QUANTIFYING VARIATION AND CHANGE IN WORD TYPES

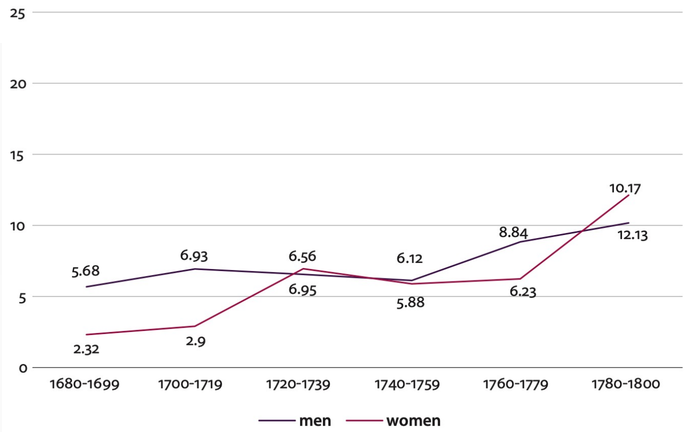
HSSH Brown Bag Seminar, 21 February 2023

Tanja Säily

HISCOP PROJECT

- *Historical Sociolinguistics Meets Construction Grammar: The Case of Productivity in English*
  - Academy of Finland, 2020–2023
  - Funded researcher: **Tanja Säily**
  - Collaborators: **Martin Hilpert**, **Jukka Suomela**, Florent Perek, Turo Vartiainen
- Aim: extend CxG by drawing on historical sociolinguistics
  - What do speakers have to know to be able to use a language? Social aspects largely missing so far
  - Focus on productivity of constructions in historical text corpora

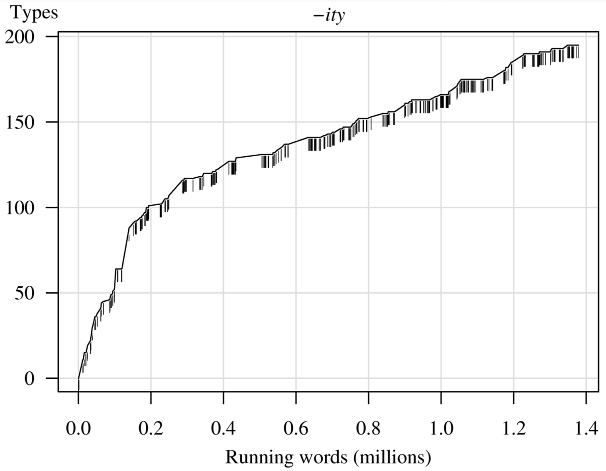
ON FREQUENCY



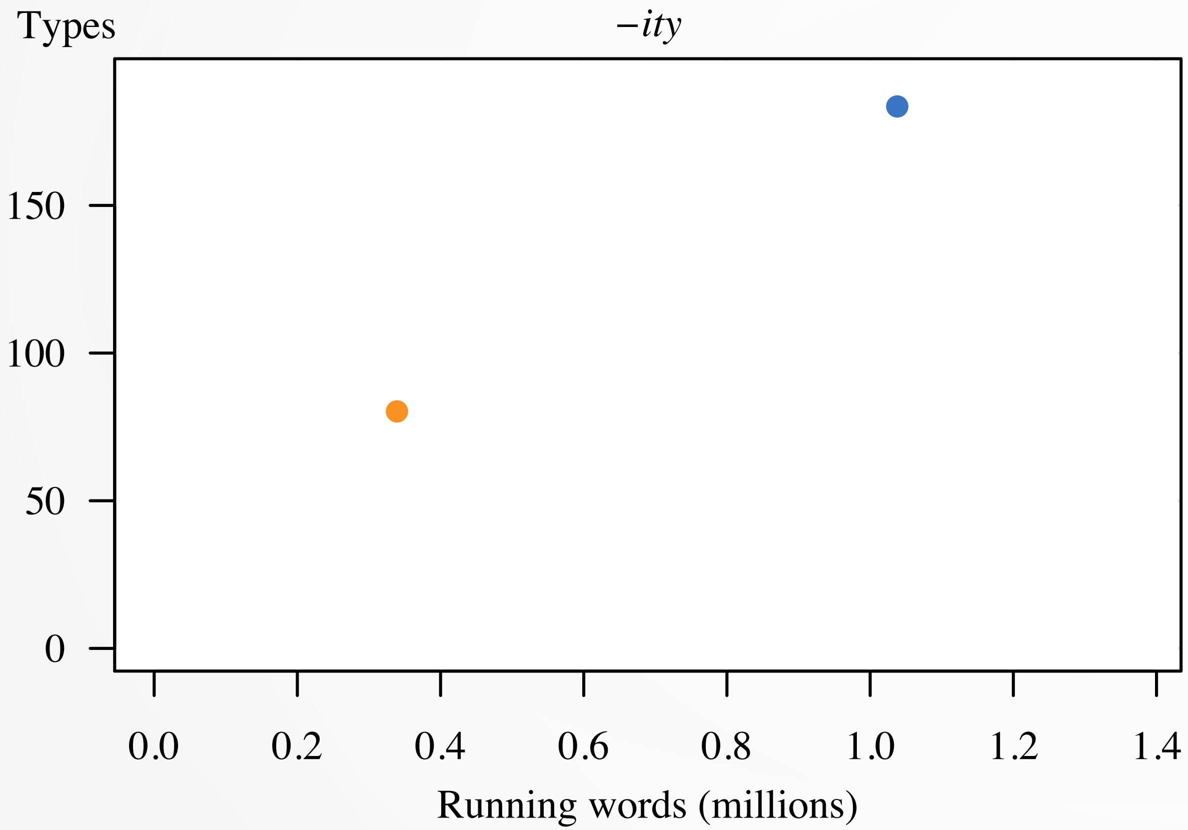
- Frequency of occurrence: an important way of assessing and comparing the prevalence of linguistic features across digitized texts
  - How do linguistic features spread over time, which social groups lead the change?
- Calculate **normalized frequency:**
  - Divide text corpus into subcorpora by social group and time period
  - Count all occurrences (*tokens*) of the feature in each subcorpus
  - Normalize the count by the number of running words in the subcorpus
- But some features can be realized through many different word *types*
  - e.g. nominal suffix -*ity*: *ability, absurdity, acclivity, acidity, activity*, …
  - The more different types, the more productively the feature is being used

MORPHOLOGICAL PRODUCTIVITY

- The readiness with which an element enters into new combinations (Bolinger 1948)
- **Quantitative measures** (e.g. Baayen 1993; Cowie & Dalton-Puffer 2002):
  - Number of different words containing the morpheme in a corpus (**types**)
  - Number of types occurring only once in the corpus (**hapax legomena**)
  - Number of types not occurring in previous periods (**new types**)
- **Problem**: Difficult to compare across (sub)corpora
  - Different amounts of data from different periods & groups
  - Type-based measures grow nonlinearly with corpus size → **normalization not justifiable**

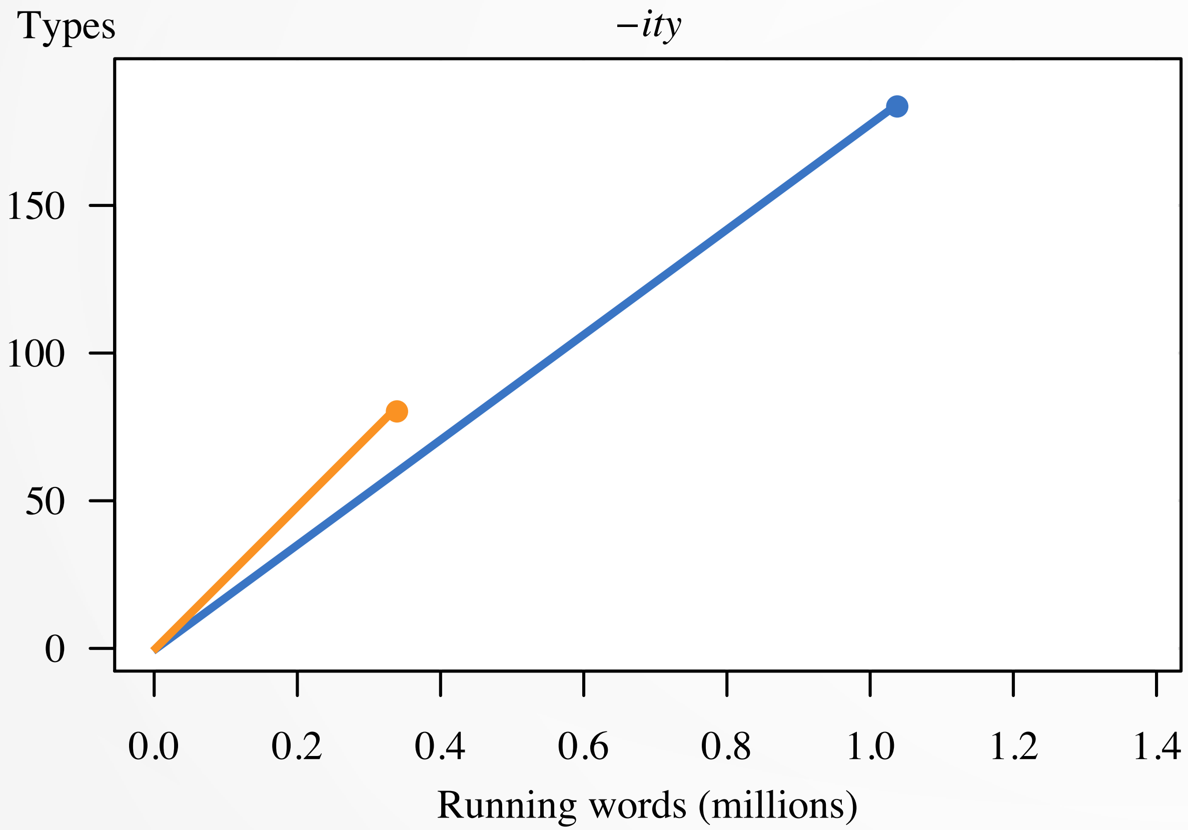


NORMALIZATION



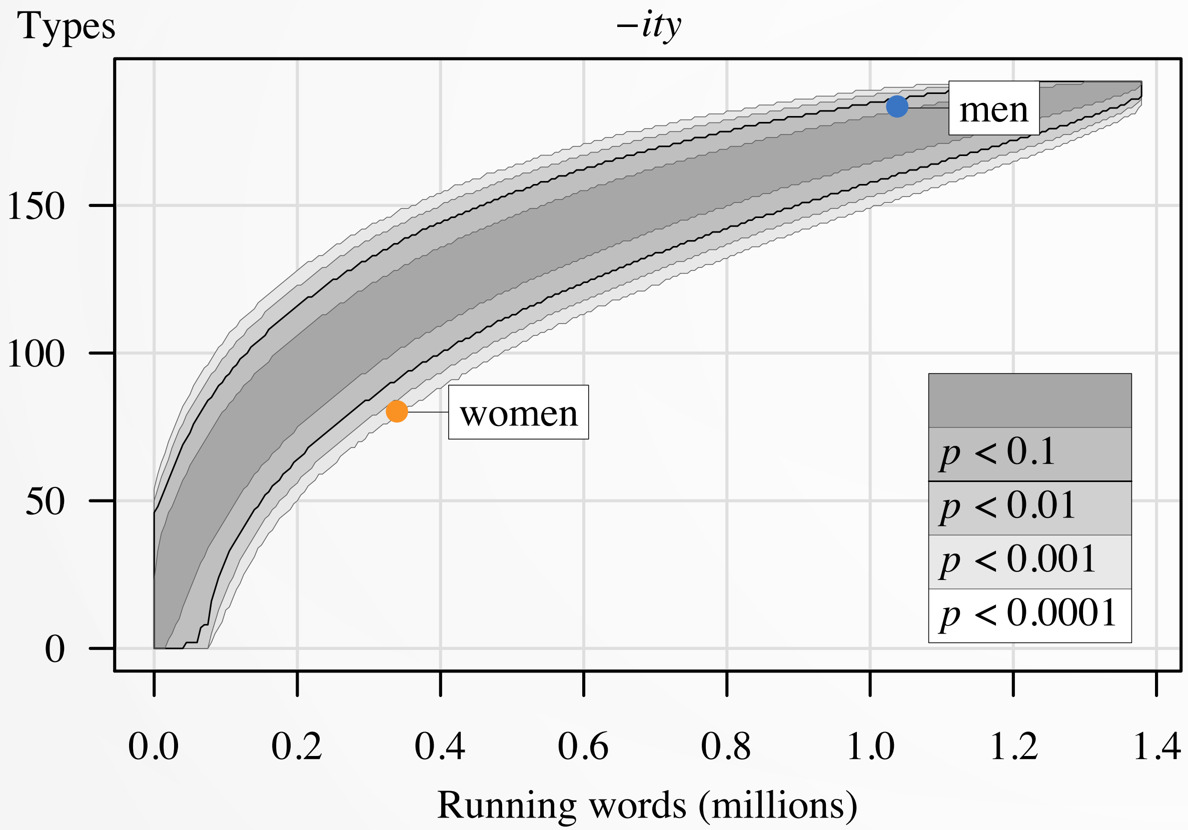
- Who uses comparatively more -*ity* types, **men** or **women**?

NORMALIZATION



- Who uses comparatively more -*ity* types, **men** or **women**?
- Normalization says women, but…

SÄILY & SUOMELA (2009, 2017)

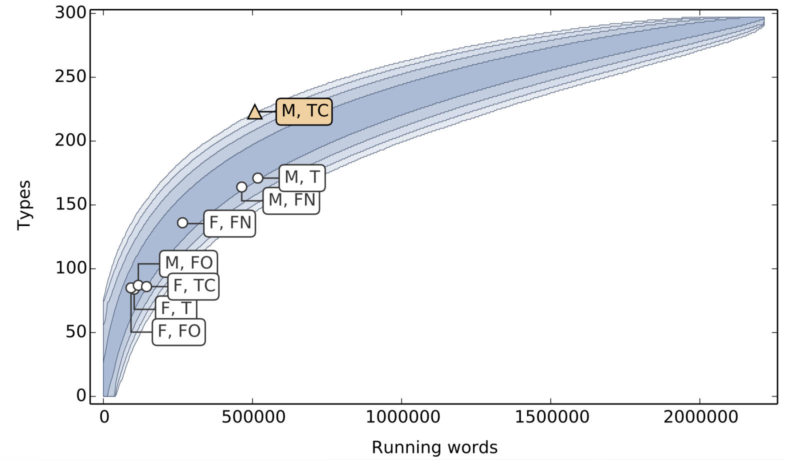


- Compare each subcorpus with subcorpora of equal size, randomly sampled from the corpus as a whole
- Automatically provides a measure of statistical significance
- **Problems**:
  - Comparisons over time still difficult; *x*-axis = corpus size, not time period
  - Only measures variation within a morpheme, not between morphemes

-ITY AND *-NESS*

- **Nominal suffixes**, usually derive abstract nouns from adjectives
  - e.g. *productive* → *productivity* or *productiveness*
- -*ness* native, -*ity* borrowed from French (+ Latin) in Middle English
  - More sociolinguistic variation in the productivity of -*ity* (Säily 2014); prestige, learnedness
- Early Modern English: large-scale expansion of vocabulary
  - -*ity* gains ground on -*ness* in all registers, starting from written registers and spreading towards speech-related ones
    - – Rodríguez-Puente (2020); Rodríguez-Puente et al. (2022)

-ITY AND *-NESS* IN C17–18 PERSONAL LETTERS



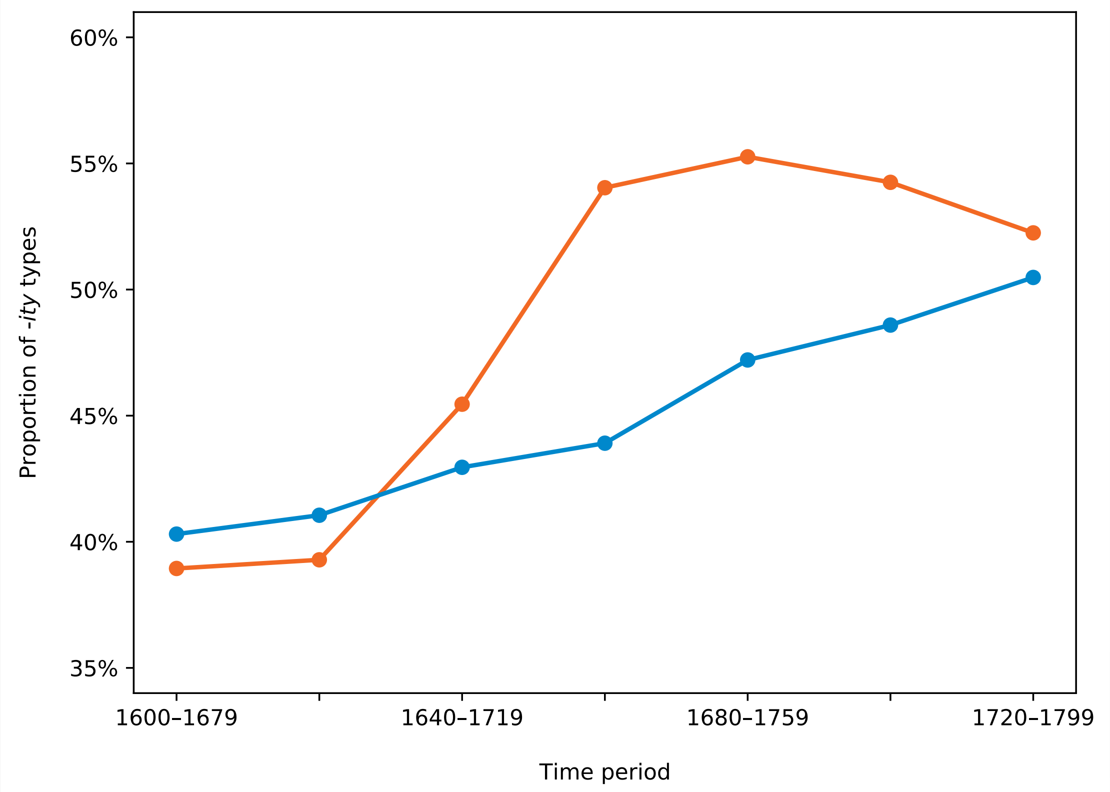
- Säily (2014): **external** factors
  - Productivity of *-ity* increases, *-ness* remains stable (*Corpora of Early English Correspondence*, type frequencies)
  - Gender: women lag behind in the use of *-ity* in C17, difference disappears in C18
    - – Exception: difference remains in letters to close friends (cf. Wolfson 1990)
- Now: analyse suffix competition (cf. Rodríguez-Puente et al. 2022), add **internal** factors
  - Hilpert (2013): a number of language-internal factors connected to change in the productivity of the V-*ment* construction (*Oxford English Dictionary*, 1250–2000)
    - – We will analyse some of the same factors

SUFFIX COMPETITION

ANALYSING SUFFIX COMPETITION

- **Problems with existing method**:
  - Comparisons over time difficult; *x*-axis = corpus size, not time period
  - Only measures variation within a morpheme, not between morphemes
- Towards a solution:
  - Force **time on the *x*-axis** and see what it requires from the method
  - Compare competing morphemes as if they formed a **linguistic variable**
    - – Calculate proportion of -*ity* types out of all -*ity* and -*ness* types

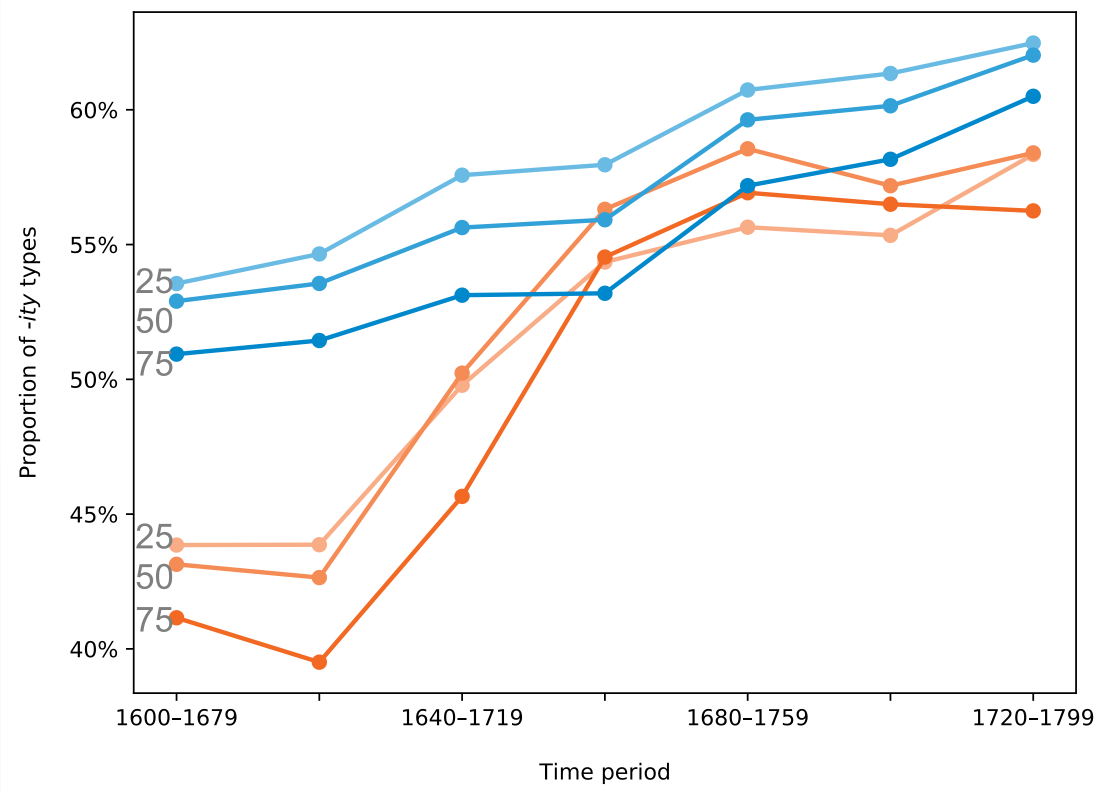
FIRST ATTEMPT



- **Blue** = men, **orange** = women
- 80-year sliding window,
  20-year intervals

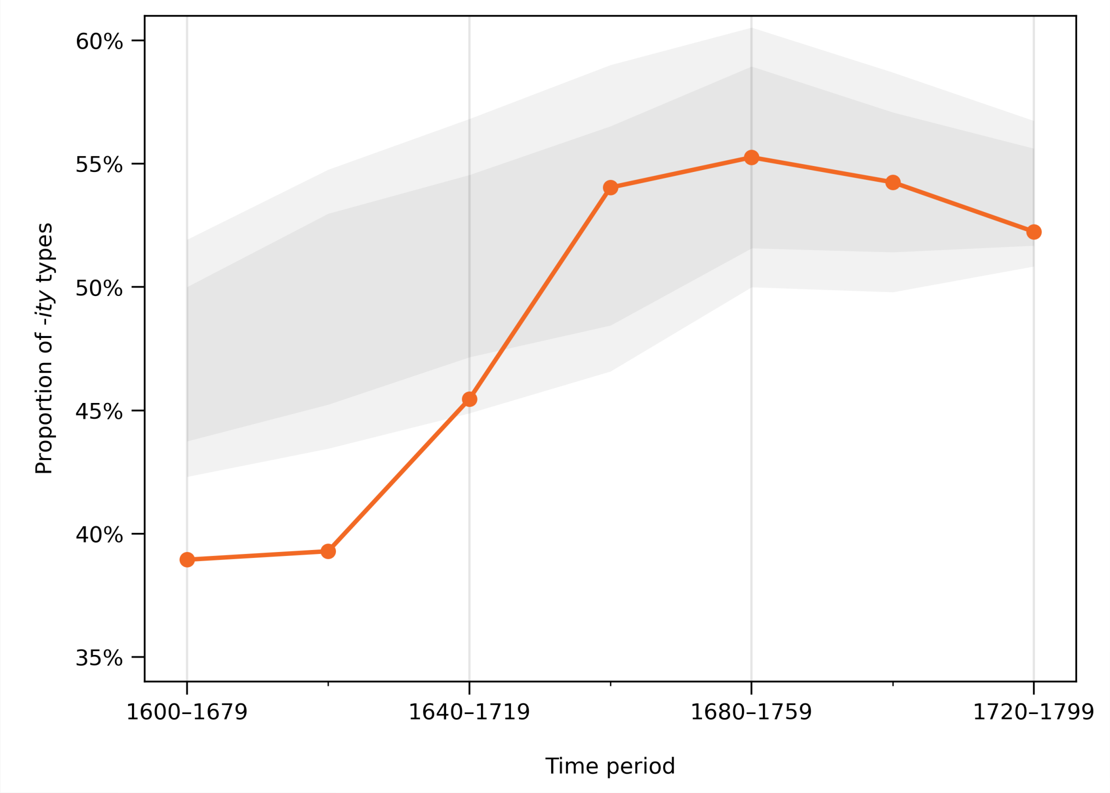
- **Problems**:
  - Different amounts of data from genders → comparability?
    - – Turns out that *proportions of types* grow nonlinearly with corpus size, too! ☹
  - Statistical significance?

TAKE SAMPLES OF EQUAL SIZE FROM GENDER-BASED SUBCORPORA



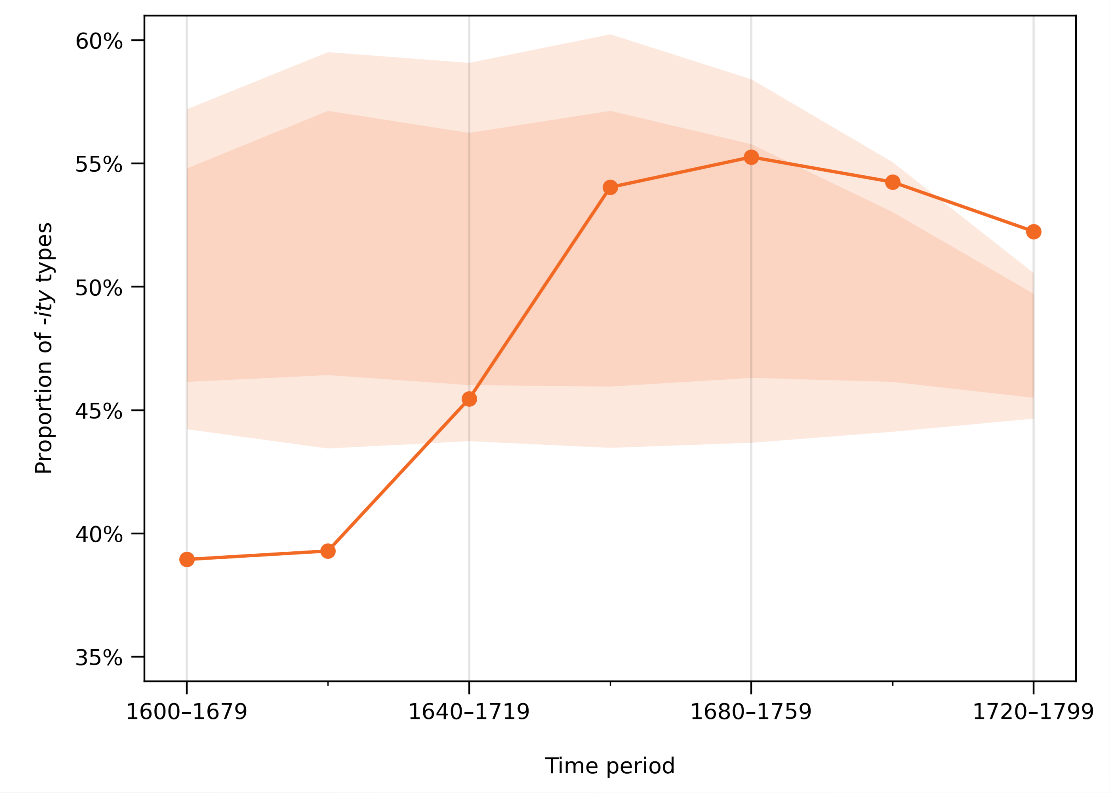
- 3 corpus sizes: a total of 25/50/75 -*ity*/-*ness* types
- Proportion of -*ity* increases over time
  - **Men**: steady growth
  - **Women**: lag behind at first, then quickly catch up
    - – But is this statistically significant?

SIGNIFICANCE OF GENDER DIFFERENCES



- Compare e.g. women of each period with randomly composed subcorpora of the same period
- Women = orange, random = grey

SIGNIFICANCE OF CHANGE OVER TIME



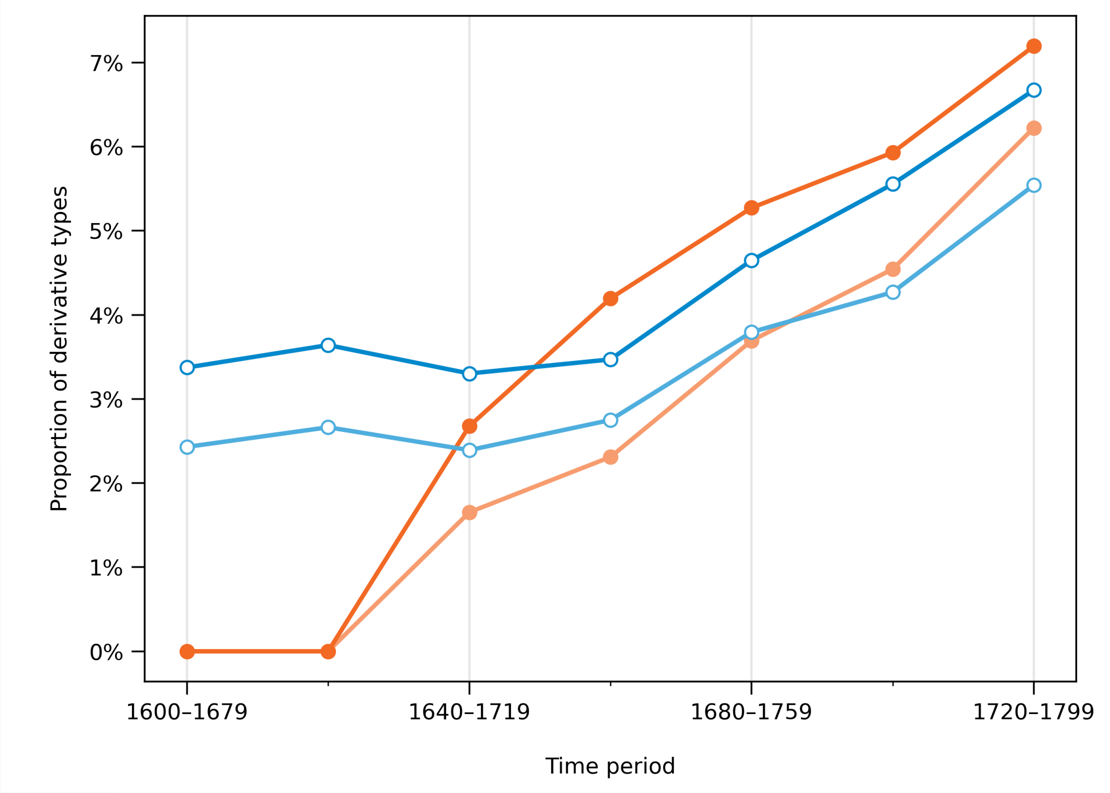
- Compare e.g. women of each period with randomly composed subcorpora of women of all periods
- Women = orange
- Säily et al. (in preparation)

INTERNAL FACTORS

FACTORS ANALYSED

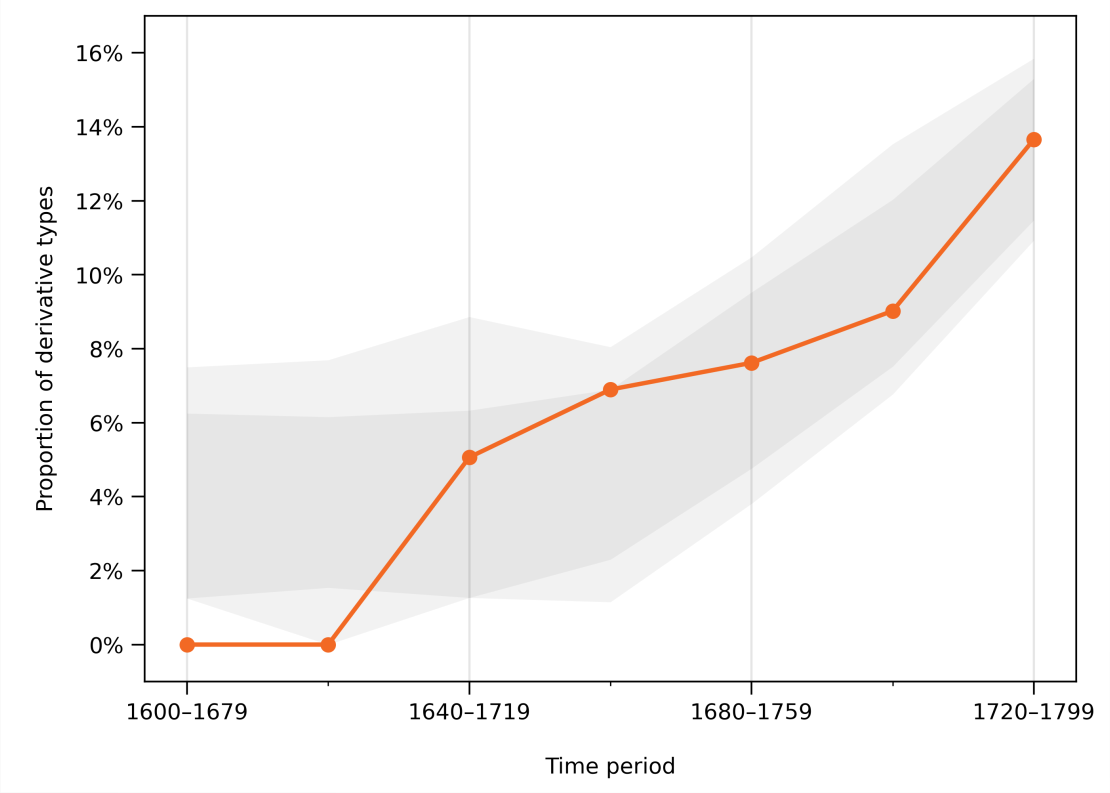
- **Etymology** (borrowing / derivative); OED
  - e.g. *ability* borrowing, *oddity* derivative
- **Base POS** (usually adjective but others possible as well); OED
  - e.g. *ability*: *able* ADJ, *authorshipness*: *authorship* NOUN
- **Branching structure** (binary / left / right); Hilpert (2013)
  - e.g. [*odd–ity*] binary, [[*un–couth*]*–ness*] left, [*non–*[*conform–ity*]] right

ETYMOLOGY



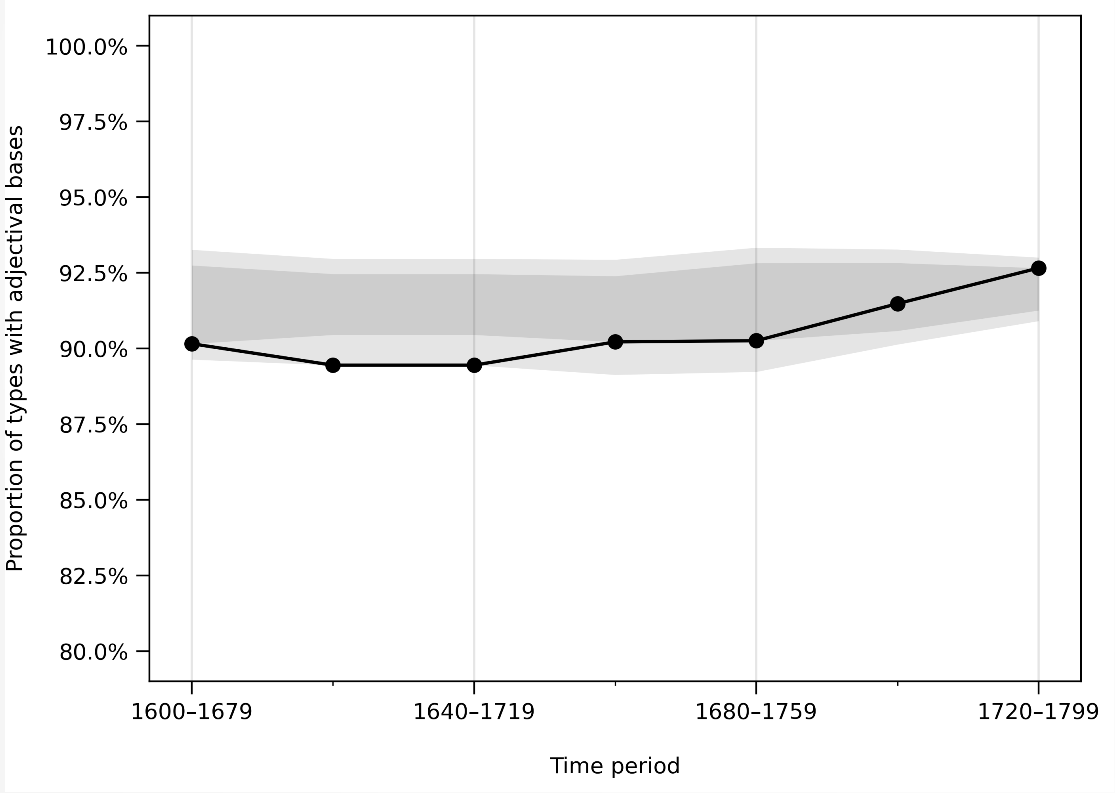
- *-ity*: women lag behind during C17
  - Then quickly catch up with men, and the proportion of **derived** types only really starts to grow when women join men in using them

ETYMOLOGY



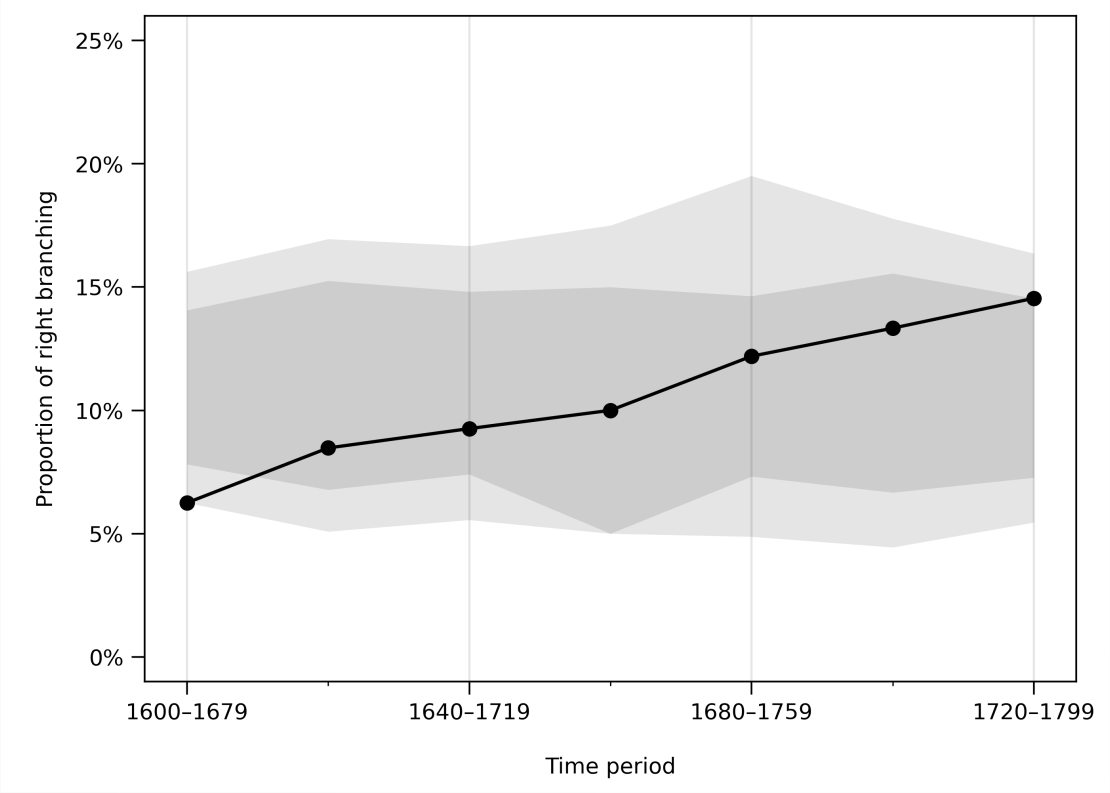
- *-ity*: women lag behind during C17
  - Then quickly catch up with men, and the proportion of **derived** types only really starts to grow when women join men in using them
  - 1st period: lag statistically significant ($p < 0.02$)

BASE POS



- No statistically confirmable trends by gender
- *-ity*: slight increase in share of **adjectival** bases over time
  - Last period: most *-ity* types with non-adjectival bases are earlier borrowings or right-branching

BRANCHING STRUCTURE



- No statistically confirmable trends by gender
- -*ness*: slight increase in share of **right-branching** types over time

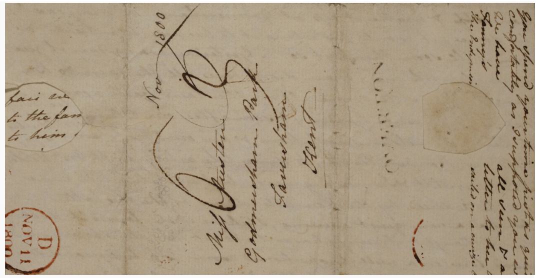
RESULTS

- Etymology
  - *-ity*: share of types derived within English increases over time, women lag behind in C17; *-ness*: no change
- Base POS
  - *-ity*: share of adjectival bases increases over time; *-ness*: no clear change
- Branching structure
  - *-ity*: no clear change; *-ness*: share of right-branching, prefixed types increases over time

CONCLUSIONS

- Results support and refine earlier findings
  - **Male-led increase in the productivity of *-ity*** also in relation to *-ness*, more information on diachronic development
- Internal factors, too, point towards increasing productivity of *-ity*
  1. Increase in the share of types originally derived within English
  2. Increase in the share of adjectival bases (types with other bases tend to be borrowed)
  - CxG: 2 surprising – increase in productivity expected to entail use in more contexts, not fewer
- Quantifying variation and change in word types is hard but worth it!

REFERENCES



- Baayen, R.H. 1993. On frequency, transparency and productivity. G. Booij & J. Van Marle (eds.), *Yearbook of Morphology 1992*, 181–208. Kluwer.
- Bolinger, D. 1948. On defining the morpheme. *Word 4:*18–23.
- Cowie, C. & C. Dalton-Puffer. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. J.E. Díaz Vera (ed.), *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*, 410–437. Rodopi.
- Hilpert, M. 2013. *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. CUP.
- Rodríguez-Puente, P. 2020. Register variation in word-formation processes: The development of *-ity* and *-ness* in Early Modern English. *International Journal of English Studies* 20(2):145–167.
- Rodríguez-Puente, P., T. Säily & J. Suomela. 2022. New methods for analysing diachronic suffix competition across registers: How *-ity* gained ground on *-ness* in Early Modern English. *International Journal of Corpus Linguistics* 27(4): 506–528*.*
- Säily, T. 2014. *Sociolinguistic Variation in English Derivational Productivity: Studies and Methods in Diachronic Corpus Linguistics*. Société Néophilologique.
- Säily, T., M. Hilpert & J. Suomela. In preparation. New approaches to investigating change in derivational productivity. *Proc. ICAME 42*.
- Säily, T. & J. Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. A. Renouf & A. Kehoe (eds.), *Corpus Linguistics: Refinements and Reassessments*, 87–109. Rodopi.
- Säily, T. & J. Suomela. 2017. *types*2: Exploring word-frequency differences in corpora. T. Hiltunen, J. McVeigh & T. Säily (eds.), *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*. VARIENG. https://varieng.helsinki.fi/series/volumes/19/saily_suomela/