STS: SigDiff
The statistical significance of differences

Computational Approaches to Collocations

Vienna, July 2002

Stefan Evert

Collocation identification

- Extract candidate pairs from corpus
  - adjacent word pairs (or pairs within window)
  - Adj+N pairs from NP chunks
  - Obj+V, Subj+V & PP+V from parse trees
- Rank candidates by "association scores"
  - true collocations should obtain high scores
  - using association measures (AMs)
- *N*-best list of highest-ranking candidates

Test data: AdjN set

- Adjective+Noun pairs
- Corpus of German law texts (800k words)
- Extraction of candidates
  - adjacent adjective+noun pairs, lemmatised
- Criteria for manual identification of TPs
  - intuitive notion of "typical" combinations
  - marked by two annotators
- Frequency threshold: $f \geq 2$

Test data: PNV set

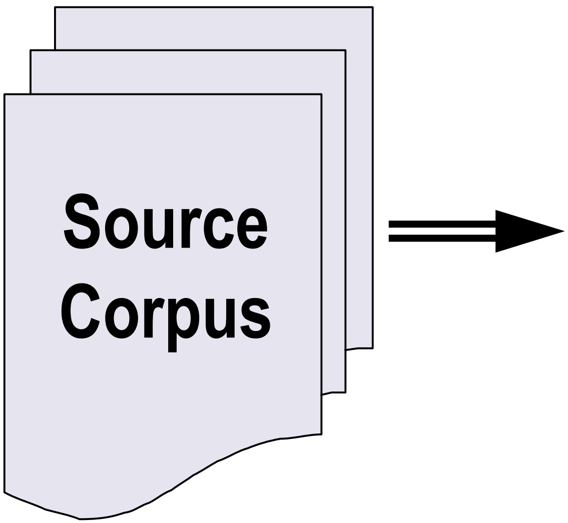
- PP(Prep,N)+Verb pairs
- Newspaper corpus (*FR*, 8M words)
- Extraction of candidates
  – PP chunks and verbs co-occurring in sentence
  – verbs are lemmatised, but *not* the PPs
- Criteria for manual identification of TPs
  – idiomatic expression *or* support verb construction
- Frequency threshold: $f \geq 3$

Test data: summary

AdjN data	
total	11 087
$f \geq 2$	4 652
TPs ($f \geq 2$)	15.84% = 737

PNV data	
total	294 534
$f \geq 3$	14 654
TPs ($f \geq 3$)	6.41% = 939

Evaluation procedure



candidate list

Candidate pair	t-score
abgefaßt Entscheidung	1.371
abgefaßt Urteil	1.984
abgegeben Erklärung	2.195
abgegeben Gutachten	1.403
abgegeben Stimme	4.789
abgegeben Willenserklärung	1.723
abgekürzt Form	1.988
abgelaufen Haushaltsplan	1.414
abgelaufen Zeit	1.388
abgelegt Zeugnis	1.414
abgelehnt Gesetzesvorlage	1.414
abgelehnt Mitglied	1.391
abgelehnt Richter	2.619
abgeordnet Richter	1.406
abgesandt Handelsbrief	1.414
abgeschlossen Geschäft	2.789
abgeschlossen Hochschulstudium	1.997
abgeschlossen Raum	1.400
abgeschlossen Staatsvertrag	1.411
abgeschlossen Verfahren	2.914
...	...

Evaluation procedure

ranked cand. list

Rank	Candidate pair	t-score
1.	zuständig Behörde	17.391
2.	mündlich Verhandlung	17.239
3.	entsprechend Anwendung	17.171
4.	personenbezogen Datum	13.853
5.	deutsch Mark	13.656
6.	gesetzlich Vertreter	13.387
7.	bürgerlich Gesetzbuch	13.171
8.	geltend Vorschrift	12.832
9.	erst Rechtszug	11.841
10.	schwer Fall	10.912
11.	andere Ehegatte	10.868
12.	gentechnisch Arbeit	9.842
13.	zuständig Stelle	9.629
14.	elterlich Sorge	9.614
15.	juristisch Person	9.459
16.	sofortig Beschwerde	9.122
17.	beweglich Sache	8.998
18.	deutsch Bundespost	8.979
19.	bezeichnet Art	8.898
20.	andere Teil	8.857
...	...	...

Evaluation procedure: *N*-best lists

14 true positives

6 false positives

⇒ precision:
14/20 = 70%

total:
737 TPs

⇒ recall:
14/737 = 1.9%

Rank	Candidate pair	t-score
1.	zuständig Behörde	17.391
2.	mündlich Verhandlung	17.239
3.	entsprechend Anwendung	17.171
4.	personenbezogen Datum	13.853
5.	deutsch Mark	13.656
6.	gesetzlich Vertreter	13.387
7.	bürgerlich Gesetzbuch	13.171
8.	geltend Vorschrift	12.832
9.	erst Rechtszug	11.841
10.	schwer Fall	10.912
11.	andere Ehegatte	10.868
12.	gentechnisch Arbeit	9.842
13.	zuständig Stelle	9.629
14.	elterlich Sorge	9.614
15.	juristisch Person	9.459
16.	sofortig Beschwerde	9.122
17.	beweglich Sache	8.998
18.	deutsch Bundespost	8.979
19.	bezeichnet Art	8.898
20.	andere Teil	8.857
...	...	...

Footnote: the F-measure

- F-measure balances precision and recall
- a heuristic solution from information retrieval
- not useful for the evaluation of AMs (often: high precision, but fairly low recall)
- Yeh (2000): More accurate tests for the statistical significance of result differences is mainly concerned with the F-measure and hence not relevant in this context

Evaluation procedure: *N*-best lists

N = 100	MI	chi-sq.	t-score	log-l.	freq.
Precision	23.00%	37.00%	57.00%	65.00%	51.00%
Recall	3.12%	5.02%	7.73%	8.82%	6.91%

N = 500	MI	chi-sq.	t-score	log-l.	freq.
Precision	23.00%	34.00%	42.00%	42.80%	40.60%
Recall	15.60%	23.07%	28.49%	29.04%	27.54%

N = 1000	MI	chi-sq.	t-score	log-l.	freq.
Precision	21.70%	28.80%	32.90%	35.10%	30.70%
Recall	29.44%	39.08%	44.64%	47.63%	41.66%

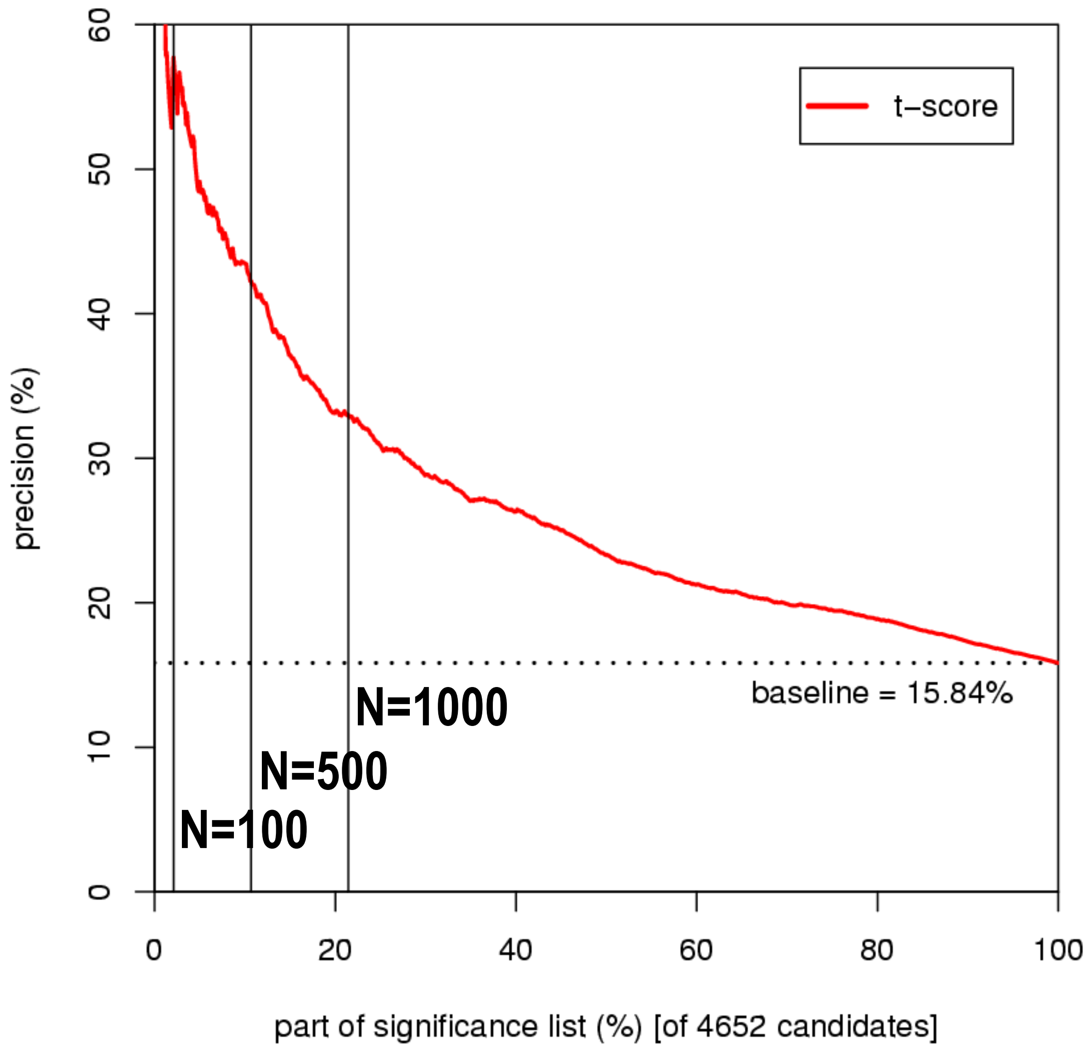
Evaluation procedure: *N*-best lists

N = 100	MI	chi-sq.	t-score	log-l.	freq.
Precision	23.00%	37.00%	**57.00%**	65.00%	51.00%
Recall	3.12%	5.02%	7.73%	8.82%	6.91%

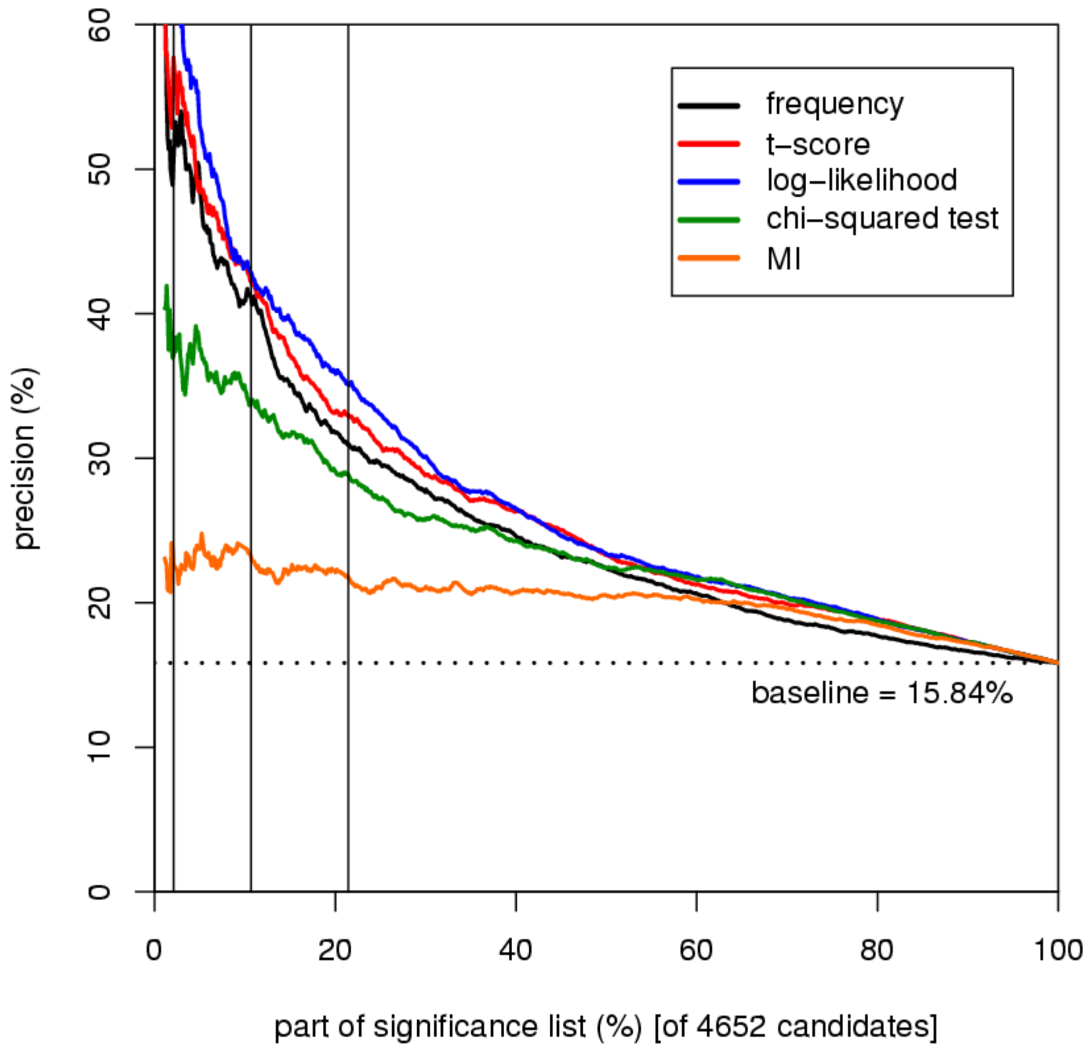
N = 500	MI	chi-sq.	t-score	log-l.	freq.
Precision	23.00%	34.00%	**42.00%**	42.80%	40.60%
Recall	15.60%	23.07%	28.49%	29.04%	27.54%

N = 1000	MI	chi-sq.	t-score	log-l.	freq.
Precision	21.70%	28.80%	**32.90%**	35.10%	30.70%
Recall	29.44%	39.08%	44.64%	47.63%	41.66%

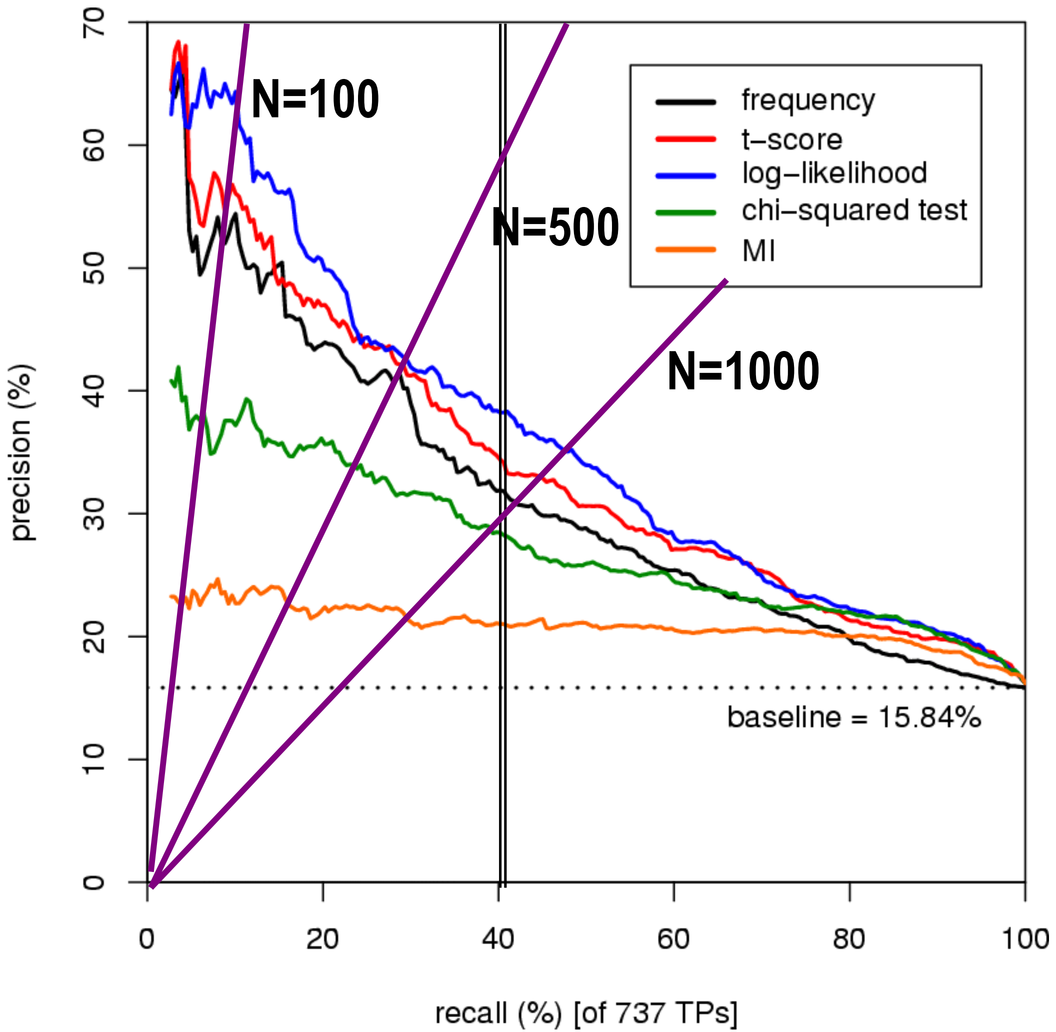
Evaluation procedure: Precision plot



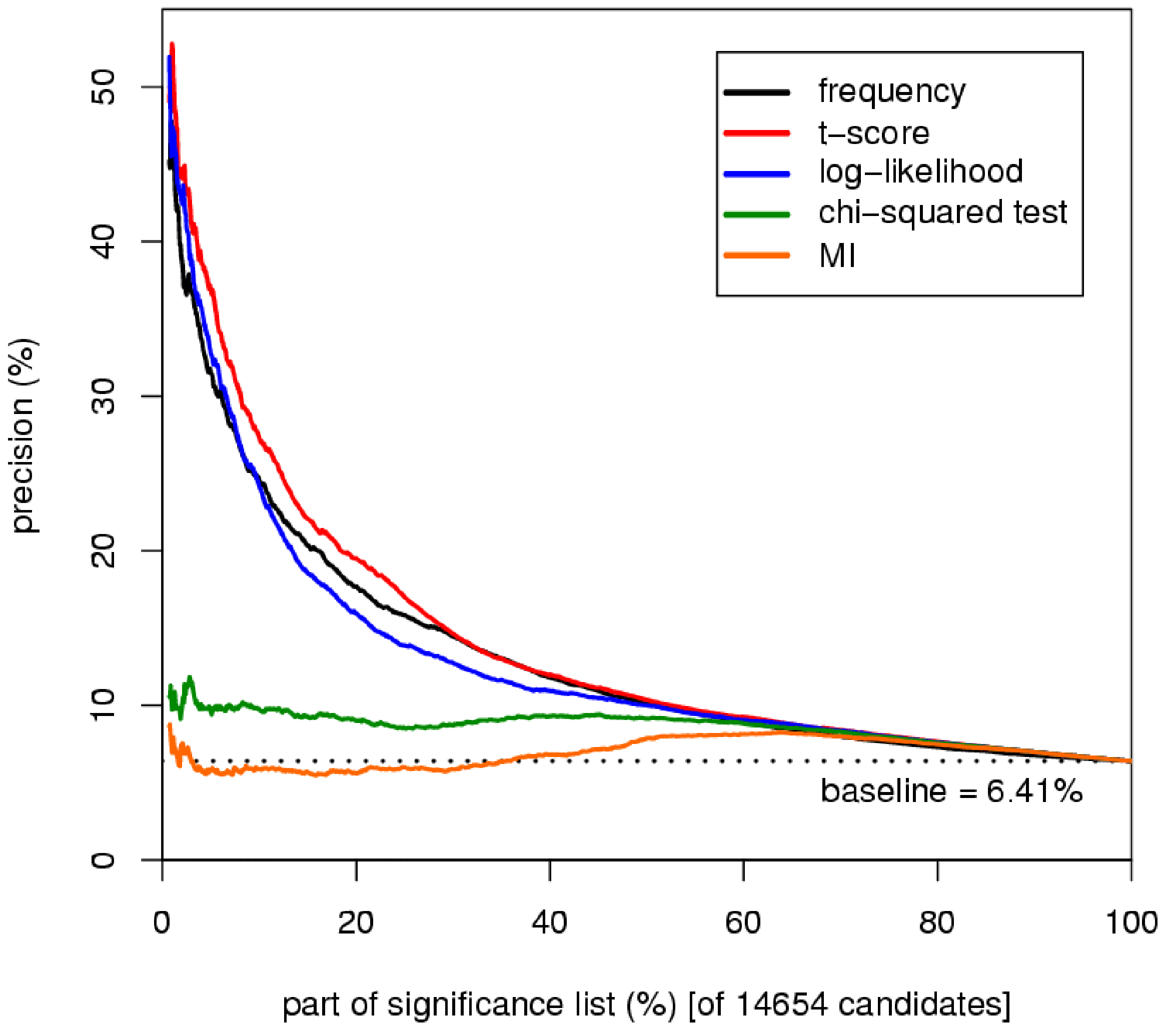
Evaluation procedure: Precision plot



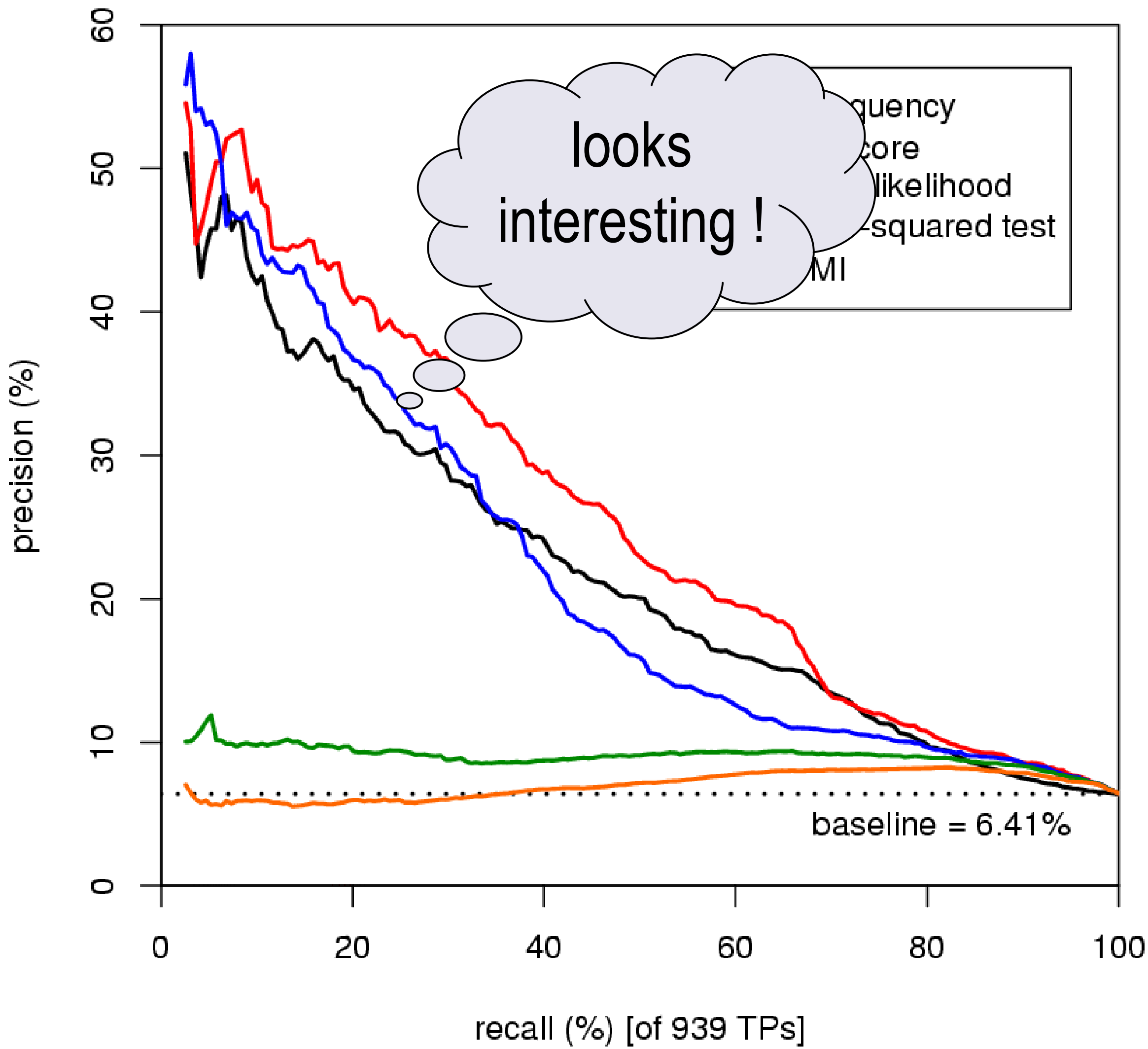
Precision against recall



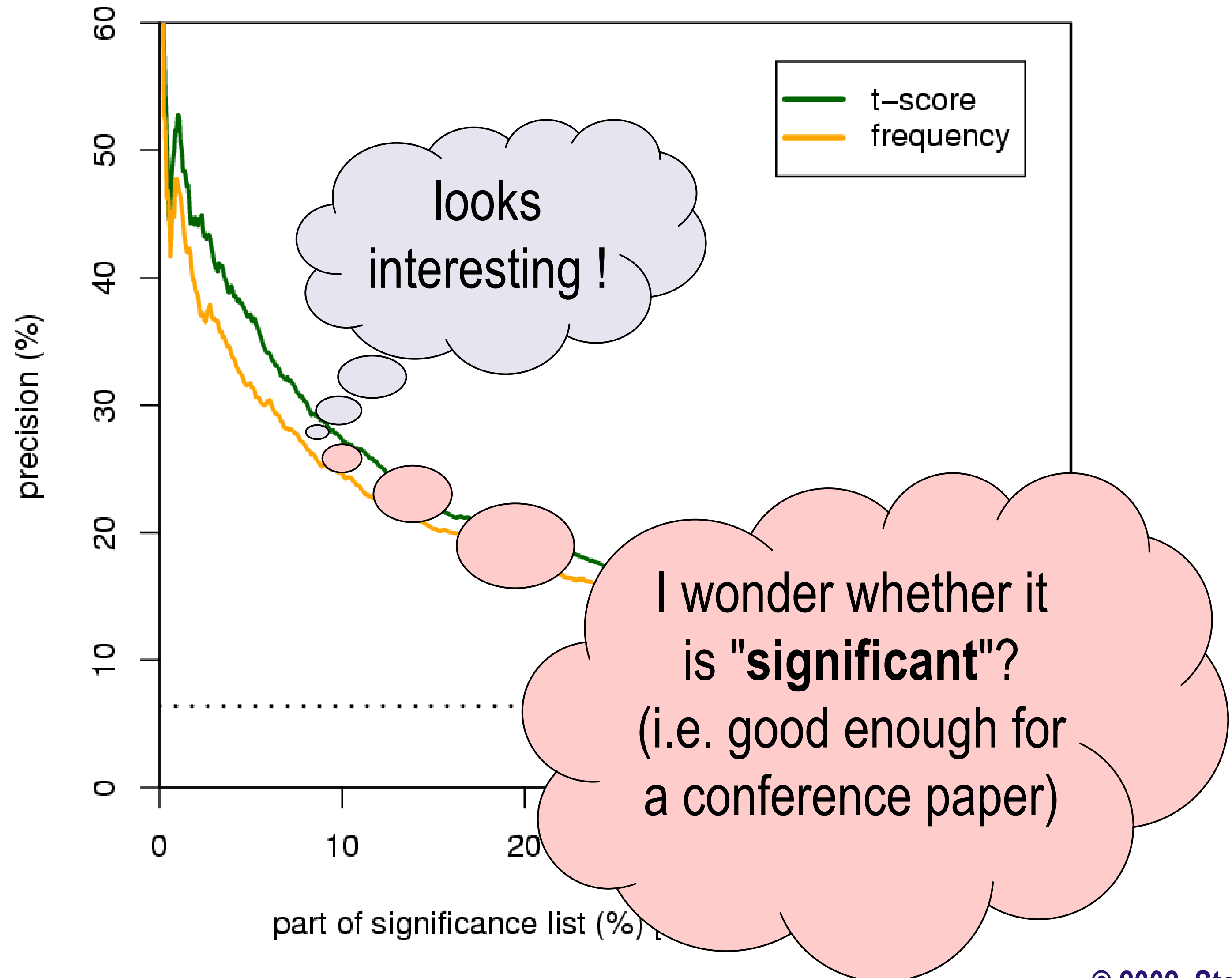
Precision graphs: PNV data



Precision against recall: PNV



Differences between two AMs



What is statistical significance?

- significant = meaningful? substantial?
- Kilgarriff (2001): Comparing Corpora cites a statistics textbook

> None of the null hypotheses we have considered with respect to goodness of fit can be *exactly* true, so if we increase the sample size (and hence the value of χ^2), we would ultimately reach the point when all null hypotheses would be rejected. All that the χ^2 test can tell us, then, is that the sample size is too small to reject the null hypothesis.
>
> (Owen/Jones: *Statistics*, 1977, p. 359)

Significance vs. relevance

- we must distinguish between significance and relevance
- **significance** = Could the observed difference be due to chance, or is there a systematic effect, however small?
- **relevance** = Is the difference large enough to be of interest for our application?
- we consider only significance in this STS (⟶ Directions for a combined approach)

Reminder: significance tests (ST)

- **null hypothesis** H_0:
  observed differences are due to chance
- **alternative hypothesis** H_1:
  anything else
- interesting cases = rejection of H_0
  (i.e. sufficient evidence against H_0)
- **significance level** = confidence of rejection
  (not the ST's confidence in its decision!)

Reminder: significance tests (ST)

- ST can make two kinds of errors:
  - **type I** error = unjustified rejection of H_0
  - **type II** error = failure to reject H_0
- risk of type I error controlled by sig. level
- type I error is highly unlikely in our situation, even if the assumptions of a ST are not met
- **power** of ST = risk of type II error and the precise meaning of H_0 are of interest

Reminder: significance tests (ST)

- ST can make two kinds
  - **type I** error = unj
  - **type II** error = fail

does anybody need to be convinced?

- risk of type I error con sig. level
- type I error is highly unlikely in our situation, even if the assumptions of a ST are not met
- **power** of ST = risk of type II error and the precise meaning of H_0 are of interest

A closer look at H_0

- Null hypothesis = question that ST answers
  "Are observed differences due to chance?"
- What is "chance", intuitively speaking?
- What does a rejection of H_0 mean?
- How are random differences explained?
- A more intuitive and explicit question:
  "If we repeated the experiment, would measure A again perform better than B?"

A closer look at H_0

- A more intuitive and explicit question:
  "If we repeated the experiment, would measure A again perform better than B?"
- this is **not** the same as the first question!
- $\longrightarrow$ Directions, for a discussion of possible formulations of the null hypothesis

Program for this STS

Program for the rest of this STS:

Look at several STs that have been used and/or are suggested by statistics textbooks

- precise formulation of H_0 (question)
- assumptions and theoretical model
- comparative discussion of ST's power

Siegel (1956): Nonparametric Tests for the Behavioral Sciences
Agresti (1990): Categorical Data Analysis

A religious belief

- **parametric** test vs. **non-parametric** (**distribution-free**) test
- parametric tests assume specific distribution (e.g. normal) with parameters
- non-parametric tests make weaker assumptions $\longrightarrow$ more general
- when applicable, parametric tests are usually more powerful

A religious belief, but ...

- many tests which assume a distribution with parameters are considered non-parametric
- even χ^2 test! (based on normal distribution)
- precisely: parametric tests assume specific distribution of a numerical property *in the population*, while non-parametric tests at most assume that a certain proportion of the population has a particular feature
- we use only non-parametric tests

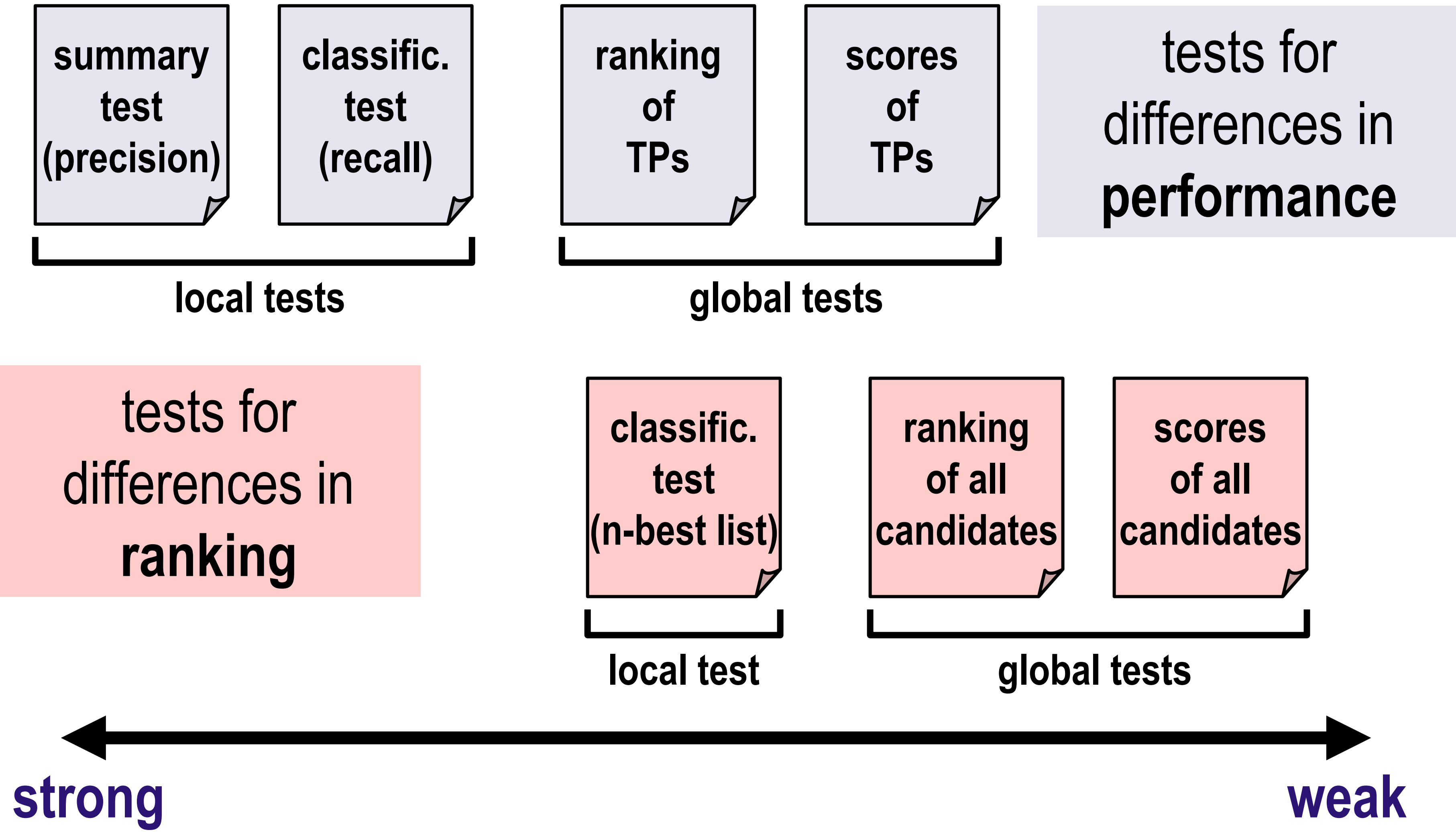
Other classification criteria

- asymptotic vs. exact test
  (less important, since we have large samples)
- continuous vs. discrete data (same as above)
- scale of measurement:
  interval $\leftrightarrow$ **ordinal** (ranking) $\leftrightarrow$ **nominal**
- **related** vs. **unrelated** samples
  (tests for independent data are usually less powerful when applied to related samples)

A practical criterion

- the **strength** of a significance test, giving a rough scale from strong to weak
- **strong** test = requires more evidence (perhaps too much) to find significant diff.
- **weak** test = detects significant differences more easily, but may overestimate them
- note that a strong test is *less* powerful
- strength is an intuitive notion $\neq$ applicability

General classification of STs



General classification of STs

nominal nominal ordinal interval

local tests

global tests

tests for differences in **performance**

tests for differences in ranking

nominal ordinal interval

local test

global tests

strong ⟷ **weak**

Pearson's χ^2 test

tbl	t-score	frequency
TPs	322	283
FPs	678	717

- number of TPs and FPs for 1000-best lists

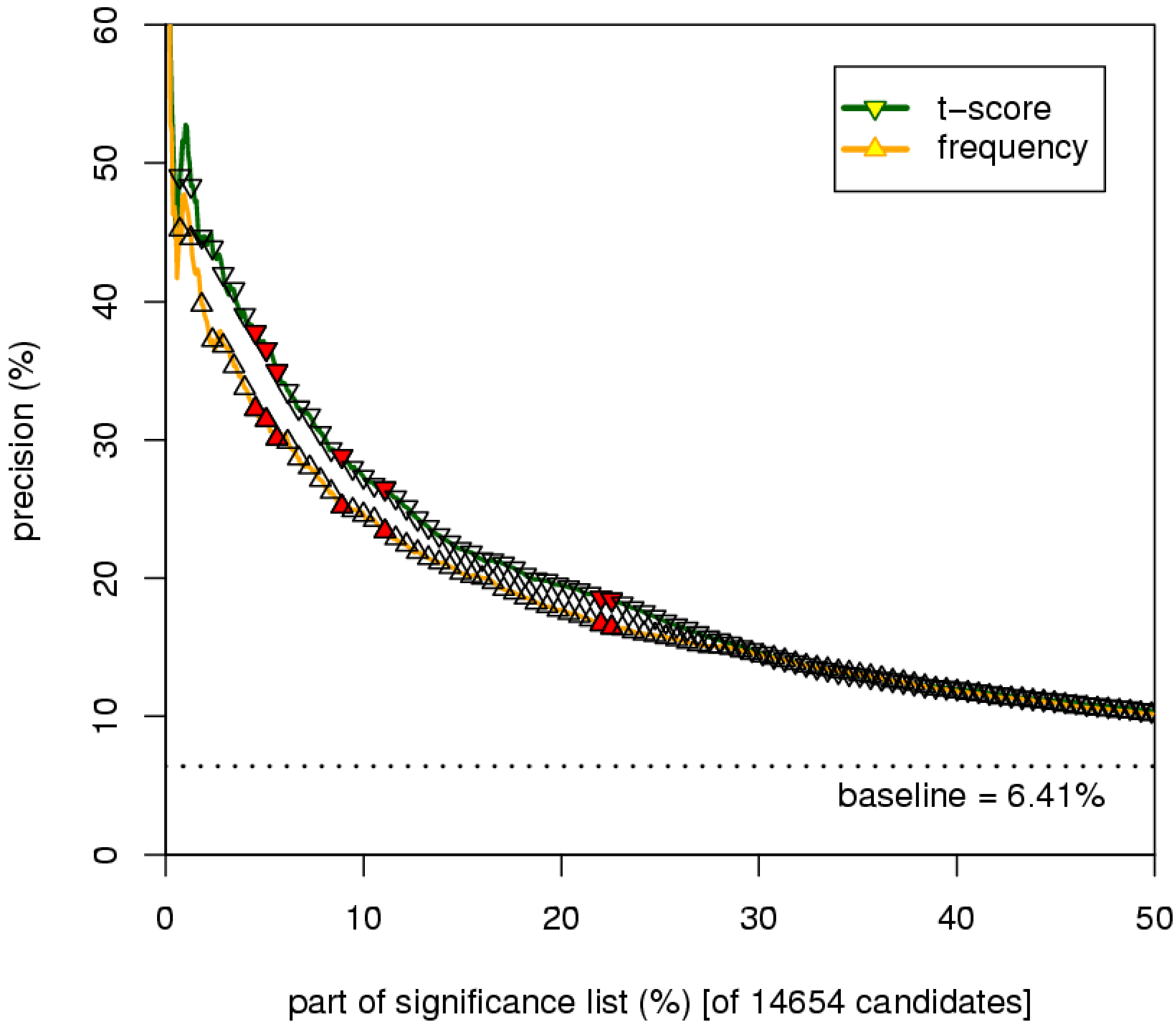
```
> chisq.test(tbl)
```

- p-value = 0.064 $\longrightarrow$ difference not significant

Pearson's χ^2 test

- we usually apply χ^2 test at 95% confidence level (significance level α=0.05)
- *either* perform χ^2 test for an "interesting difference" determined from the plot
- *or* compute χ^2 test for various *N*-best lists and add results to precision graphs
- cannot easily mark significant differences in precision-against-recall plot

Precision graph with χ^2 test



Multiple comparisons

- for introduction, see e.g. Cohen (1996): Getting What You Deserve from Data
- at 95% confidence level, as many as one in twenty test applications will randomly report a significant difference (type I error)
- therefore, even if there are no systematic differences between the two AMs, we must expect some red triangles in the plot

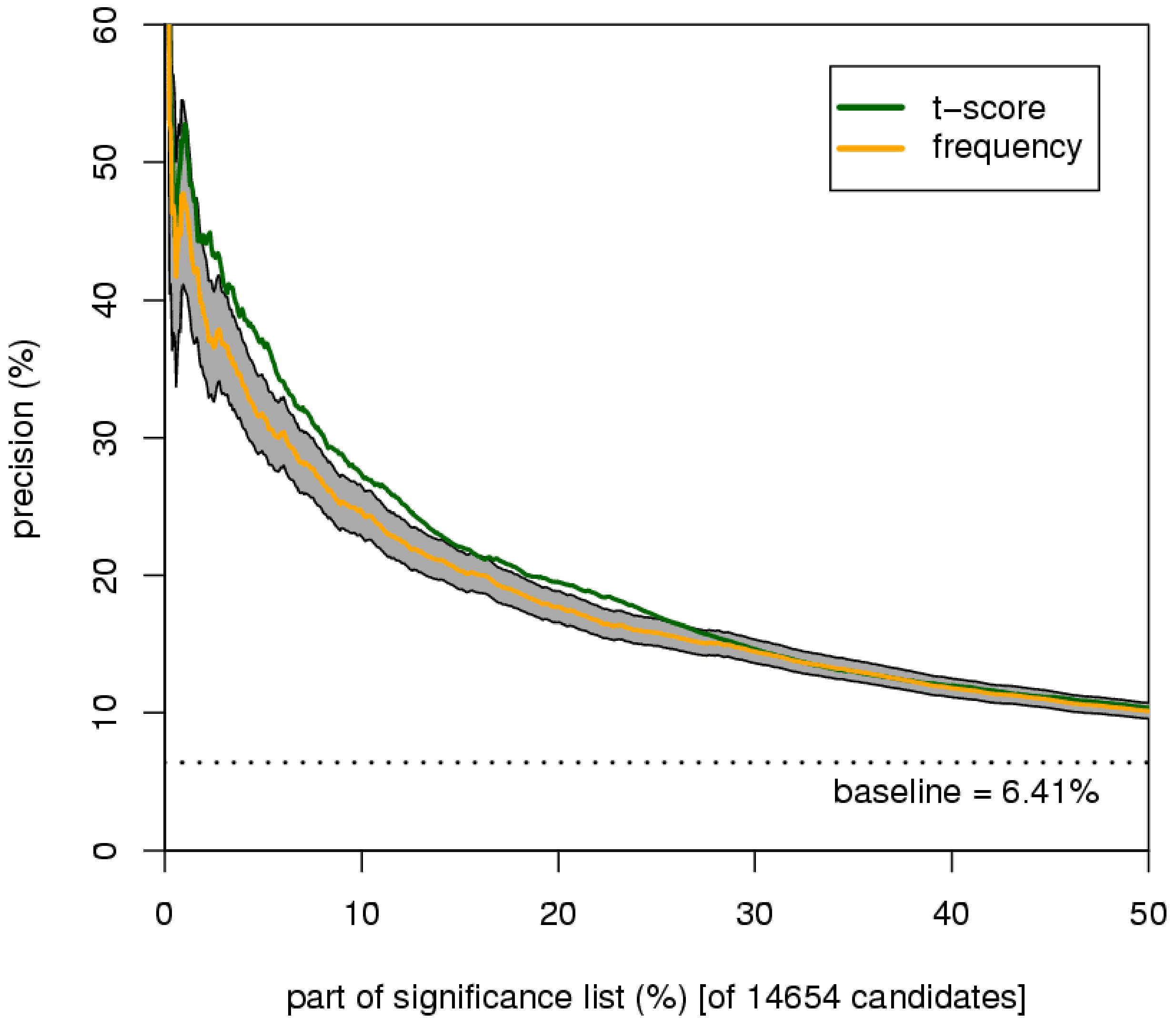
Multiple comparisons

- we *do* have multiple comparisons, but the results are highly correlated (because they are parts of the same rankings)
- it is unclear, if and how to correct for multiple comparisons ($\rightarrow$ Directions)
- no problem for pre-defined *N*-best list
- single χ^2 test for "interesting difference" $\rightarrow$ multiple comparisons by eye

Confidence intervals

- instead of markers for significant differences, we can also display confidence intervals around precision graphs
- confidence intervals mark differences that can be explained by random effects
- ranges are obtained from binomial test (at 95% confidence level) and differ slightly from results of χ^2 test

Confidence interval graphs



Pearson's χ^2 test: the details

- **theoretical model**: association measures A and B choose candidates for each *N*-best list independently;
  measure A selects TP with probability p_A
  measure B selects TP with probability p_B
- **null hypothesis** H_0: $p_A = p_B$
- *N*-best lists are assumed to represent independent samples

Pearson's χ^2 test: problems

- χ^2 test assumes that p_A and p_B are constant for the N highest-ranking candidates
  $\rightarrow$ not consistent with precision graphs
- N-best lists are in truth related samples (re-rankings of the same candidate set)
- intuitively: AMs have fewer "opportunities" to make different choices than predicted
  $\rightarrow$ **strong** test (probably too strong)
- more appropriate: paired classification test

McNemar's test

tbl	– t-score	+ t-score
– freq	610	46
+ freq	7	276

+ = in 1000-best list **–** = not in 1000-best list

- ideally: all TPs in 1000-best list (possible!)
- H_0: differences between AMs are random

McNemar's test

tbl	– t-score	+ t-score
– freq	610	46
+ freq	7	276

+ = in 1000-best list **–** = not in 1000-best list

```
> mcnemar.test(tbl)
```

- p-value < 0.001 $\rightarrow$ highly significant

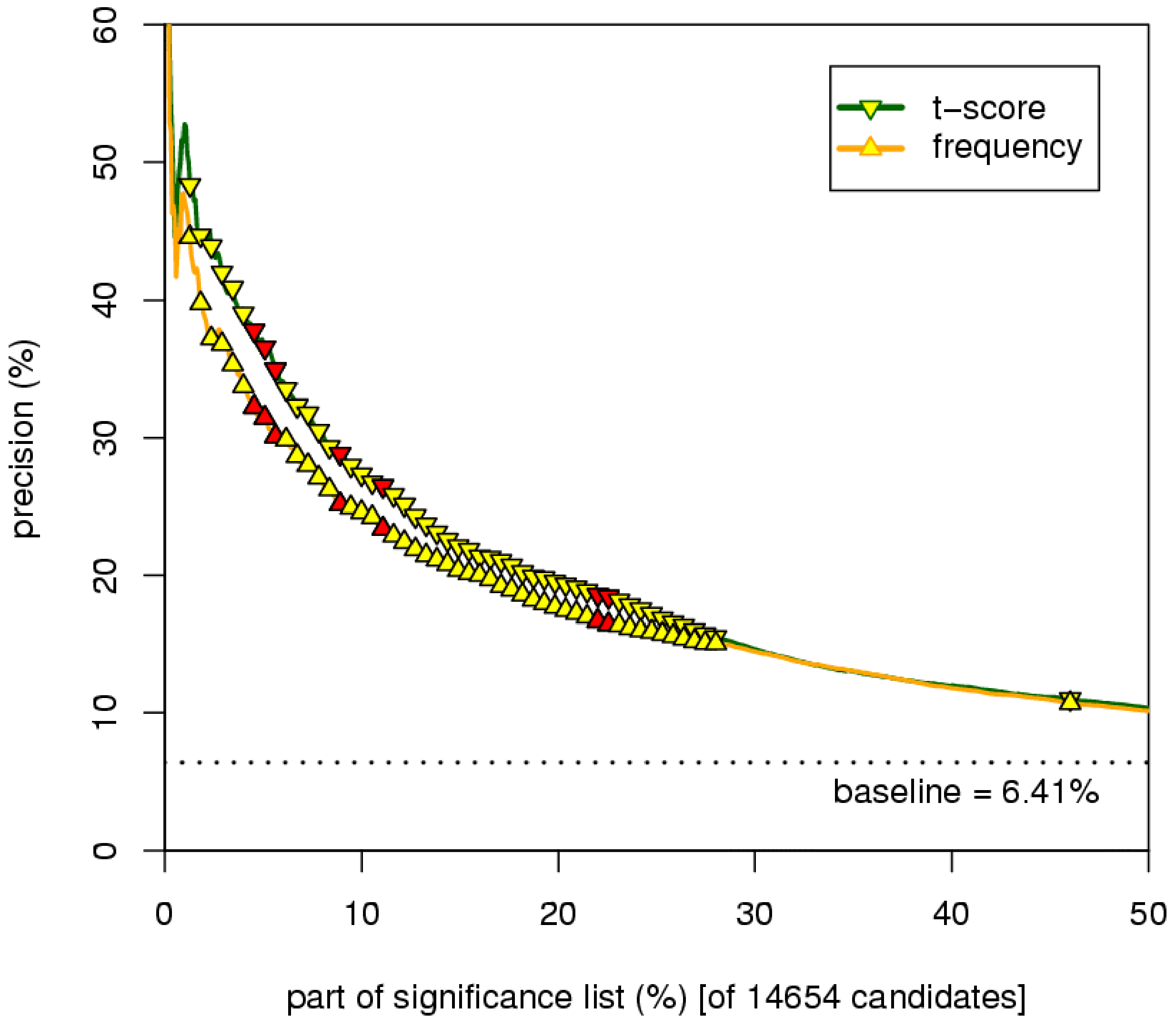
McNemar's test: discussion

- McNemar's test only considers data where the two association measures differ
- null hypotheses: when A and B differ, A is just as likely as B to make the right decision
- no further assumptions about classification
  $\longrightarrow$ seems to be the most appropriate ST
- McNemar uses normal approximation; substitute binomial distribution for exact test

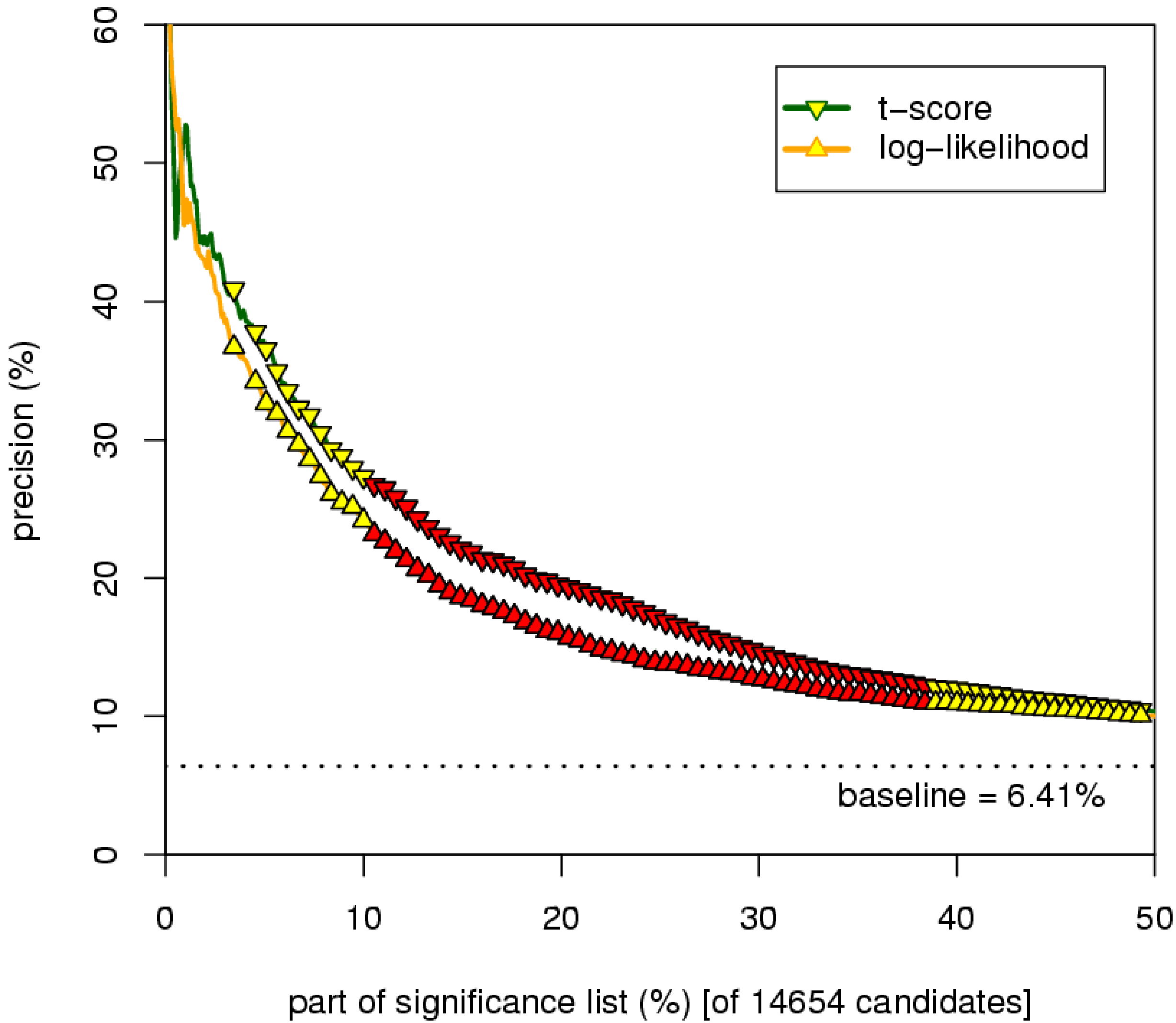
McNemar's test: discussion

- McNemar's test might consider differences for a few exotic cases significant
- even if the AMs perform equally badly for the candidates common to both *N*-best lists
- hence, McNemar is likely to overestimate differences and is a very **weak** test
- idea: use McNemar as **lower threshold**, △ and χ^2 test as **upper threshold** ▲

Precision graphs with both tests



Precision graphs with both tests



Interpretation of the combined test

- **lower threshold** (McNemar):
  when the AMs differ, A is systematically better than B
- **upper threshold** (χ^2 test):
  A *always* makes systematically better choices than B, rather than agreeing with B's mistakes
- McNemar only considers choices for TPs, whereas χ^2 test considers *all* choices

Local vs. global tests

- *still* have problem of multiple comparisons
- esp. McNemar's test has high risk of type I error for two very similar AMs
- multiple comparisons are a problem for all local tests, based on single *N*-best lists
- try global tests, which compare full rankings (ranking of TPs for test of performance)
- STs for related ordinal data (ranking tests)

Rank correlation coefficients

- e.g. Spearman's rank correlation or Kendall rank correlation (Siegel, 1956)
- test whether high-ranking TPs from AM A are also assigned high ranks by B
- problem: if measure A ranks TPs much higher than measure B, but in the same order, tests report a strong correlation
- not useful for our purposes

Walsh sign test (cf. Siegel, 1956)

- answers the question whether measure A systematically ranks TPs higher than B
- ranks for each TP are compared: **+** / **–**
- H_0: **+** and **–** occur equally often
- problem: imagine measures A and B, where A puts each TP exactly one rank higher than B
- only **+**'s $\rightarrow$ A considered significantly better
- but there will be no difference in performance

Wilcoxon signed ranks test (Siegel)

- considers size of differences between ranking
- paired test: compares rankings for each TP
- H_0: the (absolute values of) positive differences (A ranks higher than B) are on average as large as those of negative differences (B ranks higher than A)
- this corresponds to our intuition that a large difference in ranks is more important

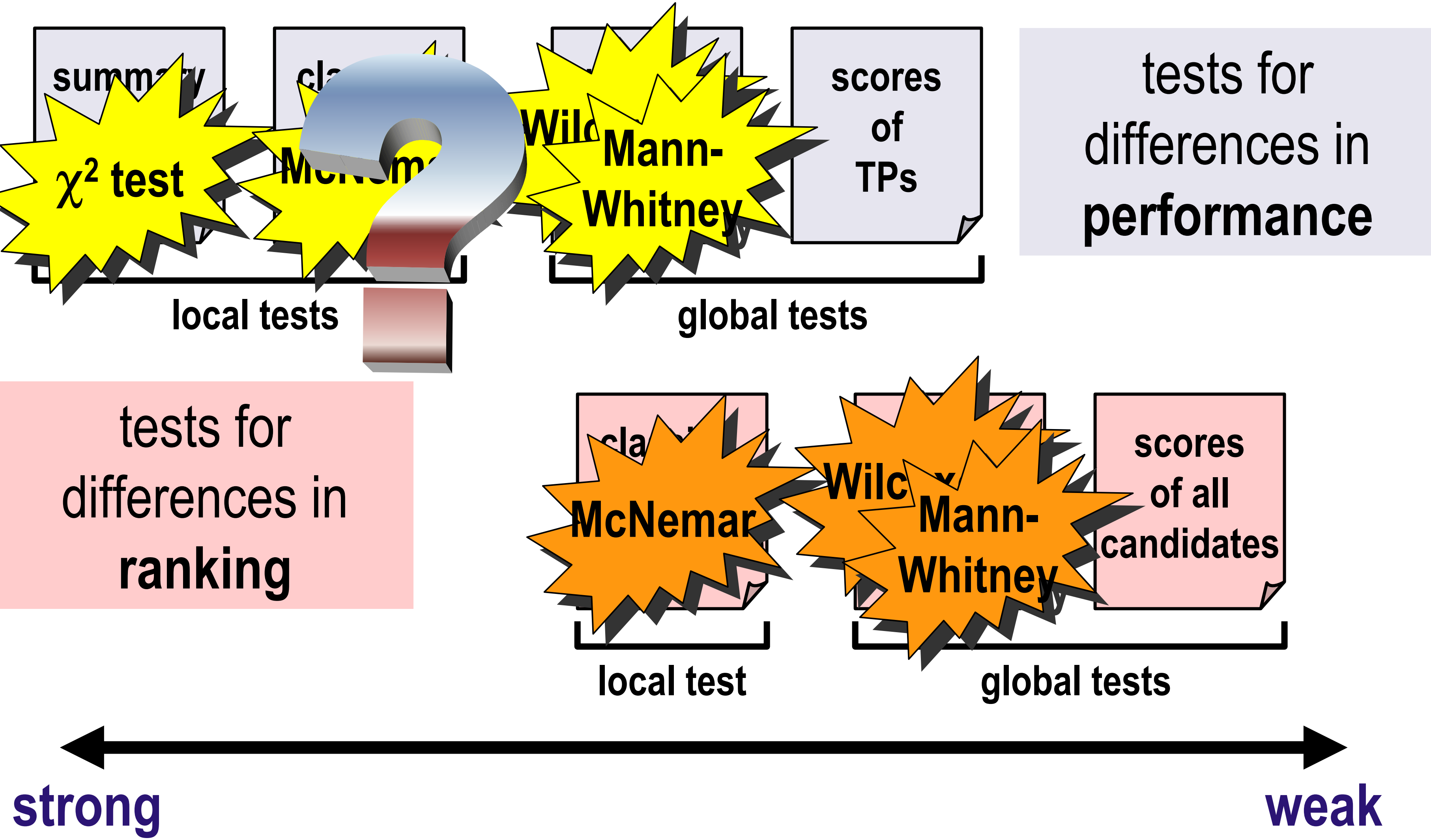
Wilcoxon signed ranks test (Siegel)

- note that Wilcoxon test is based on ranking of (the absolute values of) rank differences
- need two parallel lists (vectors) giving the ranks of all TPs according to A and B, e.g. `rank.TP.tscore` and `rank.TP.freq`
- ```
  > wilcox.test(rank.TP.tscore,
                rank.TP.freq, paired=TRUE)
  ```
- p-value < 0.001 $\rightarrow$ highly significant
- should we compare actual rank differences?
  ```

Mann-Whitney test (Siegel, 1956)

- similar to Wilcoxon, but tests whether measures A and B rank TPs equally high on average (for unrelated samples)
- can be computed with same R function
- ```
  > wilcox.test(rank.TP.tscore,
                rank.TP.freq)
  ```
- p-value < 0.022 $\rightarrow$ not significant
- seems to be stronger than Wilcoxon test, less sensitive to small systematic rank differences
  ```

Summary


χ^2 test
Mann-Whitney
scores of TPs
local tests
global tests
tests for differences in **performance**
tests for differences in **ranking**
McNemar
Mann-Whitney
scores of all candidates
local test
global tests
strong
weak

Directions for the future

- fill the gap between χ^2 test and McNemar (generally: related and unrelated samples)
- the problem of multiple comparisons
- significance and relevance
- What is the question?

Filling the gap

- STs for unrelated samples seem too strong, tests for related samples seem too weak
- goal: ST that estimates how often A could (& should) have made a better choice than B
- or are we mixing confidence & relevance?
- perhaps answer two questions separately:

  How many differences are there between A and B

  Is A systematically better on these differences?

Multiple comparisons, again

- correcting for multiple comparisons is an open question for local STs
- global STs do not have this problem, but:
- status of global tests is not entirely clear (i.e. their strength compared to local tests)
- practical problem: global tests require manual annotation of entire candidate set
- there are still multiple comparisons in a pairwise evaluation of *k* AMs

Significance and relevance

- combine significance and relevance for a practical evaluation of AMs
- null hypothesis: no relevant difference
- alternative: there *is* a relevant difference, e.g. measure A is at least 50% better than B
- need to define what "50% better" means
- can we apply non-parametric tests?
- similar to estimation of confidence intervals

What is the question?

- all STs we have considered test the null hypothesis, that differences are due to chance
- STs differ in what "random differences" are
- make H0 more explicit (esp. for Wilcoxon and Mann-Whitney test) and compare to intuition
- our intuitive question was:
  "If we repeated the experiment, would measure A again perform better than B?"

What is "repeat the experiment" ?

- which parameters may change?
  - type of collocation & precise definition
  - domain & text type
  - pre-processing & extraction methods
  - text source (e.g. newspaper vs. newsgroups)
  - size of source corpus
  - different segement of same source corpus
- can we obtain empirical results?