Sociolinguistic variation in morphological productivity

in 18th-century English

Introduction

- I study variation and change in the productivity of two affixes from Early Modern to present-day English
- -*ness* and -*ity*: roughly synonymous, derive abstract nouns from adjectives
  - *prescriptive* + -*ness* → *prescriptiveness*
    prescriptive + -*ity* → *prescriptivity*
- -*ness* native, -*ity* borrowed from French in the Middle Ages (later reinforced through Latin)
  - Sociolinguistically interesting: English 'diglossia'
- Do e.g. men and women use these differently?

Morphological productivity

- “The statistically determinable readiness with which an element enters into new combinations” (Bolinger 1948: 18)

- Also: the extent of use of the element

Easiest measure: Type frequency

- The number of different words of a particular morphological category found in a corpus
- ! Some of the words could be centuries old
  - Lexicalisation: *business*
- ! *-ity*: some of the words originally borrowings, e.g. *generosity* < Latin *generōsitās*
  - If the base exists in English, the user can still form the word from its components: *generous* + *-ity*

Related concepts

- **Token frequency**: the number of all words of a particular morphological category found in a corpus
  - Each occurrence is counted, even if the same word has been encountered before
  - No good as an indicator of productivity on its own
    - Easily inflated by a small number of very common types
- **Hapax frequency**: the number of words of a particular morphological category occurring only once in a corpus

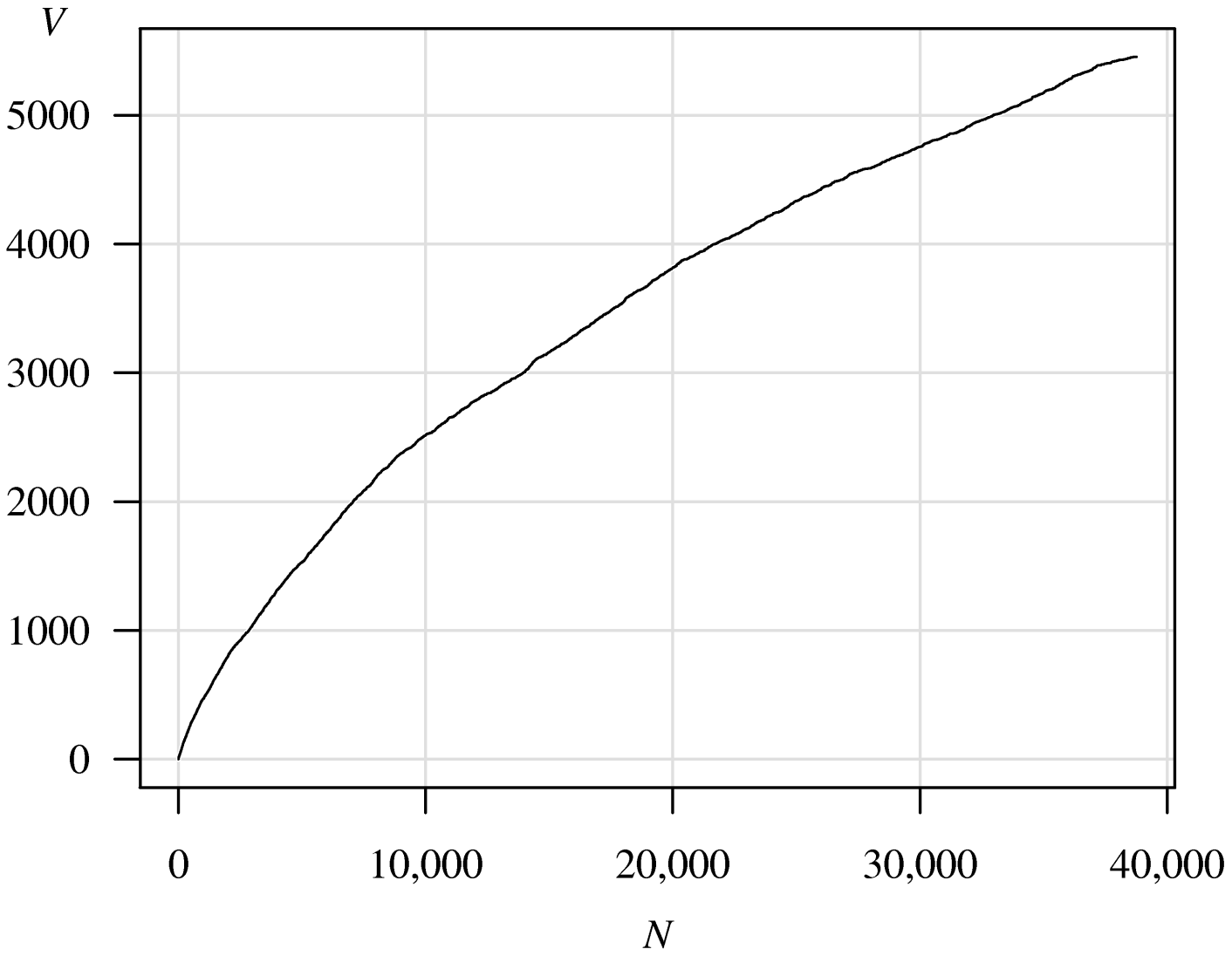
Baayen's measures (1990s →)

- Based on the above-mentioned three concepts
  - Types *V*, tokens *N*, hapaxes n_1
- **Realised productivity** *V*
  - Extent of use
- **Potential productivity** *P*
  - Category-conditioned degree of productivity
- **Expanding productivity** *P**
  - Hapax-conditioned degree of productivity

Realised productivity *V*

- *V* = type frequency
- Estimates the size of the morphological category
- Grows with token frequency, but not at the same rate → problem if we want to compare figures
  - To enable comparisons between type frequencies in subcorpora of, say, men and women, token frequencies should be the same in both subcorpora
- Comparing normalised figures (e.g. *V* per 10,000 words) will not work because *V* does not grow linearly

A type accumulation curve



Potential productivity *P*

- $P = n_1/N$
  (hapax frequency / token frequency)
- Estimates the growth rate of the morphological category
  - When we encounter a word belonging to the category, how likely is it to be new?
- Could be drawn as the tangent to the endpoint of the type accumulation curve
- Depends on the size of the corpus!

Expanding productivity *P**

- $P^* = n_1/h$
  (hapax frequency / number of hapaxes in the whole corpus; within the same corpus, n_1 counts alone are sufficient)
- Estimates the contribution of the morphological category to the overall vocabulary growth
  - When we encounter a new word, how likely is it to belong to the morphological category?
- Depends on the size of the corpus!

Problem of comparison

- All three measures depend on corpus size
- Corpus size can be defined as:
  - The number of running words
  - The token frequency of the morphological category
- How can we compare the productivity of men and women if the sizes of the subcorpora are different?

Solution 1: Gaeta & Ricca (2006)

- Improvement on *P*: compare hapax frequencies at the same token frequency
  - E.g. if more data from men than women, take a random sample from men containing the same number of tokens as the women's subcorpus, and calculate $P = n_1/N$ for the sample
- Ok, but this means getting rid of valuable data
- If we do observe a difference between *P* figures, how will we know if it is statistically significant?

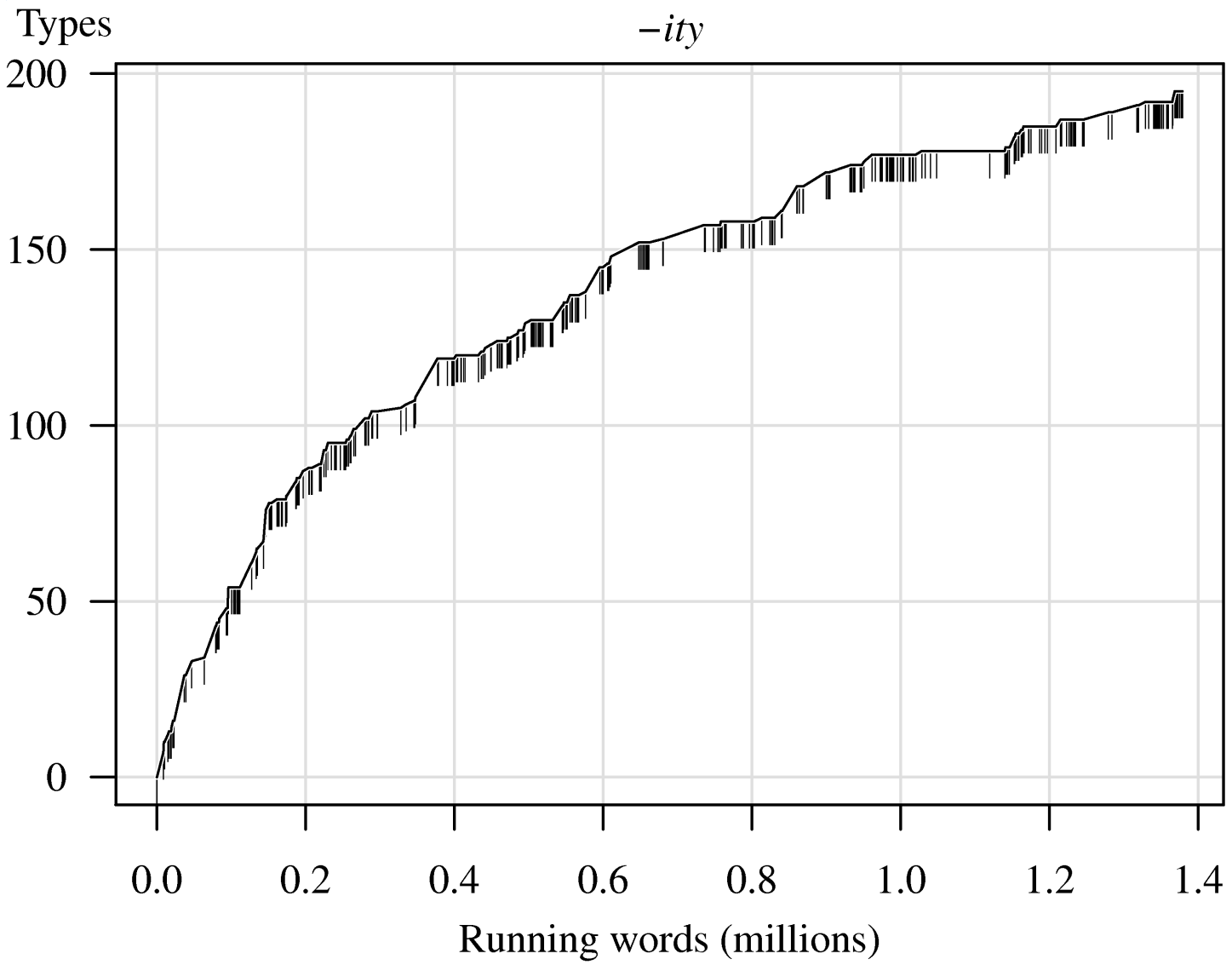
Solution 2: Säily & Suomela (2009)

- Based on type accumulation curves and permutation tests
- Type or hapax frequencies from subcorpora are compared with the whole corpus
- No simplifying assumptions, no discarded data
- Easy to determine statistical significance

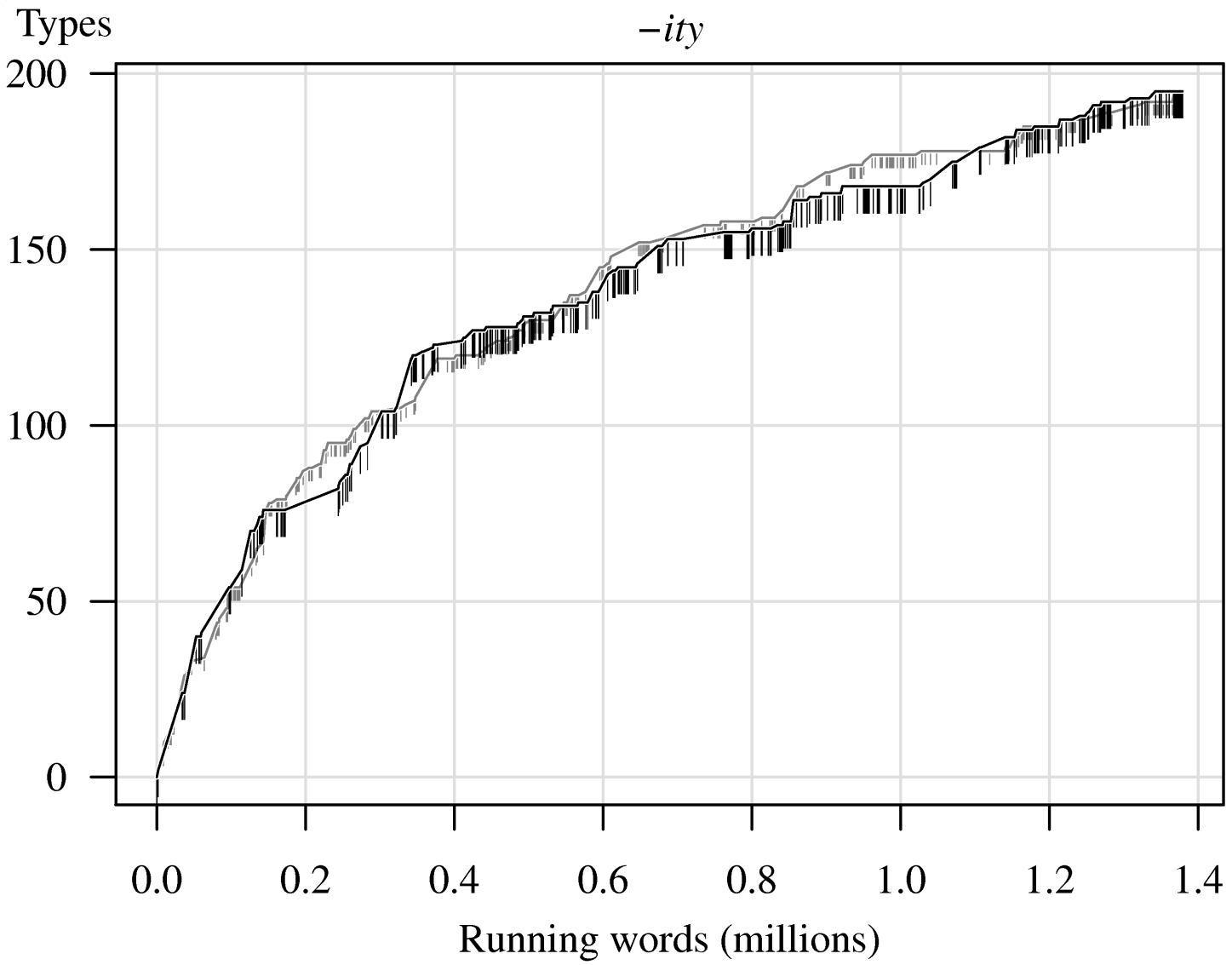
How to do it

- Divide the corpus into samples large enough to preserve discourse structure (e.g. 1 text = 1 sample)
- Pick a sample randomly and calculate the number of types in it
  - Plot the sample on a figure with the size of the sample on the *x* axis and type frequency on the *y* axis
- Pick another sample, add it to the previous one, and calculate the combined number of types
  - Plot the result on the same figure
- Repeat until all samples have been picked

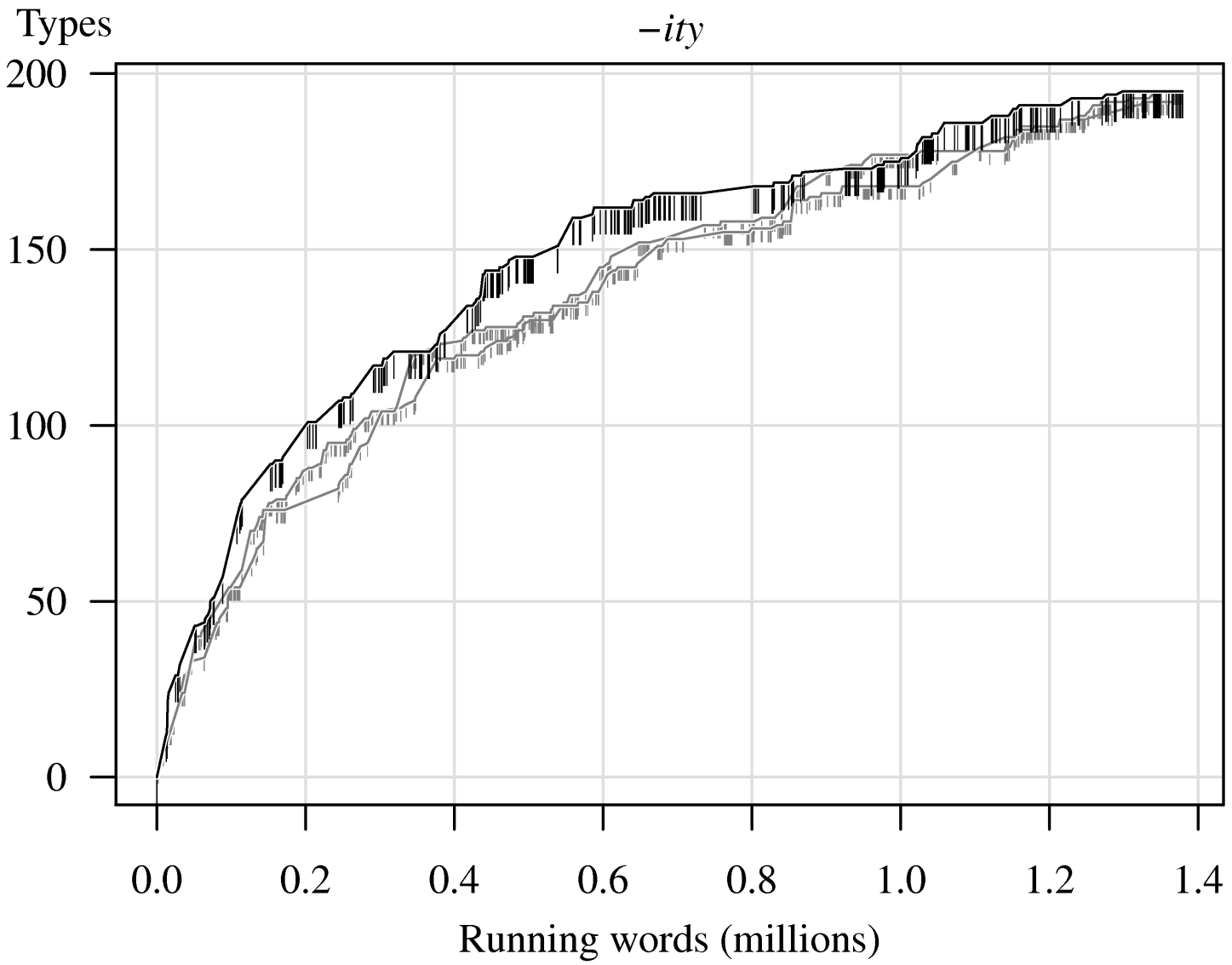
A random type accumulation curve



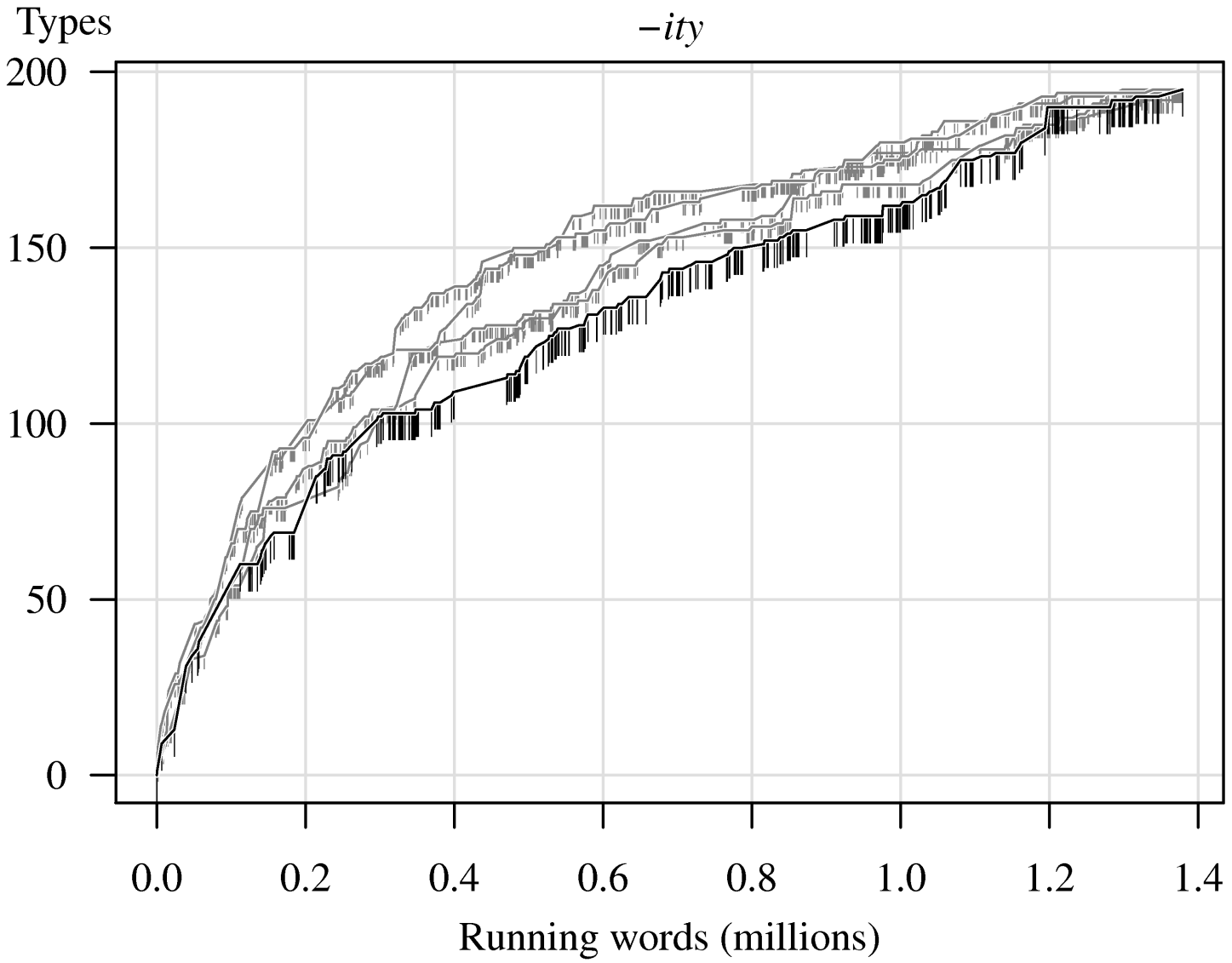
Then: repeat the process…



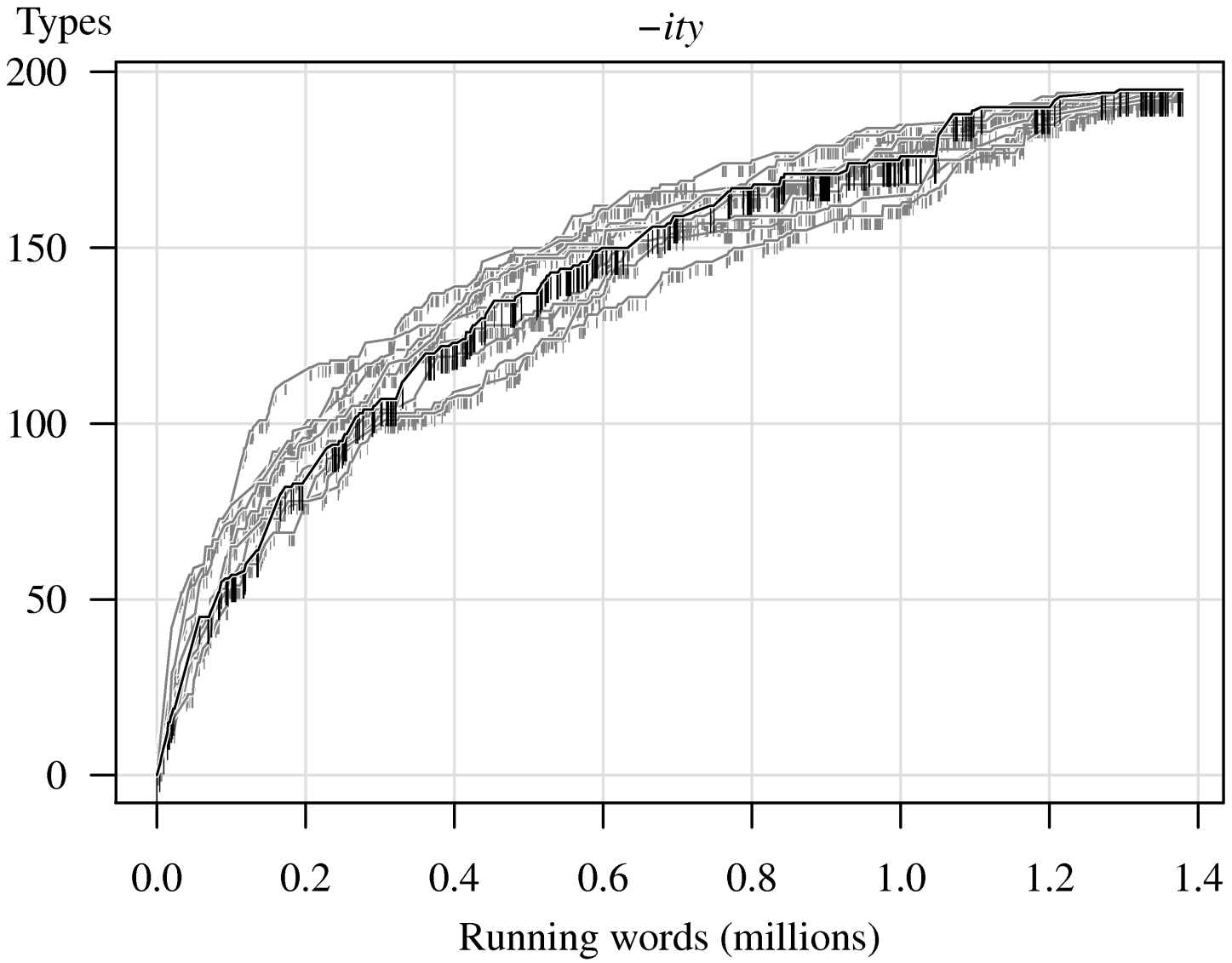
Then: repeat the process…



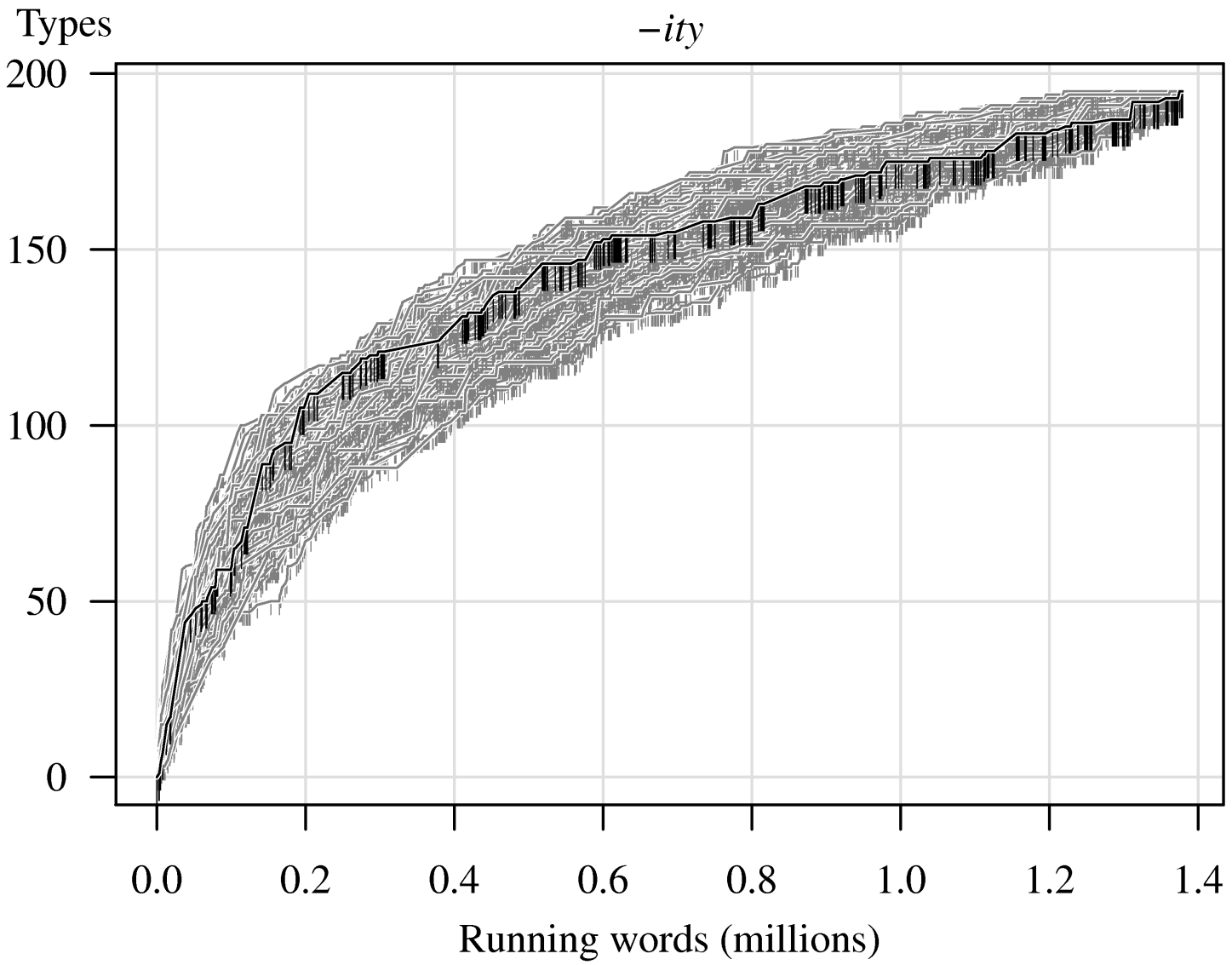
Then: repeat the process…



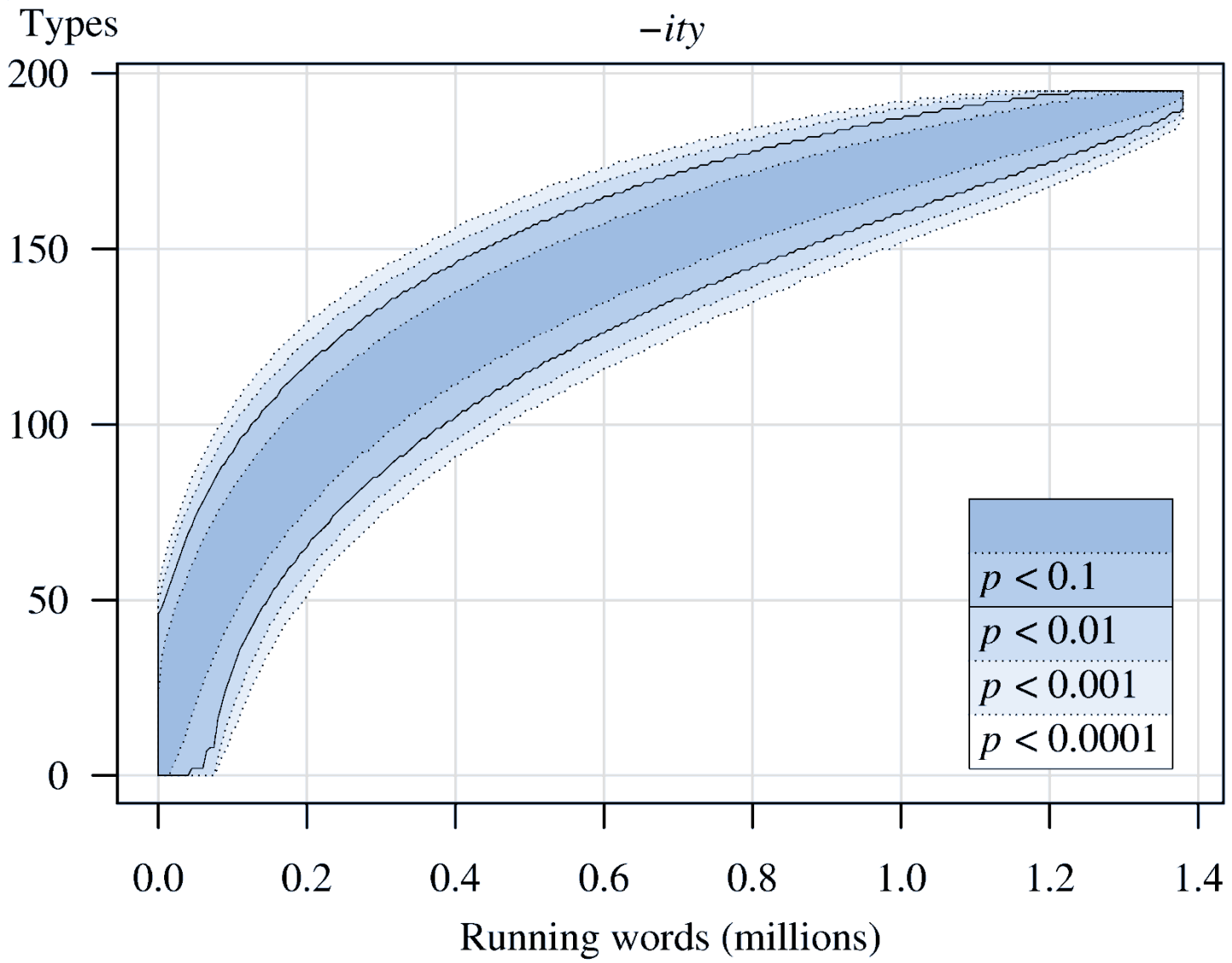
Then: repeat the process…



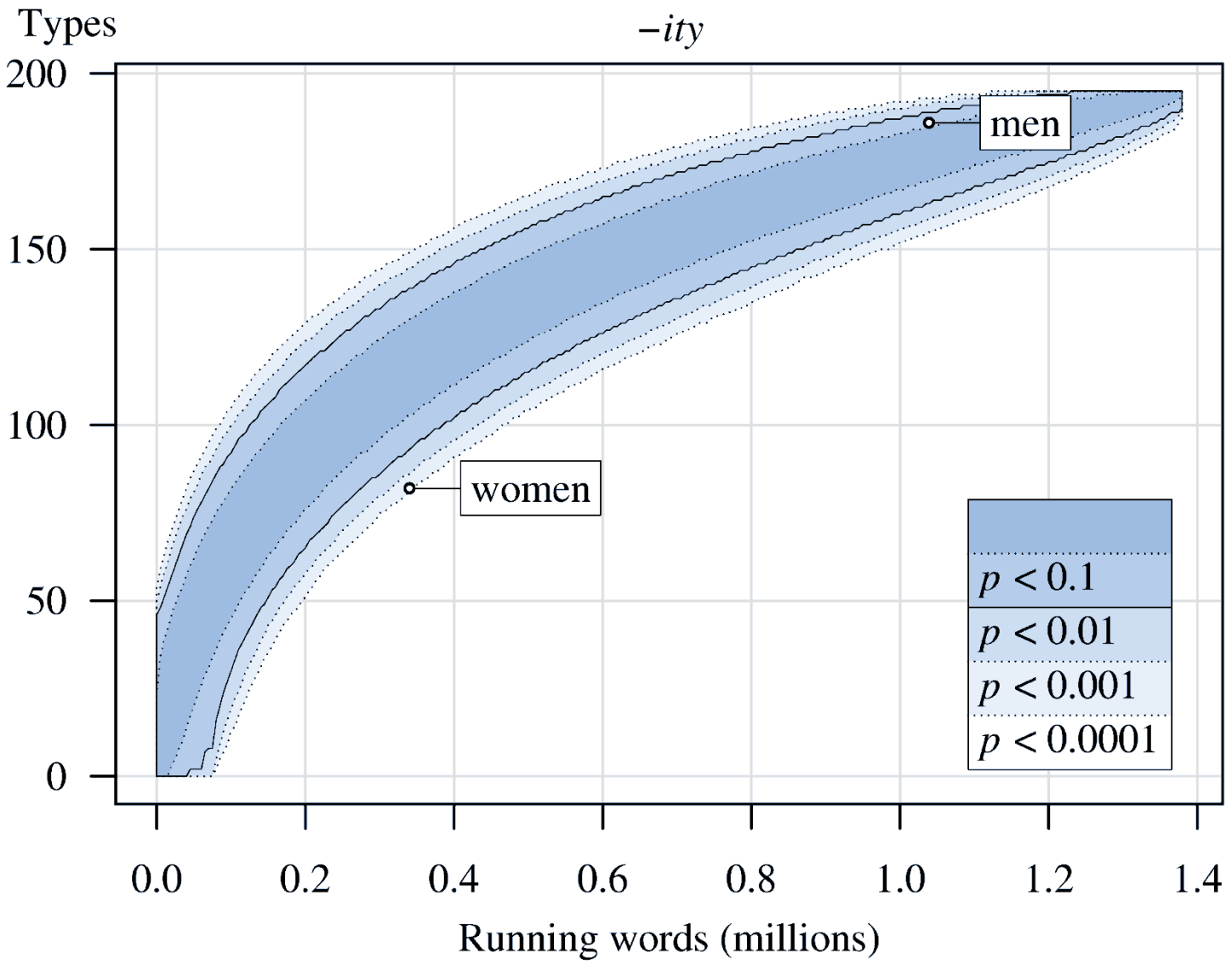
Then: repeat the process…



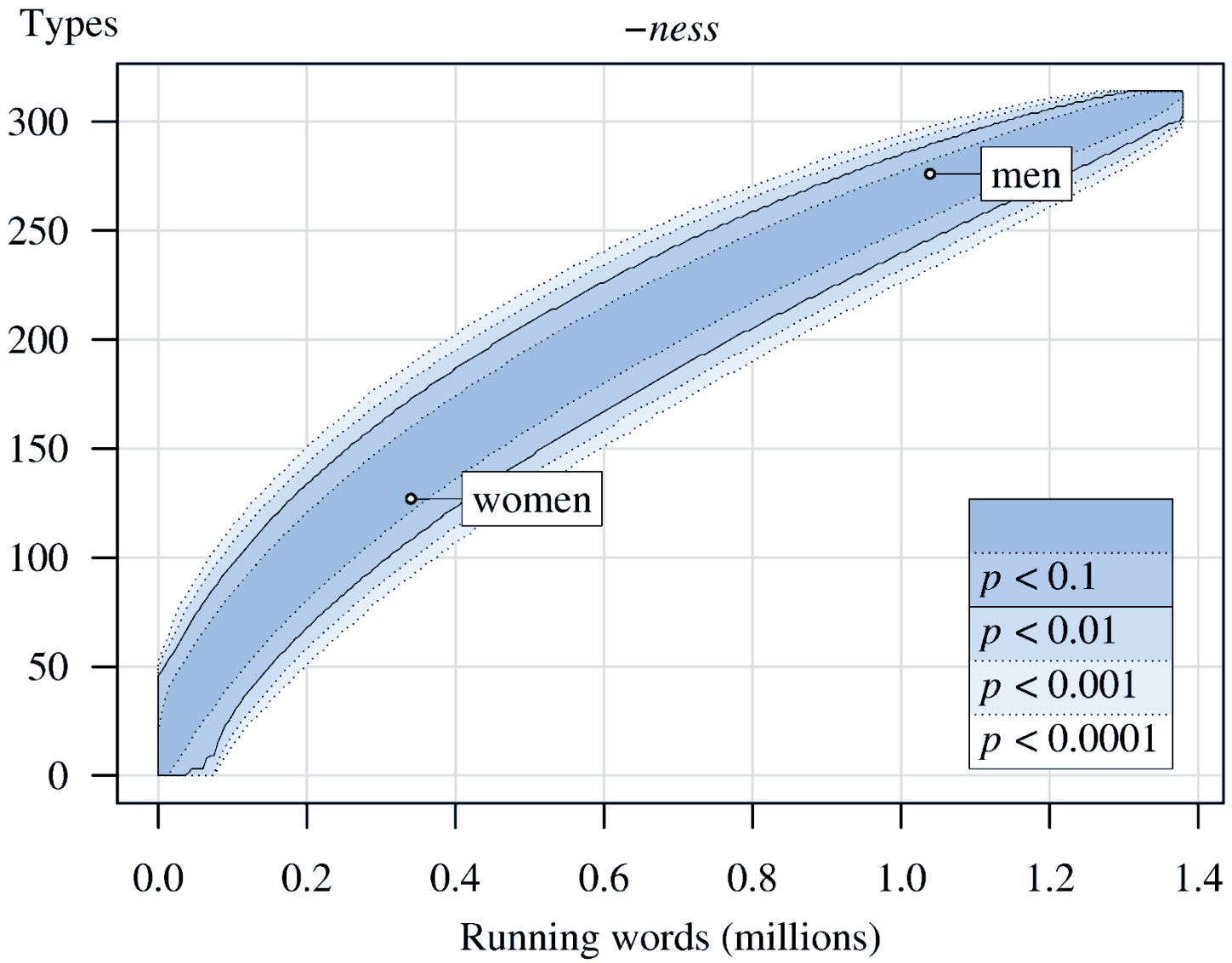
… a million times!



Plot subcorpora and compare



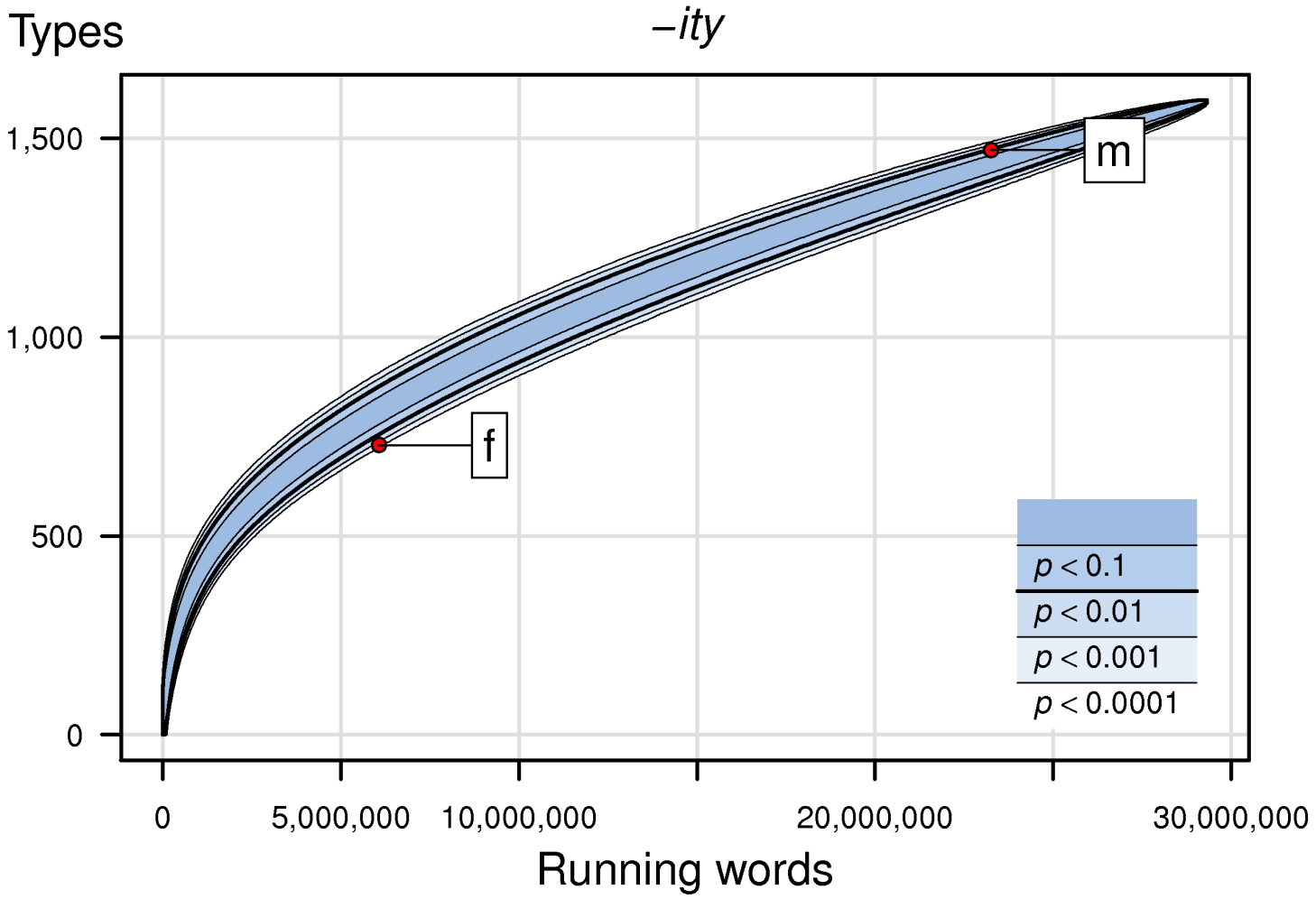
The same for -*ness* types



Results

- The frequency of women's *-ity* types is statistically significantly low compared to the corpus as a whole
  - *Corpus of Early English Correspondence* (CEEC), 17th century
- *-ness*: no statistically significant differences
- Measure used here: **type frequency** as a function of the number of **running words**
  - Type frequency as a function of token frequency: results similar but less significant
- PDE: similar results – stable gendered styles?
  - *British National Corpus* (BNC)

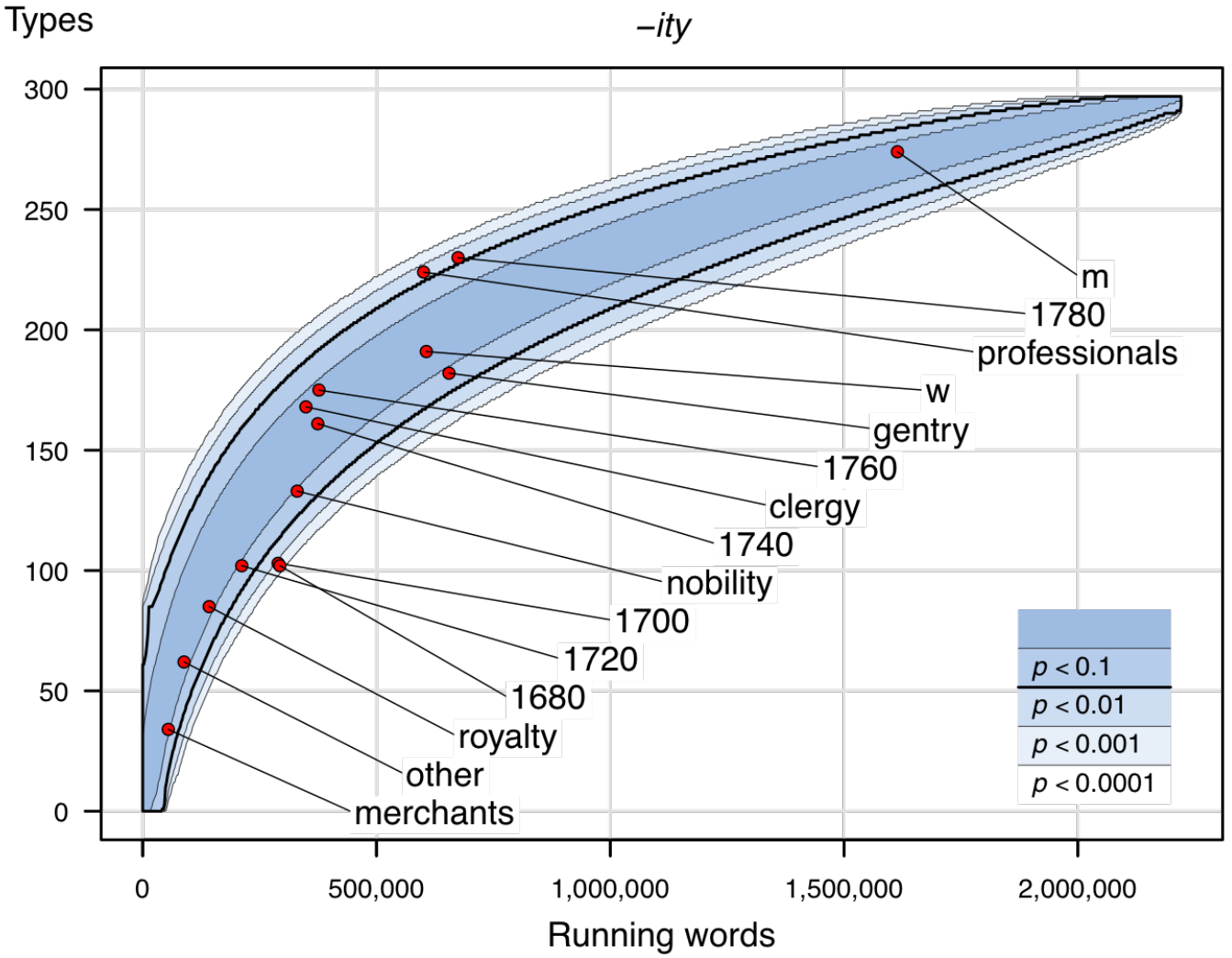
BNC, informative written texts



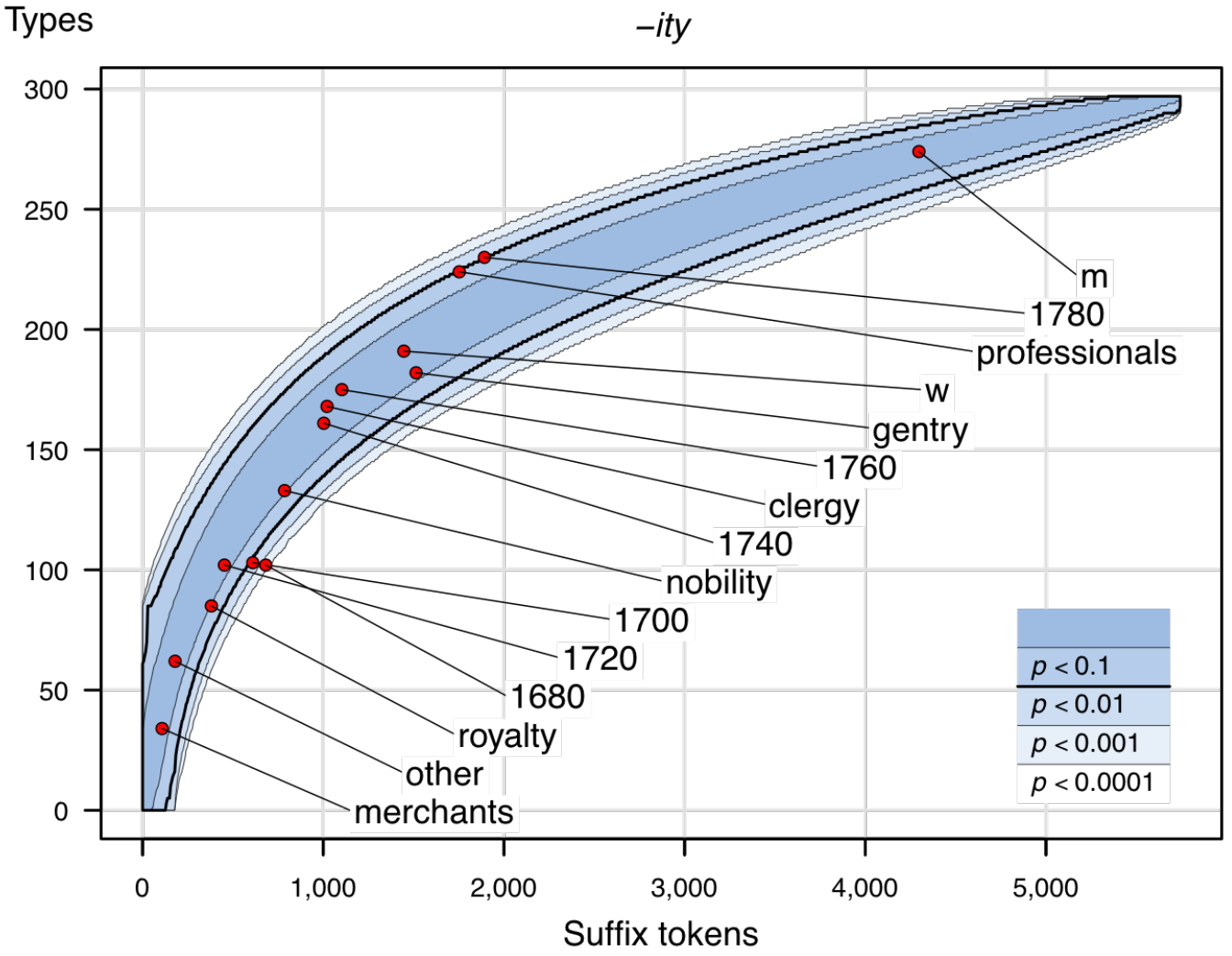
18th century: CEECE

- *Corpus of Early English Correspondence Extension* (CEECE), 1680–1800
- Initial results on *-ity*
  - No gender difference!
  - Productivity seems to increase over time
- Initial results on *-ness*
  - Nothing stands out as statistically significant
  - However, if we use token frequency rather than the number of running words on the *x* axis:
    - Productivity significantly low with royalty, high with clergy

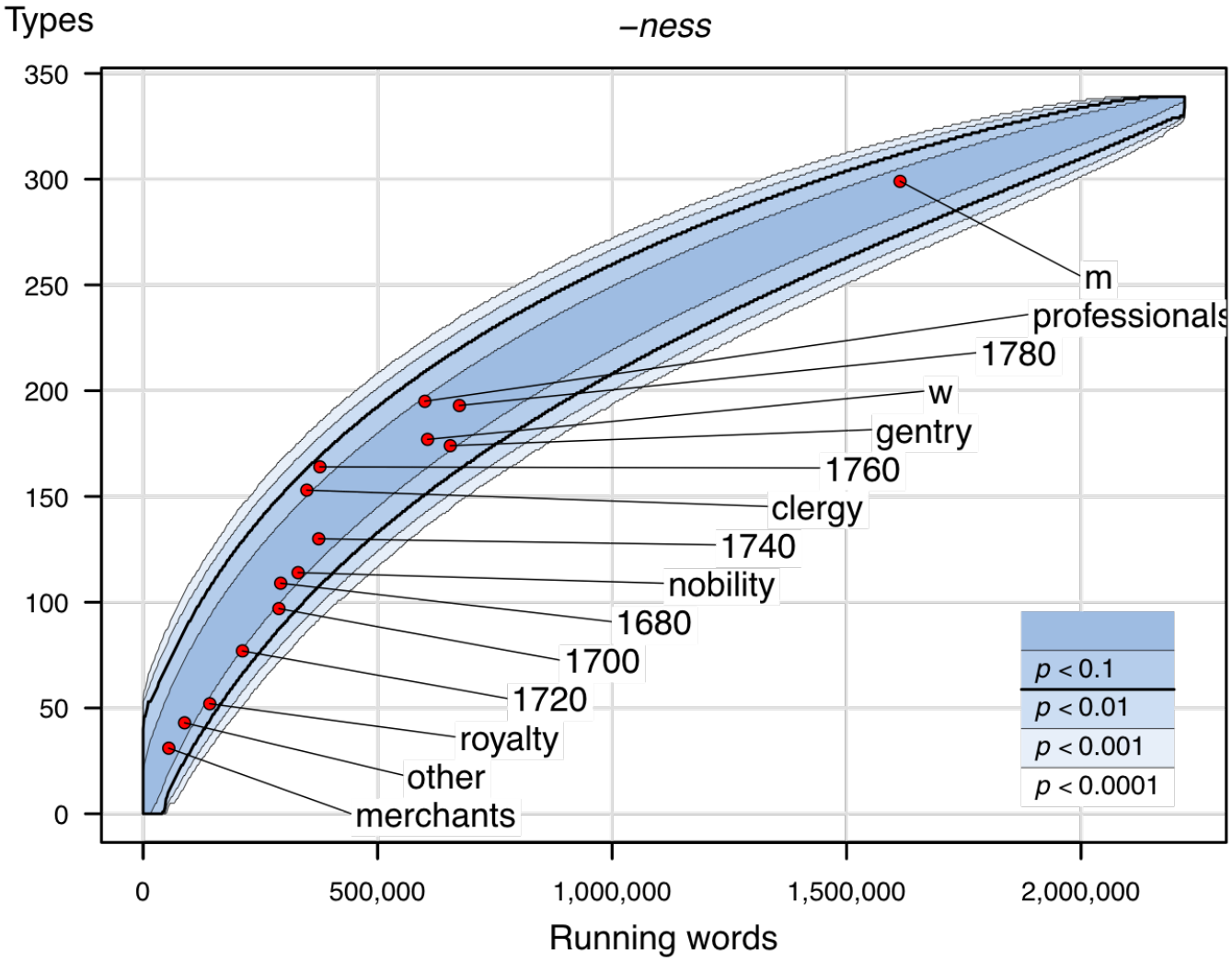
-ity types vs. running words



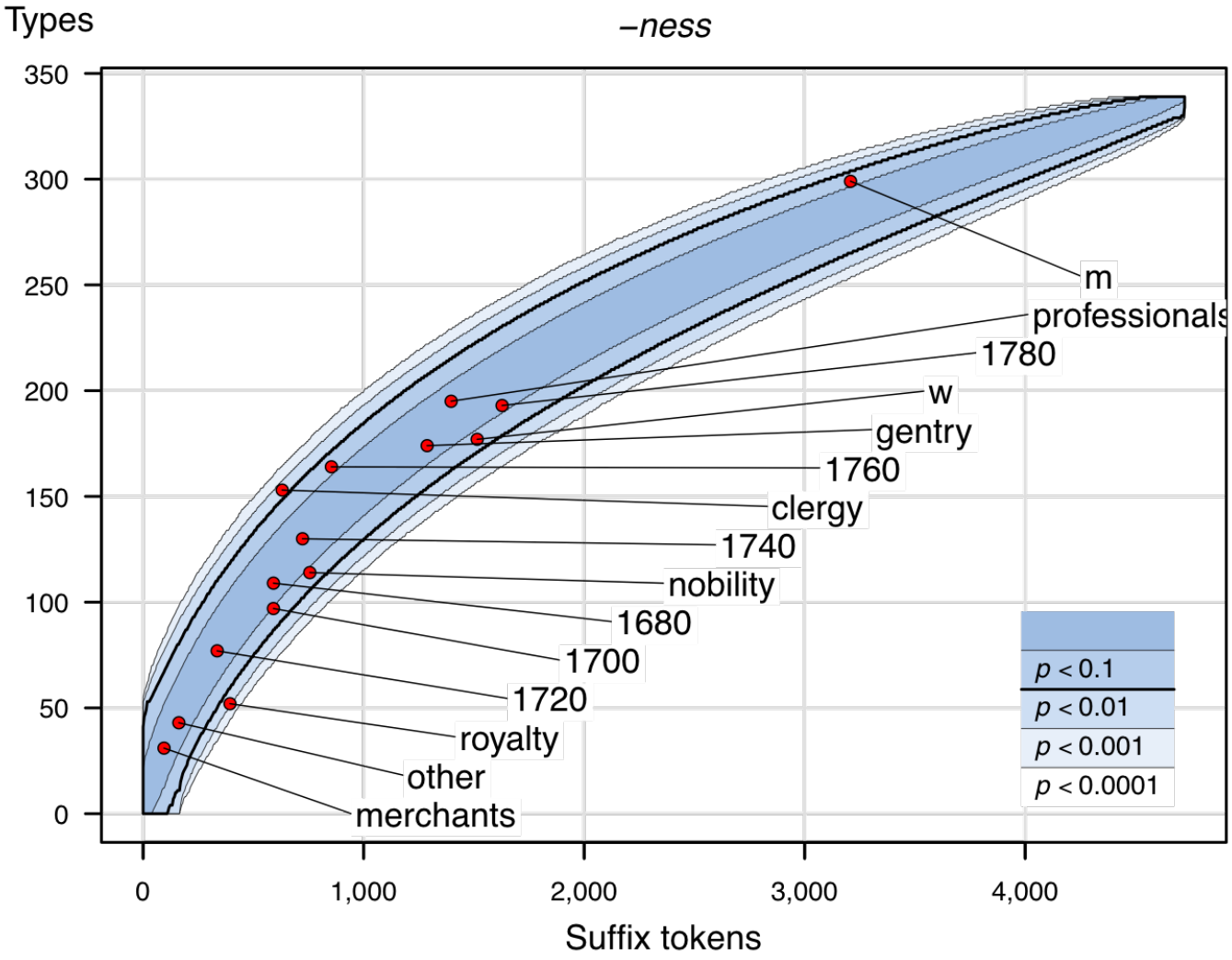
-ity types vs. tokens



-*ness* types vs. running words



-ness types vs. tokens



Examples: Royalty

[…] I must humbly beg of your Majesty to present my most respectful duty to the Queen, and if I might presume to request it of your Majesty, my most affectionate love to my sisters who, I trust, will ever join me in prayer to Heaven for your **happiness** & that of the Queen.

GEORG3A_010, Prince Edward to the King, 1785

[…] Your Majesty's **goodness** and **kindness** towards me gives me great hopes of this fortunate event soon taking place. My petition now is […]

GEORG3A_071, Prince Augustus to the King, 1795

Examples: Clergy

[…] He wrote me, on the occasion, 2 or 3 pages of most manly ***inside-outness*** & impartiality, such as hardly ever came, I believe, from any man but himself. […]

TWINING_044, Thomas Twining to his brother, 1788

[…] I beleive I drank too much wine last night at Hurstbourne; I know not how else to account for the shaking of my hand to day; – You will kindly make allowance therefore for any **indistinctness** of writing by attributing it to this venial Error. […]

AUSTEN_027, Jane Austen to her sister, 1800?

Examples: Professionals

[…] In hopes therefore of giving to my Book some marks of **originality** or at least of Novelty, I shall e'en go and allay my Thirst of Knowledge at the Source, […]

BURNEY_005, Charles Burney to Montagu North, 1770

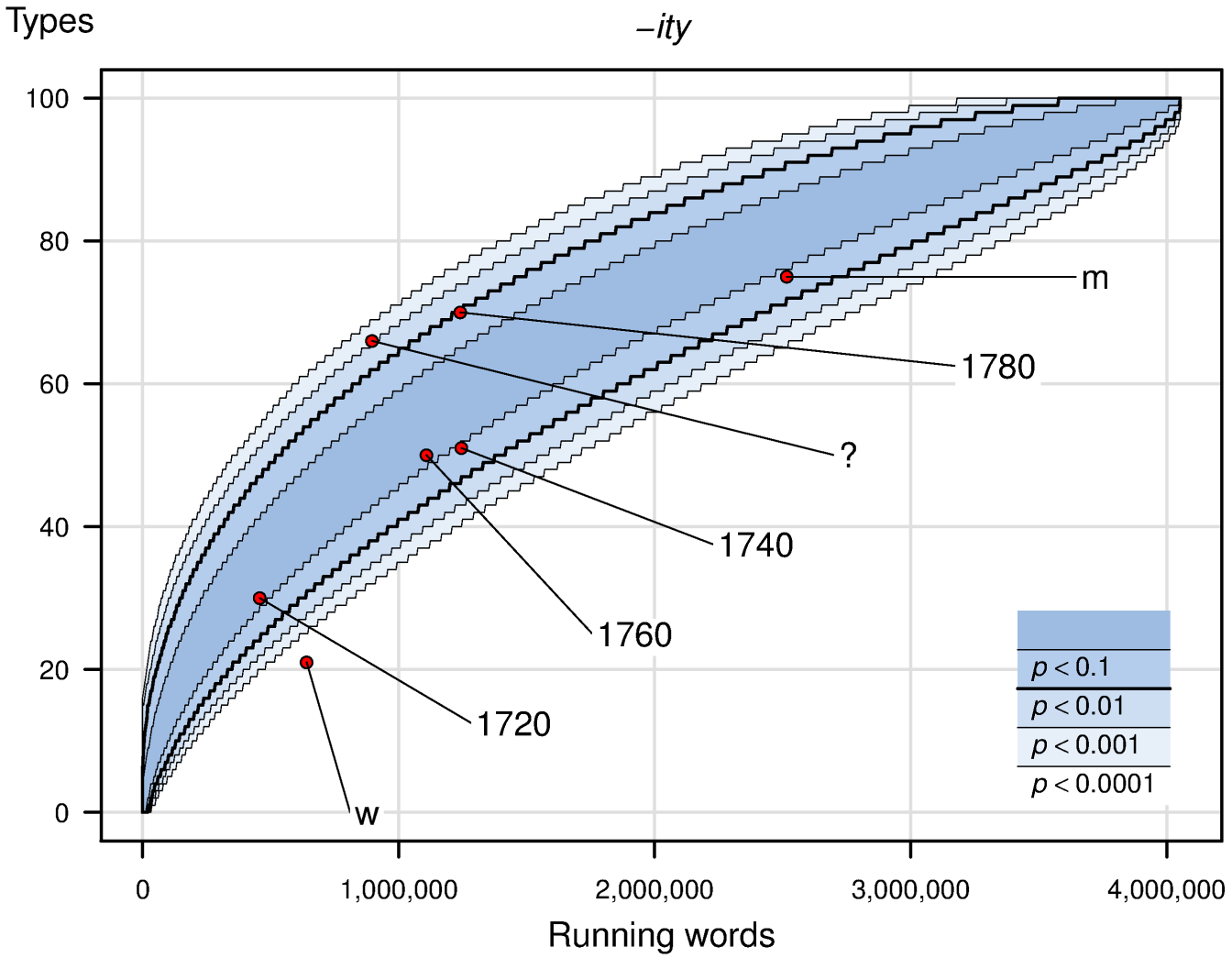
[…] Most of my fraternity would as soon shorten the noses of their children because they were said to be too long, as thus dock their compositions in compliance with the opinion of others. I beg that when my life shall be written hereafter my Authorship's **ductility** of temper may not be forgotten. […]

COWPERW_063, William Cowper to Walter Bagot, 1789

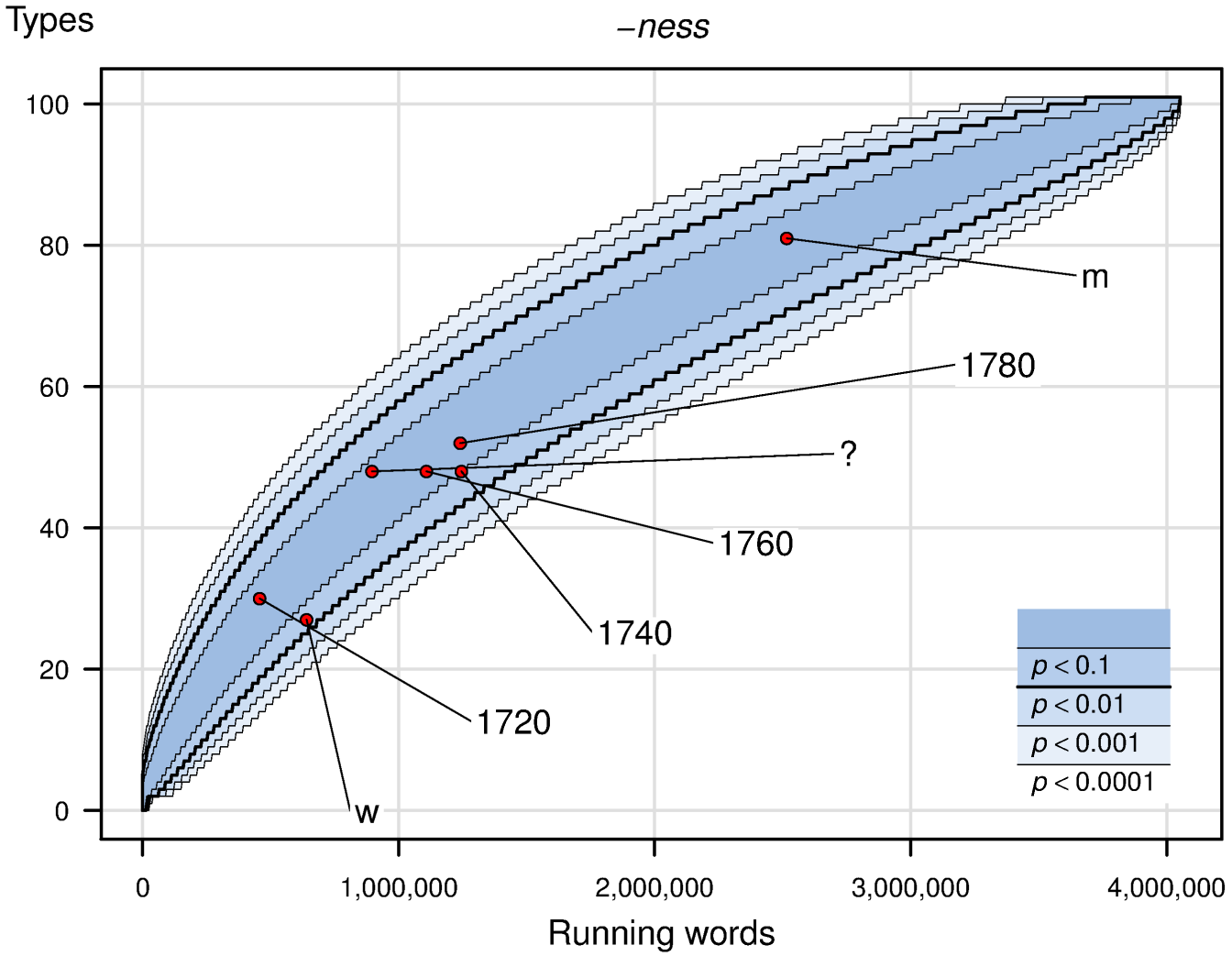
18th century: OBC

- *Old Bailey Corpus* (OBC), 1730–1800
  - Courtroom discourse, recorded by scribes and printed
- Initial results on *-ity*
  - Productivity significantly low with women
  - Productivity seems to increase over time
- Initial results on *-ness*
  - No statistically significant differences
- Unlike CEECE, results in line with CEEC and BNC
  - Supports the idea of stable gendered styles
    - 18th-century literati apparently different

Old Bailey Corpus, 1730–1800



Old Bailey Corpus, 1730–1800



Conclusion

- Study of productivity variation now possible even in small, unlemmatised historical corpora
  - Testing sociolinguistic hypotheses in the long diachrony
- Method is robust and assumption-free
  - Could be used as a benchmark for parametric models
- Possible improvements
  - Show both measures of corpus size simultaneously
  - Apply post-hoc analysis to results
  - Develop interactive tool
- Hapax legomena still unusable in smaller corpora…

References

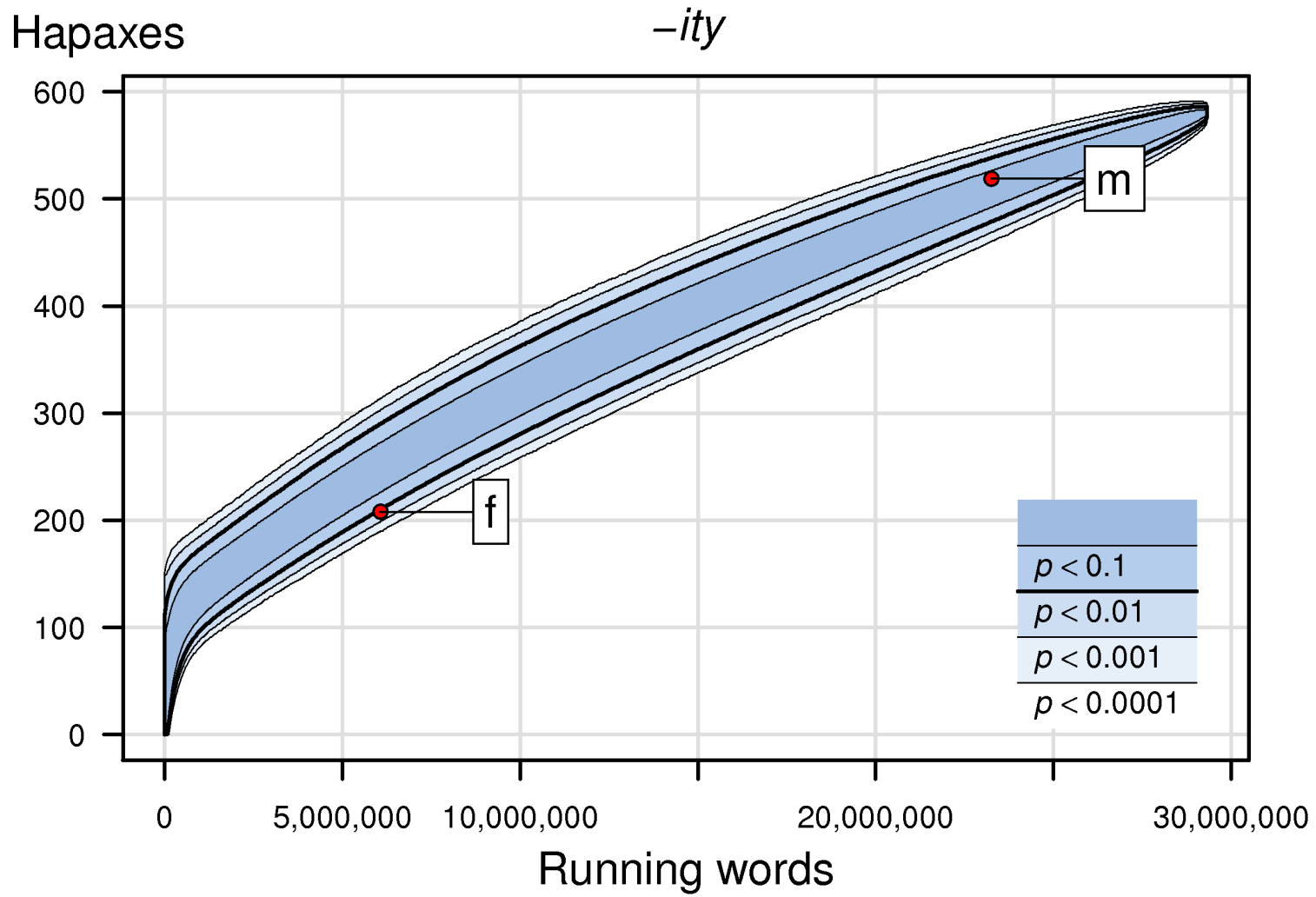
Baayen, R. H. 1993. On frequency, transparency and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.

Bolinger, Dwight L. 1948. On defining the morpheme. *Word* 4. 18–23.

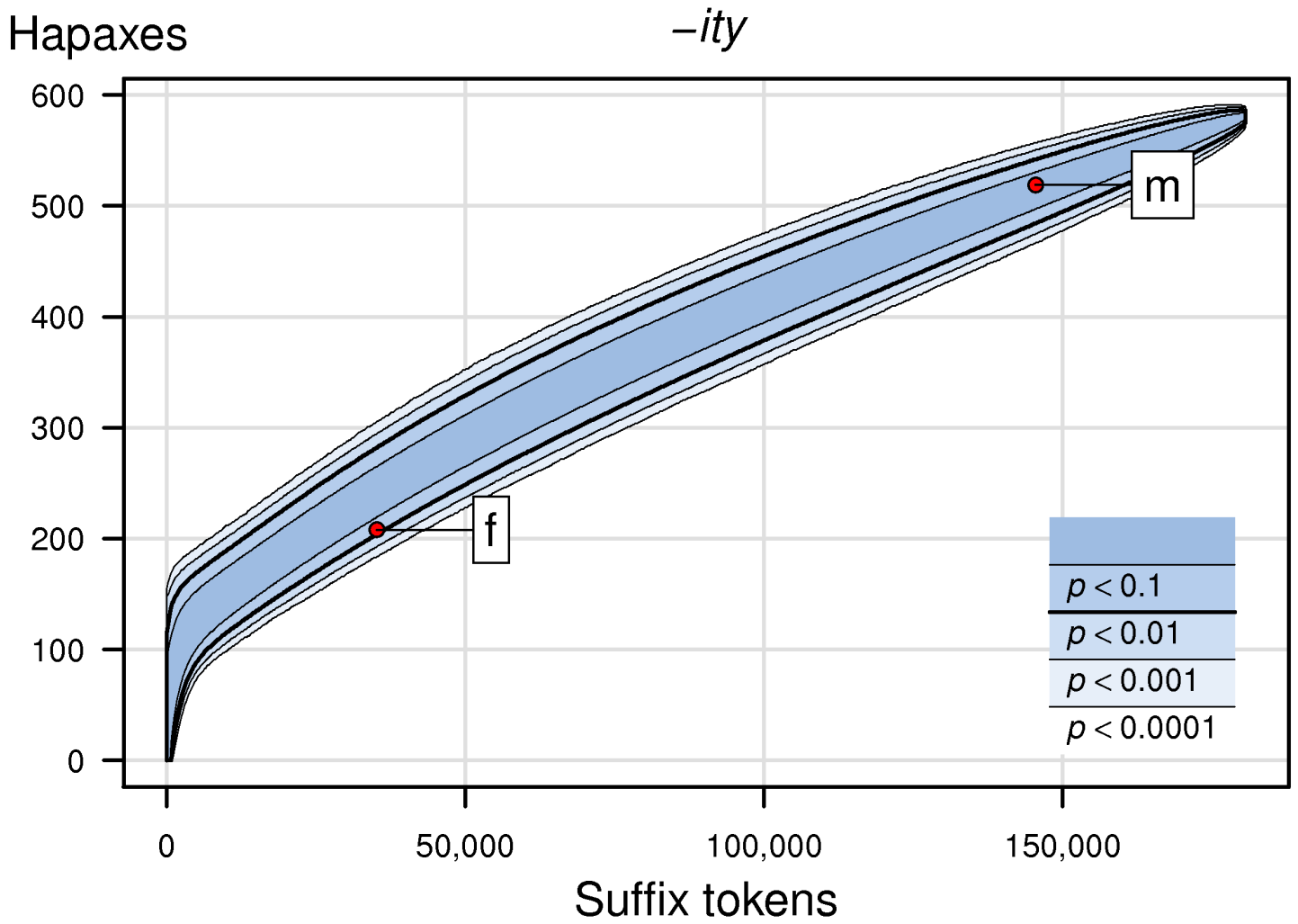
Gaeta, Livio & Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1). 57–89.

Säily, Tanja & Jukka Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. In Antoinette Renouf & Andrew Kehoe (eds.), *Corpus linguistics: Refinements and reassessments* (Language and Computers: Studies in Practical Linguistics 69), 87–109. Amsterdam: Rodopi.

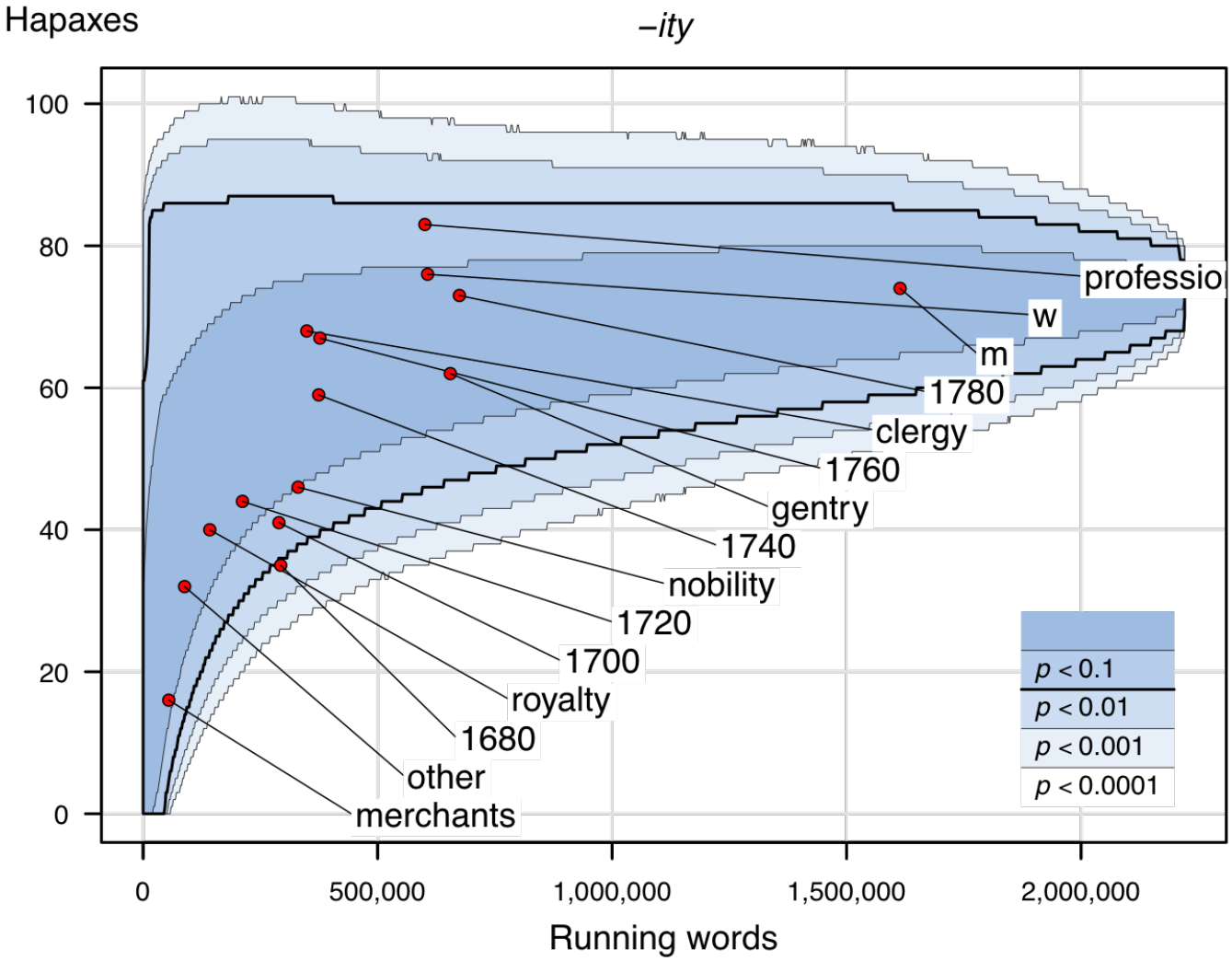
-ity hapaxes, BNC



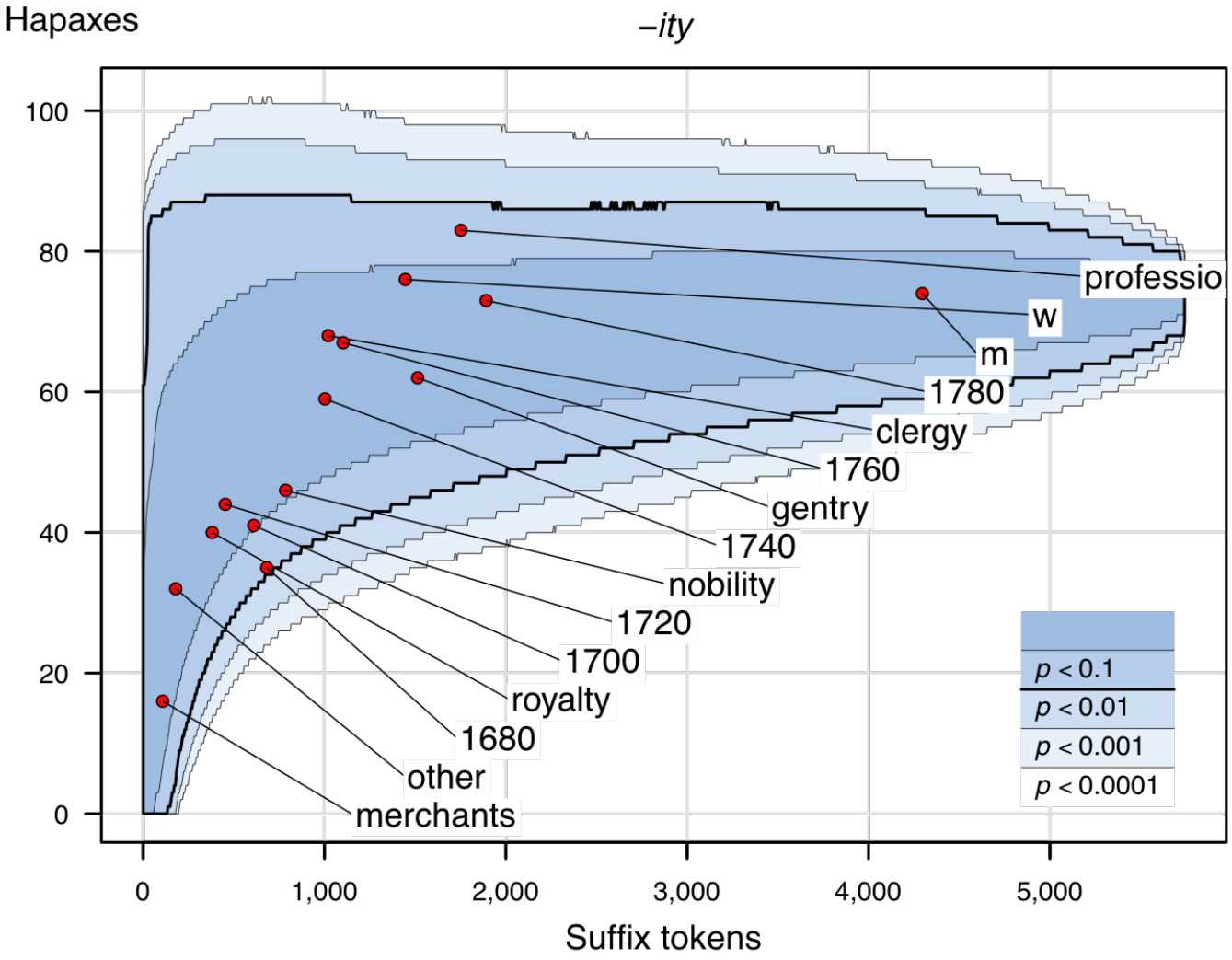
-ity hapaxes, BNC



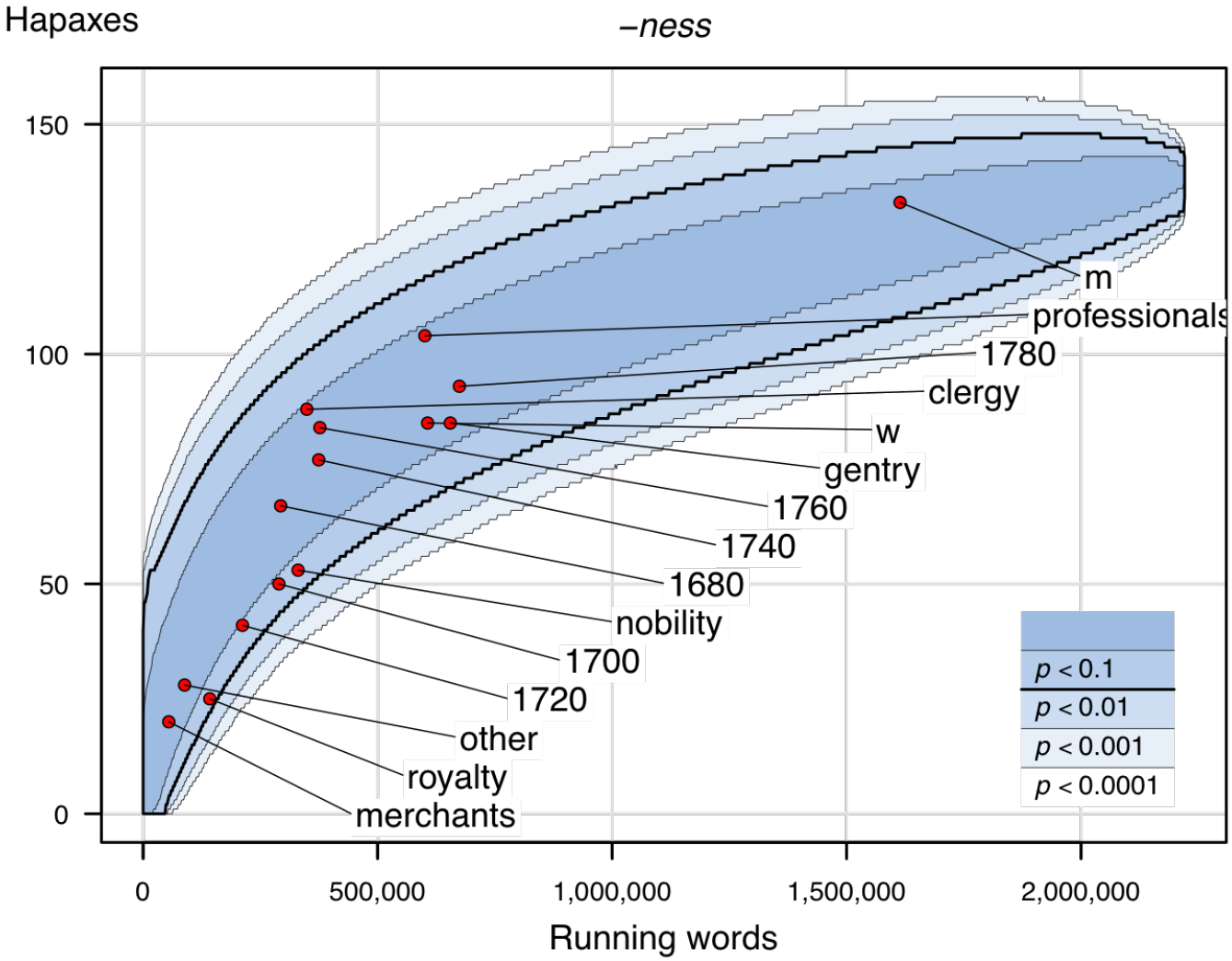
-ity hapaxes, CEECE



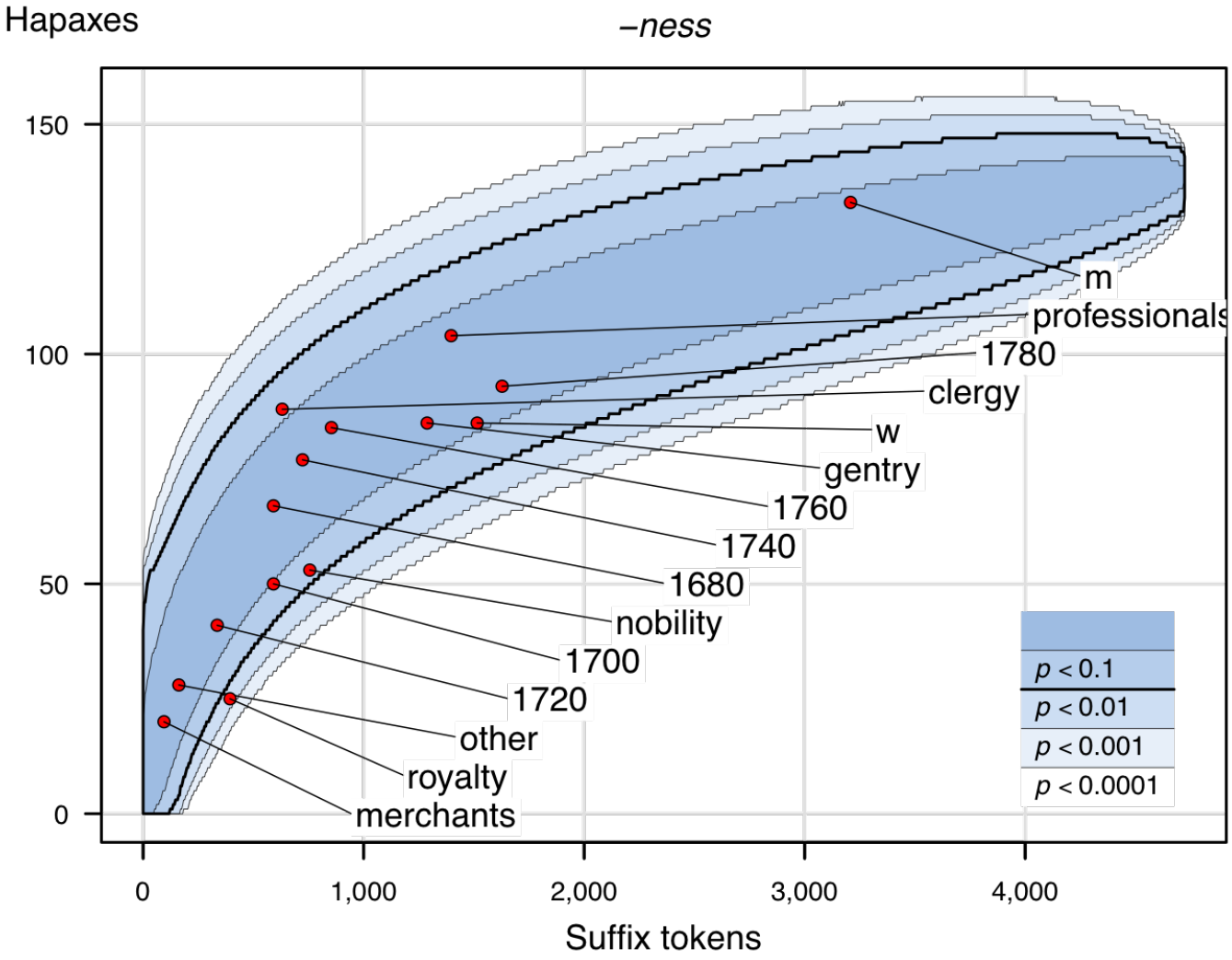
-ity hapaxes, CEECE



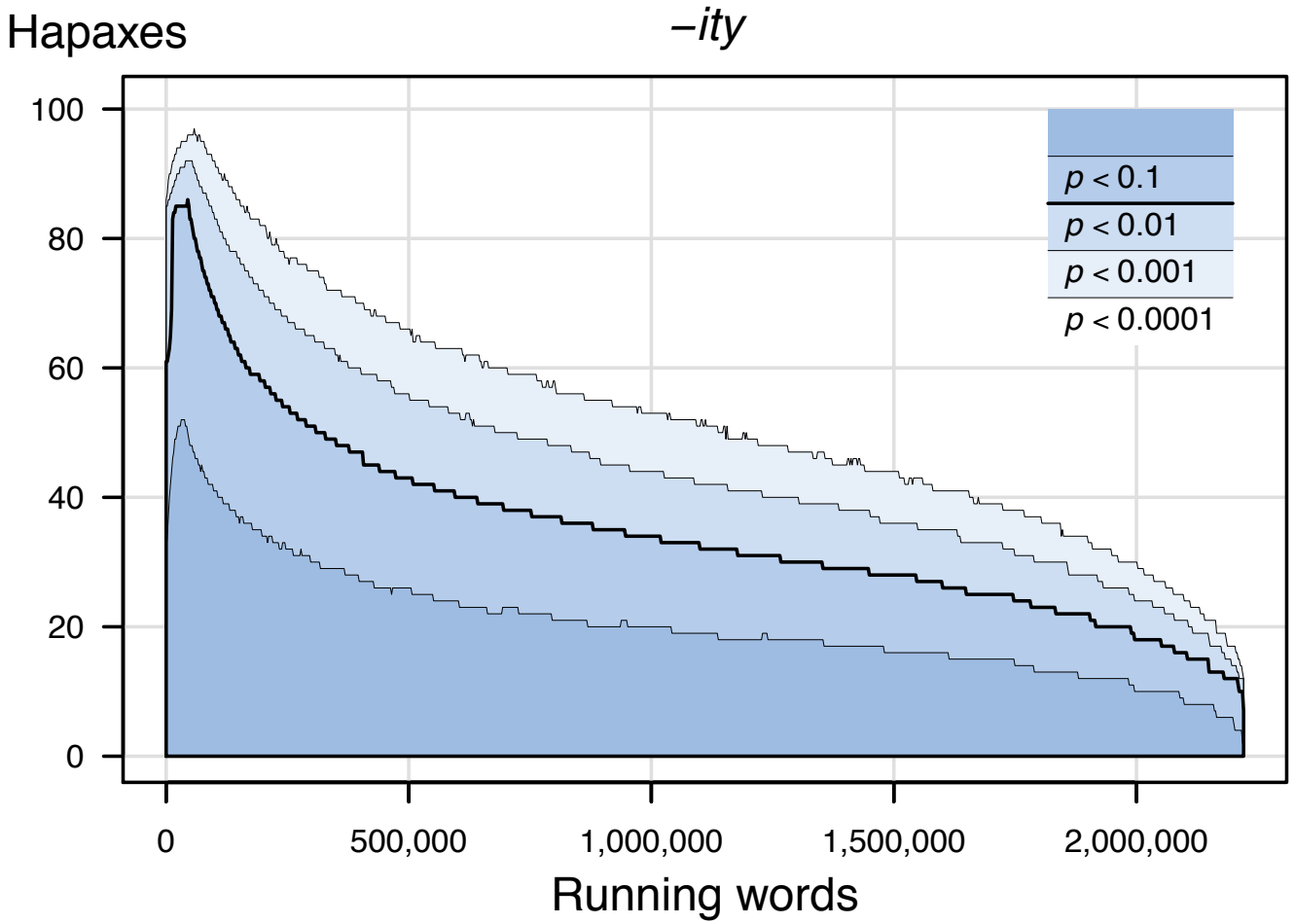
-*ness* hapaxes, CEECE



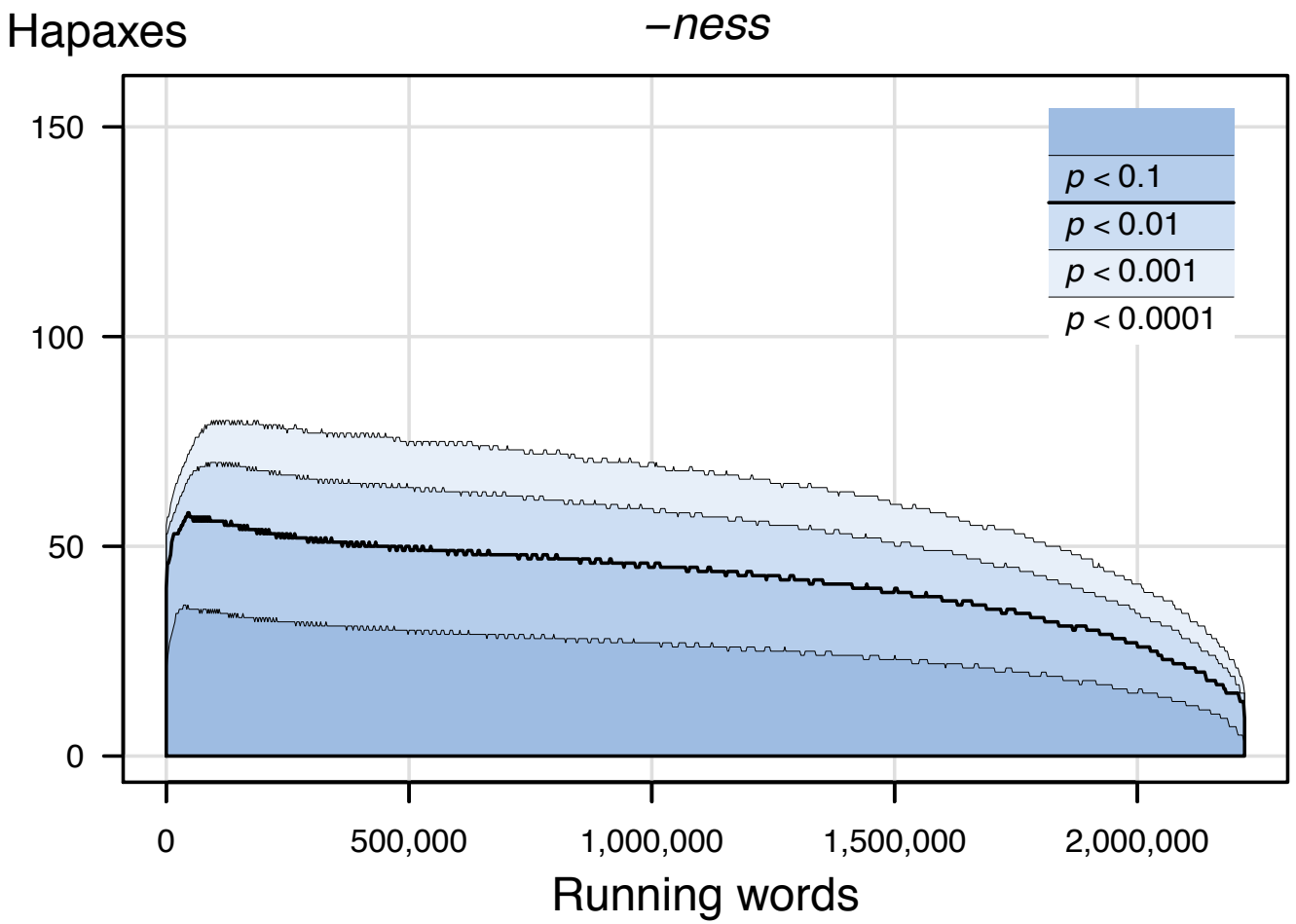
-*ness* hapaxes, CEECE



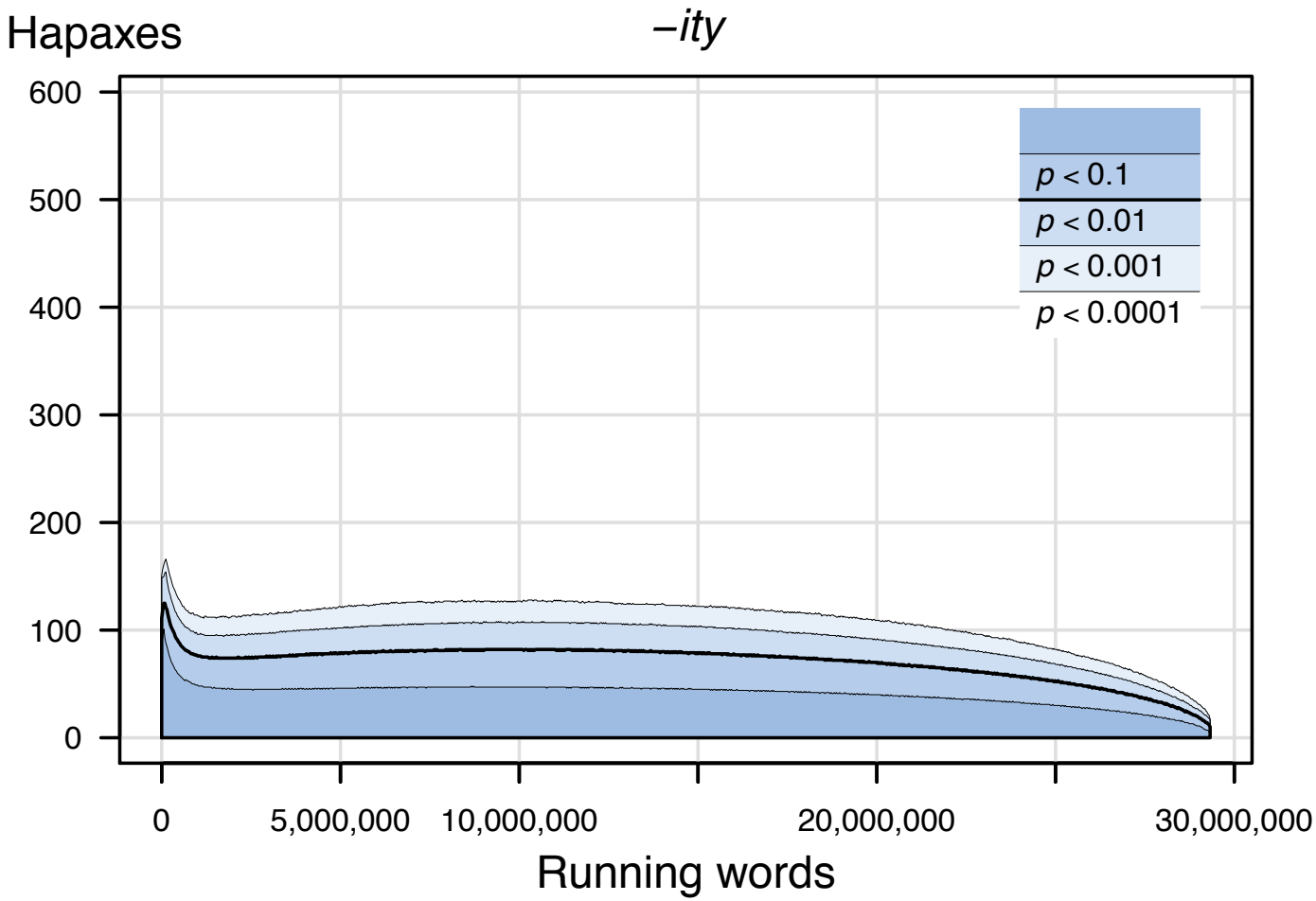
Confidence intervals, CEECE



Confidence intervals, CEECE



Confidence intervals, BNC



Confidence intervals, BNC