Adjusting p-values for heterogeneity in collocation analysis

Jefrey Lijffijt, University of Bristol
Tanja Säily, University of Helsinki

University of BRISTOL

UNIVERSITY OF HELSINKI

Why collocations

- “You shall know a word by the company it keeps!” (Firth 1962 [1957]: 11)

- **Language description**, lexicography, language learning, distributional **semantics**, NLP (Evert 2005: 22–27)…

- Collocations & key words: staples of statistical analysis in corpus linguistics

How to find collocations

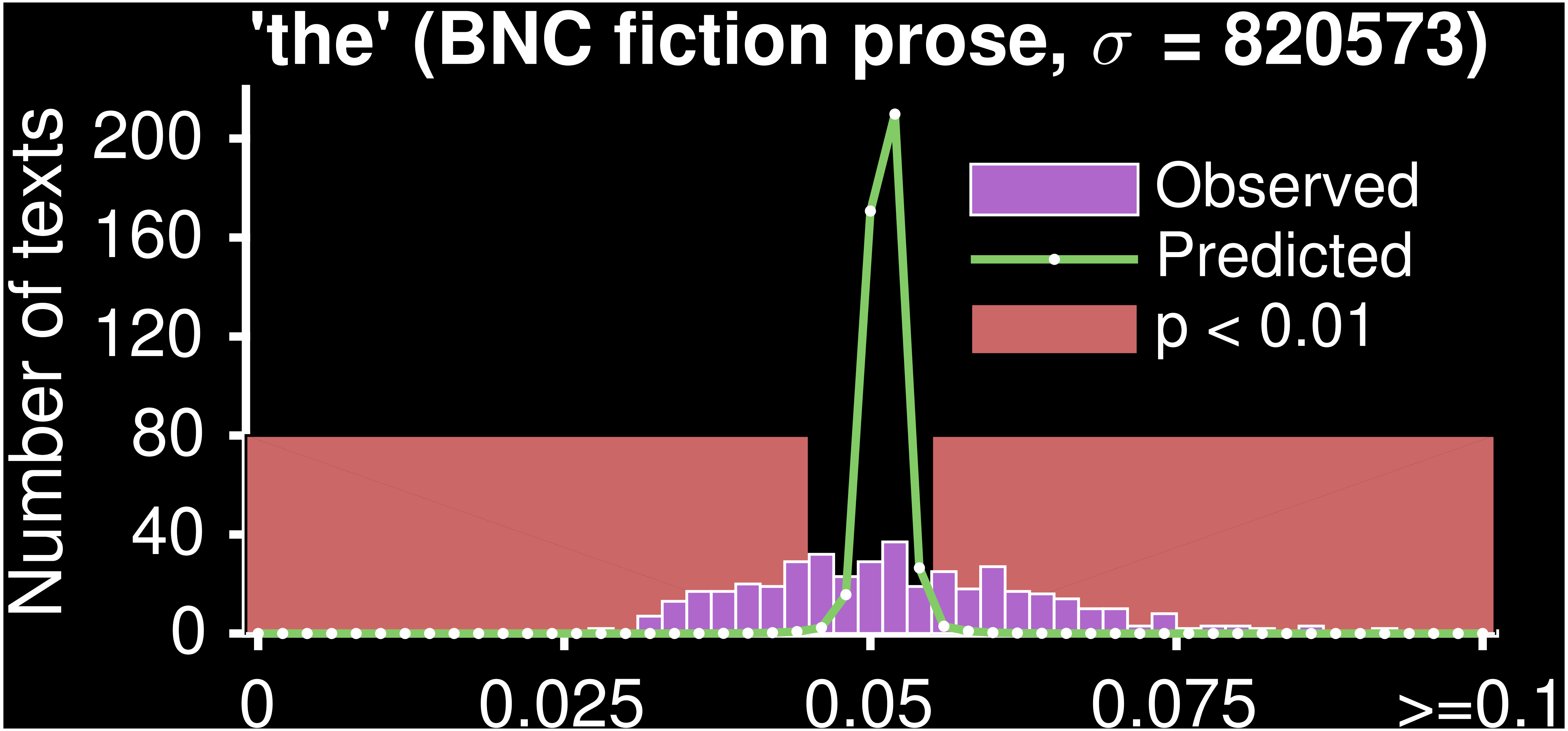
- Statistical test
    - chi-square, t-test, Fisher, …
- Association measure
    - Mutual information, conditional probability, ΔP, …
- We refer to any such quantity as a *statistic*

Burstiness / dispersion

- Rare words occur in very few texts
- Frequent words also have poor dispersion
    - Their frequency variation is not as expected
    - The occurrences are related

  –> Occurrences are statistically dependent

Dispersion for single words



Dispersion for single words

- If we want to compare word frequencies across (sub-)corpora, we can account for dispersion
    - Data representation matters
        - Report frequencies per text, not per corpus
    - Use a suitable test (t-test, Wilcoxon rank-sum…) [our advice, see Lijffijt et al. forthcoming]

Dispersion & collocations

- Dispersion matters a lot for words
- How about finding collocations?
- Is the problem less severe or worse?

Dispersion & collocations

- We derive collocation statistics from a 2x2 table
- For example, whether B occurs after A
  - (Holds for any position)
- B after A may be frequent in a corpus, but occur in only one text
- Is it a collocation then?

	A occurs	A does not occur
B occurs	X_1	X_3
B does not occur	X_2	X_4

What now?

- Is it possible to account for poor dispersion?
- Not if we apply log-likelihood ratio or MI directly
  - The word counts should not be pooled
  - No appropriate test exists
- One idea for fixing this
  - Bootstrapping

Bootstrapping

- Resample the corpus
    - E.g., if the corpus has 100 texts
    - Generate 100 numbers between 1 and 100
    - Texts with these indices form a ‘random’ corpus
        - There will be duplicates and some exclusions
- Compute the statistic every time to get a confidence interval

Bootstrapping

- Get a confidence interval from the random corpora
- Problem
  - Instead of 1 statistic per collocate, we get very many
  - And we have loads of collocates to look at
  - What now?

p^2 (p-squared)

- We define p^2 as

  the smallest value $p = \gamma$ obtained with probability $1-\gamma$

- For example, $p^2 \leq 0.01$ if there is ≥ 99% probability (under resampling) that $p \leq 0.01$

- Like h-index

Estimation algorithm

```
p_in = sort(p_values,’descend’)
n = length(p_in)
i = 1
while (i <= n && p_in(i) >= i/n) {
      i = i + 1
}
if (i > 1) {
      p2_index = min(p_in(i - 1), i/n)
} else {
      p2_index = i/n
}
```

it is neither algorithmically complicated, nor difficult to compute

1/100: $p = 0.179$: fine, **$p^2 \leq 0.179$** $(0.01 \leq 0.179)$

2/100: $p = 0.155$: fine, **$p^2 \leq 0.155$** $(0.02 \leq 0.155)$

3/100: $p = 0.152$: fine, **$p^2 \leq 0.152$** $(0.03 \leq 0.152)$

4/100: $p = 0.104$: fine, **$p^2 \leq 0.104$** $(0.04 \leq 0.104)$

5/100: $p = 0.096$: fine, **$p^2 \leq 0.096$** $(0.05 \leq 0.096)$

6/100: $p = 0.086$: fine, **$p^2 \leq 0.086$** $(0.06 \leq 0.086)$

7/100: $p = 0.078$: fine, **$p^2 \leq 0.078$** $(0.07 \leq 0.078)$

8/100: $p = 0.075$: stop, **$p^2 = 0.078$** $(0.08 > 0.075)$

9/100: $p = 0.042$

10/100: $p = 0.033$

11/100: $p = 0.031$

...

98/100: $p = 0.000$

99/100: $p = 0.000$

100/100: $p = 0.000$

p^2 (p-squared)

- We define p^2 as

  the smallest value $p = \gamma$ obtained with probability $1-\gamma$

- Some nice properties
  - Single statistic, easy to read and can be used to rank
  - If the null is true, p^2 → 0.5 if the data size grows
    - Unlike p, which is always uniformly random on [0, 1]

Case study: *teacher(s)*

- BNC, demographically sampled spoken section, 417 hits
- Frequency of node+collocate ≥ 5, window 5L + 5R:
  116 collocate candidates
- Log-likelihood ratio test: 75 significant left-hand collocates, 61 right ($p \leq 0.01$)
  - p^2: only 14 left, 12 right —> less noise
  - e.g. *now* (right): $p = 0.0002$, $p^2 = 0.1163$
    - Occurs 7 times in 5 different texts

1. KB7 348 We didn't take a lot [pause] I mean she was a history **teacher** so **now** you know why I didn't learn a lot of history cos all we did was giggle.

2. KCA 2723 They changed all the **teachers** round **now** because

3. KCA 2734 So they moved the **teachers** all round **now**.

4. KCS 770 they're blaming er parents are blaming school **teachers** about the kids, **now** where I live kids are running around up to eleven o'clock at night sometimes, it's not the teachers to blame it's the parents

5. KCS 1658 Oh she has enough certificates to of gone to **teachers**' training college, **now** that, I always feel although I think she's quite happy now, but for myself, for myself and I'm always er tempted by the fact that they always have twelve weeks' holiday you know, I mean in one go the teachers

6. KDW 7663 And erm [pause] now he's a supply **teacher** in [pause] **now** he's got a band or something [unclear] , I dunno.

7. KPY 158 er, that'll teach her, see she's, **teacher**'s name Sarah **now** and my names [unclear]

Significant collocates (p^2)

- Left:
  - *school*, *the*, *to*, *’s*, *of*, *a*, *your*, *and*, *our*, *she*, *my*, *by*, *one*, *teachers*
- Right:
  - *school*, *the*, *and*, *she*, *that*, *to*, *’s*, *are*, *he*, *at*, *in*, *you*

(blue = both)

Conclusion

- p^2:
  - Accounts for dispersion, reduces number of false positives
  - Requires bootstrapping as a preceding step
  - Simple algorithm, easy to read statistic, can be used to rank

References

- Evert, Stefan. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Institut für maschinelle Sprachverarbeitung, University of Stuttgart PhD dissertation. http://nbn-resolving.de/urn:nbn:de:bsz:93-opus-23714.

- Firth, J. R. 1962 [1957]. A synopsis of linguistic theory 1930–1955. *Studies in linguistic analysis*, 1–32. Oxford: Basil Blackwell.

- Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki & Heikki Mannila. Forthcoming. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqu064