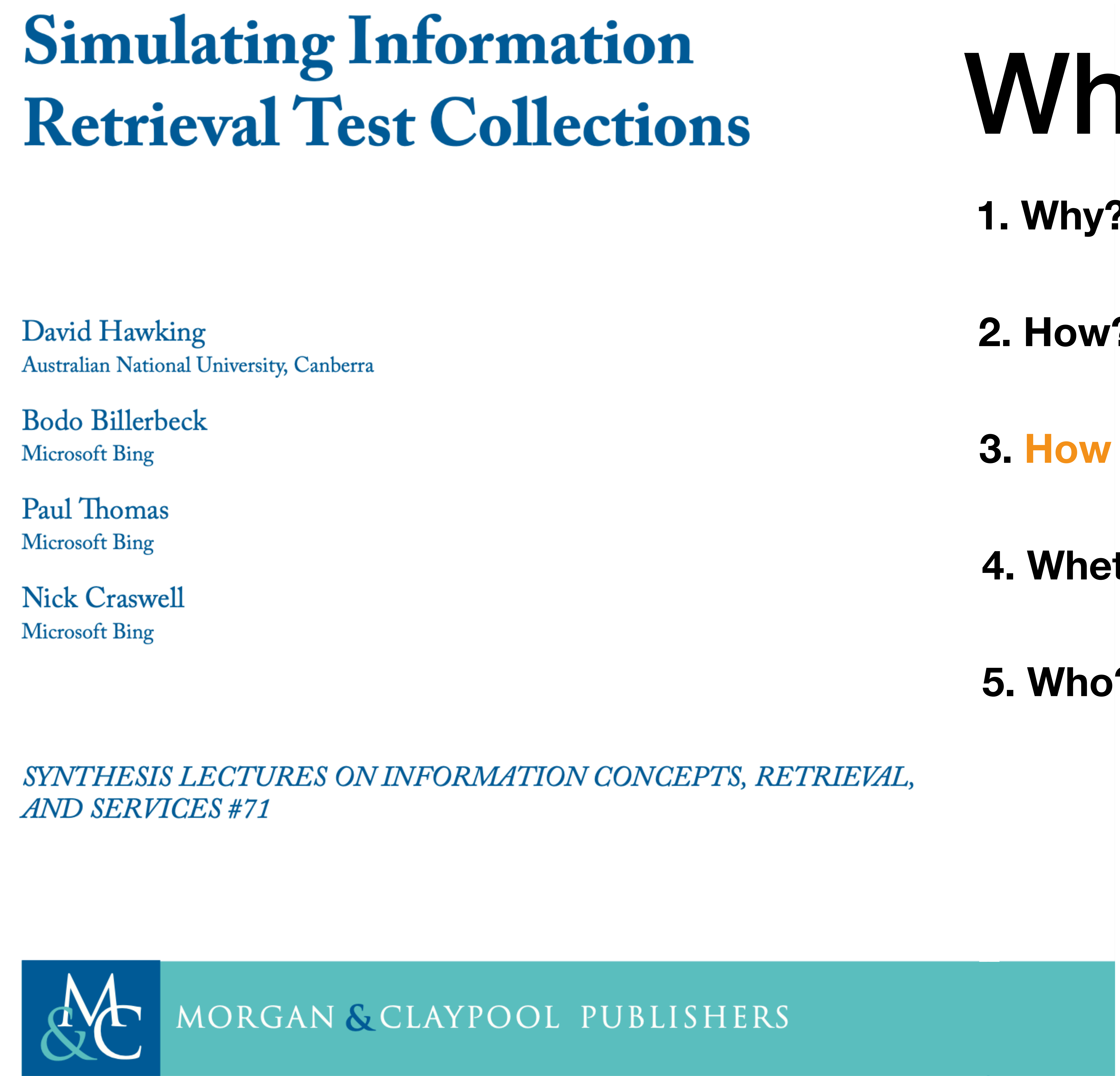
How useful are results from simulated offline IR collections?

David Hawking, ANU, Canberra, Australia
david.hawking@acm.org
(Joint work with Bodo Billerbeck, Nick Craswell and Paul Thomas)

Sim4IR Workshop, SIGIR2021, 15 July 2021



Simulating Information Retrieval Test Collections

David Hawking
Australian National University, Canberra

Bodo Billerbeck
Microsoft Bing

Paul Thomas
Microsoft Bing

Nick Craswell
Microsoft Bing

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #71

MORGAN & CLAYPOOL PUBLISHERS

What?

1. Why? — Motivations for simulation

2. How? — Simulation methods

3. How good is simulation?

4. Whether? —Risk-benefit discussion

5. Who? When?— History of simulation in IR

Why?

- Tuning, training, resource allocation on private collections
- Reproducible efficiency experimentation
  - Ability to engineer corpus properties
- Meaningful study of scalability

How?

- Language models, LDA topic models
- Markov
- Encryption - Caesar, Nomenclator
- Macro methods, e.g. Synthacorpus
- Neural methods: LSTMs, GPT-2

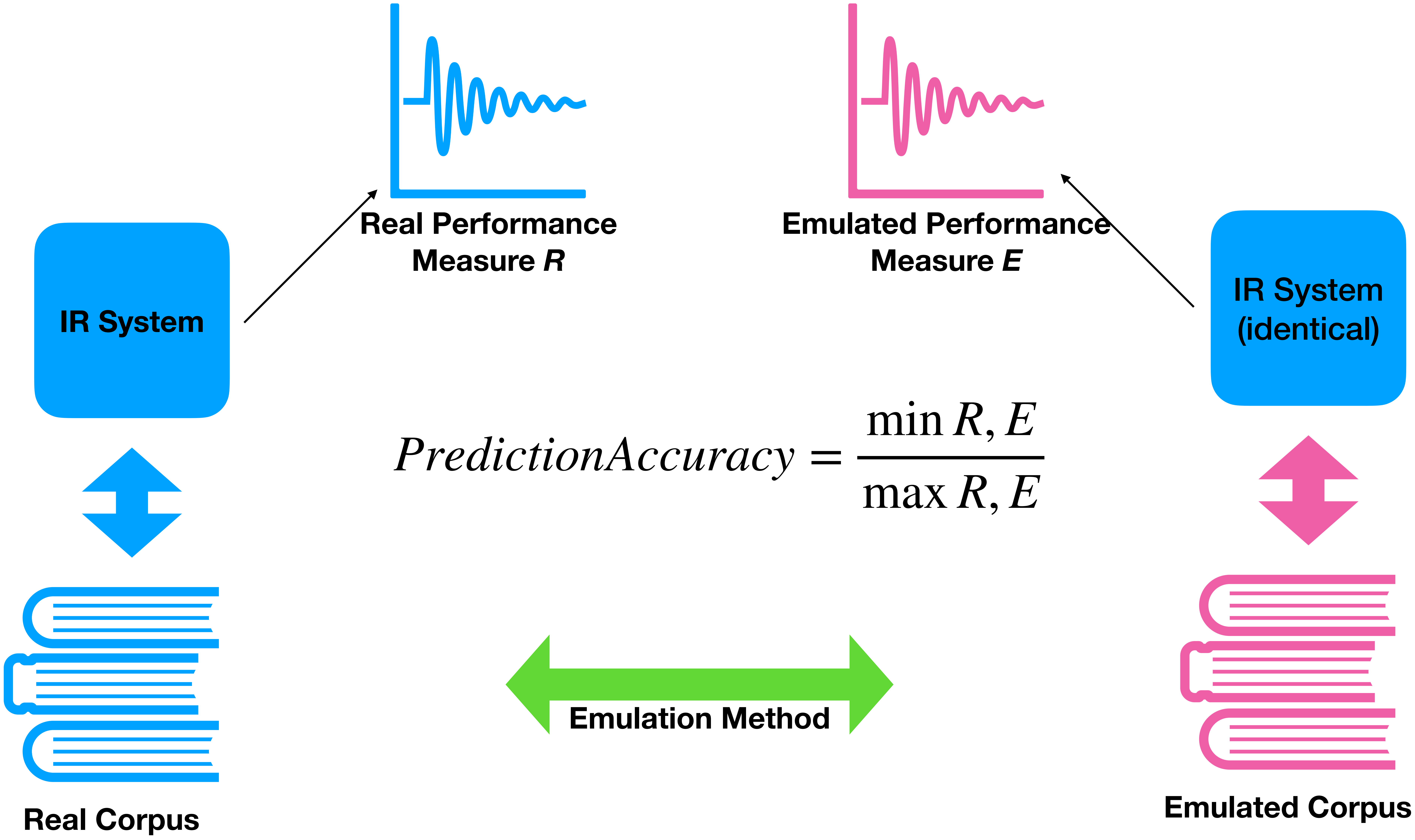
```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
```

https://bitbucket.org/davidhawking/synthacorpus

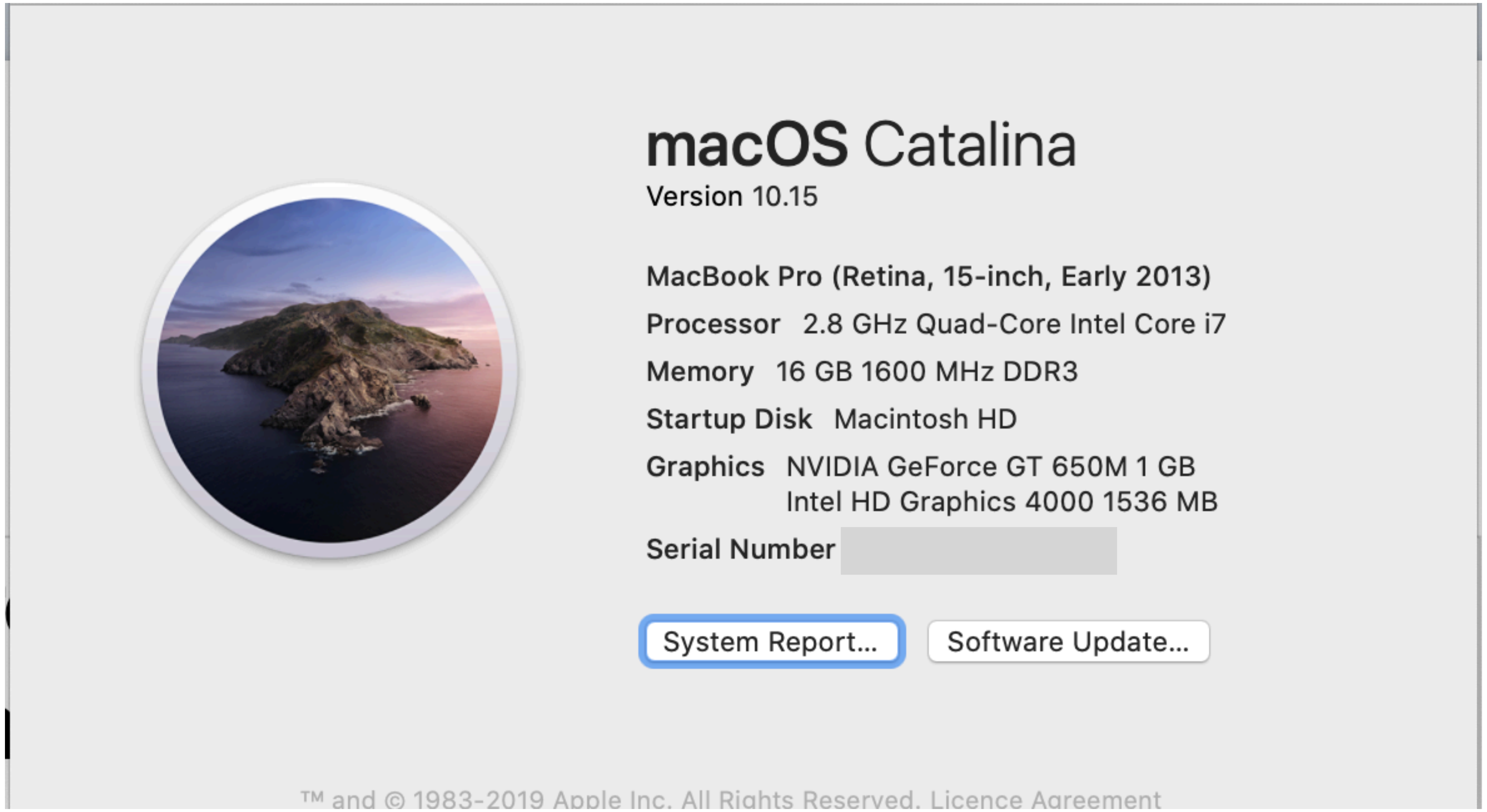
Corpus emulation with SynthaCorpus

```
Usage: emulateARealCorpus.pl <corpus_name> <tf_model> <doc_length_model> <term_repn_model>
<dependence_model> [-dependencies=neither|both|base|mimic]
    <corpus_name> is a name (e.g. TREC-AP) not a path.  We expect to find a single file called
          <corpus_name>.tsv, <corpus_name>.trec, or <corpus_name>.starc in the directory ../
Experiments/Base
    <tf_model> ::= Piecewise|Linear|Copy
          If Piecewise we'll use a 3-segment term-frequency model with 10 headpoints, 10 linear
          segments in the middle and an explicit count of singletons.  If Linear we'll approximate
          the whole thing as pure Zipf.  If Copy we'll copy the exact term frequency distribution
          from the base corpus.
    <doc_length_model> ::= dlnormal|dlsegs|dlhisto
          If dlhisto or dlsegs is given, the necessary data will be taken from the base corpus.
          Recommended: dlhisto (Unfortunately dlgamma not available in this version)
    <term_repn_model> ::= tnum|base26|bubble_babble|simpleWords|from_tsv|markov-9e?
          The Markov order is specified by the single digit, represented by '9'.
          If present, the 'e' specifies use of the end-of-word symbol.  Otherwise
          a random length will be generated for each word and it will be cut off there.
          The Markov model will be trained on the base corpus.
          If from_tsv is given, the vocab will be that of the base corpus.
          Recommended: from_tsv if appropriate or markov-5e (6e or 7e on large RAM machines)
    <dependence_model> ::= ind|ngrams[2-5]|bursts|coocs|fulldep
          Currently, only ind(ependent) and ngramsX are implemented.  Ind means that words are
          generated completely independently of each other.  Fulldep means ngrams + bursts
          + coocs.  Dependence models are only applied if the relevant files, i.e ngrams.termids,
          bursts.termids, coocs.termids, are available for the base corpus.
```

How good is an emulation?



$$PredictionAccuracy = \frac{\min R, E}{\max R, E}$$



macOS Catalina
Version 10.15

MacBook Pro (Retina, 15-inch, Early 2013)
Processor 2.8 GHz Quad-Core Intel Core i7
Memory 16 GB 1600 MHz DDR3
Startup Disk Macintosh HD
Graphics NVIDIA GeForce GT 650M 1 GB
Intel HD Graphics 4000 1536 MB
Serial Number [redacted]

™ and © 1983-2019 Apple Inc. All Rights Reserved. Licence Agreement

SSD: 768GB

Using simplified Azzopardi, de Rijke, Balog method for generating known item queries:

- **Each query set contains 1000 queries**

Let’s compare the prediction accuracy of 5 emulation methods across 8 underlying measures, 4 base corpora, and 3 retrieval systems

Which corpora?

- TREC ap
- TREC fr
- TREC patents
- WT10g

After running base corpora through `detrec` to reduce to:
• indexable words, plus
• document boundary markup: DOC and DOCNO,

after converting character encodings to UTF-8.

(Emulation methods produce the same format.)

Which retrieval systems?

- Indri (LM)
- Terrier (DFR)
- ATIRE (BM25)

Which emulation methods?

SimpleSynth

```
t26362 t368932 t64855 t33466 t044332 t62265 t23046 t78835 t843821
t264032 t23285 t996501 t909682 t221021 t016831 t832522 t885692 t68428
t159031 t98886 t7284 t982411 t327802 t344882 t73642 t685372 t289752
t404341 t5841 t64914 t27763 t674702 t378461 t999731 t847232 t467012
t752122 t614761 t327702 t563871 t73307 t843911 t064941 t802901
```

Real

```
<DOC>
<DOCNO> AP880212-0001 </DOCNO>
<TEXT>
Reports Former Saigon Officials Released from Re education
Camp More than 150 former officers of the overthrown South
Vietnamese government have been released from a re education
camp after 13 years of detention the official Vietnam News
Agency reported Saturday
...
</TEXT>
</DOC>
```

SophSynth

```
crash praisal pi in crash do kamleh ik crash nomadic vauhgan
gimbels crash oo ut boo crash de ux boo crash de ux nev crash
abu iba ma crash xa bogersonellaeg boo crash hob coatham fle
```

Caesar

```
Sfqpsut Gpsnfs Tbjhpo Pggjdjbmt Sfmfbtfe gspn Sf fevdbujpo Dbnq Npsf
uibo 261 gpsnfs pggjdfst pg uif pwfsuispxo Tpvui Wjfuobnftf hpwfsonfou
ibwf cffo sfmfbtfe gspn b sf fevdbujpo dbnq bgufs 24 zfbst pg
efufoujpo uif pggjdjbm Wjfuobn Ofxt Bhfodz sfqpsufe Tbuvsebz Uif
sfqpsu gspn Ibopj npojupsfe jo Cbohlpl eje opu hjwf tqfdjgjd gjhvsft
```

Nomenclator

```
moschorsholt biarrithem vladish esbuscovar ngau competanya padrnos
kumsisant fu derauding abori cristyn bederick vladish chalis gierkeg
herbed bullistoforceab casperimentativ estheticaly nhilunbuy carlatt
bonnoticeable competanya padrnos acri kumsisant fu derauding juicines
recurragchaa scaffold gierkeg guntumbley herbed destructuring sepate
```

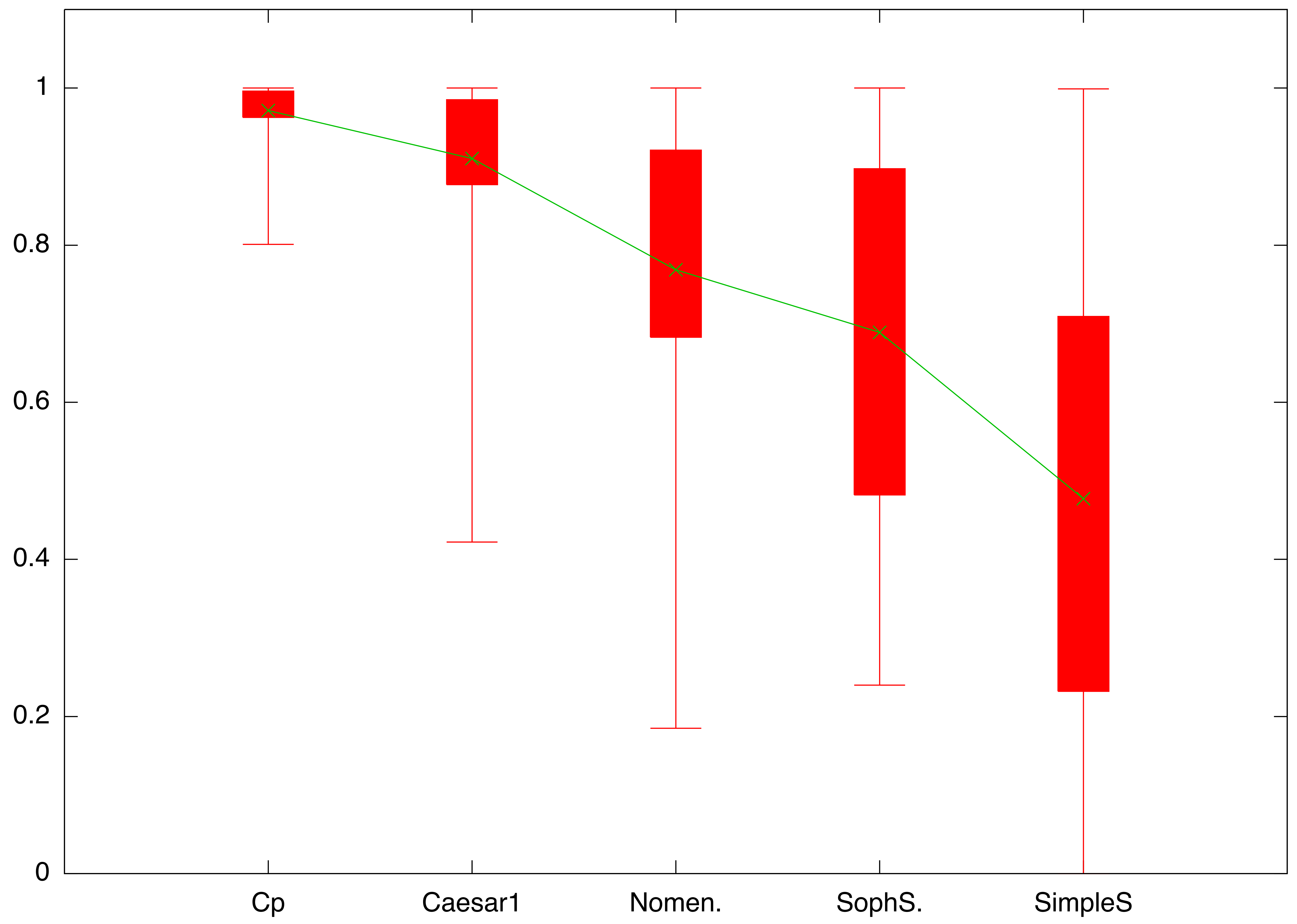
- /bin/cp — indication of noise
- Caesar substitution
- Nomenclator substitution
- SynthaCorpus Sophisticated
- SynthaCorpus Simple

Emulation method	**Preservation of confidentiality**	**Expected prediction accuracy rank**
Copy	None	1
Caesar	None	2
Nomenclator	OK in limited circumstances.	3
SophSynth	Good	4
SimpleSynth	Good	5

We emulated TREC-AP with a neural method GPT-2 — too slow to include in this experiment.

Which underlying measures?

- Indexing time
- Indexing memory use
- Query processing (QP) time (3wd, 6wd, 9wd query sets)
- Mean reciprocal rank (3wd, 6wd, 9wd query sets)



Accuracy scores averaged across all the measures, all the retrieval systems, all the query lengths, and all the corpora.

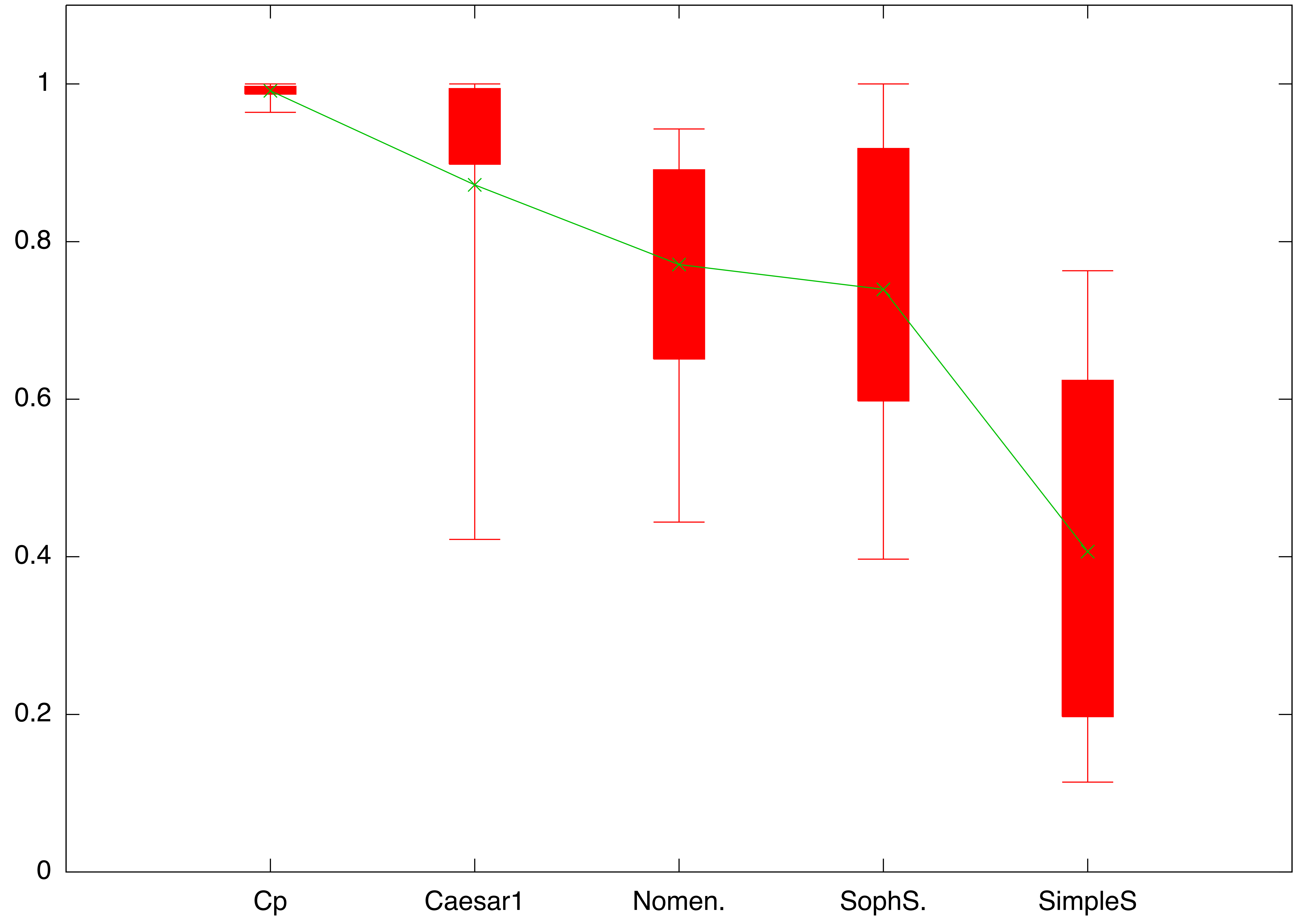
- **Solid — 2nd and 3rd quartiles**
- **Whisker — range**
- **Green — mean**

Each data point is the mean of five trials. A new query set is generated for each trial.

Cp noise due to disk layouts and variation in query sets.

Left to right decline is as expected.

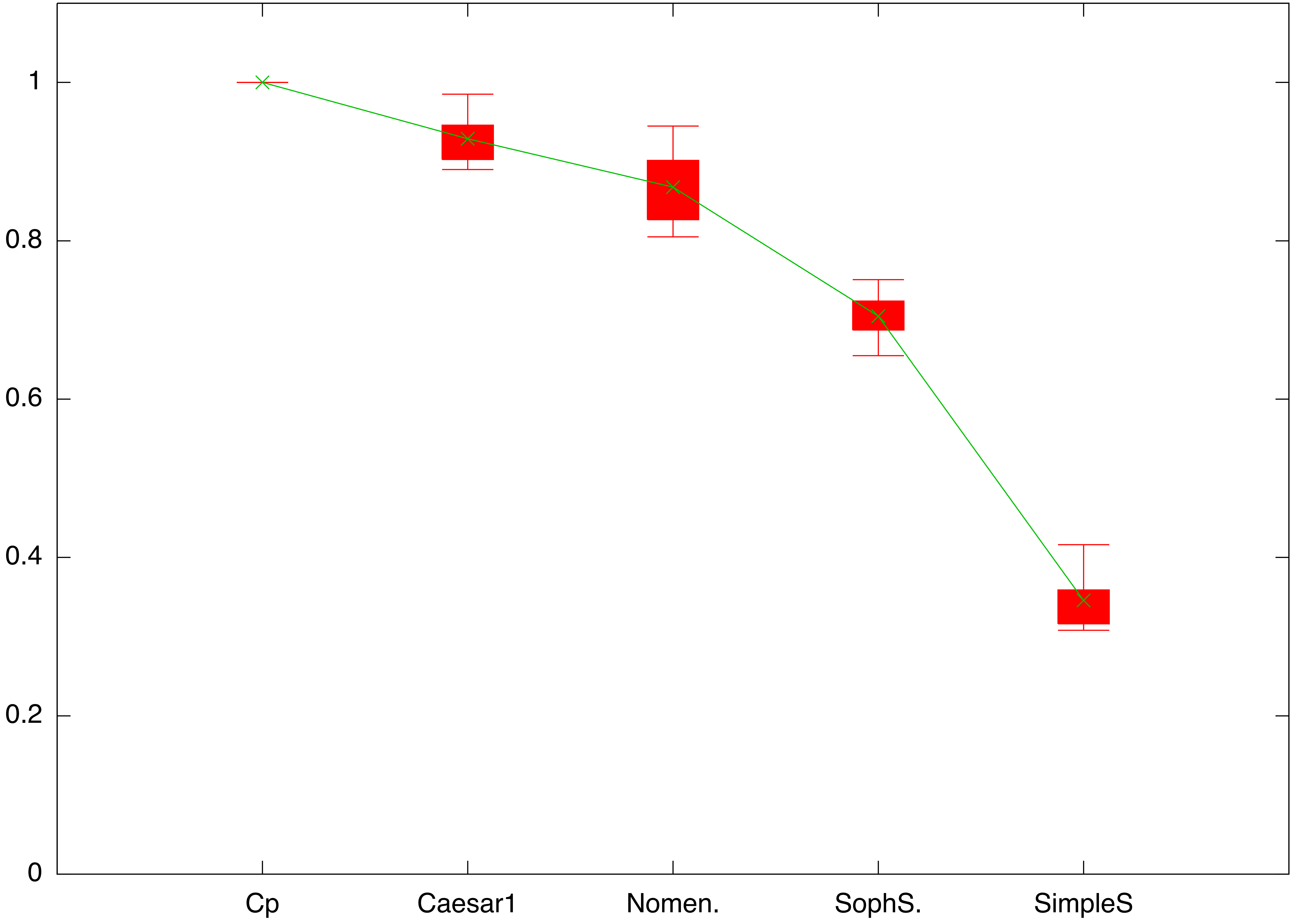
A lot of variation, even for the best emulation methods.



Accuracy scores for Indexing Time averaged across all the retrieval systems and all the corpora

Less noise for Cp because no variance due to query generation.

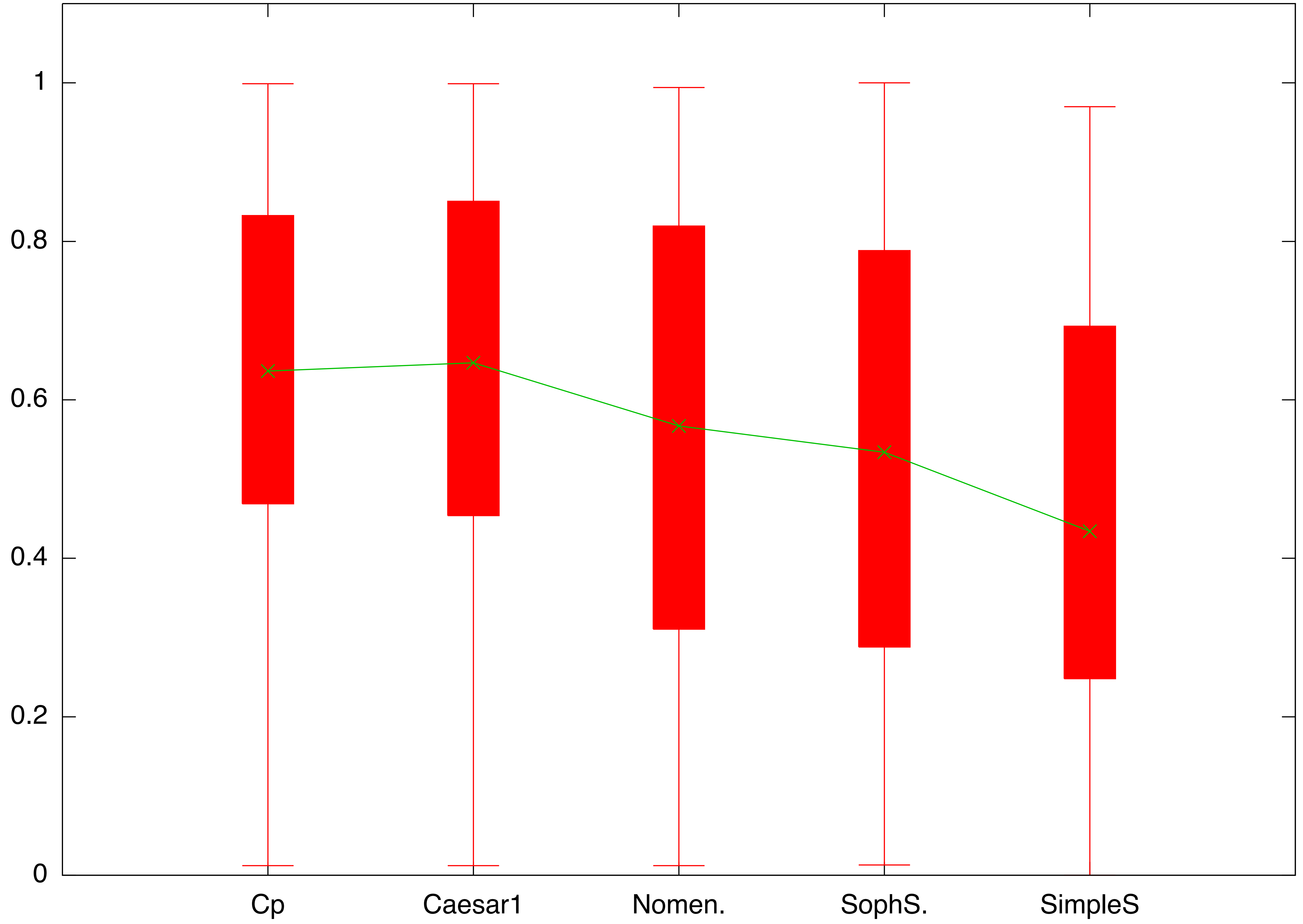
SimpleSynth is much worse, presumably because word frequency distribution is uniform.



Accuracy scores for Indexing Memory for ATIRE averaged across all the corpora

ATIRE clearly reported memory use. I couldn't see how to obtain meaningful figures for the others.

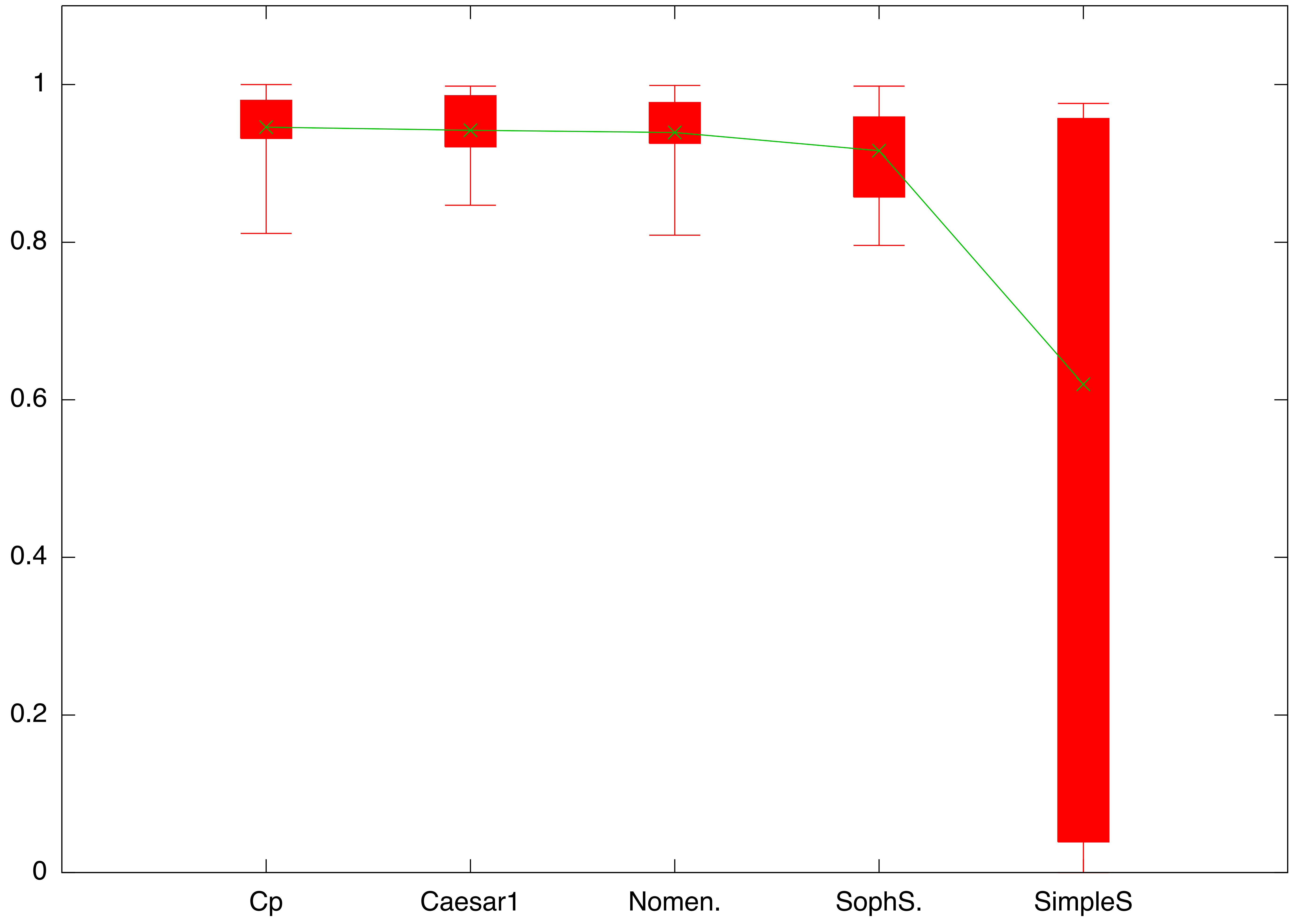
Interesting that SimpleS is so different. |V| and N are the same as for other methods, word freq. list. and term representations are very different.



Accuracy scores for Query Processing Time averaged across all the query lengths, all the retrieval systems and all the corpora

Very wide variation in prediction accuracy for time to process 1000 queries. None of the emulation methods give reliable predictions.

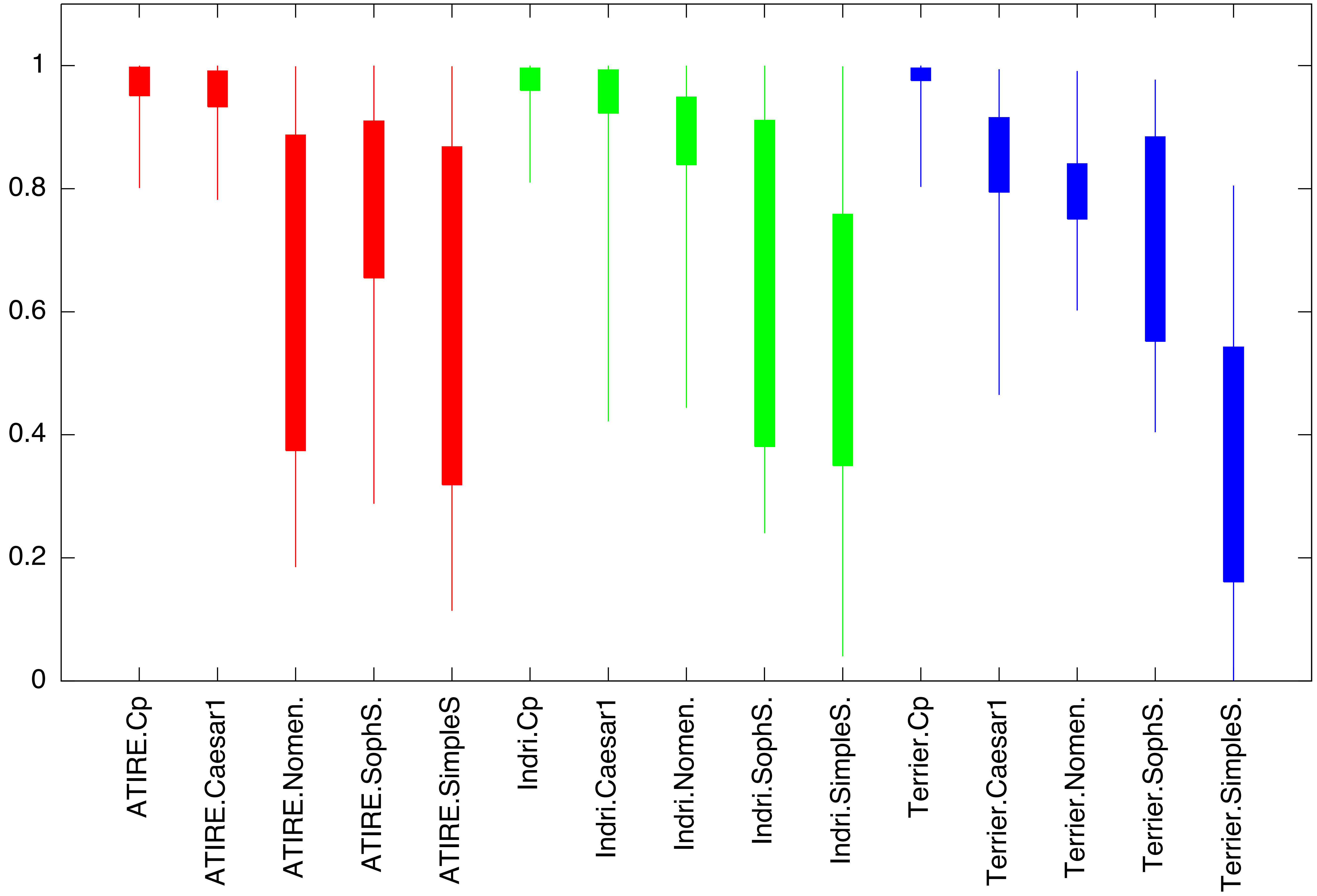
Like to make a hypothesis?



Four emulation methods give very good prediction of MRR performance.

With a uniform word freq. distribution, SimpleSynth makes it difficult to choose queries which discriminate a known item.

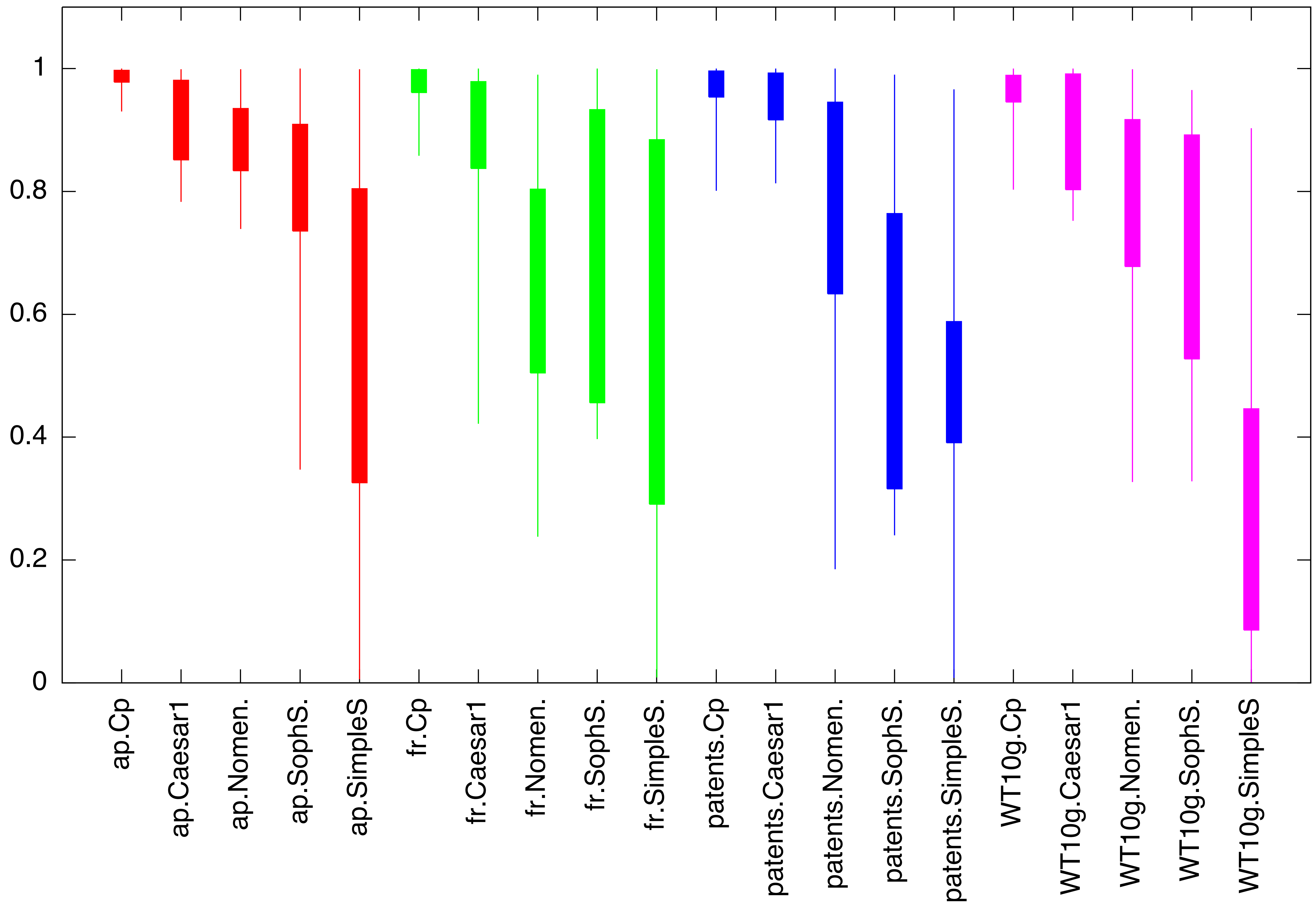
Accuracy scores for Mean Reciprocal Rank averaged across all the query lengths, all the retrieval systems and all the corpora



Accuracy scores <u>per retrieval system</u> averaged across all the query lengths, all the measures and all the corpora

Nomenclator gives worse predictions for ATIRE than for the other two.

SophSynth gives worse predictions for Indri than for the other two.



Accuracy scores per corpus averaged across all the query lengths, all the retrieval systems and all the measures

Clearly there is something of a corpus effect — see WT10g.SimpleSynth.

I have decided not to show you accuracy results for the 5 x 4 x 4 = 80 individual conditions 😇

Whether?

- Cp and Caesar are just baselines — can't be used in a condfidentiality environment.

- Opinion: SimpleSynth doesn't make good enough predictions.

- Opinion: Only Nomenclator and SophSynth make accurate enough predictions for use in practice.

- Opinion: It would be hard to crack rare words in Nomenclator, even through n-gram frequency attack, or with the availability of some plain-cypher paired text.

- Opinion: SynthaCorpus methods do not leak confidential information.

- Data Owner's Opinion: Whether Nomenclator or SynthaCorpus methods provide sufficient protection.

- SynthaCorpus provides a compact means (parameters + random seed) by which a researcher can allow reproduction of experimental results obtained on a private corpus.

- SynthaCorpus can engineer corpora with specific properties to explore and understand the behaviour of IR systems.

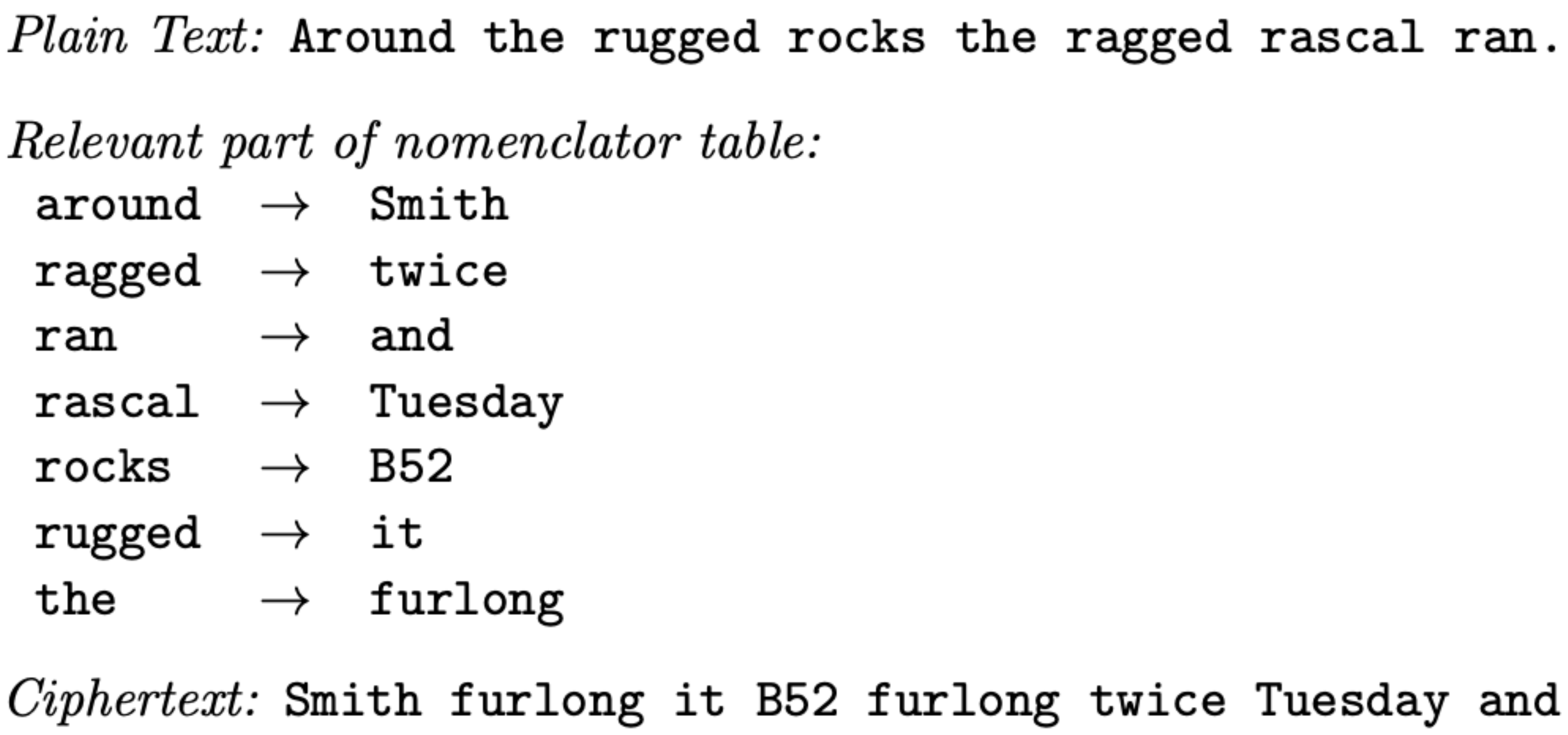
- SynthaCorpus incorporates growth models which allow realistic scaling up of a corpus, including vocabulary growth (à la Herdan / Heaps), thus permitting meaningful study of algorithmic scalability

Who? When?

- 1966 C.R. Blunt et al — simulating information storage and retrieval systems.
- 1973 M.D. Cooper — artificial corpora (tiny!) built from topic models
- 1980 J. Tague et al — simulation of document term matrix
- 1996 T. Kanungo — generation of degraded text
- 2000 E. Reiter et al — Building natural language generation systems
- 2006/7 L. Azzopardi — building simulated queries
- 2010 D.L. Chen et al — automated sportscasting
- 2011 I. Sutskever et al — generating text with recurrent neural networks. Also Karpathy, Radford et al.
- 2012/13 R. Berendsen et al — generating test collections for learning to rank
- 2016 D. Maxwell et al — simulated users

Please let me know of any other relevant work 😊

Nomenclator explanation



Plain Text: `Around the rugged rocks the ragged rascal ran.`

Relevant part of nomenclator table:

around	→	Smith
ragged	→	twice
ran	→	and
rascal	→	Tuesday
rocks	→	B52
rugged	→	it
the	→	furlong

Ciphertext: `Smith furlong it B52 furlong twice Tuesday and`