Day 2: Textual Data, Sampling, and Working with Texts

Kenneth Benoit

Essex Summer School 2011

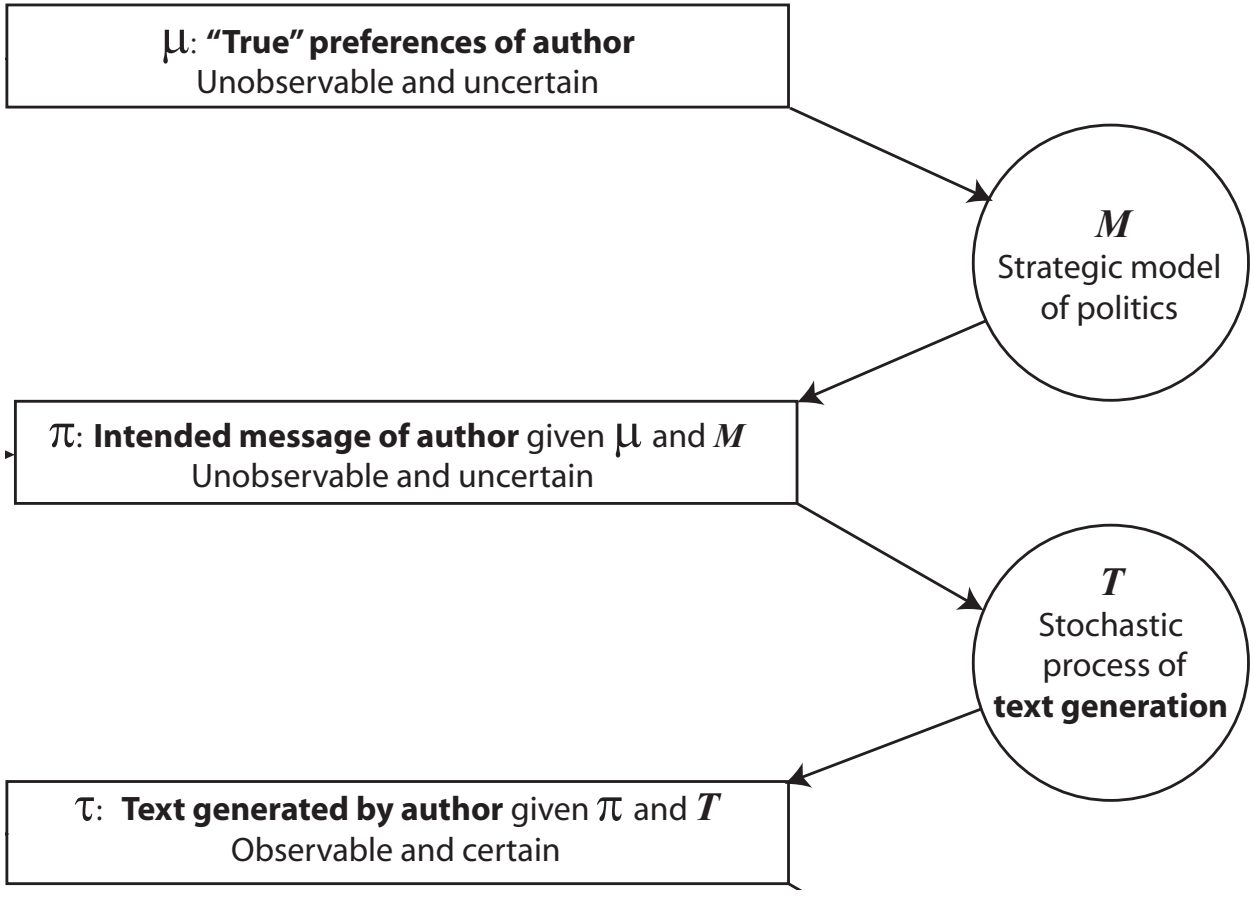
July 12, 2011

Strategies for selecting units of textual analysis

- Words
- *n*-word sequences
- pages
- paragraphs
- Themes
- Natural units (a speech, a poem, a manifesto)
- Key: depends on the research design

Sample v. "population"

- Basic Idea: Observed text is a stochastic realization
- Systematic features shape most of observed verbal content
- Non-systematic, random features also shape verbal content



Sampling strategies for selecting texts

- Difference between a sample and a population
- May not be feasible to perform any sampling
- May not be necessary to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"
- Different types of sampling vary from random to purposive
  - random sampling
  - non-random sampling
- Key is to make sure that what is being analyzed is a valid representation of the phenomenon as a whole – a question of research design

Random versus "Constructed" Sampling

- Based on a study by Riffe, Aust and Lacy (1993), who compared sampling from newspaper articles randomly versus "constructed"
- Either randomly sample 7 consecutive days, or between 2–4 consecutive weeks, and compare to "known" quantities
- Study showed that constructed sampling is much more efficient
- Why? Because cyclic variation in newspaper content occurs according to the day of the week – not every day contains equal proportions of different content

Word frequency examples

- Variations use vocabulary diversity analysis (e.g. Labbé et. al. 2004)

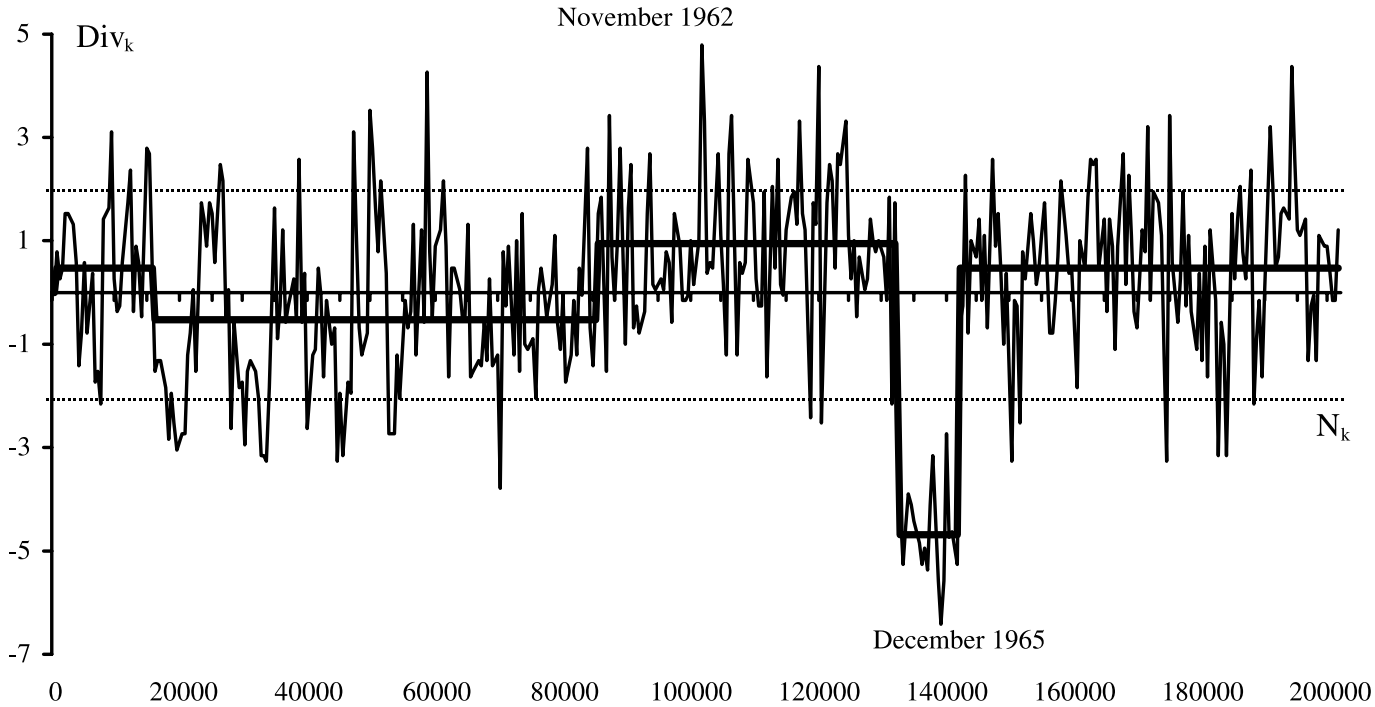
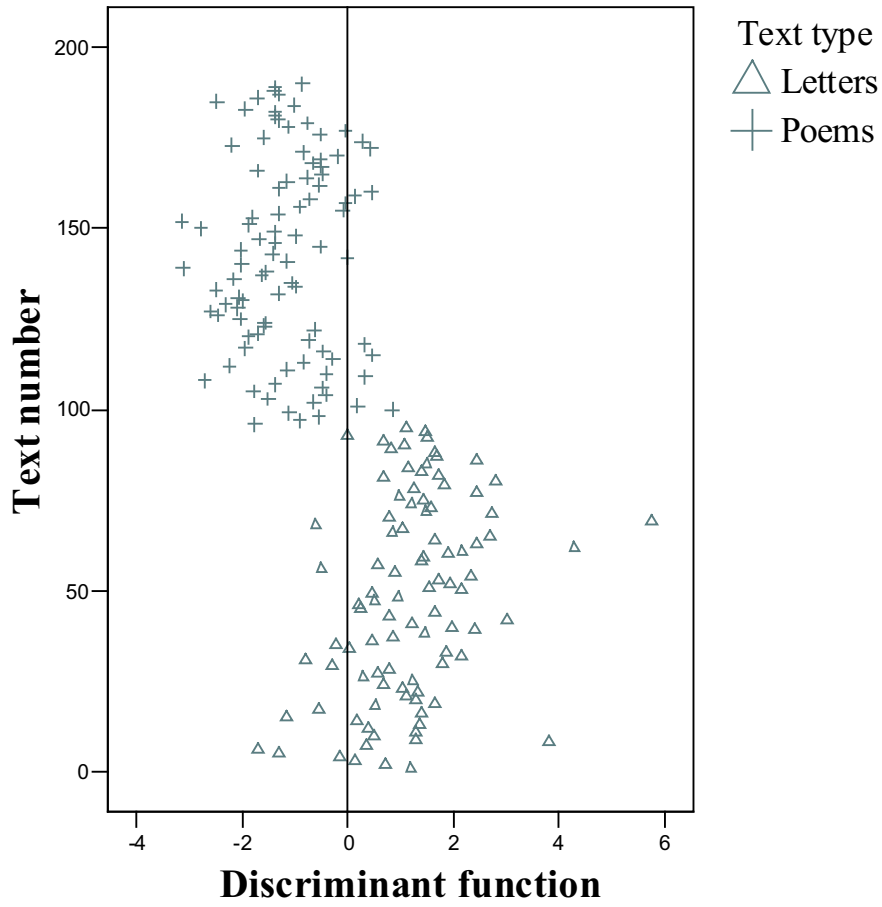


Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

Examples continued

- Word *length* (defined as number of syllables) can be indicative of genre, if not necessarily authorship (Kelih et. al. 2004)



General Issues

1. Validity: does a measurement reflect the truth of what is being measured?
2. Reliability: does repetition of a research procedure produce stable results?
3. Replicability: can a text analysis procedure be repeated at all?
4. Uncertainty: what is the variability of our estimates?
5. Precision: How exact are the estimates from our procedure?
6. Accuracy: How closely do our estimates correspond to the truth?

Practical issues working with texts

File formats How the electronic text is formatted

Conversion Converting files from one format to another

Pre-analysis text processing
- *stemming* (lemmatization)
- reducing infrequent words
- "stop lists" for most frequent words

Dataset generation How to convert text files into "datasets"

Software preview

- Jfreq
- Yoshikoder
- MaxQDA
- Stata and Wordscores library
- R and `austin` library
- Other programs