Measures of productivity and lexical diversity

Stefan Evert

Applications
- morphological productivity (Baayen 1992; Evert & Lüdeling 2005)
- vocabulary size estimation (Gani 1975; Efron & Thisted 1976)
- stylometry & authorship attribution (Juola 2006)
- (diachronic) syntactic productivity (Zeldes 2012; Bentz et al. 2014)
- readability & text difficulty (Grässer et al. 2004)
- dementia detection (Garrard et al. 2005; Le et al. 2011)
- feature in register studies (Biber 1988, Neumann & Evert 2017)
- Bayesian prior for probabilistic models, e.g. in collocation analysis

Type-token statistics
- \( N \) = sample size (# tokens)
- \( V \) = number of distinct types
- \( V_m \) = # types with \( f_i = m \) (spectrum)
- \( \tau_i \) = rel. frequency of type \( i \)

Productivity measures
- type-token ratio
  \[ \text{TTR} = \frac{N}{V} \]
- Baayen’s productivity index
  \[ P = \frac{V_i}{N} \]
- Zipf-Mandelbrot slope
  \[ \pi_j = \frac{C}{(i+b)^a} \]
- Sichel (1975)
  \[ S = \frac{V_i}{V} \]
- Honoré (1979)
  \[ H = \frac{\log N}{1 - \pi} \]

LNRE models (Baayen 2001)
- continuous approximation \( g(\pi) \) of discrete type probabilities \( \tau_i \)
- expectations & variances of frequency spectrum as integrals
- particularly convenient for the Zipf-Mandelbrot law (Evert 2004)

Simulation studies
- evaluate measures on random samples from known population
- based on expected frequency spectrum
- better: average over many samples → mean + variance

effect of sample size

Interpretation & perturbations
- lack of clear intuitive interpretation
- measures distorted by noise (e.g. typos)
- and presence of frequent lexicalized types

Significance tests & non-randomness
- problems: Zipf-Mandelbrot population, parameter estimation, non-randomness
- "echo" model (Baroni & Evert 2007)
- use document frequency (df) or substitute any repetition within lag window by type echo
- validated on extrapolation
- cross-validation (Evert et al. 2017)
- combines bootstrapping & cross-validation
- corrects for size effects and non-randomness
- confidence intervals & significance testing
- validated on Iris Murdoch (Le & al 2011)

Work in progress
- more complex LNRE models
- better parameter estimation with conf. int.
- mixture distributions → account for noise
- interpretability of productivity measures
- validation of bootstrap by LNRE simulation