types2

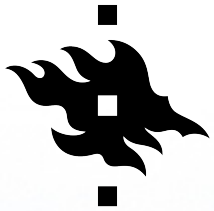
A TOOL FOR ANALYSING VARIATION IN MORPHOLOGICAL PRODUCTIVITY

TANJA SÄILY (WITH JUKKA SUOMELA)

UNIVERSITY OF HELSINKI

RESEARCH QUESTIONS

- Is there **sociolinguistic variation and change** in the productivity of *-ness* and *-ity* in the history of English?
- Are the **productivity measures** proposed in previous research valid in and applicable to sociolinguistic data of this kind?
- What are the requirements for a **usable tool** for studying variation in productivity in data of this kind?

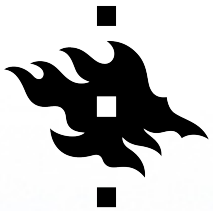


UNIVERSITY OF HELSINKI

PREVIOUS MEASURES OF PRODUCTIVITY

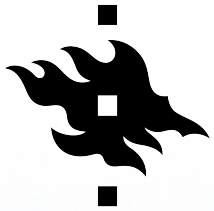
- Baayen (1992, 2009): measures based on frequencies of **types** (*V*), **tokens** (*N*) and **hapax legomena** (n_1, *h*)
  - Realised productivity *V* = type frequency
  - Potential productivity $P = n_1/N$
  - Expanding productivity $P^* = n_1/h$

- At least *V* and *P* depend non-linearly on corpus size
  → cannot **compare**, e.g., men and women if less data from women
- P^* unfeasible in non-lemmatised corpora with lots of **spelling variation**
  - Frequency of genuine hapax legomena *h* unclear

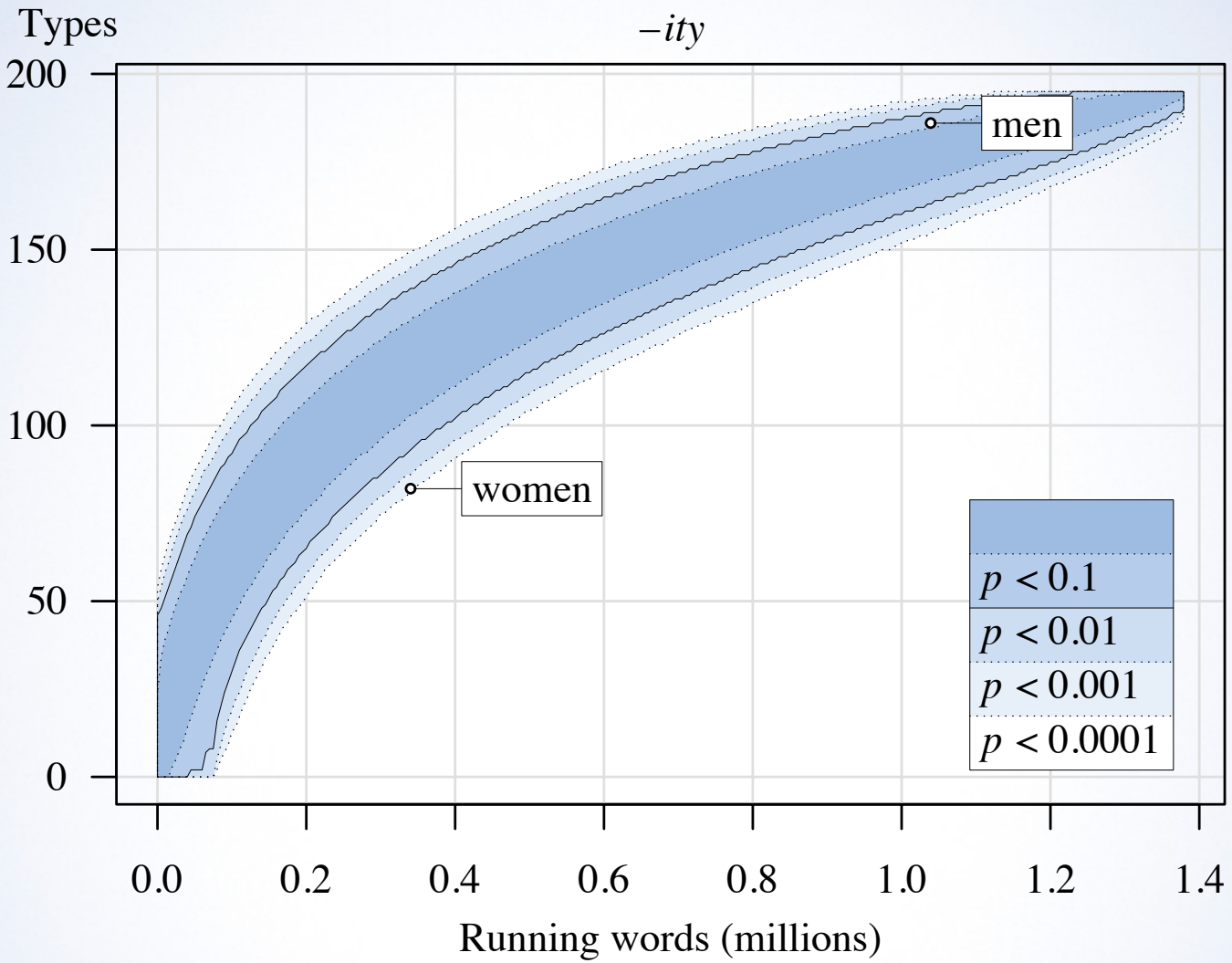


SÄILY & SUOMELA (2009)

- Method based on **accumulation curves** and **permutation testing**
- Solves problem of comparison: only compares subcorpus (e.g. women) with randomly composed subcorpora of the **same size**
  - Two measures of corpus size: running words and affix tokens
- Finds hapax legomena to be unusable in corpus (1.4M words)
  → concentrates on type frequency



UNIVERSITY OF HELSINKI

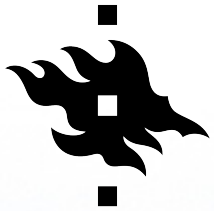


UNIVERSITY OF HELSINKI

PROBLEMS

- Exploratory analysis → multiple hypotheses tested, need to:
  - Adjust **significance** level
  - Conveniently **browse** through the results

- Method-specific requirement:
  - Easily change the measure of **corpus size** (running words vs. affix tokens)



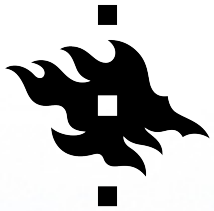
UNIVERSITY OF HELSINKI

NEW IMPLEMENTATION: *types2*

- Open-access tool (Suomela 2014)

- Facilitates **exploration**: interactive images, hyperlinks
  - Results also provided as tables and static figures

- Provides actual *p*-values, **false discovery rate control** (Benjamini & Hochberg 1995)



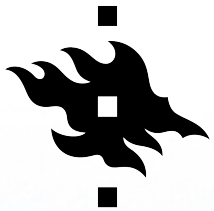
UNIVERSITY OF HELSINKI

EXAMPLE: *-ness* & *-ity*

- **Material**: *Corpus of Early English Correspondence* (CEEC), 1600–1681; CEEC *Extension* (CEECE), 1680–1800

- **Sociolinguistic subcorpora** based on:
  gender, social rank, social mobility, education, time period

- **Measure of productivity**: type frequency
  as a function of the number of running words / affix tokens



UNIVERSITY OF HELSINKI

1600–1681

UNIVERSITY OF HELSINKI

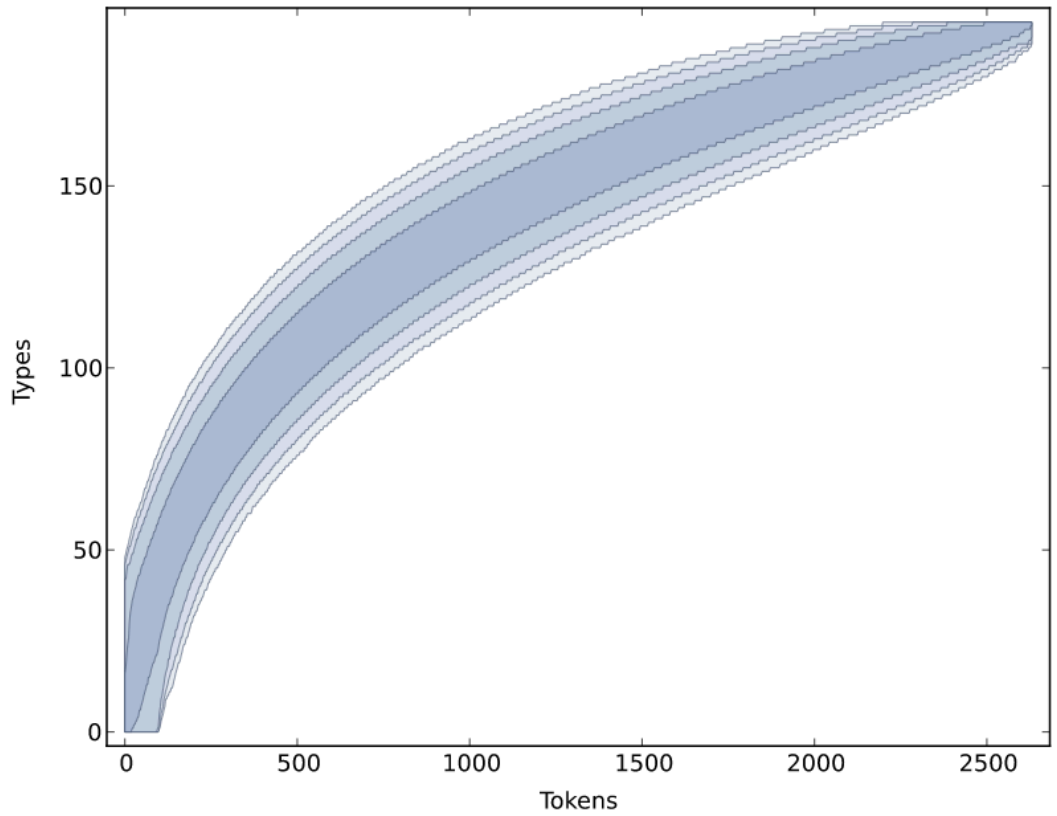
Corpus: ceec+ceece-1680-1800-person-period ...person-period-relcode
ceec-1600-1681-person-period ...person-period-relcode

ceec-1600-1681-person-period · CEEC, 1600-1681, sample = letters with a certain sender & period

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

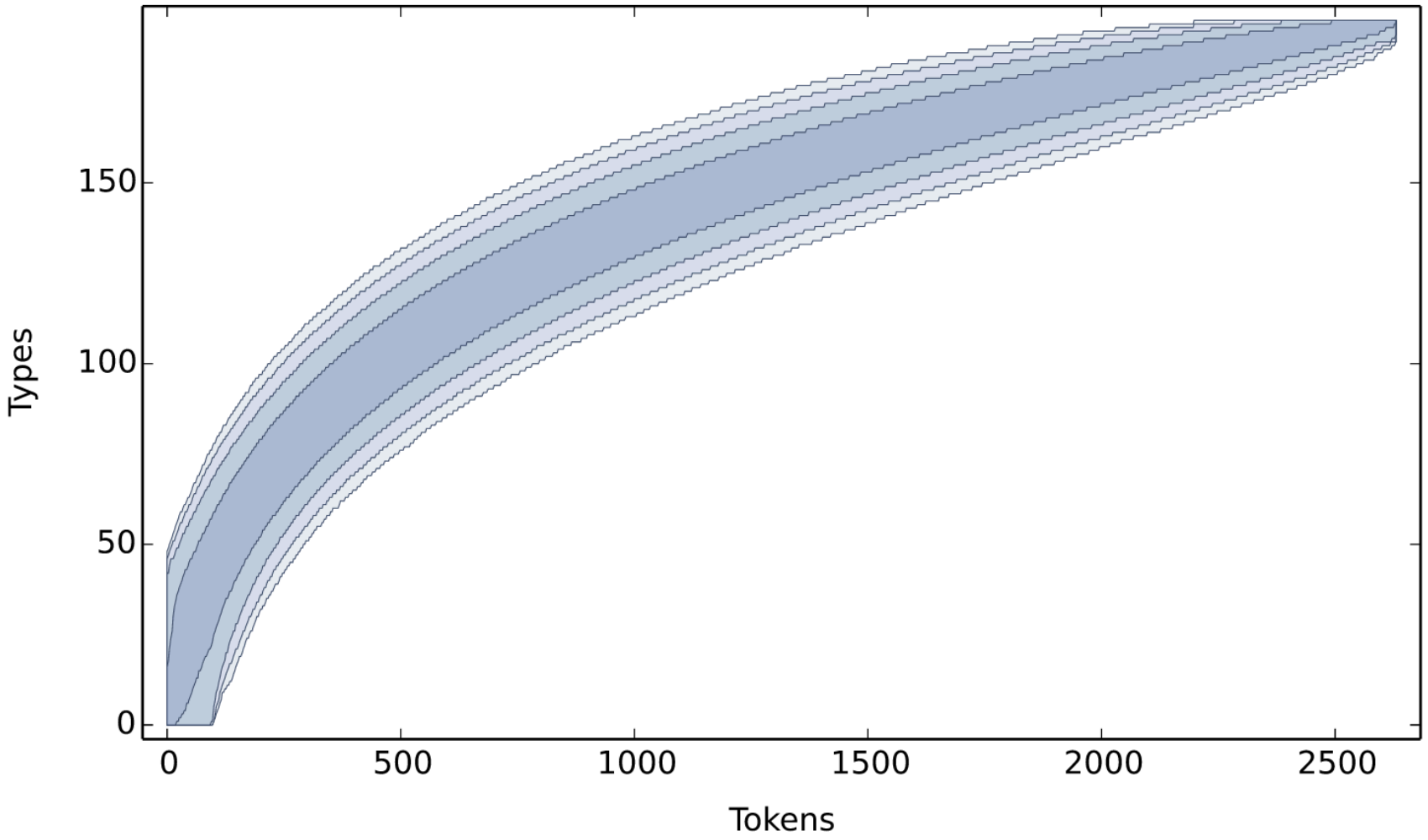


Collection: none

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

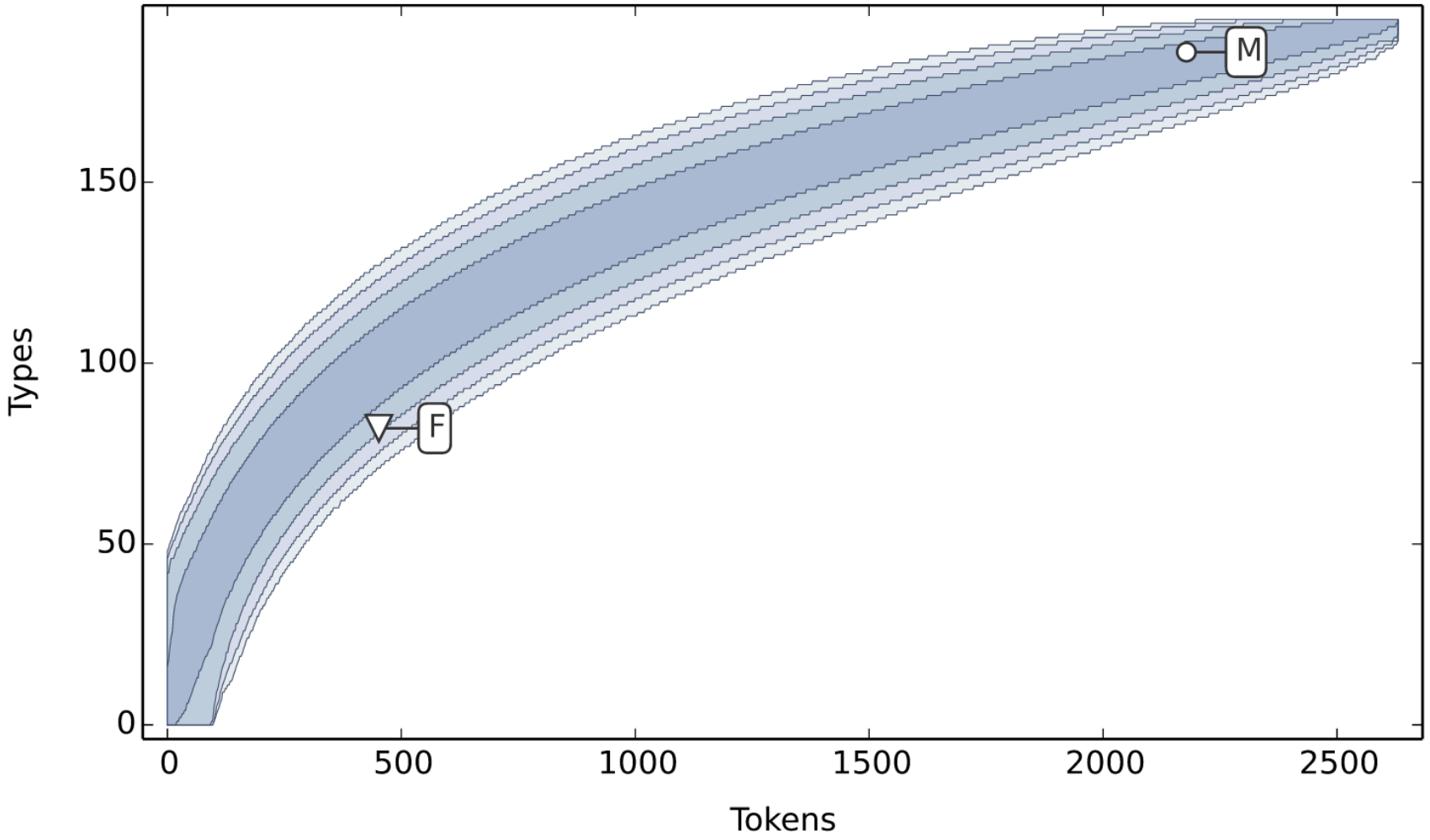


Collection: none

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

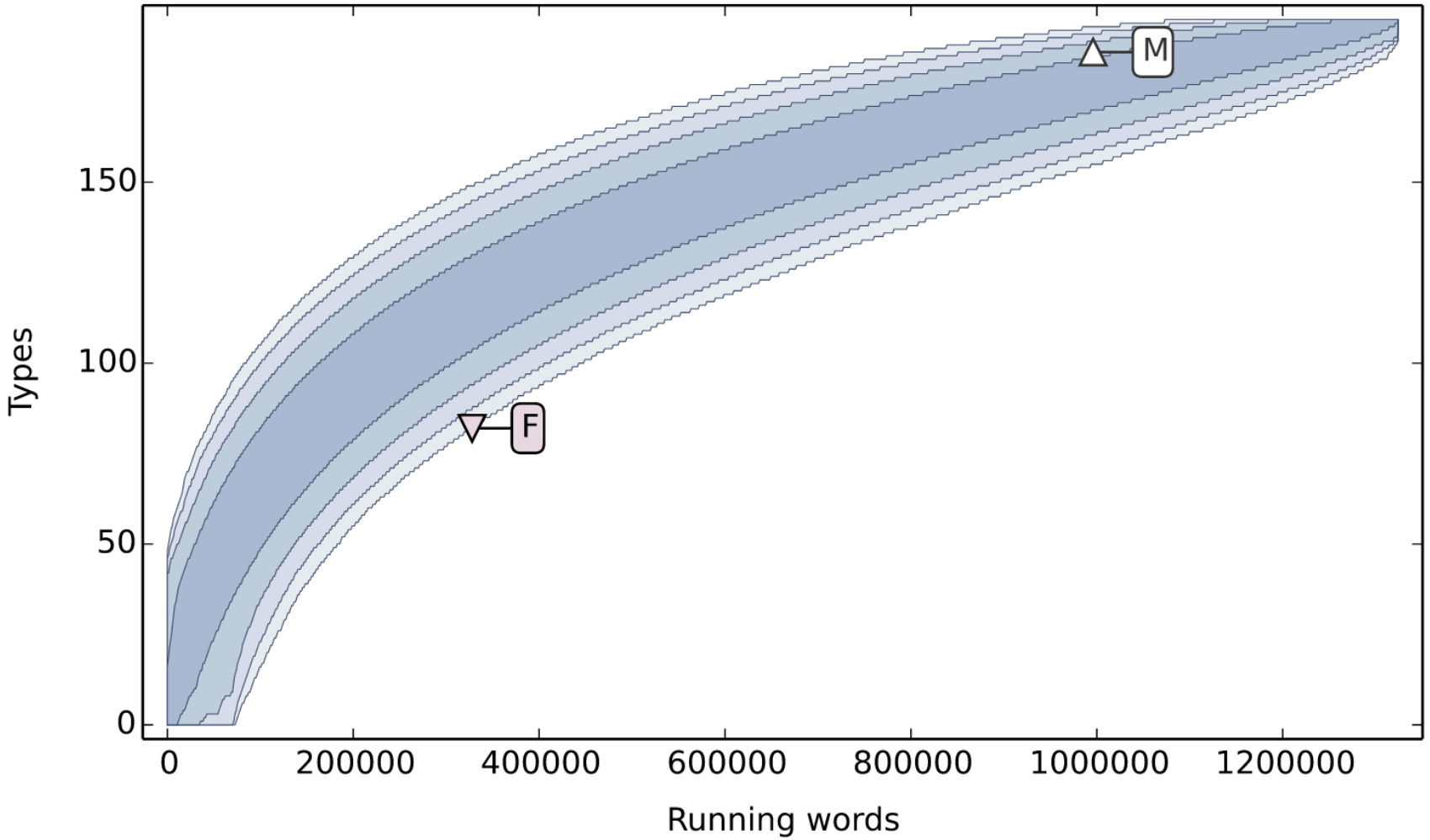


Collection: none F M

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

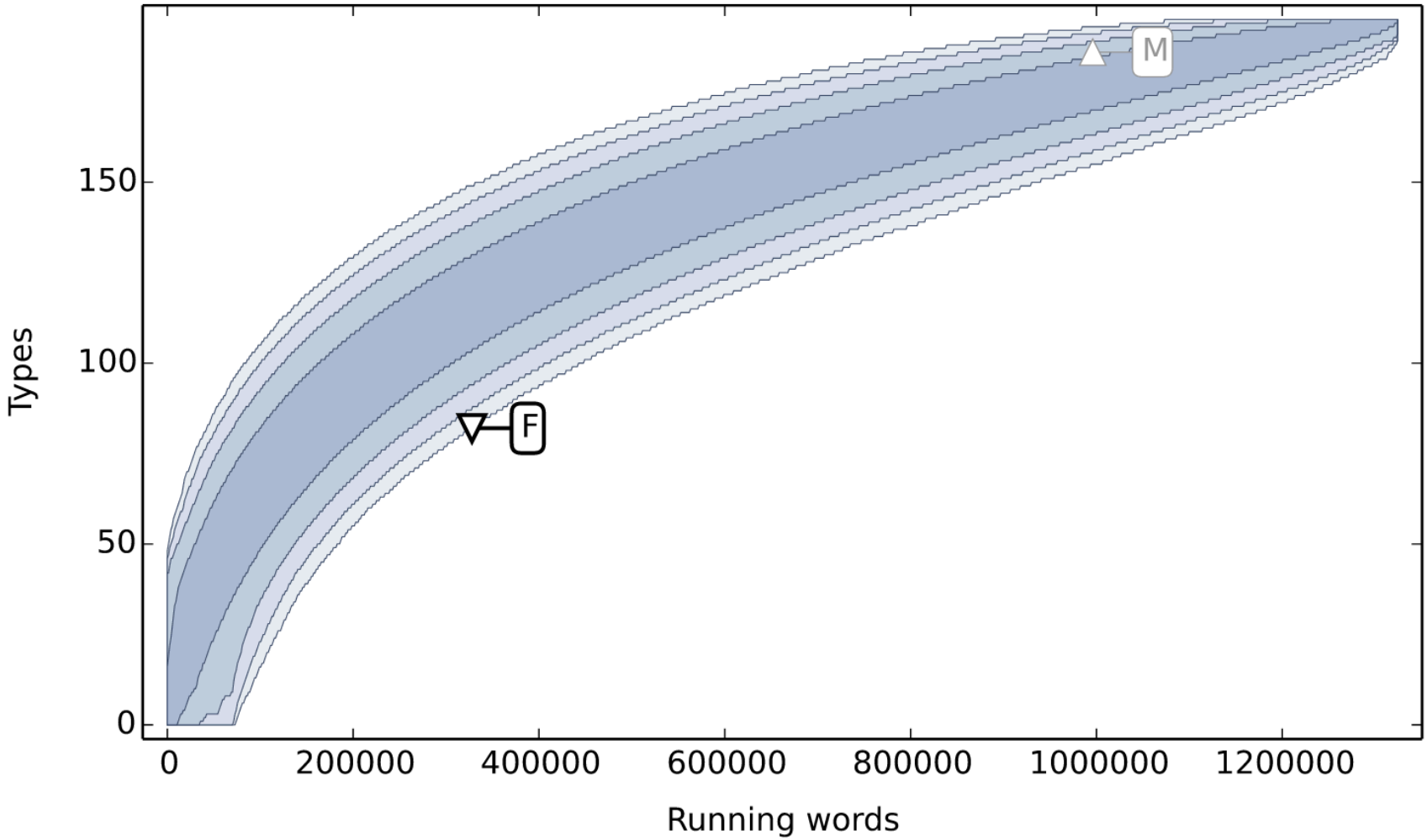


Collection: none F M

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words



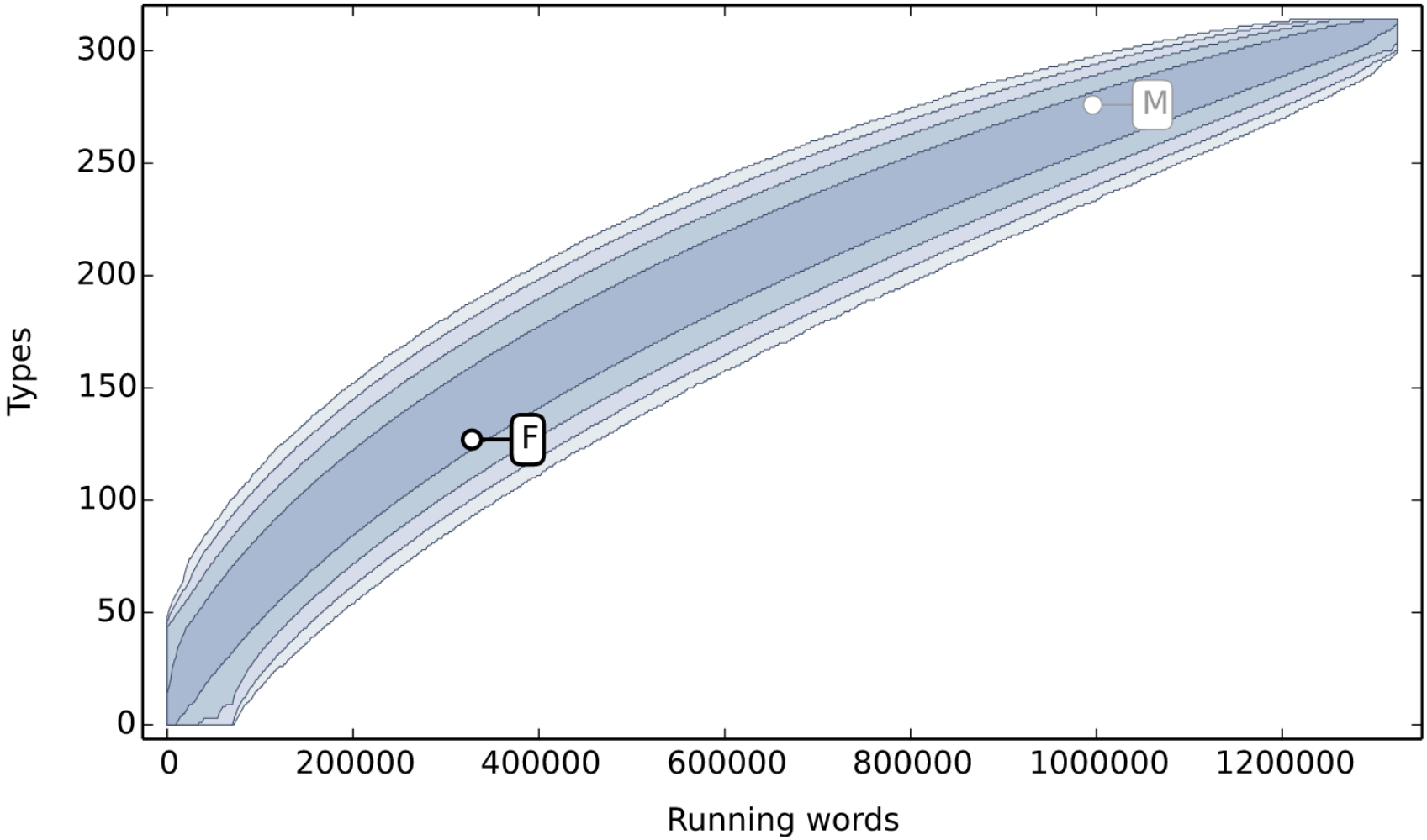
Collection: none F M

Statistics: 327923 running words 82 types 0.000103 below

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words



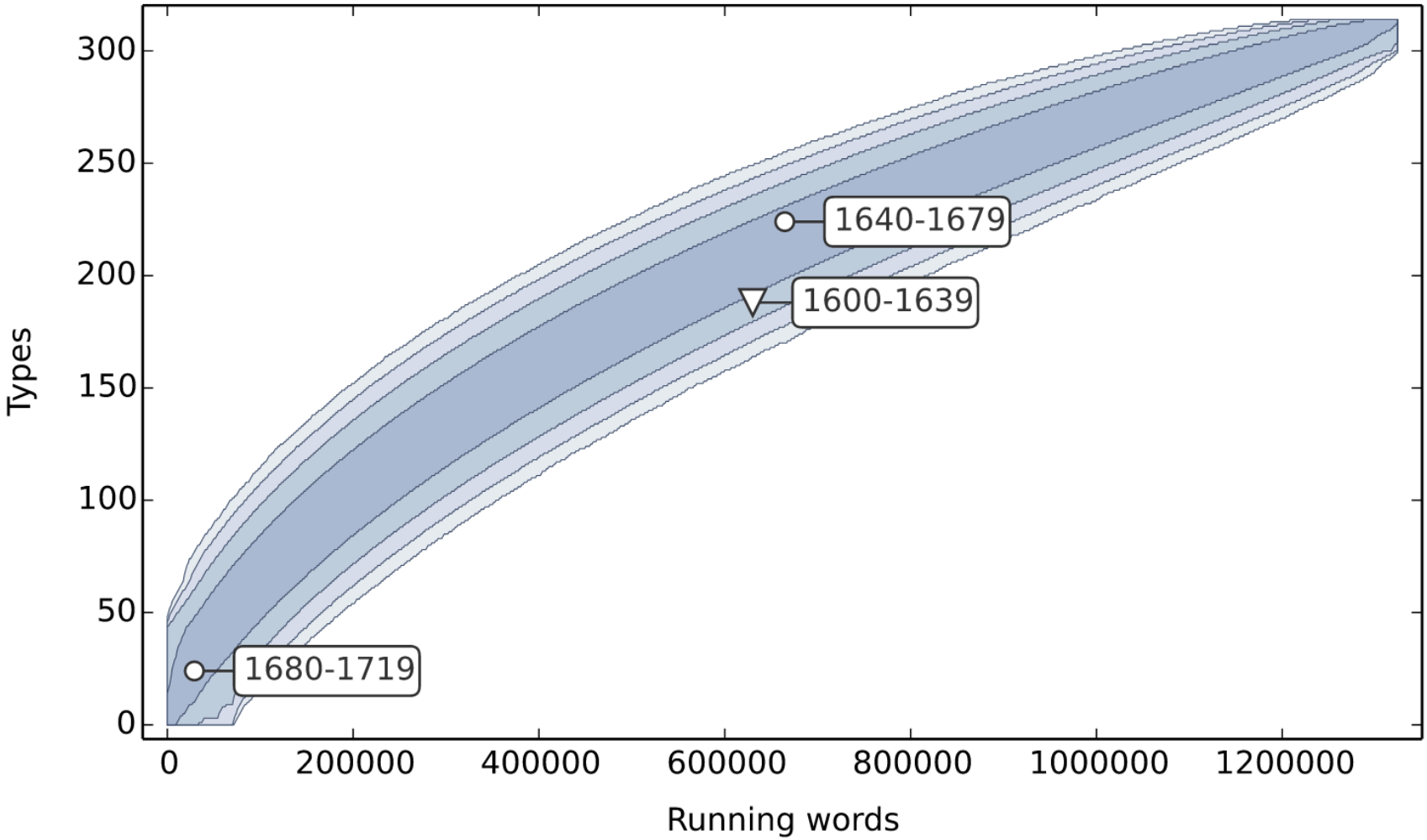
Collection: none F M

Statistics: 327923 running words 127 types 0.181273 below

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

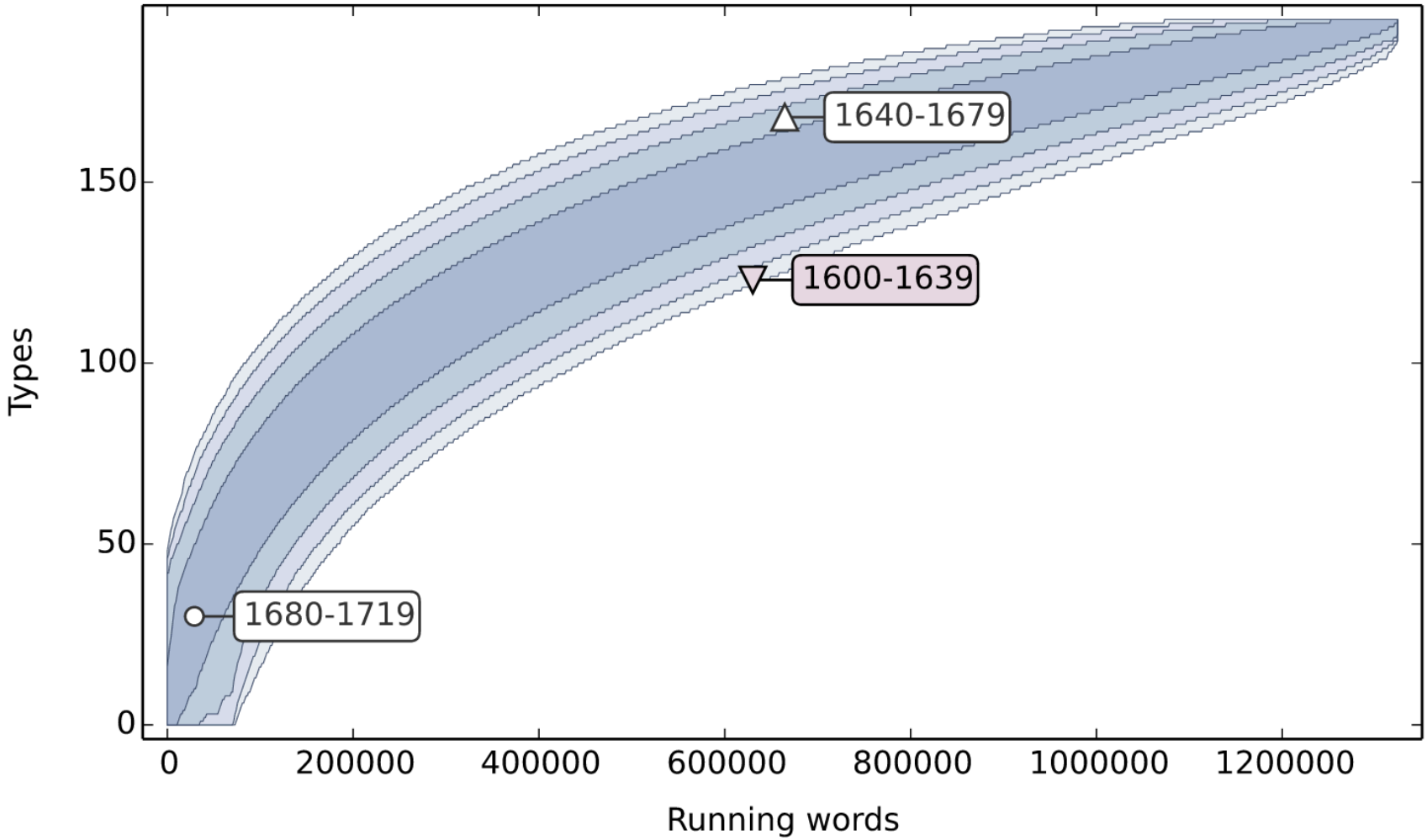


Collection: none 1600-1639 1640-1679 1680-1719

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

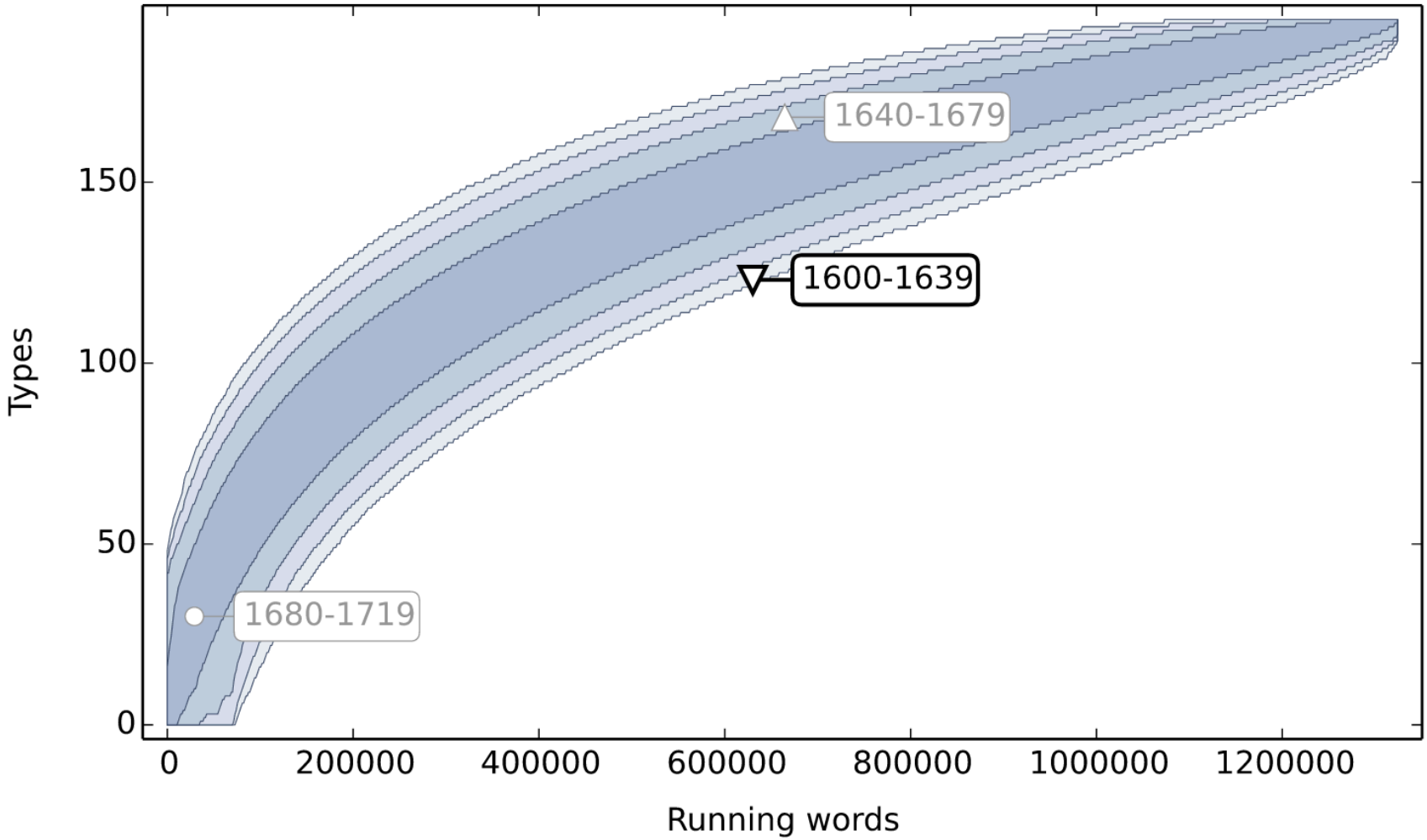


Collection: none 1600-1639 1640-1679 1680-1719

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words



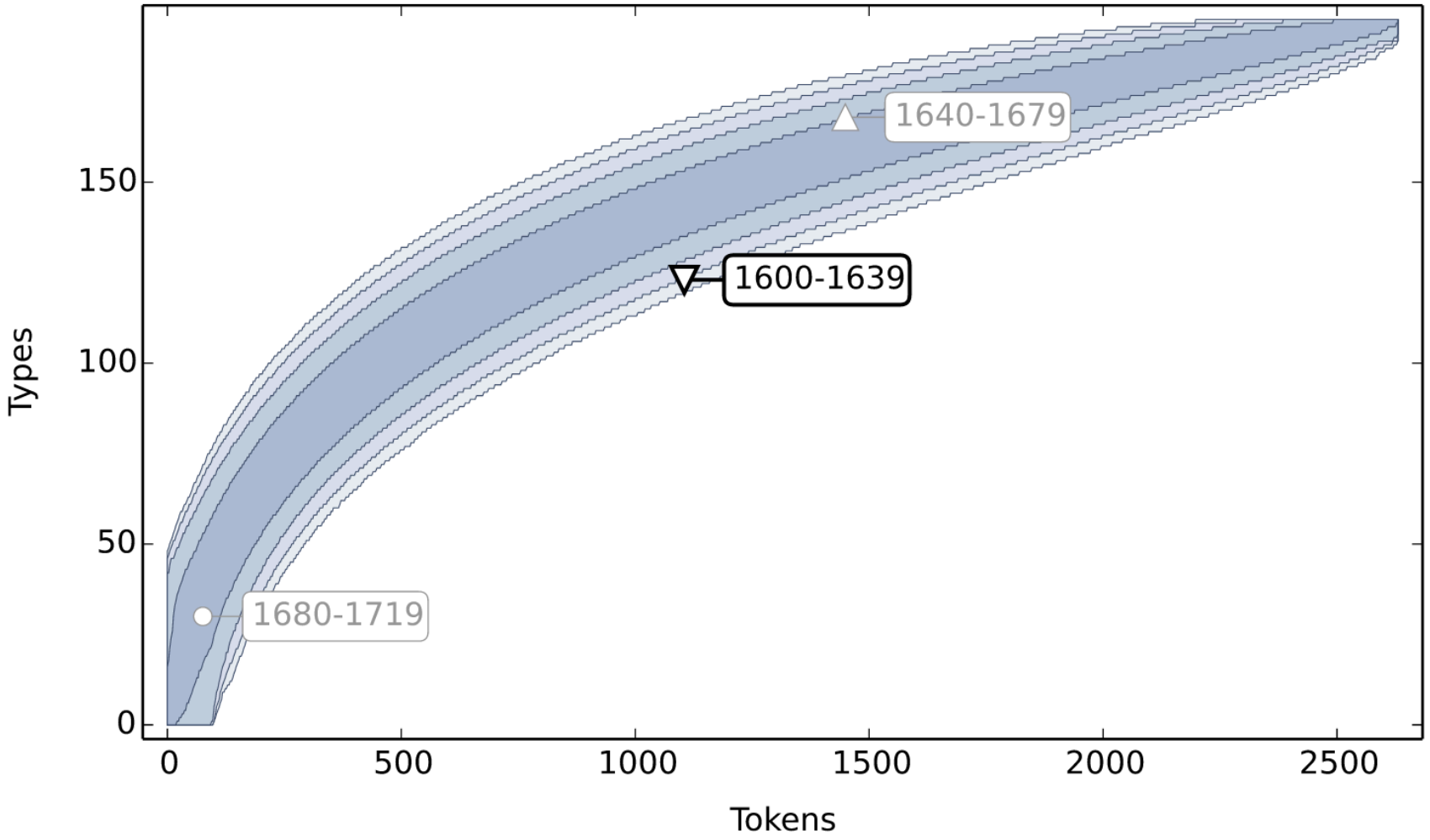
Collection: none 1600-1639 1640-1679 1680-1719

Statistics: 630158 running words 123 types 0.000229 below

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words



Collection: none 1600-1639 1640-1679 1680-1719

Statistics: 1105 tokens 123 types 0.000954 below

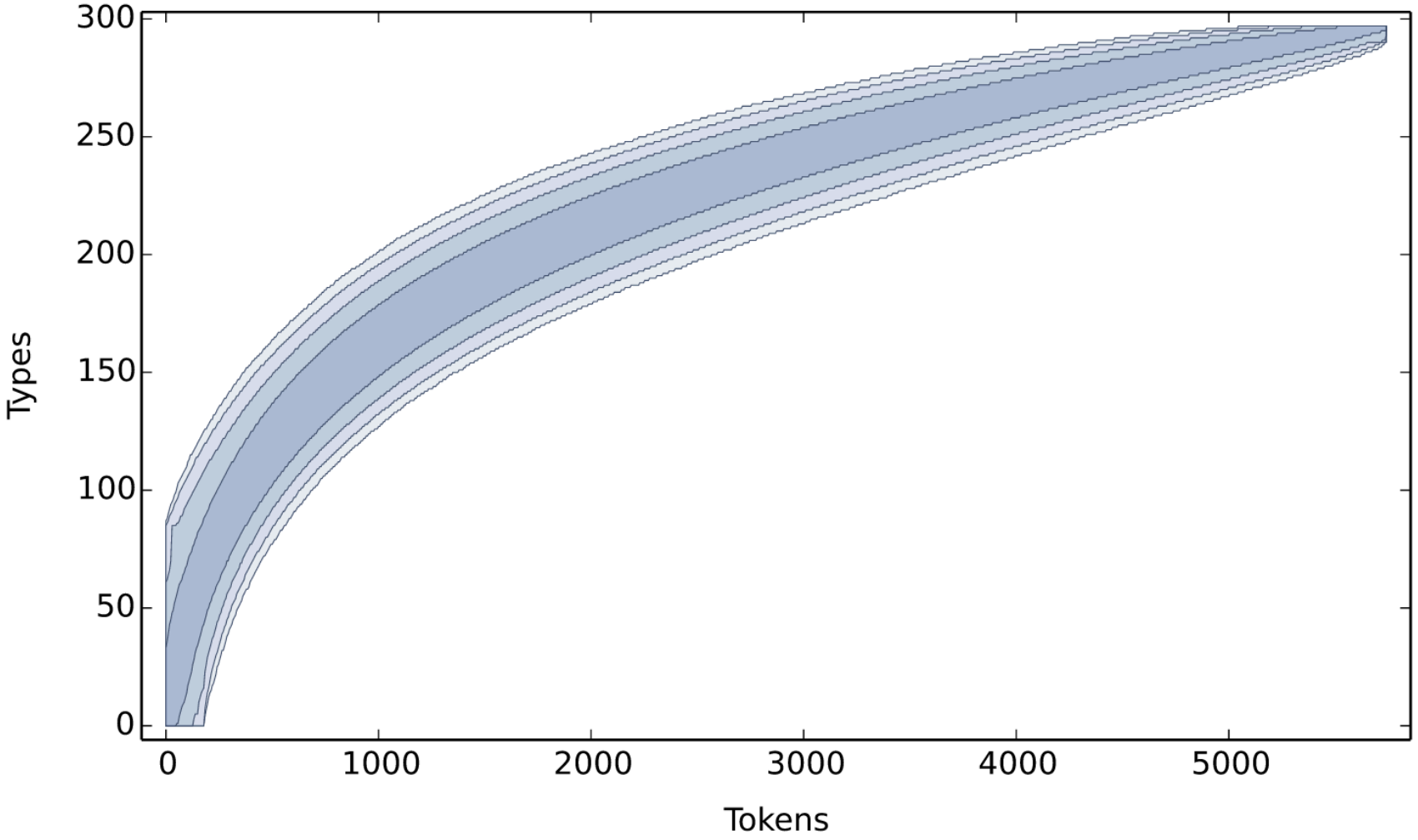
1680–1800

UNIVERSITY OF HELSINKI

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

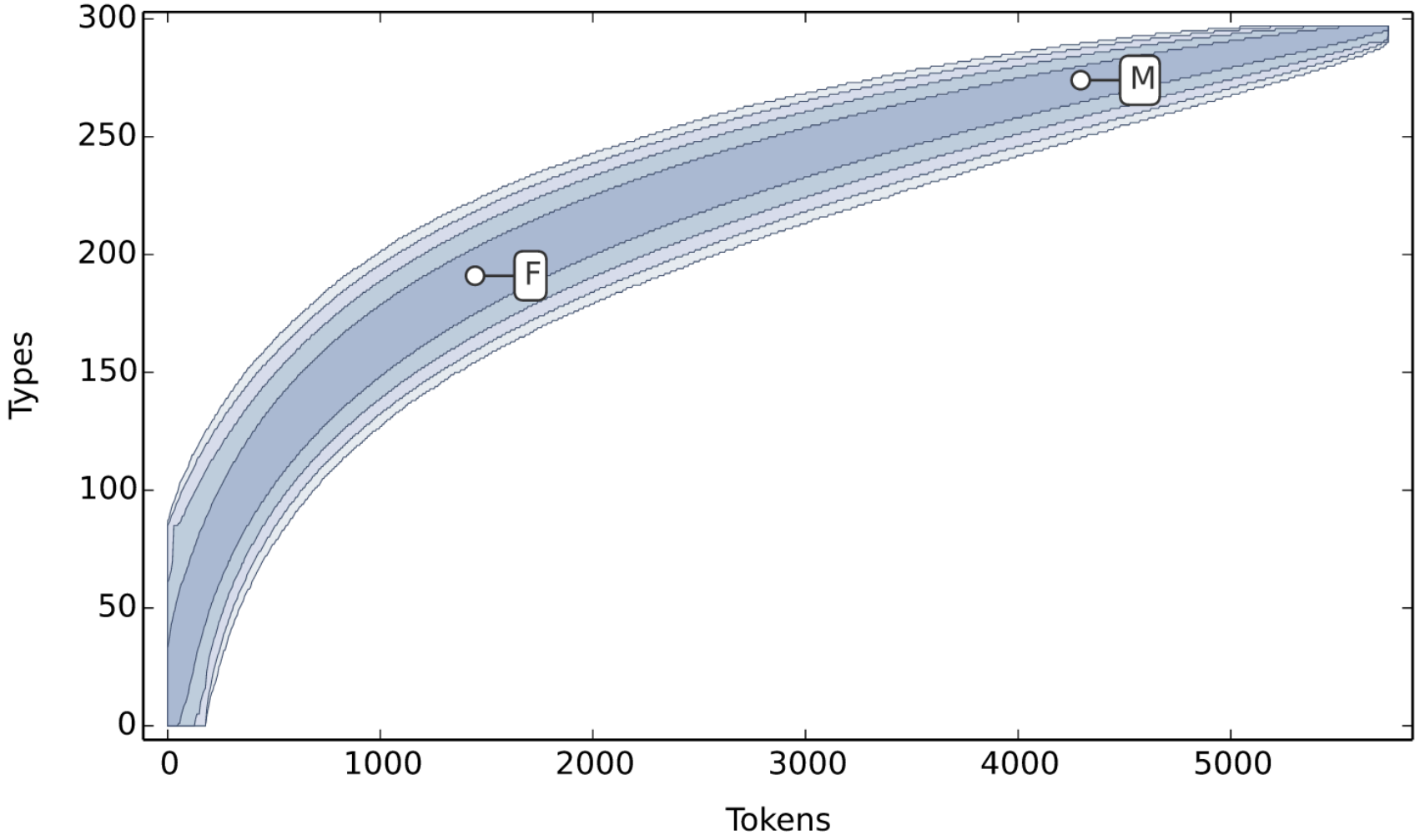


Collection: none

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

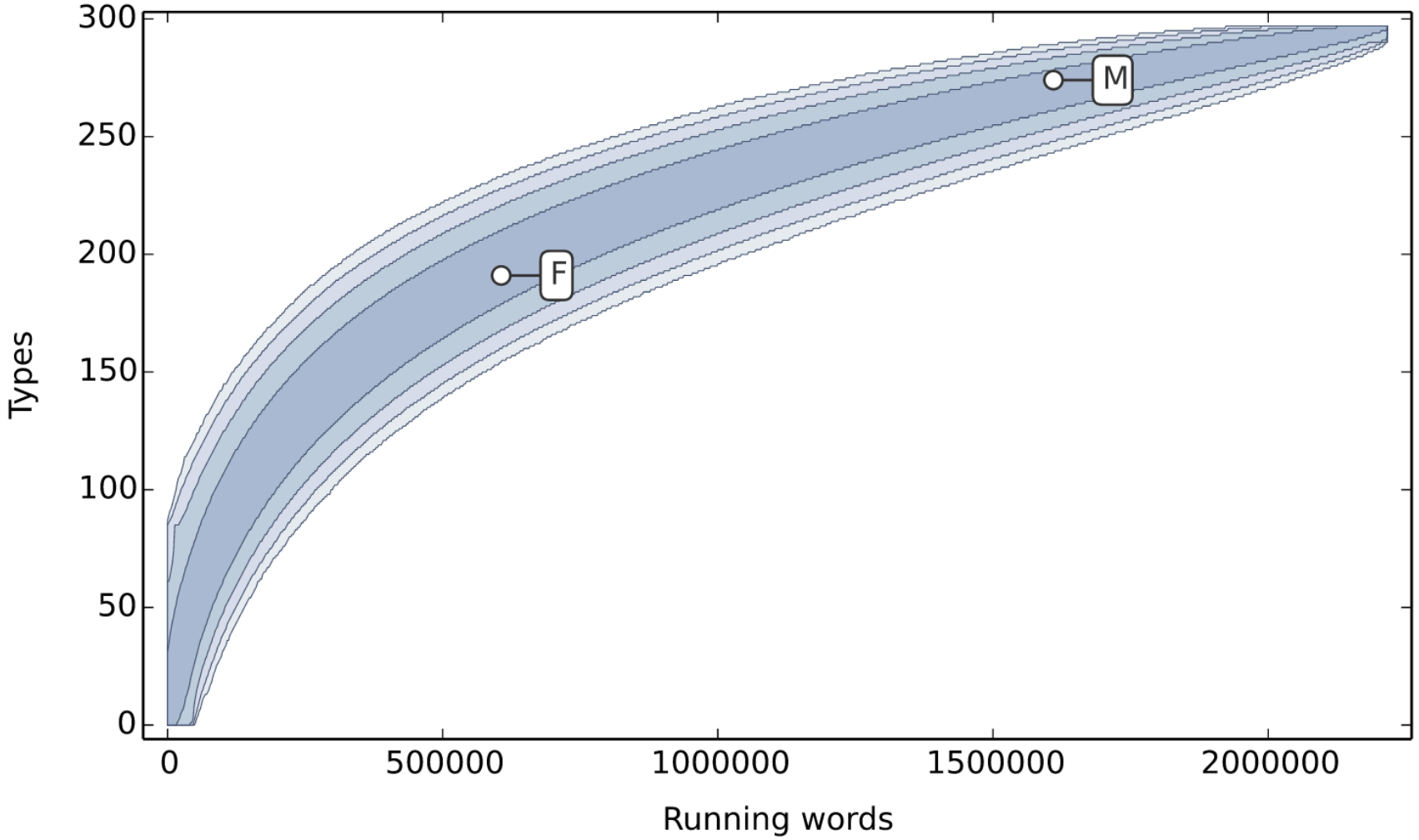


Collection: none F M

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

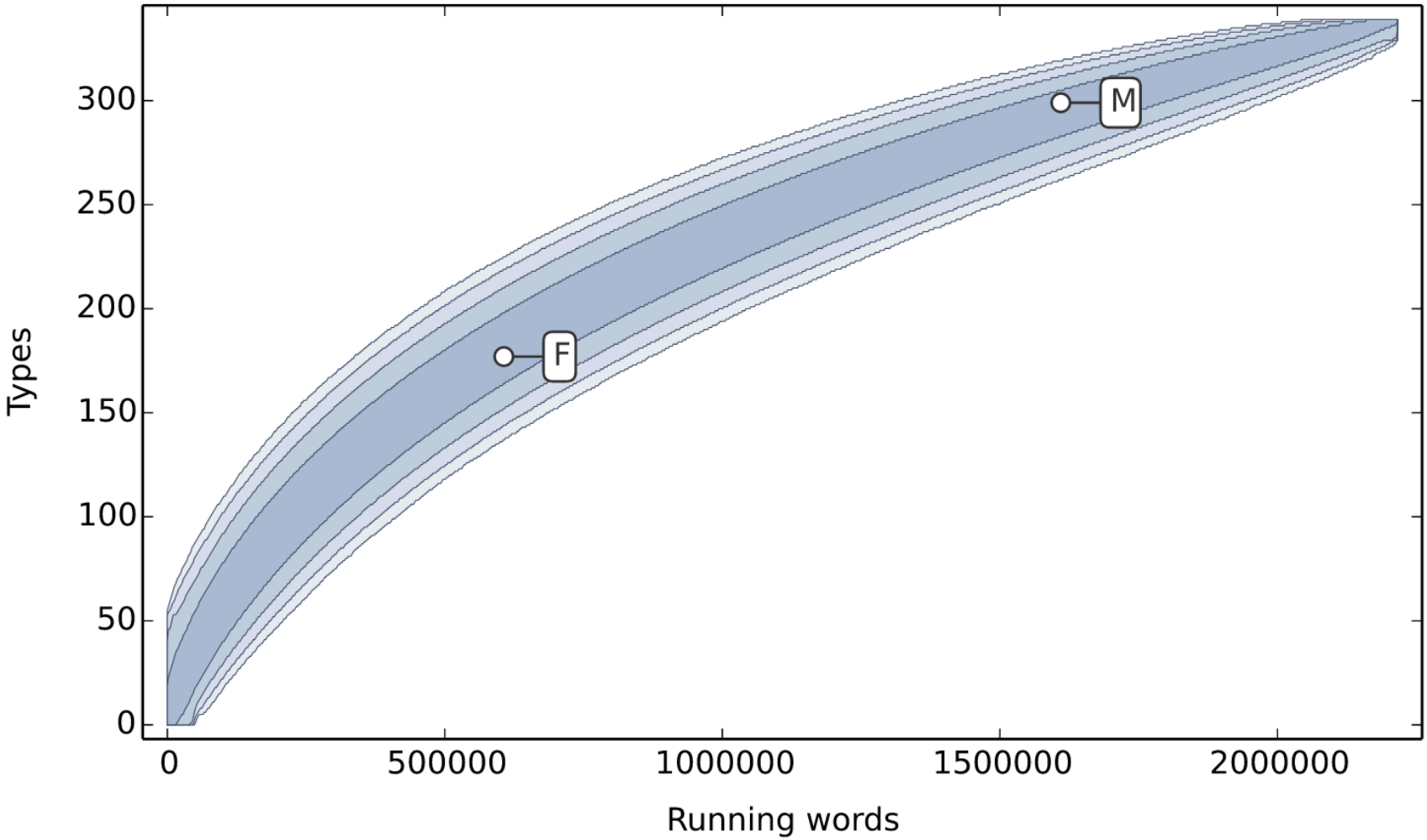


Collection: none F M

Dataset: ity [ness]

Points: none education period period-40 rank-current [sex] socmob

Axes: types/tokens [types/running words]

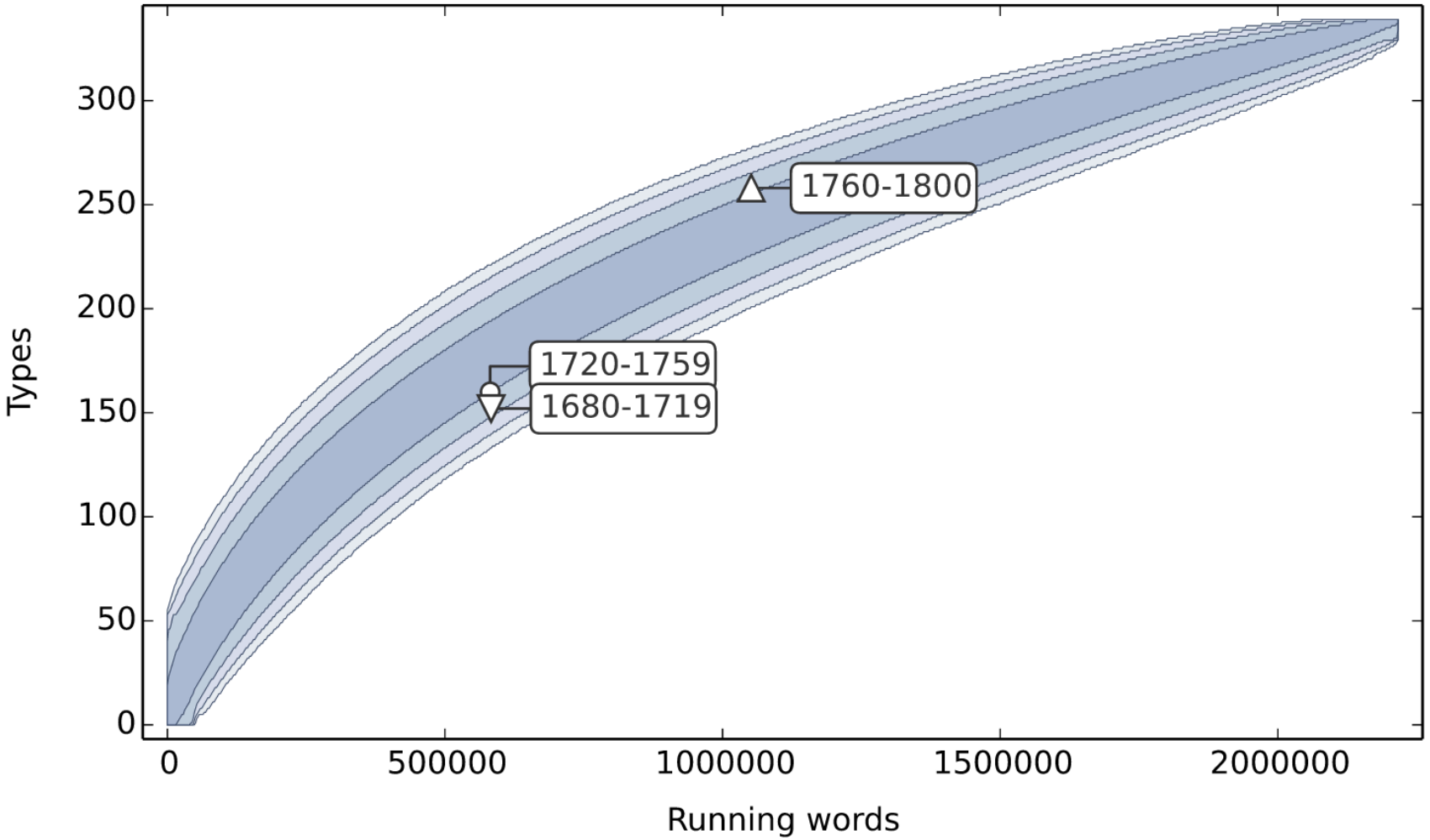


Collection: [none] F M

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

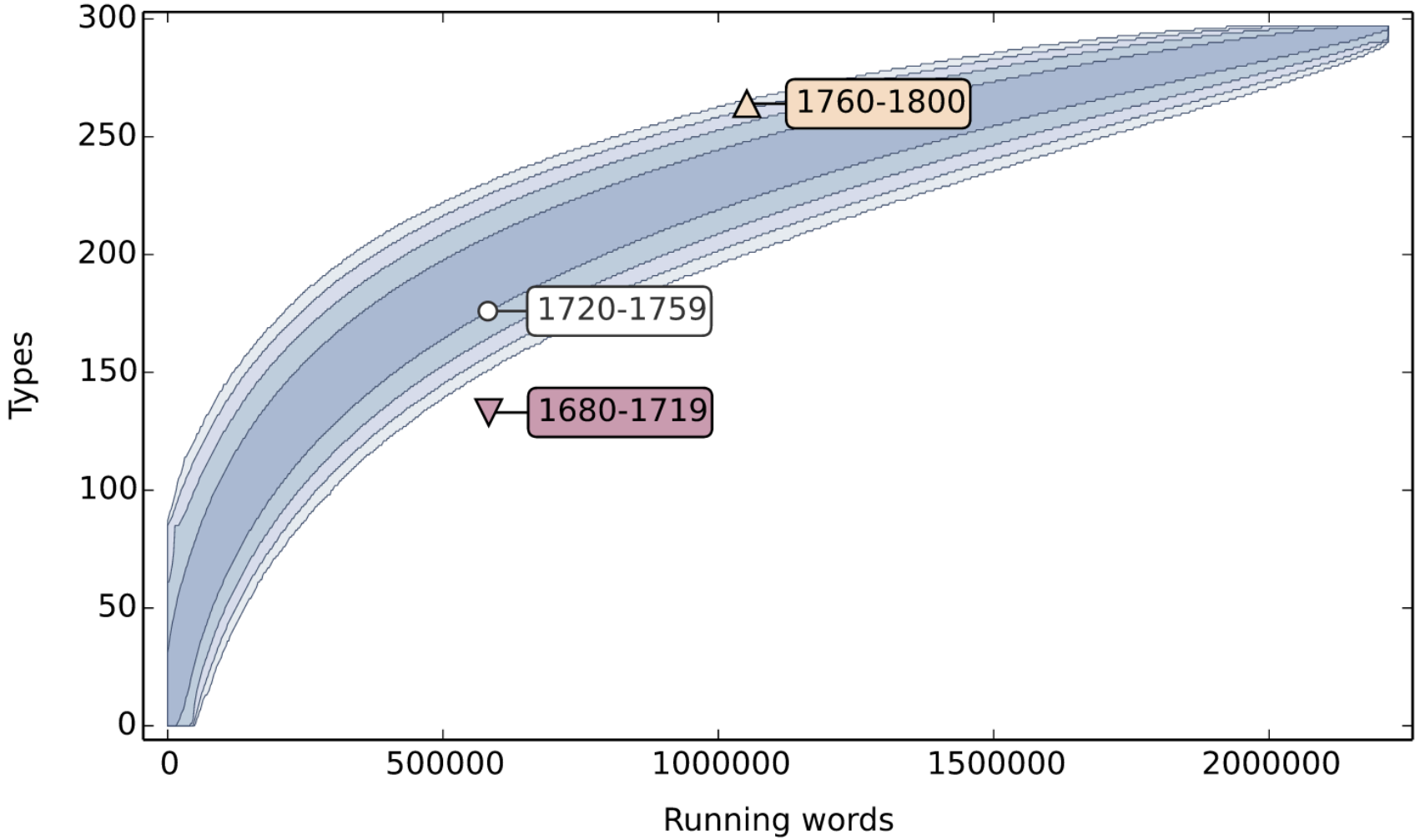


Collection: none 1680-1719 1720-1759 1760-1800

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words

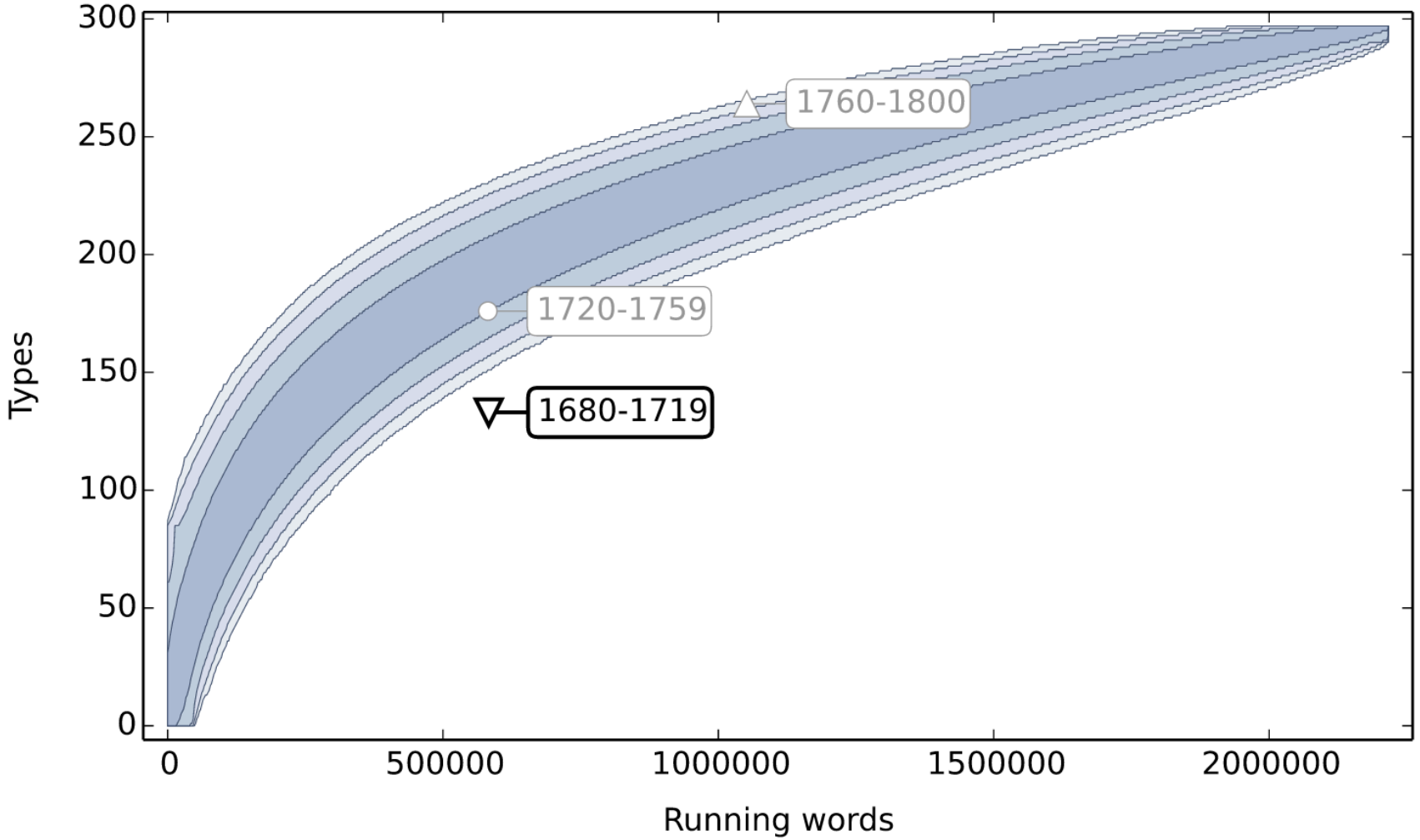


Collection: none 1680-1719 1720-1759 1760-1800

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words



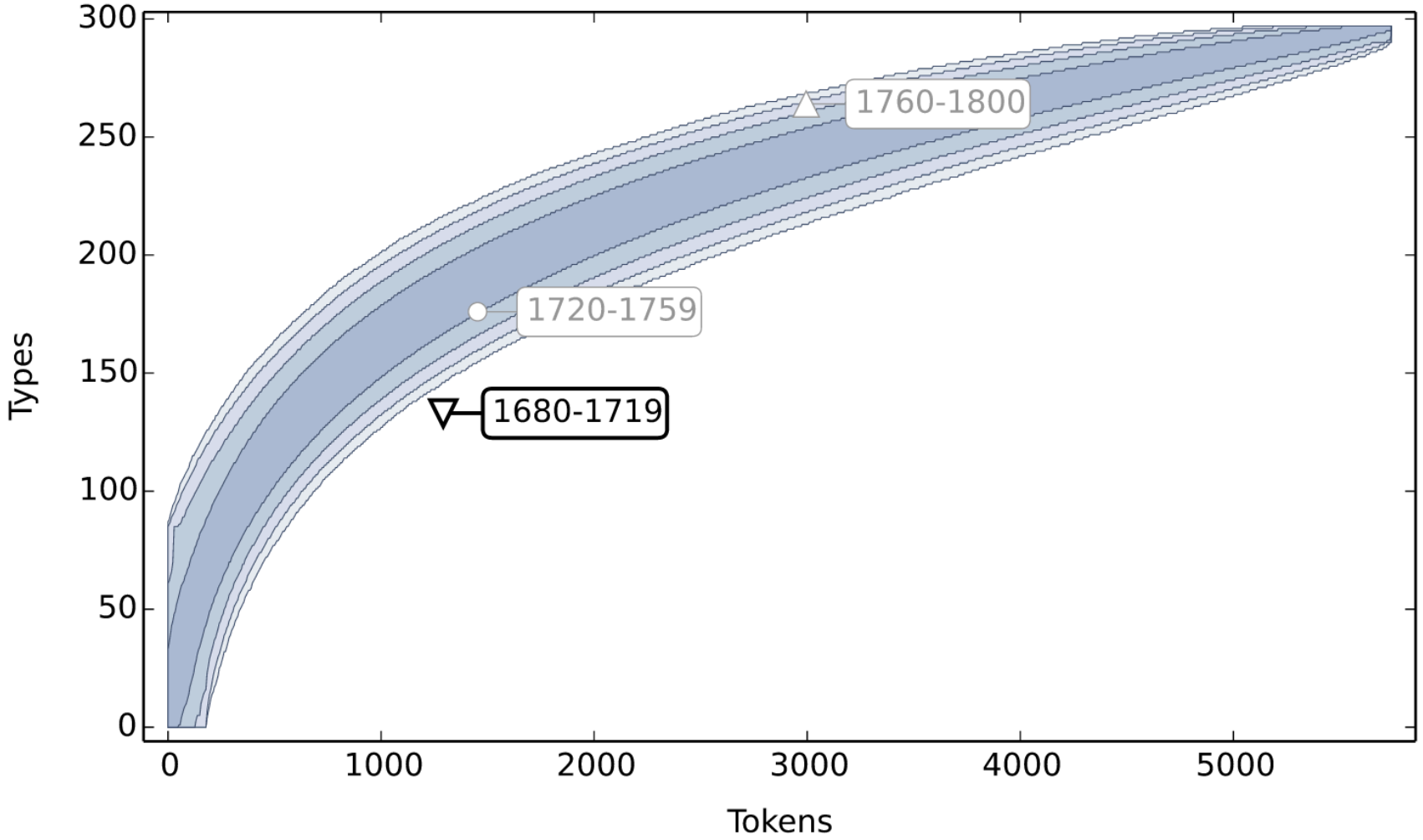
Collection: none 1680-1719 1720-1759 1760-1800

Statistics: 583086 running words 133 types 0.000000 below

Dataset: ity ness

Points: none education period period-40 rank-current sex socmob

Axes: types/tokens types/running words



Collection: none 1680-1719 1720-1759 1760-1800

Statistics: 1292 tokens 133 types 0.000000 below

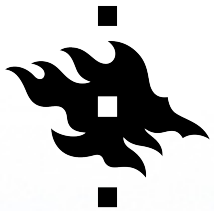
SIGNIFICANT RESULTS, 1600–1681

-ity	F, education ?	type-word	below
-ity	F	type-word	below
-ity	1600-1639	type-word	below
-ity	1600-1639	type-token	below

UNIVERSITY OF HELSINKI

SIGNIFICANT RESULTS, 1680–1800

-ity	1680-1719	type-word	below
-ity	1680-1719	type-token	below
-ity	1760-1800	type-word	above
-ity	1760-1800	type-token	above
-ness	rank R	type-token	below

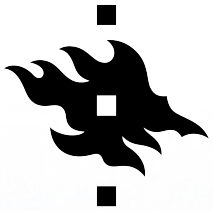


UNIVERSITY OF HELSINKI

SUMMARY OF RESULTS

- **Productivity of -*ity* increases** throughout the 17th and 18th centuries
- **Gender** variation in the use of -*ity*
  - 17th century: men use -*ity* more productively than women
- Social **rank** may also be a factor

(Säily & Suomela 2009, Säily forthcoming)



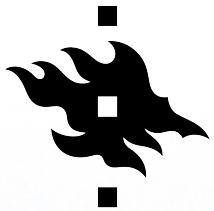
UNIVERSITY OF HELSINKI

CONCLUSION

1. Sociolinguistic variation in morphological productivity? **Yes**
2. Previous measures applicable? **Partly**
3. Requirements for tool? **Exploration, multiple measures**

- *types*2: both **exploratory** and **confirmatory** analysis
  - Can use both types and hapaxes in large corpora (Säily 2011)
    ≈ Baayen's realised and potential productivity (*V*, *P*)

- Future work: link to more metadata, actual corpus texts
  → facilitate **interpretation** of results



UNIVERSITY OF HELSINKI

REFERENCES

- Baayen, R. H. 1992. Quantitative aspects of morphological productivity. *Yearbook of Morphology 1991*, 109–149. Dordrecht: Kluwer.
- Baayen, R. H. 2009. Corpus linguistics in morphology: Morphological productivity. *Corpus Linguistics: An International Handbook*, 899–919. Berlin: Mouton de Gruyter.
- Bell, A. 1984. Language style as audience design. *Language in Society* 13(2): 145–204.
- Benjamini, Y. & Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57(1): 289–300.
- *Corpora of Early English Correspondence*. Compiled by T. Nevalainen, H. Raumolin-Brunberg et al. at the University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
- Säily, T. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7(1): 119–141.
- Säily, T. Forthcoming. *Sociolinguistic Variation in English Derivational Productivity: Studies and Methods in Diachronic Corpus Linguistics*. PhD dissertation, University of Helsinki.
- Säily, T. & J. Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. *Corpus Linguistics: Refinements and Reassessments*, 87–109. Amsterdam: Rodopi.
- Suomela, J. 2014. *types*2: Type and hapax accumulation curves. Computer program. ZENODO. doi:10.5281/zenodo.9868, www.iki.fi/suo/types2

UNIVERSITY OF HELSINKI