Aalto University
School of Science

UNIVERSITY OF HELSINKI

DAMMOC
DATA MINING TOOLS
for changing modalities of communication

Chi-square test considered harmful: Better methods for testing the significance of word frequencies

Jefrey Lijffijt *, *Tanja Säily* **, Terttu Nevalainen **

* Department of Information and Computer Science, Aalto University

** Department of Modern Languages, University of Helsinki

Comparing frequencies across corpora

- Traditional approach: create cross-table of frequencies

Word	Freq in *S*	Freq in *T*
I	2,805	2,445
Total	162,000	162,000

- **Is this statistically significant?**

- $p_{\textit{Log-likelihood ratio test}} = 0.000000541$

  → Significant overuse in corpus *S*

Bag-of-words model (log-likelihood ratio test, χ^2-test, Fisher's exact test, binomial test)

- **Assume all words are independent**

- Easy to use (2x2 table)
- Mathematically simple

Word	Freq in *S*	Freq in *T*
I	2,805	2,445
Total	162,000	162,000

- *However:* texts have structure!

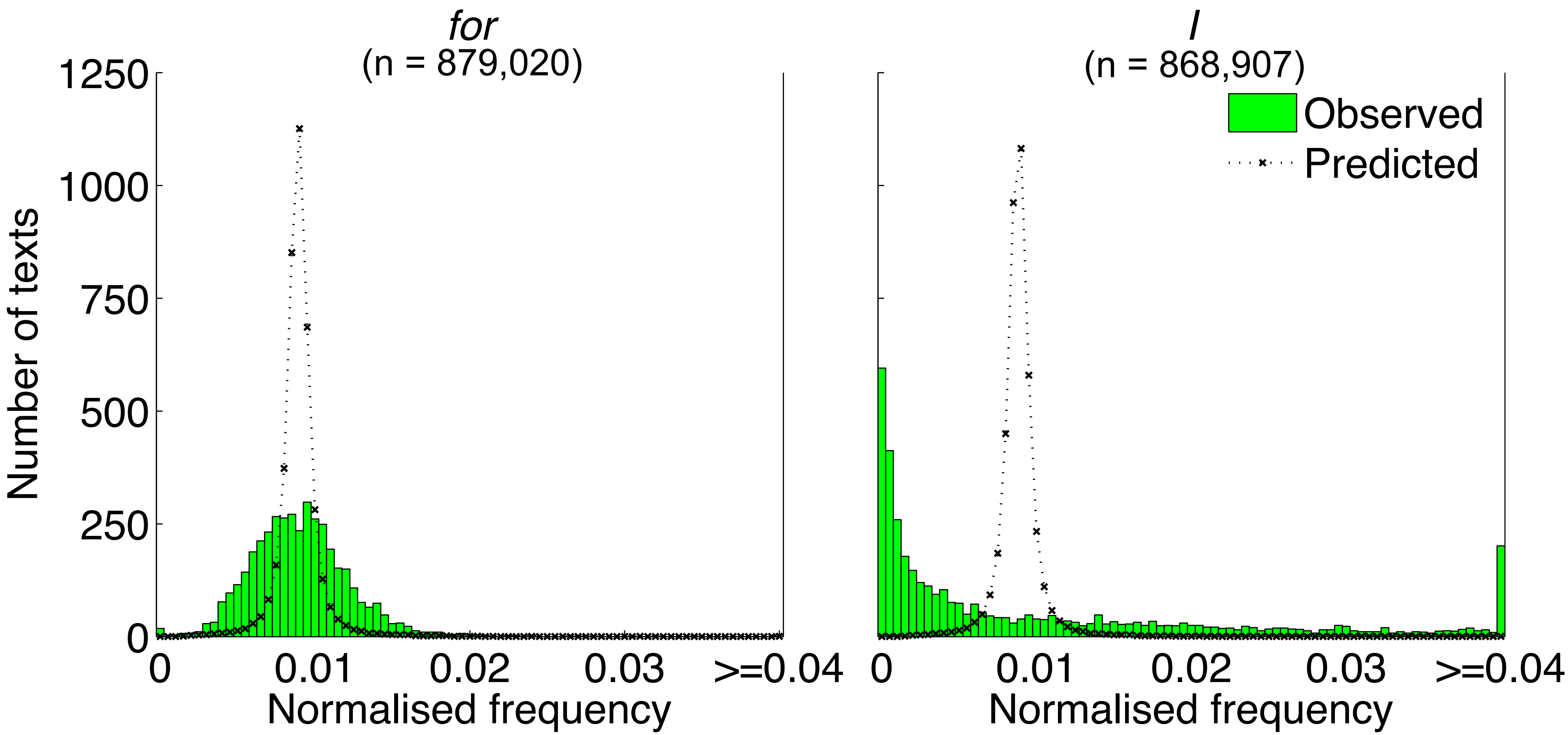
- Core questions:
  - Can we provide more realistic models?
  - Does it matter *when comparing corpora*?

Previous critiques [of bag-of-words based tests]

- Too many results, bad assumptions (Kilgarriff 2001)
- Arbitrary results, null hypothesis is false (Kilgarriff 2005)
- Unit of sampling ≠ unit of measurement (Evert 2006)
- Too many results ← burstiness of words (Lijffijt et al. 2011)

Bag-of-words model makes poor predictions

Data: British National Corpus, 4049 texts



There exist other tests (Bootstrap test, Wilcoxon rank-sum test)

- Cross-table of frequencies

Word	Freq in *S*	Freq in *T*
I	2,805	2,445
Total	162,000	162,000

- $p_{\textit{Log-likelihood ratio test}} = 0.000000541$
- $p_{\textit{Bootstrap test}} = 0.280$

- High *p* → maybe not so significant after all !

Bootstrap test (Lijffijt et al. *forthcoming*)

- Produce *N* random corpora using resampling
  - $S_1, \ldots, S_N$ and $T_1, \ldots, T_N$
  - P-value based on comparing random samples
- $$p_1 = \frac{\sum_{i=1}^{N} H(freq(q,S_i) \leq freq(q,T_i))}{N}, \quad H(x \leq y) = \begin{cases} 1, & x < y \\ 0.5, & x = y \\ 0, & x > y \end{cases}$$
- $$p_2 = \frac{1 + N \cdot 2 \cdot \min(p_1, 1 - p_1)}{1 + N}$$

Some new experiments (Lijffijt et al. *forthcoming*)

- Experimental set-up:

1. Use a reasonably homogeneous corpus
2. Pick a word with sufficient frequency (≥ 50)
3. Assign half of the texts to *S* and the other half to *T*
4. Compute p-value

- Repeat 3 & 4 many times

- **The resulting p-values should be uniform in [0,1]**

Some new experiments (Lijffijt et al. *forthcoming*)

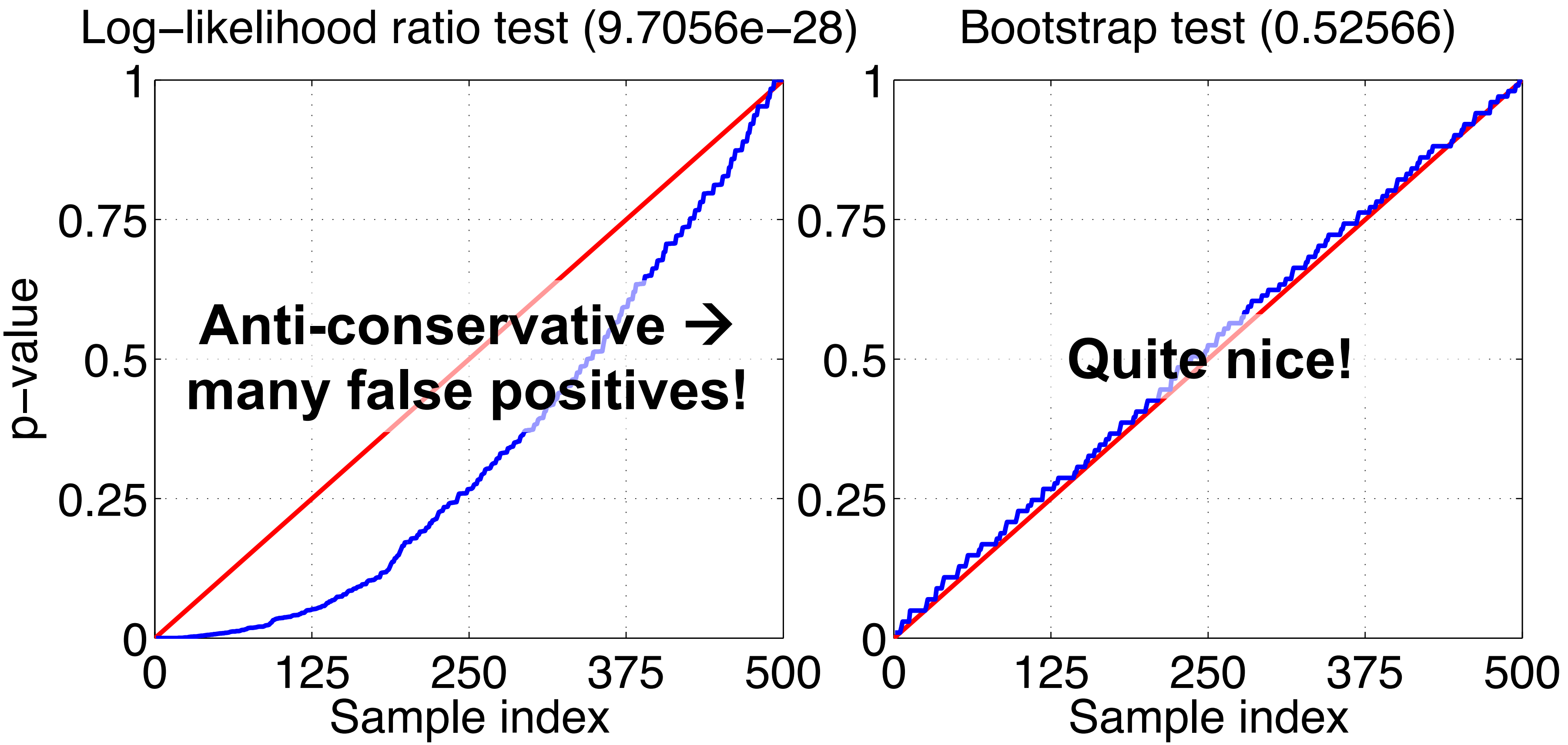
- **The resulting p-values should be uniform in [0,1]**

- We can test this uniformity using a statistical test
  - Kolmogorov-Smirnov test

- If p-values too high → test is *conservative* (low power)
  - Results in many false negatives

- If p-values too low → test is *anti-conservative*
  - Results in many false positives

Experimental result for *would (n = 2590)*


Log–likelihood ratio test (9.7056e–28)

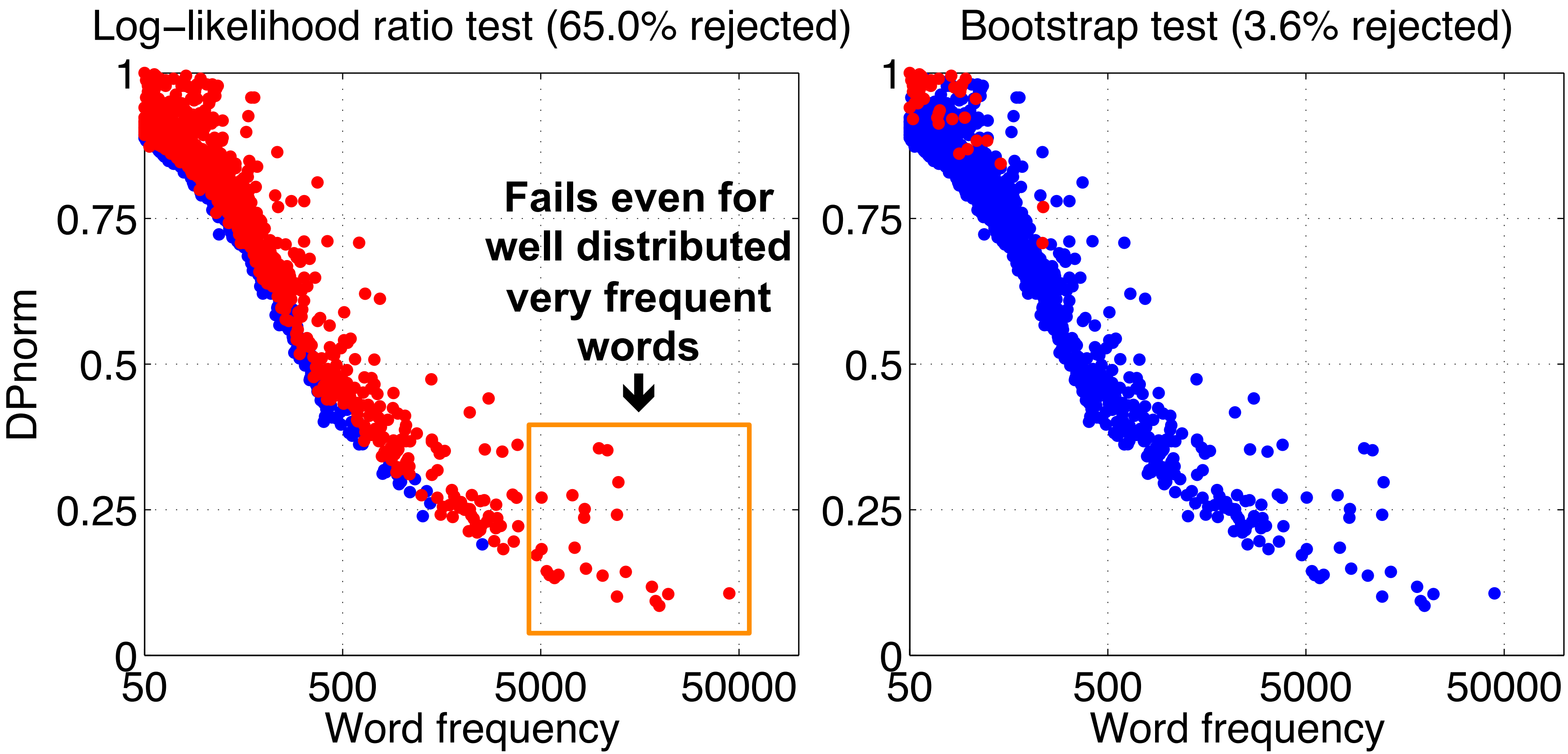
Bootstrap test (0.52566)

Anti-conservative → many false positives!

Quite nice!

Data: British National Corpus, fiction prose subcorpus, 405 texts

We did this for all words (freq ≥ 50)



Log–likelihood ratio test (65.0% rejected)

Bootstrap test (3.6% rejected)

Fails even for well distributed very frequent words

Data: British National Corpus, fiction prose subcorpus, 405 texts

Case study on gender variation

- Are there **differences between male and female writing** in our material in terms of word frequencies?
  - Cf. Lijffijt et al. (forthcoming)

- Do these differences depend on the **audience** at which the writing is aimed?
  - Bell (1984)

- Both bootstrap and log-likelihood ratio (LL) tests used
  - Significance threshold 0.05; FDR control → 0.0004454

Material

- ***British National Corpus***, prose fiction genre (Lee 2001)
- 2,000-word samples, equal number of texts (81) and words (162,000) for each subcorpus:
  - **Women** writing for **any** audience
    - male, *female*, mixed-gender, unknown
  - **Women** writing for a **mixed-gender** audience
  - **Men** writing for **any** audience
    - *male*, female, mixed-gender, unknown
  - **Men** writing for a **mixed-gender** audience
- Words lowercased, tagging and punctuation ignored

Words overused by WOMEN (both bootstrap and LL tests)

ANY AUDIENCE

Word	$\text{Freq}_{\text{Male}}$	$\text{Freq}_{\text{Female}}$
be	623	810
her	1,239	2,566
herself	50	164
male	0	17
she	1,378	2,884

MIXED-GENDER AUDIENCE

Word	$\text{Freq}_{\text{Male}}$	$\text{Freq}_{\text{Female}}$
blouse	0	9
cow	0	10
families	0	12
her	1,077	2,119
herself	45	131
she	1,398	2,367
sheets	0	9

Words overused by MEN (both bootstrap and LL tests)

ANY AUDIENCE

Word	$\text{Freq}_{\text{Male}}$	$\text{Freq}_{\text{Female}}$
calls	17	2
frank	19	0
funny	22	4
knows	42	11
military	10	0
policeman	31	0
wheel	14	0

MIXED-GENDER AUDIENCE

Word	$\text{Freq}_{\text{Male}}$	$\text{Freq}_{\text{Female}}$
below	38	7
sin	10	0
slowly	56	21

Log-likelihood ratio test: Misleading results

- Words under analysis: significant according to LL but bootstrap p-value > 0.05, most frequent first
- Overuse by **women**
  - Mostly proper nouns:
    tom, *jack*, *henry*, *sam*, *helen*, … (mixed-gender audience)
  - Many are poorly dispersed = high $\mathrm{DP}_{\mathrm{norm}}$ (Lijffijt & Gries 2012, Gries 2008), which could be used to prune the results
  - But some with a relatively low $\mathrm{DP}_{\mathrm{norm}}$:
    rose, *meeting*, *rain* (any audience)
    → Difficult to explain; no coherent semantic set

Log-likelihood ratio test: Misleading results

- Overuse by **men**
  - *I*, *my* (both any & mixed-gender audience)
    → Contradicts previous research: women expected to use more (e.g. Argamon et al. 2003, Rayson et al. 1997)
  - *car*, *boy*, *mrs*, *island* (any audience)
    → Could be (wrongly) seen as audience/genre markers
  - *john*, *says*, *wrote*, *dogs* (mixed-gender audience)
    → E.g. verb use could seem interesting
  - Also many infrequent and/or poorly dispersed proper nouns

Discussion

- **Male and female writing** do differ from each other in our material in terms of word frequencies
  - Most conspicuous difference: women's overuse of feminine personal pronouns (independent of audience)
- There are also **audience-related** key words
  - Female-to-female writing: *be*, *male*
  - Male-to-male(?) writing: *knows*, *funny*, …
- The **log-likelihood ratio test** yields 30–50 times as many significant results as the bootstrap test
  - Many of these are poorly dispersed
  - Some could be (mis)taken as linguistically interesting

Conclusion

Presentation available at
http://users.ics.aalto.fi/lijffijt/

- Bag-of-words tests harmful for key word analysis
  - Assume word-level independence
    → Too optimistic, lots of work to prune manually
  - Not always easy to tell which results are genuinely significant
- We recommend the **bootstrap test**
  - Assumes text-level independence
    → More reasonable, fewer results to wade through
  - Performs better than other such tests (Lijffijt et al. forthcoming)

  ! Statistically significant ≠ linguistically interesting
- Software developers: please incorporate bootstrapping!
  - Already available in R

References (1/2)

- Argamon, S., M. Koppel, J. Fine & A.R. Shimoni. 2003. “Gender, genre, and writing style in formal written texts”. *Text* 23(3), 321–346.
- Bell, A. 1984. “Language style as audience design”. *Language in Society* 13, 145-204.
- *The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.
- Evert, S. 2006. “How random is a corpus? The library metaphor”. *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177–190.
- Gries, S.Th. 2008. “Dispersions and adjusted frequencies in corpora”. *International Journal of Corpus Linguistics* 13(4), 403–437.
- Kilgarriff, A. 2001. “Comparing corpora”. *International Journal of Corpus Linguistics* 6(1), 97–133.
- Kilgarriff, A. 2005. “Language is never, ever, ever, random”. *Corpus Linguistics and Linguistic Theory* 1(2), 263–276.

References (2/2)

- Lee, D.Y.W. 2001. “Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle”. *Language Learning & Technology* 5(3), 37–72.
- Lijffijt, J. & S.Th. Gries. 2012. “Correction to ‘Dispersions and adjusted frequencies in corpora’”. *International Journal of Corpus Linguistics* 17(1), 147–149.
- Lijffijt, J., T. Nevalainen, T. Säily, P. Papapetrou, K. Puolamäki & H. Mannila. Forthcoming. “Significance testing of word frequencies in corpora”. Submitted to *International Journal of Corpus Linguistics*.
- Lijffijt, J., P. Papapetrou, K. Puolamäki & H. Mannila. 2011. “Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping”. In *Proceedings of ECML-PKDD 2011*, 341–357. Berlin: Springer.
- Rayson, P., G. Leech & M. Hodges. 1997. “Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus”. *International Journal of Corpus Linguistics* 2(1), 133–152.

Examples: Female writing

- As **she** walked into his cabin, **she** could smell the faint elusive fragrance that was uniquely his, a blend of soap, shower gel, and the heady musk of clean warm **male**.

(H7W 1756; female to female)

- I should like Alida, **she** thought, I should **be** kind to **her** — I will **be** kind to **her**.

(AD1 506; female to female)

- **She** knew them all; **she** was devastated for them and their **families**, who would **be** left husbandless and fatherless.

(AEA 19; female to mixed)

Examples: Male writing

- Certainly the Pentagon **knows** it’s already under investigation, but Hawkins didn’t want anyone to know that he was pointing fingers in certain directions.

  (CKC 3394; male to male)

- The **funny** thing is, he’s not very chatty or friendly; people say he’s a very shy man.

  (HWP 2341; male to unknown)

- He smiled tightly and waved a hand at the **slowly** diminishing figure on the hillside far **below**.

  (GUG 390; male to mixed)