

Variation in noun and pronoun frequencies

Gendered drift or a corpus artefact?

2010-05-27

Tanja Säily (VARIENG, University of Helsinki)
with Terttu Nevalainen (VARIENG) and Harri Siirtola (TAUCHI)

Introduction

- **Noun ratio** (proportion of nouns out of all words) **declines** over time – true or a corpus artefact?
 - Key concern in historical linguistics: comparable data over time
- **Genre evolution** a known issue to corpus compilers (e.g., Nevalainen and Raumolin-Brunberg 1993)
 - Grammatically annotated corpora → studies on genre consistency now possible

Noun & pronoun ratios

- Indicators of **content type / focus** (e.g., Biber & Finegan 1989)
 - ‘Nouny’ texts more informational
 - ‘Pronouny’ texts more involved
- Associated with **gendered styles** (Rayson et al. 1997, Argamon et al. 2003)
 - Women use fewer nouns and more personal pronouns than men in PDE

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Material: PCEEC

- *Parsed Corpus of Early English Correspondence* (2006), 2.2 Mw
 - 4,969 personal letters written c.1415–1681
 - Based on good, original-spelling editions
- Useful for **historical sociolinguistics**
 - Metadata about letters, writers, recipients
- 3 versions: plain text, **tagged**, parsed
 - Penn-Helsinki corpus annotation scheme

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Research questions

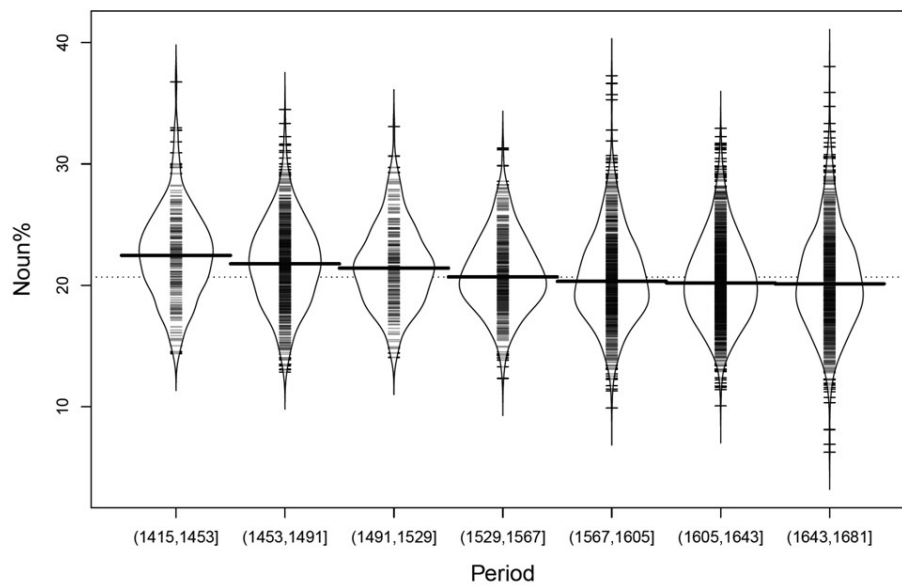
- Is there significant **variation** in noun and pronoun frequencies in this corpus, either over time or across writers?
- Is the variation **linguistic**, or an **artefact** of the corpus or its annotation scheme?

DAMMOC

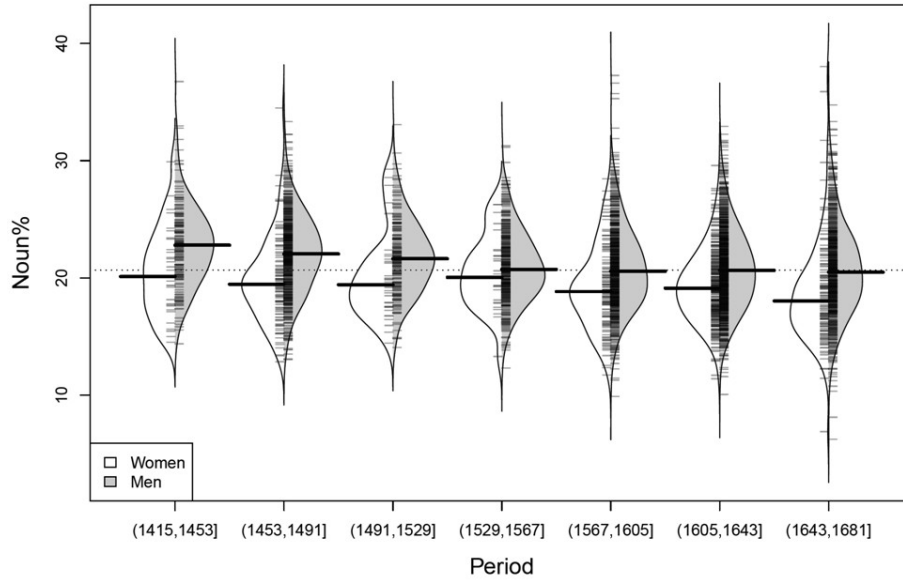
Tanja Säily, VARIENG

2010-05-27

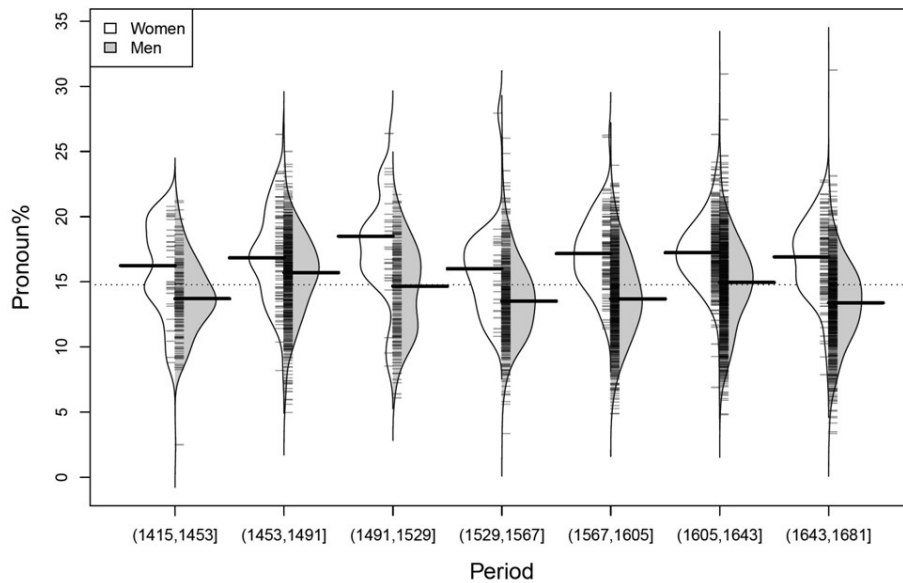
Variation in noun ratio over time



Sociolinguistic variation in noun ratio over time: gender



Sociolinguistic variation in pronoun ratio over time: gender



Results

- Trend towards fewer nouns over the centuries
 - English letter-writing becomes less focused on information over time?
- Women use more pronouns and fewer nouns than men in every subperiod
 - Gendered styles similar to PDE?

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Artefact of annotation?

- Categorisation & tokenisation issues, e.g.:
 - Adverbs conservatively tagged as nouns
 - *likewise*_ADJ+N tagged identically to *gentleman*_ADJ+N
 - When reflexive pronoun written as two words, *self* tagged as a noun
 - *him*_PRO *self*_N
 - Frequency dependent on time period

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Artefact of annotation?

- Solution: new version, ReCEEC
 - Reannotated: *likewise_ADV*, *him self_PRO*, etc.
- Main results remained the same
- Specific percentages of nouns & pronouns changed slightly (cf. Hardie 2007)

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Artefact of corpus design?

- We experimented with computationally identifying and excluding deviant data
 - Certain individuals (Dorothy Osborne...)
 - First subperiod somewhat skewed
 - Very long letters behaved differently
- Again, main results remained the same

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Examples

WILLIAM CELY (FAMILY SERVANT) TO GEORGE CELY, 1481
Off_P the_D vj_NUM **packys_NS ffell_N** beth_BEP v=c= NUM
xxxviiij_NUM cast_VAN small_ADJ **tale_N wynter_N ffellys_NS**
off_P **London_NPR** markyd_VAN wyth_P **ynccke_N - ,** the_D
marke_Nys_BEP a_D C_N - , and_CONJ certeyne_ADJ
somer_N ffellys_NS - , the_D **marke_Nys_BEP** off_P
them_PRO a_D **O_N - ,** whych_WD vj_NUM **packys_NS**
+ge_PRO muste_MD receyue_VB

DOROTHY OSBORNE (FUTURE WIFE) TO WILLIAM TEMPLE, 1652
S=r= _N **You_PRO** may_MD please_VB to_TO lett_VB **my_PRO\$**
Old_ADJ Servant_N <paren> _CODE as_P **you_PRO** call_VBP
him_PRO </paren> _CODE know_VB , , that_C **I_PRO**
confesse_VBP **I_PRO** owe_VBP much_Q to_P **his_PRO\$**
merritts_NS , , and_CONJ the_D many_Q Obligations_NS
his_PRO\$ kindenesse_N and_CONJ Civility's_NS has_HVP
layde_VBN upon_P **mee_PRO**; , .

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Linguistic variation?

- Extralinguistic factors play a part
 - Topic, education/literacy, letter length, ...
- Next step: look at the role of the recipient
 - Gender (M–M > M–F > F–M > F–F?)
 - Social status, relationship to sender

DAMMOC

Tanja Säily, VARIENG

2010-05-27

Conclusion

- Main results not a corpus artefact
 - Proportion of nouns decreases; women use more pronouns & fewer nouns than men
 - Tested by reannotating the corpus and excluding deviant data
- Genre fairly consistent
- Still: user, know thy corpus!

DAMMOC

Tanja Säily, VARIENG

2010-05-27

References

- Argamon, S., M. Koppel, J. Fine & A.R. Shimoni. 2003. "Gender, genre, and writing style in formal written texts". *Text* 23(3): 321–346.
- Biber, D. & E. Finegan. 1989. "Drift and the evolution of English style: A history of three genres". *Language* 65(3): 487–517.
- Hardie, A. 2007. "Part-of-speech ratios in English corpora". *IJCL* 12(1): 55–81.
- Nevalainen, T. & H. Raumolin-Brunberg. 1993. "Early Modern British English". *Early English in the Computer Age*, ed. M. Rissanen, M. Kytö & M. Palander-Collin, 53–73. Berlin: Mouton de Gruyter.
- PCEEC = *Parsed Corpus of Early English Correspondence*, tagged version. 2006. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.
- 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". *IJCL* 2(1): 133–152.

DAMMOC

Tanja Säily, VARIENG

2010-05-27