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?
PREVIOUS MEASURES OF PRODUCTIVITY

• Baayen (1992, 2009): measures based on frequencies of types ($V$), tokens ($N$) and hapax legomena ($n_1$, $h$)
• Realised productivity $V = \text{type frequency}$
• 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
• 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
Running words (millions)

Types

-ity

Women

Men

$p < 0.1$

$p < 0.01$

$p < 0.001$

$p < 0.0001$

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)
NEW IMPLEMENTATION: \texttt{types2}

• Open-access tool (Suomela 2014)

• Facilitates \textit{exploration}: interactive images, hyperlinks
  • Results also provided as tables and static figures

• Provides actual $p$-values, \textbf{false discovery rate control}
  (Benjamini & Hochberg 1995)
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
1600–1681
Corpus: ceec+ceee-1680-1800-person-period ...person-period-relcode

Dataset: ity ness

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

Axes: types/tokens types/running words
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
Statistics: 1105 tokens 123 types 0.000954 below
1680–1800
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
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
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)
CONCLUSION

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

- *types2*: 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
REFERENCES


• 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/


