Variation in morphological productivity in the BNC:
Sociolinguistic and methodological considerations

Tanja Säily, University of Helsinki
9 October 2009

In collaboration with Dr. Jukka Suomela,
Helsinki Institute for Information Technology HIIT
Introduction

- *-ness and -ity*
  - Roughly synonymous suffixes
  - Typically form abstract nouns from adjectives: *productive → productiveness, productivity*

- Sociolinguistics
  - Do men and women use these suffixes differently in present-day English?

- Methodology
  - Are hapax-based productivity measures valid?
Material

- **British National Corpus (BNC)**
  - 100 million words: ~90% written, ~10% spoken

- Demographically sampled spoken component (BNC-DS)
  - 4.2 million words from early 1990s
  - Gender known for 88% of the data, social class for 62% (2.6 million words)

- Written component (BNC-W)
  - 88 million words, 1960s–1990s
  - Gender known for 51% of the data (45 Mw)
Methods

- How to measure productivity?
  - Count the number of different words (types)
  - Count the number of words occurring only once (hapax legomena, or hapaxes)
    - Approximating ‘new’ words

- Comparing type counts from subcorpora
  - Normalisation problematic, establishing statistical significance likewise
  - Permutation testing: take samples in random order and see how types accumulate, 1M times
CEEC

- *ity* types vs. running words

$p < 0.1$

$p < 0.01$

$p < 0.001$

$p < 0.0001$
Sociolinguistics: Related work

- Productivity of *-ity* significantly low in 17th-century letters written by women
  - *Corpus of Early English Correspondence* (CEEC), Säily & Suomela (2009)
  - *-ity* ‘learned’, etymologically foreign; women less well educated than men → less able to use *-ity*?

- Women favour pronouns over common nouns
  - Rayson et al. 1997 (BNC-DS), Argamon et al. 2003 (BNC-W), Säily et al. forthcoming (CEEC)
Sociolinguistics: BNC-DS

- Productivity of both -*ity* and -*ness* significantly low in women’s speech
  - Expected result
    - Women’s style more interactive
  - -*ity*: difference just about significant
  - -*ness*: gender difference tied to social class
Tanja Säily, Variation in morphological productivity in the BNC
BNC-DS - *ness* types vs. running words

\[ m \text{ C2+DE} \quad f \text{ C2+DE} \]

- \( p < 0.1 \)
- \( p < 0.01 \)
- \( p < 0.001 \)
- \( p < 0.0001 \)
Productivity of -ity (but not -ness) significantly low in women’s writing

- Holds for both imaginative (BNC-W_{imag}) and informative (BNC-W_{inf}) texts

- Result for -ity expected; negative result for -ness requires more research

- Semantics of -ness? ‘Embodied attribute/trait’ goes well with interactive writing style

  - Could also apply to 17th-century results
BNC-$W_{\text{imag}}$ - *ity* types vs. running words

- $p < 0.1$
- $p < 0.01$
- $p < 0.001$
- $p < 0.0001$
BNC-$W_{\text{inf}}$ - *ity* types vs. running words

- $p < 0.1$
- $p < 0.01$
- $p < 0.001$
- $p < 0.0001$
Methodology: Related work

- Baayen (e.g., 1993)
  - Category-conditioned degree of productivity
    \[ P = \frac{n_1}{N} \]
  - Hapax-conditioned degree of productivity
    \[ P^* = \frac{n_1}{h} \text{ (or, within the same corpus, just } n_1) \]

- CEEC: hapax accumulation curves
  (Säily & Suomela 2009)
  - Confidence intervals too wide
- *ity* hapaxes vs. running words

![Graph showing variation in morphological productivity in the BNC with CEEC.](image)

**Legend**:
- $p < 0.1$
- $p < 0.01$
- $p < 0.001$
- $p < 0.0001$
Methodology: BNC study

- BNC-W: hapax accumulation curves
  - More data $\rightarrow$ narrower confidence intervals
    - Results look similar to type accumulation curves but less significant
  - However, the number of hapaxes does not grow linearly with either corpus size or the number of suffix tokens
    - Comparing $P$ figures can be unreliable unless the sizes of the subcorpora / numbers of suffix tokens are of a similar magnitude
- *ity* hapaxes vs. running words
- *ity* hapaxes vs. suffix tokens

**BNC-W$$^{\text{inf}}$$**
Conclusion

- There can be sociolinguistic variation in morphological productivity.

- There seem to be gendered speech styles and writing styles in English (possibly relatively stable over centuries).

- There is no perfect solution for measuring productivity as of yet.
References


References (cont.)

