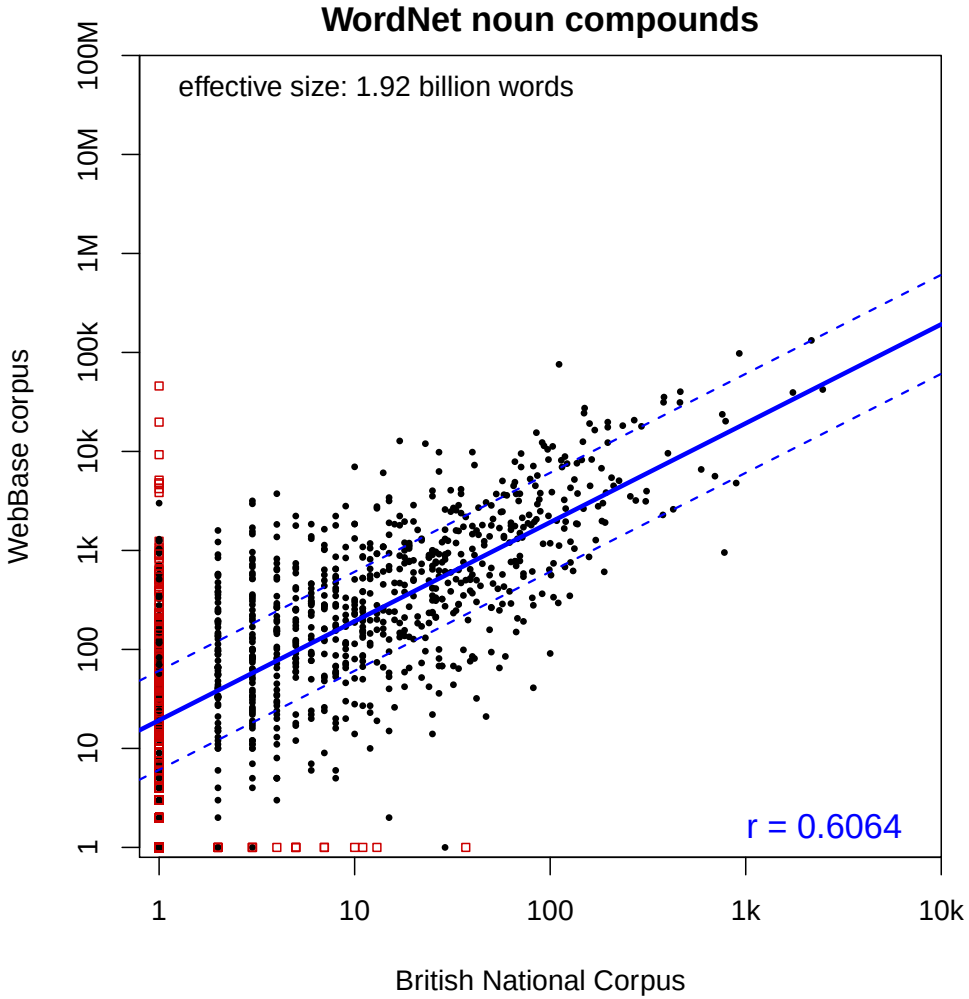
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



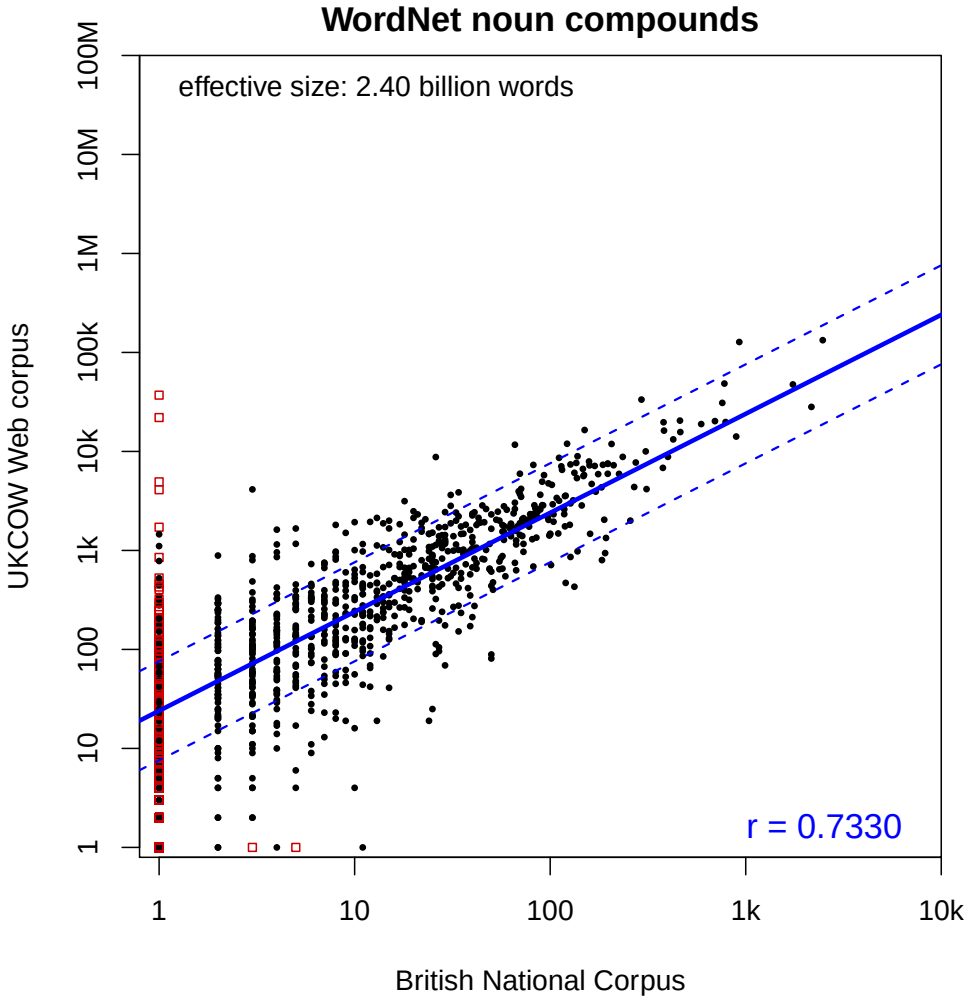
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



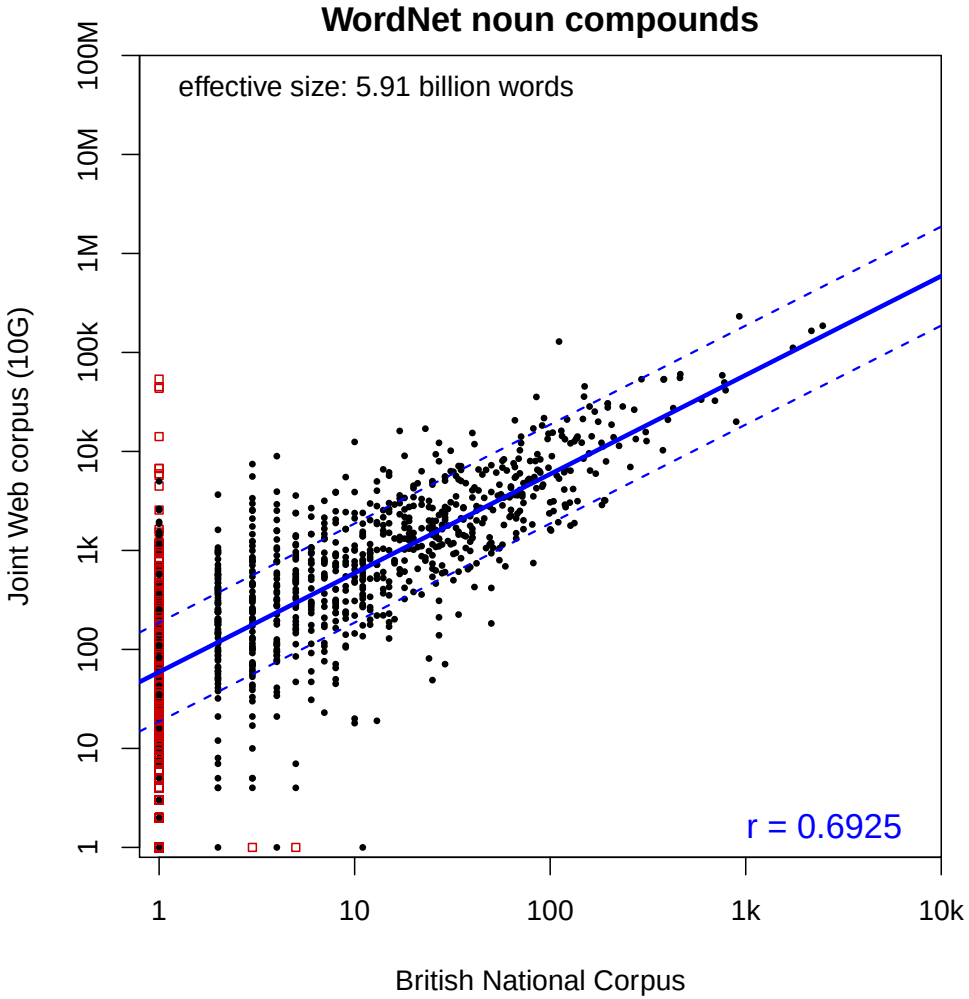
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



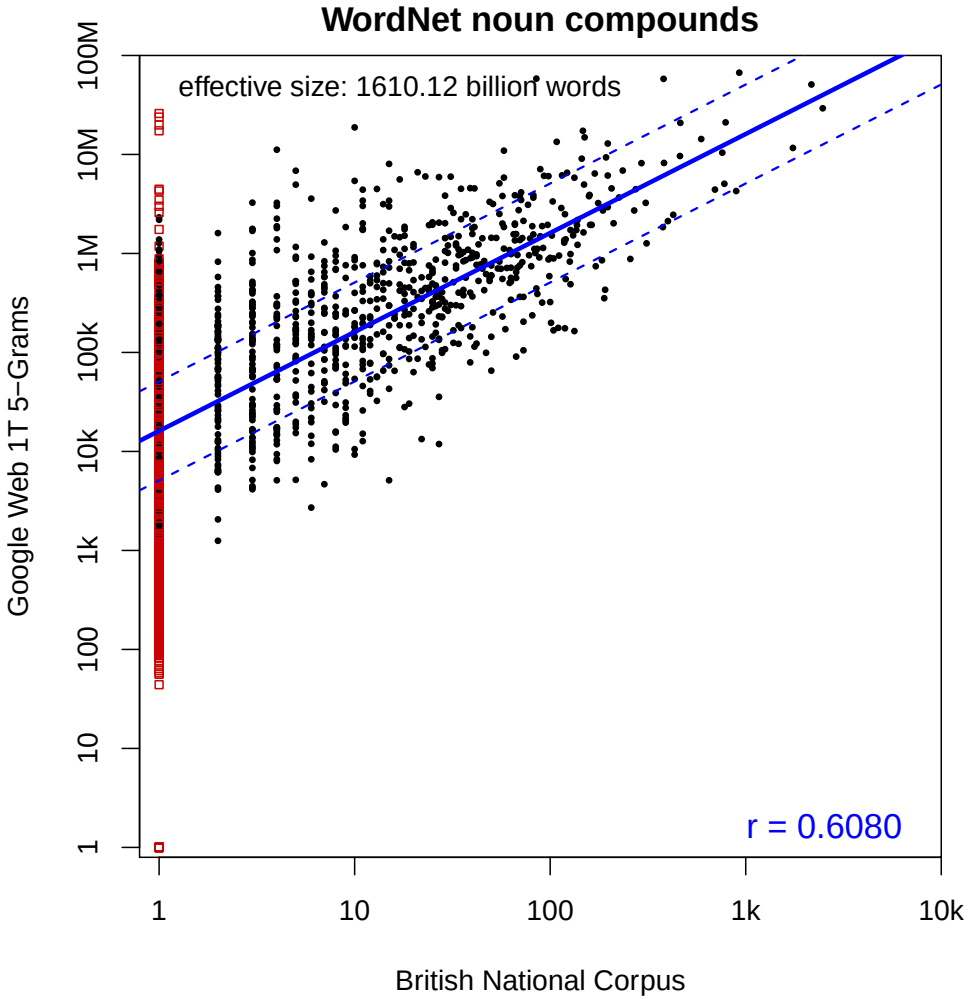
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



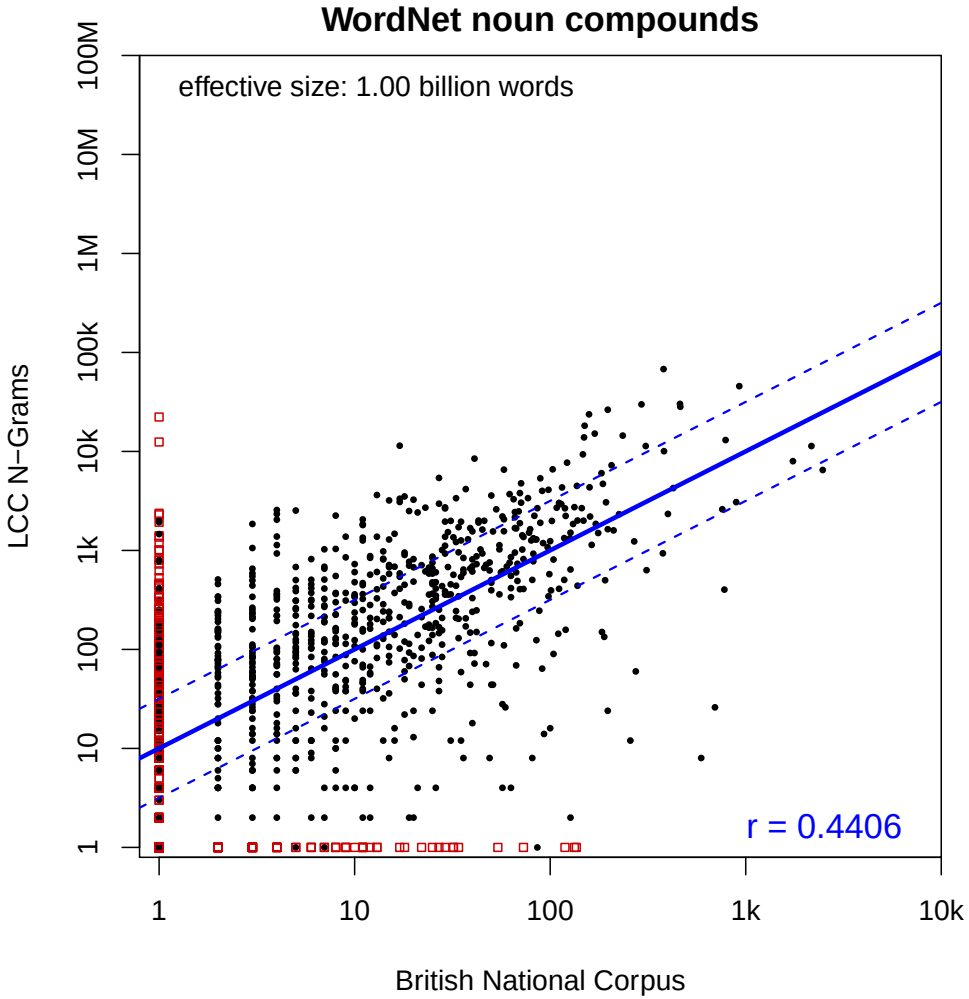
(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



(dashed lines indicate acceptable frequency difference within one order of magnitude)

Frequency comparison: compound nouns (WordNet)



(dashed lines indicate acceptable frequency difference within one order of magnitude)

MWE Identification & Collocations

Collocations

- Collocation: frequent co-occurrence within short span of up to 5 words (Firth 1957; Sinclair 1966, 1991)
  - plays important role in lexicography, corpus linguistics, language description, word sense disambiguation, ...
  - key feature for MWE identification
  - collocation database is also a sparse representation of a distributional semantic model (term-term matrix)

Collocations

- Collocation: frequent co-occurrence within short span of up to 5 words (Firth 1957; Sinclair 1966, 1991)
  - plays important role in lexicography, corpus linguistics, language description, word sense disambiguation, ...
  - key feature for MWE identification
  - collocation database is also a sparse representation of a distributional semantic model (term-term matrix)
- Web1T5 only provides exact co-occurrence frequencies for immediately adjacent bigrams (e.g. `* day` and `day *`)

Collocations

- Collocation: frequent co-occurrence within short span of up to 5 words (Firth 1957; Sinclair 1966, 1991)
  - plays important role in lexicography, corpus linguistics, language description, word sense disambiguation, ...
  - key feature for MWE identification
  - collocation database is also a sparse representation of a distributional semantic model (term-term matrix)

- Web1T5 only provides exact co-occurrence frequencies for immediately adjacent bigrams (e.g. `* day` and `day *`)
- Approximate counts for distance n from $n+1$-gram table

  `day ? ? *` and `* ? ? day`

  ➥ **quasi-collocations**

Quasi-collocations database

- Web1T5-Easy: pre-compiled database of quasi-collocations
  - brute-force, multi-pass algorithm
  - runtime approx. 3 days on server with 16 GiB RAM
- Flexible collocational span L4, . . . , L1 / R1, . . . , R4
  - separate count for each collocate and position
  - co-occurrence frequency in user-defined span and association scores are calculated on the fly
  - benefits from tight integration of Perl & SQLite
- Standard association measures: X^2, G^2, t, MI, Dice

Quasi-collocations demo

Collocates of “corpus” (f=5137372)

50 matches in 0.20 seconds

collocate	t-score	frequency	expected	span distribution (left, right)						
christi	1582.37	2504283	198.3		00%	01%	01%	97%	01%	00%
tx	794.93	639346	3725.8		00%	14%	02%	00%	16%	67%
habeas	720.32	518962	52.8		00%	00%	99%	00%	00%	00%
texas	629.04	411495	7978.1		06%	09%	02%	00%	22%	61%
columbus	429.55	186575	1034.0		48%	16%	36%	00%	00%	00%
dallas	390.37	156254	1943.7		00%	00%	00%	00%	70%	30%
writ	372.46	138960	116.1		98%	00%	00%	01%	00%	00%
callosum	368.99	136174	8.8		01%	00%	00%	98%	01%	00%
m	327.51	146346	21058.1		45%	46%	08%	00%	00%	00%
hotels	287.67	114198	16985.0		11%	15%	16%	00%	52%	05%
luteum	275.98	76176	5.7		02%	00%	00%	97%	01%	00%
oh	265.20	80036	5009.5		03%	04%	93%	00%	00%	00%

Evaluation on English VPC extraction task

(Baldwin 2008)

- English **verb-particle constructions** (VPC) consisting of head verb + one obligatory prepositional particle
  - *hand in*, *back off*, *wake up*, *set aside*, *carry on*, ...

Evaluation on English VPC extraction task

(Baldwin 2008)

- English **verb-particle constructions** (VPC) consisting of head verb + one obligatory prepositional particle
  - *hand in*, *back off*, *wake up*, *set aside*, *carry on*, ...
- Data set of 3,078 candidate VPC types
  - extracted from written part of BNC with combination of tagger-, chunker-, and parser-based methods

Evaluation on English VPC extraction task

(Baldwin 2008)

- English **verb-particle constructions** (VPC) consisting of head verb + one obligatory prepositional particle
  - *hand in*, *back off*, *wake up*, *set aside*, *carry on*, ...
- Data set of 3,078 candidate VPC types
  - extracted from written part of BNC with combination of tagger-, chunker-, and parser-based methods
- Manually annotated as compositional / non-compositional
  - baseline: **14.3%** non-compositional VPC (440 / 3078)
  - compositional: *carry around*, *fly away*, *refer back*, ...
  - further distinction of transitive/intransitive VPC not used

Evaluation on English VPC extraction task

(Baldwin 2008)

- English **verb-particle constructions** (VPC) consisting of head verb + one obligatory prepositional particle
  - *hand in*, *back off*, *wake up*, *set aside*, *carry on*, ...
- Data set of 3,078 candidate VPC types
  - extracted from written part of BNC with combination of tagger-, chunker-, and parser-based methods
- Manually annotated as compositional / non-compositional
  - baseline: **14.3%** non-compositional VPC (440 / 3078)
  - compositional: *carry around*, *fly away*, *refer back*, ...
  - further distinction of transitive/intransitive VPC not used
- Evaluation: candidate ranking based on each corpus
  - surface co-occcurrence (L0,R3) + POS filter (except Web1T5)
  - standard association measures: G^2, t, MI, Dice, X^2, f
  - precision/recall graphs; overall quality: average precision (AP)

Evaluation on English VPC extraction task

(Baldwin 2008)

verb	particle	X^2	TP
bandy	about	4857.75	−
chat	away	1105.58	−
chat	on	7.30	−
fish	out	1535.60	+
improve	by	576.06	−
move	forth	231.54	−
play	around	394.98	−
scrape	through	278.53	−
shuffle	out	85.15	−
start	over	75.88	+
transfer	to	13934.37	−
weld	on	30.83	−

Evaluation on English VPC extraction task

(Baldwin 2008)

verb	particle	$X^2 \downarrow$	TP
talk	about	950906.9	−
lean	forward	510113.9	−
want	to	477739.8	−
sort	out	406072.7	+
base	on	398035.3	−
depend	on	330956.6	−
sit	down	329143.2	+
go	to	289818.9	−
slow	down	282418.3	+
lag	behind	257224.2	+
be	by	242827.7	−
set	aside	242238.1	+

Evaluation on English VPC extraction task

(Baldwin 2008)

verb	particle	$X^2 \downarrow$	TP
talk	about	950906.9	−
lean	forward	510113.9	−
want	to	477739.8	−
sort	out	406072.7	+
base	on	398035.3	−
depend	on	330956.6	−
sit	down	329143.2	+
go	to	289818.9	−
slow	down	282418.3	+
lag	behind	257224.2	+
be	by	242827.7	−
set	aside	242238.1	+

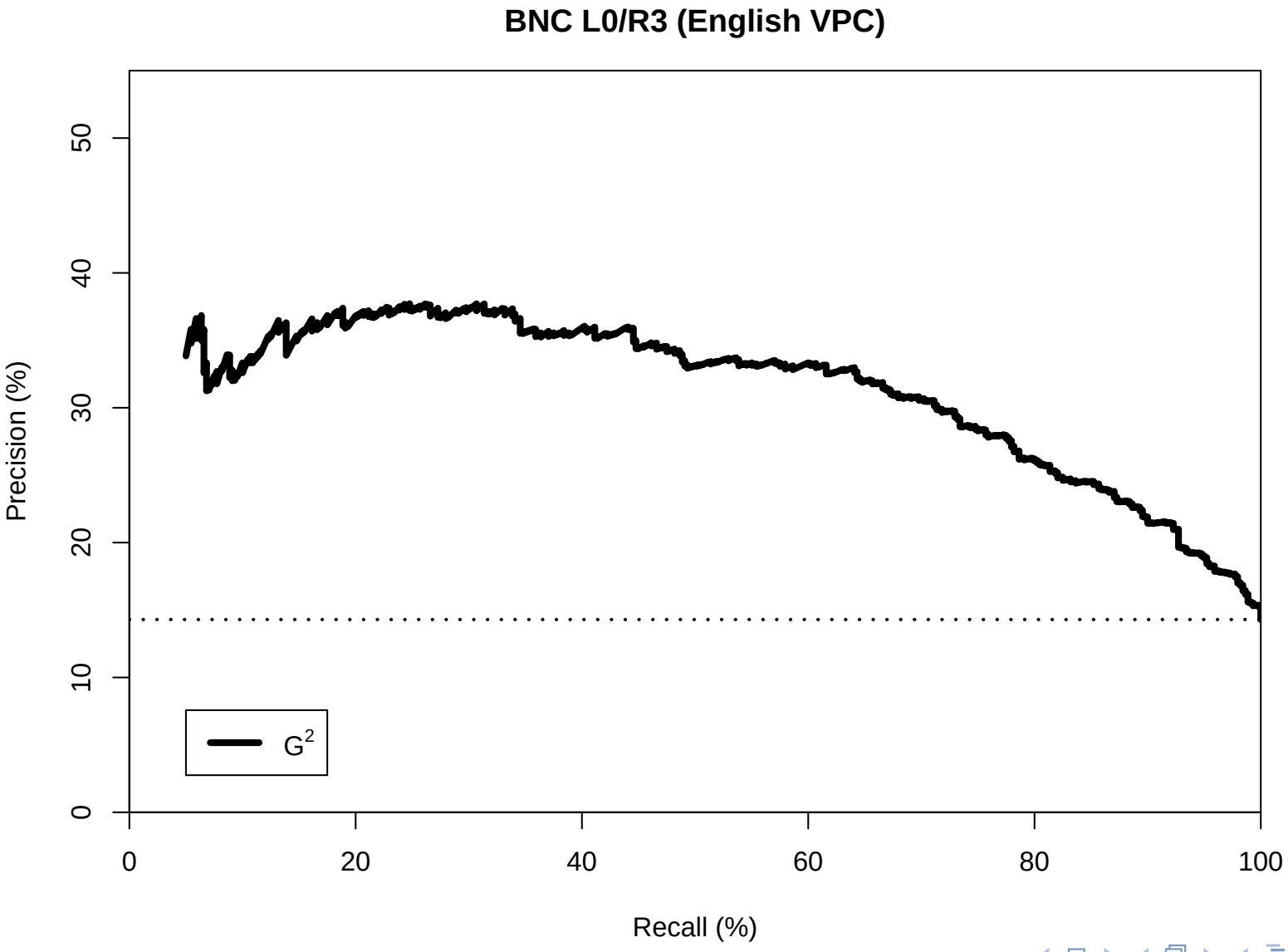
n-best list ($n = 12$)

$P = \frac{5}{12} = 41.7\%$

$R = \frac{5}{440} = 1.1\%$

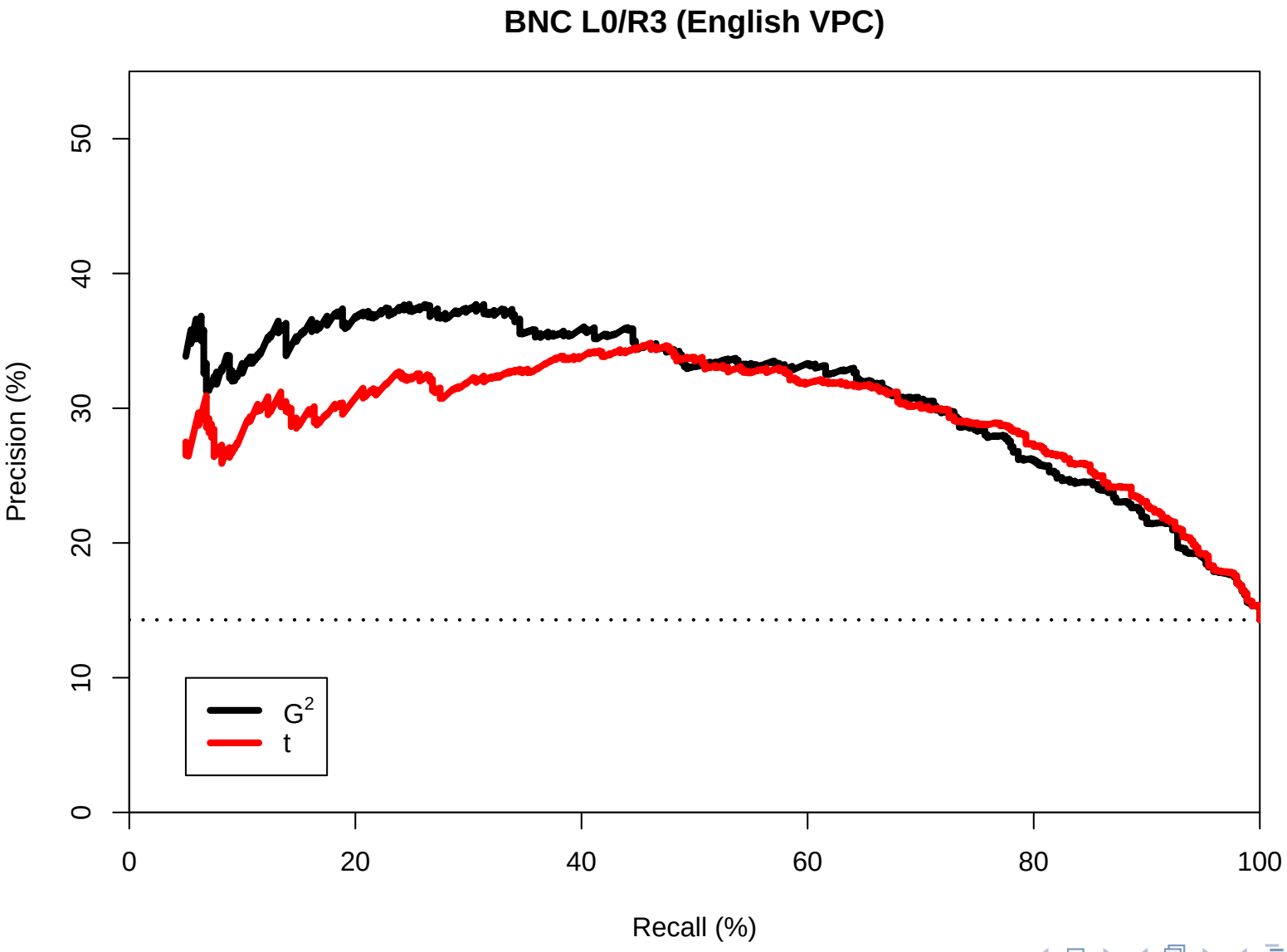
Evaluation on English VPC extraction task

(Baldwin 2008)



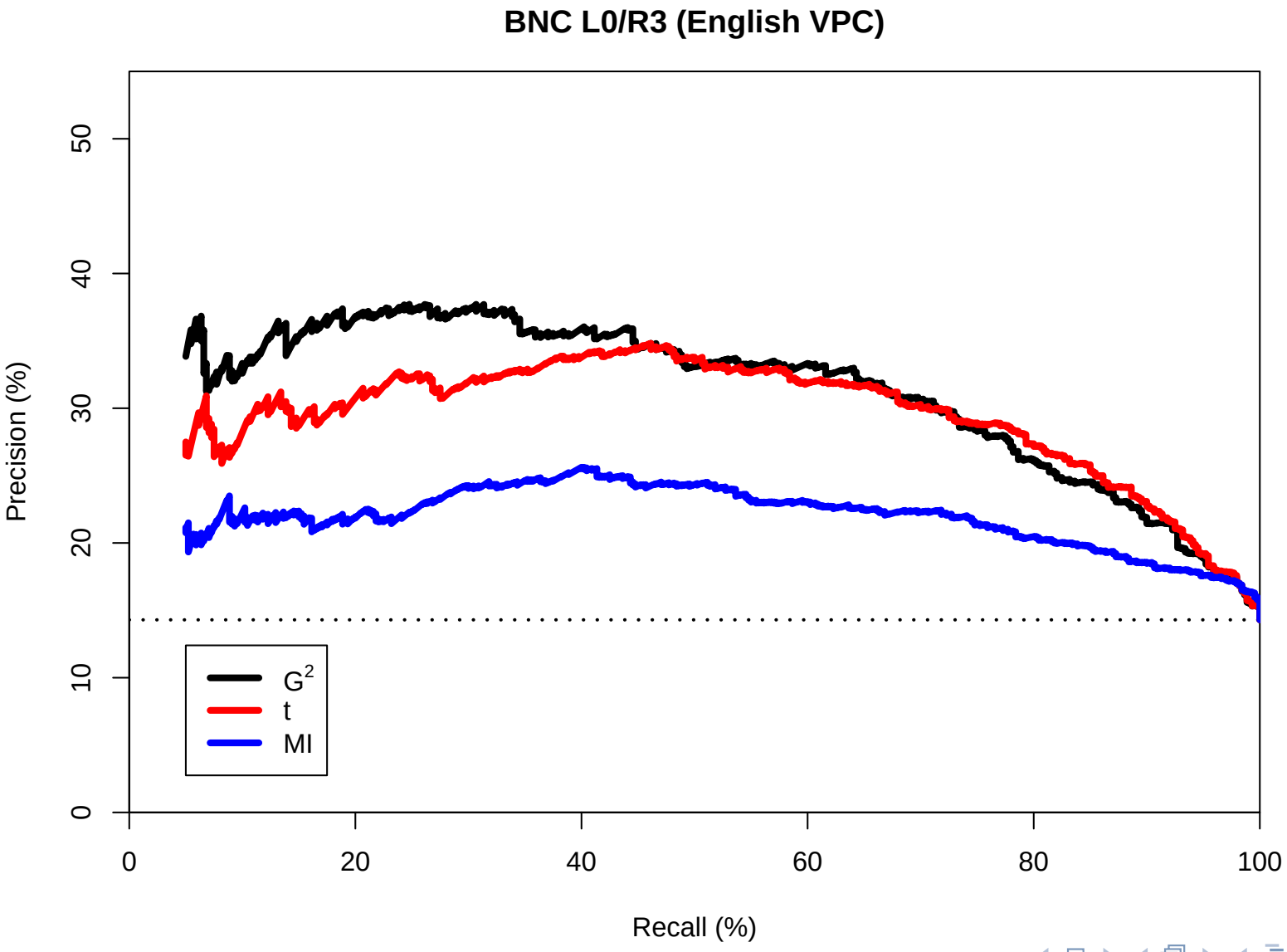
Evaluation on English VPC extraction task

(Baldwin 2008)



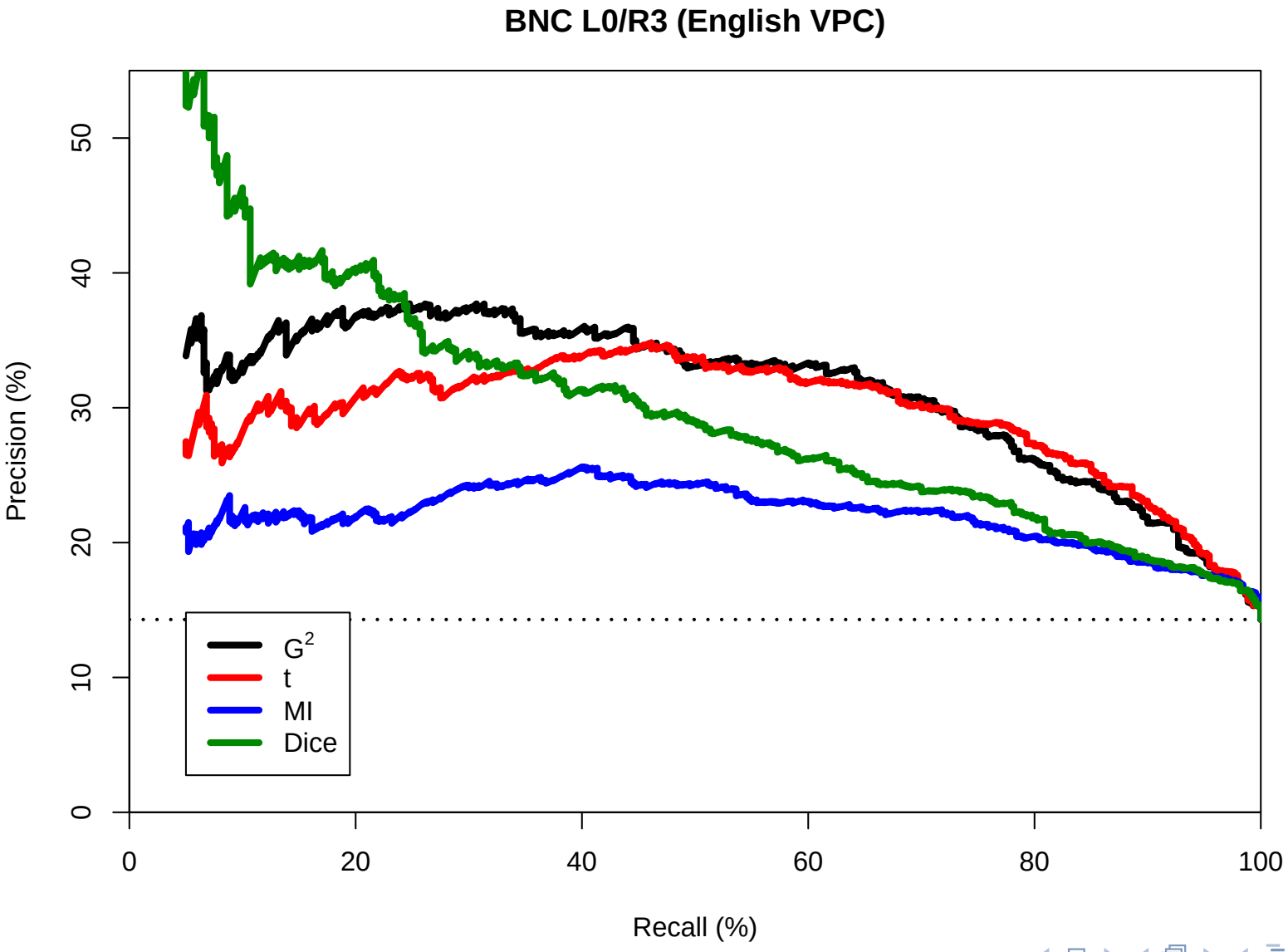
Evaluation on English VPC extraction task

(Baldwin 2008)



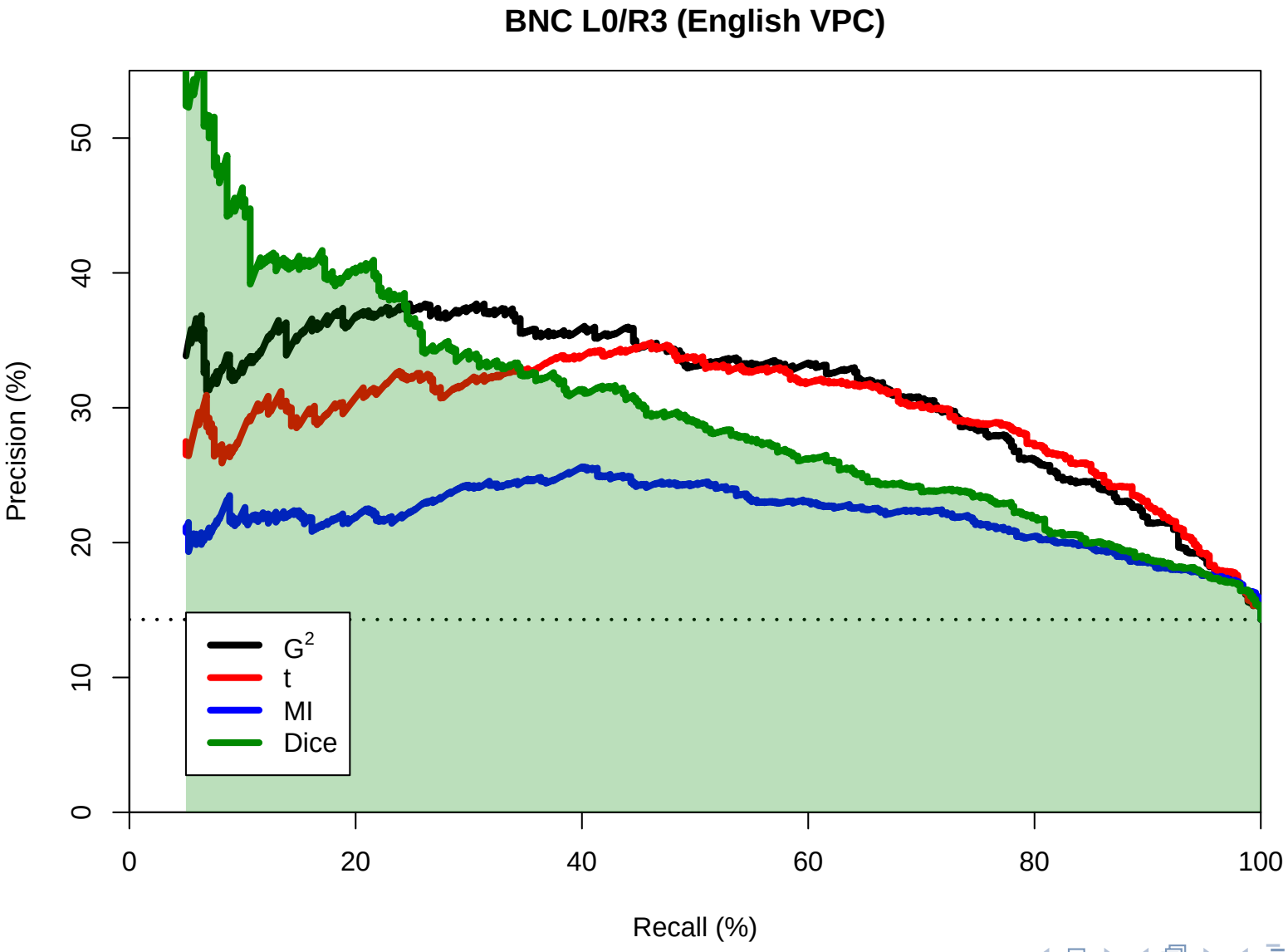
Evaluation on English VPC extraction task

(Baldwin 2008)



Evaluation on English VPC extraction task

(Baldwin 2008)

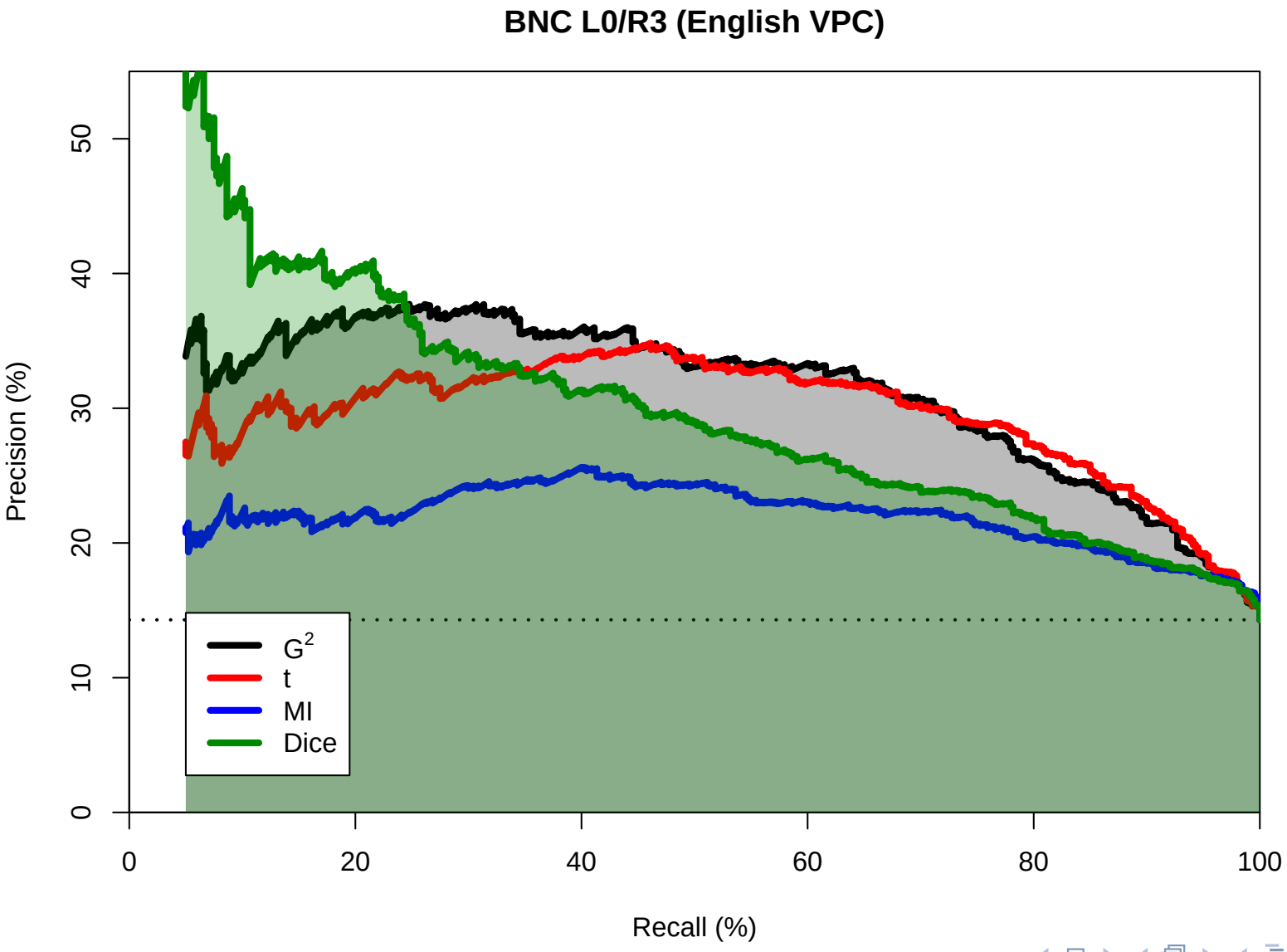


Average Precision

Dice: 30.13%

Evaluation on English VPC extraction task

(Baldwin 2008)



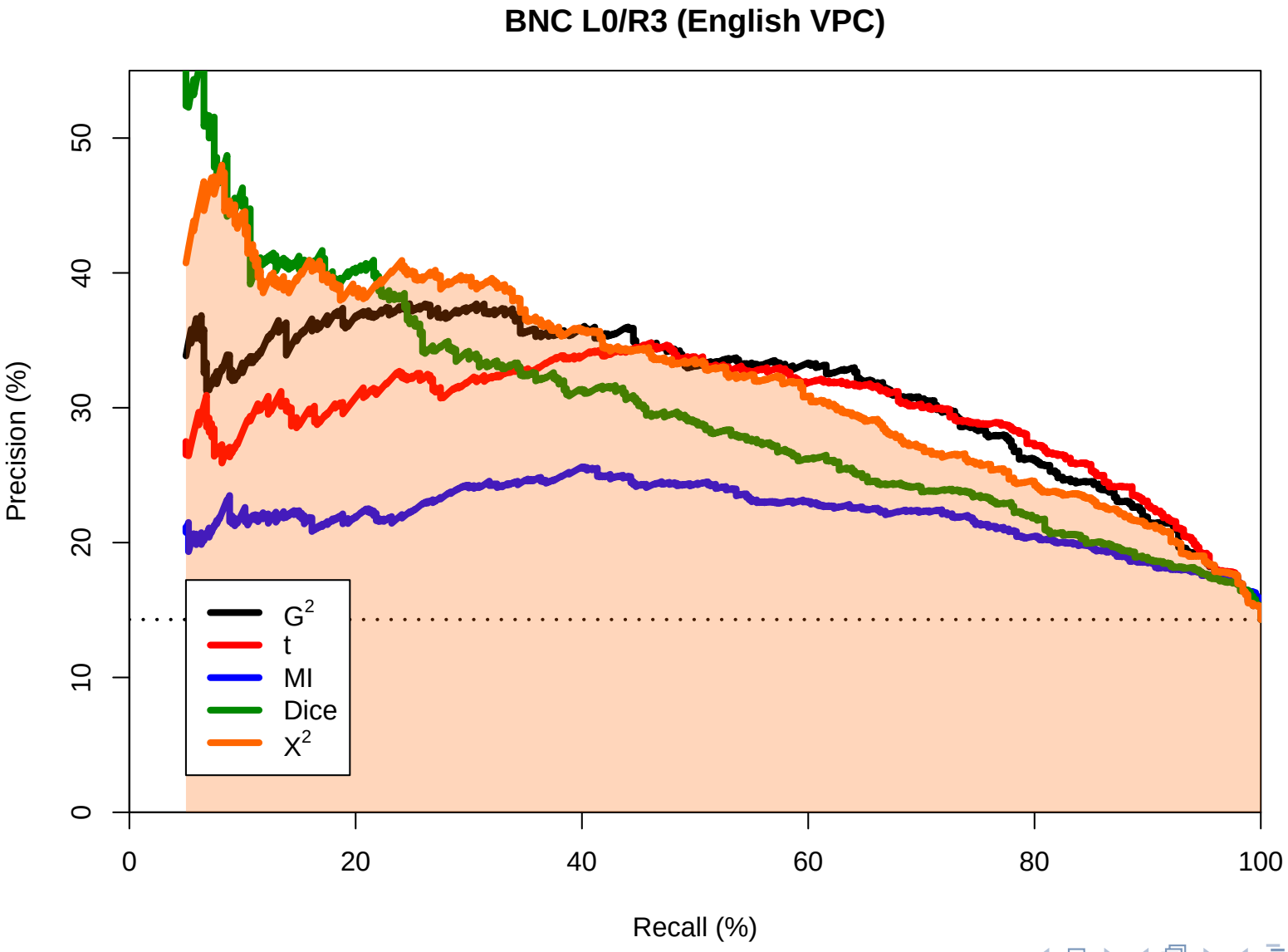
Average Precision

G^2: 31.06%

Dice: 30.13%

Evaluation on English VPC extraction task

(Baldwin 2008)



Average Precision

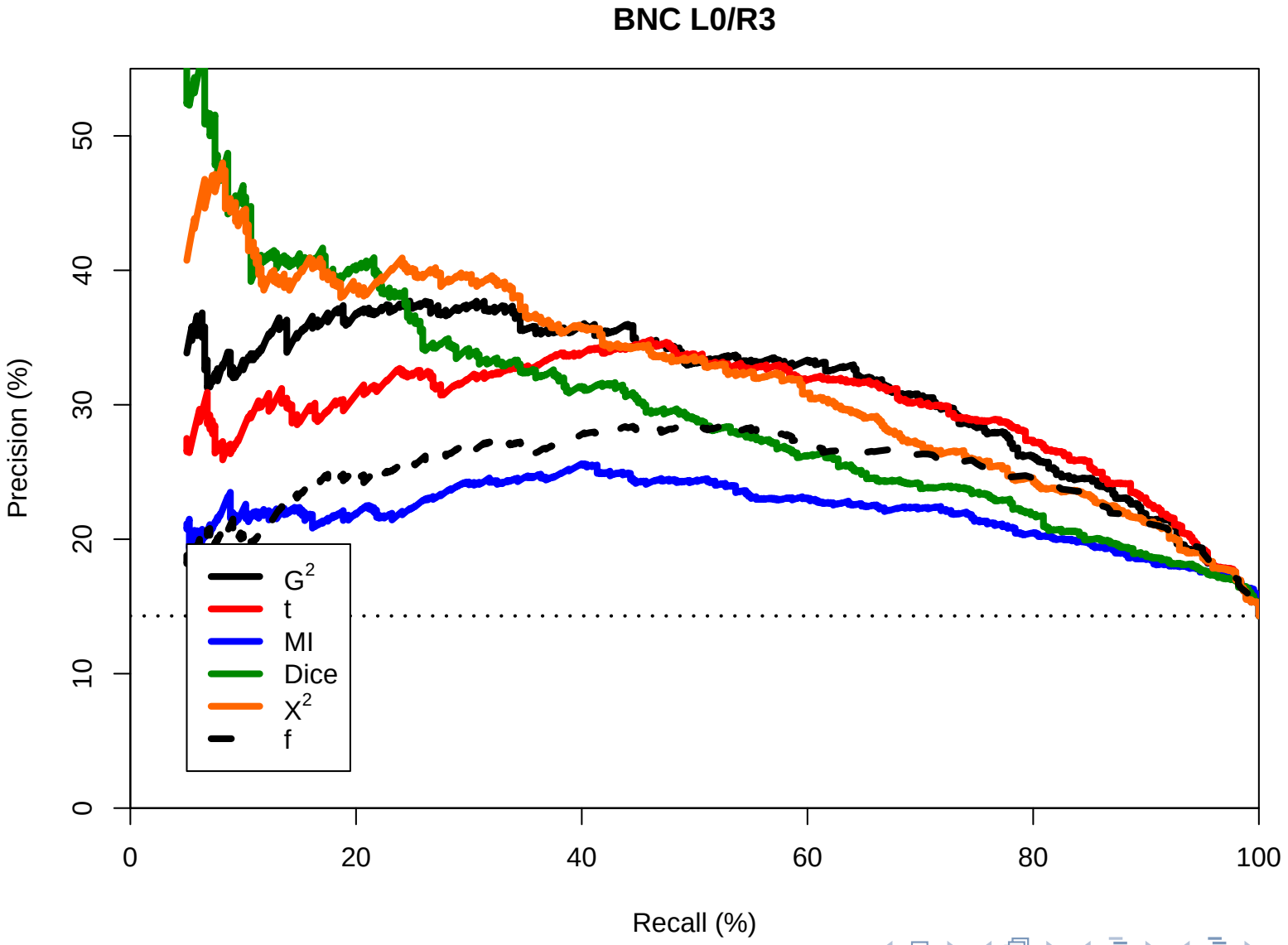
G^2: 31.06%

Dice: 30.13%

X^2: 32.12%

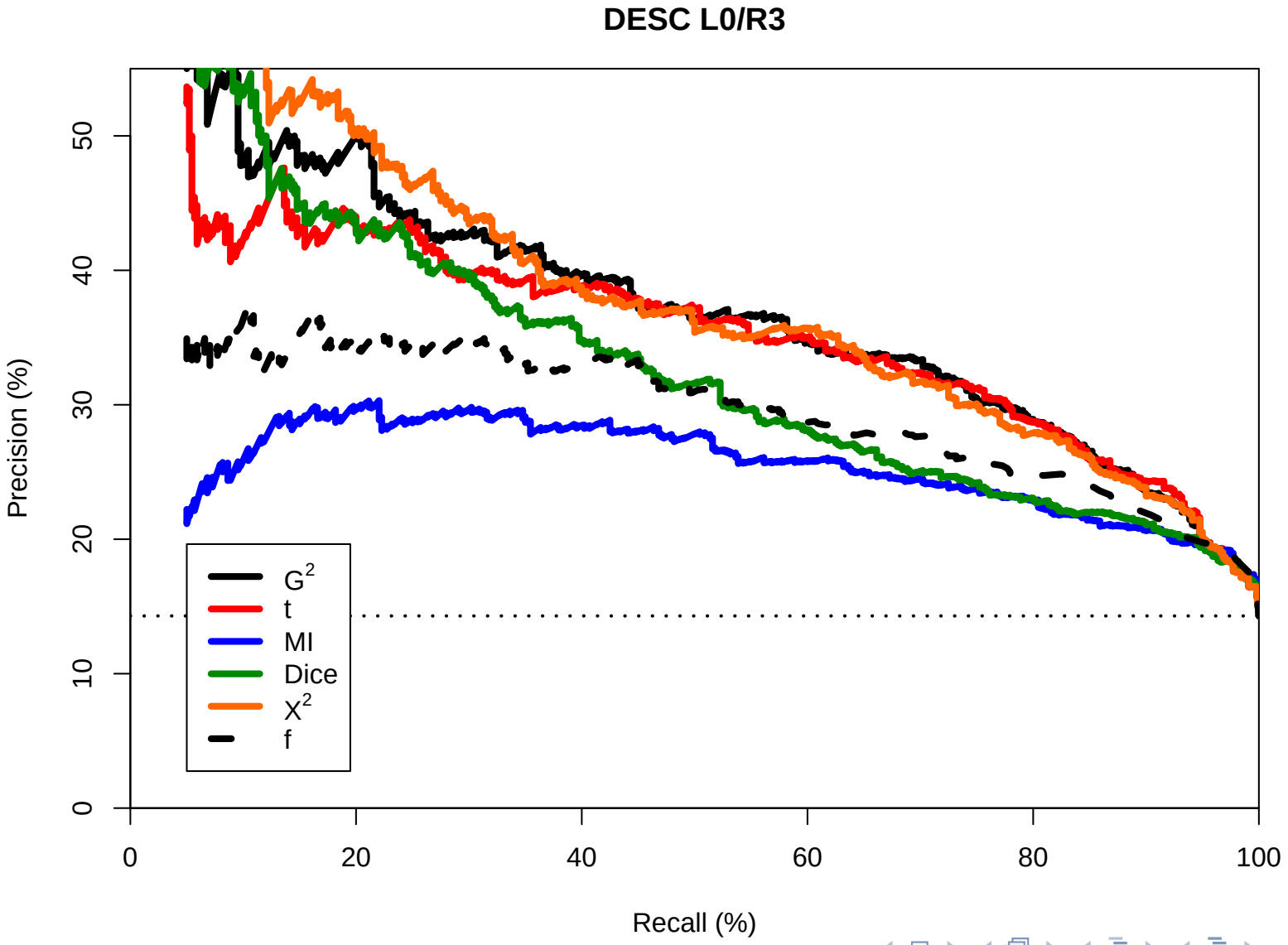
Evaluation on English VPC extraction task

(Baldwin 2008)



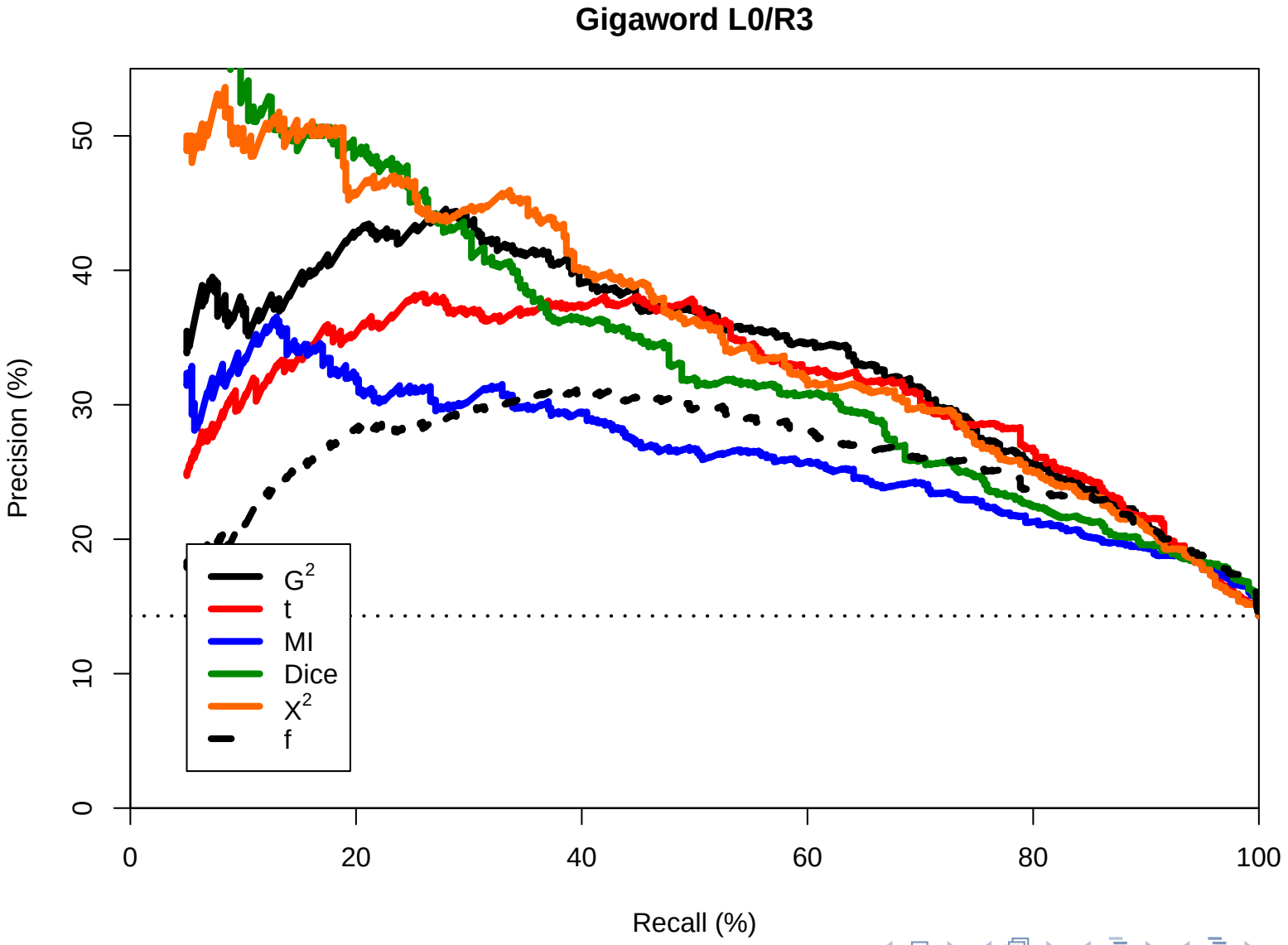
Evaluation on English VPC extraction task

(Baldwin 2008)



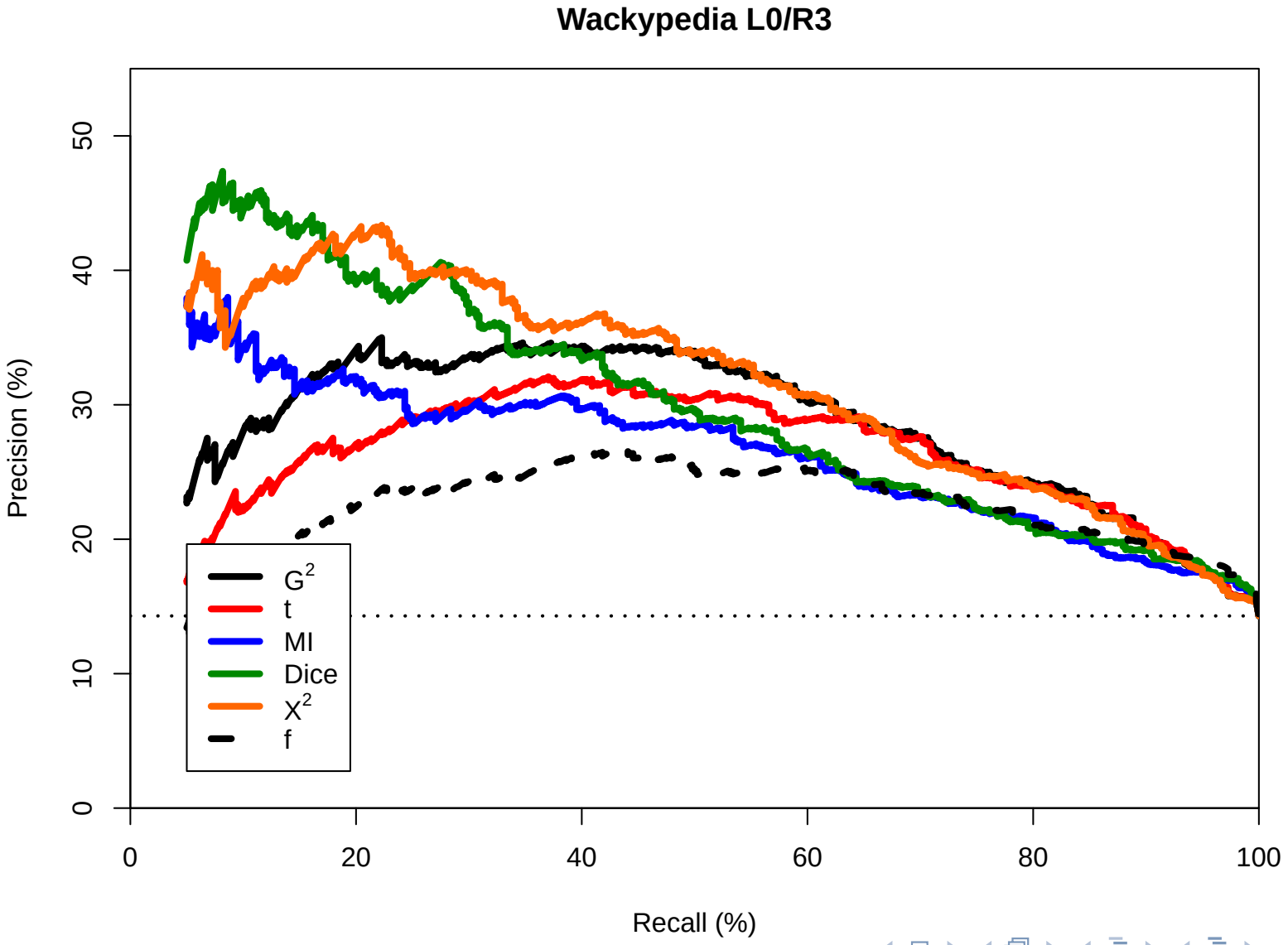
Evaluation on English VPC extraction task

(Baldwin 2008)



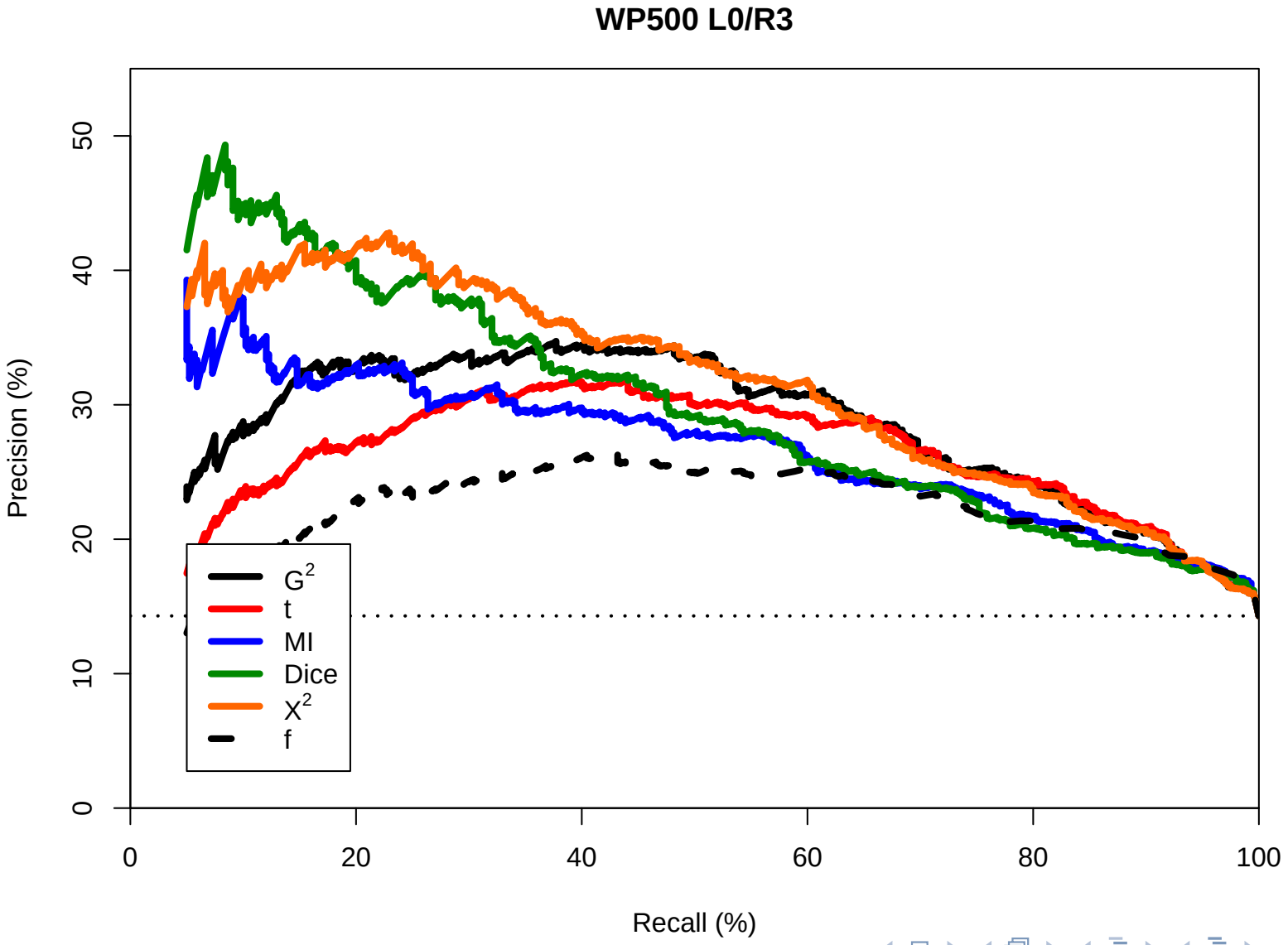
Evaluation on English VPC extraction task

(Baldwin 2008)



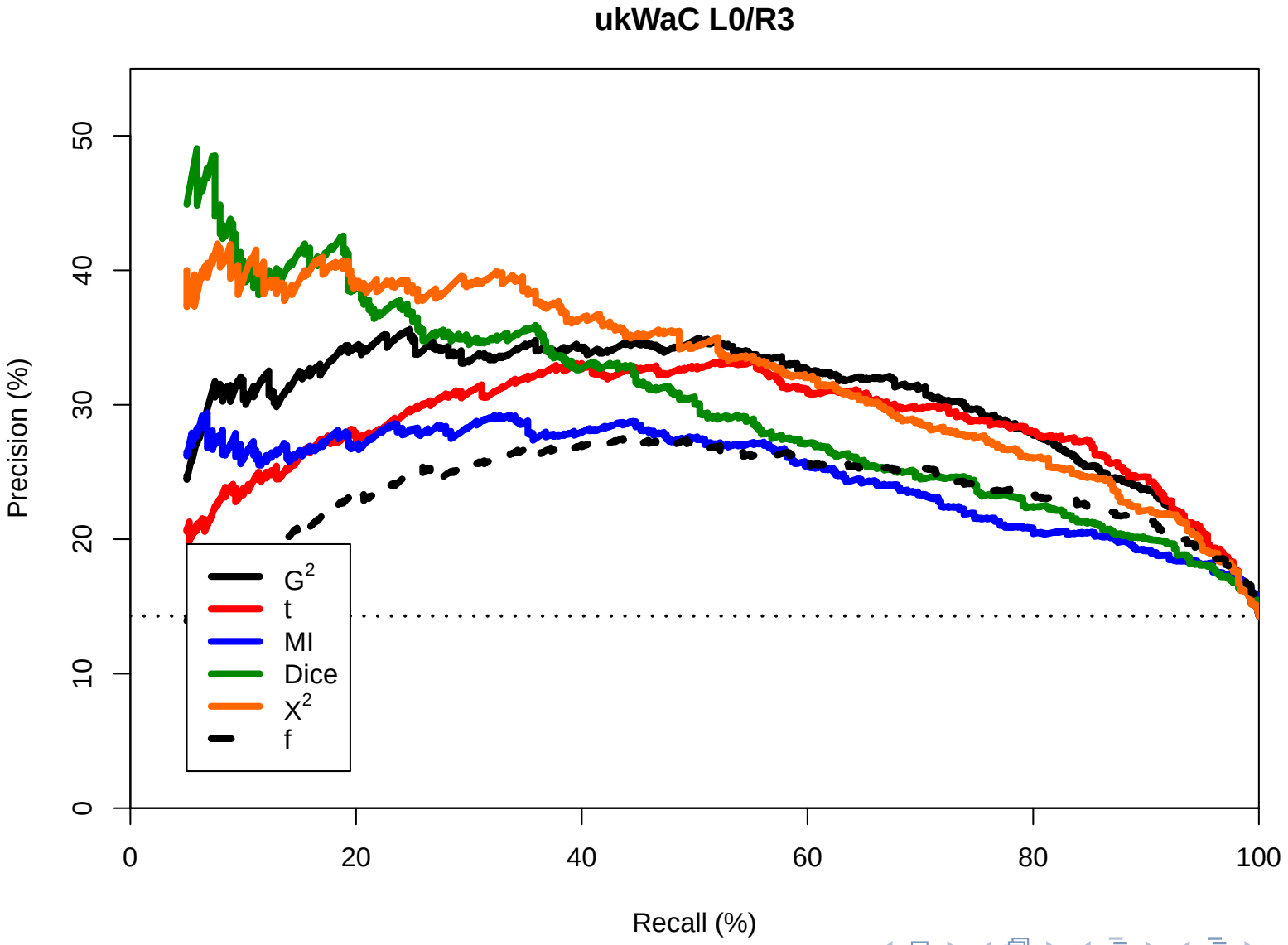
Evaluation on English VPC extraction task

(Baldwin 2008)



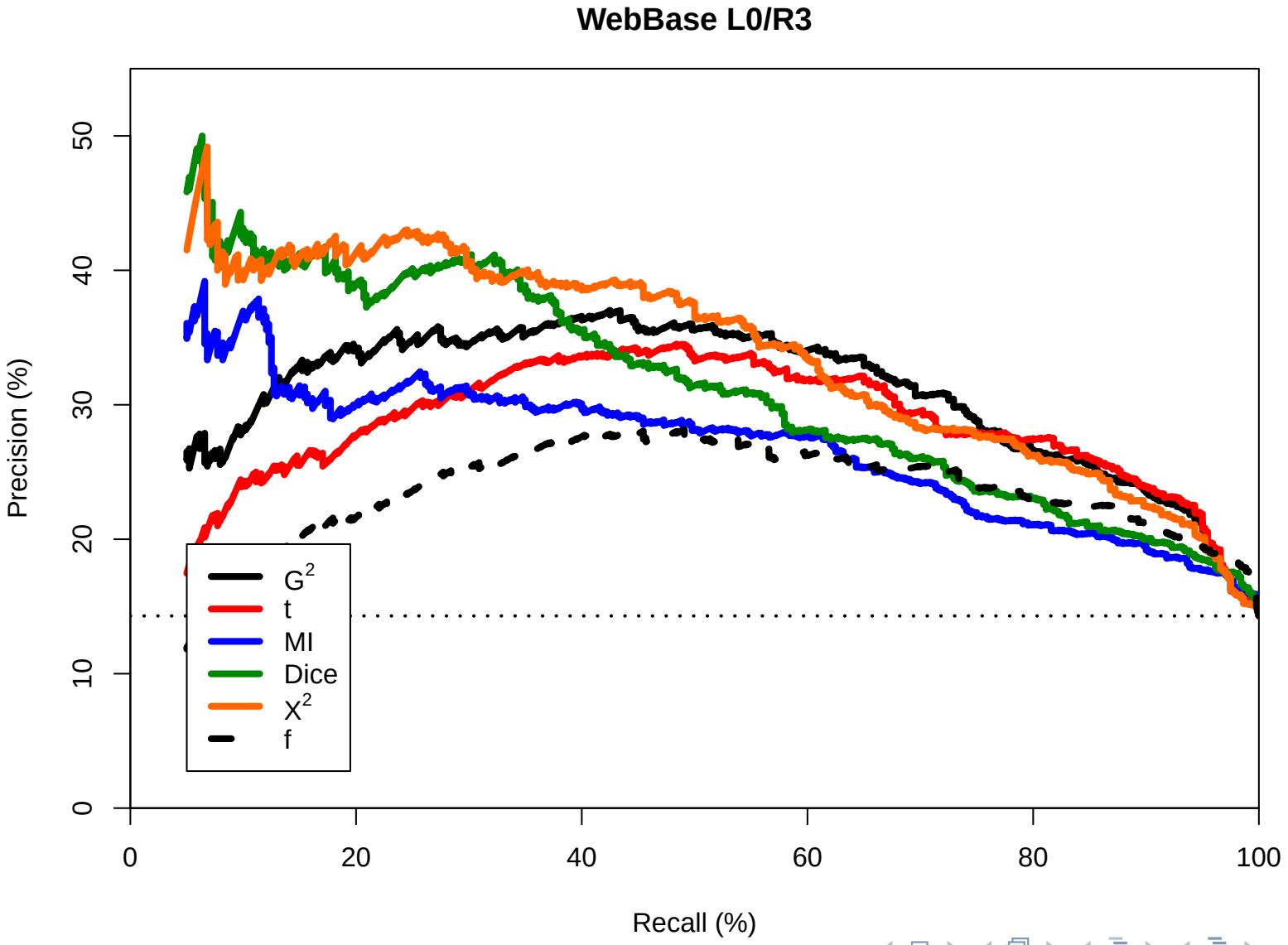
Evaluation on English VPC extraction task

(Baldwin 2008)



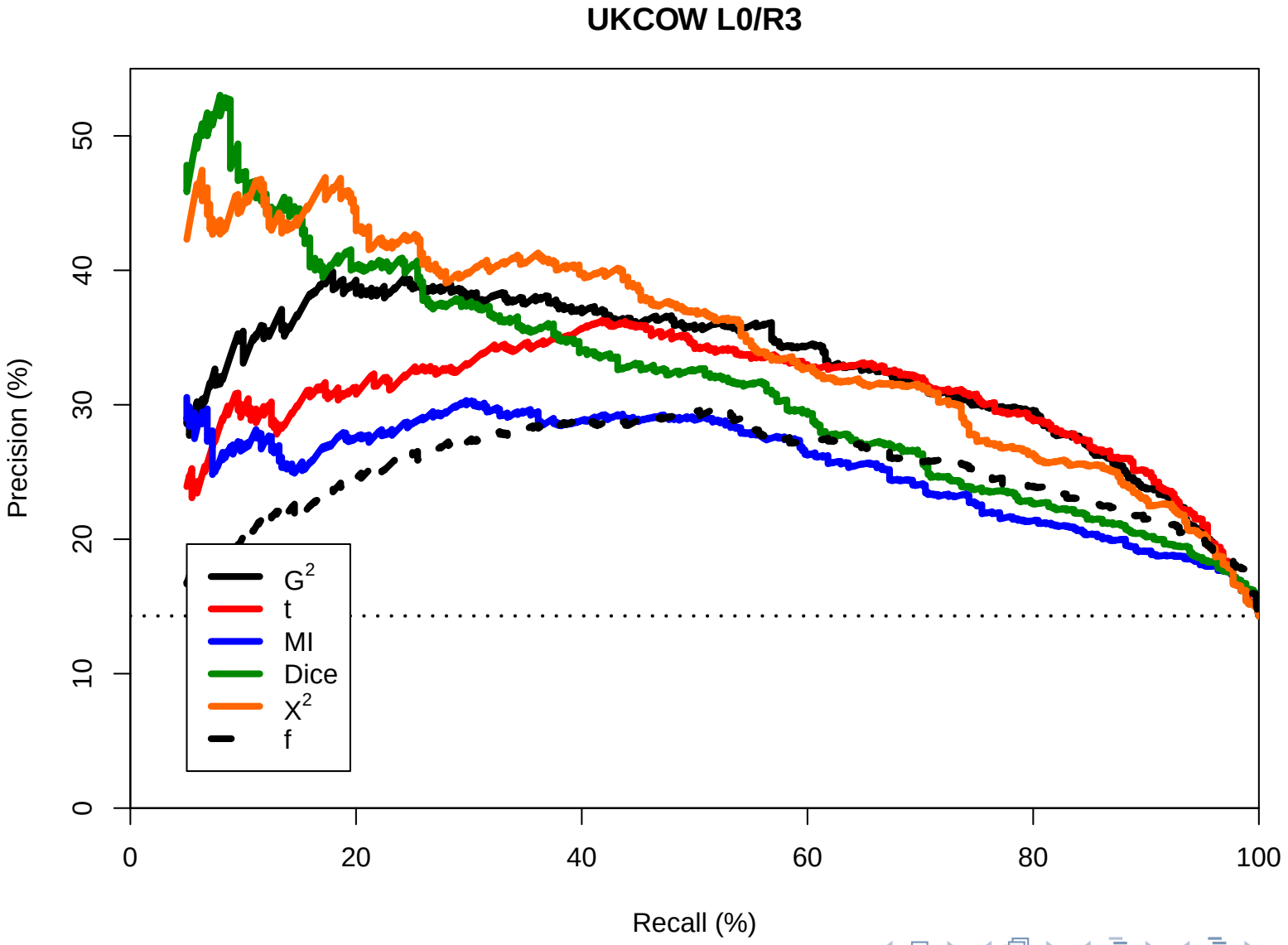
Evaluation on English VPC extraction task

(Baldwin 2008)



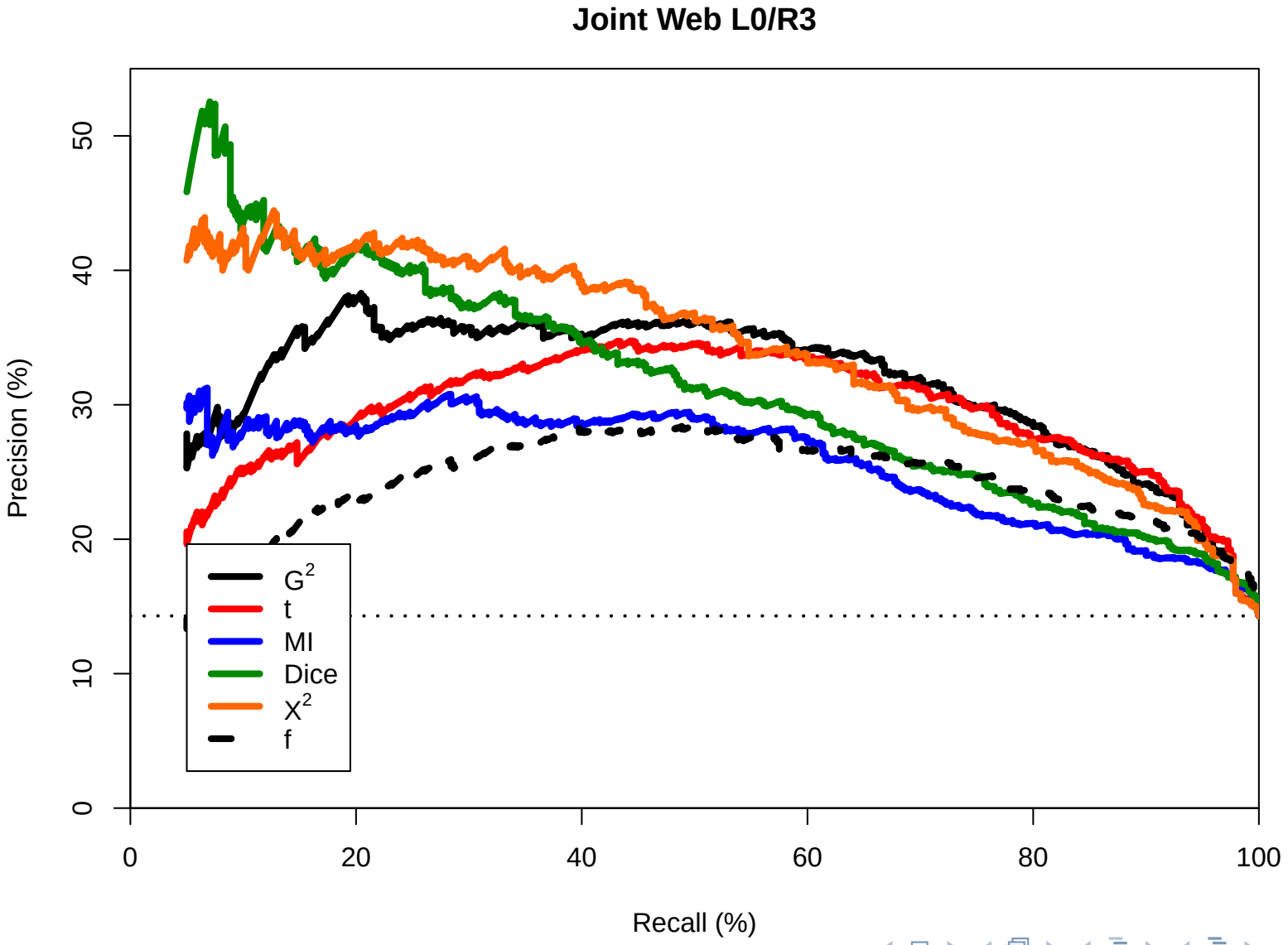
Evaluation on English VPC extraction task

(Baldwin 2008)



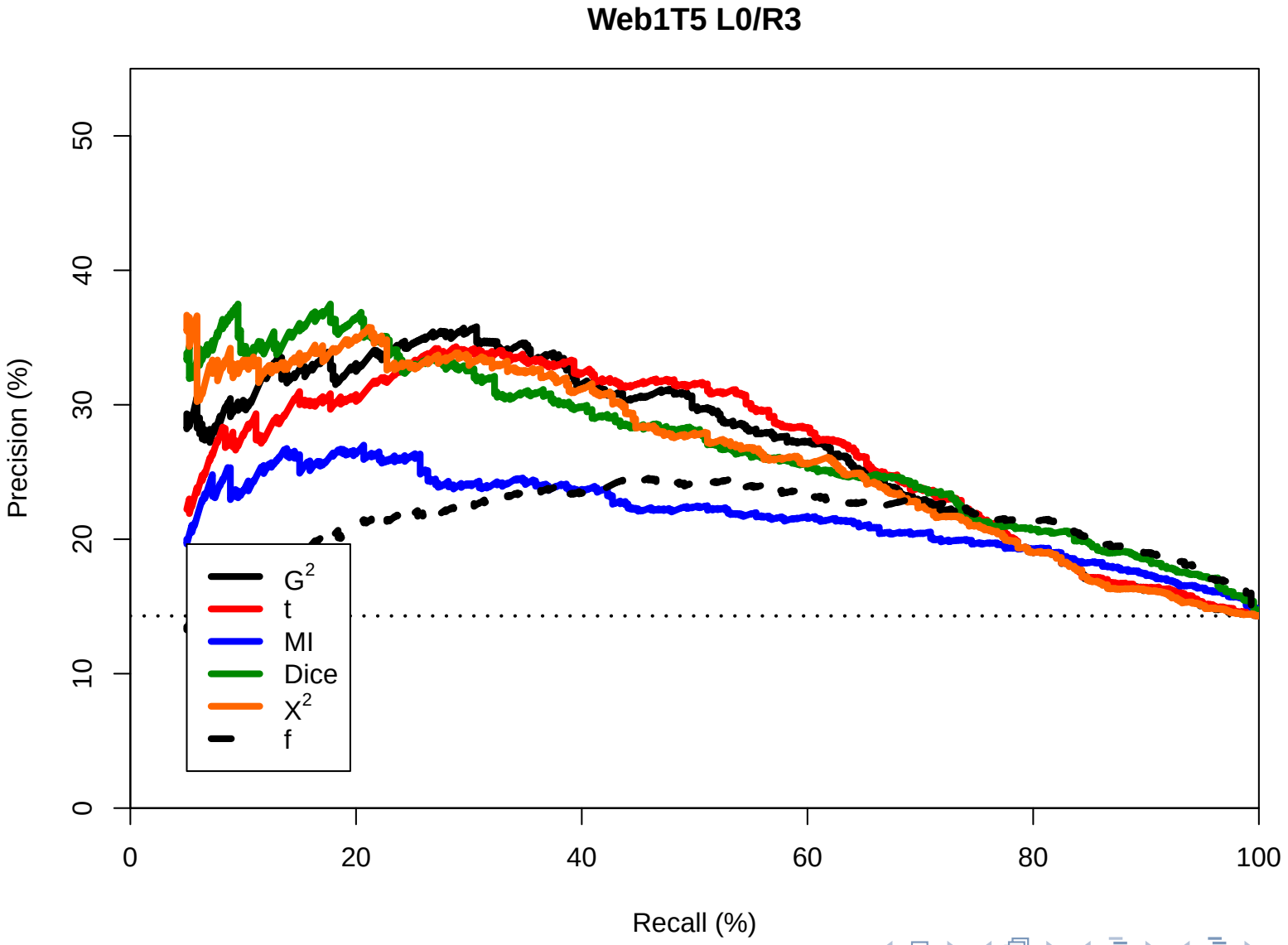
Evaluation on English VPC extraction task

(Baldwin 2008)



Evaluation on English VPC extraction task

(Baldwin 2008)



Evaluation on English VPC extraction task

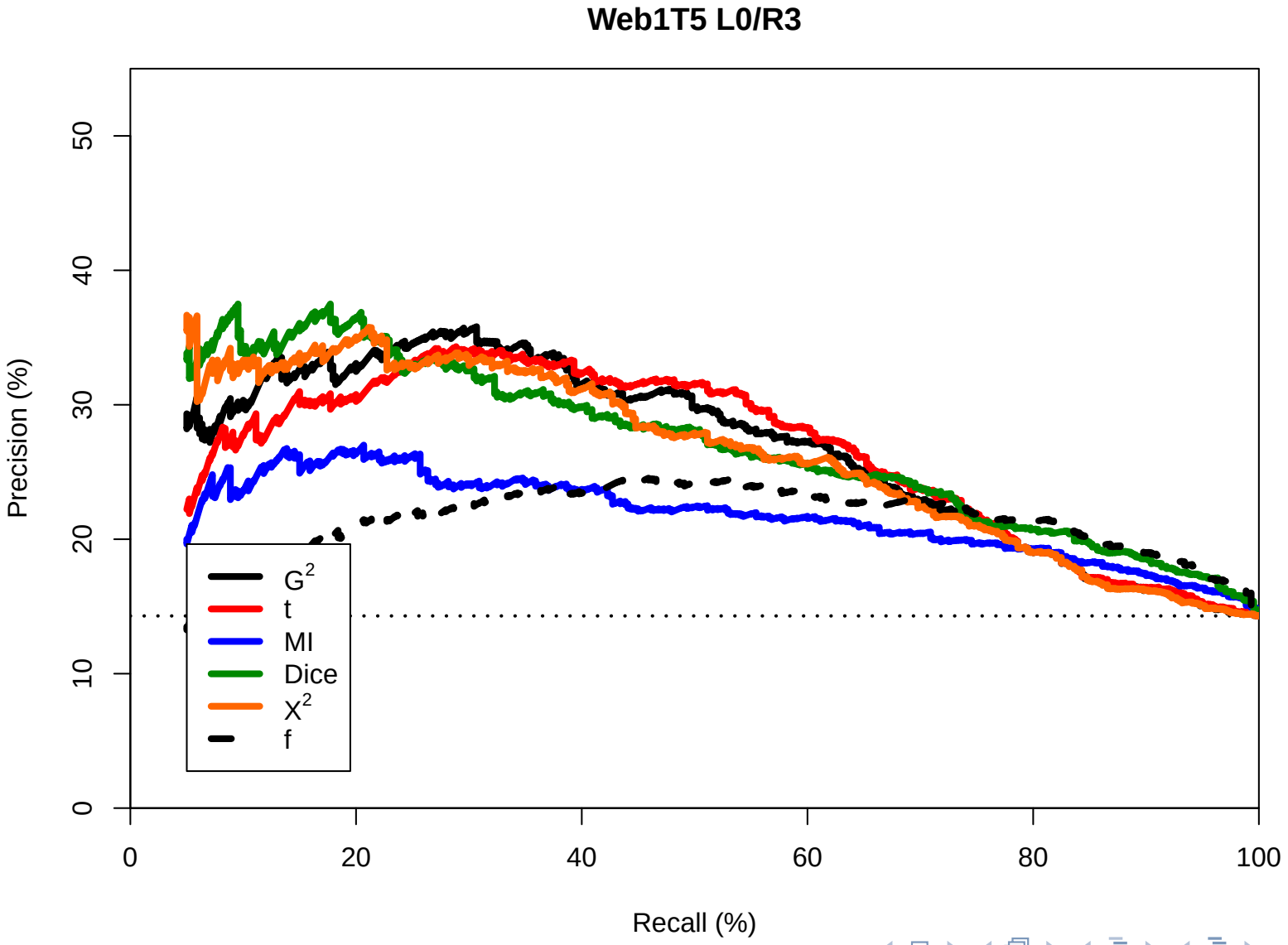
(Baldwin 2008)

Why does Web1T5 perform so badly in this task despite its size? Possible explanations include:

- Co-occurrence counts are underestimated for larger windows because of frequency threshold in n-gram database
  ➜ quasi-collocations as poor approximation
- No part-of-speech annotation available to filter candidates

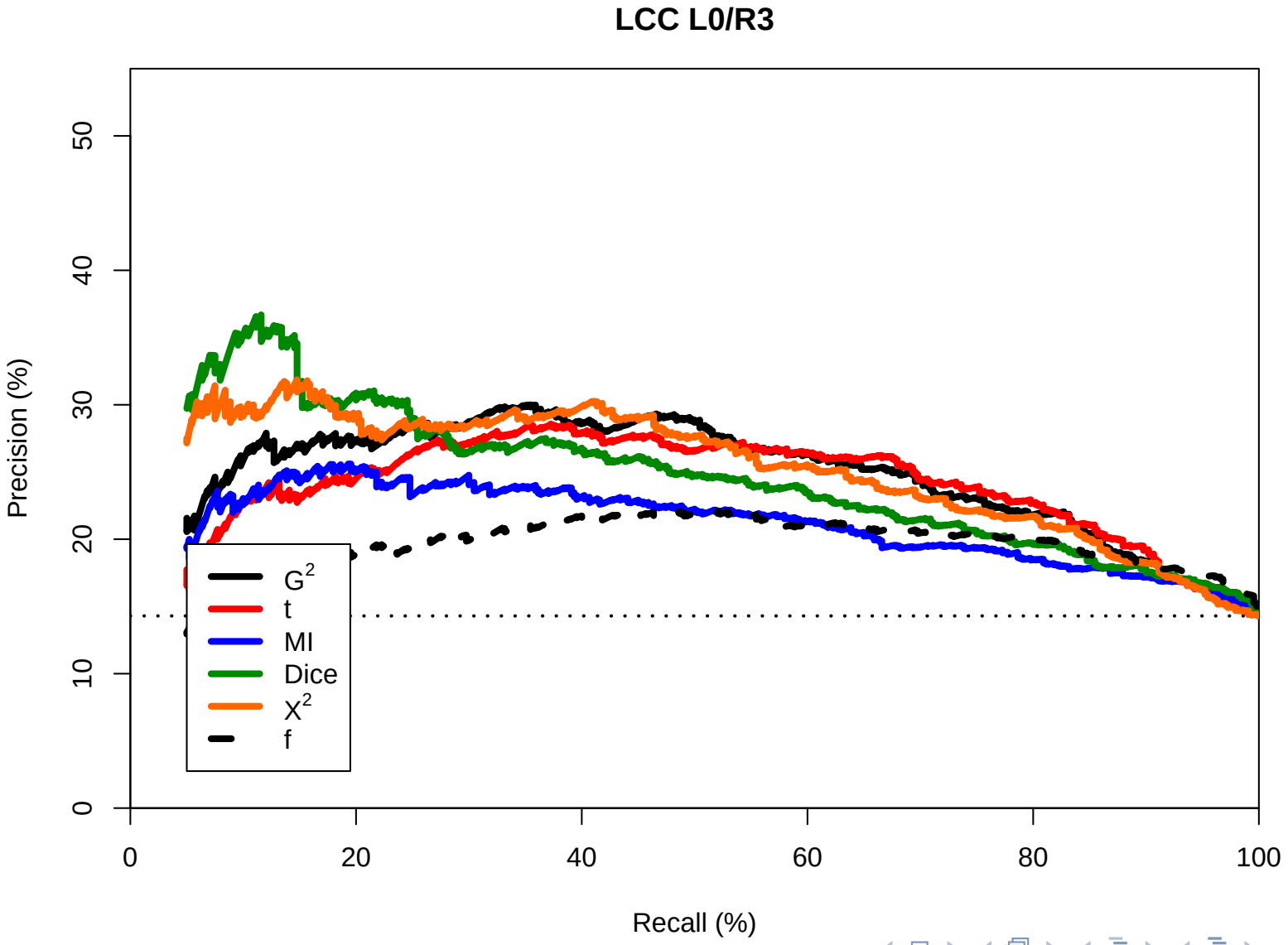
Evaluation on English VPC extraction task

(Baldwin 2008)



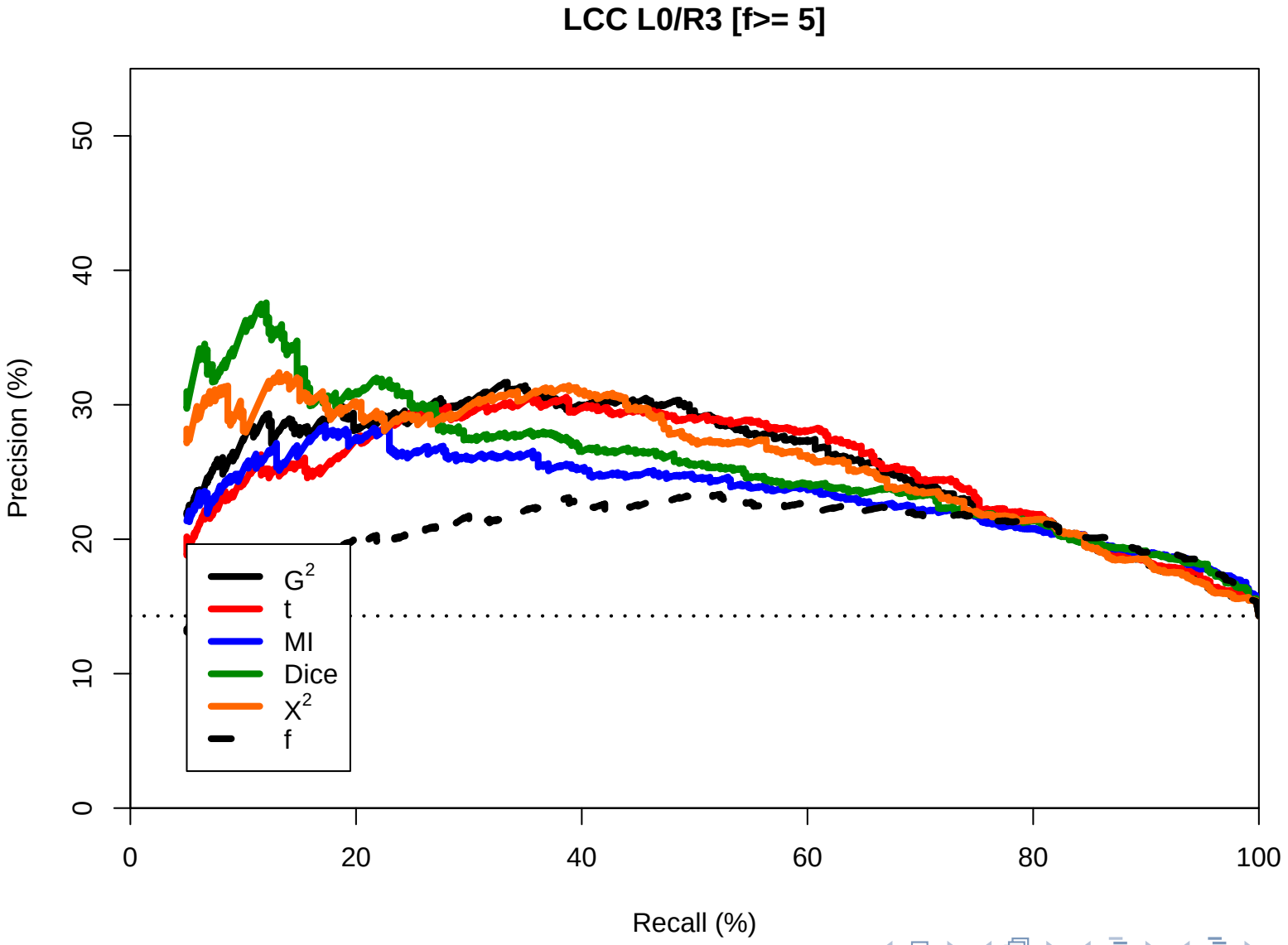
Evaluation on English VPC extraction task

(Baldwin 2008)



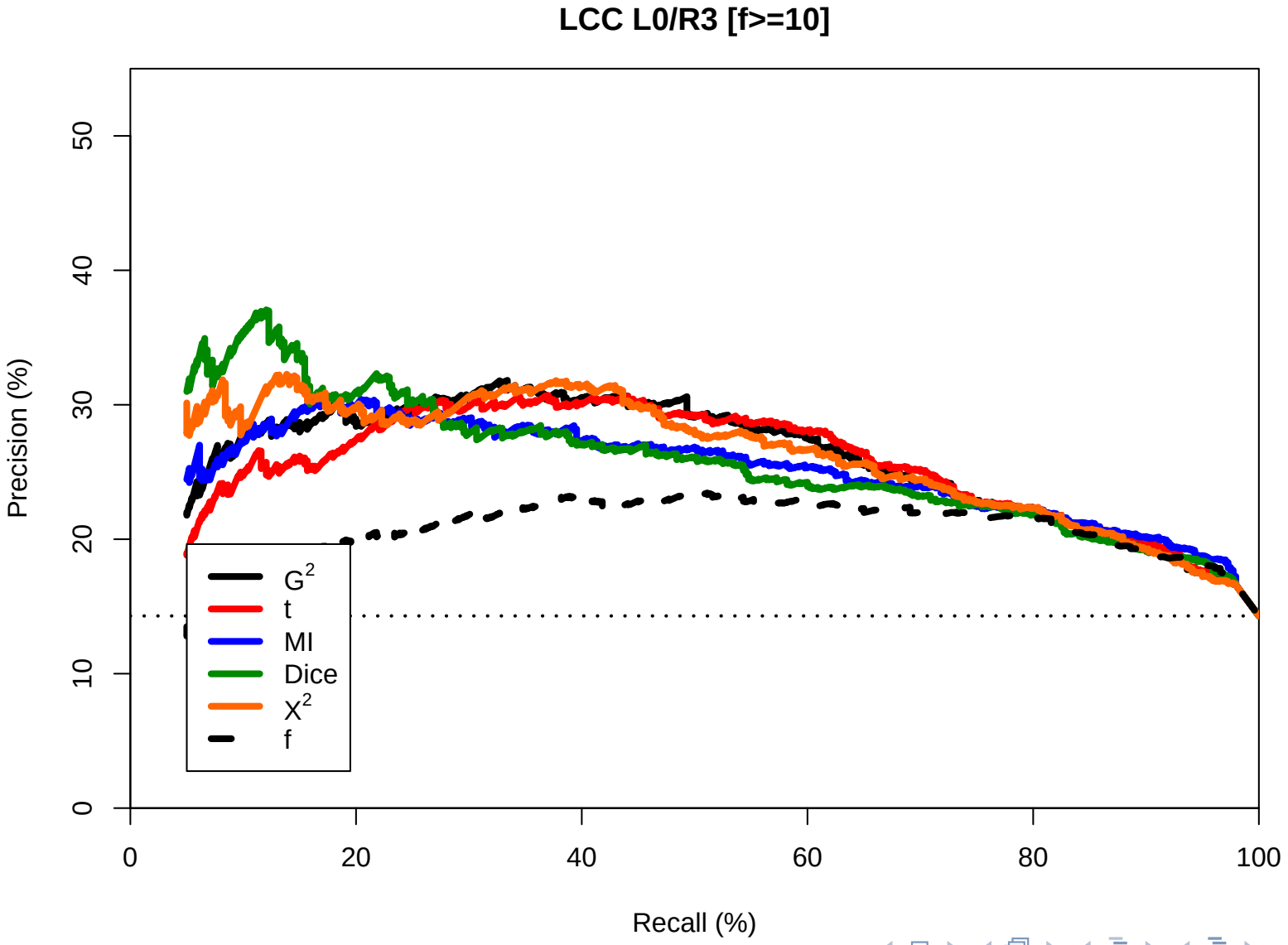
Evaluation on English VPC extraction task

(Baldwin 2008)



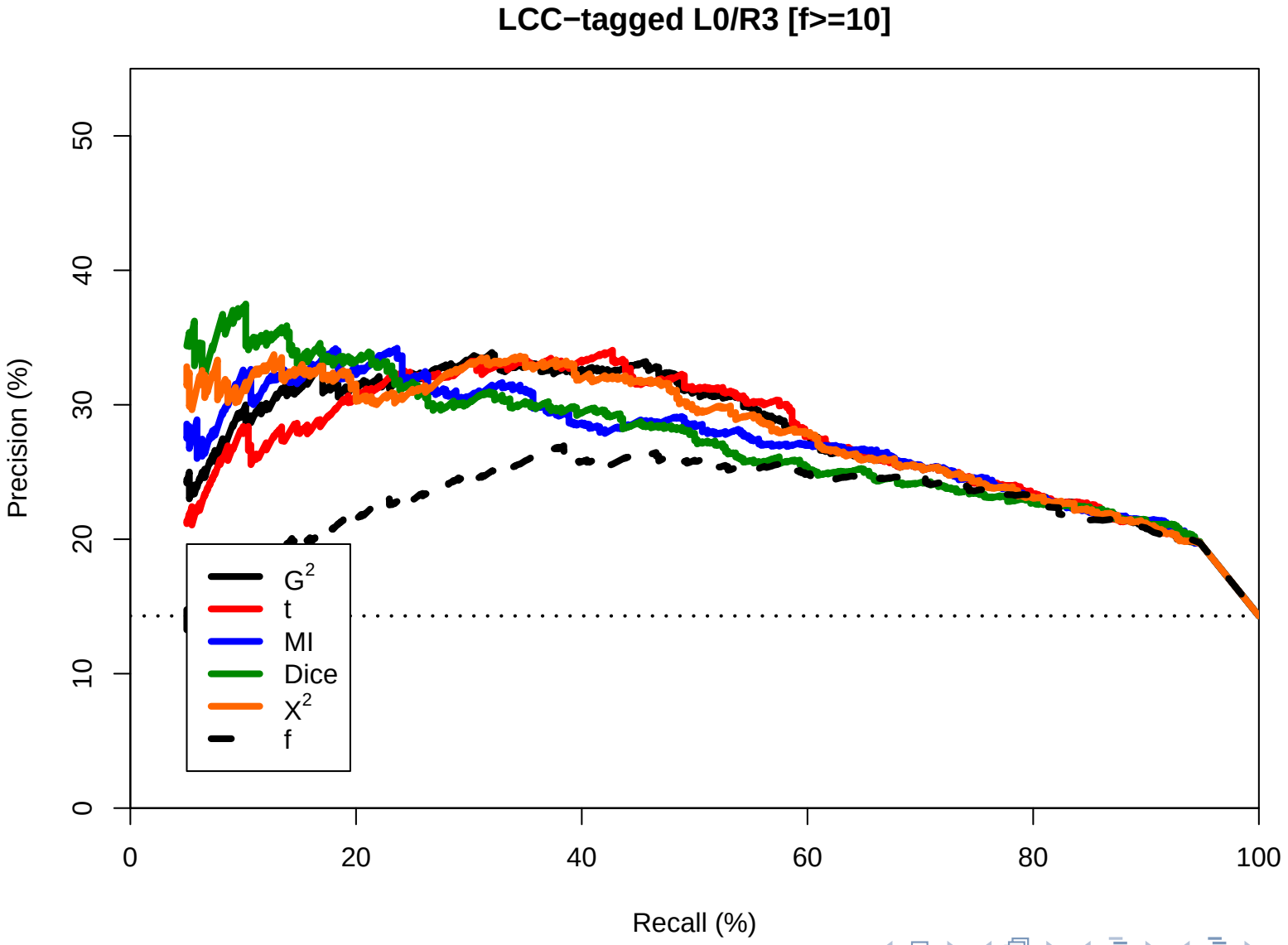
Evaluation on English VPC extraction task

(Baldwin 2008)



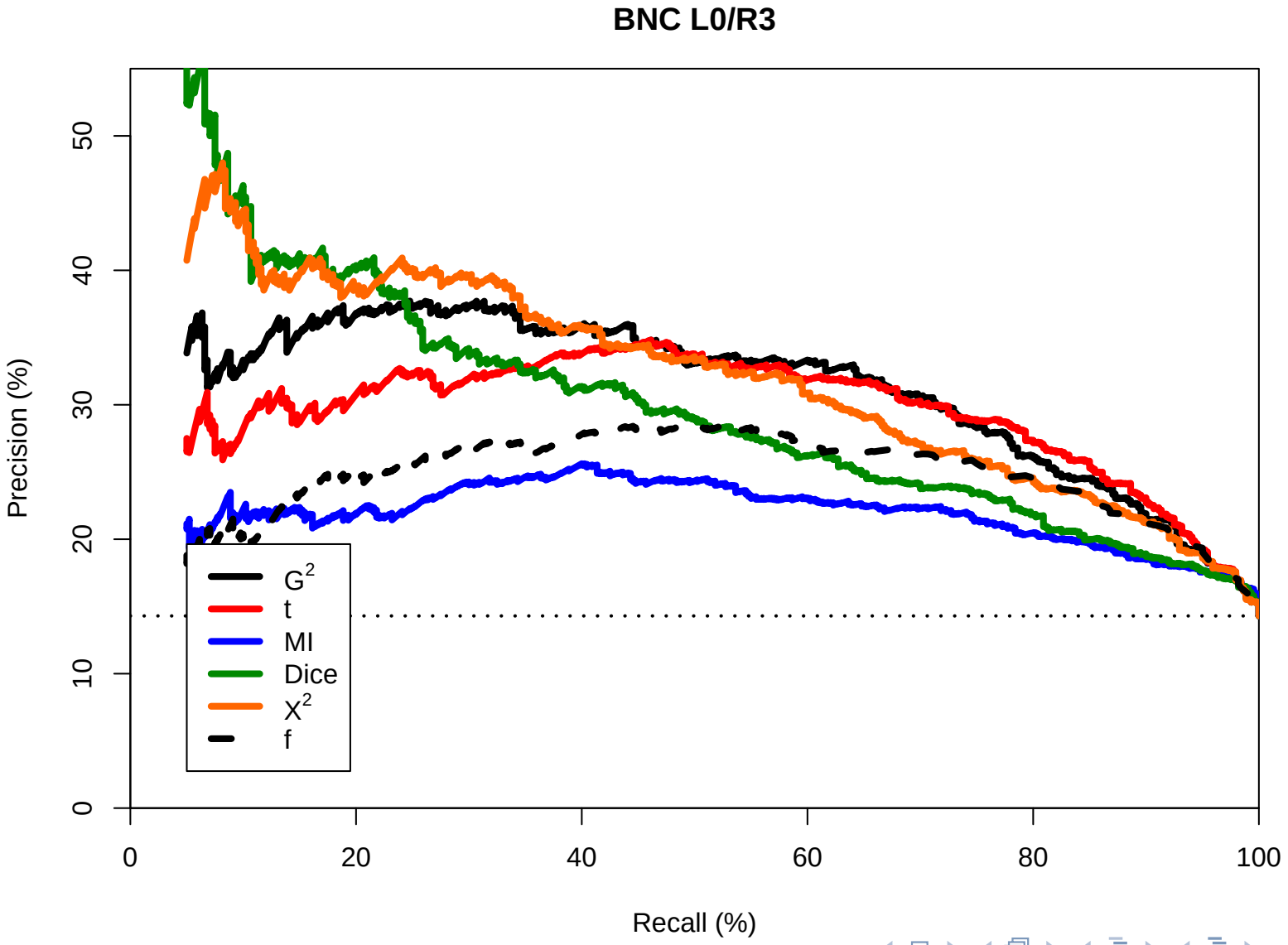
Evaluation on English VPC extraction task

(Baldwin 2008)



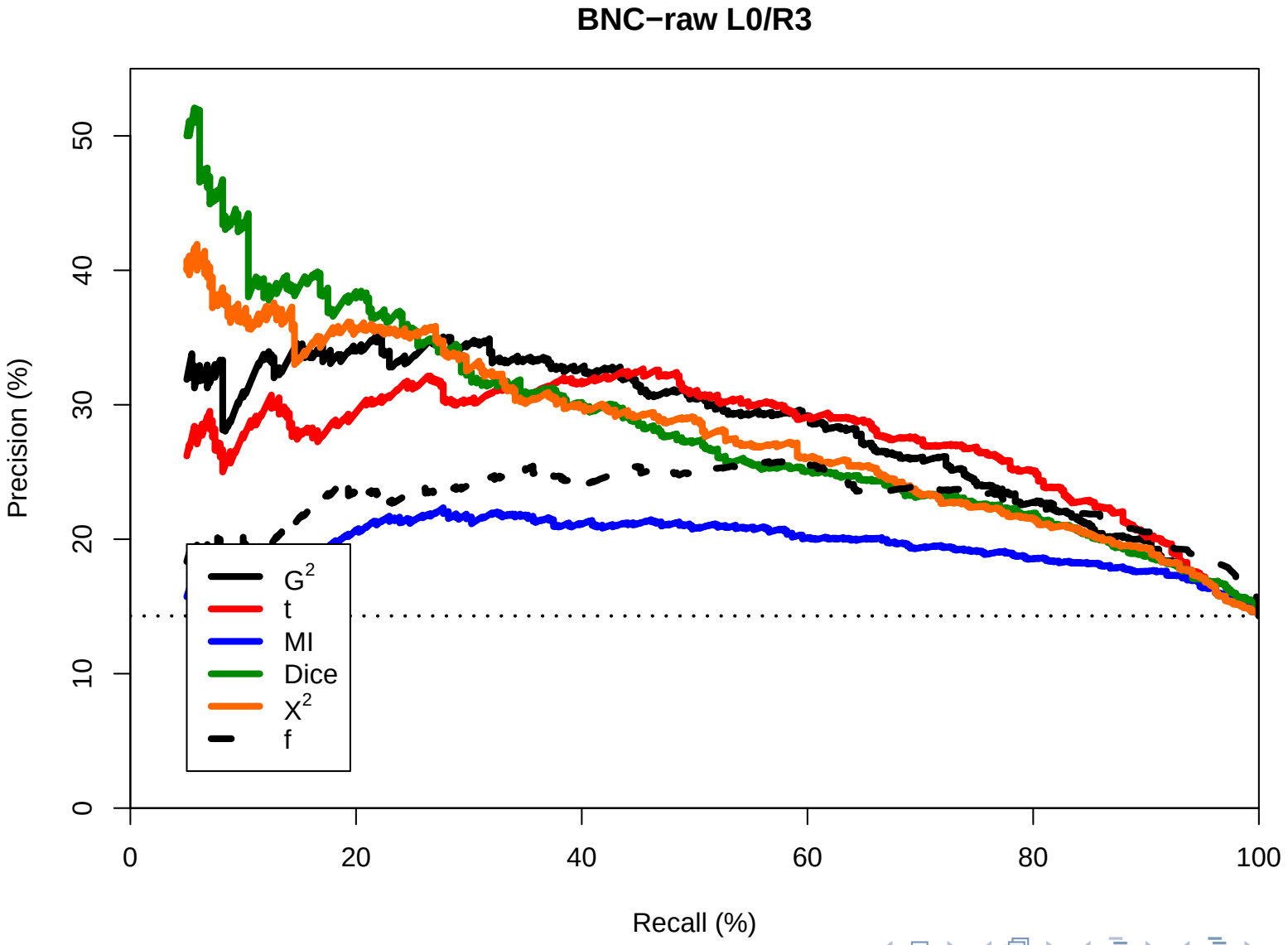
Evaluation on English VPC extraction task

(Baldwin 2008)



Evaluation on English VPC extraction task

(Baldwin 2008)



Evaluation on BBI collocation identification task

(Benson *et al.* 1986)

- Identification of lexical collocations as habitual, recurrent word combinations (Firth 1957)
  - essential for advanced learners (➜ idiomatic English)
  - different from lexicalised MWE
  - semi-compositional or fully compositional
  - no clear-cut linguistic criteria or tests available

Evaluation on BBI collocation identification task

(Benson *et al.* 1986)

- Identification of lexical collocations as habitual, recurrent word combinations (Firth 1957)
  - essential for advanced learners (➜ idiomatic English)
  - different from lexicalised MWE
  - semi-compositional or fully compositional
  - no clear-cut linguistic criteria or tests available
- Gold standard: BBI combinatory dict. (Benson *et al.* 1986)
  - BBI was compiled manually based on lexicographer intuitions
  - lexical collocations automatically extracted from scanned BBI, manually checked for 224 selected node words (➜ Bartsch224)

injury *n.* 1. to inflict (an) ~ on 2. to receive, suffer, sustain an ~ 3. a fatal; minor, slight; serious, severe ~ 4. bodily ~; an internal ~ 5. an ~ to (an ~ to the head) 6. (misc.) to add insult to ~

Evaluation on BBI collocation identification task

(Benson *et al.* 1986)

- Candidate extraction and ranking
  - for each of the 224 nodes, extract all co-occurrences with the 7,711 words that occur as collocates somewhere in the BBI
  - different spans: syntactic, L3/R3, L5/R5, L10/R10, sentence
  - full list of candidates ranked according to standard association measures (not per individual node) ➜ precision-recall graphs
  - composite: AP50 = average precision up to 50% recall, selecting best measure for each data set

Evaluation on BBI collocation identification task

(Benson *et al.* 1986)

- Candidate extraction and ranking
  - for each of the 224 nodes, extract all co-occurrences with the 7,711 words that occur as collocates somewhere in the BBI
  - different spans: syntactic, L3/R3, L5/R5, L10/R10, sentence
  - full list of candidates ranked according to standard association measures (not per individual node) ➜ precision-recall graphs
  - composite: AP50 = average precision up to 50% recall, selecting best measure for each data set
- Dictionary-based evaluation problematic (Evert 2004, 139f)
  - provides lower bound on n-best precision
  - coverage of native speaker intuitions by corpus data
  - may be biased against recent corpora, Web texts, etc.

Evaluation on BBI collocation identification task

(Benson *et al.* 1986)

- Candidate extraction and ranking
  - for each of the 224 nodes, extract all co-occurrences with the 7,711 words that occur as collocates somewhere in the BBI
  - different spans: syntactic, L3/R3, L5/R5, L10/R10, sentence
  - full list of candidates ranked according to standard association measures (not per individual node) ➜ precision-recall graphs
  - composite: AP50 = average precision up to 50% recall, selecting best measure for each data set
- Dictionary-based evaluation problematic (Evert 2004, 139f)
  - provides lower bound on n-best precision
  - coverage of native speaker intuitions by corpus data
  - may be biased against recent corpora, Web texts, etc.
- Manual validation of ranked collocates for selected nodes
  - work in progress, using custom Web-based annotation tool

Evaluation on BBI collocation identification task

(Benson *et al.* 1986)

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 .

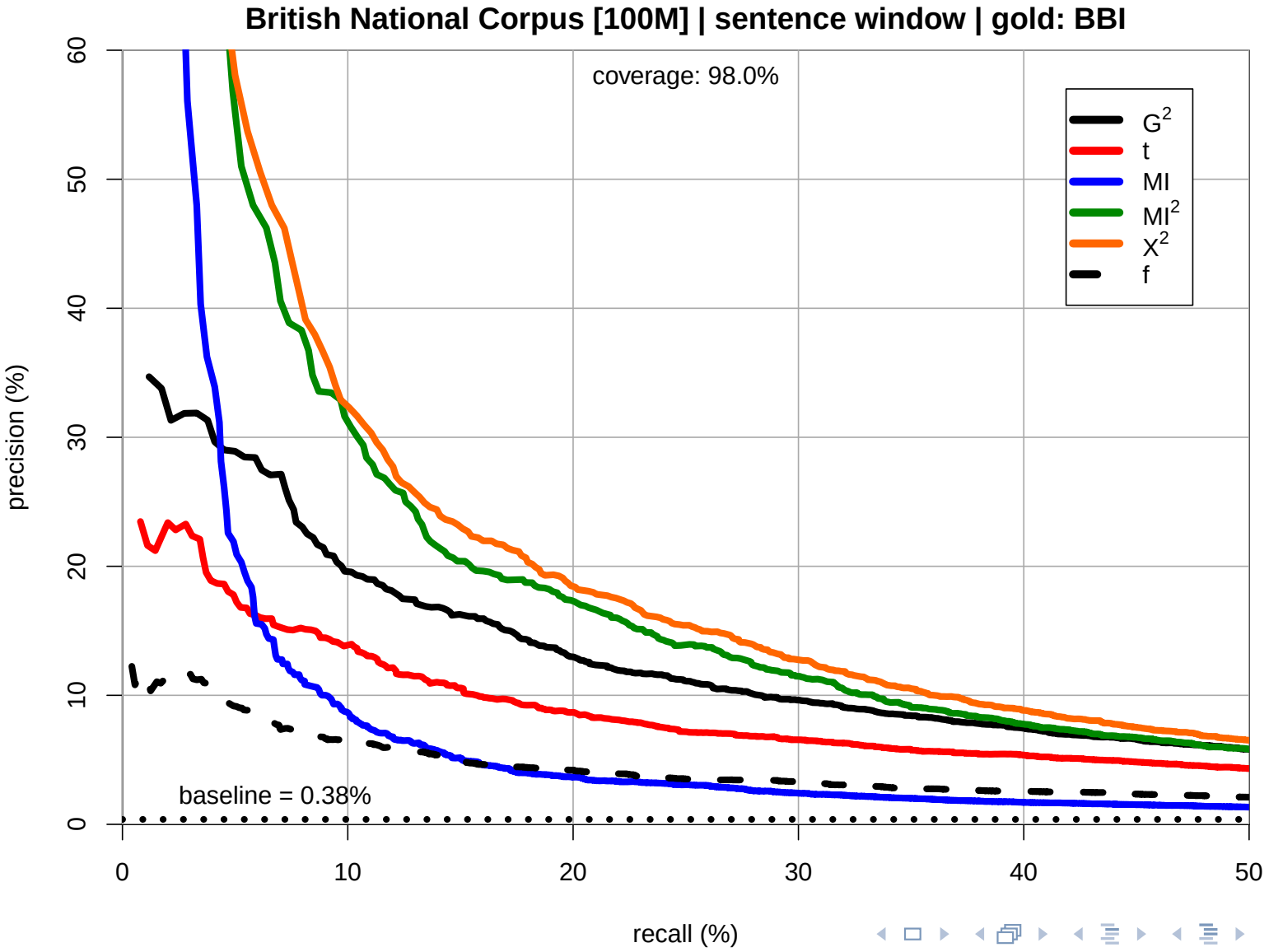
Rather than viewing them as epistemically-morally-politically pernicious forms of hasty generalization by contrast , say , with Gadamerian pre-judgements or putatively more benign practices of categorization , Fricker **argues plausibly** for a " neutral " sense of stereotype which catches their frequent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

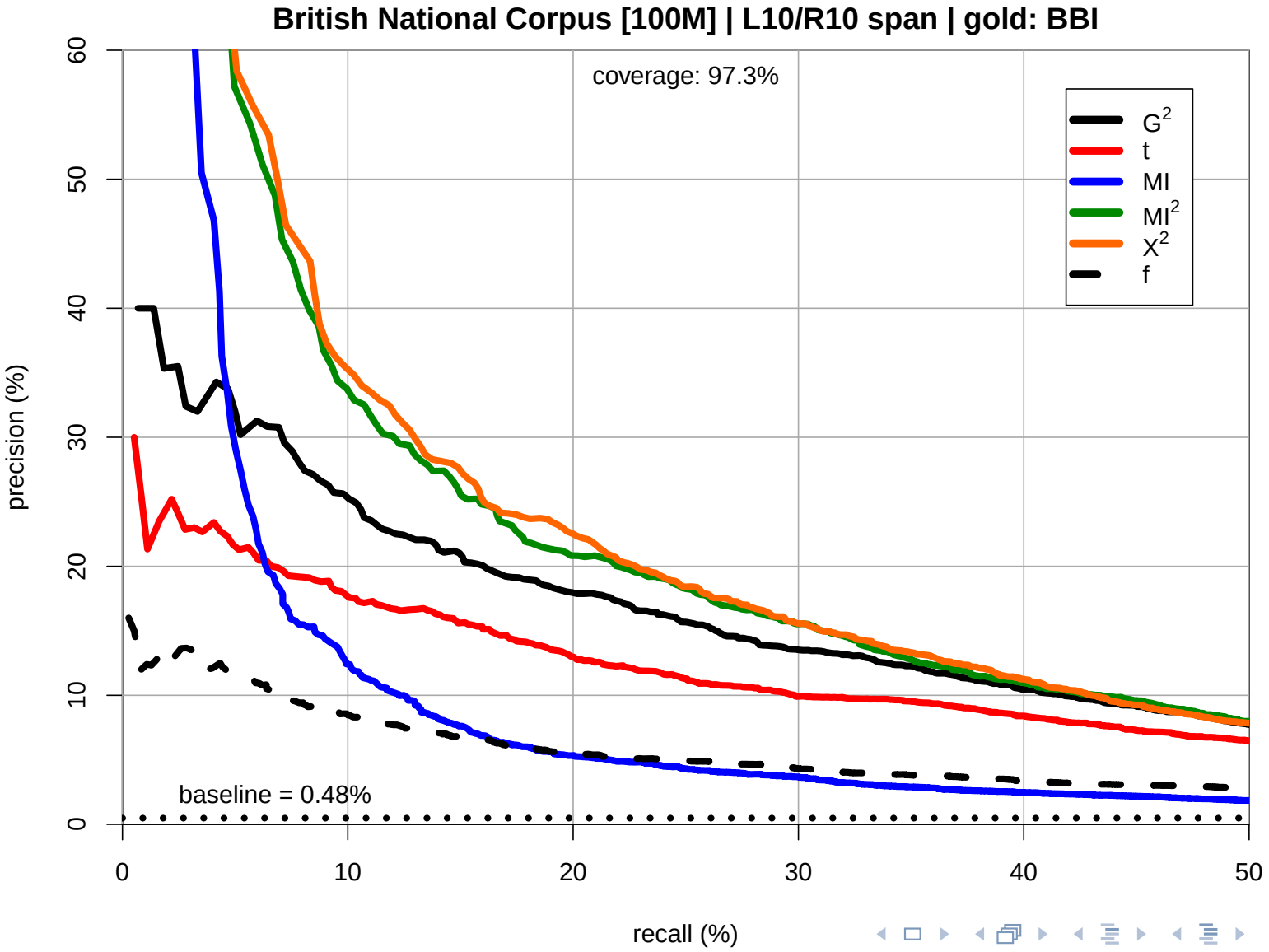
Evaluation on BBI: collocational span

(Benson *et al.* 1986)



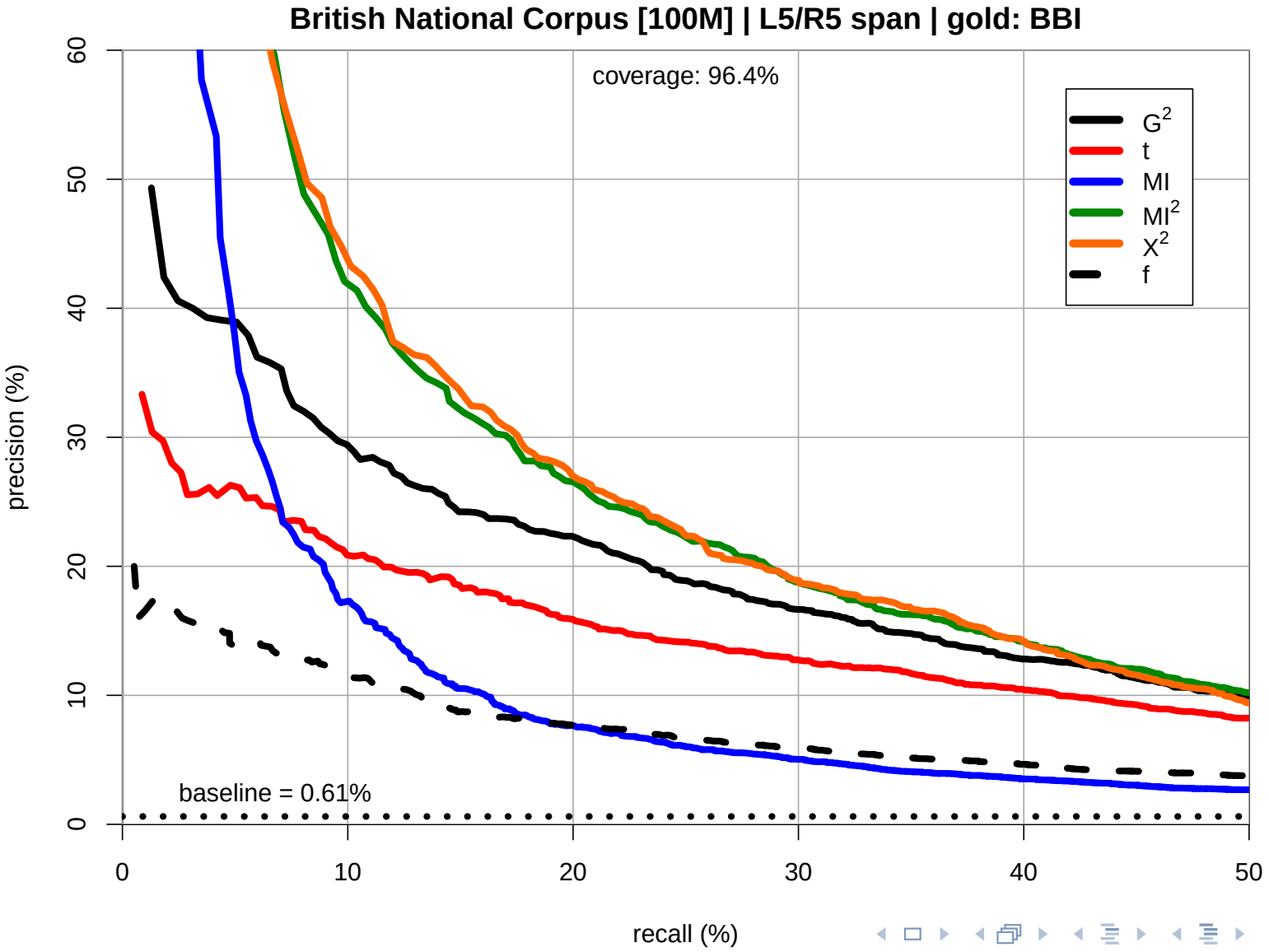
Evaluation on BBI: collocational span

(Benson *et al.* 1986)



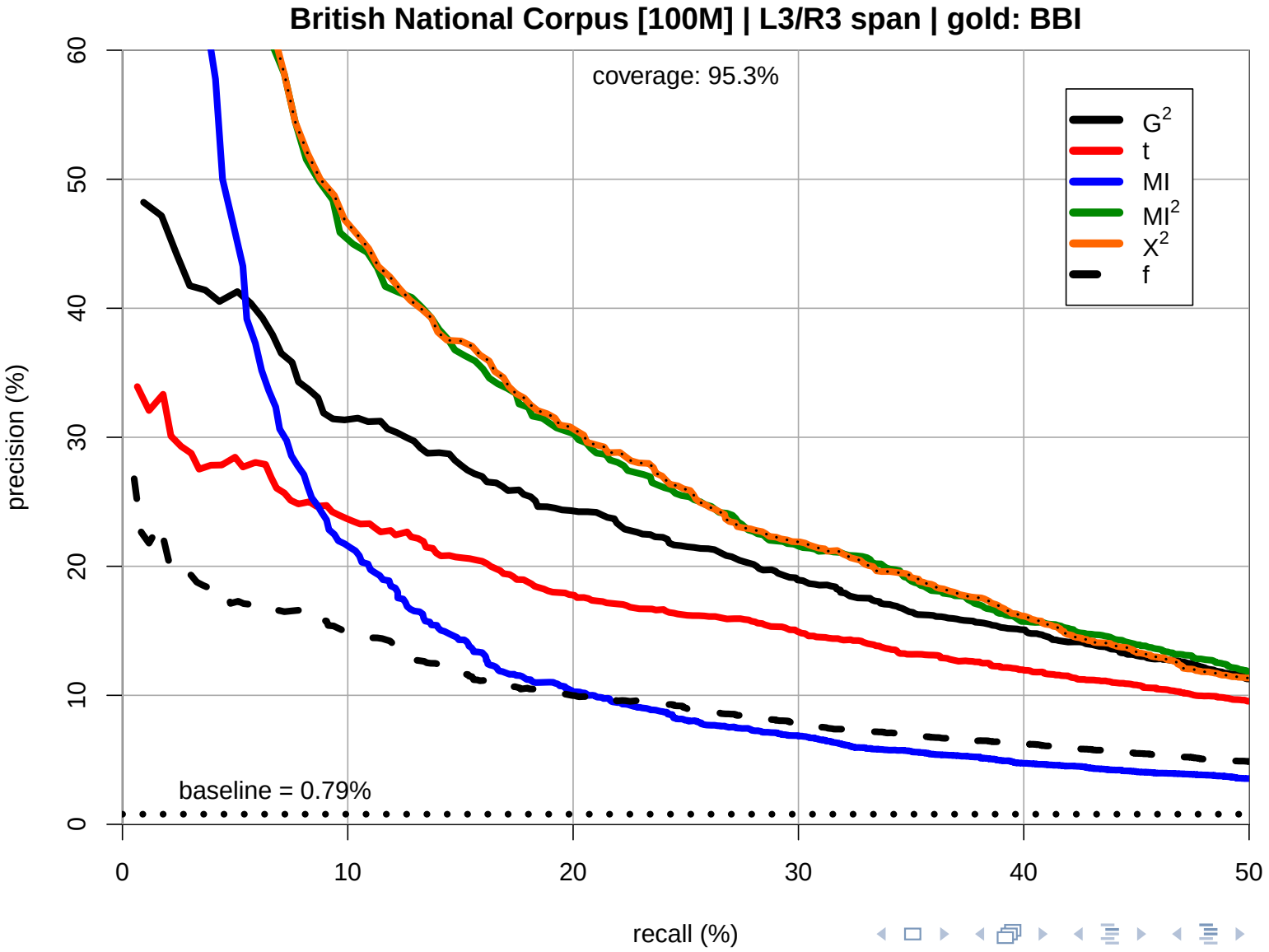
Evaluation on BBI: collocational span

(Benson *et al.* 1986)



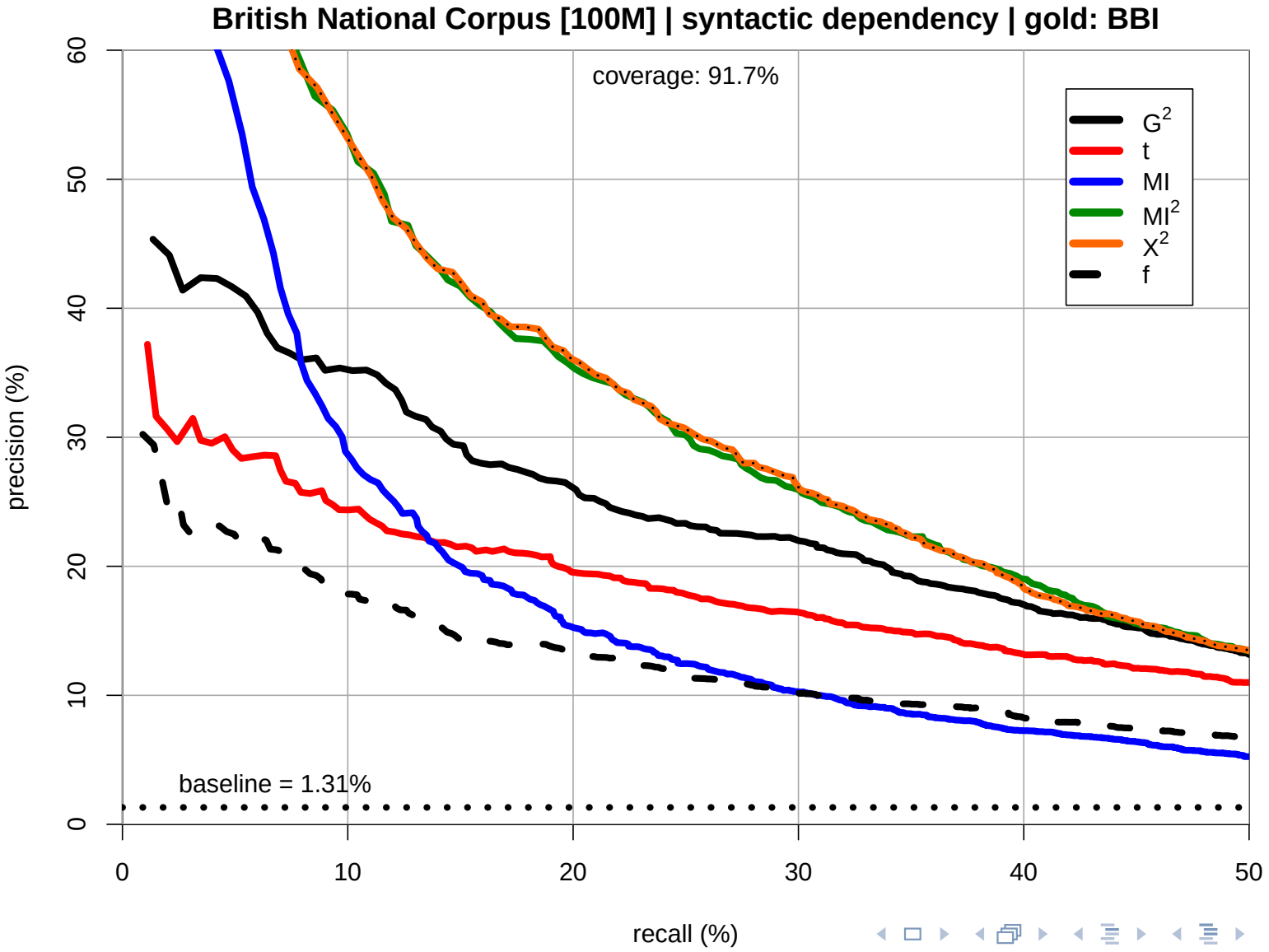
Evaluation on BBI: collocational span

(Benson *et al.* 1986)



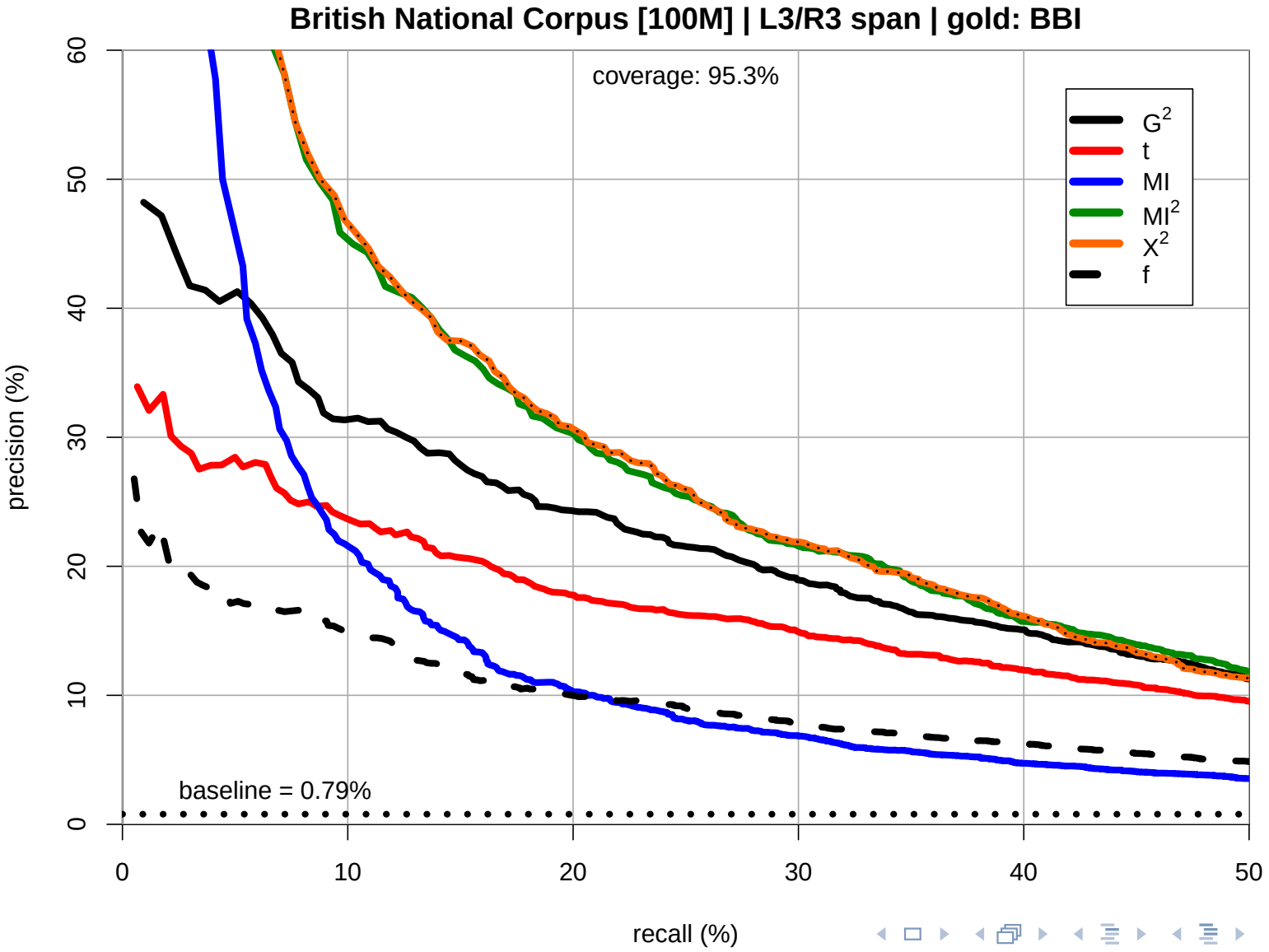
Evaluation on BBI: collocational span

(Benson *et al.* 1986)



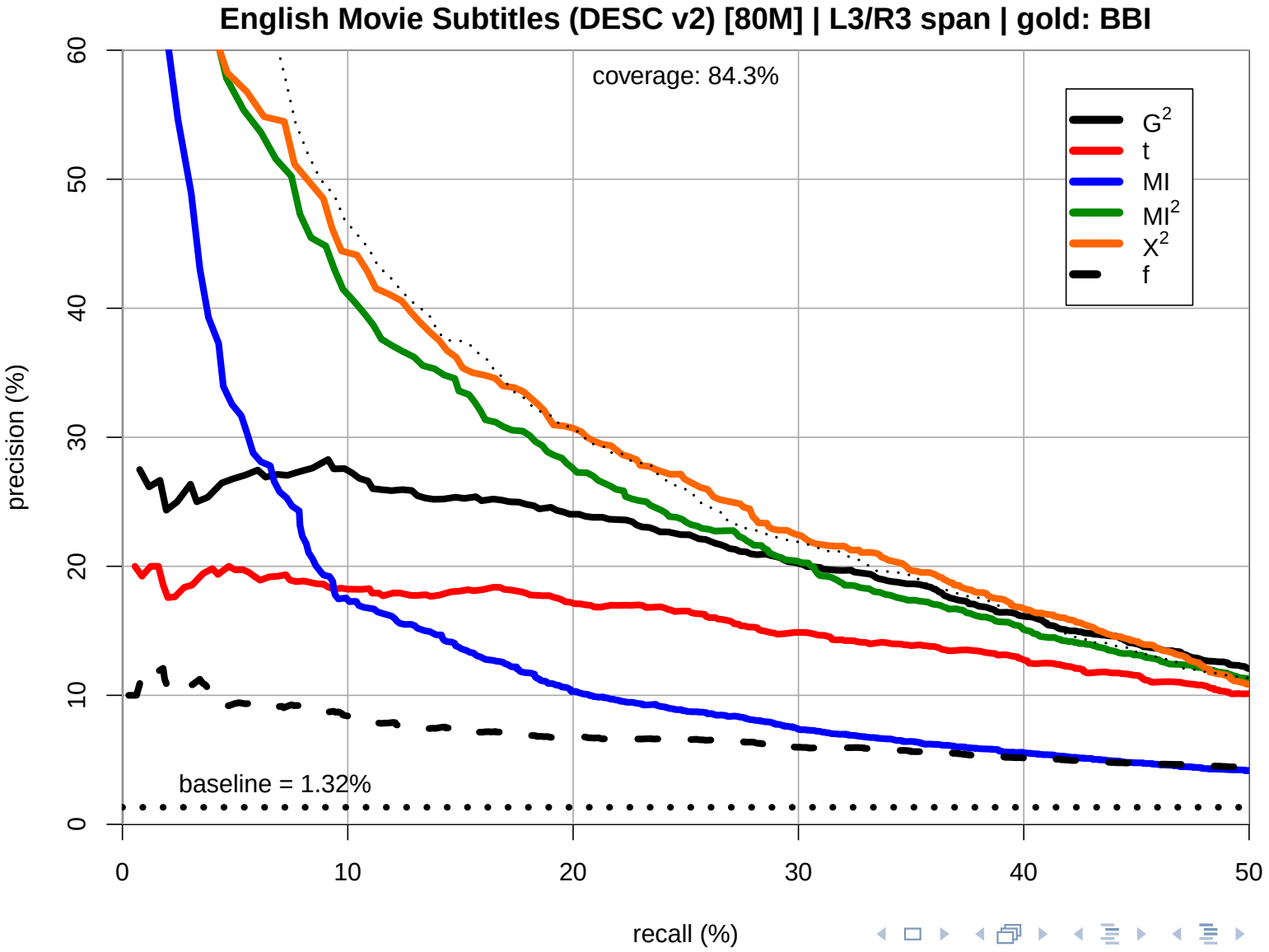
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



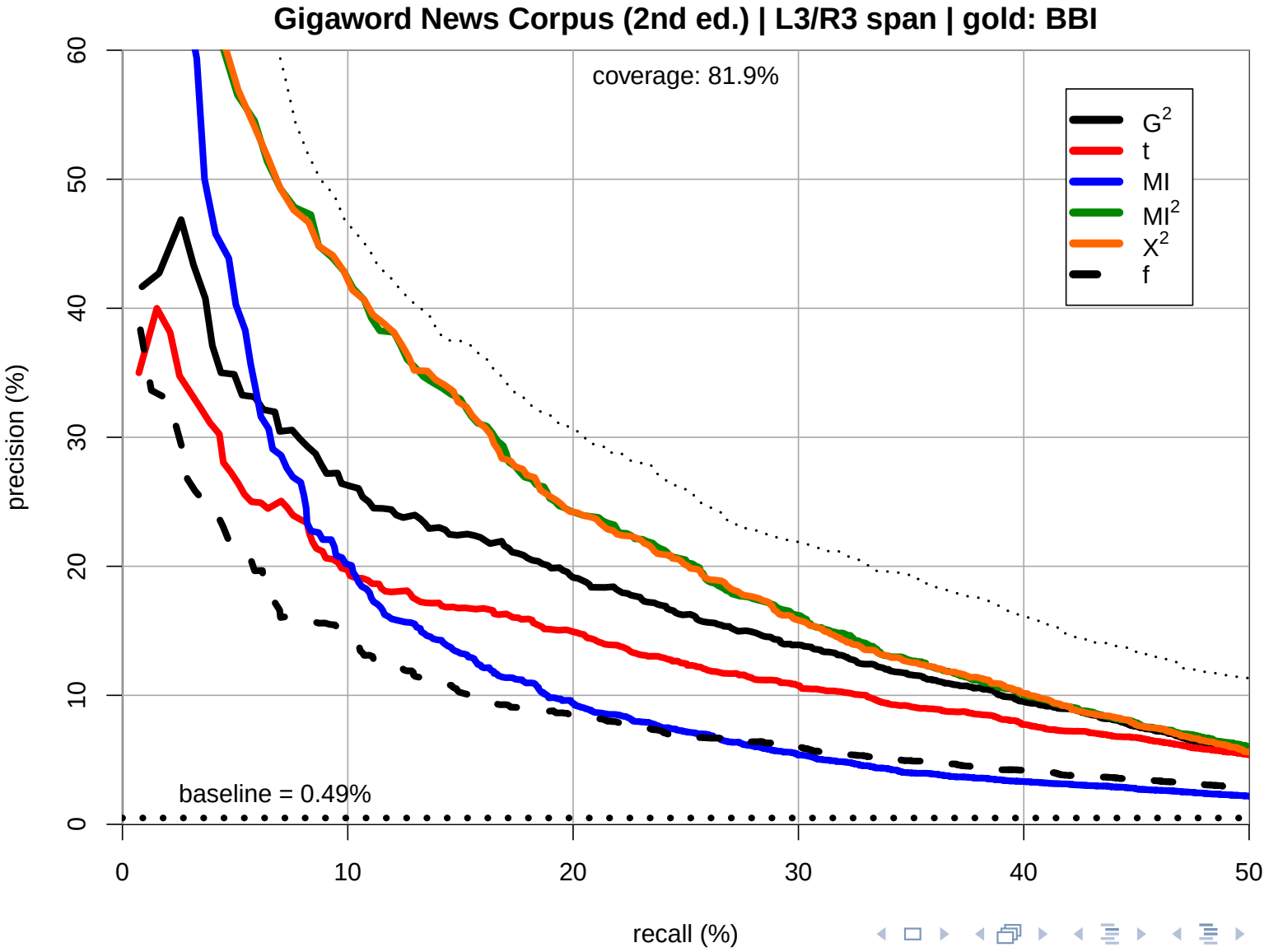
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



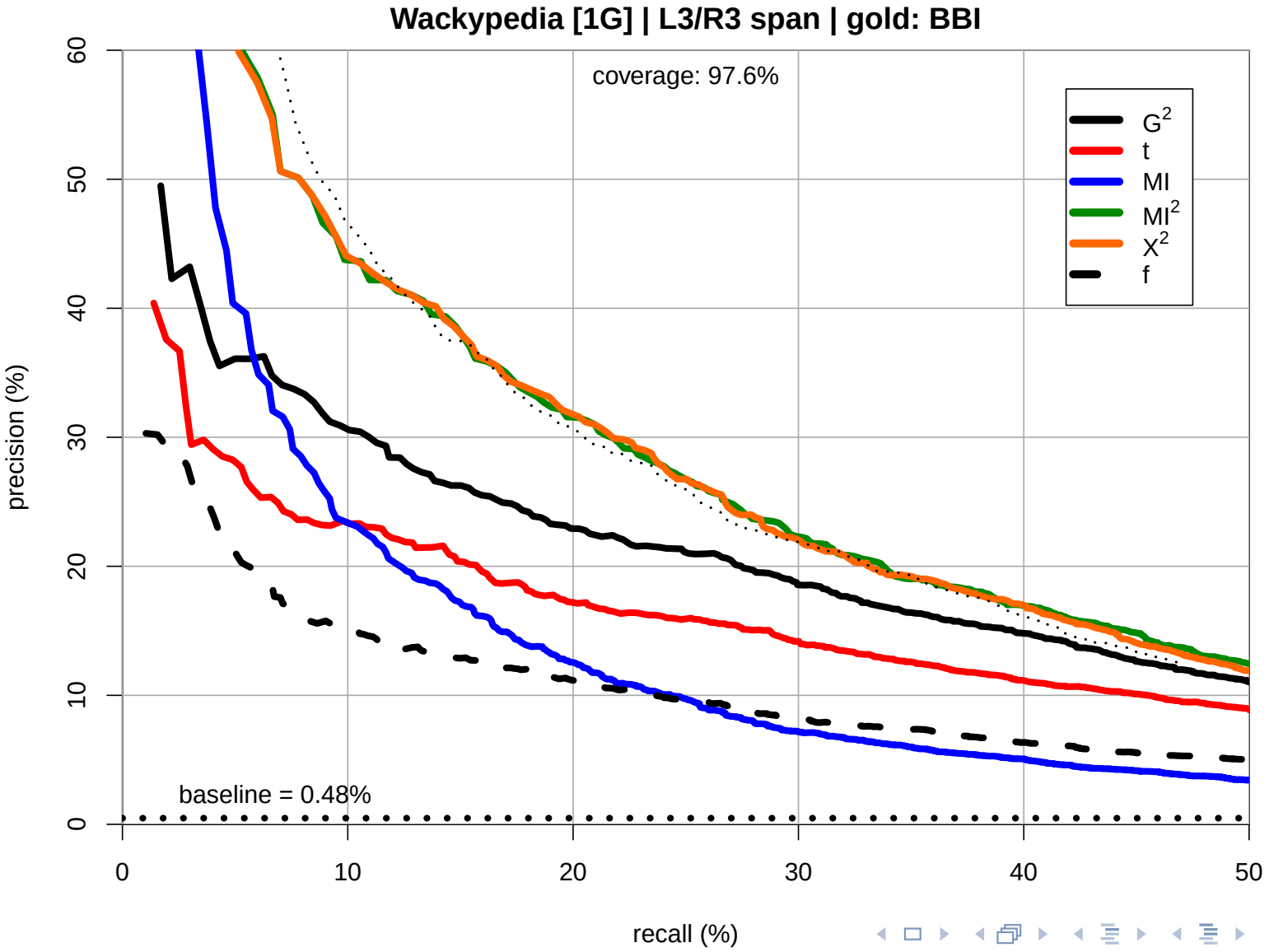
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



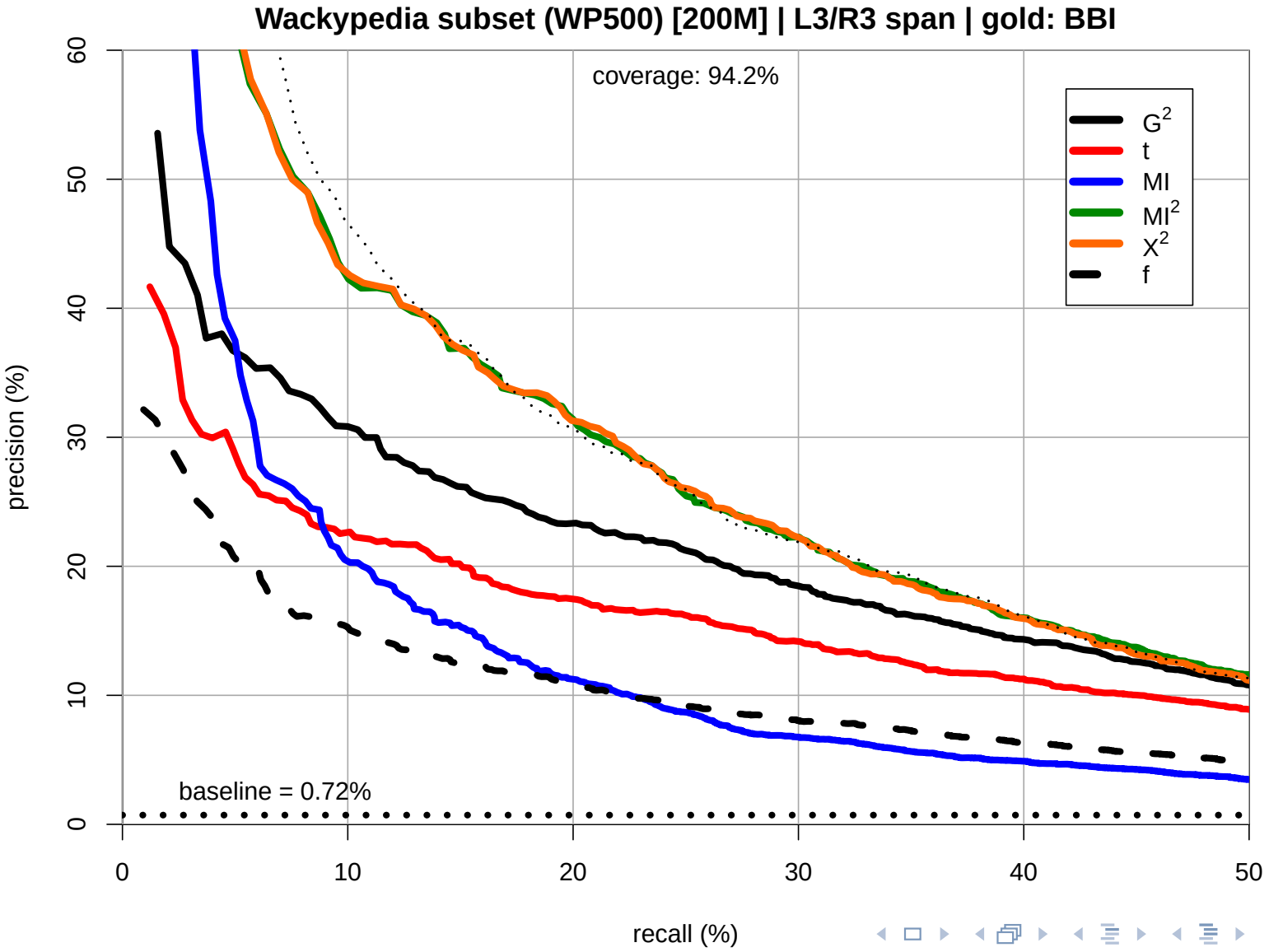
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



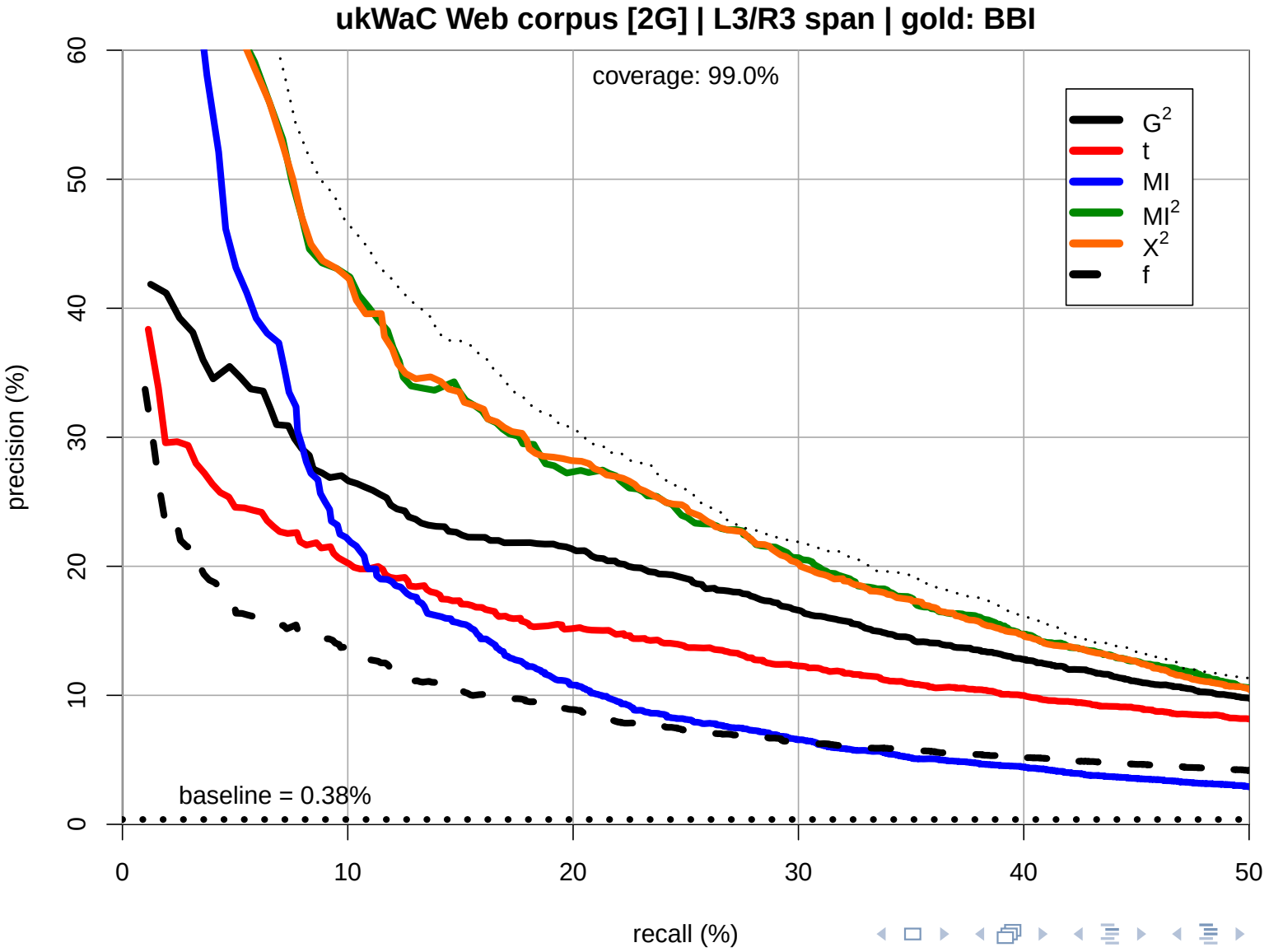
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



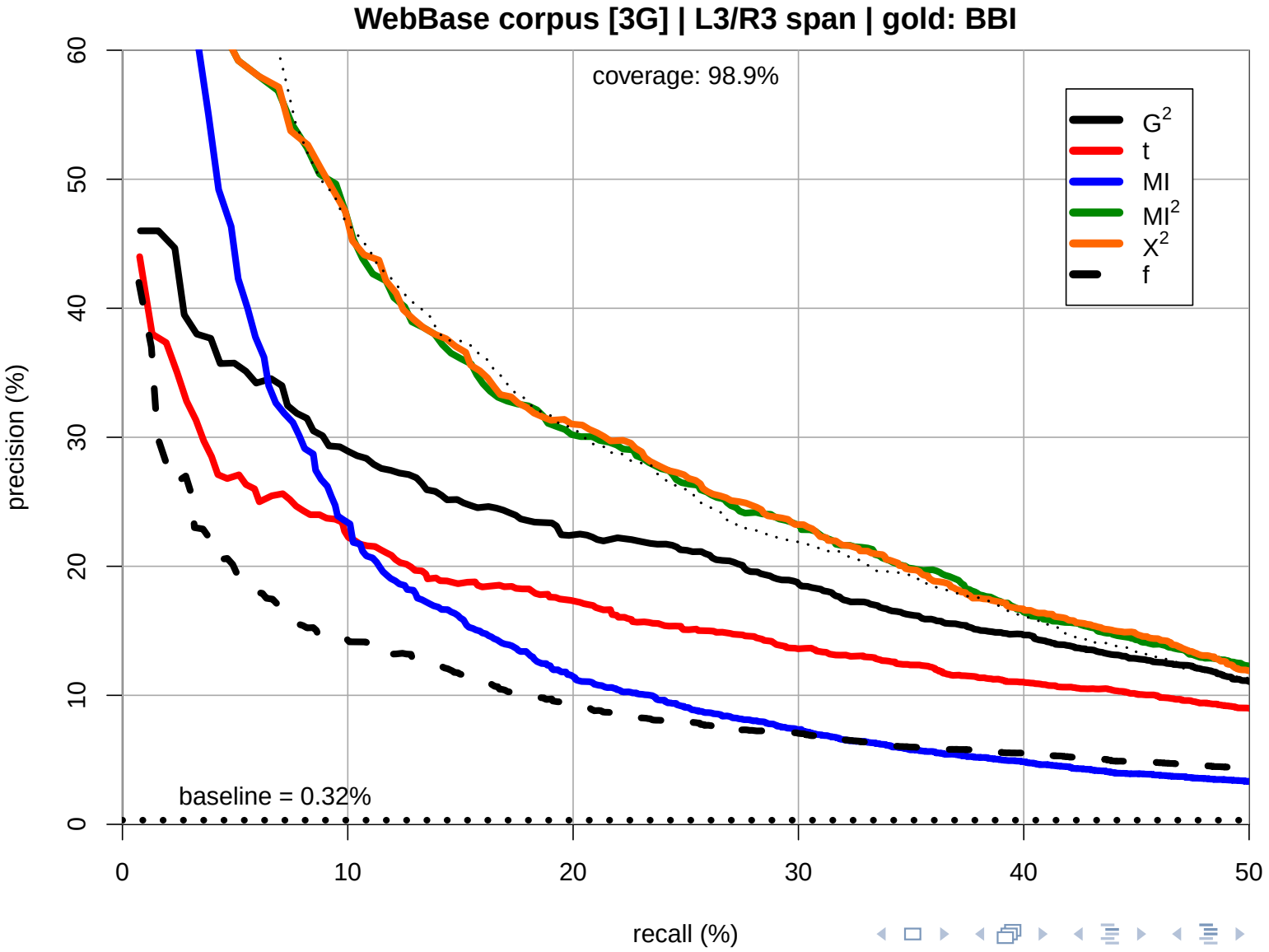
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



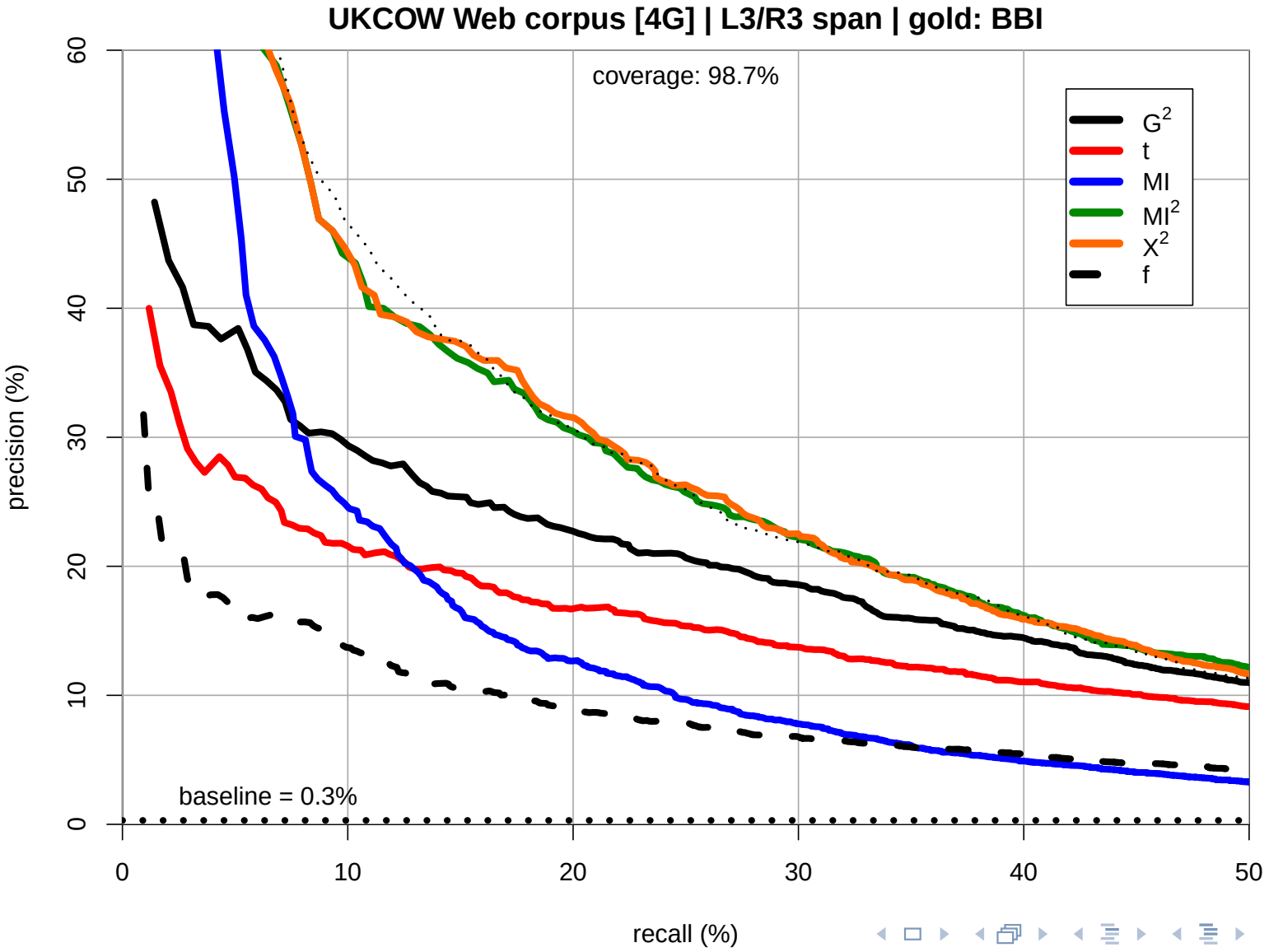
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



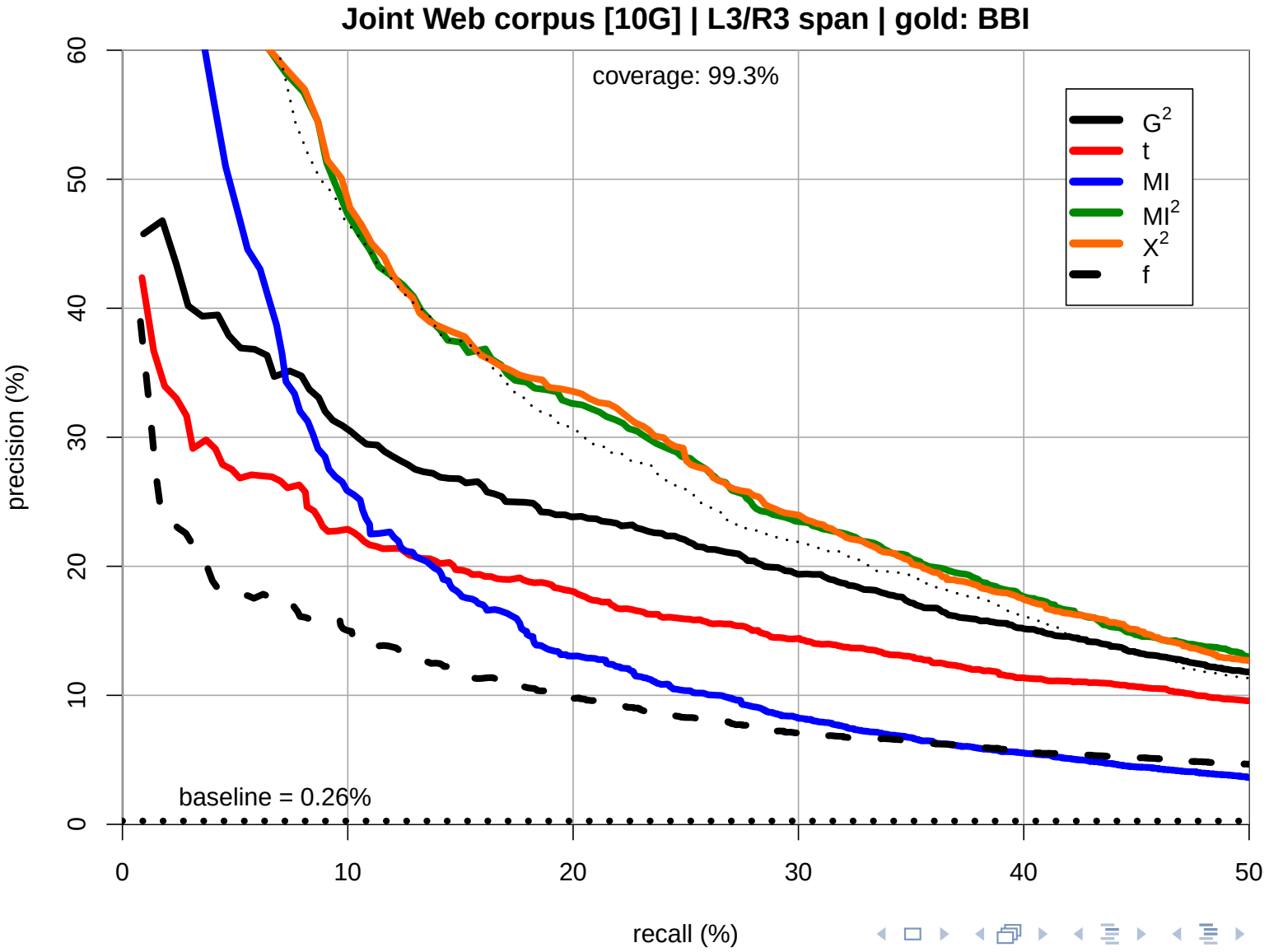
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



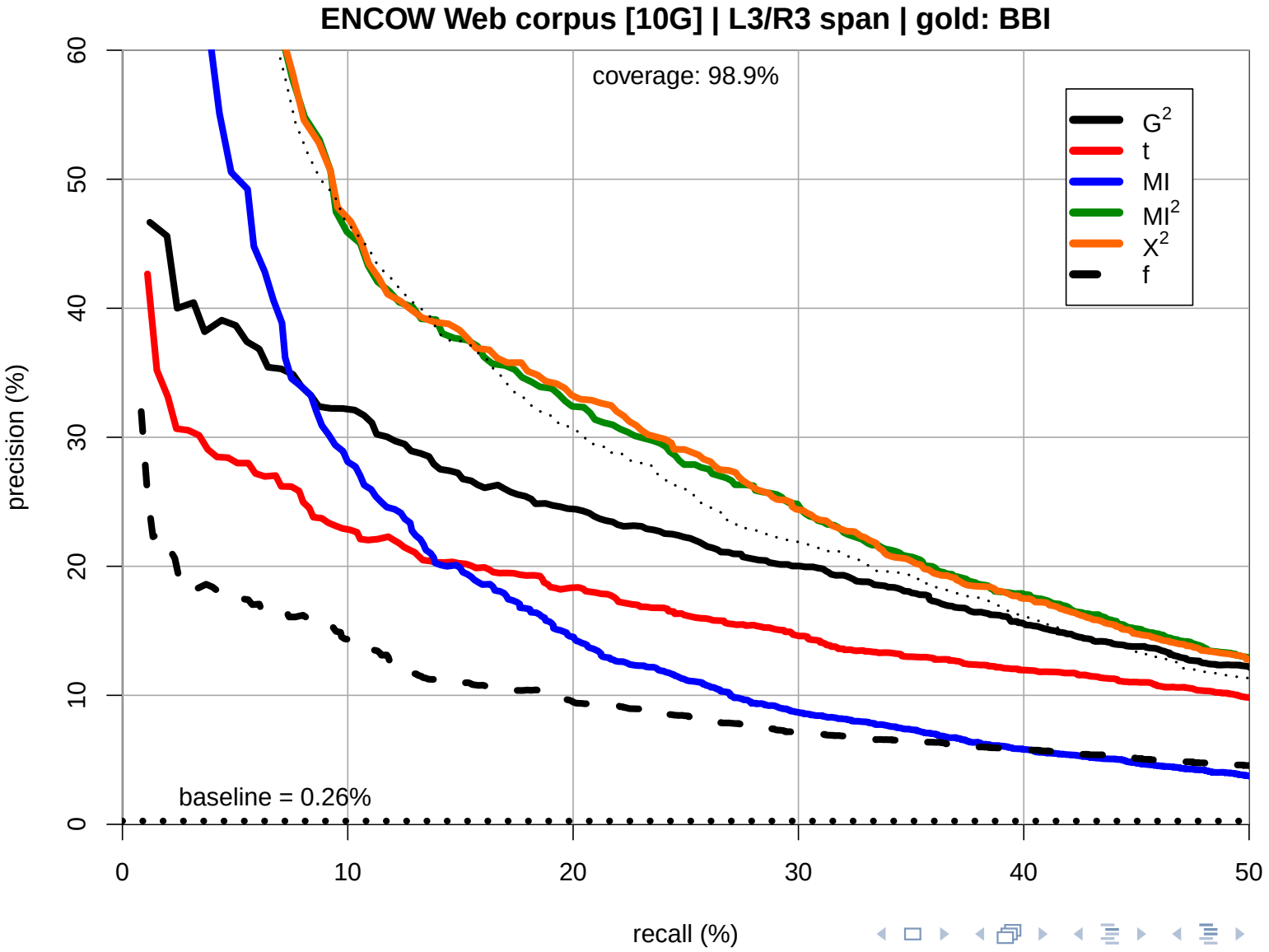
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



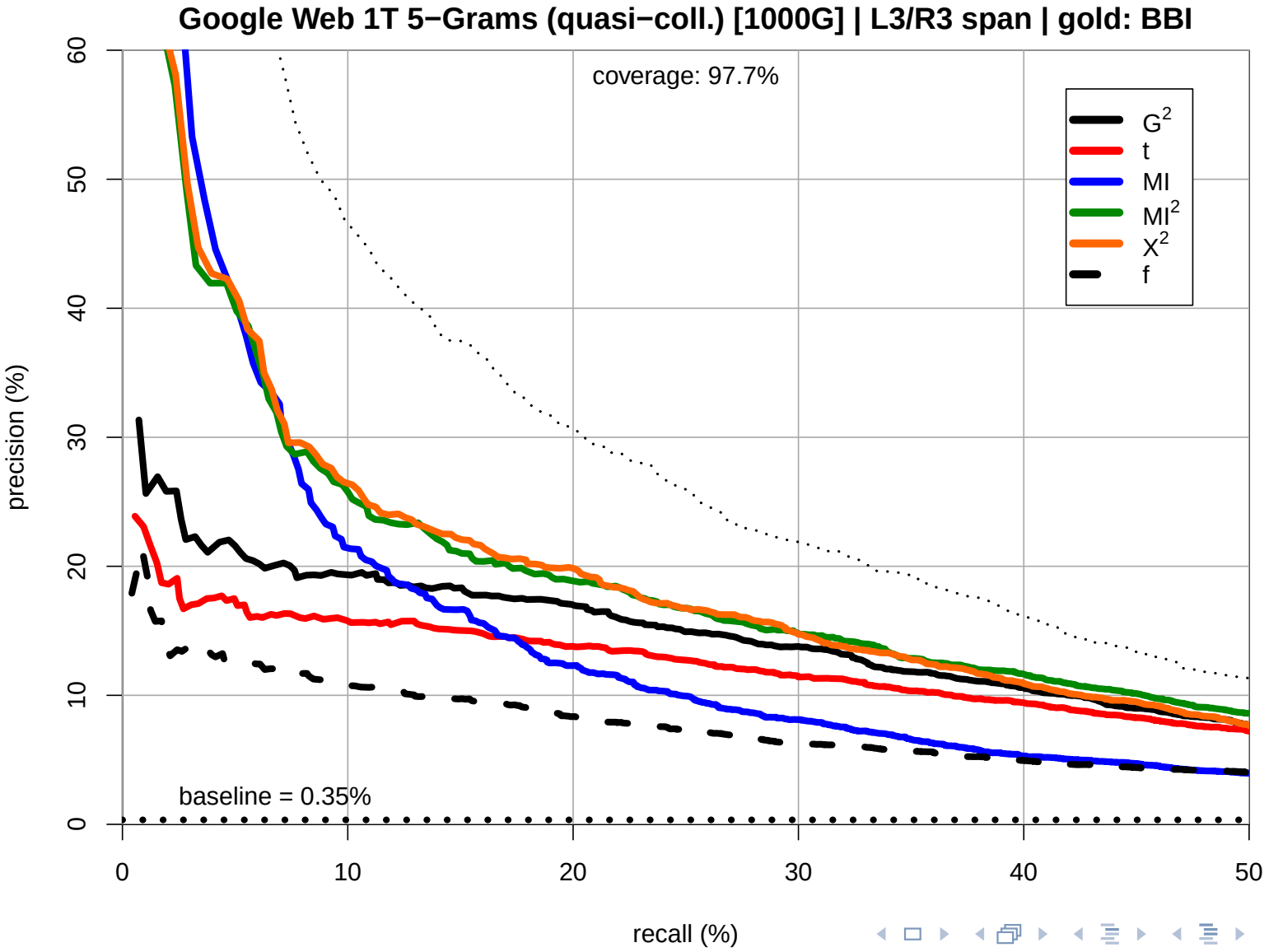
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



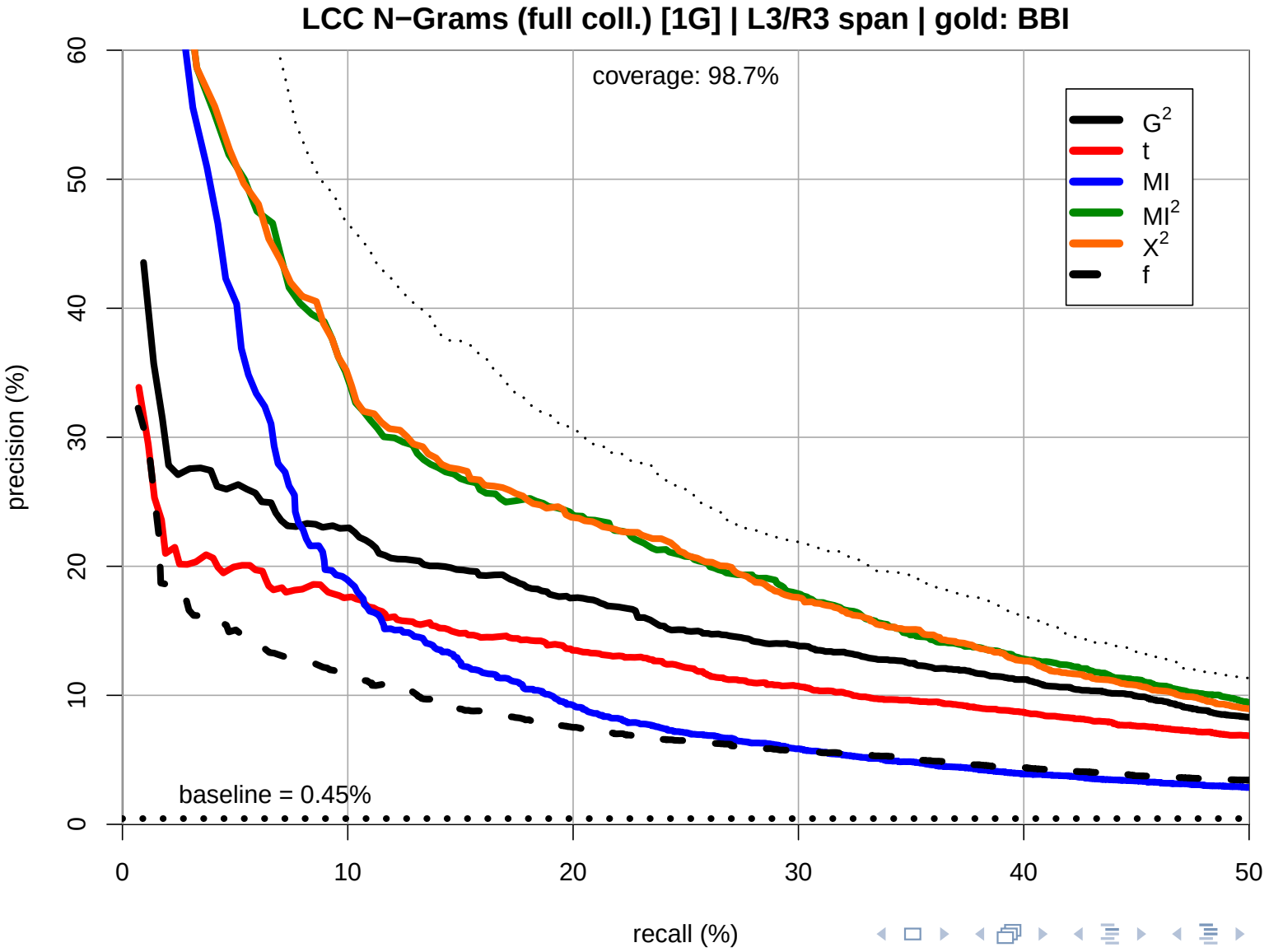
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



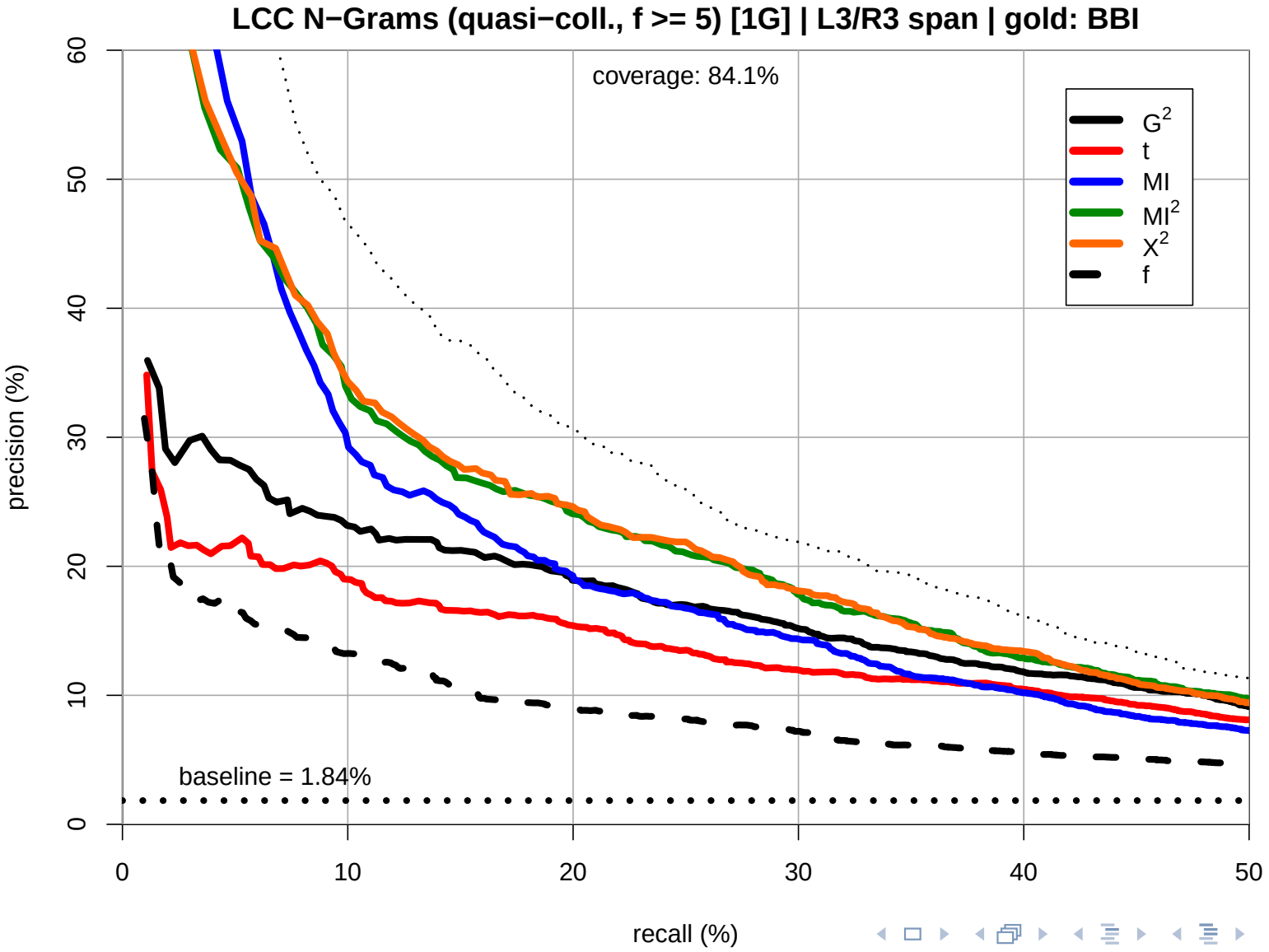
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



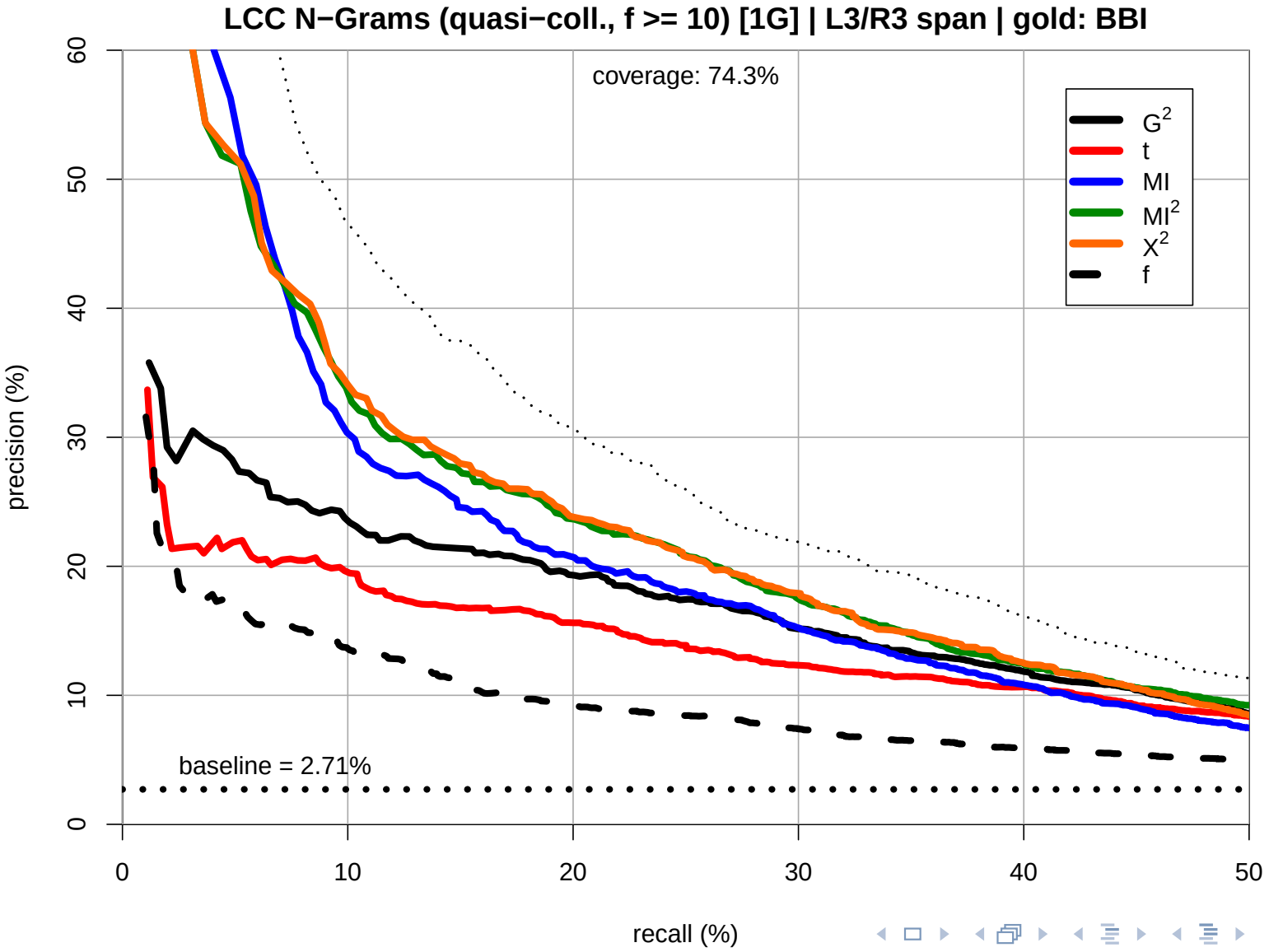
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



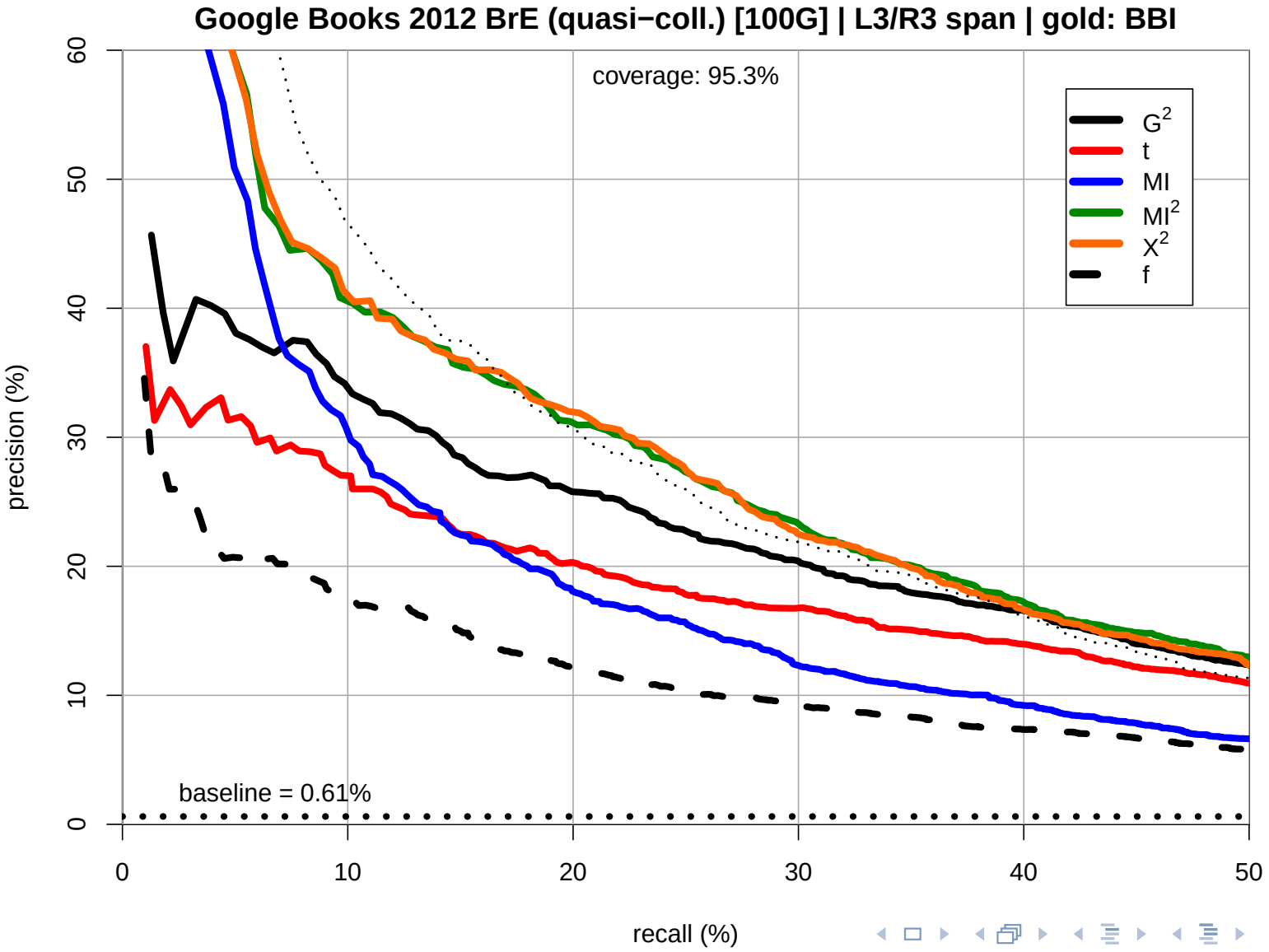
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



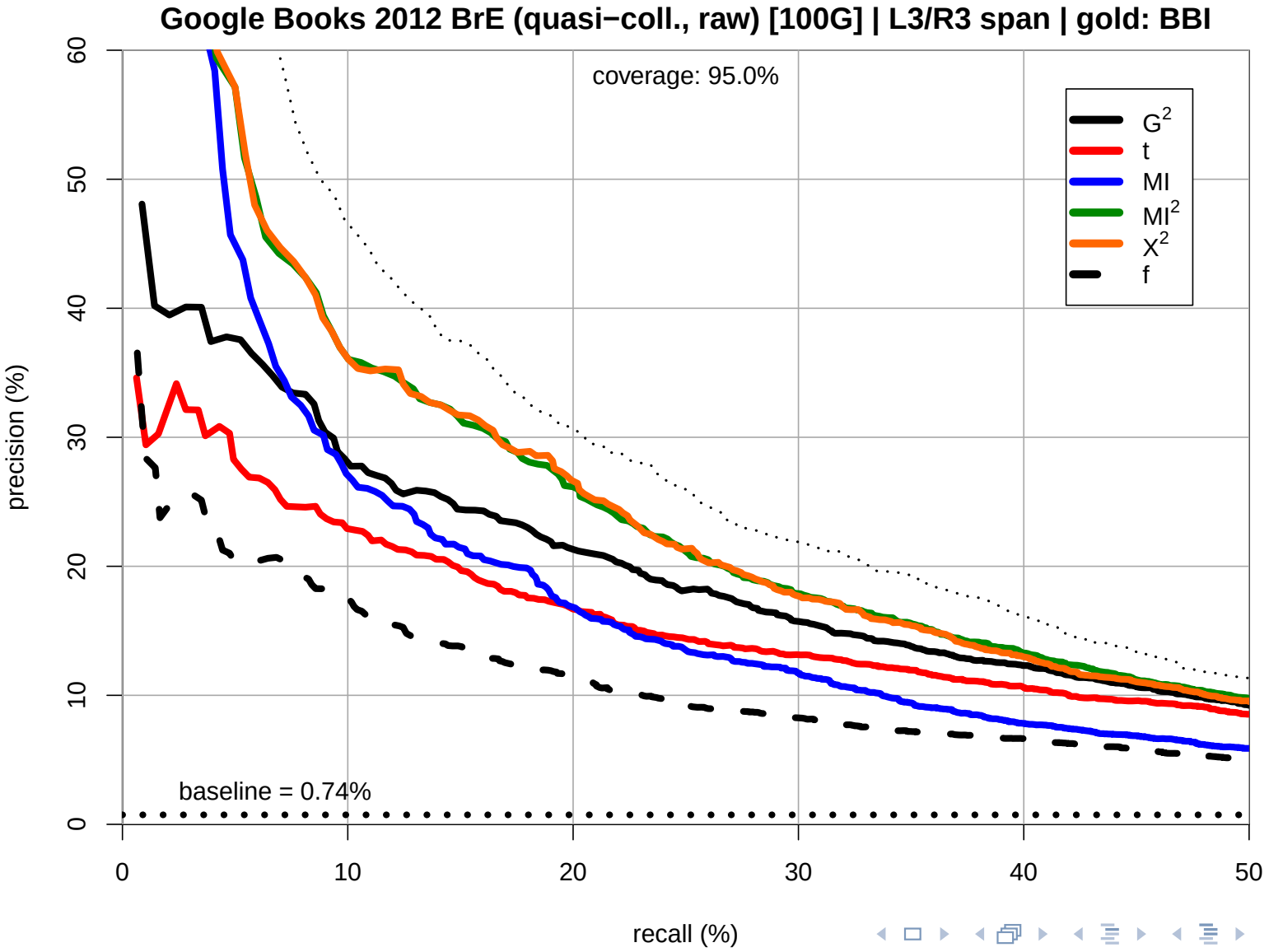
Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



Evaluation on BBI: different corpora (L3/R3)

(Benson *et al.* 1986)



Replication on Oxford Collocations Dictionary (2nd ed.)

(McIntosh *et al.* 2009)

- Criticism against BBI
  - lexicographer intuitions may be incomplete
  - outdated (published 1986, native speakers primed in 1950s)
  - biased against recent corpora

bucket *n.* 1. a coal; fire; ice; water ~ 2. a metal; wooden ~ 3. an empty; full ~ 4. (misc.) (colloq.) to kick the ~ ('to die')

Replication on Oxford Collocations Dictionary (2nd ed.)

(McIntosh *et al.* 2009)

- Criticism against BBI
  - lexicographer intuitions may be incomplete
  - outdated (published 1986, native speakers primed in 1950s)
  - biased against recent corpora
- Additional gold standard: OCD2 (McIntosh *et al.* 2009)
  - corpus-based + manual validation (1st ed. was based on BNC)
  - up to date, covers much broader range of collocates

bucket *n.* 1. a coal; fire; ice; water ~ 2. a metal; wooden ~ 3. an empty; full ~ 4. (misc.) (colloq.) to kick the ~ ('to die')

bucket *noun*

- ADJ. **empty, full | galvanized, metal, plastic | leaky | champagne, ice | water | slop, waste** *a slop bucket full of scraps of food* **| mop | coal | fire** *The sand had spilt from the fire bucket.*
- VERB + BUCKET **fill** *She filled the bucket with fresh water.* **| carry | empty, pour, throw** *She poured the bucket of dirty water down the drain.*
- BUCKET + VERB **be filled with/full of sth, contain sth, hold sth | overflow**
- PREP. **in a/the~ | ~of** *a bucket of oats for the horses*
- PHRASES **a bucket and spade** *The children ran down to the beach with their buckets and spades.* **mop and bucket** *The cleaner put down his mop and bucket and sat down.*

Replication on Oxford Collocations Dictionary (2nd ed.)

(McIntosh *et al.* 2009)

- Compilation of gold standard
  - automatic extraction from XML version of the dictionary
  - collocates explicitly marked ➜ no manual validation necessary
  - Bartsch224 nodes accepted as headwords and as collocates (OCD follows Hausmann's (1989) base-collocate distinction)

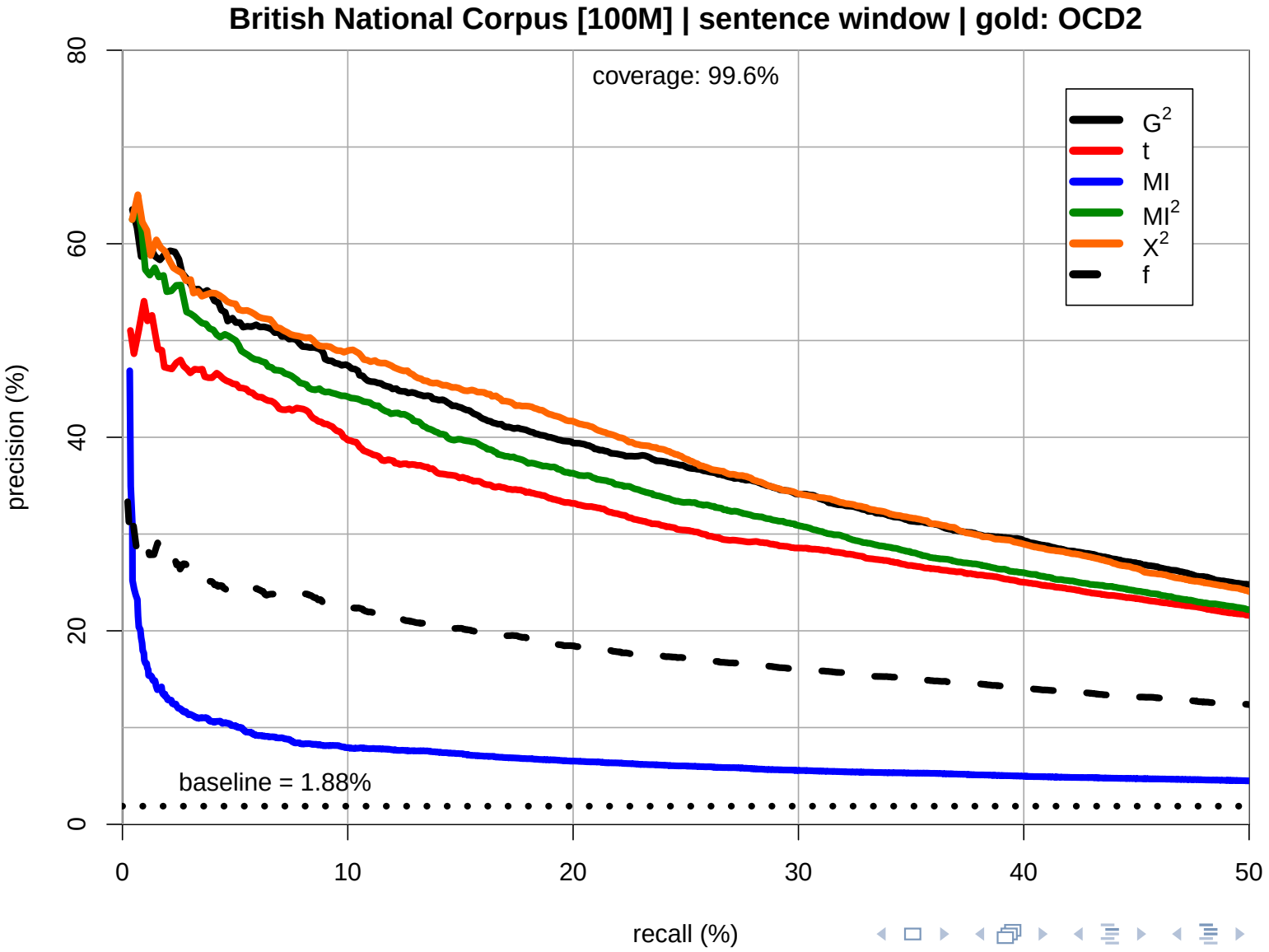
Replication on Oxford Collocations Dictionary (2nd ed.)

(McIntosh *et al.* 2009)

- Compilation of gold standard
  - automatic extraction from XML version of the dictionary
  - collocates explicitly marked ➜ no manual validation necessary
  - Bartsch224 nodes accepted as headwords and as collocates (OCD follows Hausmann's (1989) base-collocate distinction)
- Candidate extraction and ranking
  - same candidate data as for BBI experiments (lazy researcher)
  - ➥ incomplete coverage of OCD2 gold standard: 4,636 out of 18,515 collocations discarded (= 25.0%)

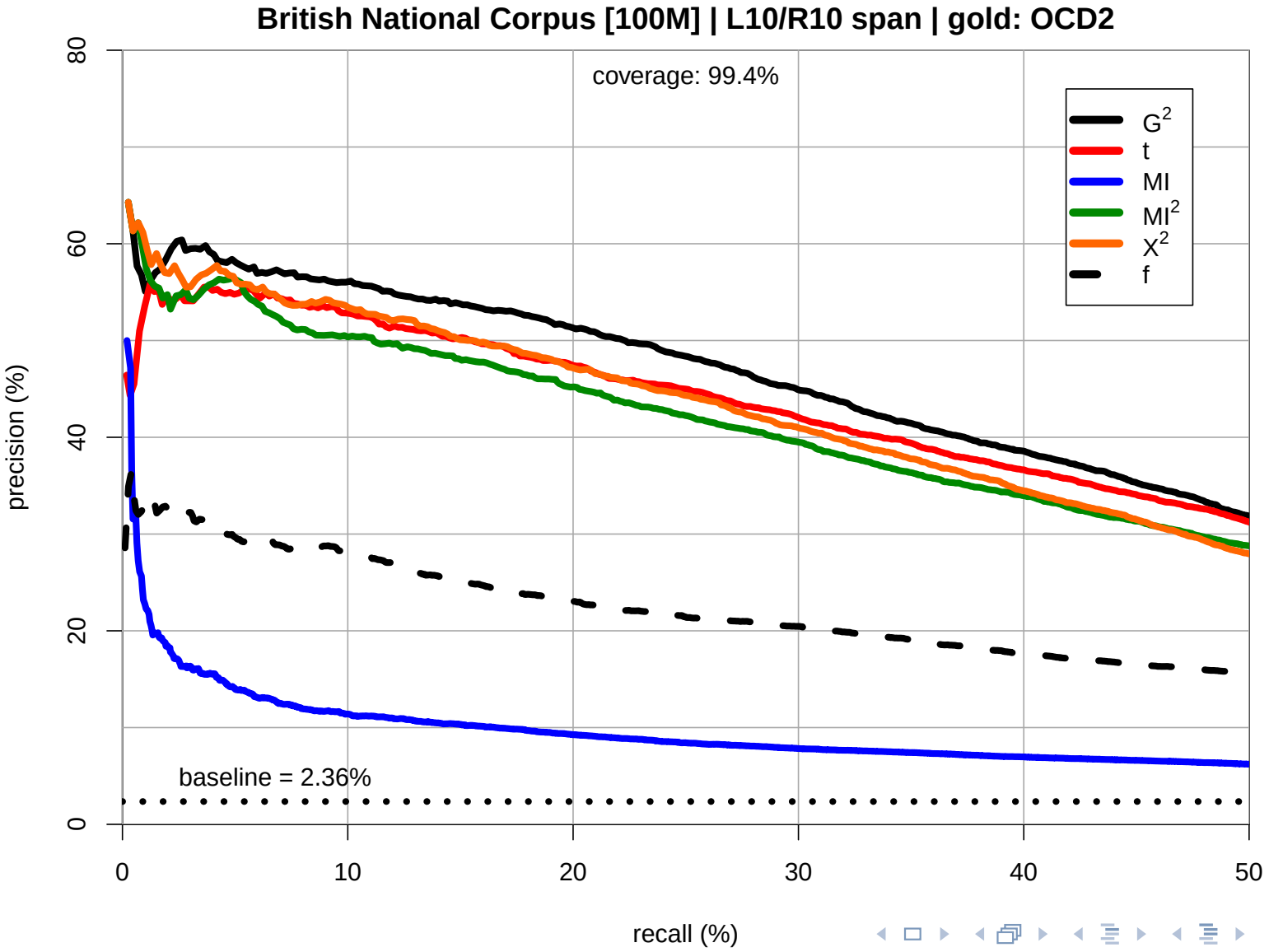
Evaluation on OCD2: collocational span

(McIntosh *et al.* 2009)



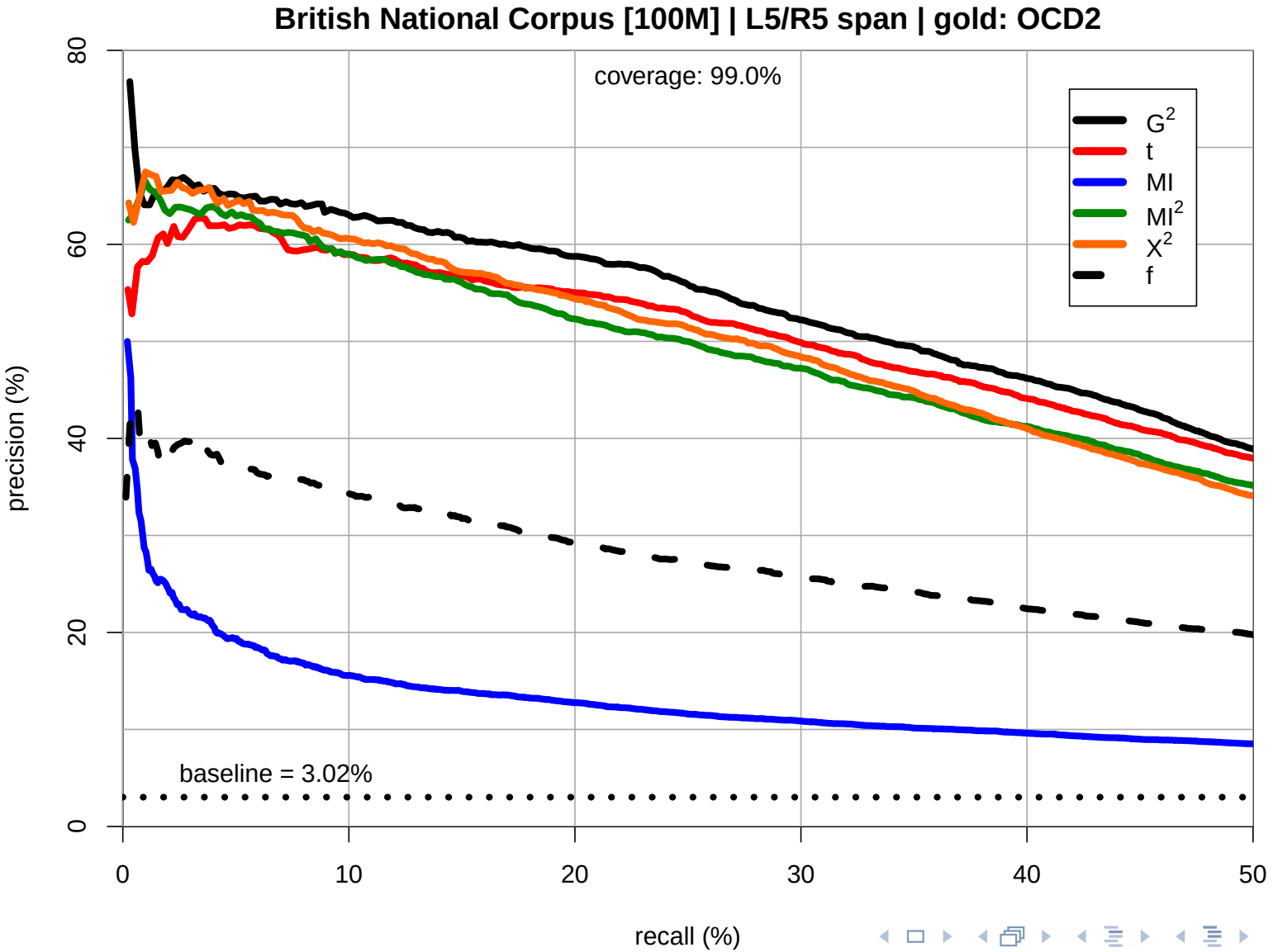
Evaluation on OCD2: collocational span

(McIntosh *et al.* 2009)



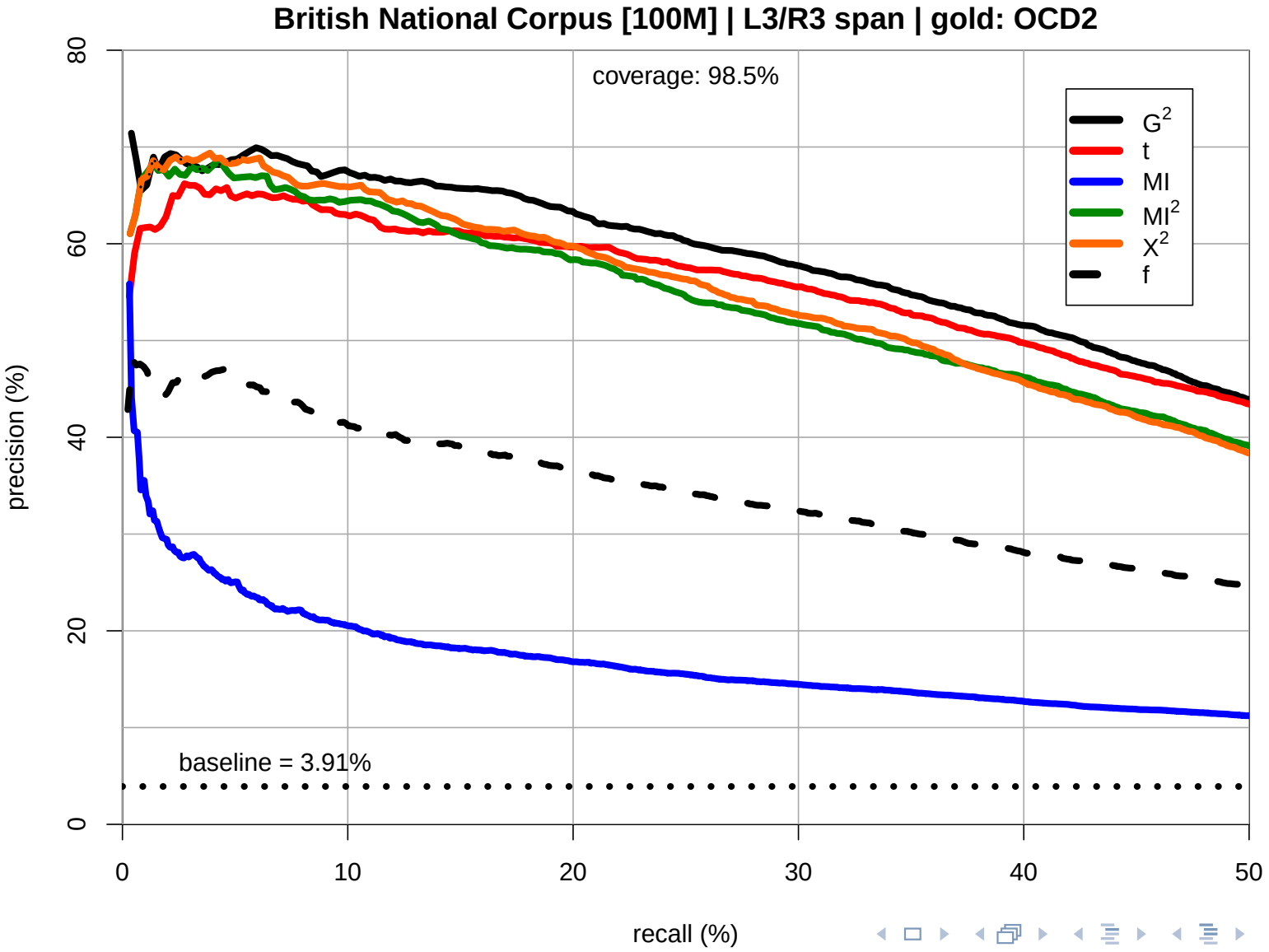
Evaluation on OCD2: collocational span

(McIntosh *et al.* 2009)



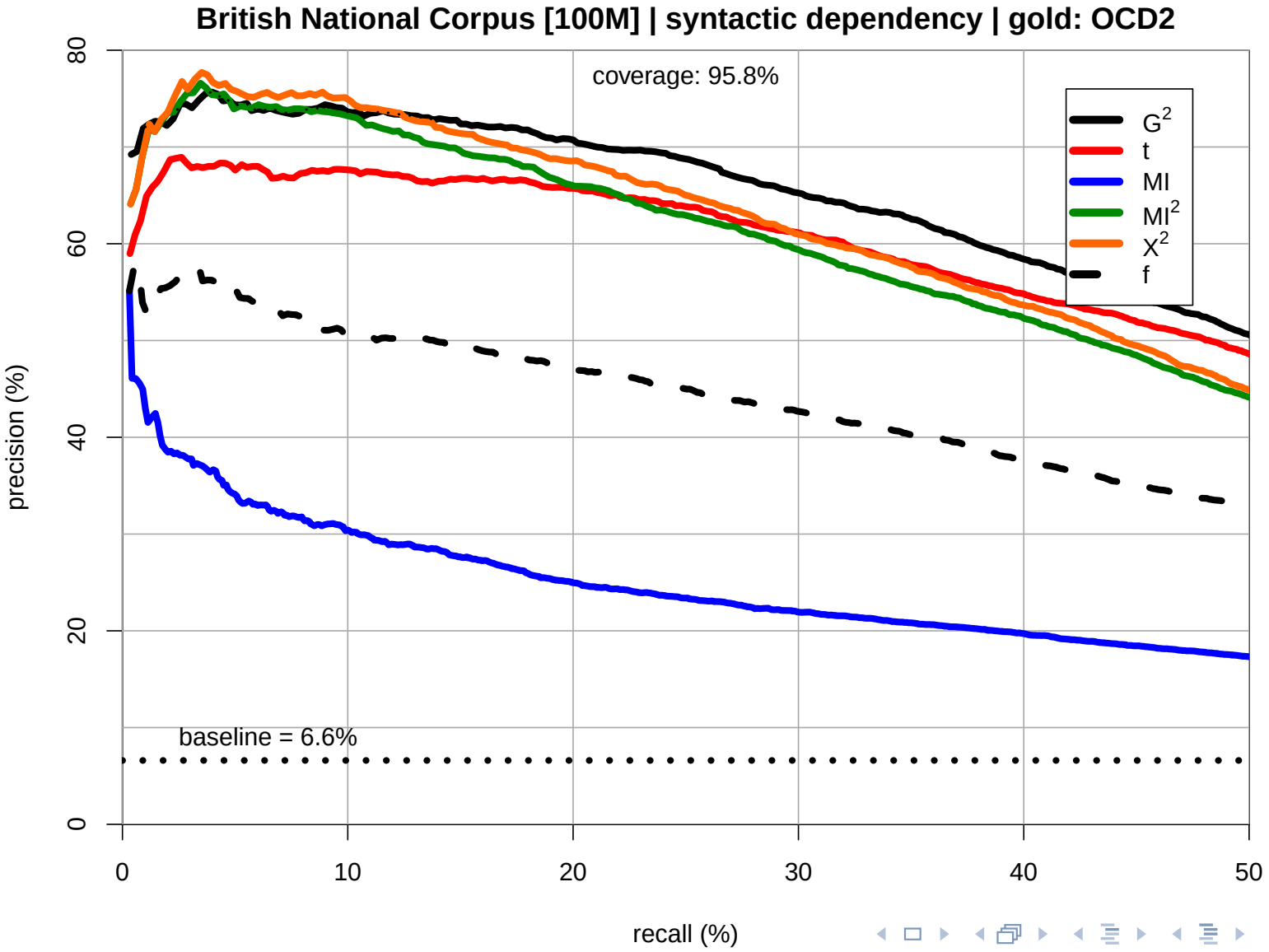
Evaluation on OCD2: collocational span

(McIntosh *et al.* 2009)



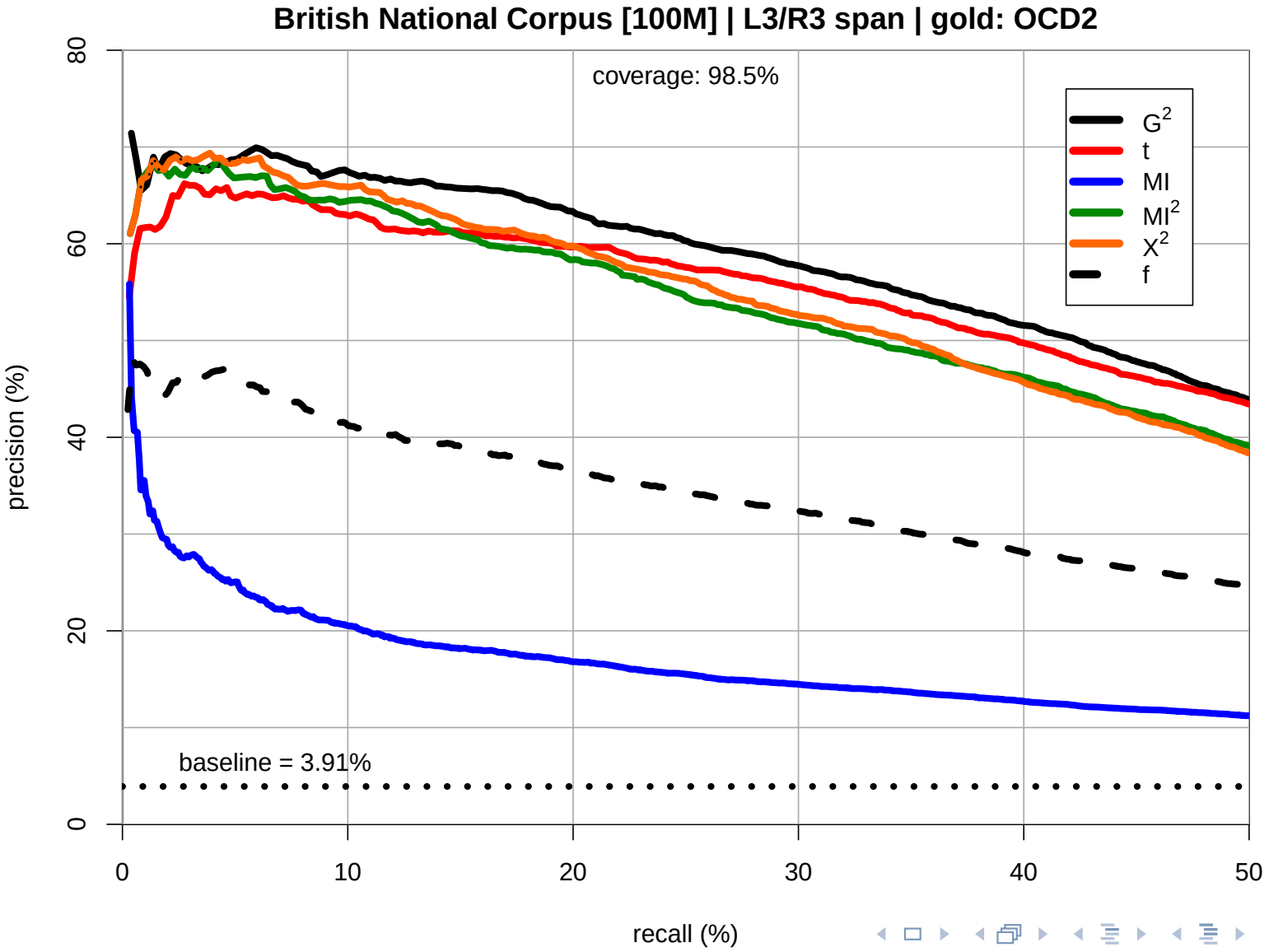
Evaluation on OCD2: collocational span

(McIntosh *et al.* 2009)



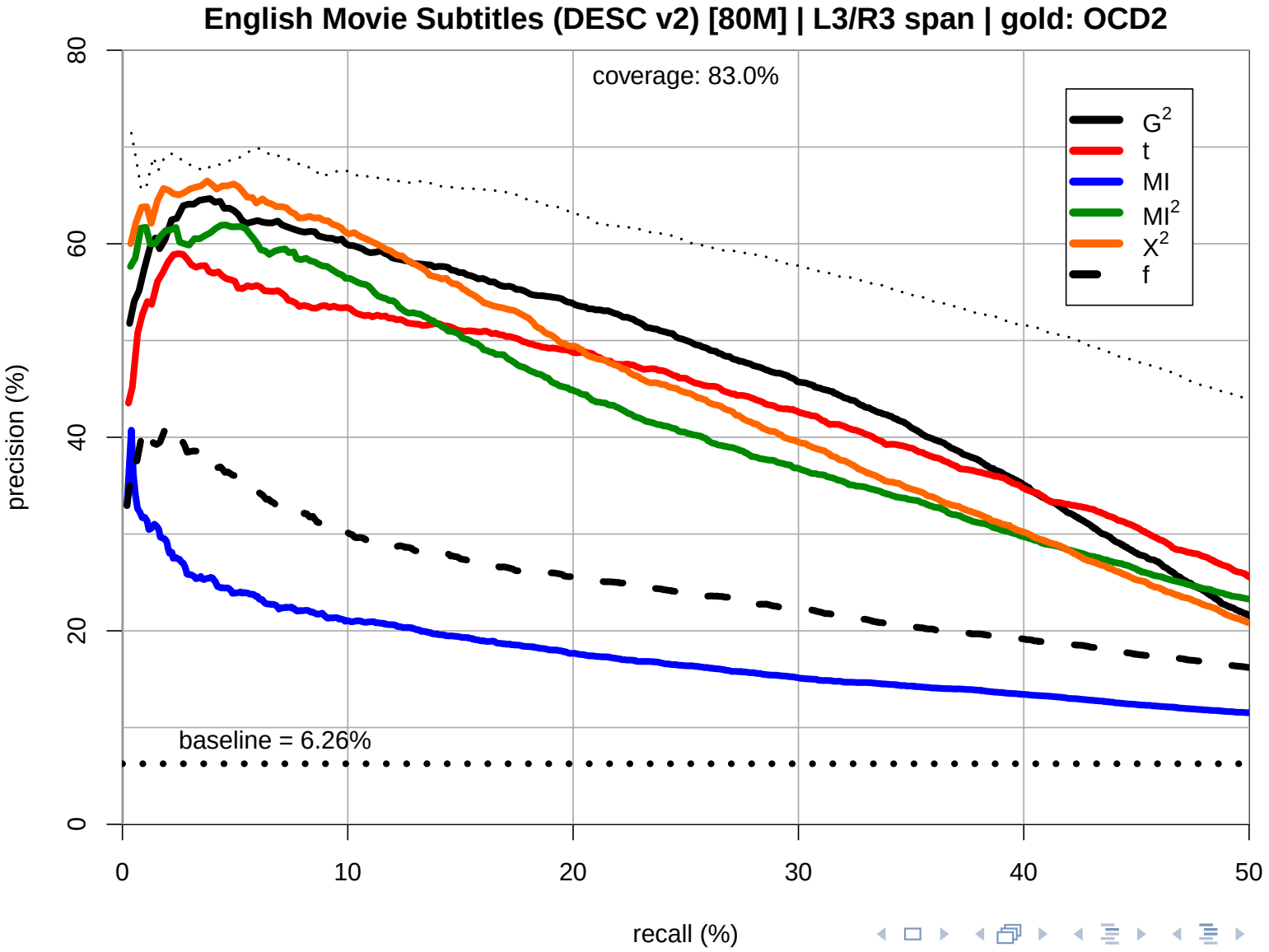
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



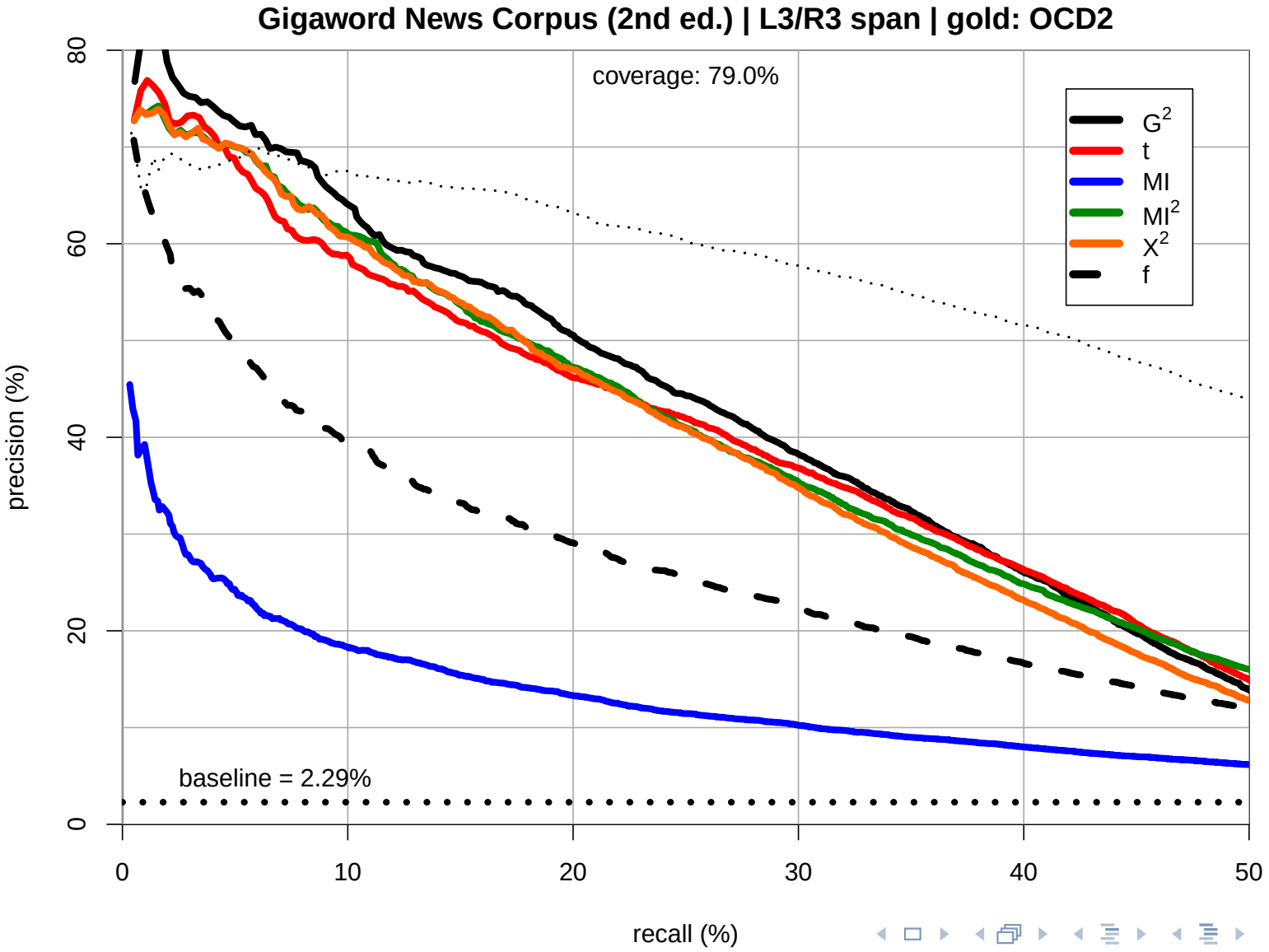
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



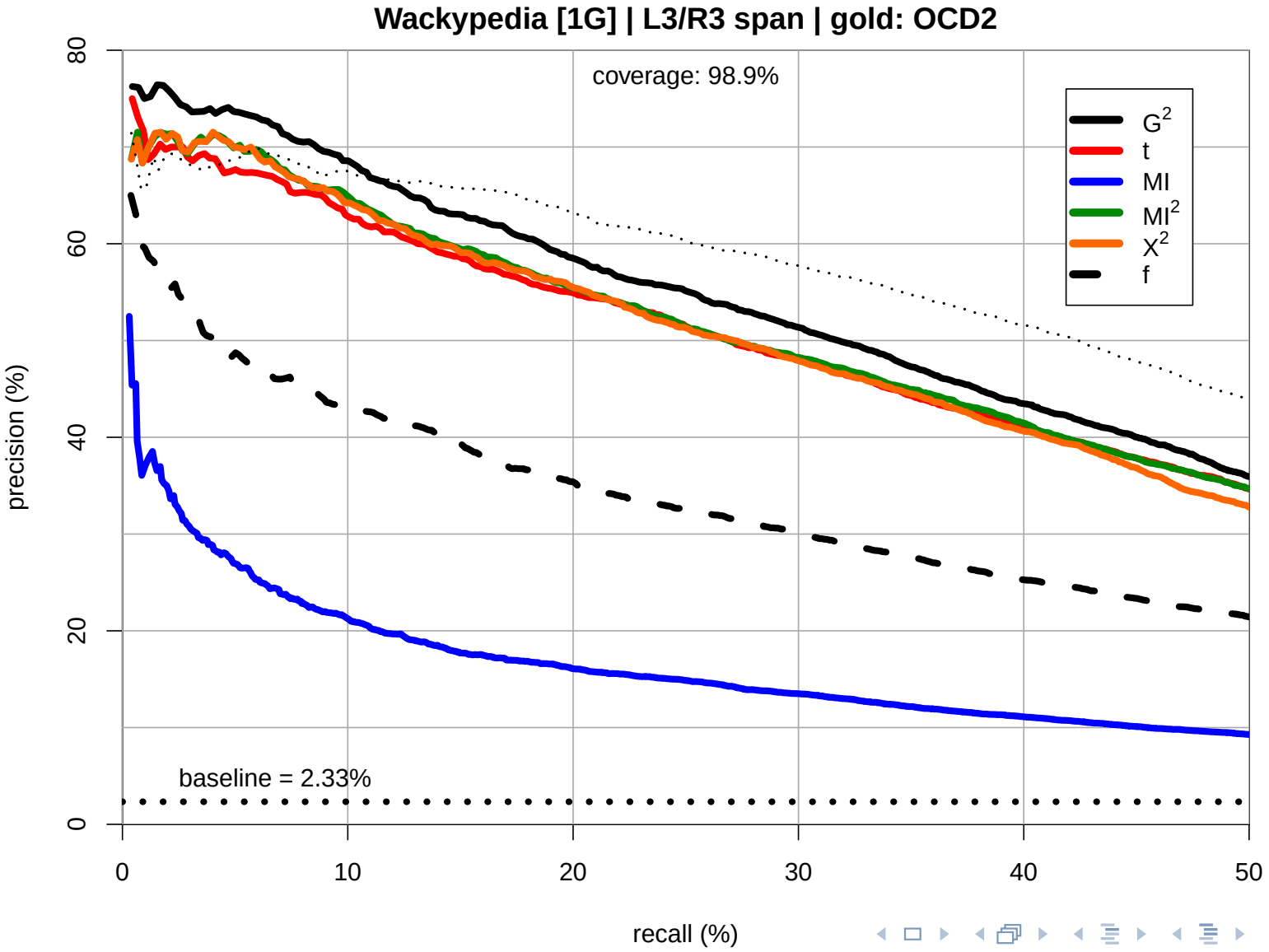
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



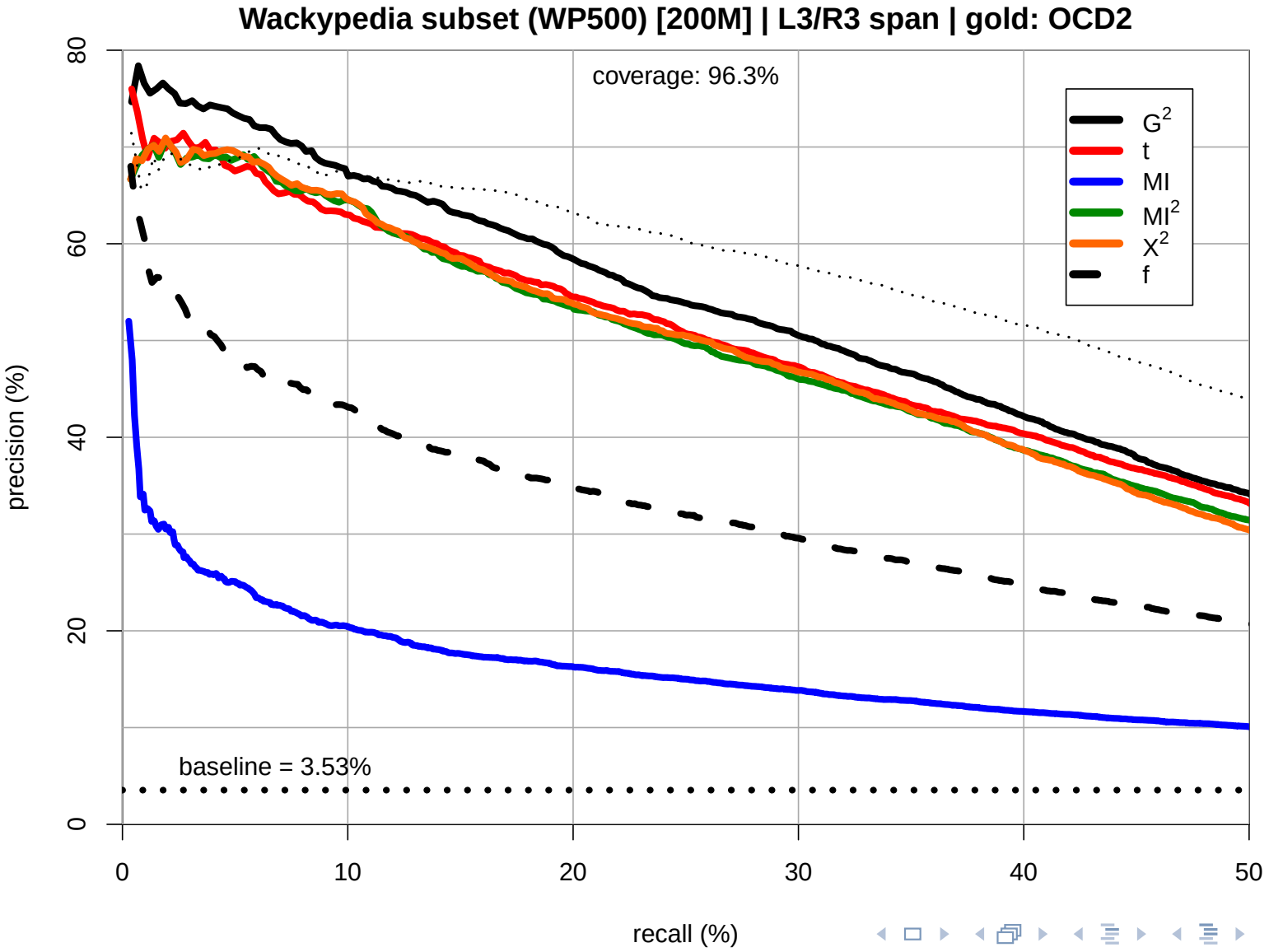
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



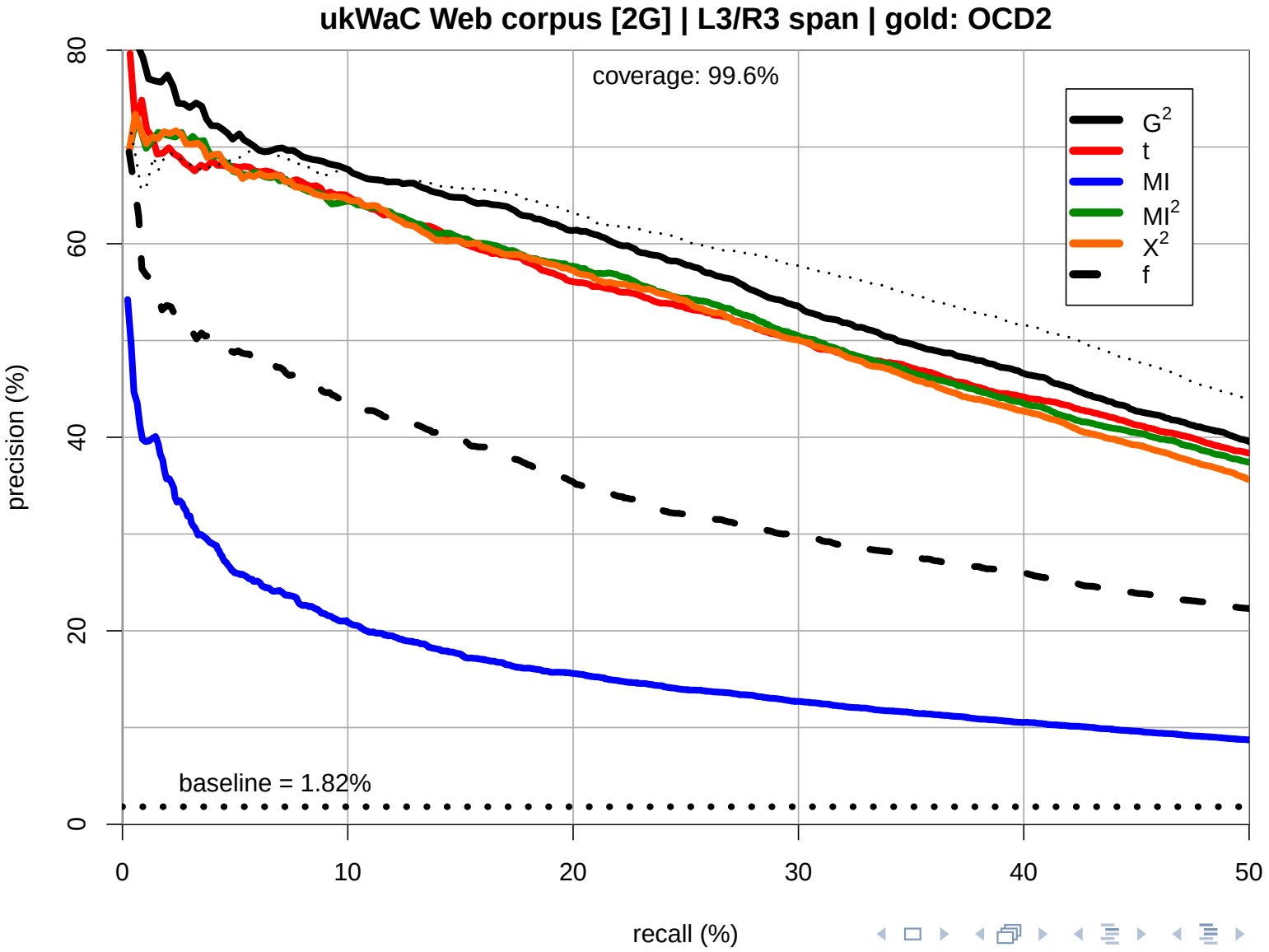
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



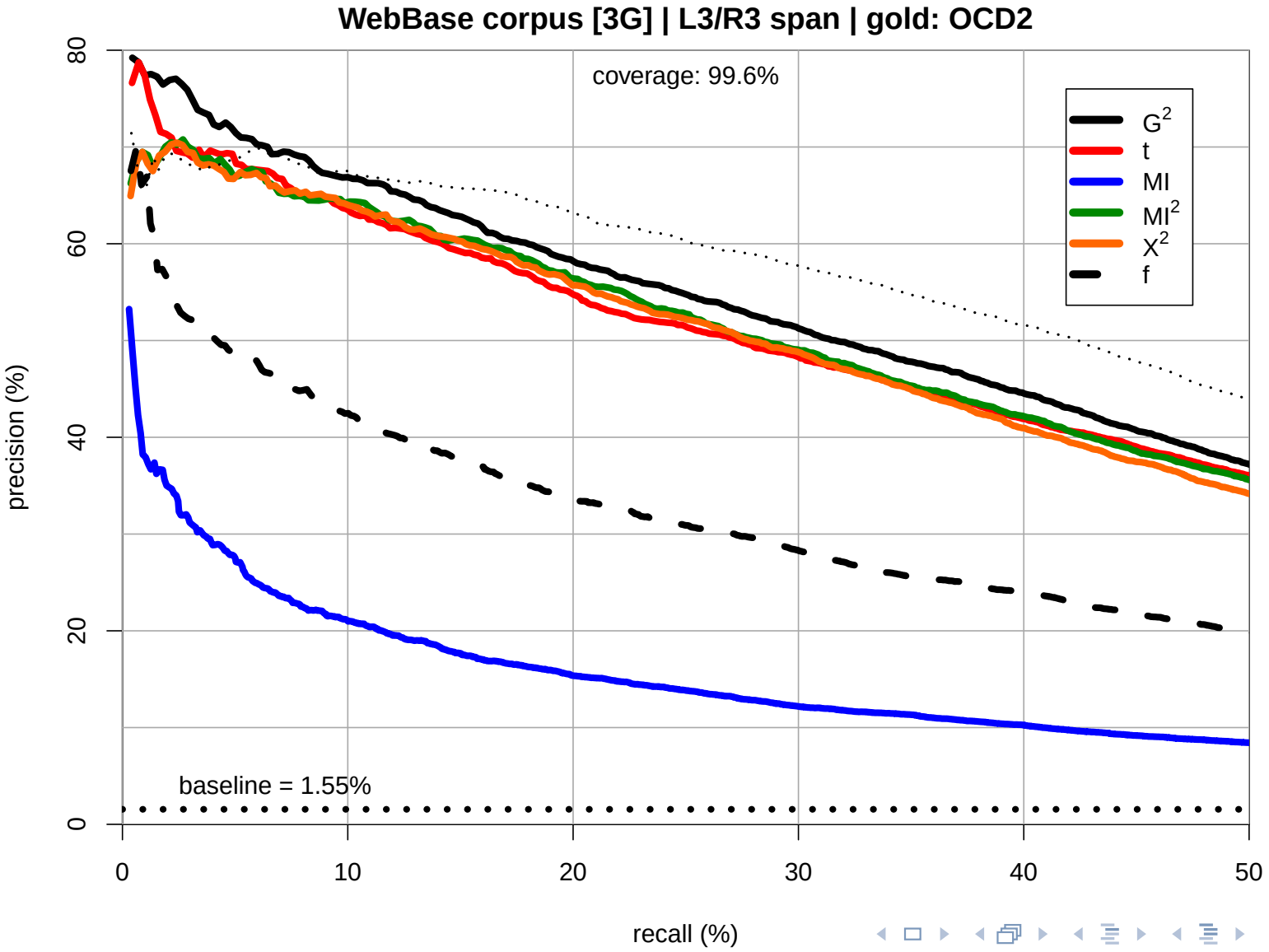
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



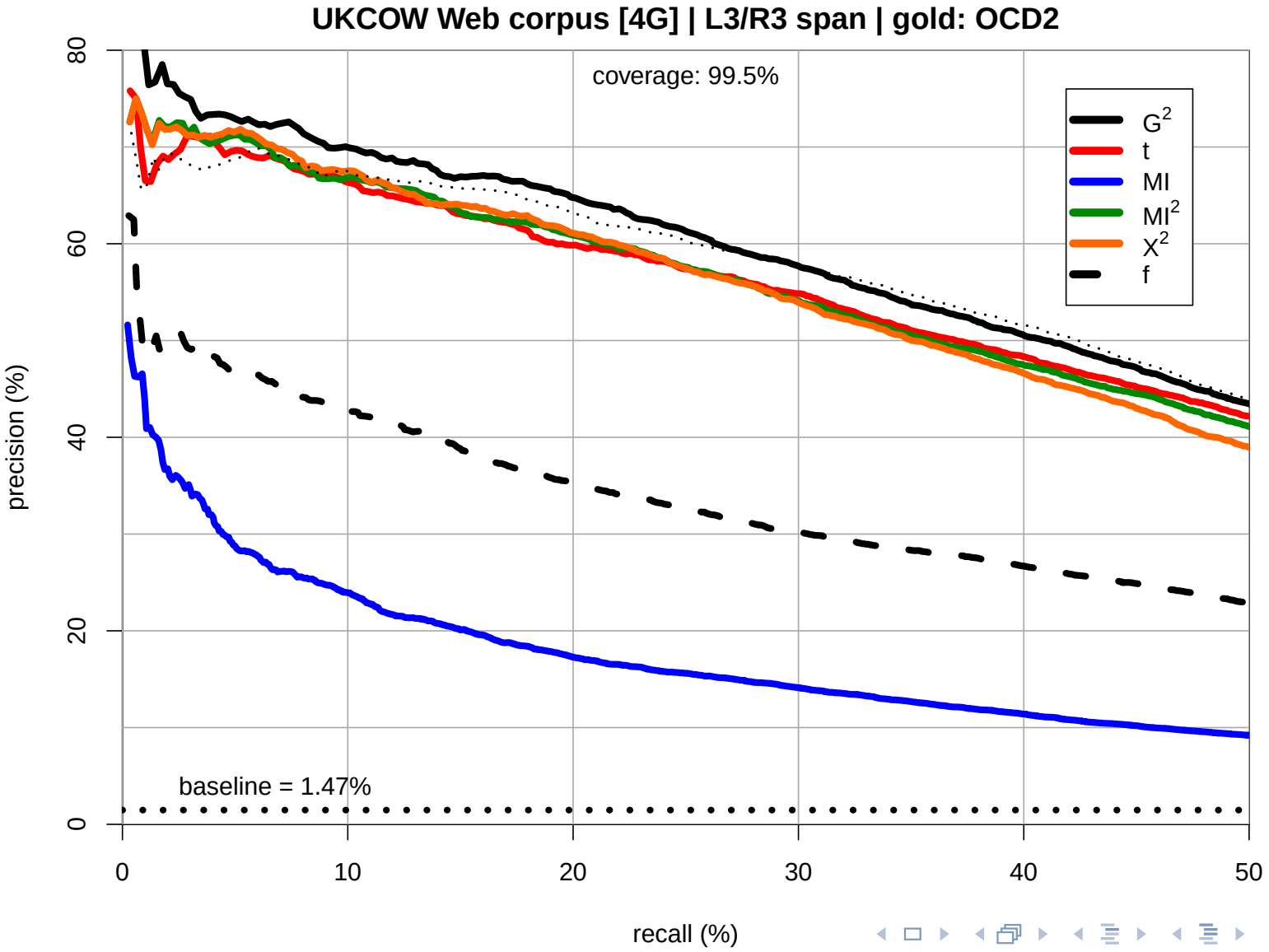
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



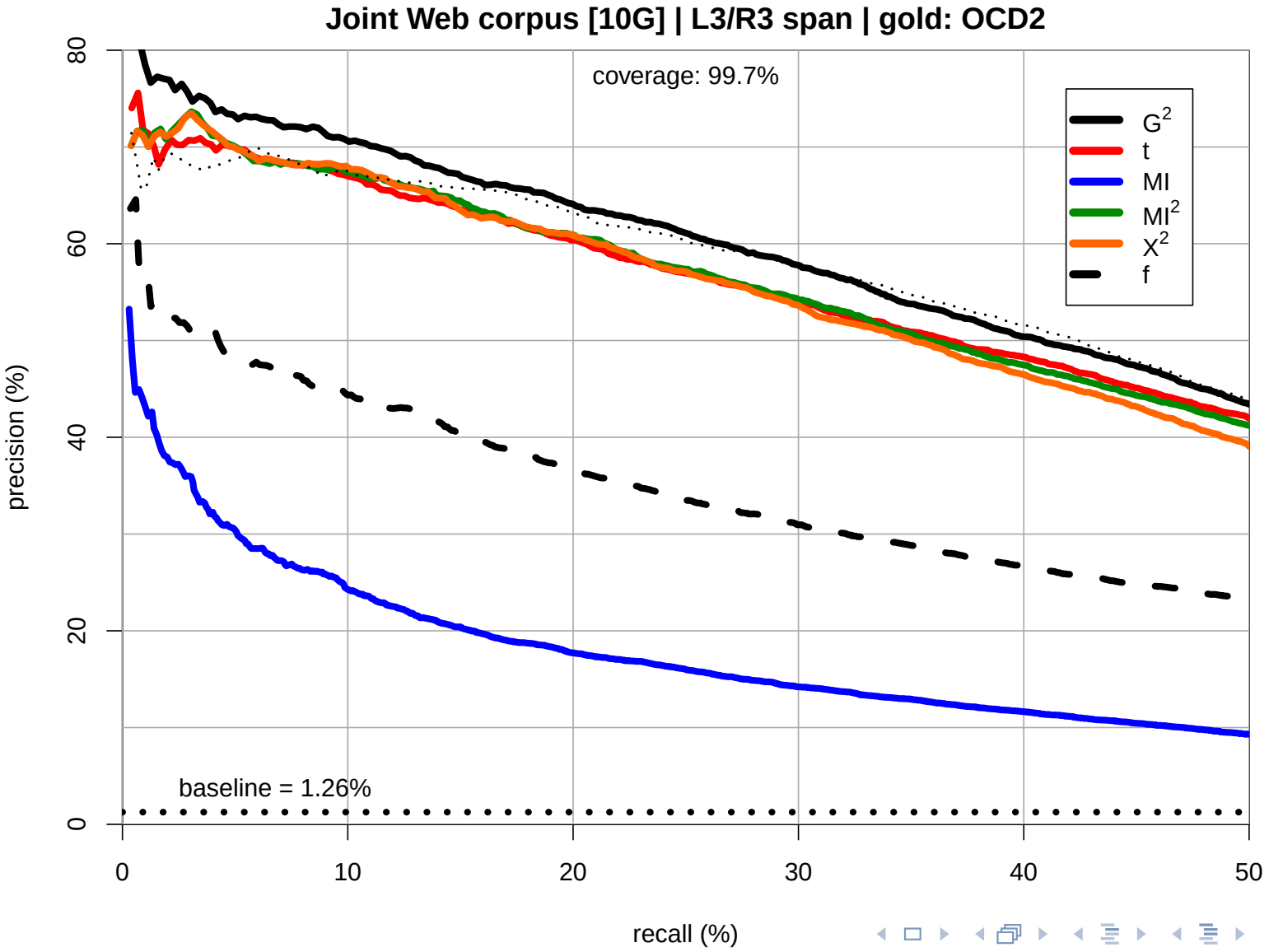
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



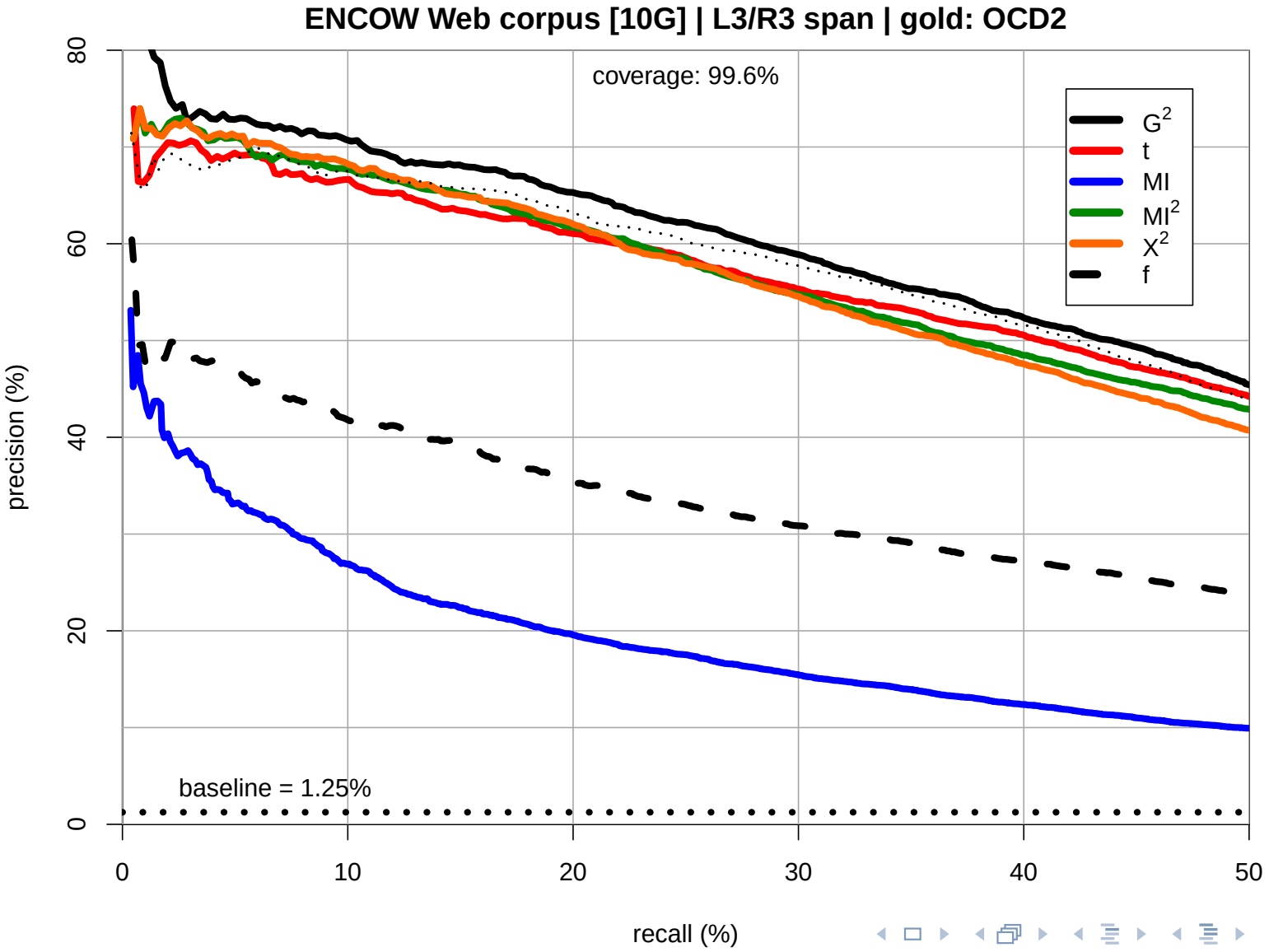
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



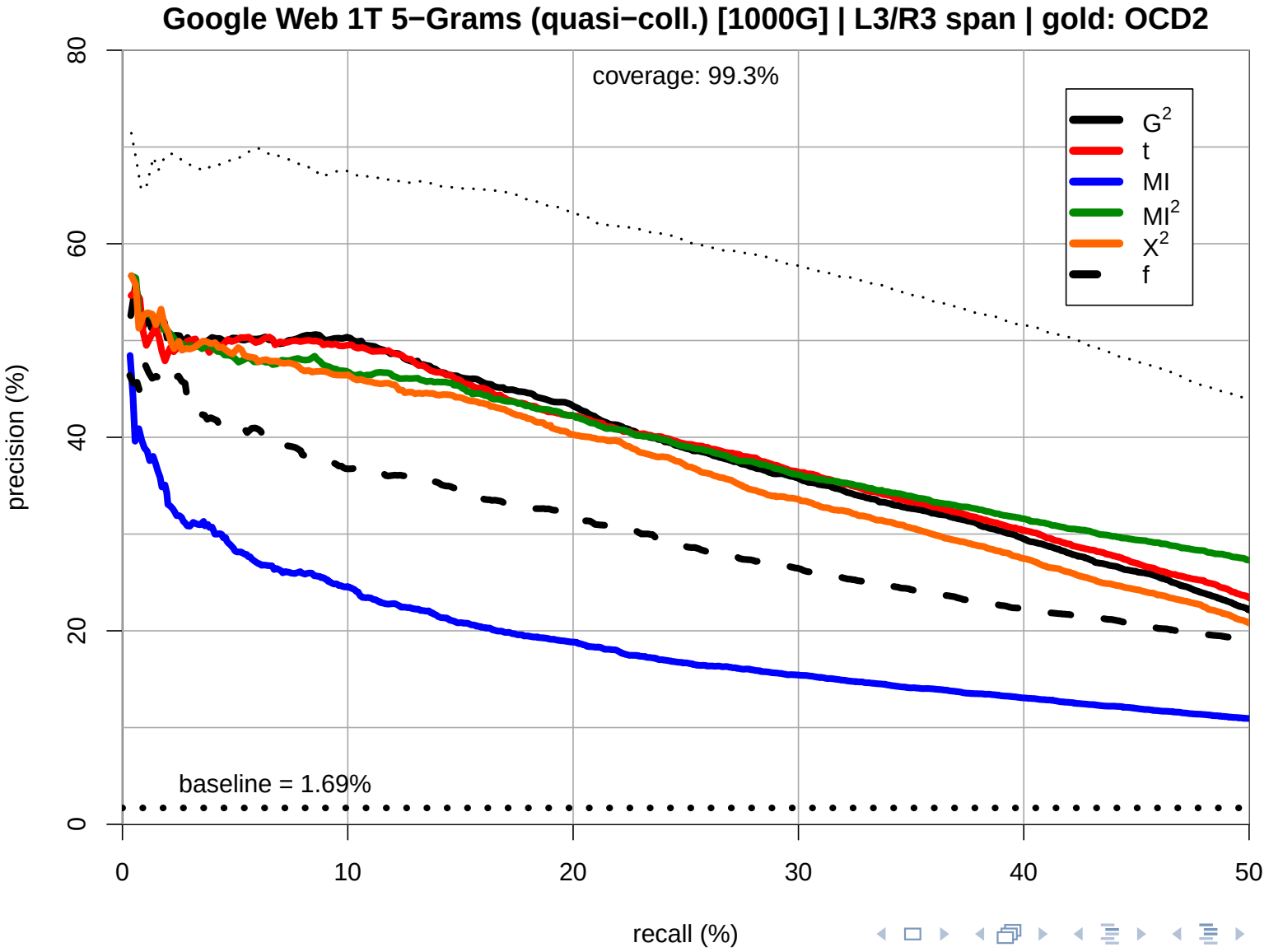
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



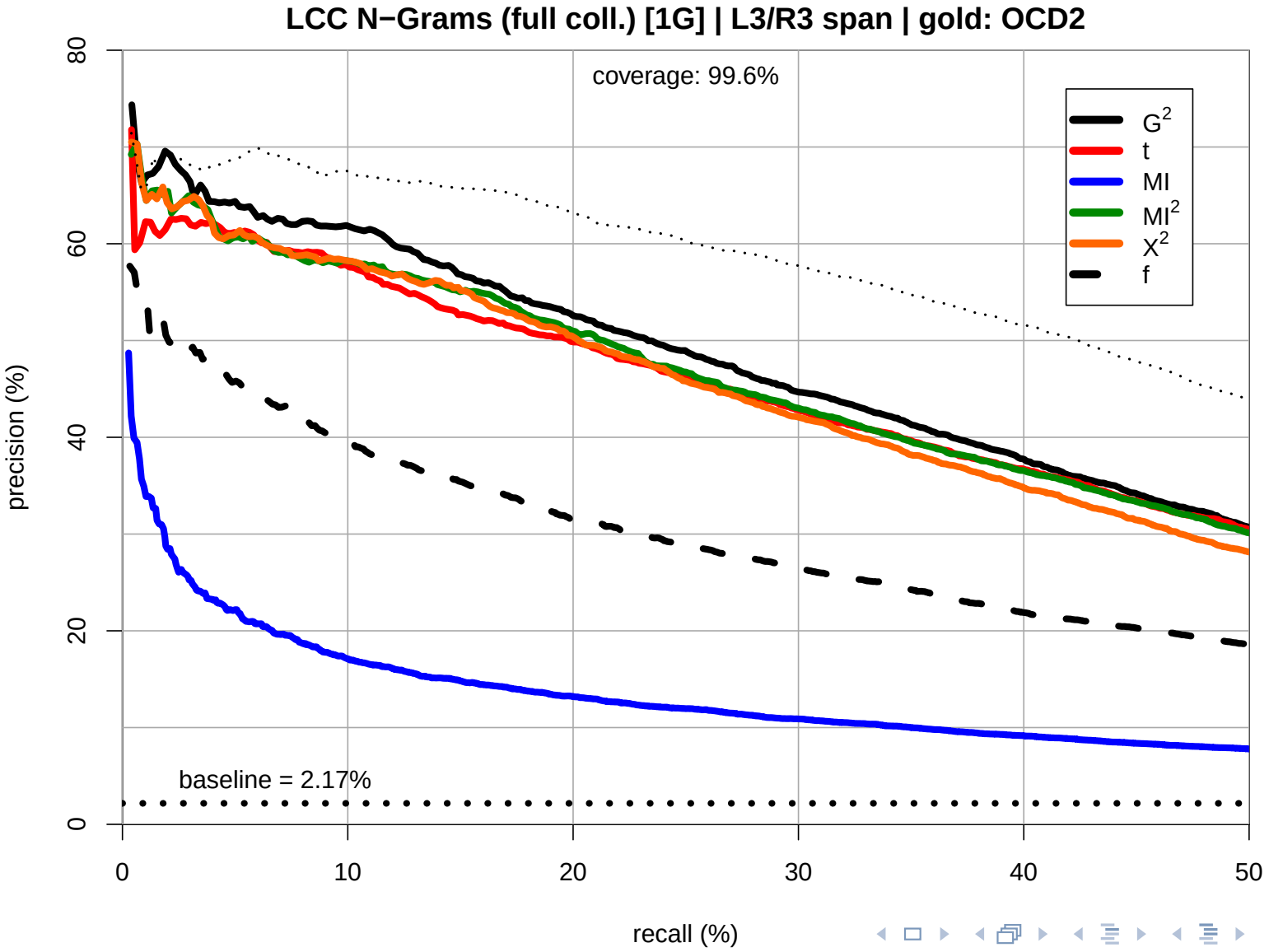
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



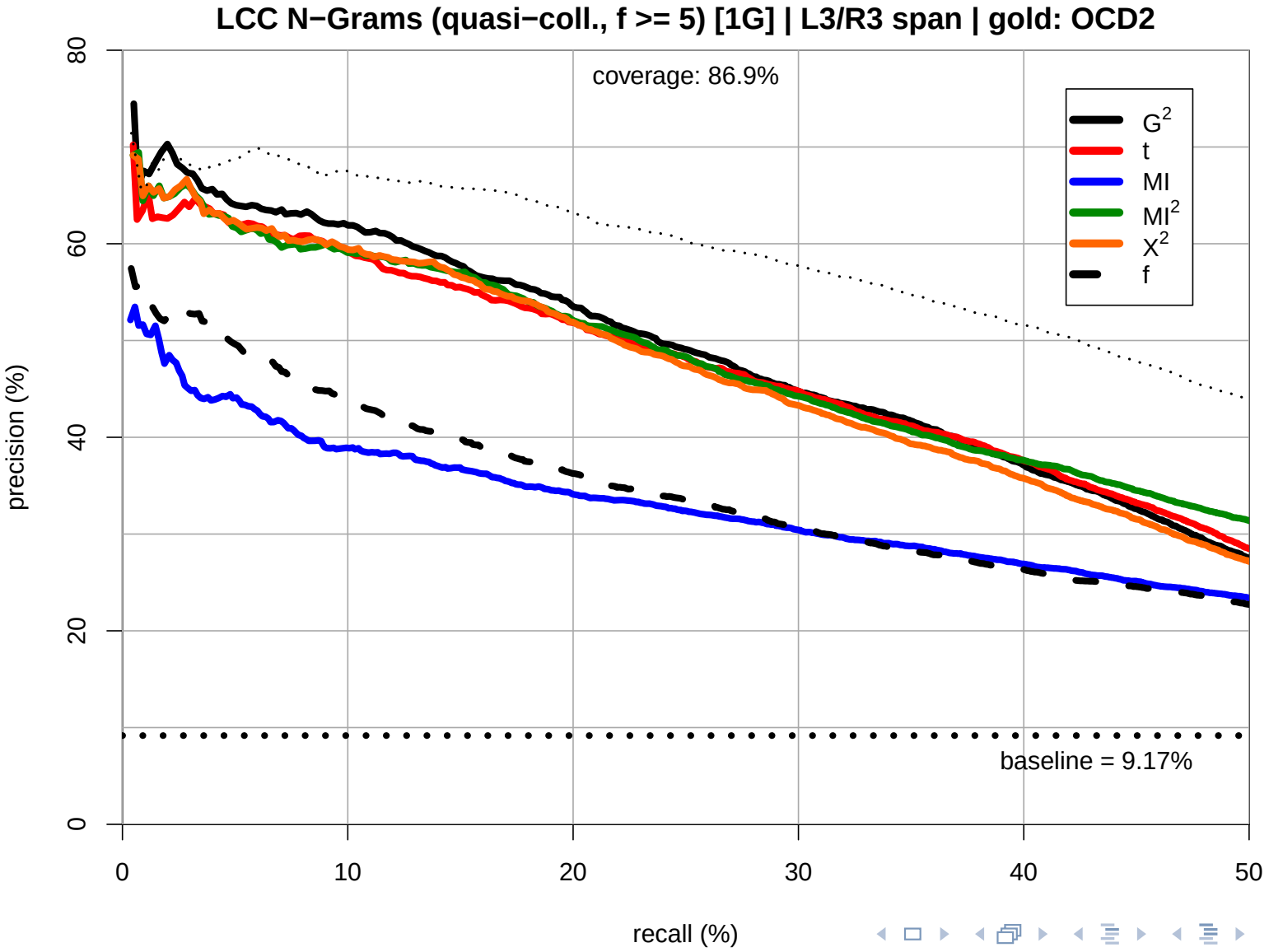
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



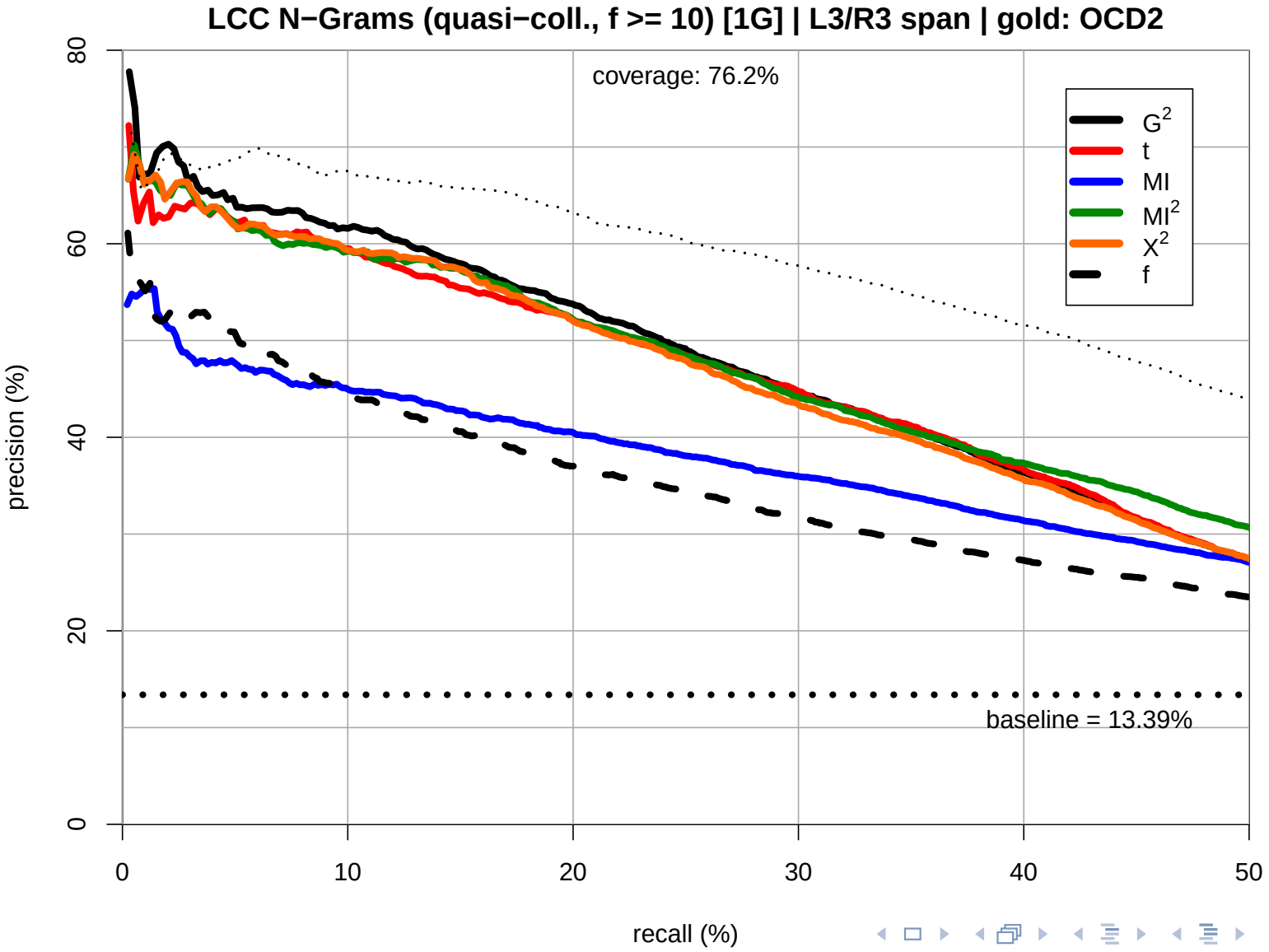
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



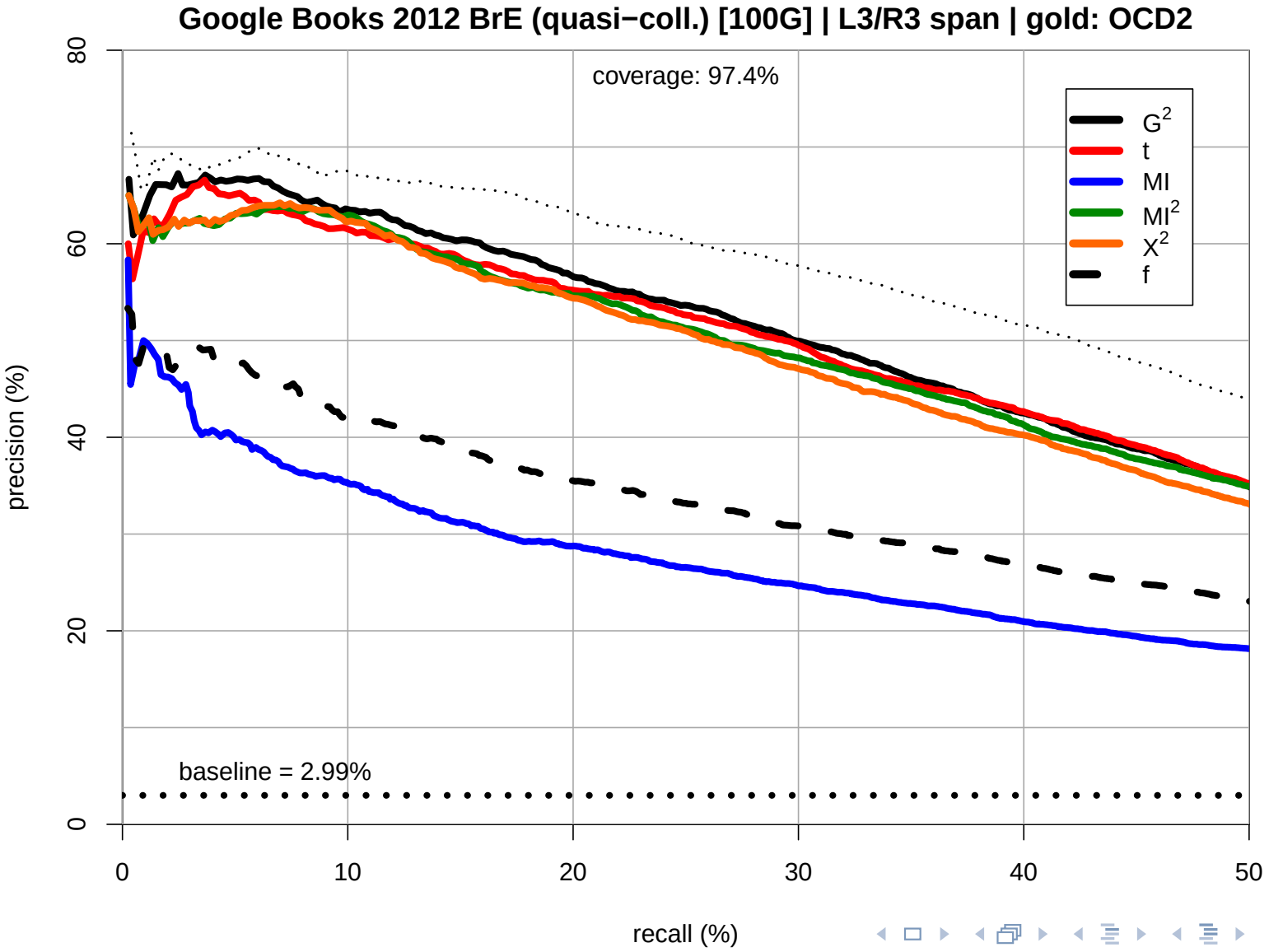
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



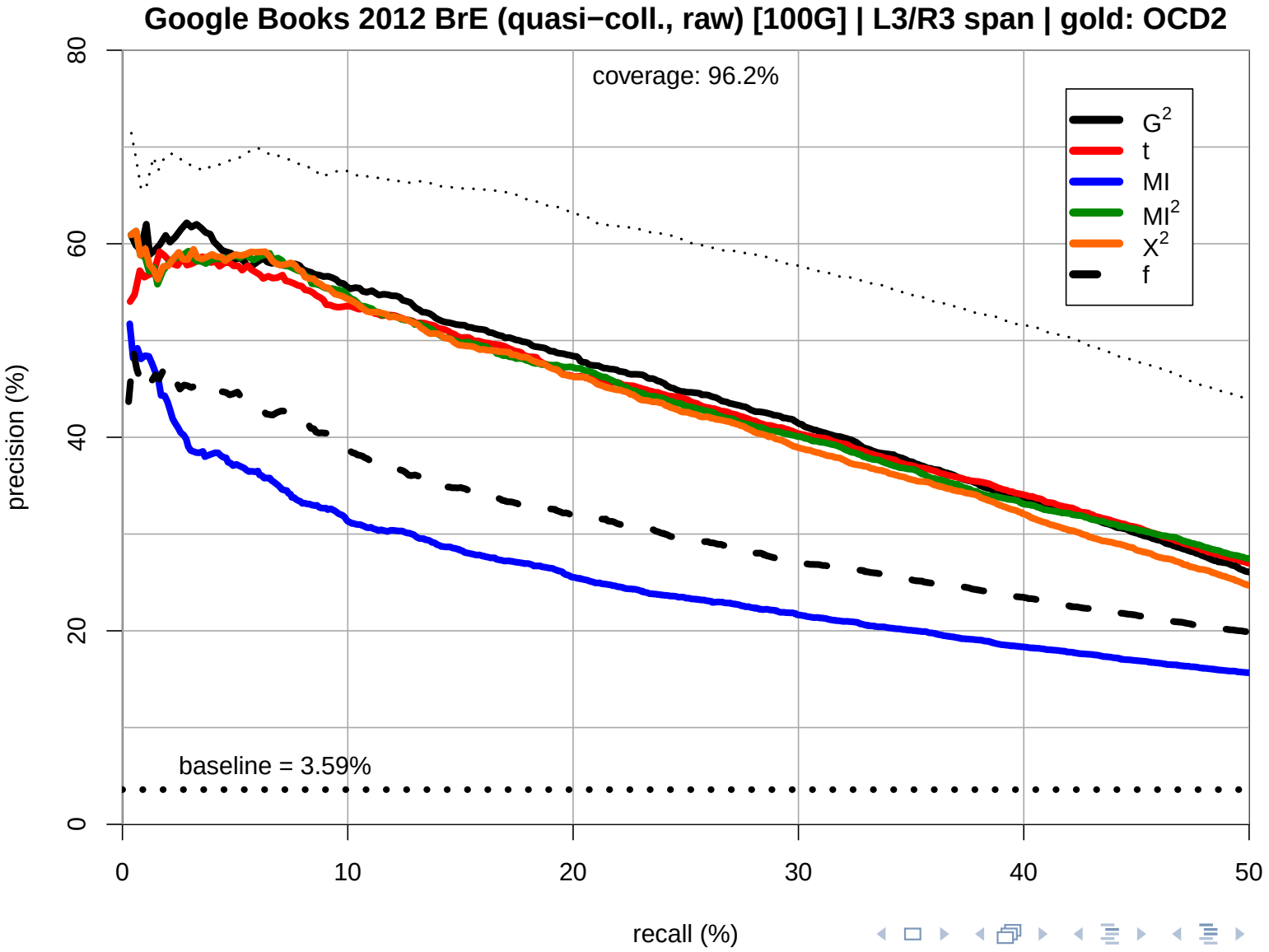
Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



Evaluation on OCD2: different corpora (L3/R3)

(McIntosh *et al.* 2009)



Distributional Semantics

Distributional semantics

- Distributional hypothesis (Harris 1954): meaning of a word can be inferred from its distribution across contexts

  "You shall know a word by the company it keeps!"

  — (Firth 1957)

Distributional semantics

- Distributional hypothesis (Harris 1954): meaning of a word can be inferred from its distribution across contexts

  "You shall know a word by the company it keeps!"

  — (Firth 1957)

- Reality check: What is the mystery word?
  - He handed her her glass of XXXXX.
  - Nigel staggered to his feet, face flushed from too much XXXXX.
  - Malbec, one of the lesser-known XXXXX grapes, responds well to Australia's sunshine.
  - I dined off bread and cheese and this excellent XXXXX.
  - The drinks were delicious: blood-red XXXXX as well as light, sweet Rhenish.

Distributional semantics

- Distributional hypothesis (Harris 1954): meaning of a word can be inferred from its distribution across contexts

  "You shall know a word by the company it keeps!"
  — (Firth 1957)

- Reality check: What is the mystery word?
  - He handed her her glass of XXXXX.
  - Nigel staggered to his feet, face flushed from too much XXXXX.
  - Malbec, one of the lesser-known XXXXX grapes, responds well to Australia's sunshine.
  - I dined off bread and cheese and this excellent XXXXX.
  - The drinks were delicious: blood-red XXXXX as well as light, sweet Rhenish.
- XXXXX = claret
  - all examples from BNC (carefully selected & slightly edited)

Distributional semantics

- A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

Distributional semantics

- A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	**51**	**20**	**84**	**0**	**3**	**0**
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

$$\text{sim}(\textbf{???}, \textbf{knife}) = 0.770$$

Distributional semantics

- A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	**12**	**17**	**3**	**2**	**9**	**27**
banana	11	2	2	0	18	0

$$\text{sim}(\textbf{???}, \textbf{pig}) = 0.939$$

Distributional semantics

- A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	**52**	**58**	**4**	**4**	**6**	**26**
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

$$\text{sim}(\textbf{???}, \textbf{cat}) = 0.961$$

Distributional semantics

- A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

??? = dog

Distributional semantics with Web1T5

- Basis of distributional semantic model (DSM): term-term **co-occurrence matrix** of collocational profiles
  - very sparse: e.g. 250k $\times$ 100k matrix with 24.2 billion cells, but only 245.4 million cells ($\approx$ 1%) have nonzero values

Distributional semantics with Web1T5

- Basis of distributional semantic model (DSM): term-term **co-occurrence matrix** of collocational profiles
  - very sparse: e.g. 250k $\times$ 100k matrix with 24.2 billion cells, but only 245.4 million cells ($\approx$ 1%) have nonzero values
- We've already computed collocational profiles
  - 32 GiB collocations database = sparse co-occurrence matrix
  - export matrix with 25k target words (rows) and 50k high-frequency word forms as features (columns)

Distributional semantics with Web1T5

- Basis of distributional semantic model (DSM): term-term **co-occurrence matrix** of collocational profiles
  - very sparse: e.g. 250k $\times$ 100k matrix with 24.2 billion cells, but only 245.4 million cells ($\approx 1\%$) have nonzero values
- We've already computed collocational profiles
  - 32 GiB collocations database = sparse co-occurrence matrix
  - export matrix with 25k target words (rows) and 50k high-frequency word forms as features (columns)
- DSM implemented in **R** (with wordspace package)
  - column-compressed sparse matrix
  - log G^2 weights, L_2-normalized, angular distance (= cosine), 500 latent dimensions + 50 skipped (randomized SVD)
  - parameter settings according to Lapesa and Evert (2014)
  - needs 10 GiB RAM and less than an hour

DSM with Web1T5: nearest neighbours

Neighbours of **linguistics** (cosine angle):

- ☞ sociology (24.6), sociolinguistics (24.6), criminology (29.5), anthropology (30.8), mathematics (31.2), phonetics (33.1), phonology (33.2), philology (33.2), literatures (33.5), gerontology (35.3), proseminar (35.5), geography (35.8), humanities (35.9), archaeology (35.9), science (36.5), ...

DSM with Web1T5: nearest neighbours

Neighbours of **linguistics** (cosine angle):

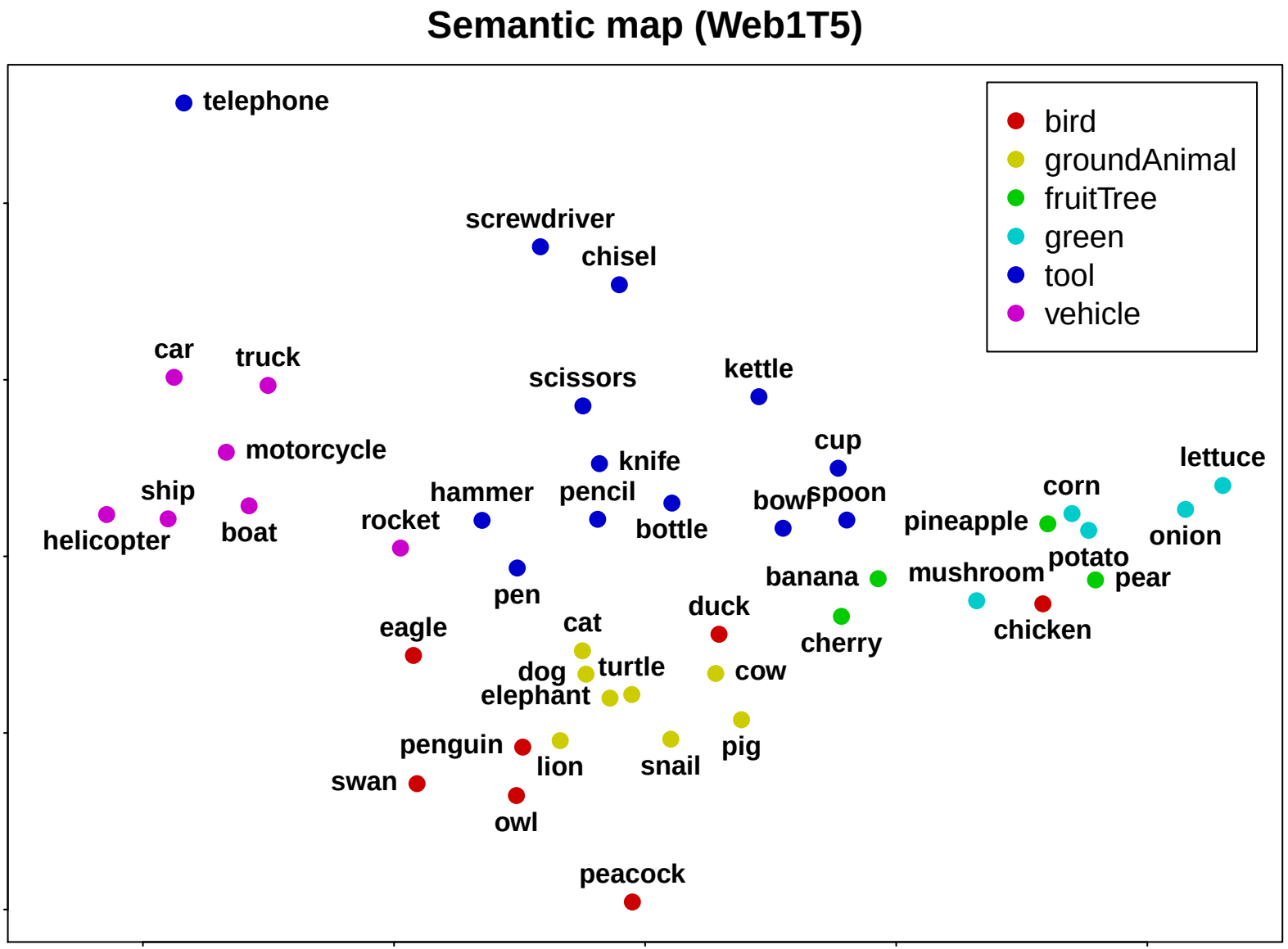
- ☞ sociology (24.6), sociolinguistics (24.6), criminology (29.5), anthropology (30.8), mathematics (31.2), phonetics (33.1), phonology (33.2), philology (33.2), literatures (33.5), gerontology (35.3), proseminar (35.5), geography (35.8), humanities (35.9), archaeology (35.9), science (36.5), ...

Neighbours of **spaniel** (cosine angle):

- ☞ terrier (23.0), schnauzer (26.5), pinscher (27.0), weimaraner (28.3), keeshond (29.1), pomeranian (29.4), pekingese (29.6), bichon (30.1), vizsla (30.5), labradoodle (30.6), apso (31.1), spaniels (32.0), frise (32.0), yorkie (32.1), sheepdog (32.3), dachshund (32.4), retriever (32.7), whippet (32.9), havanese (33.1), westie (34.5), mastiff (34.6), dandie (34.7), chihuahua (34.9), dinmont (35.0), elkhound (35.0), ...

DSM with Web1T5: semantic map

(data from ESSLLI 2008 shared task on concrete noun categorization)



Evaluating distributional similarity

- Correlation with human smilarity ratings
  - RG65: Rubenstein and Goodenough (1965)
  - WordSim-353: Finkelstein *et al.* (2002)

Evaluating distributional similarity

- Correlation with human smilarity ratings
  - RG65: Rubenstein and Goodenough (1965)
  - WordSim-353: Finkelstein *et al.* (2002)
- Multiple-choice tasks
  - TOEFL synonym questions
  - SAT analogy questions

Evaluating distributional similarity

- Correlation with human smilarity ratings
  - RG65: Rubenstein and Goodenough (1965)
  - WordSim-353: Finkelstein *et al.* (2002)
- Multiple-choice tasks
  - TOEFL synonym questions
  - SAT analogy questions
- Decision tasks for consistent/inconsistent prime
  - SPP: Semantic Priming Project (Hutchison *et al.* 2013)
  - GEK: Generalized Event Knowledge (Ferretti *et al.* 2001; McRae *et al.* 2005; Hare *et al.* 2009)

Evaluating distributional similarity

- Correlation with human smilarity ratings
  - RG65: Rubenstein and Goodenough (1965)
  - WordSim-353: Finkelstein *et al.* (2002)
- Multiple-choice tasks
  - TOEFL synonym questions
  - SAT analogy questions
- Decision tasks for consistent/inconsistent prime
  - SPP: Semantic Priming Project (Hutchison *et al.* 2013)
  - GEK: Generalized Event Knowledge (Ferretti *et al.* 2001; McRae *et al.* 2005; Hare *et al.* 2009)
- Noun clustering vs. semantic classification
  - AP: Almuhareb/Poesio (Almuhareb 2006)
  - Battig: Battig/Montague norms (Van Overschelde *et al.* 2004)
  - ESSLLI: basic-level concrete nouns (ESSLLI '08 shared task)

Evaluating distributional similarity

- Correlation with human smilarity ratings
  - RG65: Rubenstein and Goodenough (1965)
  - WordSim-353: Finkelstein *et al.* (2002)
- Multiple-choice tasks
  - TOEFL synonym questions
  - SAT analogy questions
- Decision tasks for consistent/inconsistent prime
  - SPP: Semantic Priming Project (Hutchison *et al.* 2013)
  - GEK: Generalized Event Knowledge (Ferretti *et al.* 2001; McRae *et al.* 2005; Hare *et al.* 2009)
- Noun clustering vs. semantic classification
  - AP: Almuhareb/Poesio (Almuhareb 2006)
  - Battig: Battig/Montague norms (Van Overschelde *et al.* 2004)
  - ESSLLI: basic-level concrete nouns (ESSLLI '08 shared task)
- Psycholinguistic norms & experiments
  - free association norms
  - property norms
  - priming effects (ΔRT)

Evaluating distributional similarity

- Correlation with human smilarity ratings
  - ☞ RG65: Rubenstein and Goodenough (1965)
  - ☞ WordSim-353: Finkelstein *et al.* (2002)
- Multiple-choice tasks
  - ☞ TOEFL synonym questions
    - SAT analogy questions
- Decision tasks for consistent/inconsistent prime
  - ☞ SPP: Semantic Priming Project (Hutchison *et al.* 2013)
  - ☞ GEK: Generalized Event Knowledge (Ferretti *et al.* 2001; McRae *et al.* 2005; Hare *et al.* 2009)
- Noun clustering vs. semantic classification
  - ☞ AP: Almuhareb/Poesio (Almuhareb 2006)
  - ☞ Battig: Battig/Montague norms (Van Overschelde *et al.* 2004)
  - ☞ ESSLLI: basic-level concrete nouns (ESSLLI '08 shared task)
- Psycholinguistic norms & experiments
  - free association norms
  - property norms
  - priming effects (ΔRT)

Correlation with human similarity ratings

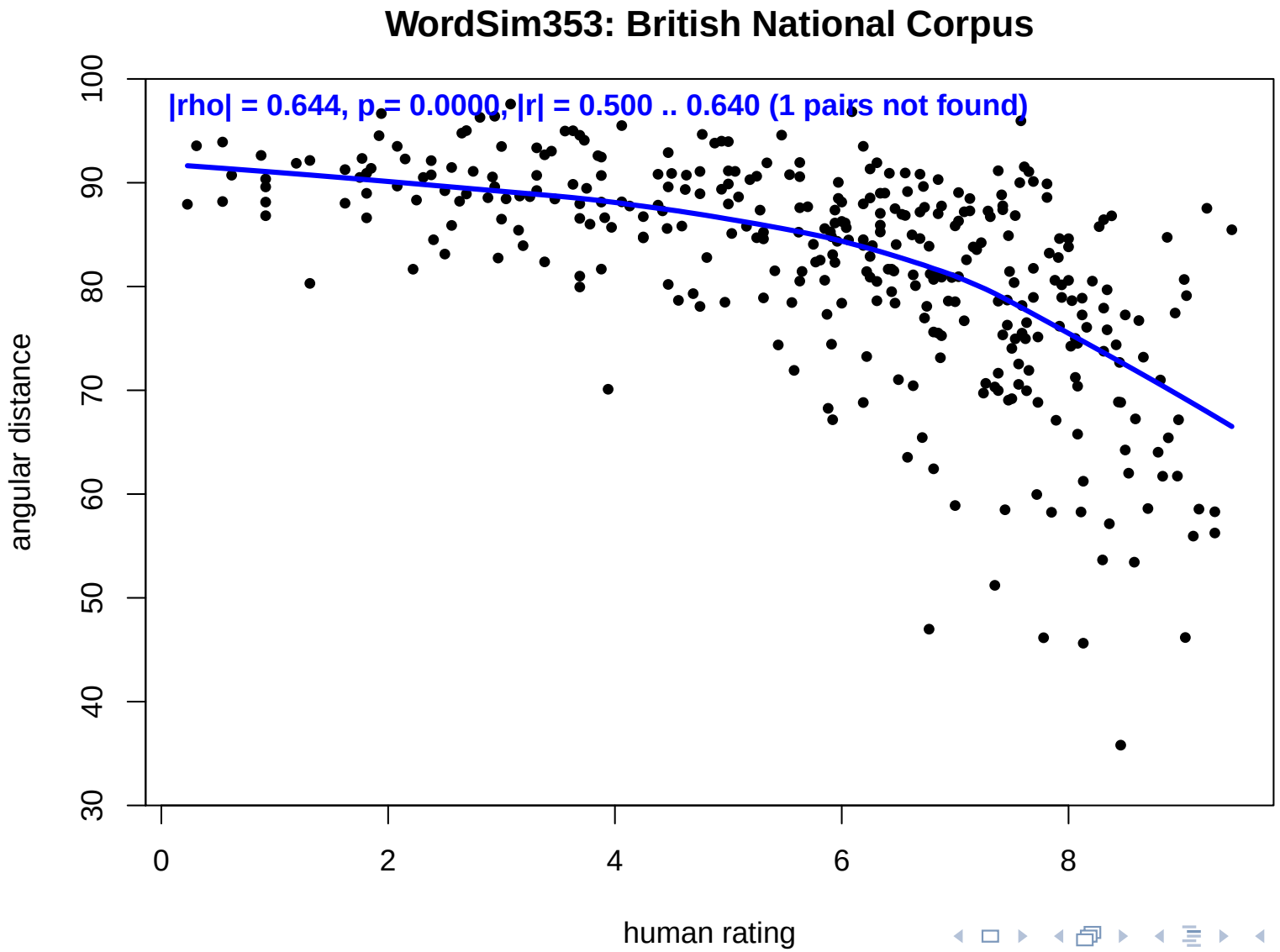
- Direct comparison with semantic similarity ratings (WordSim-353, Finkelstein *et al.* 2002)
  - 353 noun-noun pairs with "relatedness" ratings
  - rated on scale 0–10 by 16 test subjects
  - closely related: *money*/*cash*, *soccer*/*football*, *type*/*kind*, ...
  - unrelated: *king*/*cabbage*, *noon*/*string*, *sugar*/*approach*, ...
  - NB: not all "nouns" are nouns in a traditional sense (*five*, *live*, *eat*, *stupid*, ...)

Correlation with human similarity ratings

- Direct comparison with semantic similarity ratings (WordSim-353, Finkelstein *et al.* 2002)
  - 353 noun-noun pairs with "relatedness" ratings
  - rated on scale 0–10 by 16 test subjects
  - closely related: *money*/*cash*, *soccer*/*football*, *type*/*kind*, ...
  - unrelated: *king*/*cabbage*, *noon*/*string*, *sugar*/*approach*, ...
  - NB: not all "nouns" are nouns in a traditional sense (*five*, *live*, *eat*, *stupid*, ...)
- Correlation with DSM distances for different corpora
  - DSM parameters: term-term matrix, L4/R4 surface window, 30k feature terms, log G^2 weighting, cosine similarity, SVD to 500 dimensions / 50 skipped (Lapesa and Evert 2014)
  - quantitative measure: Spearman's rank correlation ρ (robust against non-linearities)

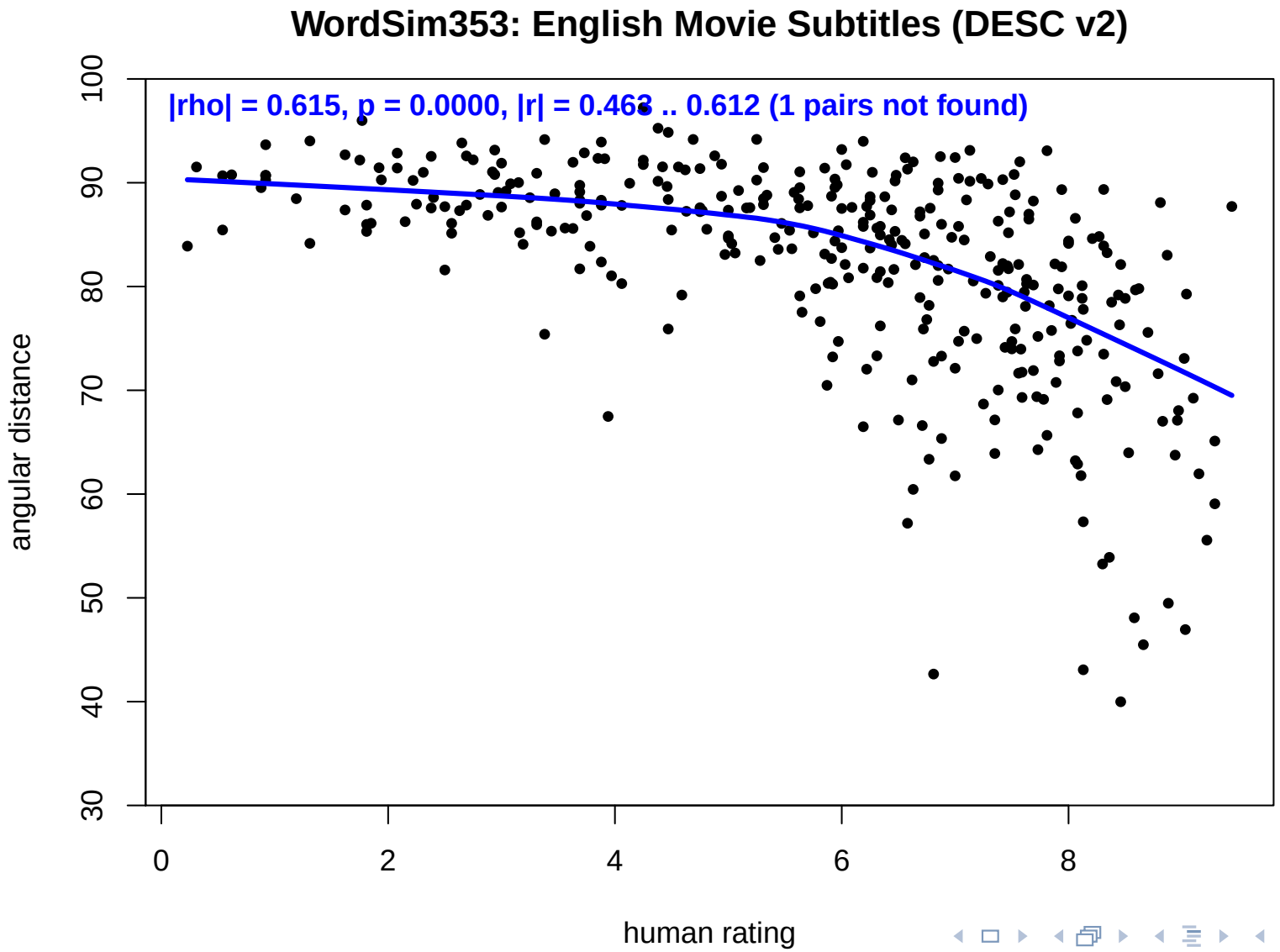
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



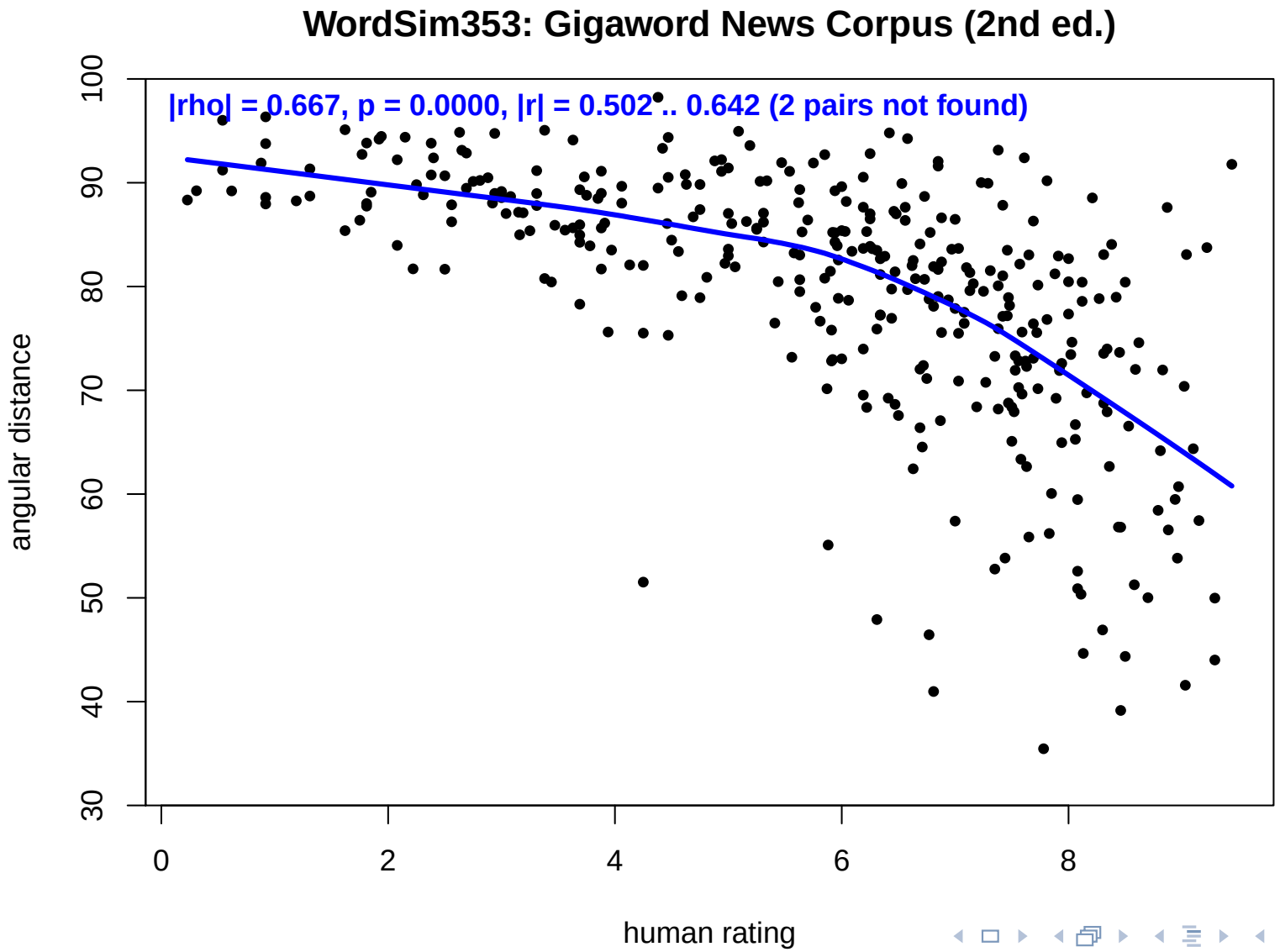
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



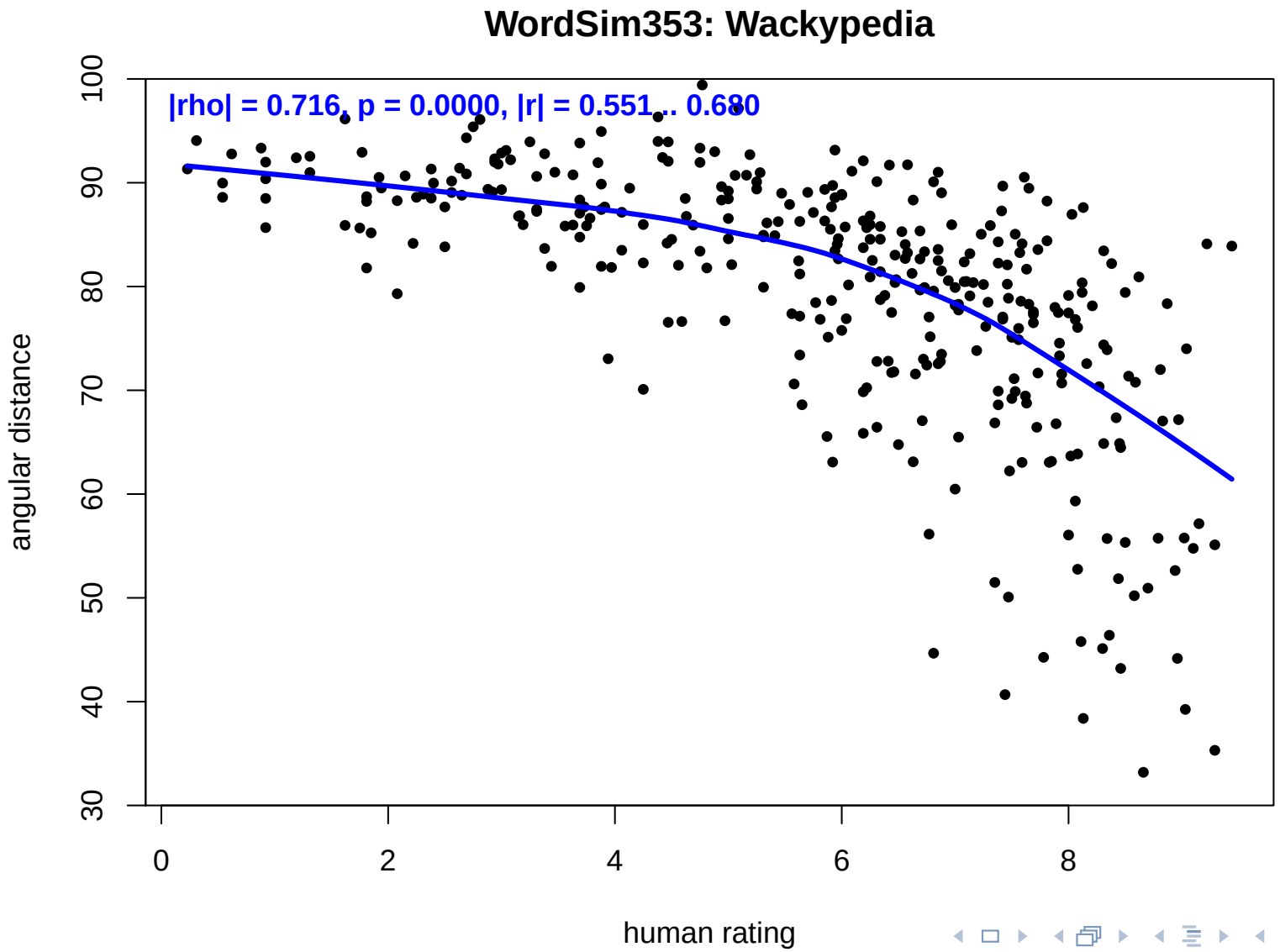
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



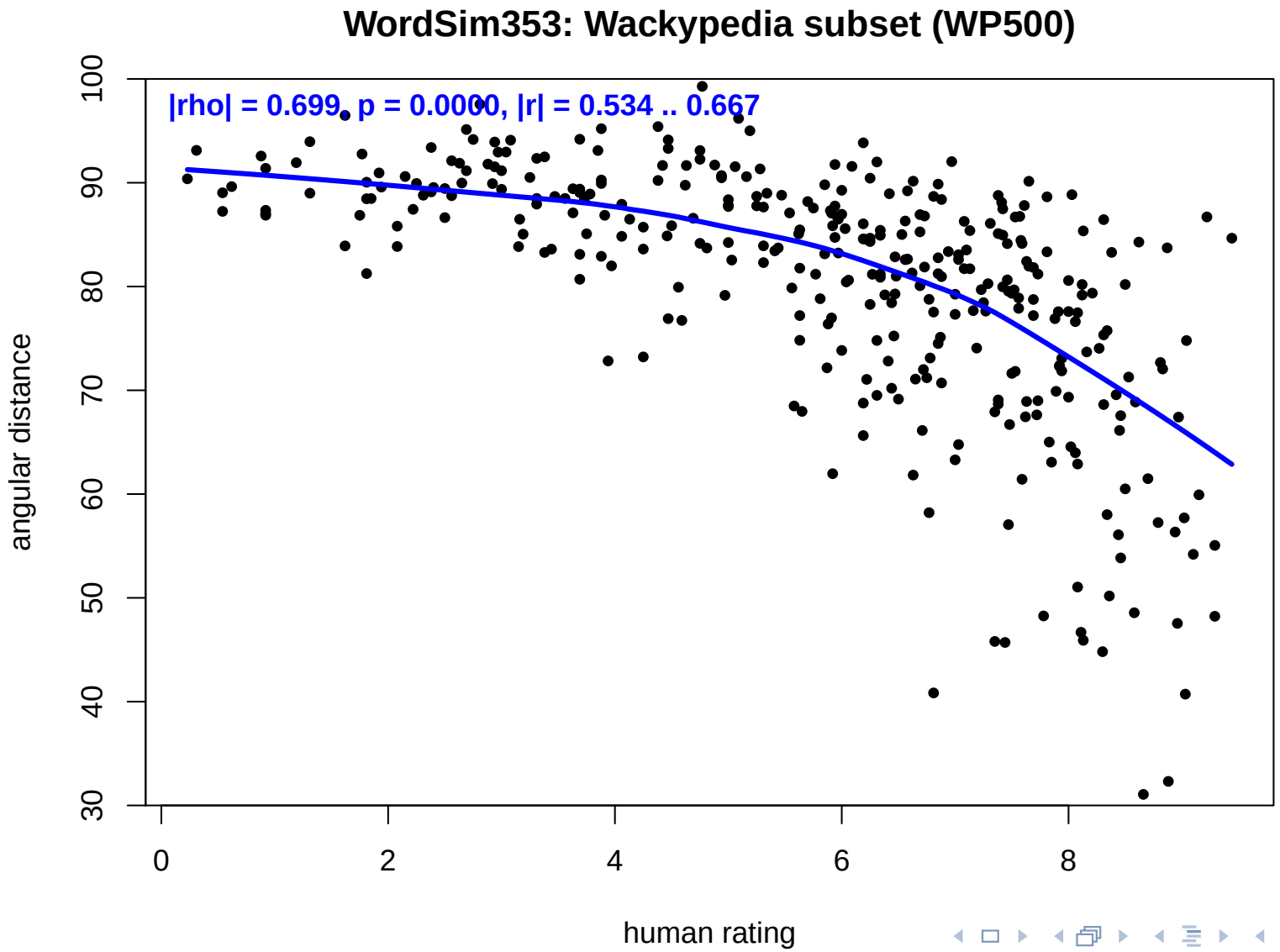
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



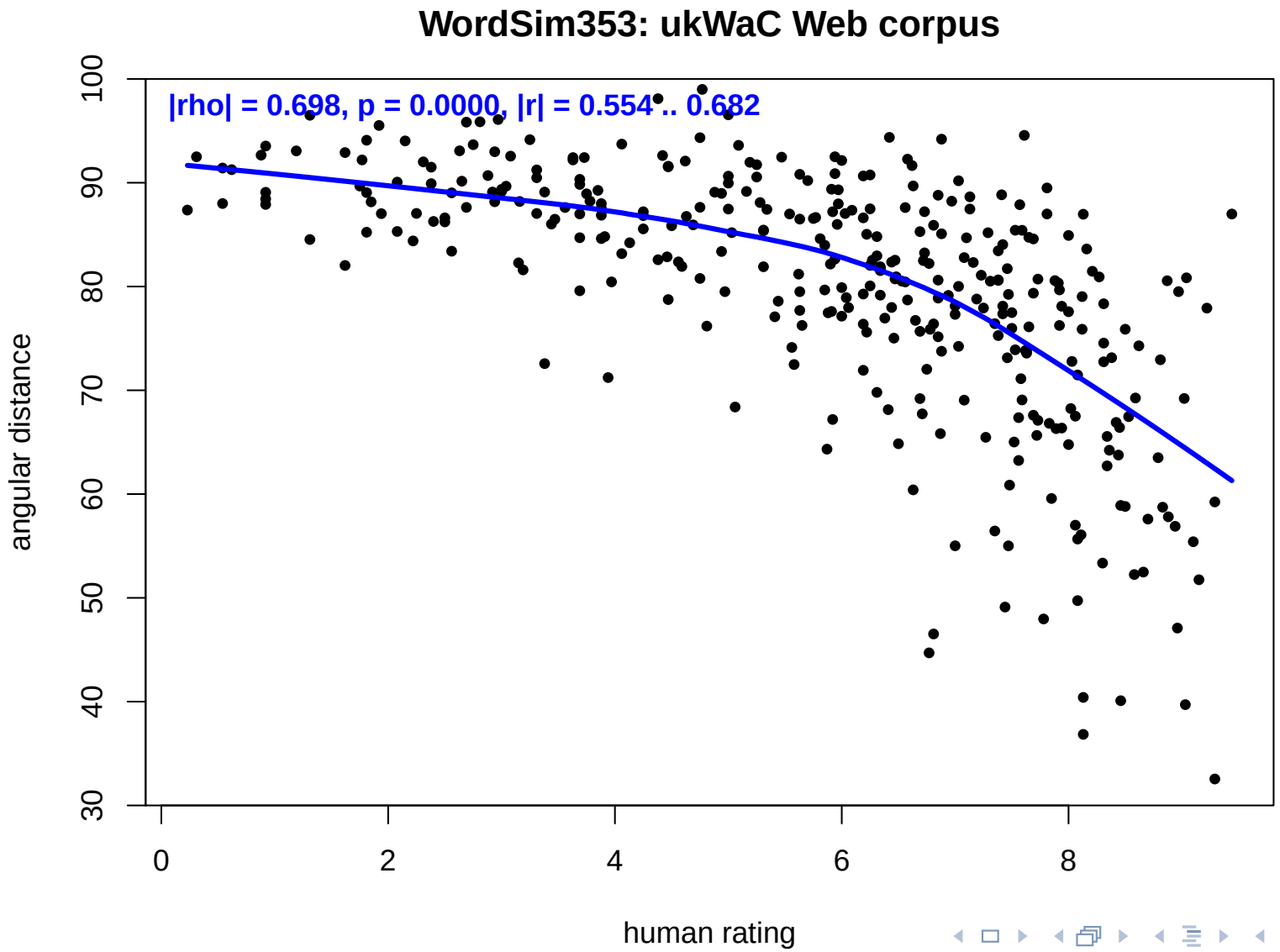
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



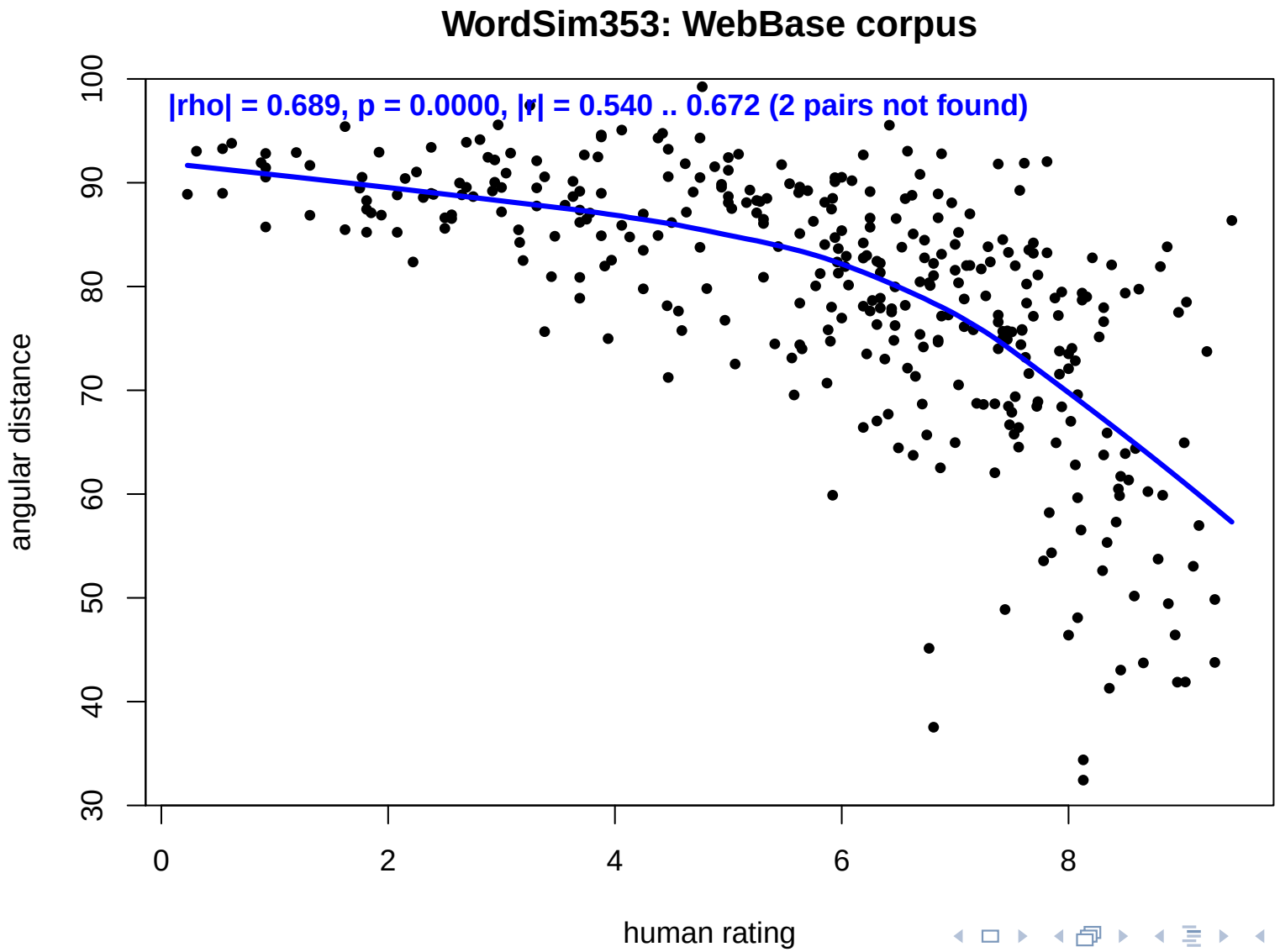
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



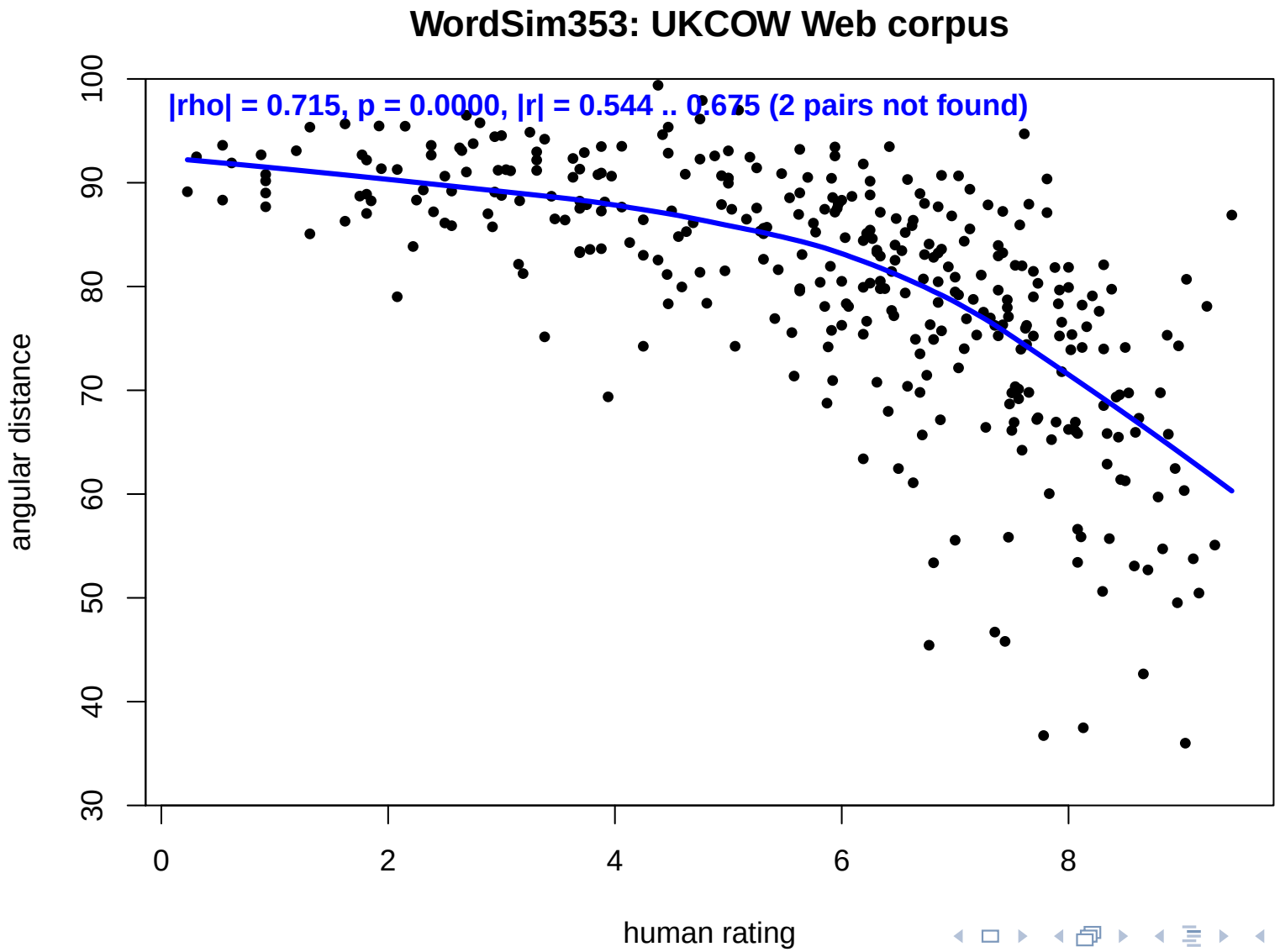
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



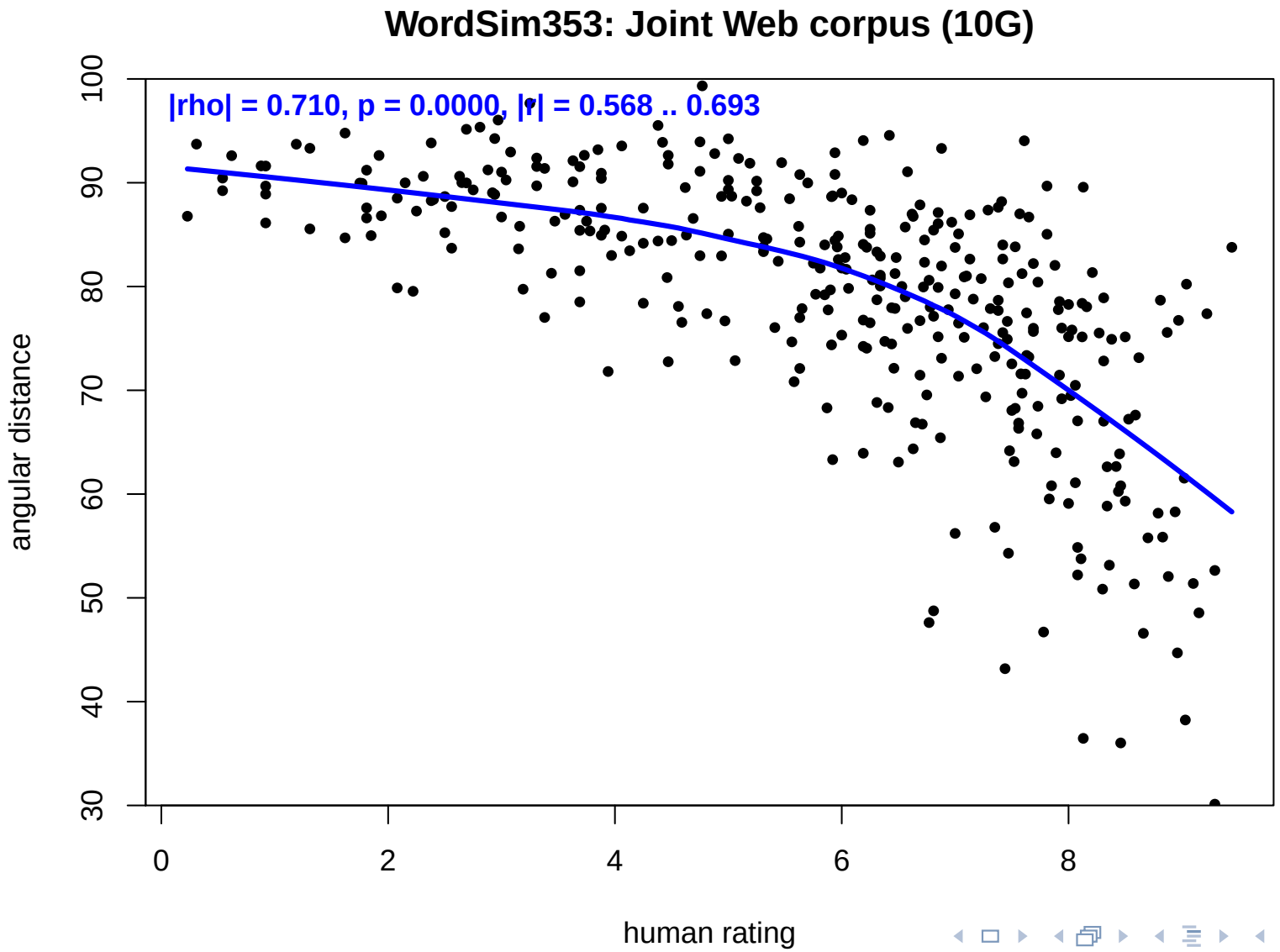
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



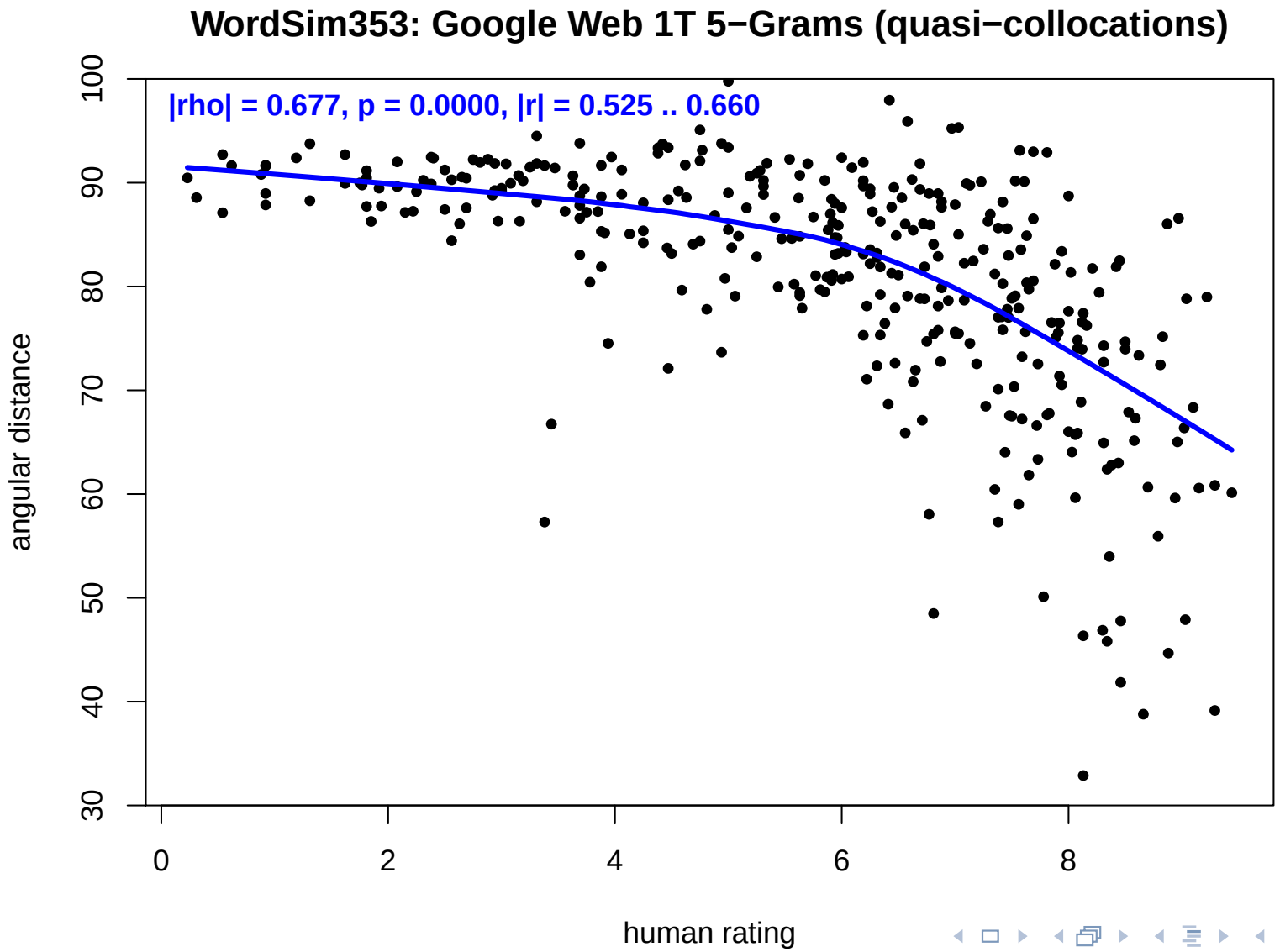
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



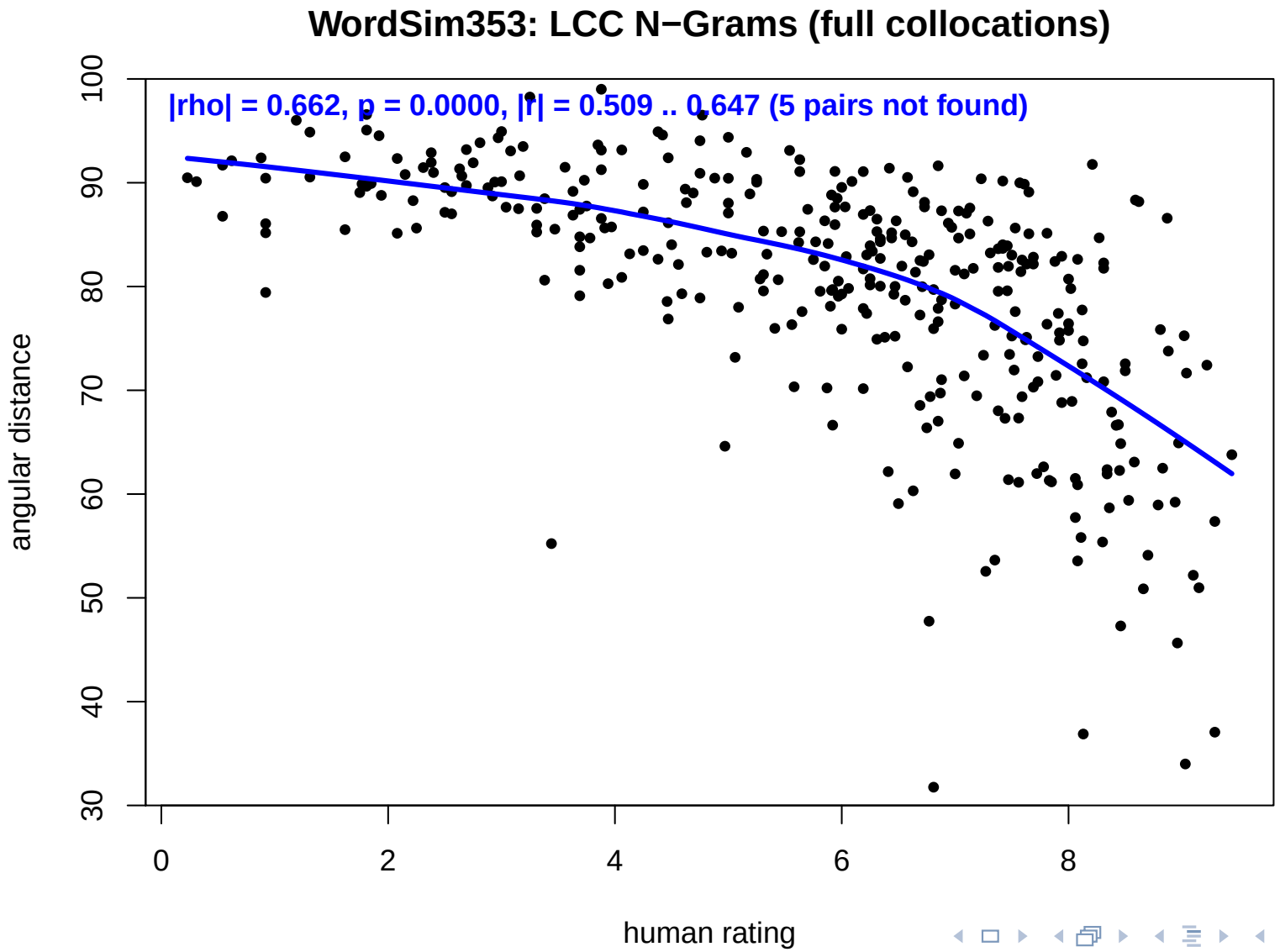
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



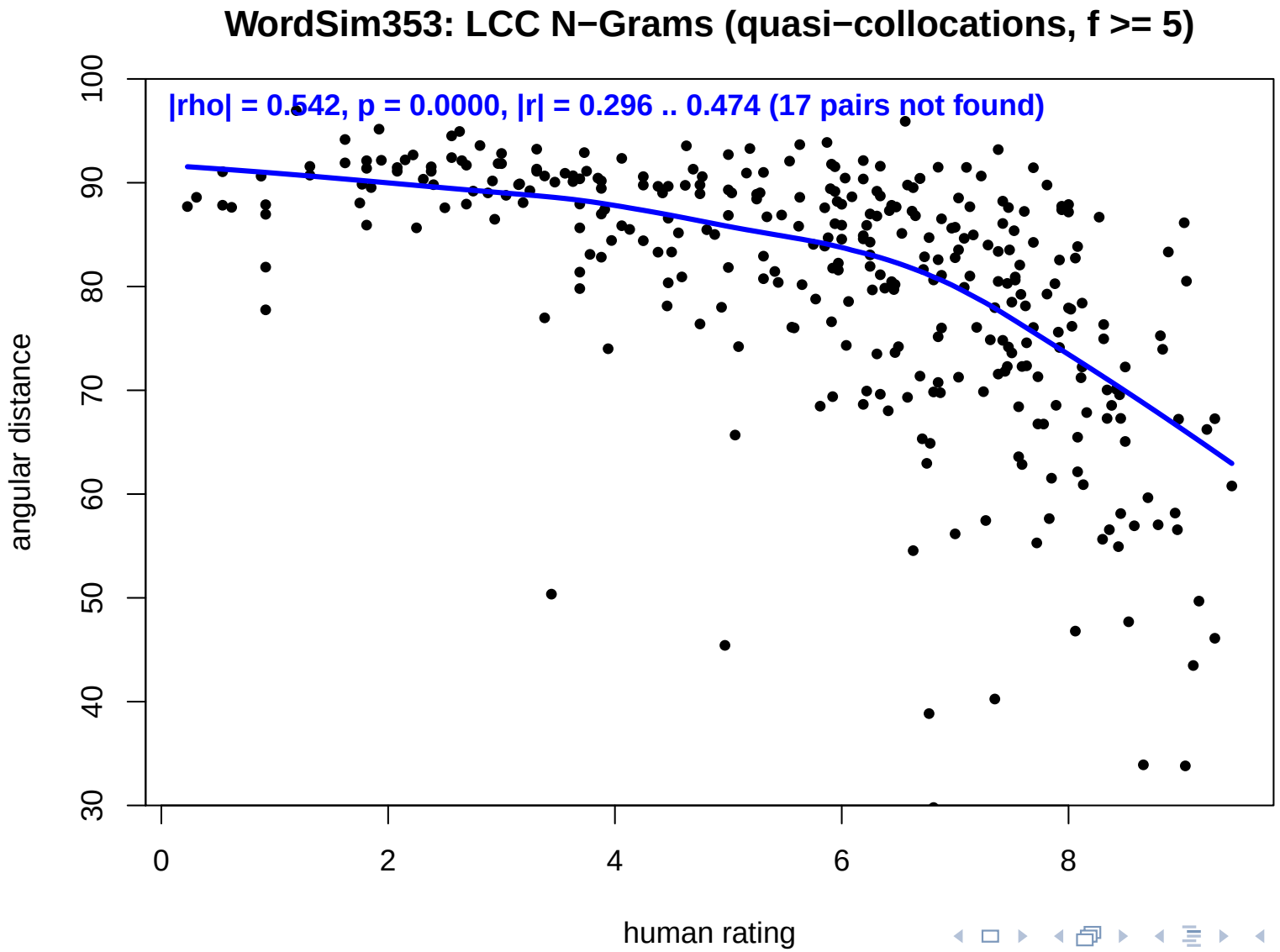
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



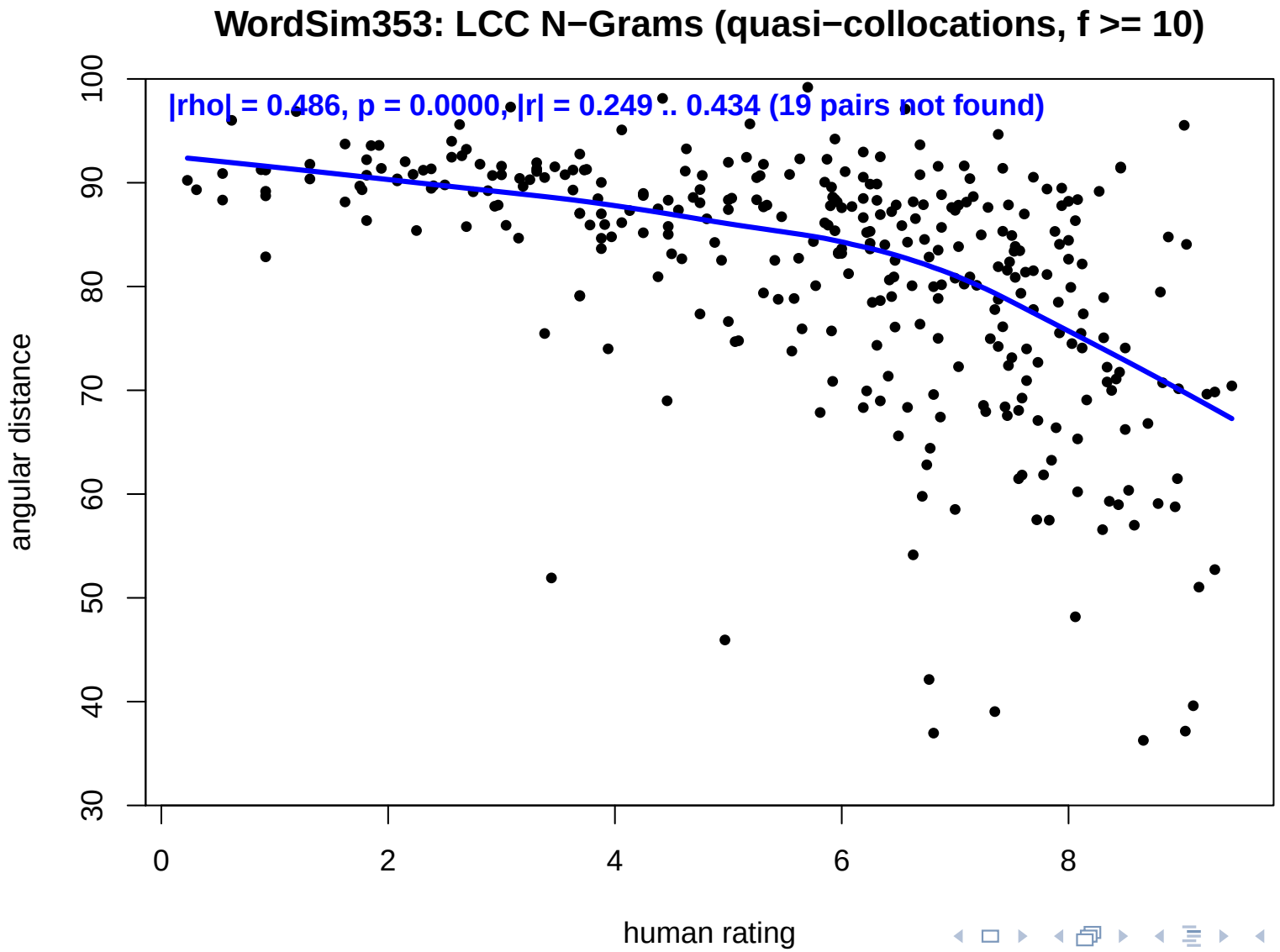
Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



Correlation with human relatedness ratings

(Finkelstein *et al.* 2002)



Why is Web1T5 so terrible despite its size?

Insufficient boilerplate removal & de-duplication:

```
from * to *
```

Why is Web1T5 so terrible despite its size?

Insufficient boilerplate removal & de-duplication:

`from * to *`	
from collectibles to cars	9,443,572
from collectables to cars	8,844,838
from time to time	5,678,941
from left to right	793,957
from start to finish	749,705
from a to z	572,917
from year to year	486,669
from top to bottom	372,935

Why is Web1T5 so terrible despite its size?

Insufficient boilerplate removal & de-duplication:

`from * to *`	
from collectibles to cars	9,443,572
from collectables to cars	8,844,838
from time to time	5,678,941
from left to right	793,957
from start to finish	749,705
from a to z	572,917
from year to year	486,669
from top to bottom	372,935

"Traditional" Web corpora are much better:

Google	≈ 121,000,000 hits
Google.de	≈ 119,600,000 hits
Web 1T 5-Grams	18,288,410 hits
ukWaC	**3 hits**
BNC	0 hits

Why is Web1T5 so terrible?

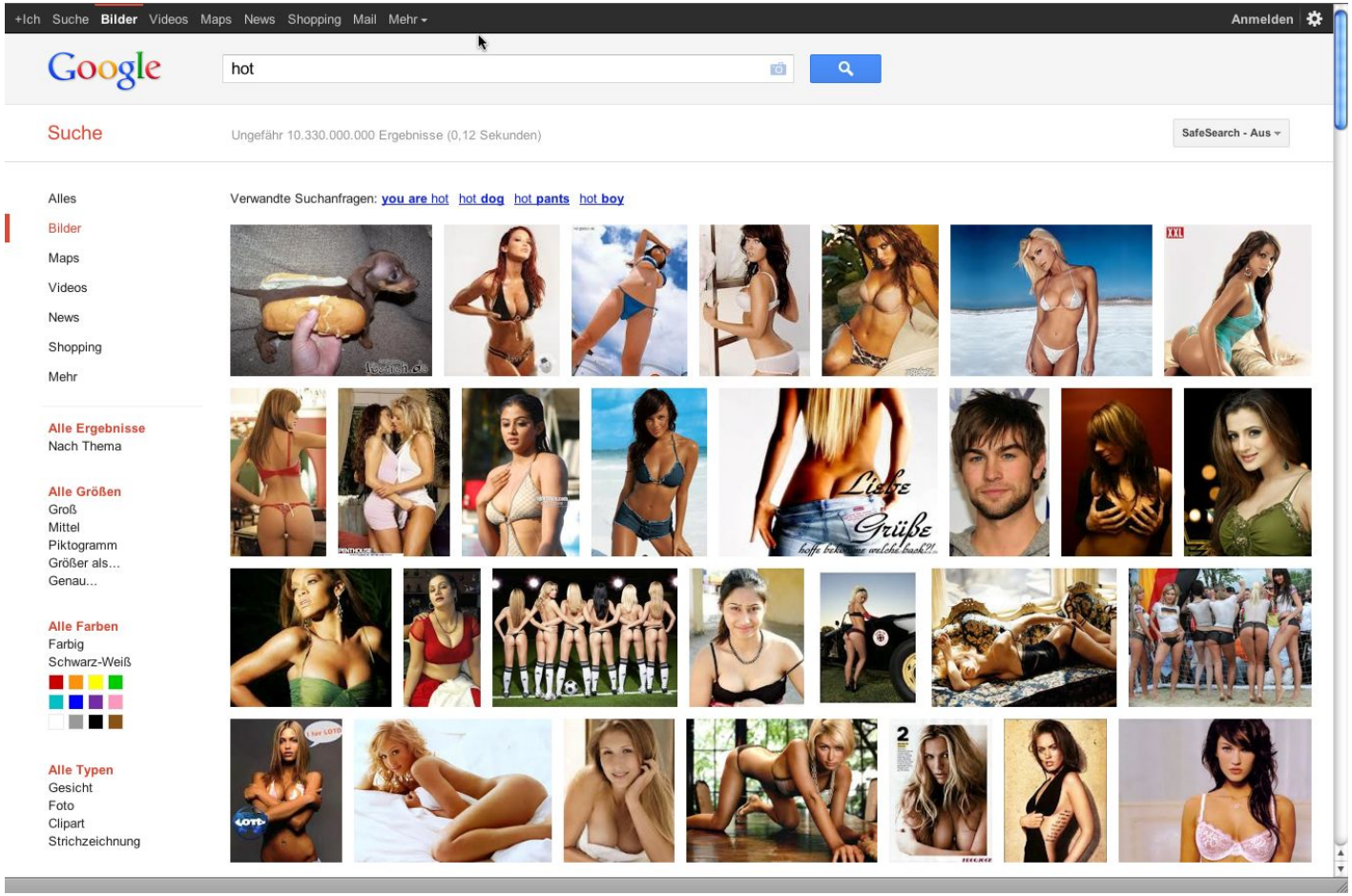
Which words are semantically similar to **hot** (in DSM)?

- I hope there are no minors in the room!

Why is Web1T5 so terrible?

Which words are semantically similar to **hot** (in DSM)?

- I hope there are no minors in the room!



Why is Web1T5 so terrible?

Which words are semantically similar to **hot** (in DSM)?

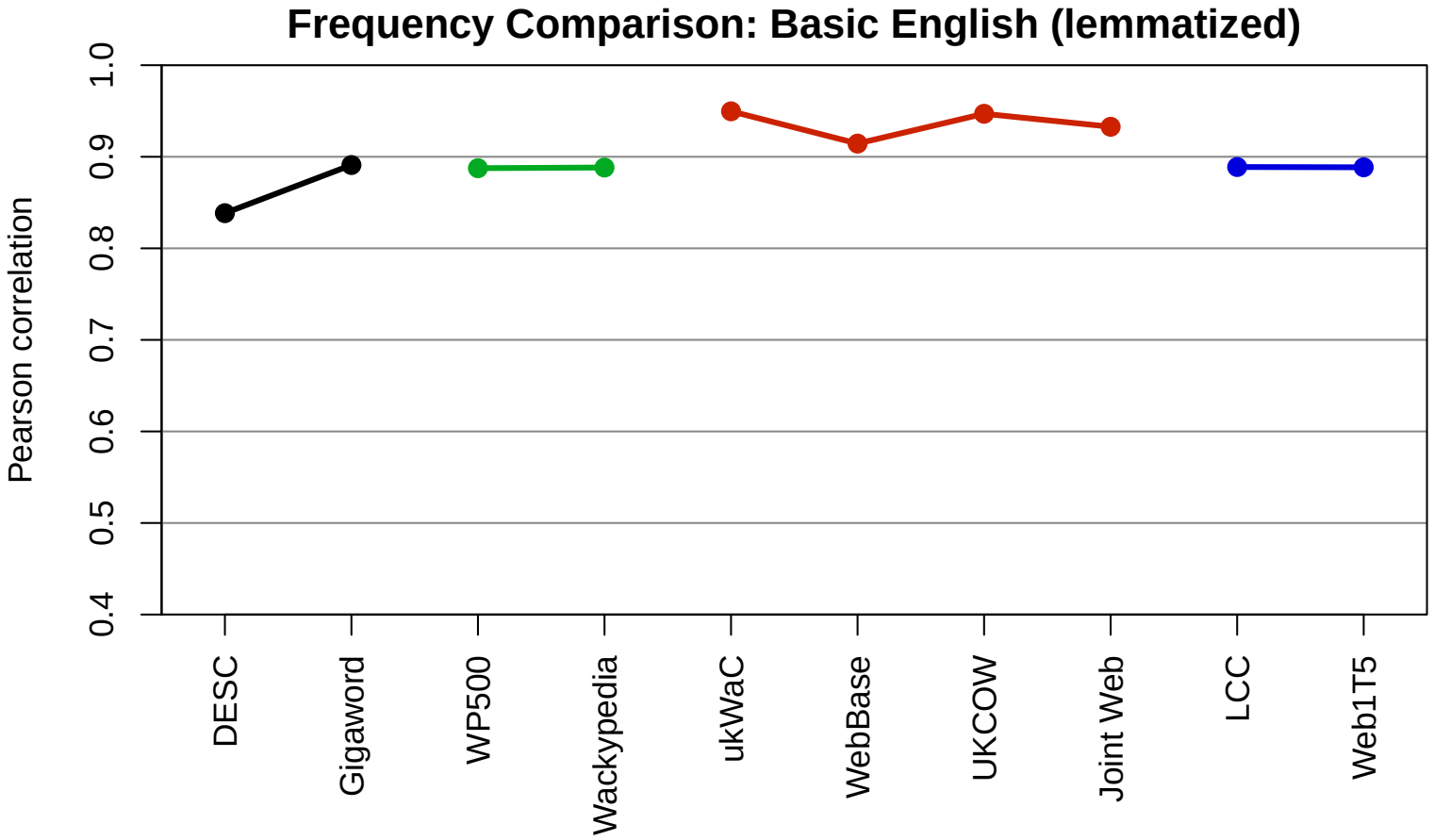
- I hope there are no minors in the room!

big (29.5), butt (31.1), ass (31.1), wet (31.2), naughty (31.6), pussy (31.6), sexy (31.6), chicks (32.0), cock (32.2), ebony (32.3), fat (32.4), girls (32.4), asian (32.7), cum (33.1), babes (33.2), dirty (33.2), bikini (33.3), granny (33.4), teen (33.8), pics (33.8), gras (34.1), fucking (34.1), galleries (34.2), fetish (34.3), babe (34.3), blonde (34.5), pussies (34.5), whores (34.6), fuck (34.6), horny (34.7)

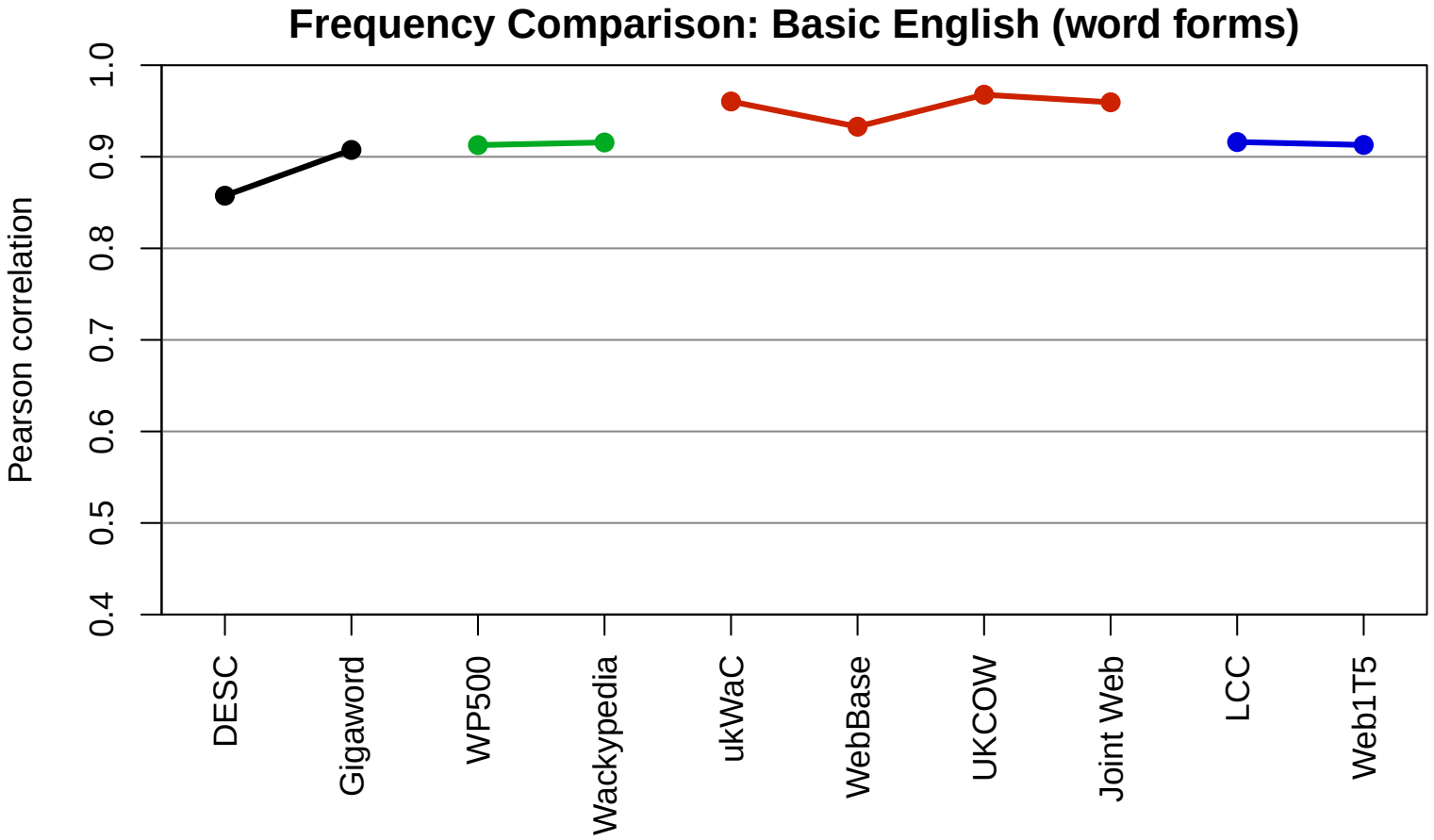
Please don't ask about cats and dogs . . .

Overview & Discussion

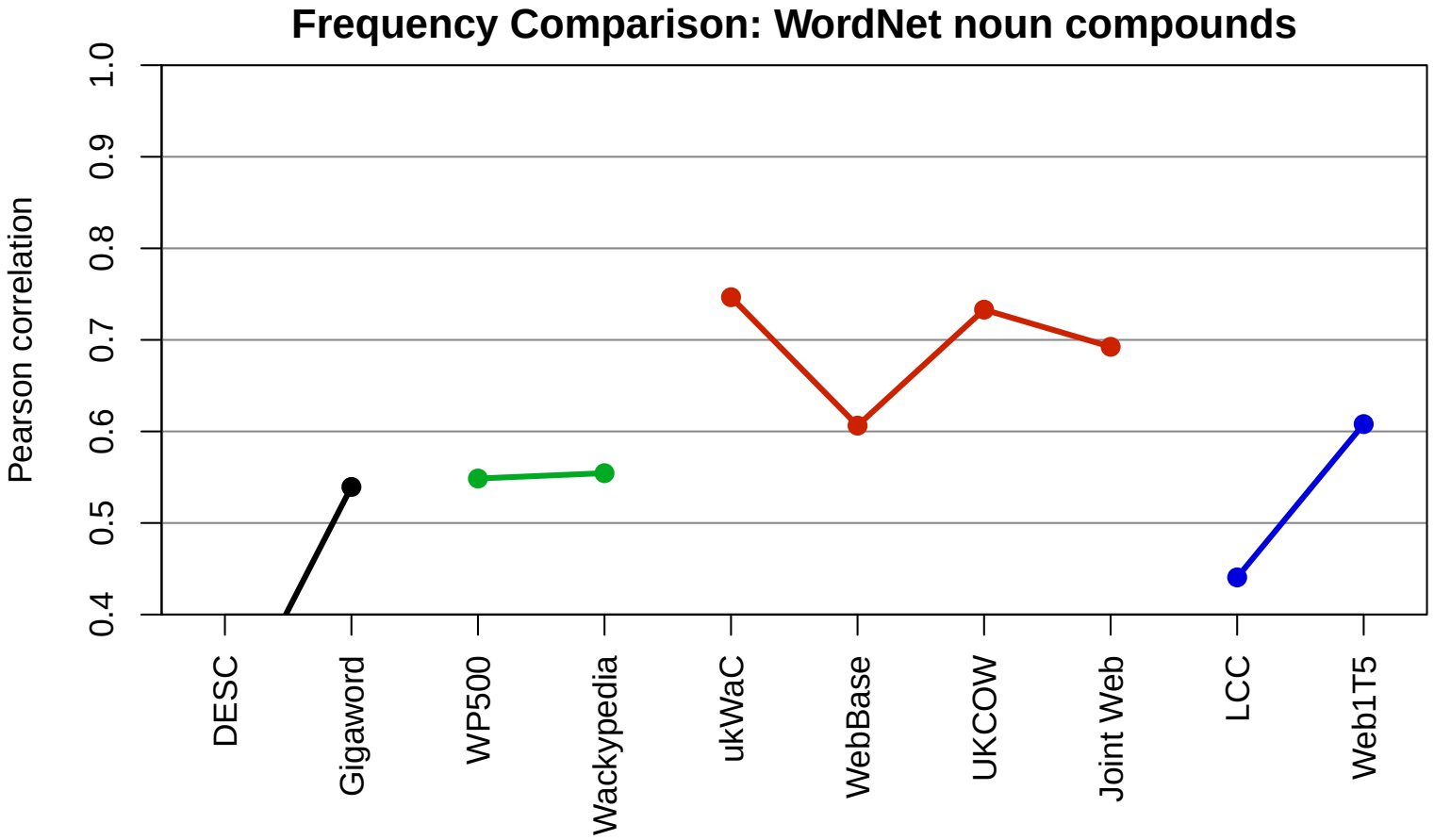
Result overview: Frequency comparison



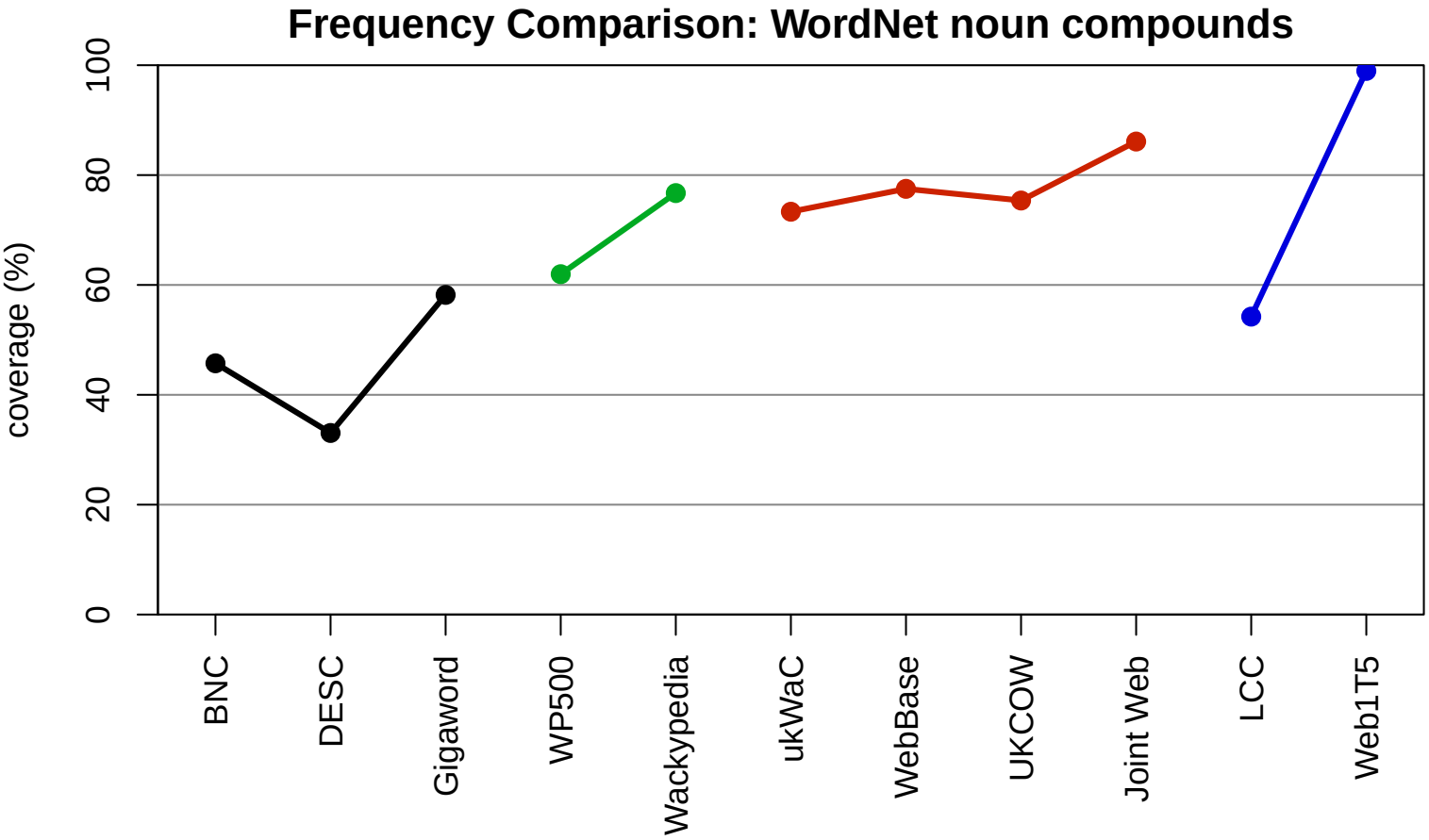
Result overview: Frequency comparison



Result overview: Frequency comparison

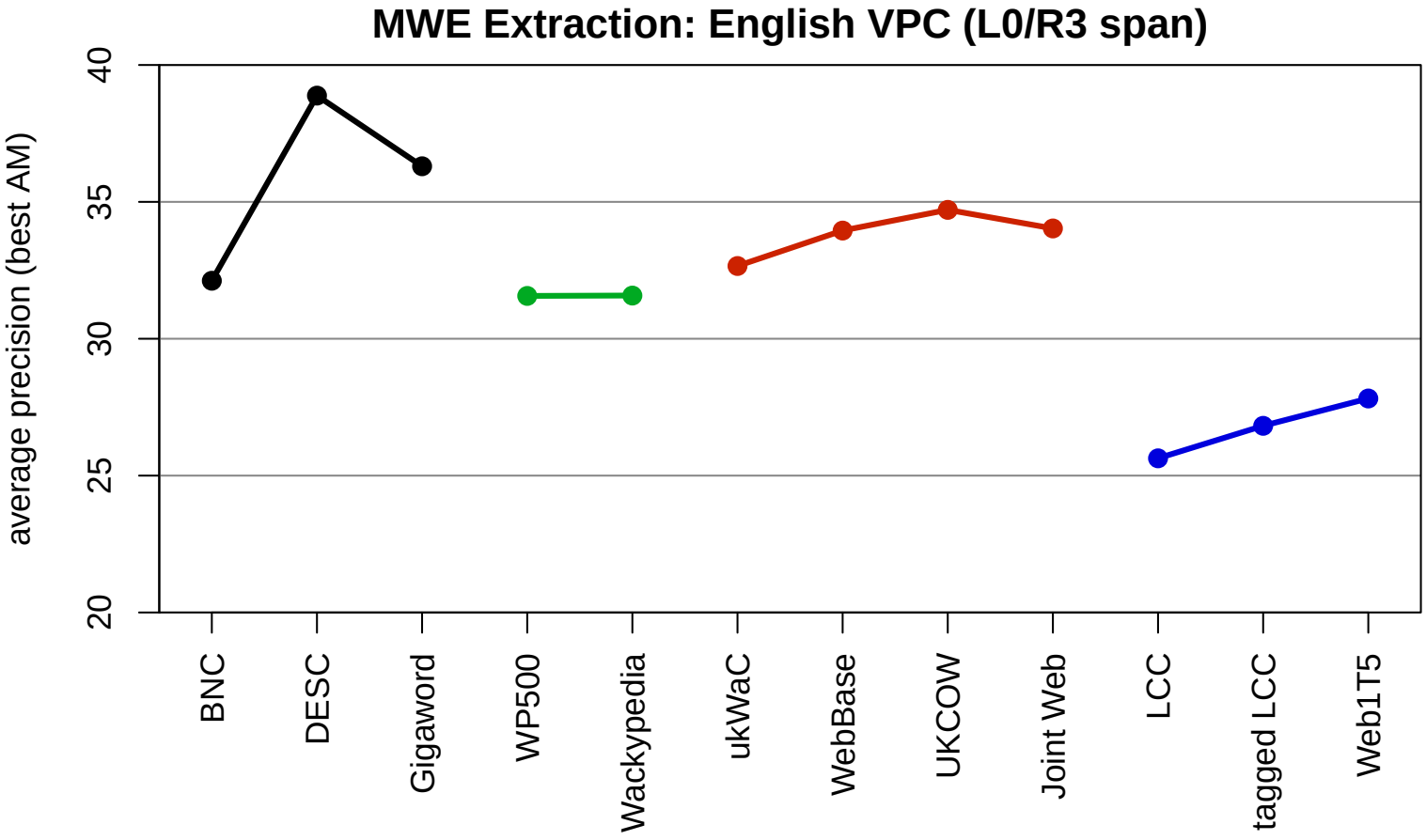


Result overview: Frequency comparison

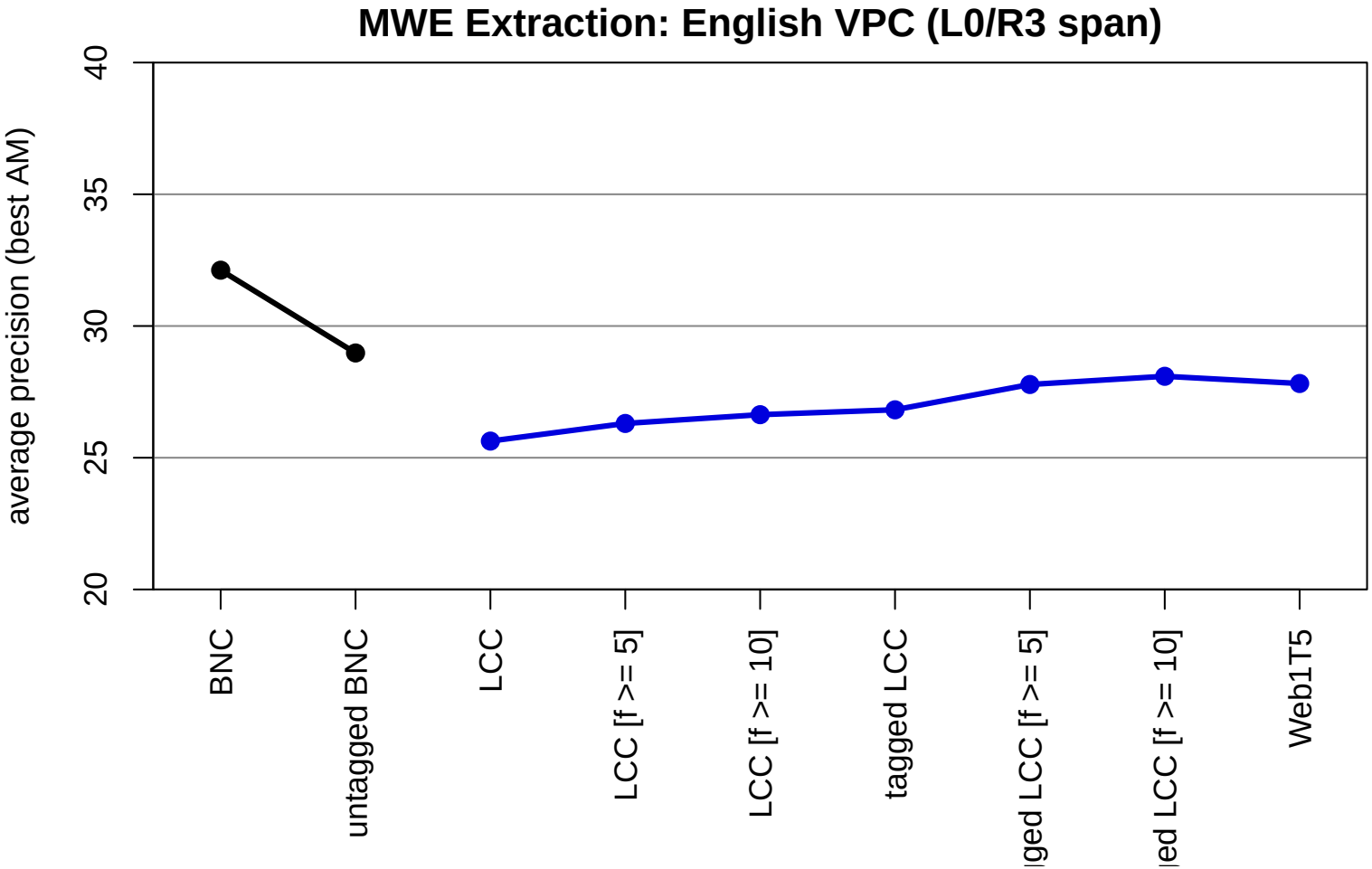


coverage analysis

Result overview: MWE identification

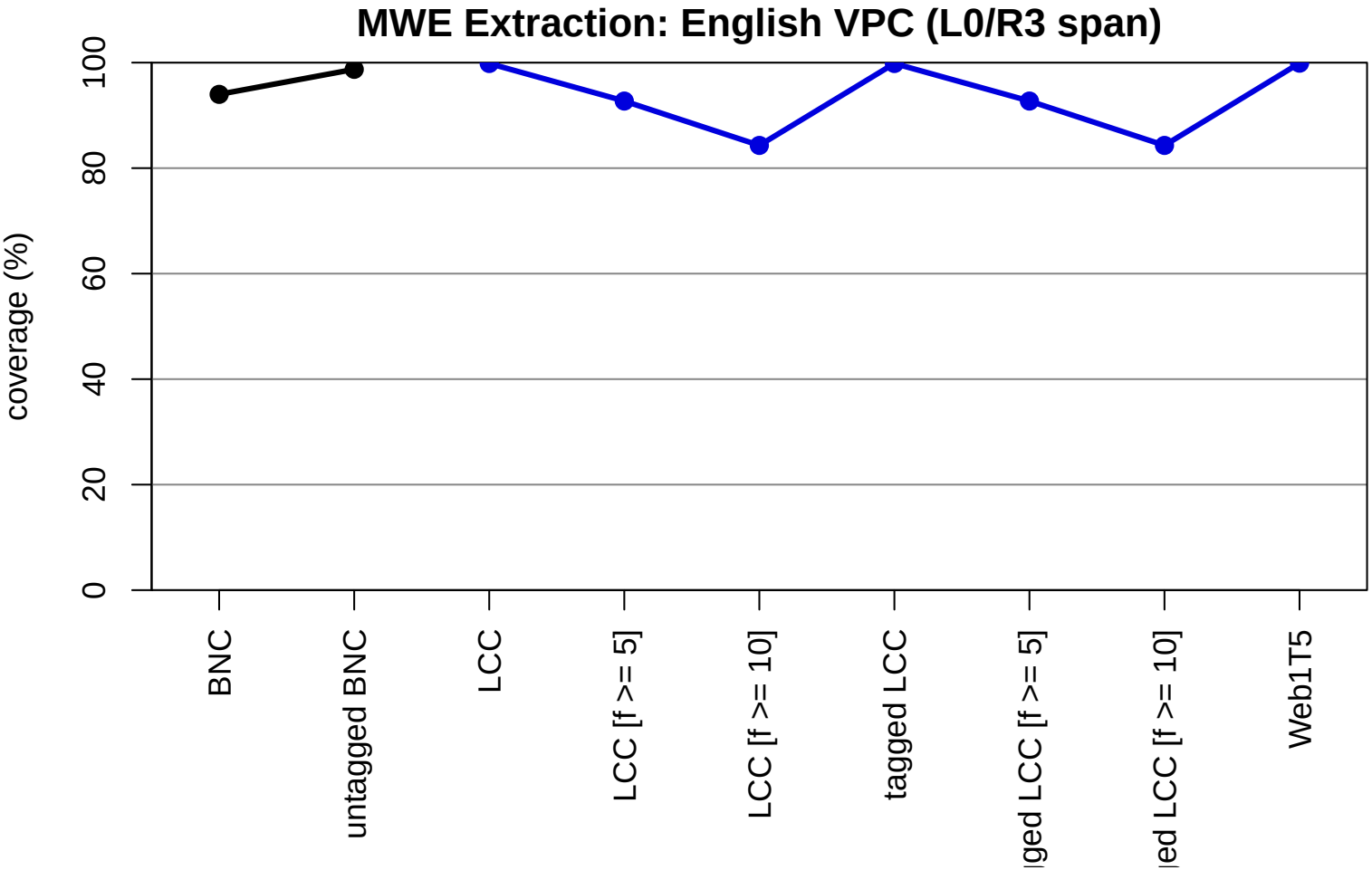


Result overview: MWE identification



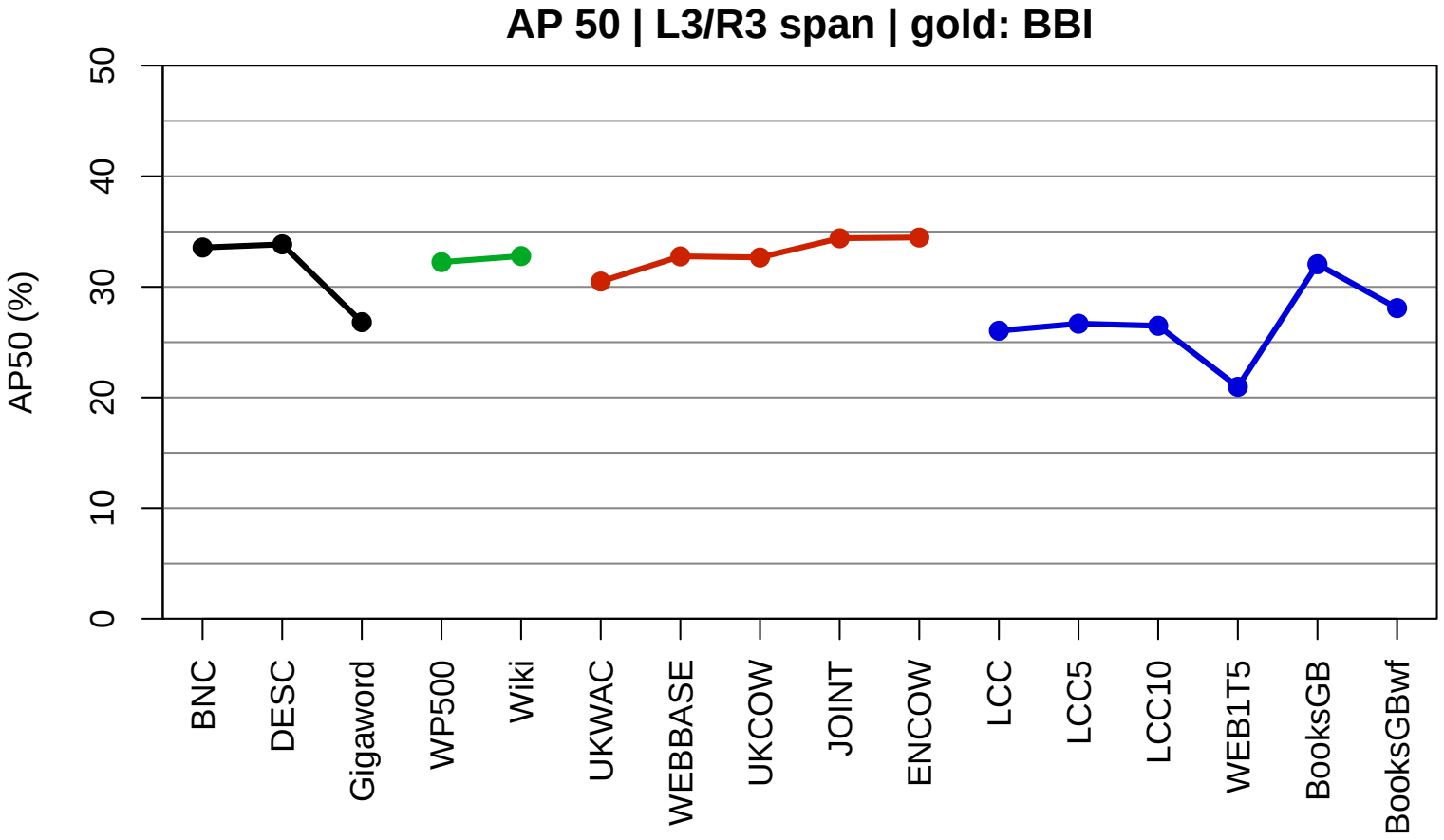
POS-tagging and frequency threshold

Result overview: MWE identification

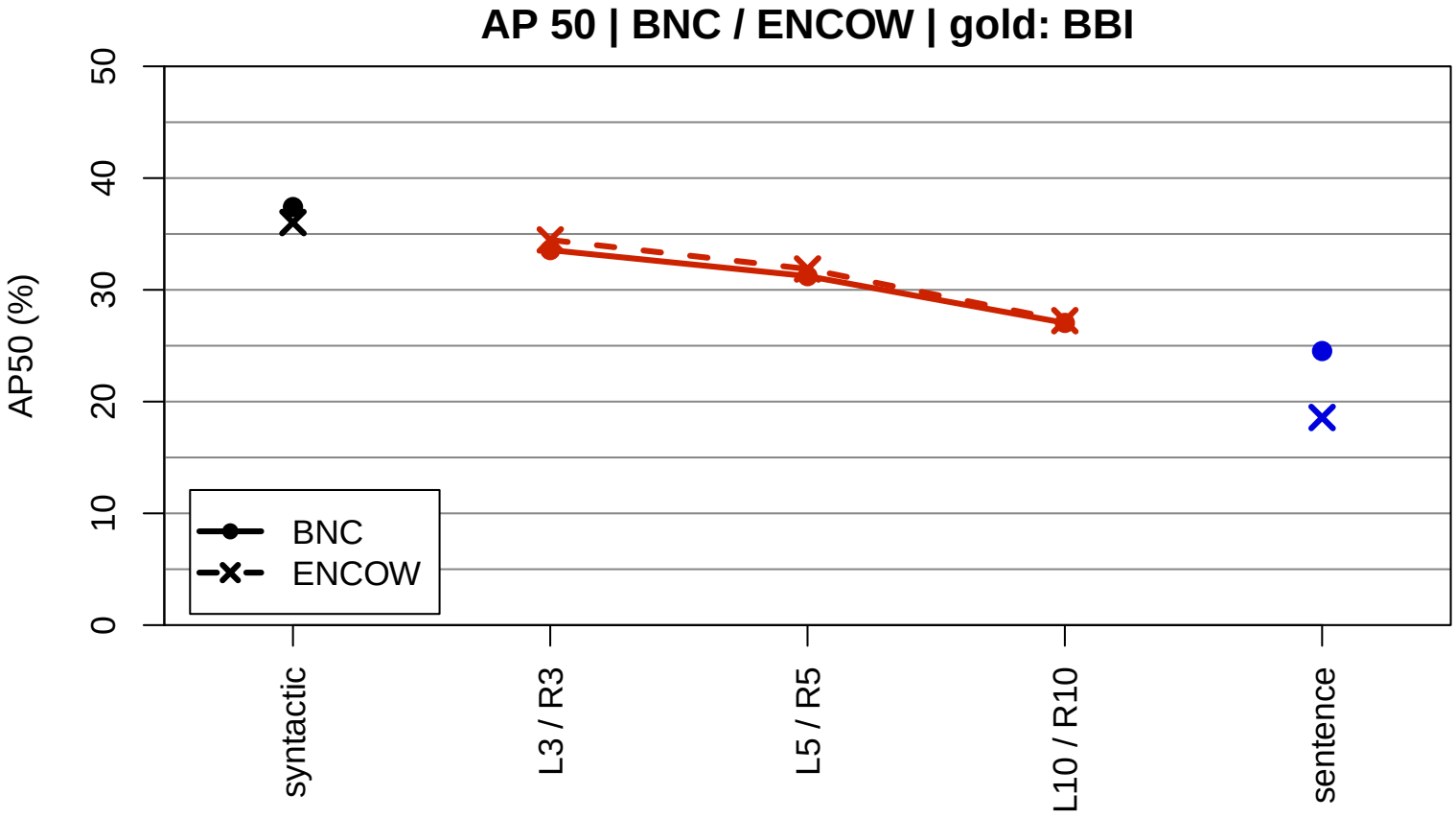


coverage analysis

Result overview: BBI collocations

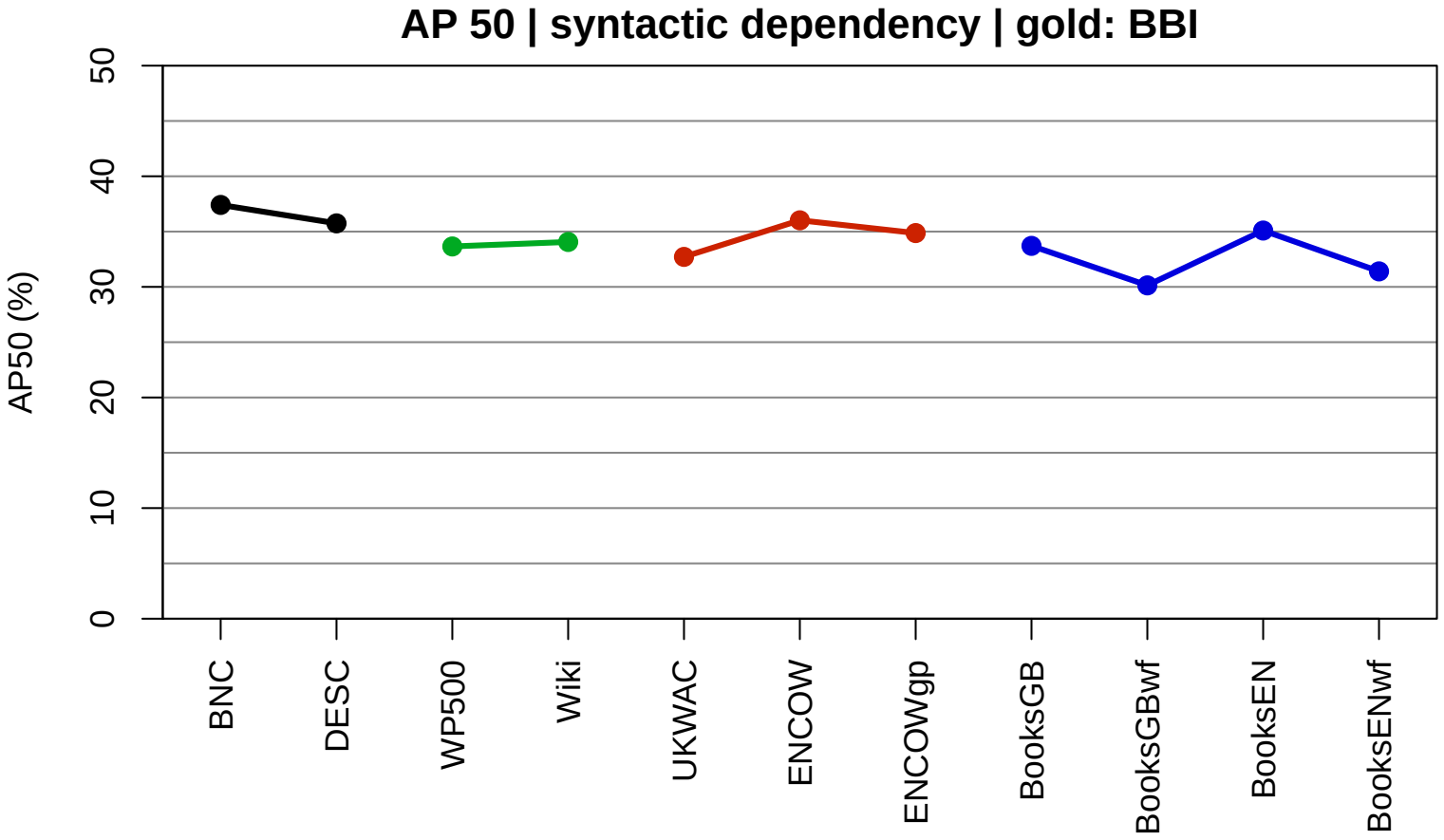


Result overview: BBI collocations

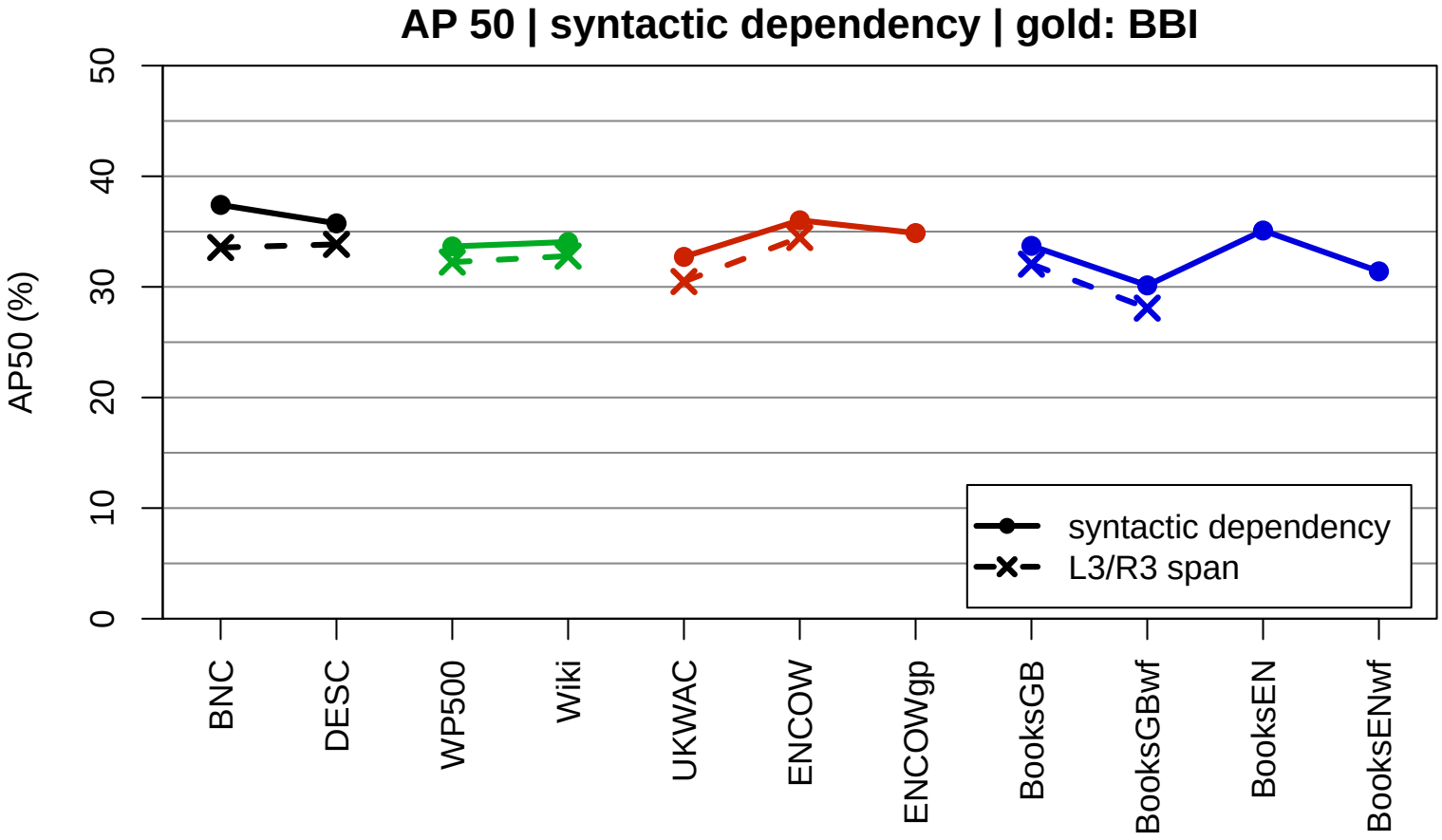


collocational span

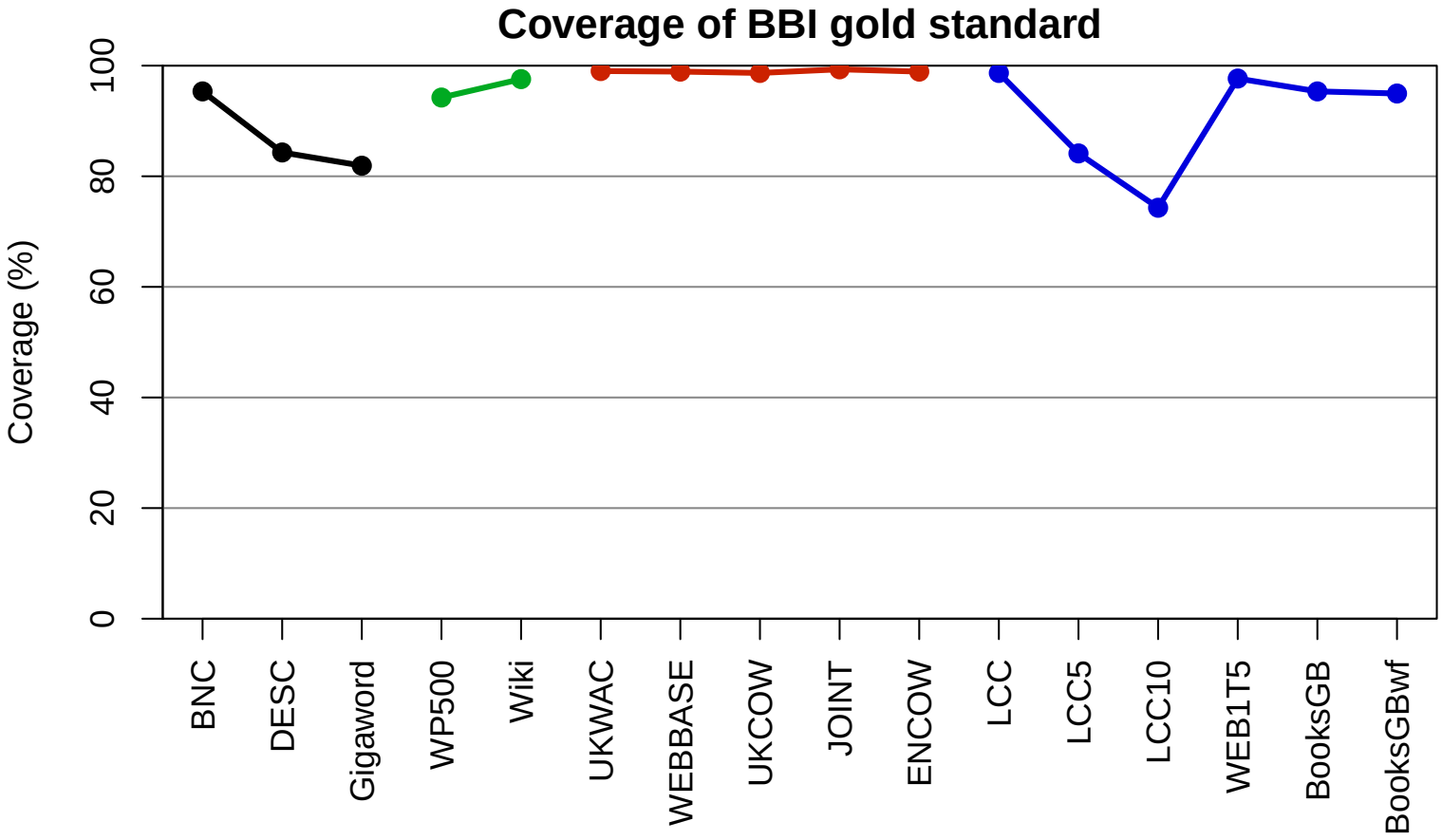
Result overview: BBI collocations



Result overview: BBI collocations

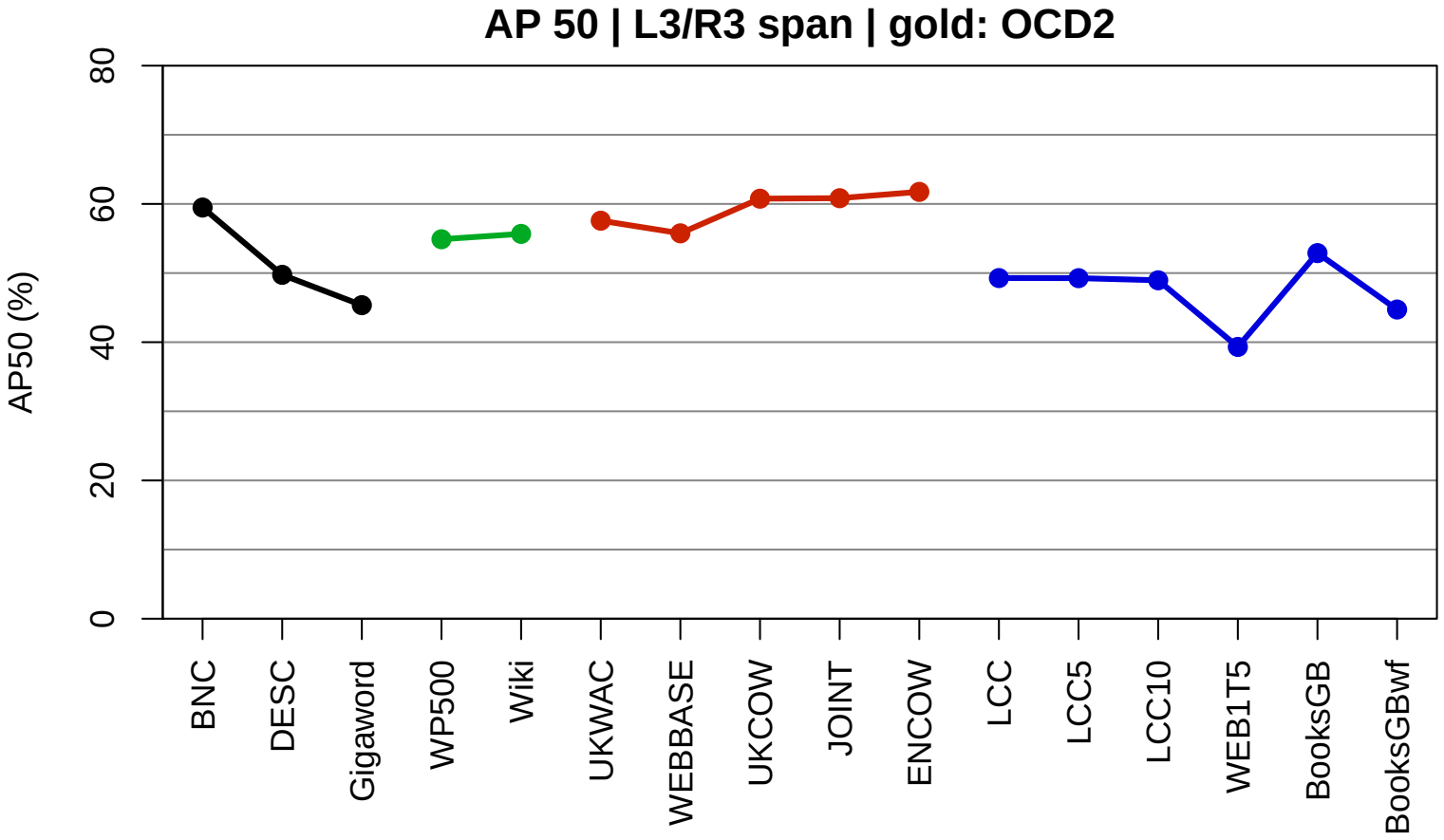


Result overview: BBI collocations

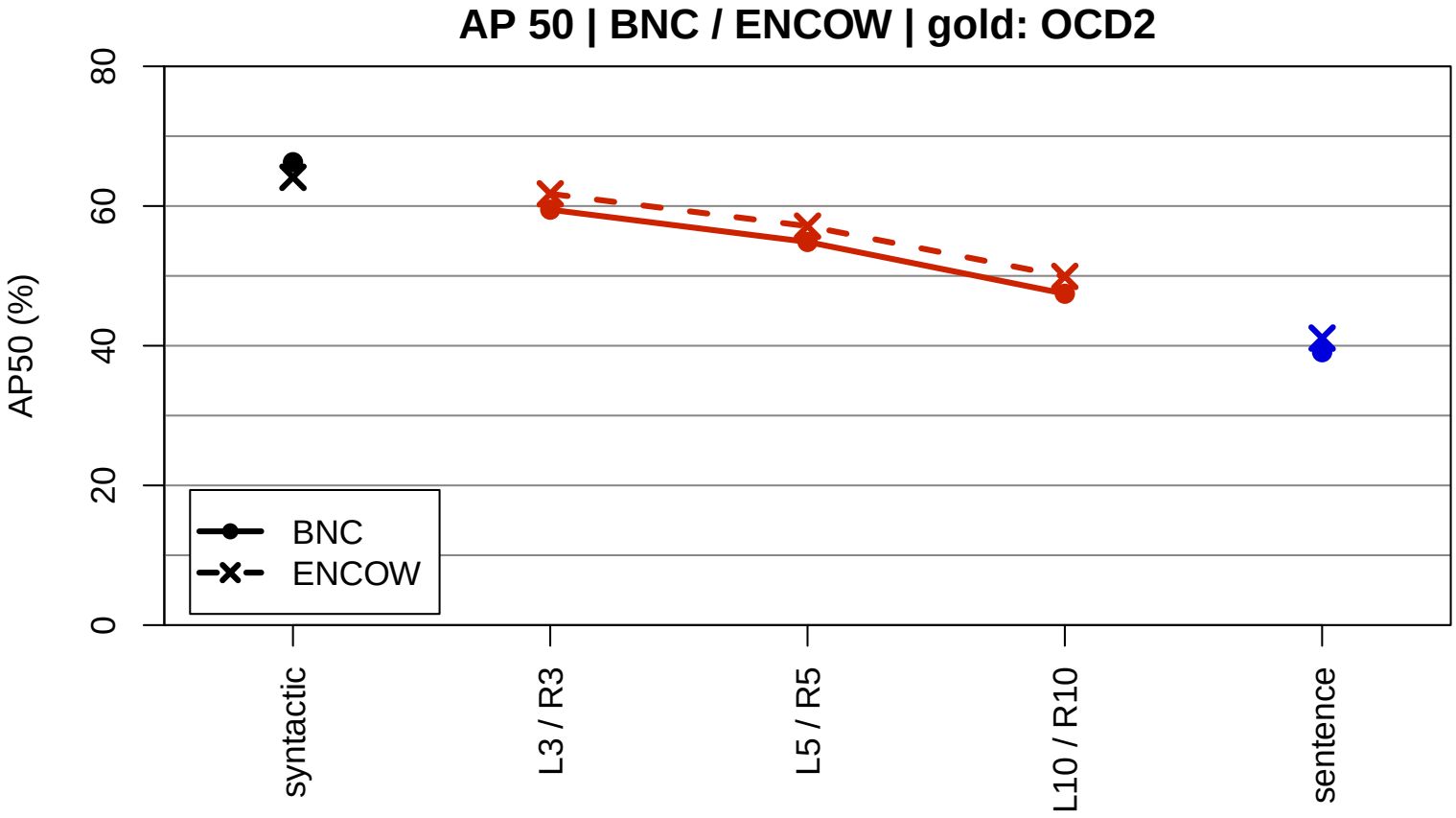


coverage analysis

Result overview: OCD2 collocations

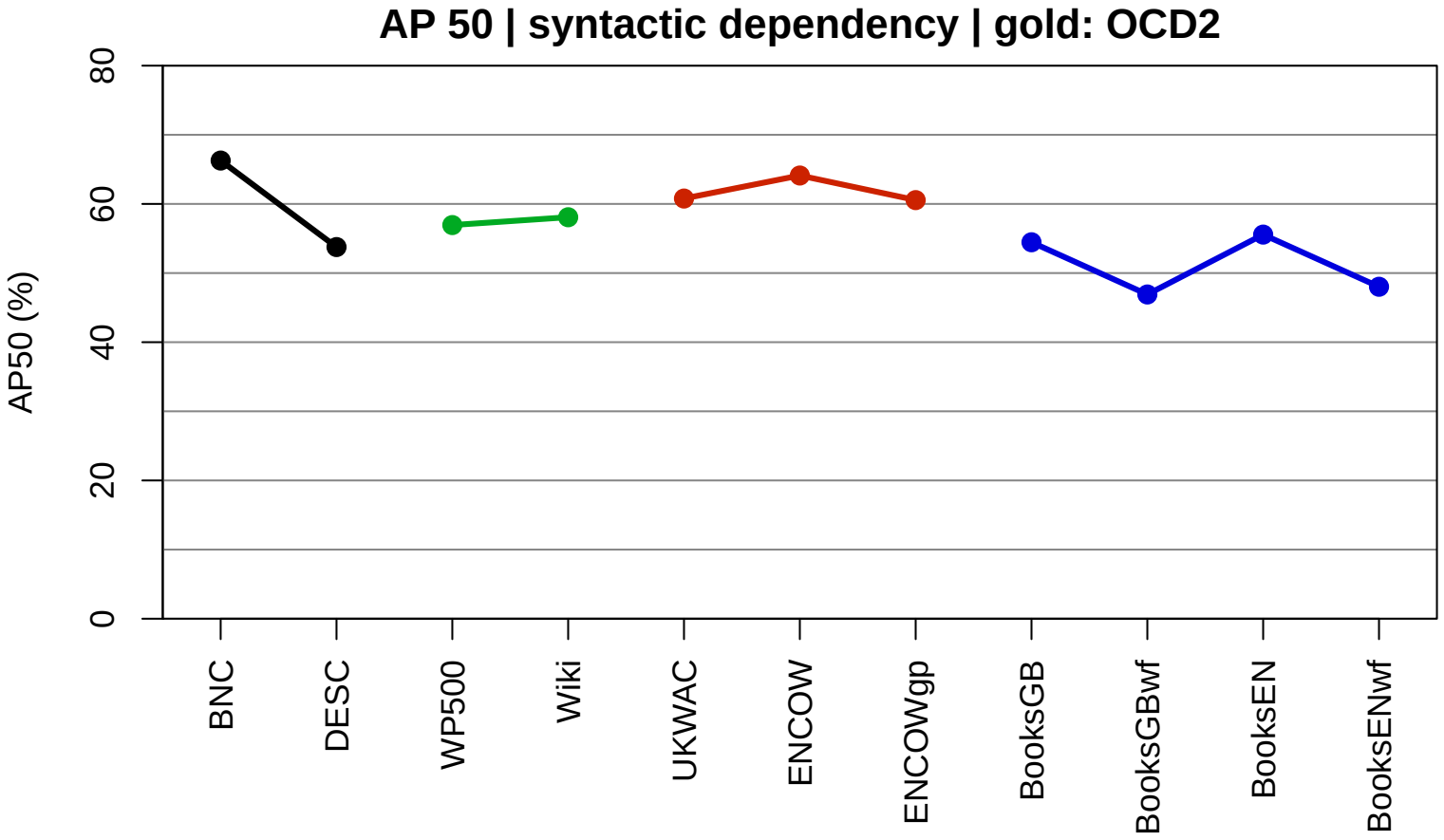


Result overview: OCD2 collocations

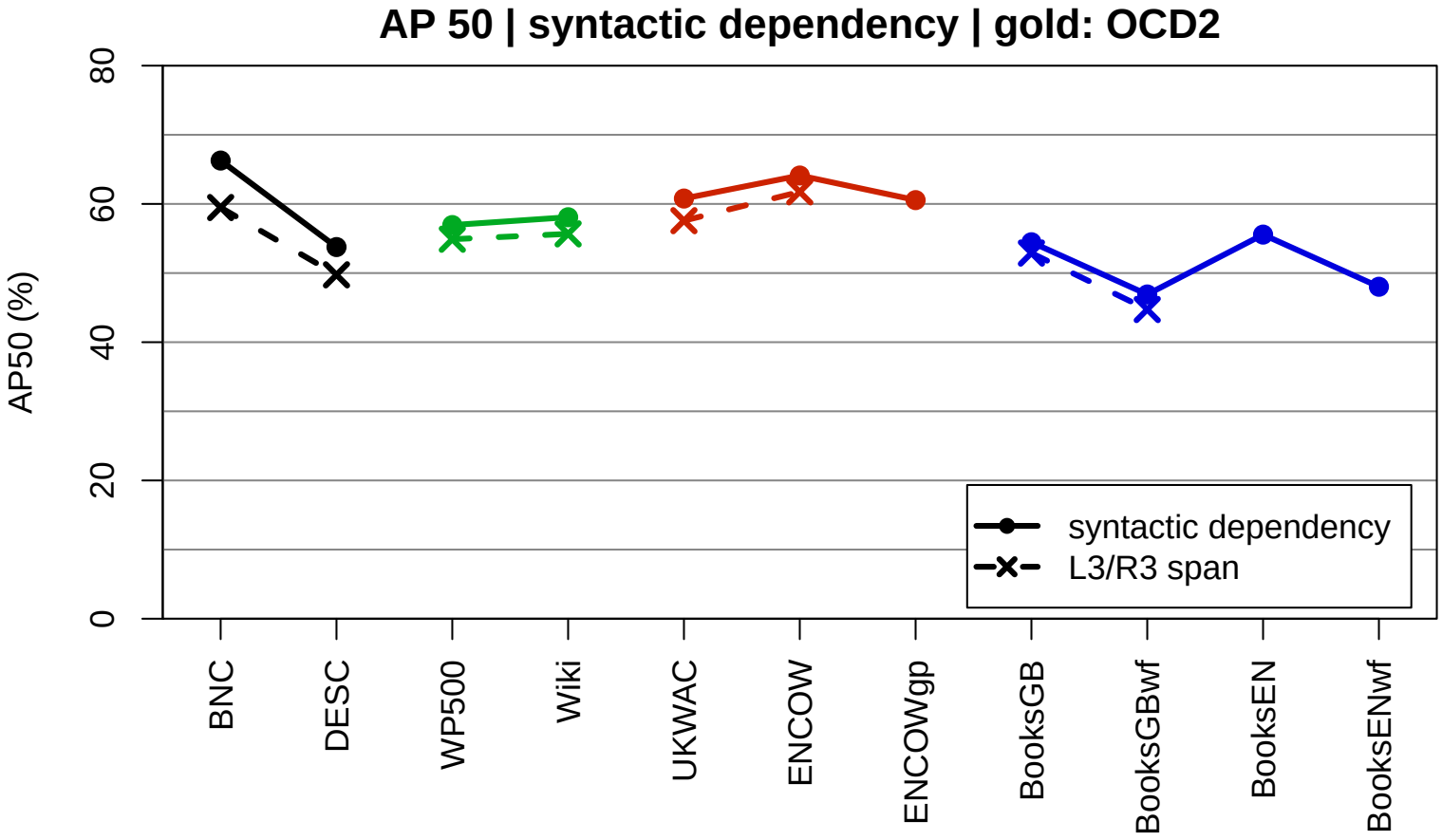


collocational span

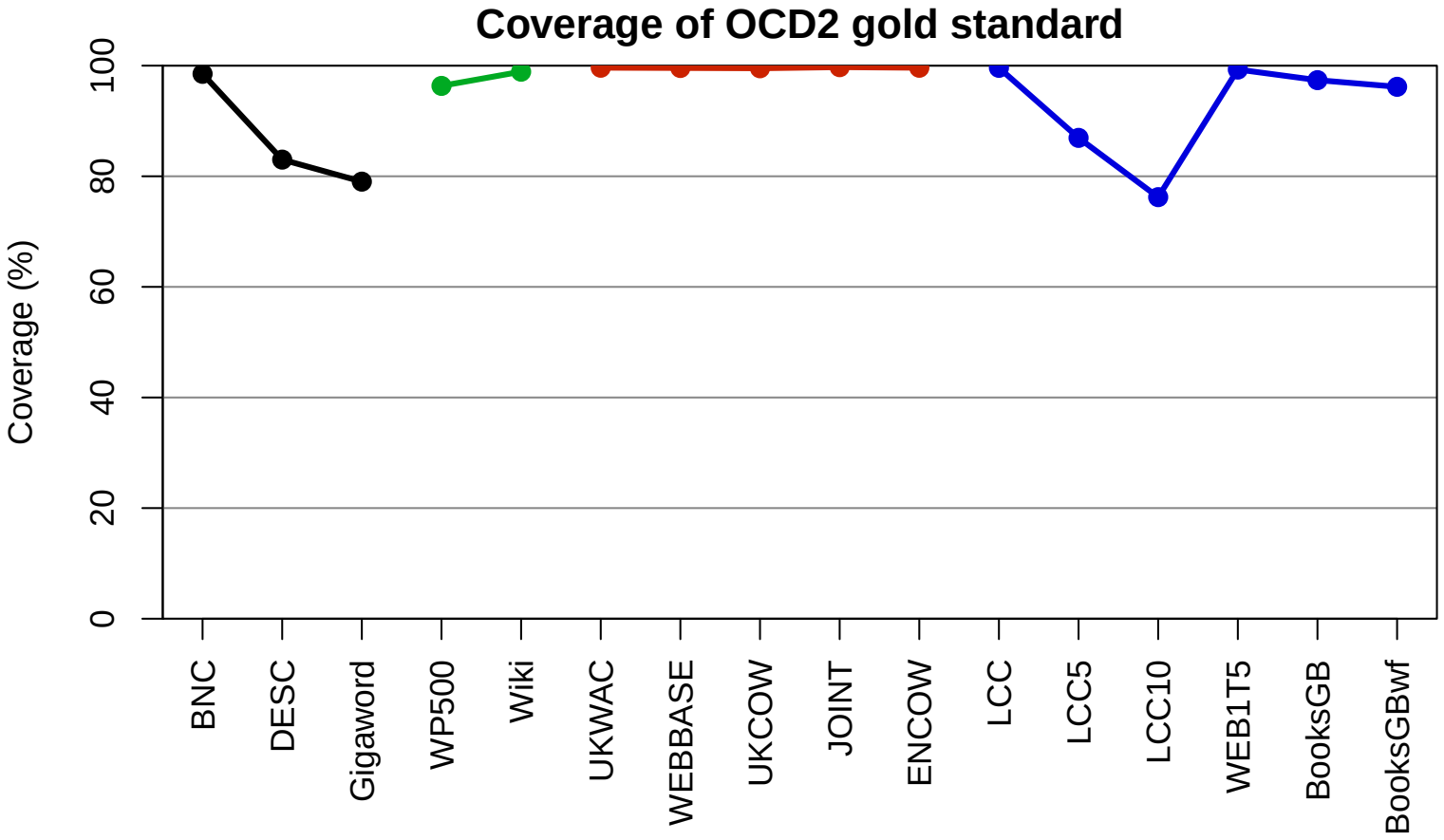
Result overview: OCD2 collocations



Result overview: OCD2 collocations

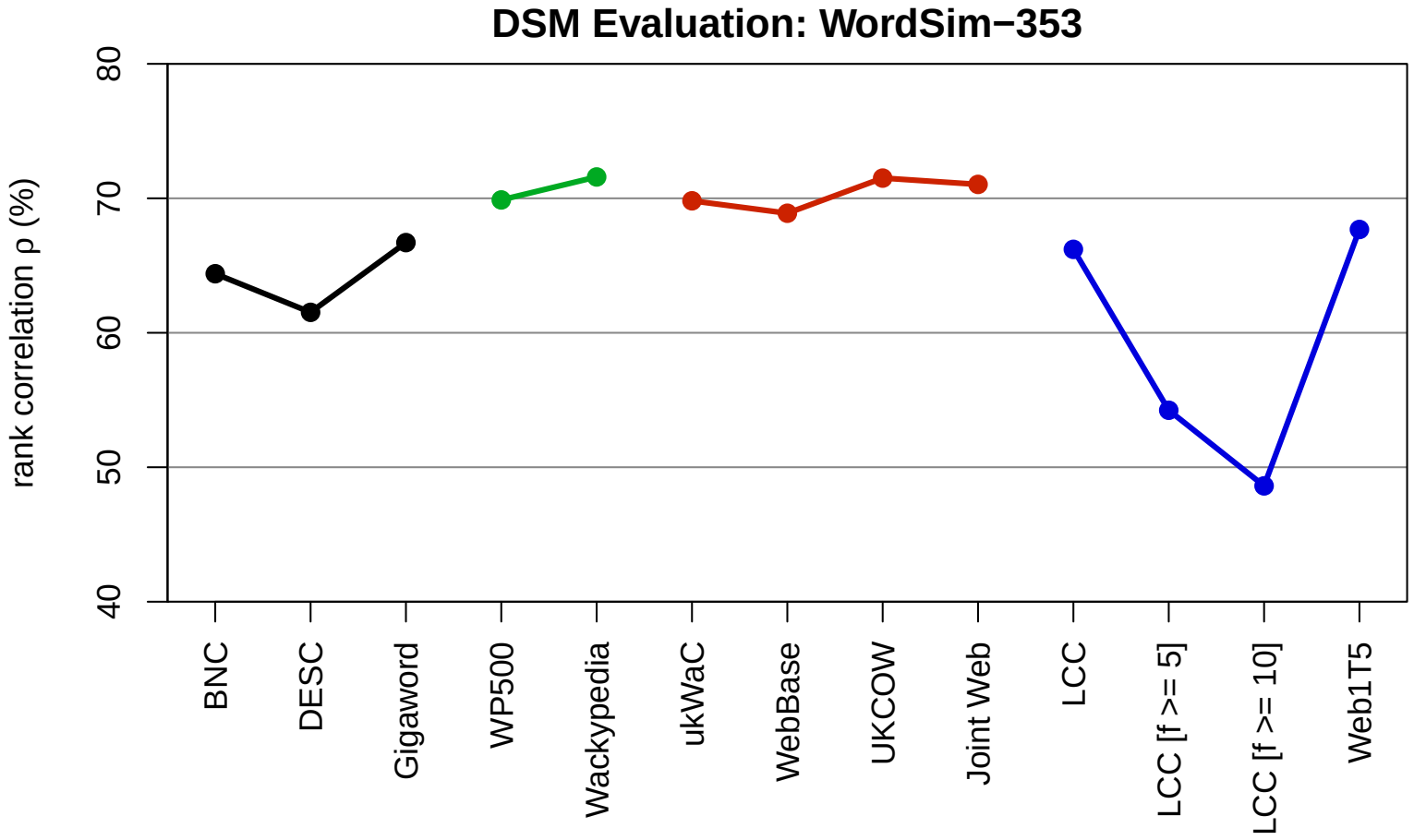


Result overview: OCD2 collocations

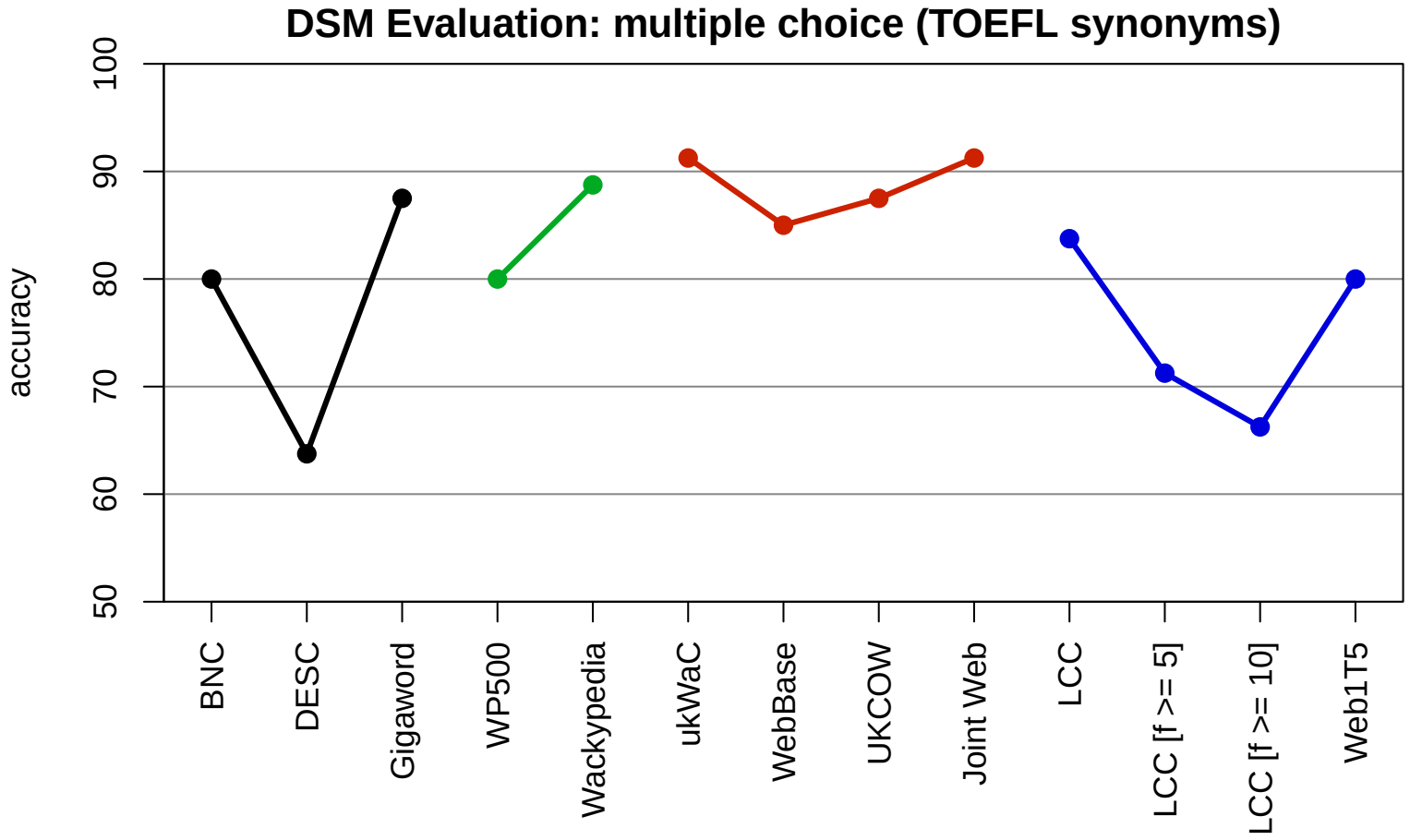


coverage analysis

Result overview: Distributional semantics

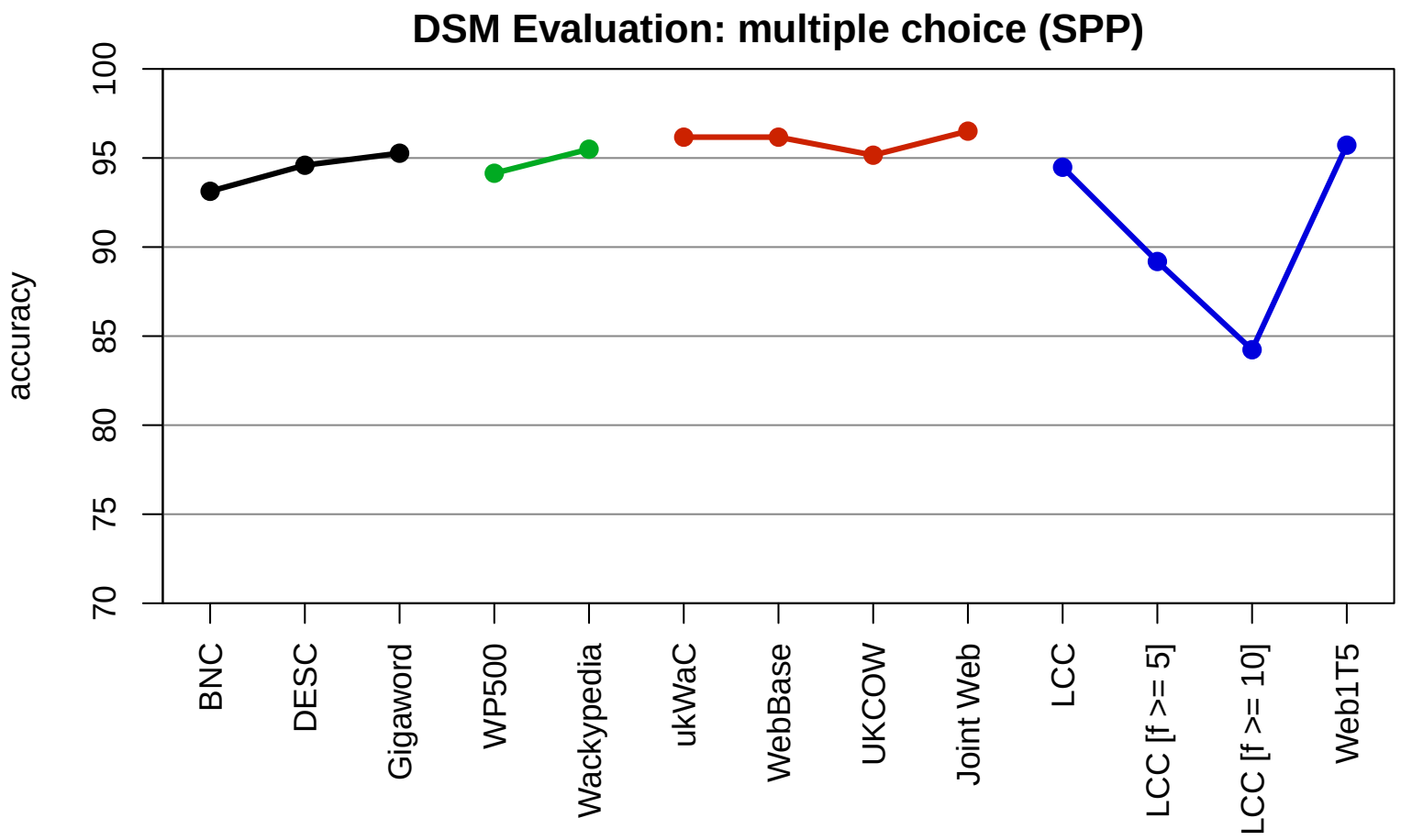


Result overview: Distributional semantics

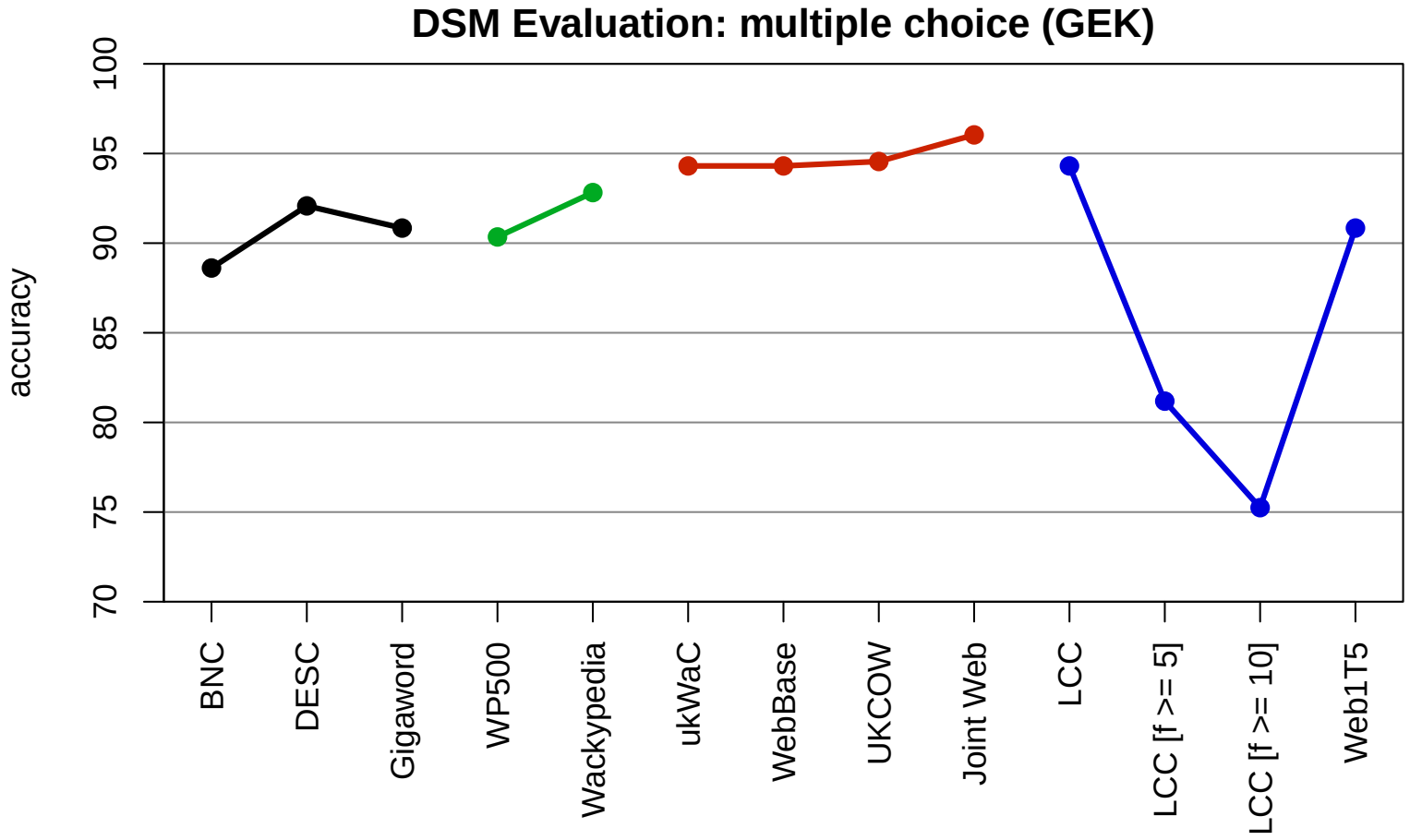


(don't take this too seriously)

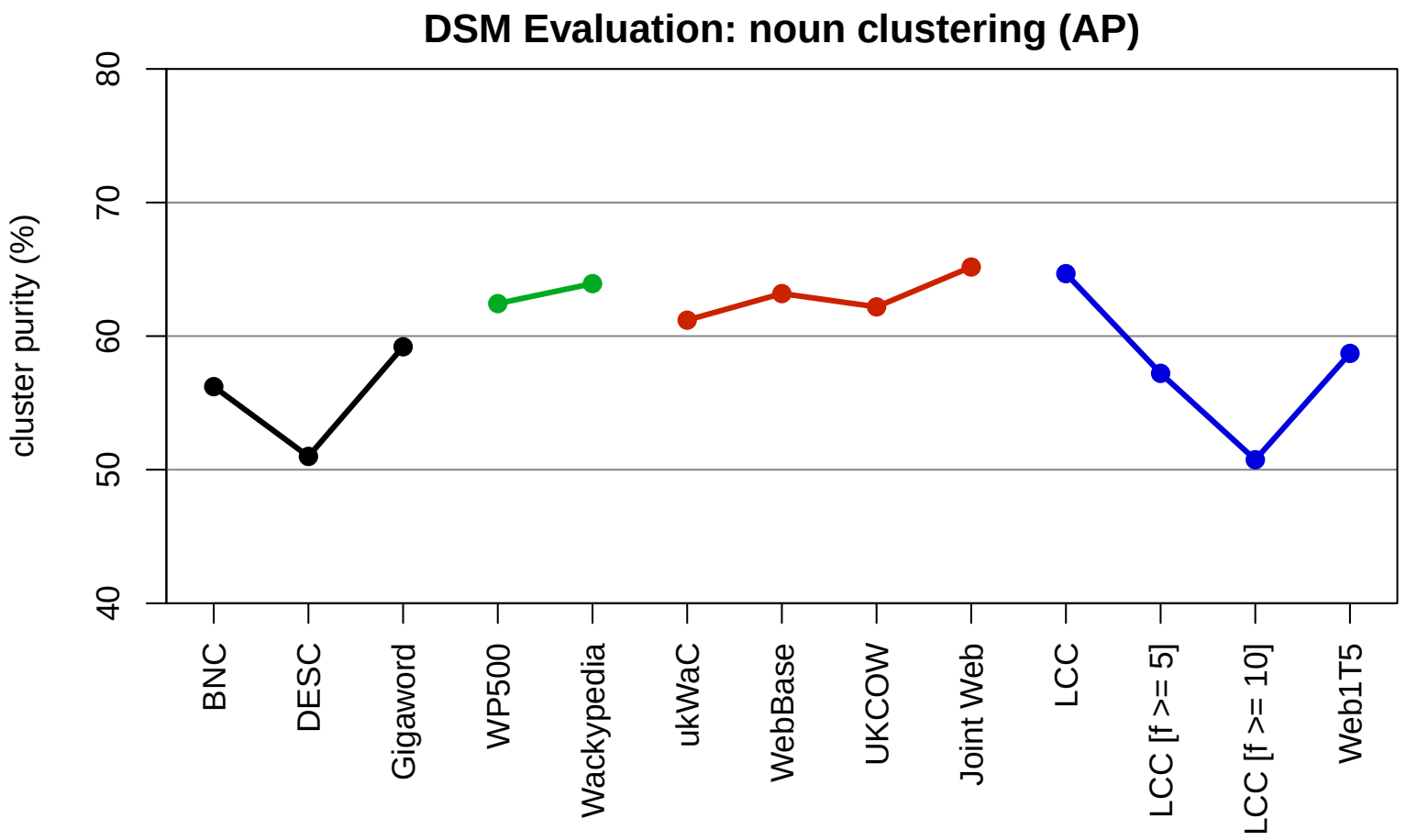
Result overview: Distributional semantics



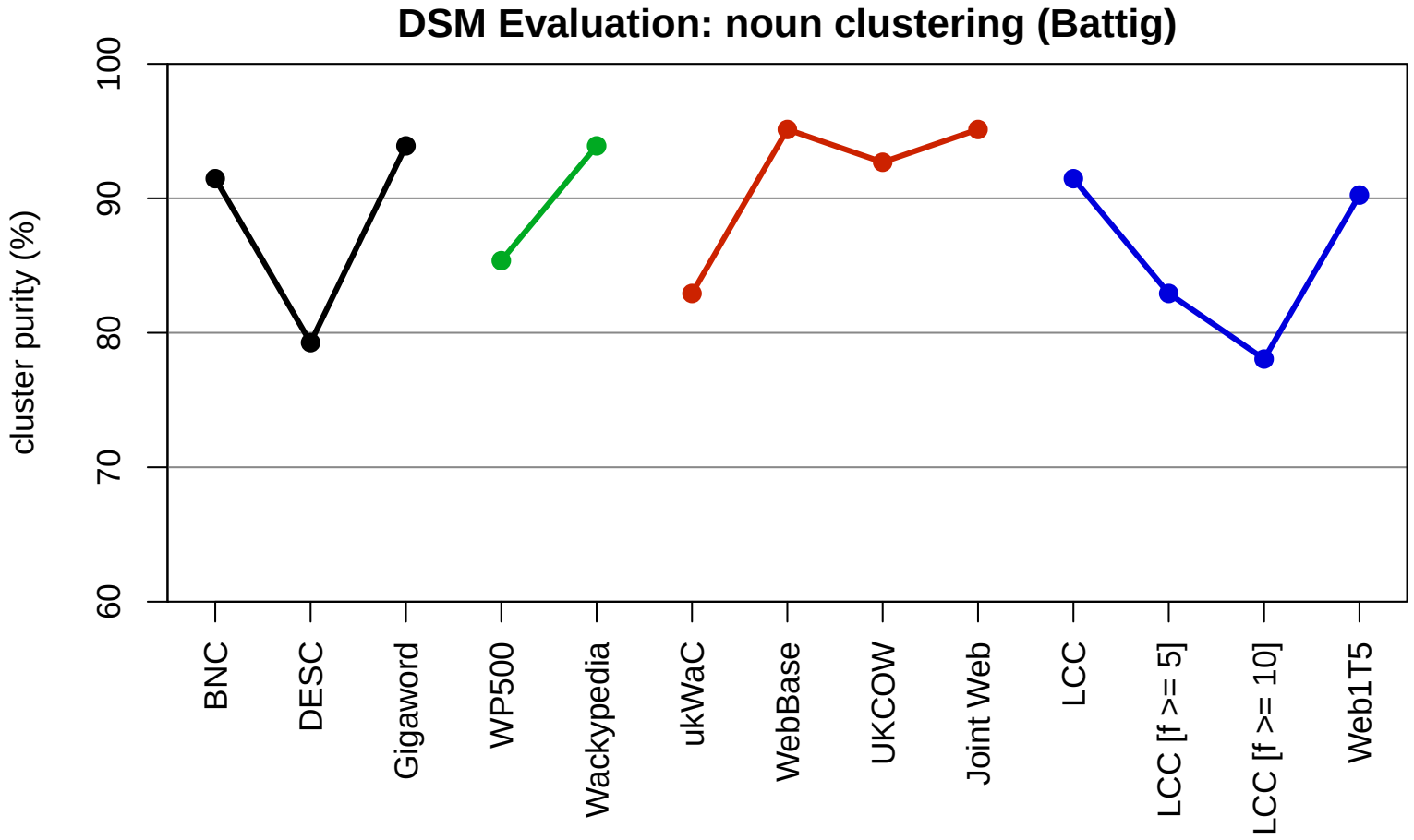
Result overview: Distributional semantics



Result overview: Distributional semantics



Result overview: Distributional semantics



(don't take this too seriously)

Thank You!

References I

Almuhareb, Abdulrahman (2006). *Attributes in Lexical Acquisition*. Ph.D. thesis, University of Essex.

Aston, Guy and Burnard, Lou (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh. See also the BNC homepage at `http://www.natcorp.ox.ac.uk/`.

Baldwin, Timothy (2008). A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 1–2, Marrakech, Morocco.

Banko, Michele and Brill, Eric (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France.

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.

References II

Bartsch, Sabine and Evert, Stefan (2014). Towards a Firthian notion of collocation. In A. Abel and L. Lemnitzer (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, number 2/2014 in OPAL – Online publizierte Arbeiten zur Linguistik, pages 48–61. Institut für Deutsche Sprache, Mannheim.

Benson, Morton; Benson, Evelyn; Ilson, Robert (1986). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam, New York.

Biemann, Chris; Bildhauer, Felix; Evert, Stefan; Goldhahn, Dirk; Quasthoff, Uwe; Schäfer, Roland; Simon, Johannes; Swiezinski, Leonard; Zesch, Torsten (2013). Scalable construction of high-quality Web corpora. *Journal for Language Technology and Computational Linguistics (JLCL)*, **28**(2), 23–59.

Brants, Thorsten and Franz, Alex (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, PA. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13`.

Church, Kenneth W. and Mercer, Robert L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, **19**(1), 1–24.

References III

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from `http://www.collocations.de/phd.html`.

Evert, Stefan (2010). Google Web 1T5 n-grams made easy (but not for the computer). In *Proceedings of the 6th Web as Corpus Workshop (WAC-6)*, pages 32–40, Los Angeles, CA.

Ferretti, Todd; McRae, Ken; Hatherell, Ann (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, **44**(4), 516–547.

Finkelstein, Lev; Gabrilovich, Evgeniy; Matias, Yossi; Rivlin, Ehud; Solan, Zach; Wolfman, Gadi; Ruppin, Eytan (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20**(1), 116–131.

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.

References IV

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*. Association for Computational Linguistics. Corpus available from `http://ebiquity.umbc.edu/resource/html/id/351`.

Hare, Mary; Jones, Michael; Thomson, Caroline; Kelly, Sarah; McRae, Ken (2009). Activating event knowledge. *Cognition*, **111**(2), 151–167.

Harris, Zellig (1954). Distributional structure. *Word*, **10**(23), 146–162. Reprinted in Harris (1970, 775–794).

Hausmann, Franz Josef (1989). Le dictionnaire de collocations. In F. J. Hausmann, O. Reichmann, H. E. Wiegand, and L. Zgusta (eds.), *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch*, Handbücher zur Sprach- und Kommunikationswissenschaft, pages 1010–1019. de Gruyter, Berlin, New York.

Hutchison, Keith A.; Balota, David A.; Neely, James H.; Cortese, Michael J.; Cohen-Shikora, Emily R.; Tse, Chi-Shing; Yap, Melvin J.; Bengson, Jesse J.; Niemeyer, Dale; Buchanan, Erin (2013). The semantic priming project. *Behavior Research Methods*, **45**(4), 1099–1114.

References V

Lapesa, Gabriella and Evert, Stefan (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, **2**, 531–545.

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, Jeju Island, Korea. Association for Computational Linguistics. Data sets available from `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`.

McIntosh, Colin; Francis, Ben; Poole, Richard (eds.) (2009). *Oxford Collocations Dictionary for students of English*. Oxford University Press.

McRae, Ken; Hare, Mary; Elman, Jeffrey L.; Ferretti, Todd (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, **33**(7), 1174–1184.

Rubenstein, Herbert and Goodenough, John B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**(10), 627–633.

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, Istanbul, Turkey. ELRA.

References VI

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

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.

Van Overschelde, James; Rawson, Katherine; Dunlosky, John (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, **50**, 289–335.