Comparing type counts

The case of women, men and *-ity* in early English letters

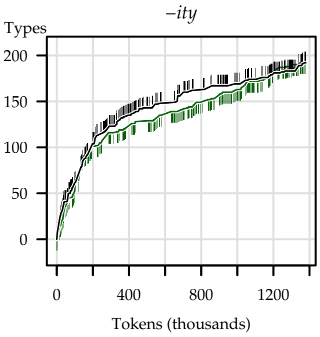
Tanja Säily and Jukka Suomela

1 Research question

- Productivity of the noun-forming suffix *-ity* (as in *generosity*) in 17th-century English letters
- Material from the *Corpus of Early English Correspondence*
  - The corpus covers the time span 1410?–1681 (2.7 million words); we use letters written between 1600–1681 (1.4 million words)
- We wish to compare the **numbers of different types** of *-ity* used by different sociolinguistic groups
- **Hypothesis**: gender is significant
  - We believe that *-ity*, a 'learned' and etymologically foreign suffix, is in this material less productive with women than with men, as 17th-century women received far less education than men
- How to compare? Only about 1/4 of the 17th-century material in the corpus was written by women
  - Take a sample of equal size from men? Problems: choosing a representative sample, loss of data
  - Normalise type counts? Problem: number of types does not grow linearly with number of tokens
- How to establish statistical significance?

2 Permutation testing

- Using a purpose-built computer program, choose a large number of **random permutations** of parts of the corpus
  - One permutation = one random reordering of the entire corpus
- For each permutation, compute a **type accumulation curve**



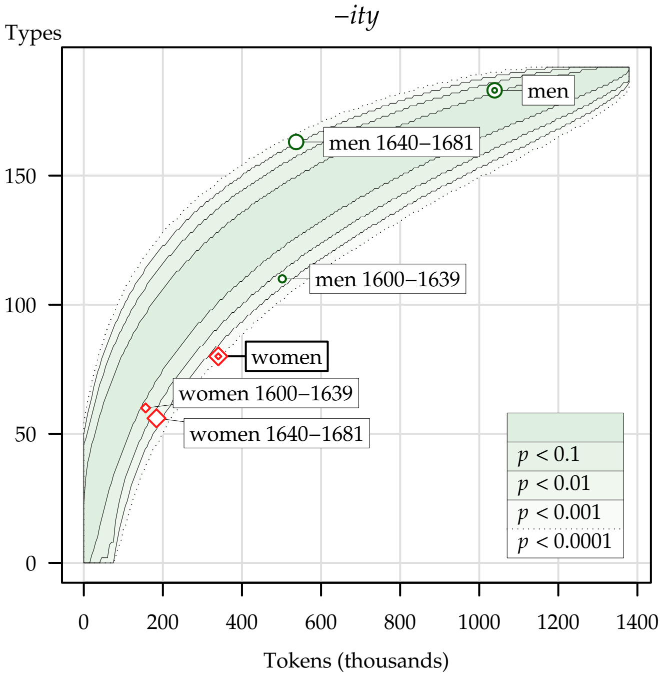
Two permutations and their type accumulation curves

Each tick mark represents the addition of one randomly selected piece of the corpus. Each piece consists of one individual's letters from a 20-year period.

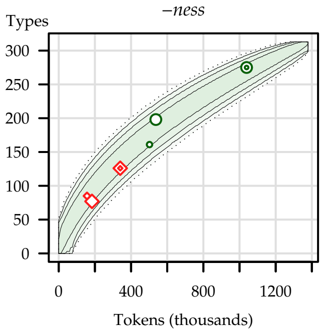
- Combine the accumulation curves and compute **nonparametric** upper and lower bounds at different levels of statistical significance
- Plot desired subcorpora (e.g., the one consisting of women's texts) on the graph to see whether their type counts differ significantly from the corpus as a whole

3 Results

- We computed upper and lower bounds of type accumulation for *-ity* using a total of one million permutations
- There were significantly few *-ity* types in the subcorpus that consists of women's letters, so our hypothesis was confirmed (p-value < 0.001)
  - It would seem that women use *-ity* much less variously than people in general in this corpus



- We also looked at the native suffix *-ness* (as in *generousness*)
- Here there was no significant difference between the subcorpus consisting of women's letters and the corpus as a whole
  - This seems natural, as little education is necessary for a person to be able to use *-ness*



References

- **Corpus of Early English Correspondence (1998)** Compiled by the Sociolinguistics and Language History project team (Nevalainen, T., H. Raumolin-Brunberg, J. Keränen, M. Nevala, A. Nurmi, M. Palander-Collin) in the Department of English, University of Helsinki. A parsed version of the corpus, published in 2006, is available through the Oxford Text Archive and ICAME.

Tanja Säily · Research Unit for Variation, Contacts and Change in English · Department of English, P.O. Box 24, FI-00014 University of Helsinki, Finland
Jukka Suomela · Helsinki Institute for Information Technology HIIT · Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Finland
Email: tanja.saily@helsinki.fi, jukka.suomela@cs.helsinki.fi

Software available at: ***http://www.cs.helsinki.fi/jukka.suomela/types/***