Study Singular “They” in Contemporary English

Bich Ngoc Do
Content

1. Introduction
2. Similar Works
3. Data Collection
4. Statistical Analysis
5. Conclusion
1. Introduction
Gender in English

- Male-oriented
  - Word: man, fireman, mailman…
  - Pronoun: Autism is complex, and each child has his own puzzle

- Gender neutralization
  - “De-Sexing the English Language” (Swift, 1972)
  - Word: fire fighter, mail carrier
  - Pronoun: he or they?
Pronoun in English

- **he vs they**
  - he: gender agreement
  - they: number disagreement
- Traditional grammarians: Epicene pronoun for third person singular
  - he, his, him, himself, his...
  - Not exist!
- Feminists protest the use of *he* in contexts which possibly involves women.
Pronoun in English

- An act of Parliament in 1850: “words importing the masculine gender shall be deemed and taken to include females”.

- Informal English:
  - Anyone who thinks *they* have been affected should contact *their* doctor.
  - One student failed *their* exam.
  - Either Mary or John should bring a schedule with *them*. 
2. Similar Work
Early research

- Mackay, 1980:
  - Used a corpus of 108 sources from scientific articles, magazine articles and textbooks...
  - The most epicene pronoun is *he*, and found no occurrences of singular *they*
Recent research

- Spoken corpora:
  - Holmes, 1998:
    - Wellington Corpus of Spoken New Zealand English (1 million words)
    - *They* is the default pronoun used in speech
  - Pauwels, 2001:
    - A part of a corpus of formal speech in Australia
    - *He* overwhelmingly dominated in the pre-reform period (1960s to late 1970s), whereas singular *they* is the most frequent epicene pronoun in the post-reform period (1990s).
Recent research

- Written corpora:
  - Baranowski, 2002
    - Two issues: The Independent (840,000 words) and San Francisco Chronicle (500,000 words)
    - *He* was no longer the preferred epicene pronoun
  - Balhorn, 2009
    - A newspaper corpus
    - *they* is used more than 60% in non-quoted texts
Limitation

- Previous work:
  - Small corpora
  - Specific genres
- My experiment:
  - Larger corpora
  - Different genres
3. Data Collection
Corpus

- Open American National Corpus (OANC)
  - 15 million words
  - Can be downloaded
  - Is tokenized and POS tagged
  - Several genres, but not equal in both size and period
<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>No. files</th>
<th>No. words</th>
</tr>
</thead>
<tbody>
<tr>
<td>charlotte</td>
<td>face to face</td>
<td>93</td>
<td>198,295</td>
</tr>
<tr>
<td>switchboard</td>
<td>telephone</td>
<td>2,307</td>
<td>3,019,477</td>
</tr>
<tr>
<td><strong>Spoken Totals</strong></td>
<td></td>
<td><strong>2,410</strong></td>
<td><strong>3,217,772</strong></td>
</tr>
<tr>
<td>Name</td>
<td>Domain</td>
<td>No. files</td>
<td>No. words</td>
</tr>
<tr>
<td>911 report</td>
<td>government, technical</td>
<td>17</td>
<td>281,093</td>
</tr>
<tr>
<td>berlitz</td>
<td>travel guides</td>
<td>179</td>
<td>1,012,496</td>
</tr>
<tr>
<td>biomed</td>
<td>technical</td>
<td>837</td>
<td>3,349,714</td>
</tr>
<tr>
<td>eggan</td>
<td>fiction</td>
<td>1</td>
<td>61,746</td>
</tr>
<tr>
<td>icic</td>
<td>letters</td>
<td>245</td>
<td>91,318</td>
</tr>
<tr>
<td>oup</td>
<td>non-fiction</td>
<td>45</td>
<td>330,524</td>
</tr>
<tr>
<td>plos</td>
<td>technical</td>
<td>252</td>
<td>409,280</td>
</tr>
<tr>
<td>slate</td>
<td>journal</td>
<td>4,531</td>
<td>4,238,808</td>
</tr>
<tr>
<td>verbatim</td>
<td>journal</td>
<td>32</td>
<td>582,384</td>
</tr>
<tr>
<td>web data</td>
<td>government</td>
<td>285</td>
<td>1,048,792</td>
</tr>
<tr>
<td><strong>Written Totals</strong></td>
<td></td>
<td><strong>6424</strong></td>
<td><strong>11,406,155</strong></td>
</tr>
<tr>
<td><strong>Corpus Totals</strong></td>
<td></td>
<td><strong>8,832</strong></td>
<td><strong>14,623,927</strong></td>
</tr>
</tbody>
</table>
Count of Epicene Pronouns

- Approximate method
- A possible candidate: a pronoun follows an *neutral gender antecedent* in 11 words or less
- Neutral gender antecedent:
  - Indefinite: *a predecessor*…
  - Definite: *the emperor*…
  - Quantifier: *every students, nobody*…
Processing Work

- Merge annotations to each document
- Import documents into a data store and index them
- Count pronouns and update result to a database
- Extract data from the database
## Data

<table>
<thead>
<tr>
<th>id</th>
<th>doc</th>
<th>genre</th>
<th>wordCount</th>
<th>year</th>
<th>theyCount</th>
<th>hesheCount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AdamsElissa.anc_00007.1.xml</td>
<td>spoken</td>
<td>1235</td>
<td>1998</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>AdamsStephanie.anc_00008.xml</td>
<td>spoken</td>
<td>966</td>
<td>1998</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>AdinolfiDavidandGail.anc_00009.xml</td>
<td>spoken</td>
<td>2573</td>
<td>1998</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ArguetaBerta-4ENG.anc_0000A.xml</td>
<td>spoken</td>
<td>4003</td>
<td>1998</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>AverittShannon.anc_00050B.xml</td>
<td>spoken</td>
<td>2477</td>
<td>1998</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>BlanchardTracy.anc_00060C.xml</td>
<td>spoken</td>
<td>1227</td>
<td>1998</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
4. Statistical Analysis
Research Questions

- Is there any differences in the distribution of singular “they” between genres?
- Is there any changes in the use of singular “they” through year?
A Quick Look at Data: Genre

Frequencies of Epicene Pronouns

- spoken
- written-fiction
- written-journal
- written-letters
- written-non-fiction
- written-technical
- written-travel guides

Observed frequencies
# A Quick Look at Data: Genre

<table>
<thead>
<tr>
<th></th>
<th>Spoken</th>
<th>Fiction</th>
<th>Journal</th>
<th>Letters</th>
<th>Non-fiction</th>
<th>Technical</th>
<th>Travel guides</th>
</tr>
</thead>
<tbody>
<tr>
<td>They</td>
<td>2387</td>
<td>3</td>
<td>835</td>
<td>14</td>
<td>40</td>
<td>323</td>
<td>45</td>
</tr>
<tr>
<td>He, she</td>
<td>1250</td>
<td>14</td>
<td>2688</td>
<td>14</td>
<td>67</td>
<td>215</td>
<td>53</td>
</tr>
</tbody>
</table>

```r
> chisq.test(ep.sum.genre)

Pearson's Chi-squared test

data:  ep.sum.genre
X-squared = 1321.473, df = 6, p-value < 2.2e-16
```
A Quick Look at Data

A lot of zero values occur in both the frequency of singular “they” and “he” or “she”!
A Quick Look at Data

- The data is hardly transformed into normal.
- It’s hard to analysis the use of singular “they” as percentage over all epicene pronouns (77% and 74% counts of “they” and “he, she” is 0 respectively).
  ⇒ “They” frequency must be analyzed separately from the rest!
Zero-Inflated Model

- Manufacturing model: To predict the number of defects on an item. However, there would be a lot of items with no defects.

- Zero-Inflated Model (ZIM):
  - Model the excess zero counts
  - Model the count values
Zero-Inflated Model

- Excess zero count $\sim$ Binomial distribution
- Count value $\sim$ Poisson distribution
Data Transformation: Frequency per Million Words

- The frequency of “they” in a document depends on its length.
- Normalize:

\[
FPM = \frac{Freq \times 1,000,000}{Length}
\]
Generalized Linear Model: Poisson Regression

Call:
glm(formula = theyFPM ~ genre, family = "poisson", data = ep)

Deviance Residuals:
   Min     1Q Median     3Q    Max
-39.36  -18.55  -18.55   -12.77  233.02

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6521804  0.0007347   9054.14 <2e-16 ***
genrewritten-fiction  -2.7009794  0.1443394    -19.27 <2e-16 ***
genrewritten-journal  -1.5038161  0.0013464   -116.90 <2e-16 ***
genrewritten-letters  -1.8681641  0.0058882    -317.27 <2e-16 ***
genrewritten-non-fiction  -1.8345906  0.0134251   -136.65 <2e-16 ***
genrewritten-technical  -2.2510391  0.0030598   -735.68 <2e-16 ***
genrewritten-travel_guides  -2.4789111  0.0093050   -266.41 <2e-16 ***
---
Signif. codes:  0 ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 9920687 on 8814 degrees of freedom
Residual deviance: 7989797 on 8808 degrees of freedom
AIC: 8007128

Number of Fisher Scoring iterations: 8
Generalized Linear Model: Negative Binomial Regression

Call:
glm.nb(formula = theyPPP ~ genre, data = ep, init.theta = 1826303.39, link = log)

Deviance Residuals:
          Min       1Q   Median       3Q      Max
-39.35    -18.55   -18.55    -12.77    232.81

Coefficients:                  Estimate Std. Error z value Pr(>|z|)
(Intercept)             6.6521804   0.0007349  9052.22  <2e-16 ***
genrewritten-fiction    -2.7809794   0.1443413  -19.27  <2e-16 ***
genrewritten-journal    -1.5038161   0.0013465  -1116.79  <2e-16 ***
genrewritten-letters    -1.8681641   0.0058885   -317.26  <2e-16 ***
genrewritten-non-fiction -1.8345906   0.0134256  -136.65  <2e-16 ***
genrewritten-technical   -2.2510391   0.0030589  -735.66  <2e-16 ***
genrewritten-travel_guides -2.4789111   0.0093052  -266.40  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1026303) family taken to be 1)

Null deviance: 9917516  on 8814  degrees of freedom
Residual deviance: 7987007  on 8808  degrees of freedom
AIC: 8004341

Number of Fisher Scoring iterations: 1
## Genre Effect: Z1

Call: 
`zeroinfl(formula = theyFM ~ genre, data = ep)`

Pearson residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.9068</td>
<td>-0.4011</td>
<td>-0.3993</td>
<td>-0.3993</td>
<td>25.0171</td>
</tr>
</tbody>
</table>

### Count model coefficients (poisson with log link):

|          | Estimate | Std. Error | z value | Pr(>|z|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 7.4473496 | 0.0073473 | 10136.43 | <2e-16 *** |
| genrewritten-fiction | -3.5762772 | 0.1443487 | -24.77 | <2e-16 *** |
| genrewritten-journal | -0.3157891 | 0.0013464 | -234.84 | <2e-16 *** |
| genrewritten-letters | 0.2729765 | 0.0058893 | 46.36 | <2e-16 *** |
| genrewritten-non-fiction | -1.7134659 | 0.0134251 | -127.63 | <2e-16 *** |
| genrewritten-technical | -1.0718371 | 0.0030888 | -350.30 | <2e-16 *** |
| genrewritten-travel_guides | -1.6702195 | 0.0098050 | -179.50 | <2e-16 *** |

### Zero-inflation model coefficients (binomial with logit link):

|          | Estimate | Std. Error | z value | Pr(>|z|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 0.19455 | 0.04109 | 4.735 | 2.19e-06 *** |
| genrewritten-fiction | -11.76076 | 324.76884 | -0.036 | 0.971 |
| genrewritten-journal | 1.64051 | 0.05945 | 27.594 | <2e-16 *** |
| genrewritten-letters | 2.68717 | 0.28795 | 9.332 | <2e-16 *** |
| genrewritten-non-fiction | 0.21086 | 0.30705 | 0.687 | 0.492 |
| genrewritten-technical | 1.63032 | 0.08778 | 18.573 | <2e-16 *** |
| genrewritten-travel_guides | 1.12456 | 0.19093 | 6.204 | 5.50e-10 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of iterations in BFGS optimization: 25
Log-likelihood: -6.433e+05 on 14 Df
**Genre Effect: Z2**

```r
zerooinfl(formula = theyPPM ~ genre, data = ep, dist = "negbin")
```

**Pearson residuals:**

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.6742</td>
<td>-0.3259</td>
<td>-0.3245</td>
<td>-0.3245</td>
<td>20.3256</td>
</tr>
</tbody>
</table>

**Count model coefficients (negbin with log link):**

|                       | Estimate | Std. Error | z value | Pr(>|z|) |
|-----------------------|----------|------------|---------|---------|
| (Intercept)           | 7.44736  | 0.02028    | 366.362 | < 2e-16 *** |
| genrewritten-fiction  | -3.87744 | 0.68194    | -5.647  | 1.55e-07 *** |
| genrewritten-journal  | -0.31577 | 0.03347    | -9.436  | < 2e-16 *** |
| genrewritten-letters  | 0.27040  | 0.18610    | 1.469   | 0.142   |
| genrewritten-non-fiction | -1.71333     | 0.15894    | -10.784 | < 2e-16 *** |
| genrewritten-technical | -1.07184      | 0.05218    | -20.542 | < 2e-16 *** |
| genrewritten-travel_guides | -1.67000      | 0.11332    | -14.737 | < 2e-16 *** |
| Log(theta)            | 0.81141   | 0.02980    | 27.138  | < 2e-16 *** |

**Zero-inflation model coefficients (binomial with logit link):**

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|---------|
| (Intercept)    | 0.1945    | 0.04109    | 4.736   | 2.18e-06 *** |
| genrewritten-fiction | -11.76086    | 324.78282  | -0.036  | 0.971   |
| genrewritten-journal | 1.64056     | 0.55445    | 27.594  | < 2e-16 *** |
| genrewritten-letters | 2.68712     | 0.28795    | 9.332   | < 2e-16 *** |
| genrewritten-non-fiction | 0.21101     | 0.30709    | 0.680   | 0.420   |
| genrewritten-technical | 1.63028     | 0.08778    | 18.573  | < 2e-16 *** |
| genrewritten-travel_guides | 1.18186     | 0.19095    | 6.205   | 5.47e-10 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Thetas = 2.3511**

Number of iterations in BFGS optimization: 32

Log-likelihood: -2.006e+04 on 15 DF
Genre Effect: Z₁ vs Z₂

- $\text{AIC}(Z₁) = 1286613.65$
- $\text{AIC}(Z₂) = 40120.18$
Genre Effect: Z₃

Call:
zeroinfl(formula = theyFPM ~ genre | 1, data = ep, dist = "negbin")

Pearson residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.4277</td>
<td>-0.4276</td>
<td>-0.4276</td>
<td>-0.4274</td>
<td>16.4089</td>
</tr>
</tbody>
</table>

Count model coefficients (negbin with log link):

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 7.44735 | 0.02029 | 366.972 | < 2e-16 *** |
| genrewritten-fiction | -3.57615 | 0.68225 | -5.242 | 1.59e-07 *** |
| genrewritten-journal | -0.31579 | 0.03347 | -9.434 | < 2e-16 *** |
| genrewritten-letters | 0.27298 | 0.18605 | 1.467 | 0.142 |
| genrewritten-non-fiction | -1.71348 | 0.15897 | -10.779 | < 2e-16 *** |
| genrewritten-technical | -1.07185 | 0.05218 | -20.543 | < 2e-16 *** |
| genrewritten-travel_guides | -1.67023 | 0.11330 | -14.741 | < 2e-16 *** |
| Log(theta) | 0.81145 | 0.02990 | 27.141 | < 2e-16 *** |

Zero-inflation model coefficients (binomial with logit link):

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 1.24614 | 0.02557 | 48.73 | < 2e-16 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.2512
Number of iterations in BFGS optimization: 33
Log-likelihood: -2.052e+04 on 9 Df
Genre Effect: Z2 vs Z3

- $\text{AIC}(Z2) = 40120.18$
- $\text{AIC}(Z3) = 41060.61$
Genre Effect: Conclusion

With $p < 0.05$:

- 78% of documents do not contain singular “they”
- Genre does have an effect on the occurrence of singular “they”: It appears a lot in spoken genre, and least in fiction genre.
Year Effect

Call:
  glm.nb(formula = theYFEM ~ year, data = cp, init.theta = 0.03067719041,
         link = log)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.7824  -0.7445  -0.7287  -0.6878   2.2926

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  383.03067   33.08994   11.57  <2e-16 ***
 year        -0.18902    0.01657  -11.41  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.0307) family taken to be 1)

    Null deviance: 3888.5  on 8228  degrees of freedom
Residual deviance: 3731.0  on 8225  degrees of freedom
(588 observations deleted due to missingness)
AIC: 42908

Number of Fisher Scoring iterations: 6

    Theta:  0.030677
Std. Err.: 0.000766
Warning while fitting theta: alternation limit reached

2 x log-likelihood: -42897.30100
Year Effect: Conclusion

With $p < 0.05$:

5. Conclusion
Conclusion

- There’s no result about preference of singular “they”.
- Singular “they” occurs a lot in spoken genre, but the reason could be that pronouns are frequently used more while speaking.
- Singular “they” shows a decrease in its trend in OANC corpus, but it may be caused by the unequal distribution of time in the corpus.
Question?