Towards the inevitable demise of *everybody*?

A multifactorial analysis of *-one*/*-body*/ ~~*-man*~~ variation in indefinite pronouns in historical American English

Emily Öhman[1], Tanja Säily[1] & Mikko Laitinen[2]
[1]University of Helsinki, [2]University of Eastern Finland

Aim

- The paradigmatic alternatives *-body* and *-one* in the pronominal quantifiers
  - *somebody / someone*
  - *anybody / anyone*
  - *everybody / everyone*
  - *nobody / no one*
- c. 200 years of AmE in the *Corpus of Historical American English* (COHA)

1) However, I don't bear **anybody** any ill will, and hope when we part it will be on good terms (COHA, 1857)
2) Something from Tom Tracy would have pleased you more than from **anyone**; but listen to me, Annie. (COHA, 1886)
3) There was but one sane inference: **someone** had taken a liberty rather gross. (COHA, 1898)
4) I thought I heard **somebody** speak. (COHA, 1828)

Previous findings

- *-one* more prestigious of the two variants (D'Arcy et al. 2013)
- *-body* preferred in spoken uses and in American English (Bolinger 1976; Quirk et al. 1985: 378; Biber et al. 1999: 352–3; D'Arcy et al. 2013: 296–298)
- Svartvik & Lindquist (1997): no convincing evidence of a variety distinction in the late 20th century
- A gradual shift towards *-one*: led by women (Nevalainen & Raumolin-Brunberg 2003; Laitinen 2018)
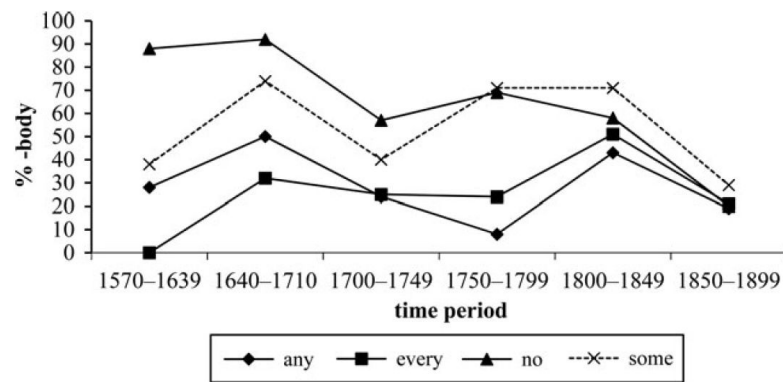- Quantifiers play a role (D'Arcy et al. 2013: 293)

FIGURE 4. Distribution of *–body* by quantifier, 1570–1899 (PPCEME, PPCMBE).

Material: COHA

- *Corpus of Historical American English* 1810–2009, c. 400 million words
- Full-text version of COHA used
- The corpus is balanced by genre across the decades, and for sub-genres and domains as well[1]
- Tagged using CLAWS7

[1] e.g. by Library of Congress classification for non-fiction; and by sub-genre for fiction – prose, poetry, drama…

- Fiction
- Magazines
- Newspapers
- Non-fiction

Prevalence of one-word spellings in COHA:
-body first

- *-body* first early 19th century
- *-one* by early 20th
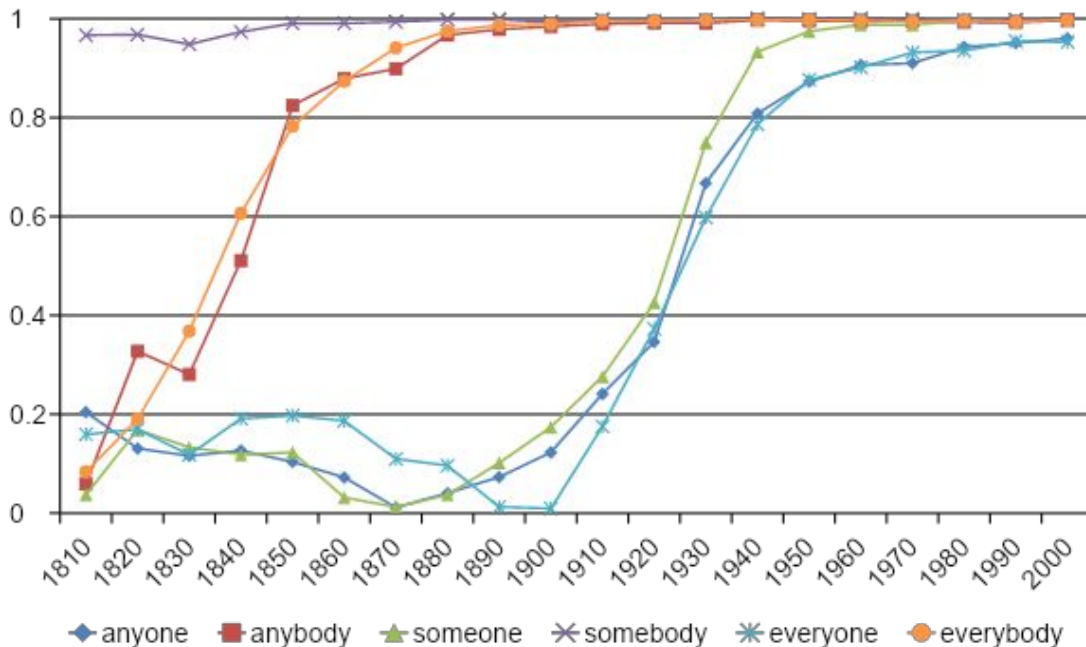
Data retrieval

- Python was used to extract pertinent data from the COHA full-text archive
- The base data was retrieved using simple regular expressions to match indefinite pronouns (words, not lemma) in the plain text and xml files yielding 111,122 eligible texts out of 112,388 texts
- A stricter search using the lemmatized and tagged wlp files (tagged as **pn**, not followed by noun, -*one* not followed by *of*) yielded only 45,276 texts (this warrants further investigation!)
- The data was manipulated using mainly Python (pandas in particular), regular expressions, and stored as .csv files
- All spelling variants were included (modern, separate, as well as misspellings such as *any buddy*)

Gender script overview

- Parish records (Wrigley et al. 2018) were used as the basis of name-based gender identification
    - For those still unknown, a mix of modern name lists and machine learning approaches was used to optimize the number of matches. Sanity checks were performed
    - 'Unknown' includes those texts where no author was indicated
   - 'Errors' are due to some texts in COHA having more than one text id
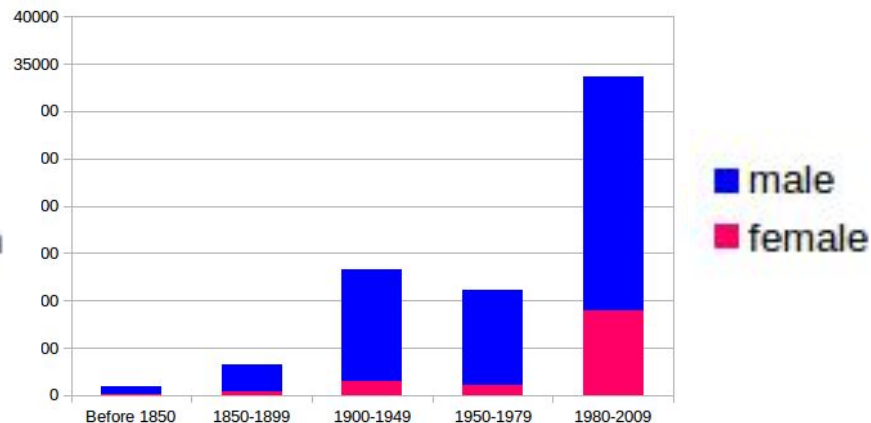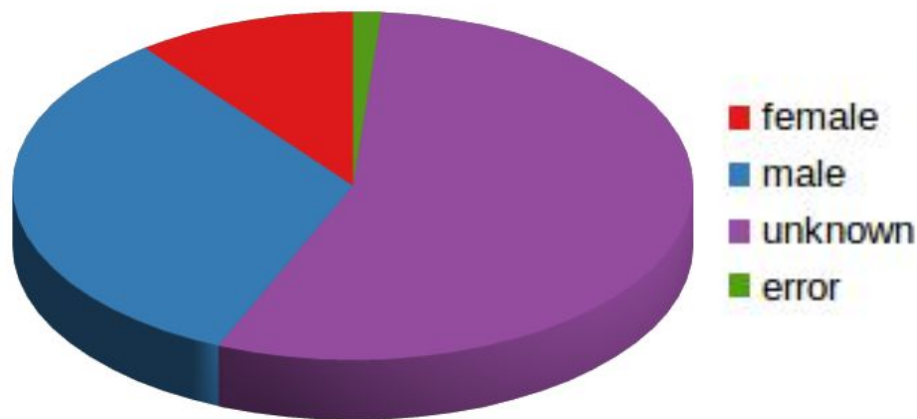
Methods

- Frequency comparisons over time
  - Normalized frequency / 10,000 words
  - % of variant (*-one*) out of variable (*-one*/*-body*)

- Visualization
  - Line graphs
    - Some with random sampling to show variability depending on corpus composition
  - Beanplots (Kampstra 2008; Säily et al. 2011)

- Multivariate: logistic regression
  - Dependent variable: *-one* (reference level: *-body*)
  - Internal factors: lemma (*any*, *some*, *every*, *no*), post-modifying preposition
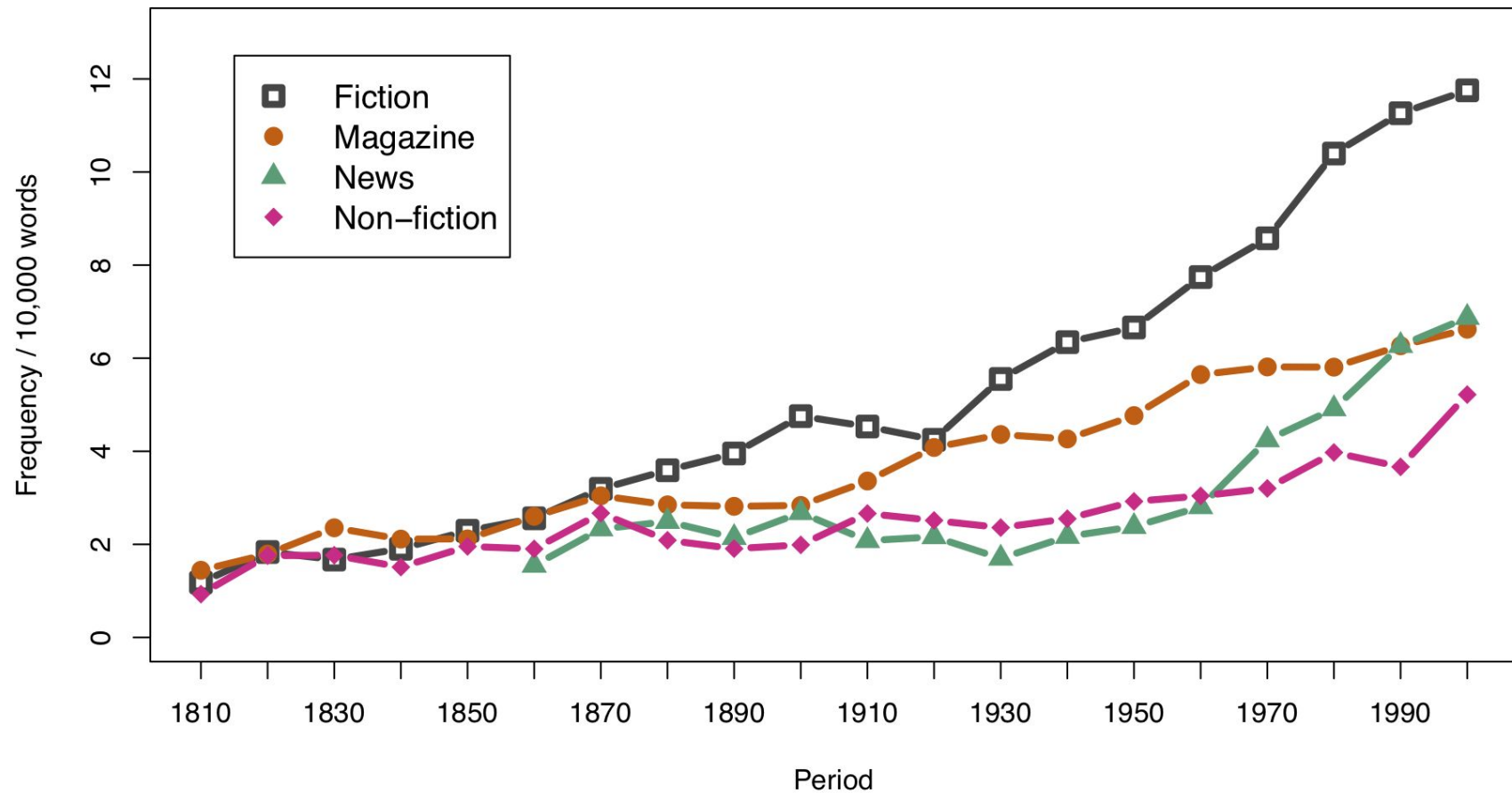 - External factors: genre, gender, year

Results

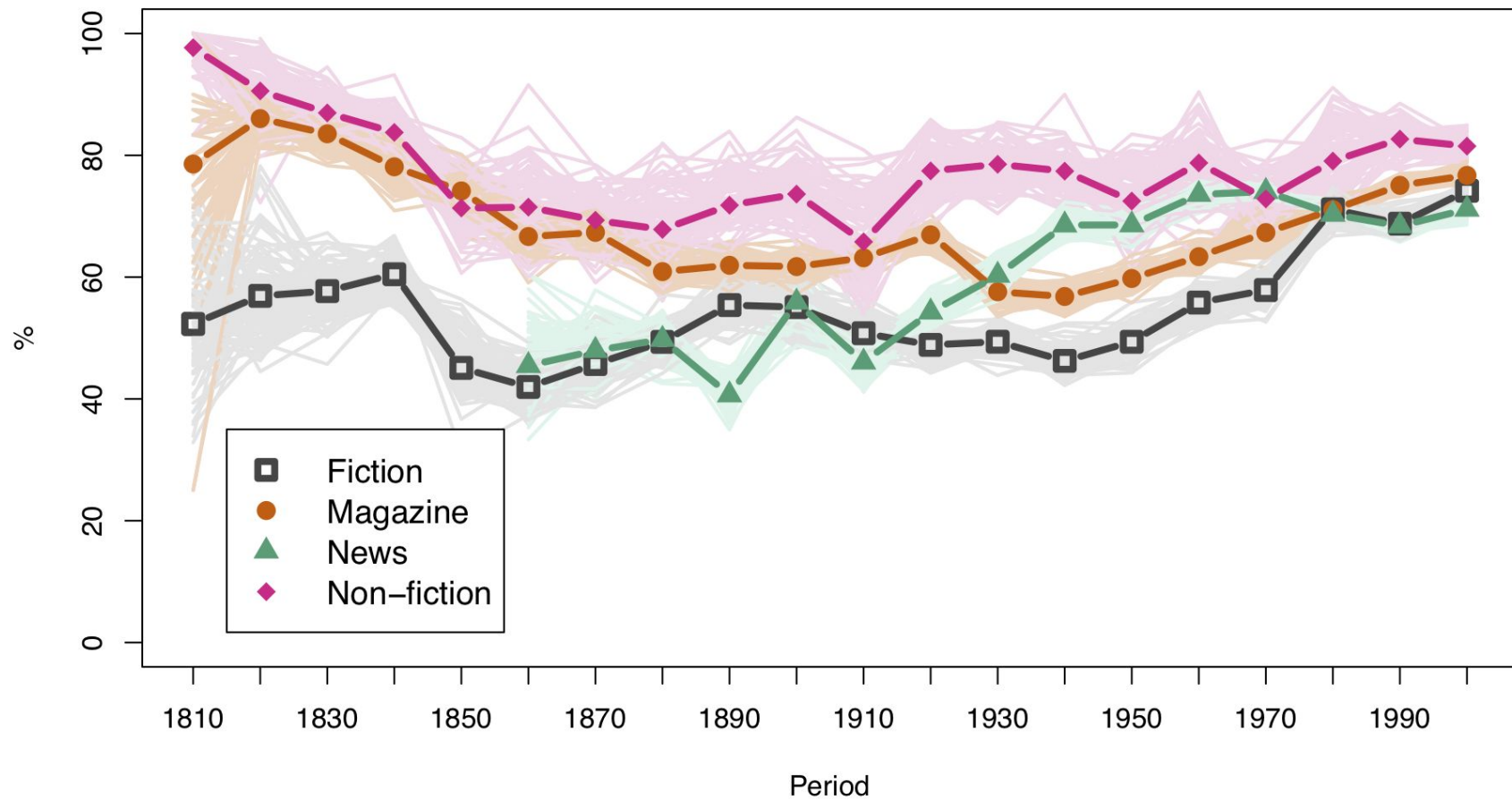# Proportion of indefinite *–one* in COHA by lemma
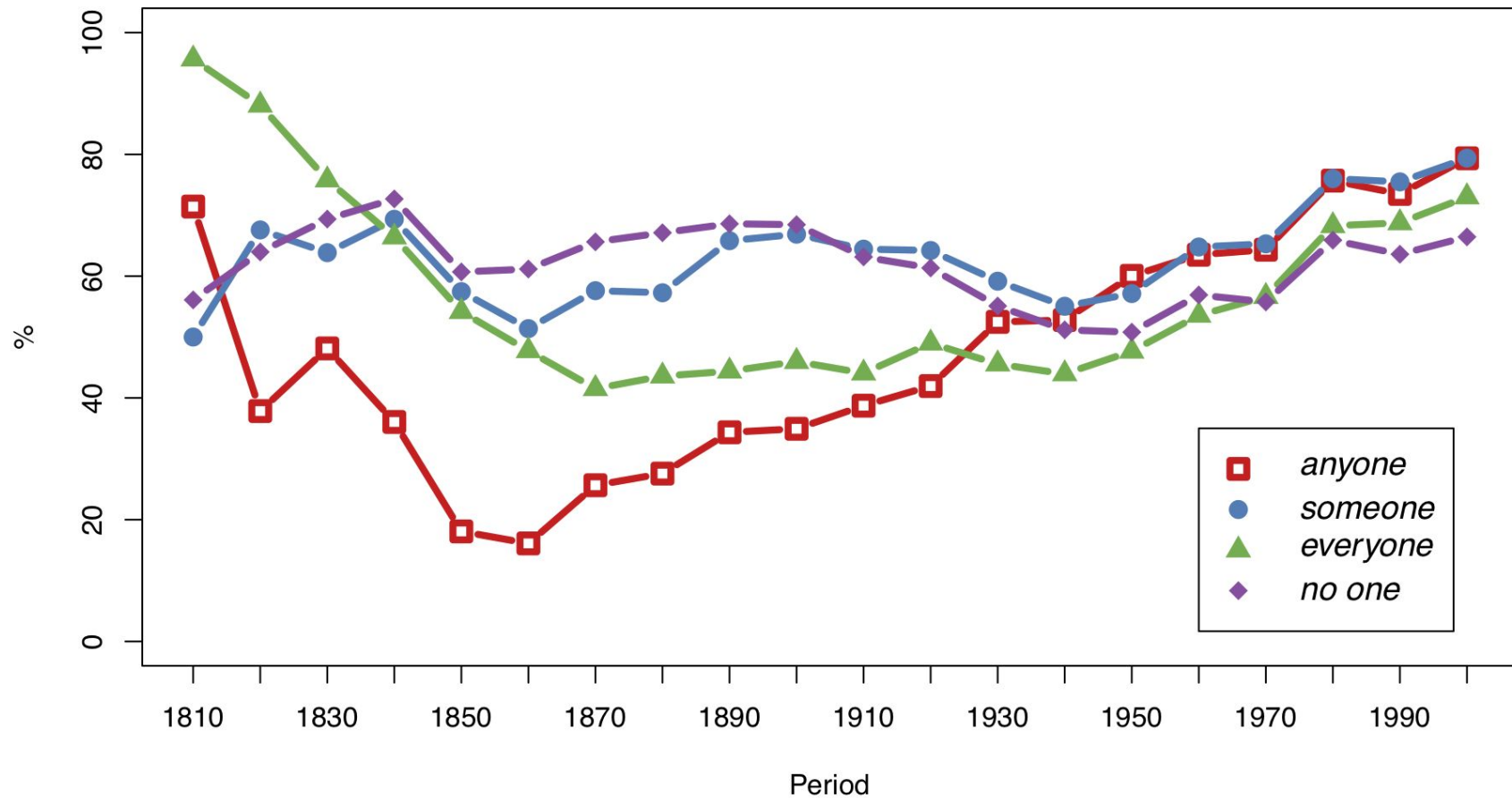
Indefinite *–one* and *–body* in COHA
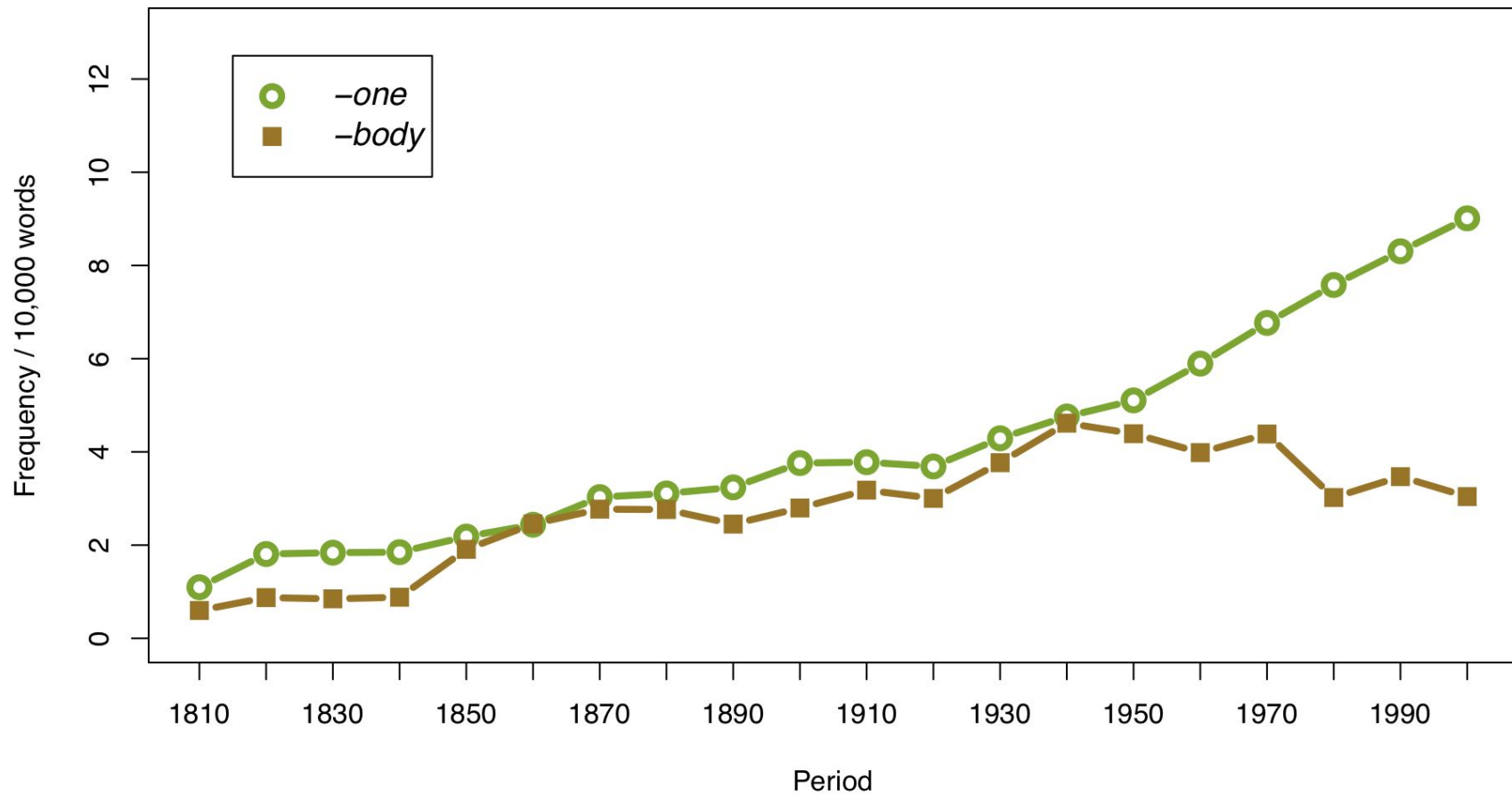
Indefinite *–one* in COHA by genre

Proportion of indefinite *–one* in COHA by genre

Results of logistic regression (lrm)

- 303,556 observations
- *-body* as reference level; coefficients for *-one*
- vif values <1.8 (no multicollinearity), but the regression only has weak predictive power
- Likelihood of *-one* in non-fiction is 2.8 when compared with fiction
- Change towards *-one* led by women
- *no*/*some* favor *-one*
- The presence of a post-modifying prep increases the likelihood of *-one* (cf. Svartvik & Lindquist 1997)
- Regression analysis replicated with post-1940 observations (182,150): no major differences in the coefficients

	Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes		
Obs	303556	LR chi2	12879.59	R2	0.056	C	0.62
body	118702	d.f.	10	g	0.495	Dxy	0.24
one	184854	Pr(> chi2)	<0.0001	gr	1.641	gamma	0.24
max |deriv|	5e-08			gp	0.113	tau-a	0.114
				Brier	0.228		

|	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	-10.4532	0.1641	-63.70	<0.0001
genre=mag	0.4613	0.0107	42.94	<0.0001
genre=news	0.3130	0.0165	18.98	<0.0001
genre=nf	1.0251	0.0175	58.42	<0.0001
gender=male	-0.2007	0.0090	-22.28	<0.0001
gender=unknown	-0.1484	0.0136	-10.91	<0.0001
year	0.0055	0.0001	65.51	<0.0001
lemma=every	-0.1033	0.0113	-9.12	<0.0001
lemma=no	0.2311	0.0110	20.99	<0.0001
lemma=some	0.4282	0.0114	37.45	<0.0001
prep=TRUE	0.3776	0.0137	27.62	<0.0001

Zooming in: *-one* and *-body* in COHA fiction

- Data coverage
    - c. **50%** of COHA is fiction

- Metadata coverage
    - **Author + gender** best known in fiction

- Genre
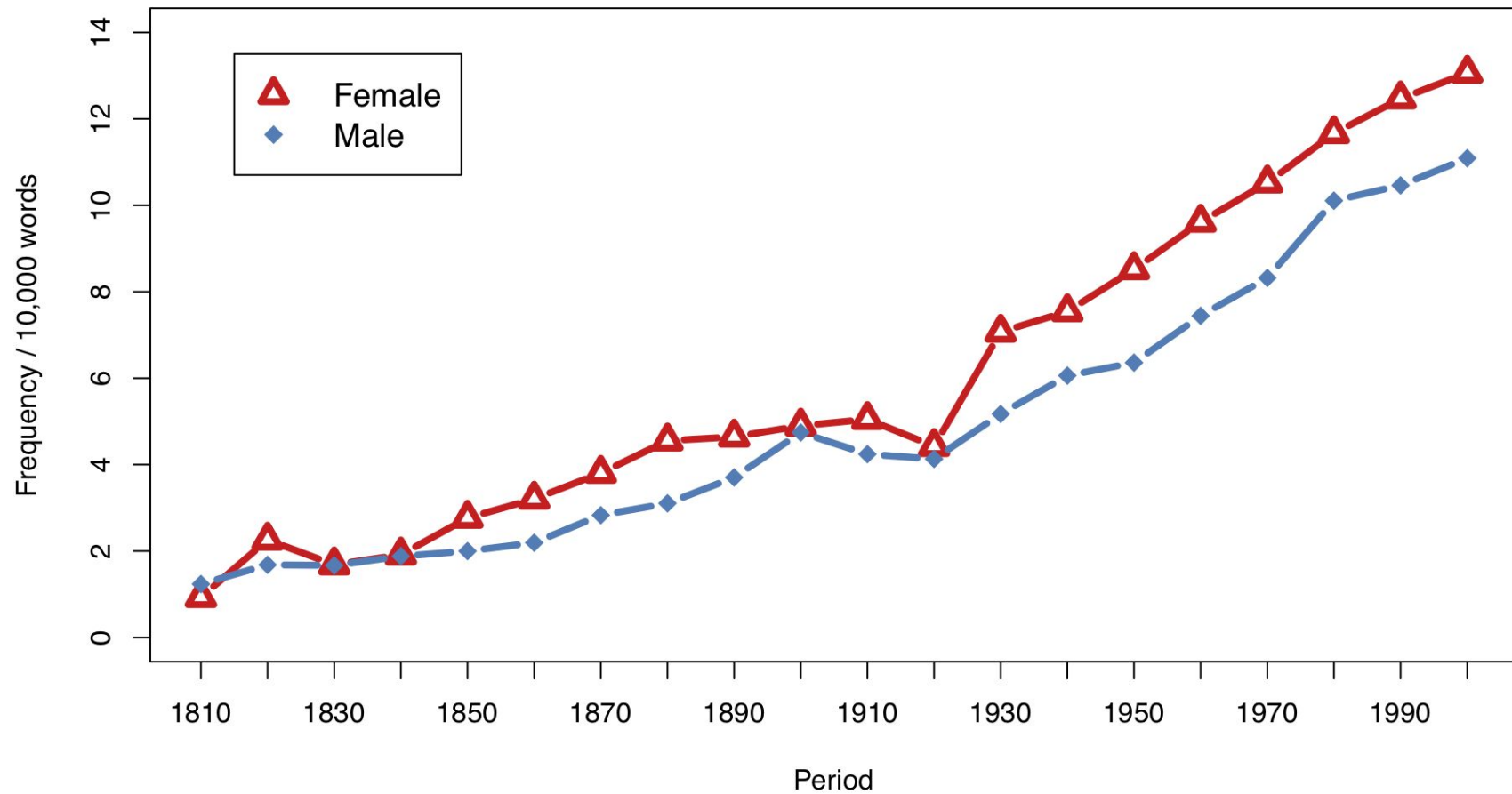   - Fiction **more oral**, good for sociolinguistic studies

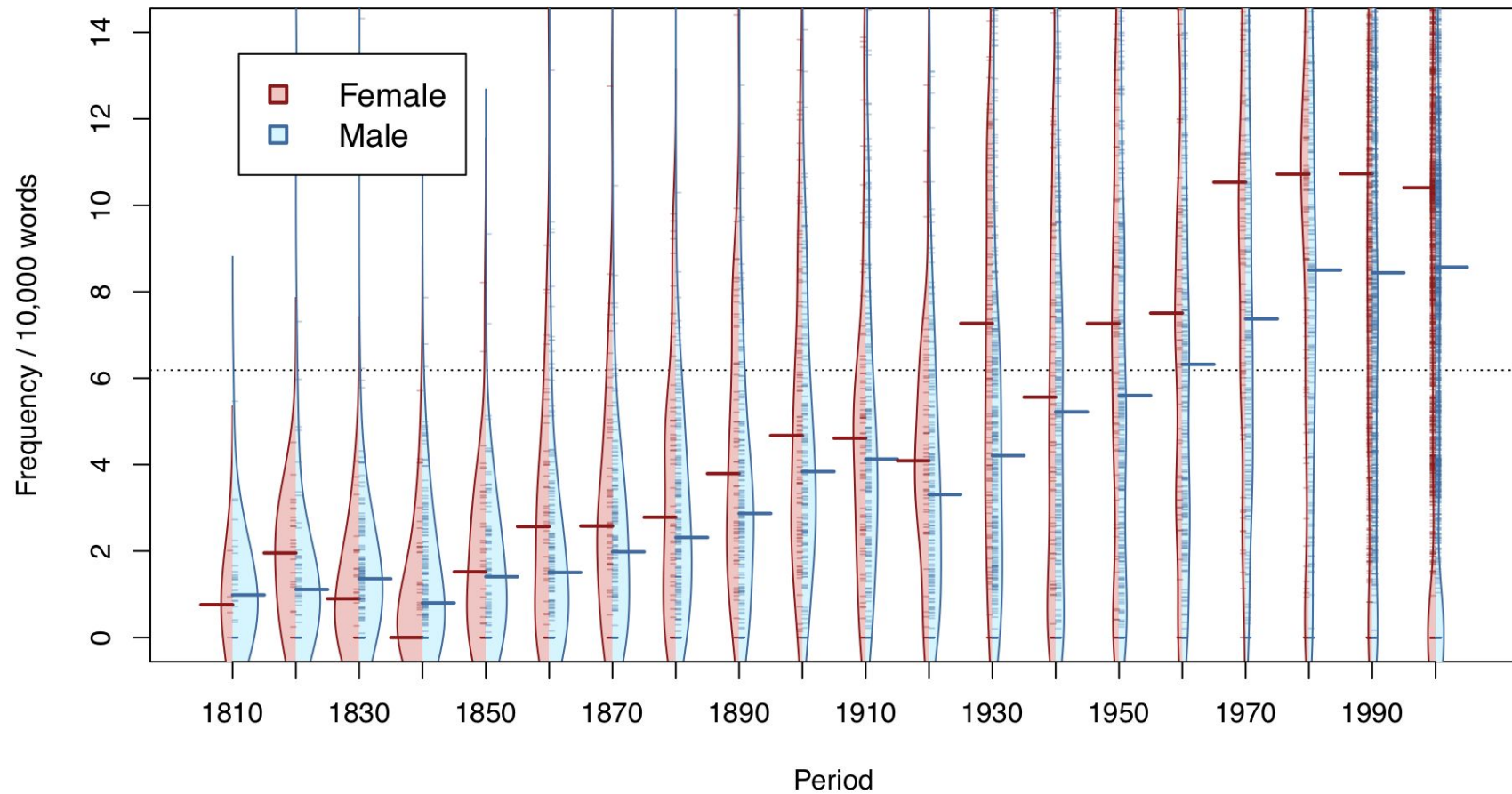Indefinite *–one* and *–body* in COHA fiction
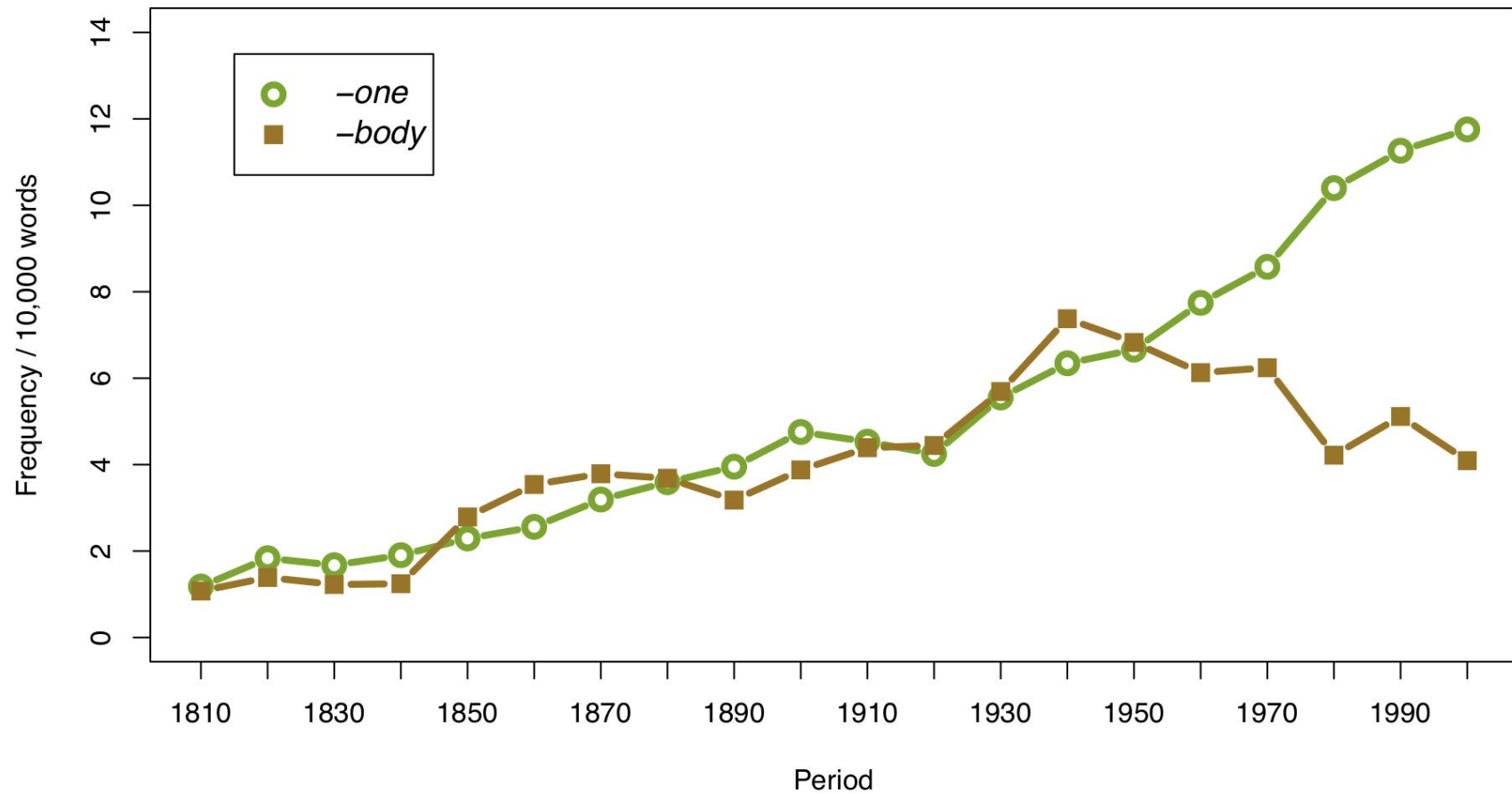
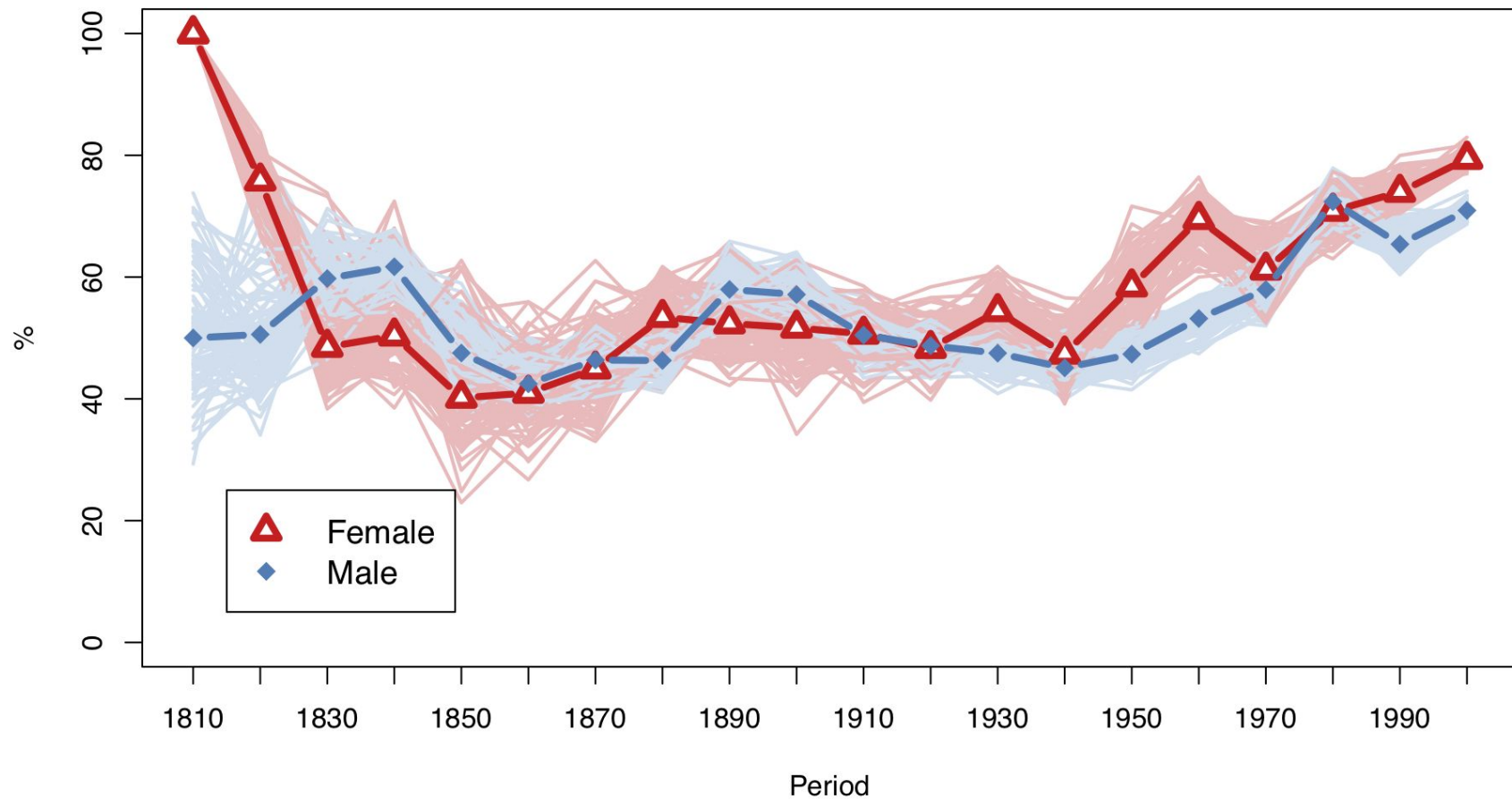Beanplot: Indefinite *–one* and *–body* in COHA fiction
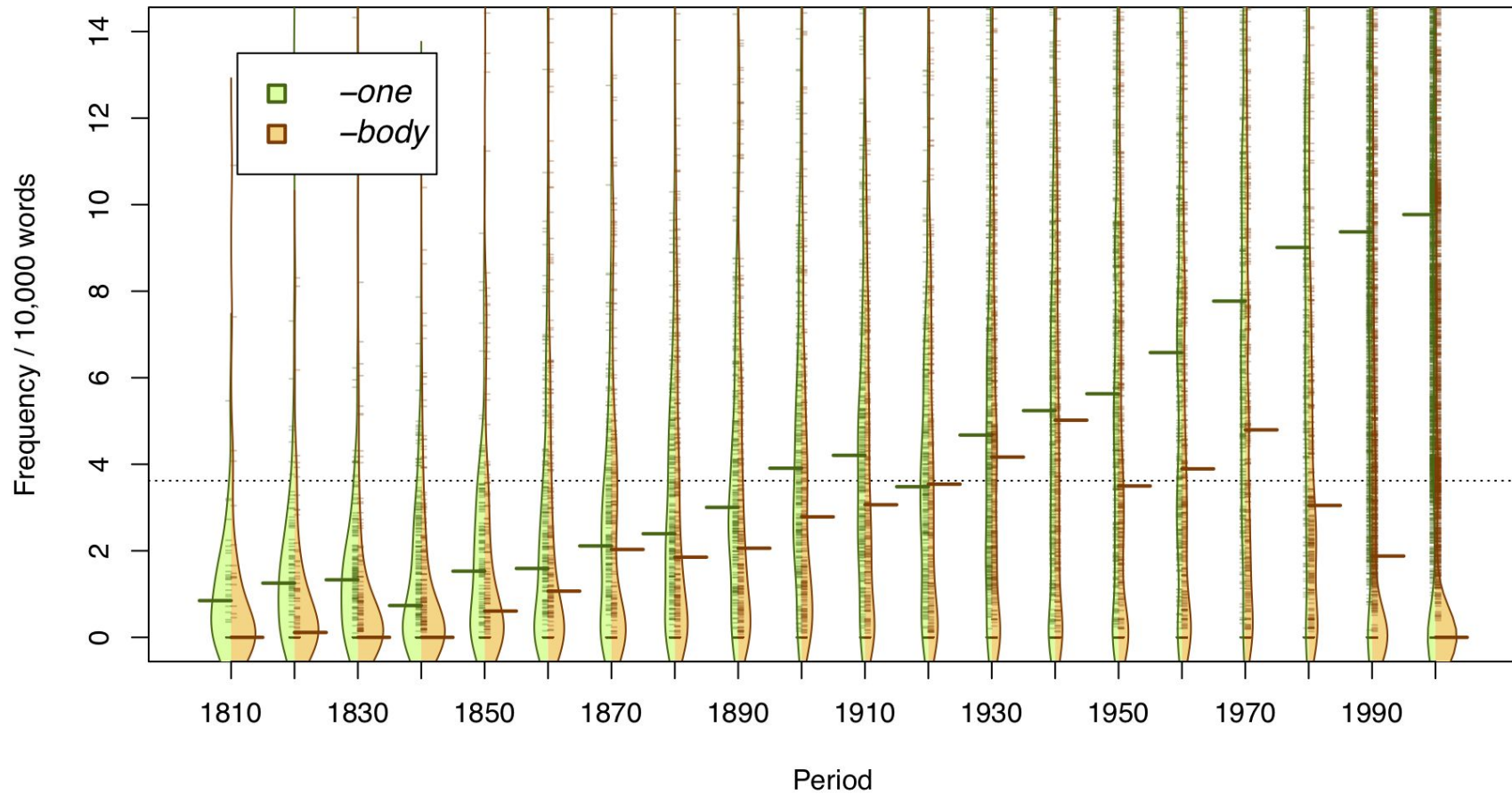
Indefinite *–one* in COHA fiction by gender

Beanplot: Indefinite *–one* in COHA fiction by gender

Proportion of indefinite *–one* in COHA fiction by gender

Discussion

- Methodological point: we have added gender metadata to texts in COHA

- Issues with COHA metadata
  - No uniform way of listing the author names
  - Some texts contain more than one text id

- COHA as material for sociolinguistic investigation?
  - Our results look promising: consistent gender difference in the use of *-one*

Conclusion

- D'Arcy et al. (2013): the demise of -*body* is accelerated starting from c. 1930s onwards in British English
- We observe a similar tendency around the same time in AmE starting from the 1940/50s
  - AmE lags behind (cf. Hundt 2009)
  - Change accelerated after large-scale mobility (cf. Raumolin-Brunberg 1998)
- Multifactorial analysis:
  - Normalized frequency of -*one* is largest in fiction, but non-fiction in particular increases the probability
  - *some*/*no* favor -*one* in AmE
  - Similar to BrE (EModE and LModE): the change is led by women

Thank you!

Contact:
emily.ohman@helsinki.fi / tanja.saily@helsinki.fi / mikko.laitinen@uef.fi

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English.* London: Longman.

Bolinger, Dwight. 1976. The *in*-group: *One* and its compounds. In P.A. Reich (ed.), *The Second LACUS Forum 1975*, 229–237. Columbia, S.C.: Hornbeam Press.

COHA = Davies, Mark. 2010–. *The Corpus of Historical American English (COHA):* 400 million words, 1810–2009.

D'Arcy, Alexandra, Bill Haddican, Hazel Richards, Sali A. Tagliamonte & Ann Taylor. 2013. Asymmetrical trajectories: The past and present of *-body/-one*. *Language Variation and Change* 25(3). 287–310.

Hundt, Marianne. 2009. Colonial lag, colonial innovation or simply language change? In Günter Rohdenburg & Julia Schlüter (eds.), *One language, two grammars? Differences between British and American English,* 13–37. Cambridge: Cambridge University Press.

Kampstra, Peter. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software* 28. Code Snippet 1.

Laitinen, Mikko. 2018. Indefinite pronouns with singular human reference: Recessive and ongoing. In Terttu Nevalainen, Minna Palander-Collin & Tanja Säily (eds.), *Patterns of change in 18th-century English: A sociolinguistic approach,* 137–158. Amsterdam: John Benjamins.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England.* London: Pearson Education.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Raumolin-Brunberg, Helena. 1998. Social factors and pronominal change in the seventeenth century: The Civil War effect? In Jacek Fisiak & Marcin Krygier (eds.), *Advances in English Historical Linguistics (1996),* 361–388. Berlin: Mouton de Gruyter.

Säily, Tanja, Terttu Nevalainen & Harri Siirtola. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26(2). 167–188.

Svartvik, Jan & Hans Lindquist. 1997. *One* and *body* language. In Udo Fries, Viviane Müller & Peter Schneider (eds.), *From Ælfric to the New York Times: Studies in English corpus linguistics*, 11–20. Amsterdam & Atlanta: Rodopi.

Wrigley, E.A., R.S. Davies, J.E. Oeppen & R.S. Schofield. 2018. *26 English parish family reconstitutions*. [Data Collection]. Colchester, Essex: UK Data Archive. doi:10.5255/UKDA-SN-853082