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

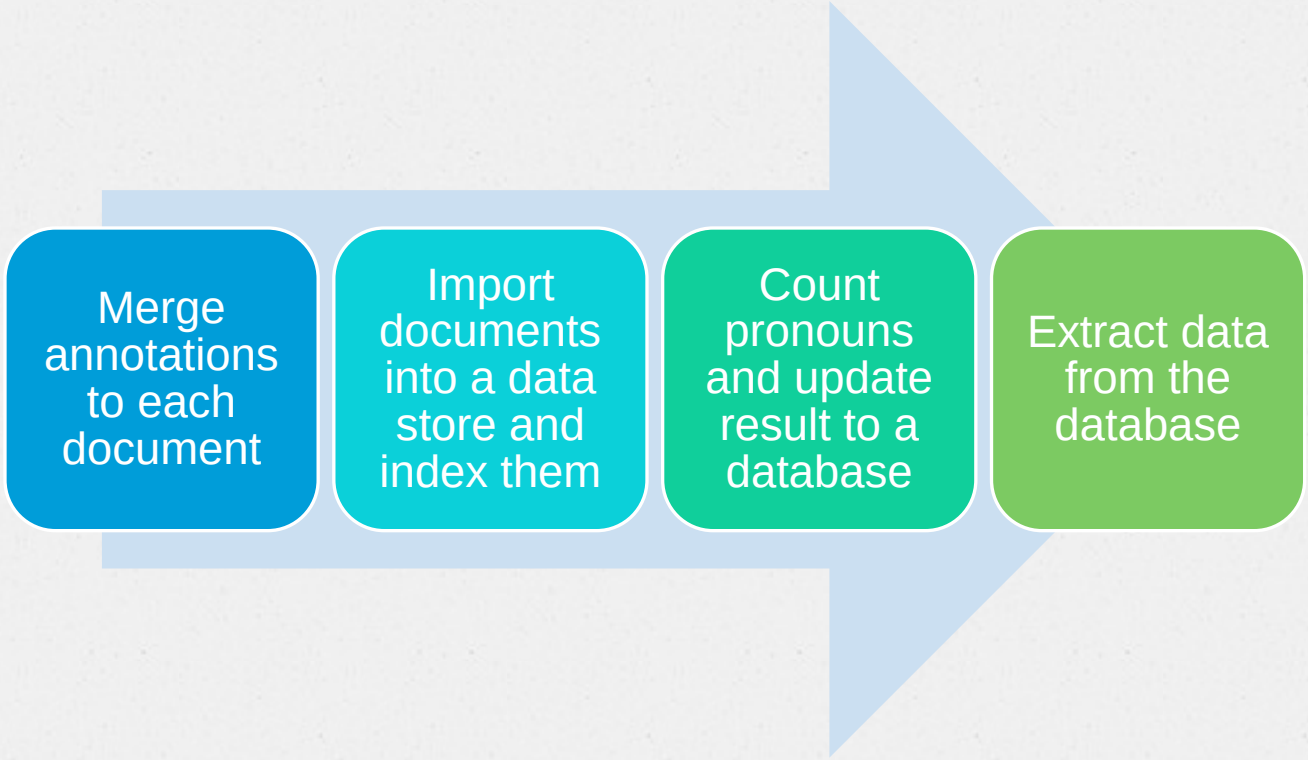
OANC

Spoken			
Name	**Domain**	**No. files**	**No. words**
charlotte	face to face	93	198,295
switchboard	telephone	2,307	3,019,477
Spoken Totals		**2,410**	**3,217,772**
Written			
Name	**Domain**	**No. files**	**No. words**
911 report	government, technical	17	281,093
berlitz	travel guides	179	1,012,496
biomed	technical	837	3,349,714
eggan	fiction	1	61,746
icic	letters	245	91,318
oup	non-fiction	45	330,524
plos	technical	252	409,280
slate	journal	4,531	4,238,808
verbatim	journal	32	582,384
web data	government	285	1,048,792
Written Totals		**6424**	**11,406,155**
Corpus Totals		**8,832**	**14,623,927**

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



Data

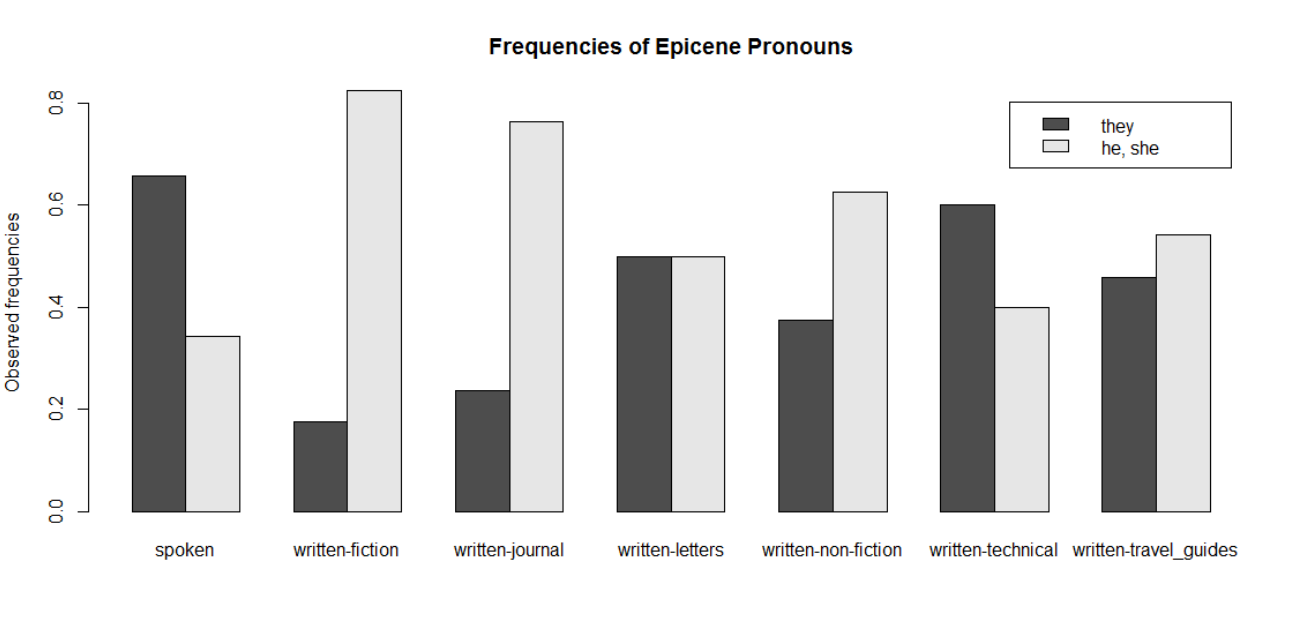
id	doc	genre	wordCount	year	theyCount	hesheCount
1	AdamsElissa.anc_00007.xml	spoken	1235	1998	0	0
2	AdamsStephanie.anc_00008.xml	spoken	966	1998	3	0
3	AdinolfiDavidandGail.anc_00009.xml	spoken	2573	1998	0	1
4	ArguetaBertila-ENG.anc_0000A.xml	spoken	4003	1998	4	7
5	AverittShannon.anc_0000B.xml	spoken	2477	1998	1	0
6	BlanchardTracy.anc_0000C.xml	spoken	1227	1998	2	1

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



A Quick Look at Data: Genre

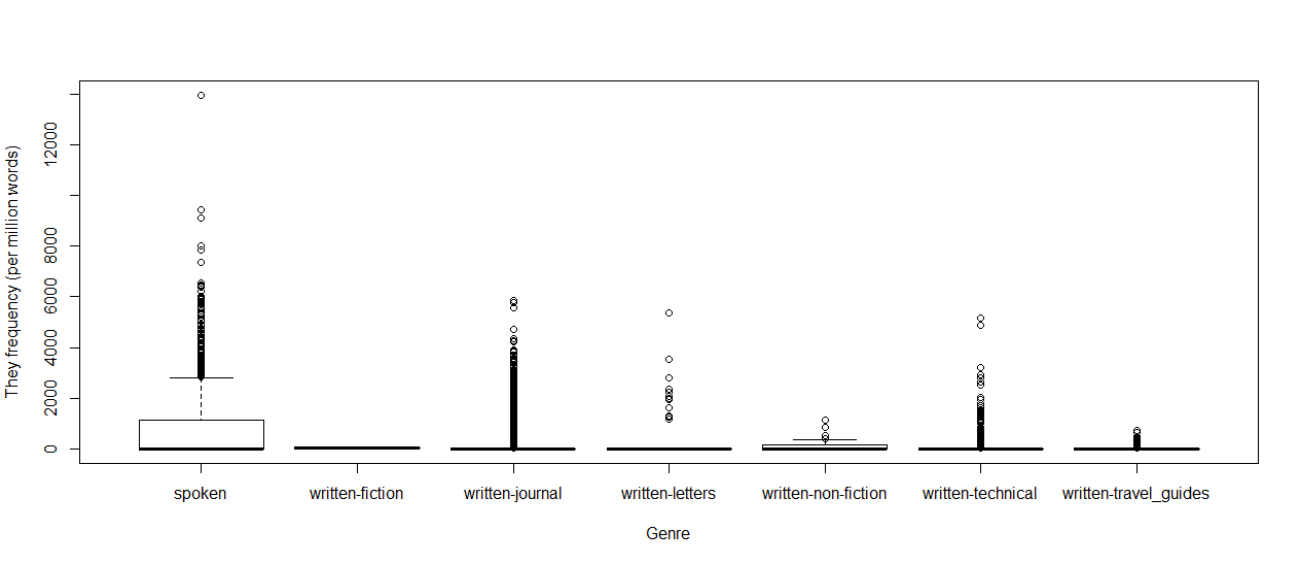
	Spoken	Fiction	Journal	Letters	Non-fiction	Technical	Travel guides
They	2387	3	835	14	40	323	45
He, she	1250	14	2688	14	67	215	53

```
> 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 ~ Binomial distribution
- Count value ~ Poisson distribution

Data Transformation: Frequency per Million Words

- The frequency of “they” in a document depends 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      -2.7809794  0.1443394   -19.27   <2e-16 ***
genrewritten-journal      -1.5038161  0.0013464 -1116.90   <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 = theyFPM ~ 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  0.0030599  -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(1826303) 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 = theyFPM ~ genre, data = ep)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.9068 -0.4011 -0.3993 -0.3993 25.0171

Count model coefficients (poisson with log link):
                             Estimate Std. Error  z value Pr(>|z|)
(Intercept)                 7.4473496  0.0007347 10136.43   <2e-16 ***
genrewritten-fiction       -3.5762772  0.1443487   -24.77   <2e-16 ***
genrewritten-journal       -0.3157891  0.0013464  -234.54   <2e-16 ***
genrewritten-letters        0.2729765  0.0058883    46.36   <2e-16 ***
genrewritten-non-fiction   -1.7134659  0.0134251  -127.63   <2e-16 ***
genrewritten-technical     -1.0718371  0.0030598  -350.30   <2e-16 ***
genrewritten-travel_guides -1.6702135  0.0093050  -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   1.18456    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

```
zeroinfl(formula = theyFPM ~ genre, data = ep, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.6742 -0.3259 -0.3245 -0.3245 20.3256

Count model coefficients (negbin with log link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                7.44736    0.02029 366.962  < 2e-16 ***
genrewritten-fiction      -3.57744    0.68186  -5.247 1.55e-07 ***
genrewritten-journal      -0.31577    0.03347  -9.433  < 2e-16 ***
genrewritten-letters       0.27340    0.18610   1.469    0.142
genrewritten-non-fiction  -1.71393    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    0.02990  27.139  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.19457    0.04109   4.736 2.18e-06 ***
genrewritten-fiction      -11.76086  324.78202  -0.036    0.971
genrewritten-journal        1.64056    0.05945  27.594  < 2e-16 ***
genrewritten-letters        2.68712    0.28795   9.332  < 2e-16 ***
genrewritten-non-fiction    0.21191    0.30709   0.690    0.490
genrewritten-technical      1.63028    0.08778  18.573  < 2e-16 ***
genrewritten-travel_guides  1.18486    0.19095   6.205 5.47e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.2511
Number of iterations in BFGS optimization: 32
Log-likelihood: -2.005e+04 on 15 Df
```

Genre Effect: Z1 vs Z2

- AIC(Z1) = 1286613.65
- AIC(Z2) = 40120.18

Genre Effect: Z3

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

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.4277 -0.4276 -0.4276 -0.4274 16.4089

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

- AIC(Z2) = 40120.18
- AIC(Z3) = 41060.61

Genre Effect: Conclusion

With $p < 0.05$:

- 78% of documents do not contains singular “they”
- Genre do have 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 = theyFPM ~ year, data = ep, 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 8226  degrees of freedom
Residual deviance: 3737.0  on 8225  degrees of freedom
  (588 observations deleted due to missingness)
AIC: 42903

Number of Fisher Scoring iterations: 6


              Theta:  0.030677
          Std. Err.:  0.000768
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -42897.301000
```

Year Effect: Conclusion

With $p < 0.05$:

- The trend of using singular “they” decreases through year (1992-2008).

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?