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

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq in S</th>
<th>Freq in T</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2,805</td>
<td>2,445</td>
</tr>
<tr>
<td>Total</td>
<td>162,000</td>
<td>162,000</td>
</tr>
</tbody>
</table>

• Is this statistically significant?

• \( p_{\text{Log-likelihood ratio test}} = 0.000000541 \)

\( \rightarrow \) Significant overuse in corpus \( S \)
Bag-of-words model (log-likelihood ratio test, $\chi^2$-test, Fisher’s exact test, binomial test)

- Assume all words are independent
- Easy to use (2x2 table)
- Mathematically simple

However: texts have structure!

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq in S</th>
<th>Freq in T</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2,805</td>
<td>2,445</td>
</tr>
<tr>
<td>Total</td>
<td>162,000</td>
<td>162,000</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq in S</th>
<th>Freq in T</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2,805</td>
<td>2,445</td>
</tr>
<tr>
<td>Total</td>
<td>162,000</td>
<td>162,000</td>
</tr>
</tbody>
</table>

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

- High $p \rightarrow$ 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

\[
\begin{align*}
p_1 &= \frac{\sum_{i=1}^{N} H(freq(q,S_i) \leq freq(q,T_i))}{N}, \quad H(x \leq y) = \begin{cases} 
1, & x < y \\
0.5, & x = y \\
0, & x > y 
\end{cases} \\
p_2 &= \frac{1 + N \cdot 2 \cdot \min(p_1, 1 - p_1)}{1 + N}
\end{align*}
\]
Some new experiments (Lijffijt et al. forthcoming)

• Experimental set-up:

1. Use a reasonably homogeneous corpus
2. Pick a word with sufficient frequency ($\geq 50$)
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 $\rightarrow$ test is conservative (low power)
  – Results in many false negatives

• If p-values too low $\rightarrow$ test is anti-conservative
  – Results in many false positives
Experimental result for *would* \((n = 2590)\)

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

Log-likelihood ratio test \((9.7056 \cdot 10^{-28})\)

Bootstrap test \((0.52566)\)

Anti-conservative \(\rightarrow\) many false positives!

Quite nice!

Data: British National Corpus, fiction prose subcorpus, 405 texts
We did this for all words (freq ≥ 50)

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 \( \rightarrow 0.0004454 \)
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

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq$_{Male}$</th>
<th>Freq$_{Female}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>be</td>
<td>623</td>
<td>810</td>
</tr>
<tr>
<td>her</td>
<td>1,239</td>
<td>2,566</td>
</tr>
<tr>
<td>herself</td>
<td>50</td>
<td>164</td>
</tr>
<tr>
<td>male</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>she</td>
<td>1,378</td>
<td>2,884</td>
</tr>
</tbody>
</table>

####MIXED-GENDER AUDIENCE

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq$_{Male}$</th>
<th>Freq$_{Female}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>blouse</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>cow</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>families</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>her</td>
<td>1,077</td>
<td>2,119</td>
</tr>
<tr>
<td>herself</td>
<td>45</td>
<td>131</td>
</tr>
<tr>
<td>she</td>
<td>1,398</td>
<td>2,367</td>
</tr>
<tr>
<td>sheets</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>
Words overused by MEN (both bootstrap and LL tests)

<table>
<thead>
<tr>
<th>ANY AUDIENCE</th>
<th>Freq\text{Male}</th>
<th>Freq\text{Female}</th>
<th>MIXED-GENDER AUDIENCE</th>
<th>Freq\text{Male}</th>
<th>Freq\text{Female}</th>
</tr>
</thead>
<tbody>
<tr>
<td>call \text{\textit{calls}}</td>
<td>17</td>
<td>2</td>
<td>\textit{below}</td>
<td>38</td>
<td>7</td>
</tr>
<tr>
<td>frank</td>
<td>19</td>
<td>0</td>
<td>\textit{sin}</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>funny</td>
<td>22</td>
<td>4</td>
<td>\textit{slowly}</td>
<td>56</td>
<td>21</td>
</tr>
<tr>
<td>knows</td>
<td>42</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>military</td>
<td>10</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>policeman</td>
<td>31</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wheel</td>
<td>14</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 \( \text{DP}_{\text{norm}} \) (Lijffijt & Gries 2012, Gries 2008), which could be used to prune the results
  – But some with a relatively low \( \text{DP}_{\text{norm}} \):
    *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

- 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)


References (2/2)


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)