Quantitative Measures of Productivity and their significance

a work-in-progress report (sorry!)

Stefan Evert

Institute of Cognitive Science
University of Osnabrück, Germany
stefan.evert@uos.de

Birmingham, 22 July 2011

Outline

What we want to measure

- Productivity: qualitative vs. quantitative
  - productivity of morphological word-formation rules (e.g. Schultink 1961; Baayen 1992; Evert and Lüdeling 2001)
  - also lexico-grammatical patterns (➜ construction grammar), collocational patterns, word senses, ...
- Vocabulary richness
  - stylometrics & register variation (Baayen 2001, 184–191)
  - authorship attribution (cf. Juola 2006)
  - Zipfian prior for statistical inference (Evert and Pipa 2010)
- Size of the (potential) vocabulary
  - How many words did Shakespeare know? (Efron and Thisted 1976) — And how many typos are there on the Internet?
  - coverage estimation of NLP grammars, dictionaries, ...
  - early indicator for Alzheimer's disease (Garrard *et al.* 2005)

Example data

- **Bare singulars** in English
  - *(go) to school*, *(live) at home*, *(do) by hand*, *(come) into effect*, *(draw) to scale*, *(fall) in line*, *(require) by law*, ...
  - some authors claim that these are lexicalised exceptions (esp. in German, cf. counter-argument by Kiss (2007))
- Corpus evidence
  - Brown corpus, spoken BNC, written BNC
  - automatic extraction of V + Prep + N sequences
  - only count nouns with $\geq 15\%$ plural occurrences in BNC
  - no manual correction (yet)
- Data extracted with CQP query
  - ```
    [class = "VERB"] @[pos = "PR[PF]"]
    [pos = "NN.*"]* [pos = "NN1" & hw = $countable]
    [: pos != "CRD|NN.*" :]
    :: match.text_mode = "spoken";
    ```
    ```

Example data

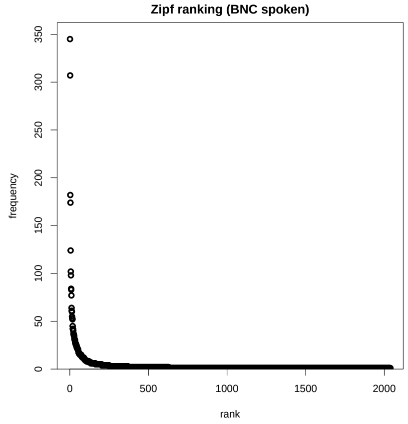
	tokens	types
Brown	1,005	651
BNC spoken	6,766	2,039
BNC written	85,750	12,876

☞ spoken BNC will be used in most of the following examples

type	f
at home	345
to school	307
at school	182
on holiday	174
in charge	124
for example	102
...	...
on trial	14
in agreement	14
...	...
on target	5
of value	5
with tax	5
...	...
per second	1
within reach	1
against noise	1
into hock	1

Corpus evidence for productivity

- evidence for productivity, type richness and vocabulary size: **type-token statistics**
- large number of types + many low-frequency types ➜ high degree of productivity
- often shown as **Zipf ranking** with typical L-shape

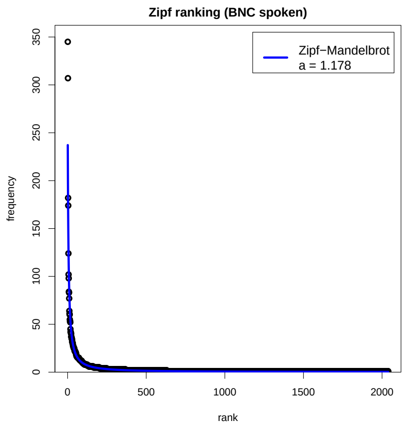


Corpus evidence: Zipf ranking

- described by **Zipf's law**
- popular Zipf-Mandelbrot version (Mandelbrot 1962)

$$f_k = \frac{C}{(k+b)^a}$$

with "slope" $a \geq 1$



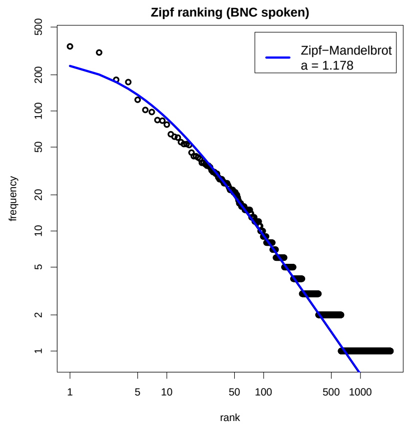
Corpus evidence: Zipf ranking

- described by **Zipf's law**
- popular Zipf-Mandelbrot version (Mandelbrot 1962)

$$f_k = \frac{C}{(k+b)^a}$$

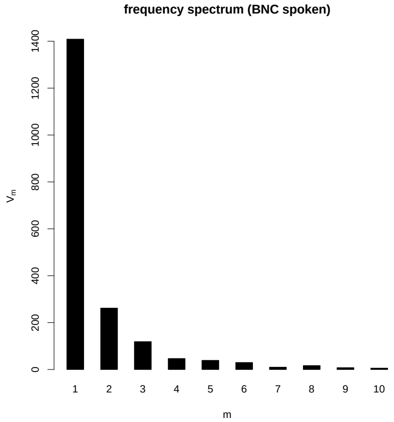
with "slope" $a \geq 1$

- easily visible as straight line in log-log plot

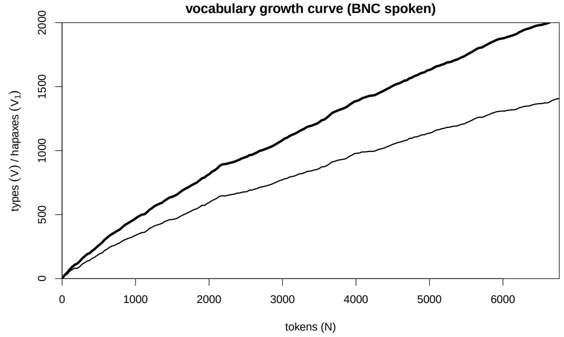


Corpus evidence: Frequency spectrum

- low-frequency types are better captured by the **frequency spectrum**
- **class size** V_m = number of types that occur m times
- V_1 = hapax legomena
- V_2 = dis legomena

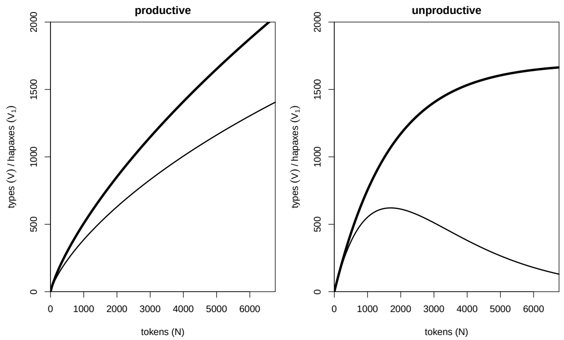


Corpus evidence: Vocabulary growth



- **vocabulary growth curve** shows how number of seen types increases across corpus (+ hapaxes, dis legomena, ...)
- plot number of seen types V against number of tokens N
- slope of VGC = how often new type is encountered

Corpus evidence: Vocabulary growth



- **vocabulary growth curve** shows how number of seen types increases across corpus (+ hapaxes, dis legomena, ...)
- plot number of seen types V against number of tokens N
- slope of VGC = how often new type is encountered
- **population size** $S = \lim_{N\to\infty} V(N)$ = potential vocabulary

Quantitative measures of productivity

(see Baayen 2001, 24–30)

- Yule (1944) / Simpson (1949)

$$K = 10\,000 \cdot \frac{\sum_m m^2 V_m - N}{N^2}$$

- Guiraud (1954)

$$R = \frac{V}{\sqrt{N}}$$

- Sichel (1975)

$$S = \frac{V_2}{V}$$

- Herdan's law (1964)

$$C = \frac{\log V}{\log N}$$

- Baayen's productivity index (slope of vocabulary growth curve)

$$\mathcal{P} = \frac{V_1}{N}$$

- TTR = token-type ratio

$$\mathrm{TTR} = \frac{N}{V}$$

- Zipf-Mandelbrot slope

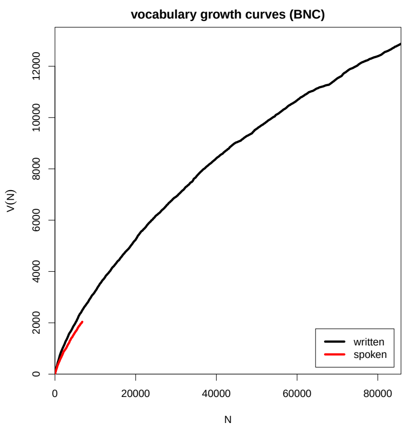
$$a$$

- population size

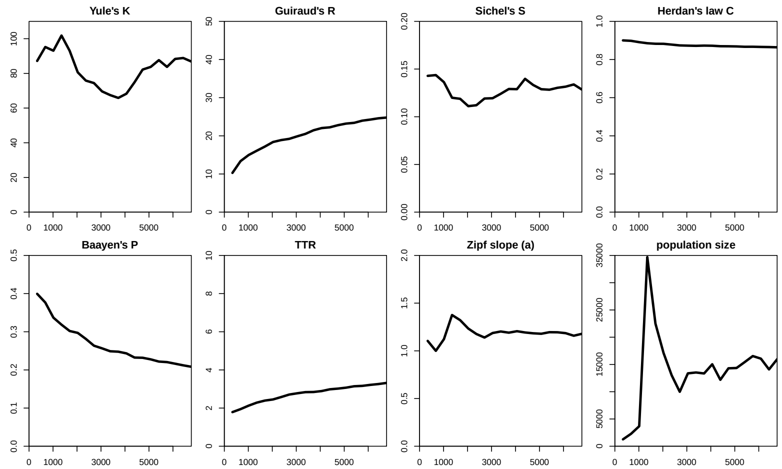
$$S = \lim_{N\to\infty} V(N)$$

Productivity measures for bare singulars in the BNC

	spoken	written
V	2,039	12,876
N	6,766	85,750
K	86.84	28.57
R	24.79	43.97
S	0.13	0.15
C	0.86	0.83
$\mathcal{P}$	0.21	0.08
TTR	3.32	6.66
a	1.18	1.27
pop. S	15,958	36,874



Are these “lexical constants” really constant?

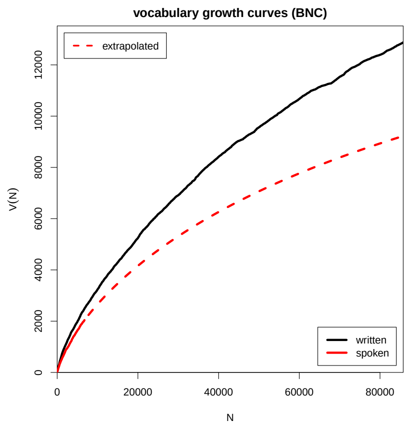


Key problems

- Comparability (➜ **corpus size**)
  - do measures depend systematically on corpus size?
- **Sampling variation**
  - significance tests for differences, confidence intervals
- Non-randomness **(➜ Baroni and Evert 2005, 2007)**
- Manual data correction
  - not feasible for large samples, e.g. 85,750 types in BNC
- Interpretation of productivity measures
  - productivity vs. vocabulary richness vs. size of vocabulary
  - does any measure match our intuition of productivity?

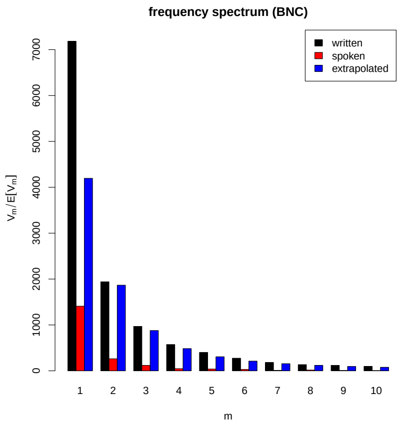
Extrapolation with LNRE models

- direct comparison of written vs. spoken BNC not possible
  - ☞ productivity measures need to be perfectly size-invariant
  - ☞ or sample size has to be adjusted (to larger sample)
- use statistical **LNRE models** (Khmaladze 1987; Baayen 2001; Evert 2004a,b) to extrapolate vocabulary growth



Extrapolation with LNRE models

- direct comparison of written vs. spoken BNC not possible
  - ☞ productivity measures need to be perfectly size-invariant
  - ☞ or sample size has to be adjusted (to larger sample)
- use statistical **LNRE models** (Khmaladze 1987; Baayen 2001; Evert 2004a,b) to extrapolate vocabulary growth
- extrapolation of frequency spectrum also possible



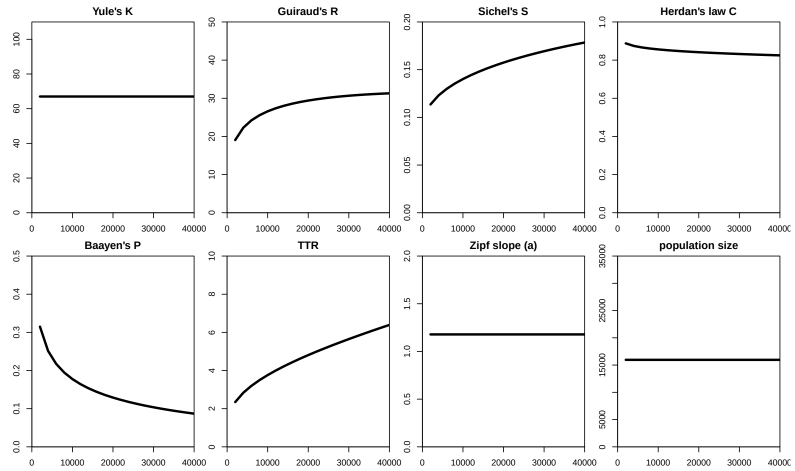
LNRE models as a methodological research tool

- LNRE models can also help us to learn more about the properties of productivity measures
- Separate variability of measures into
  1. size dependency (➜ expected spectrum for different N)
  2. sampling variation (➜ parametric bootstrap samples)

  under controlled conditions
- Quantify sampling variation ➜ significance tests, etc.
- Mature & user-friendly implementation for Gnu R in the **zipfR** package (Evert and Baroni 2007)

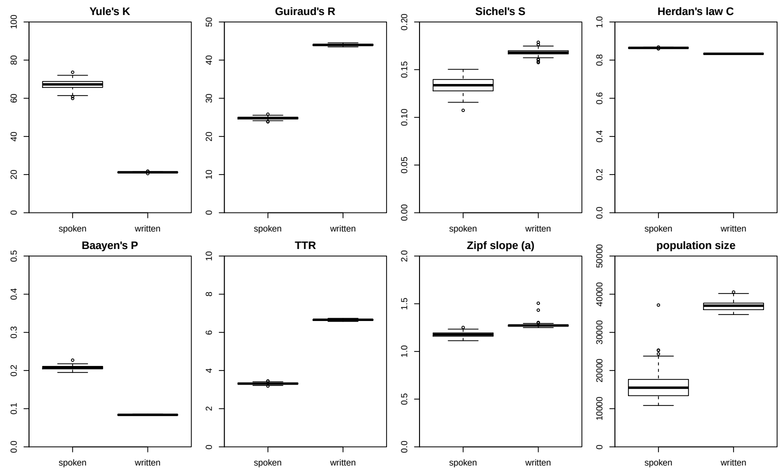
Which measures are size-invariant?

expected frequency spectrum factors out effects of sampling variation



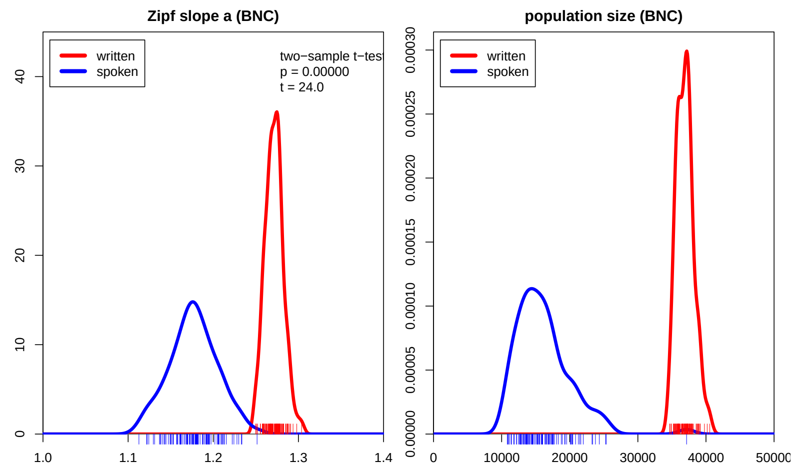
How much are measures affected by sampling variation?

are the differences between spoken and written BNC significant?



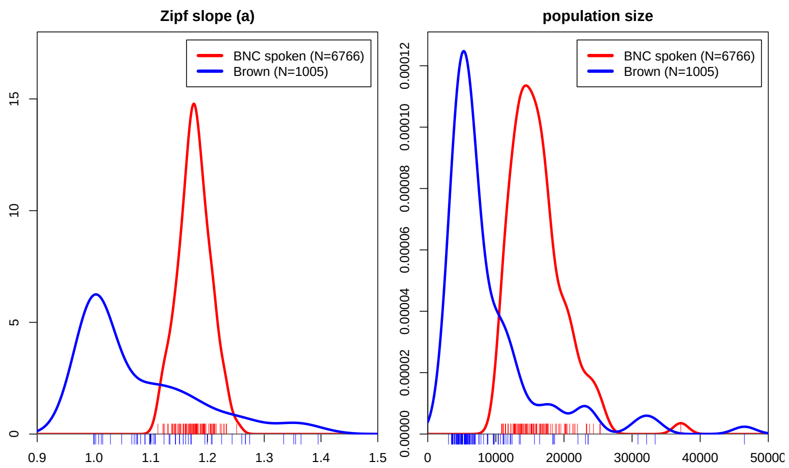
How much are measures affected by sampling variation?

Zipf slope and population size estimated from trained LNRE model



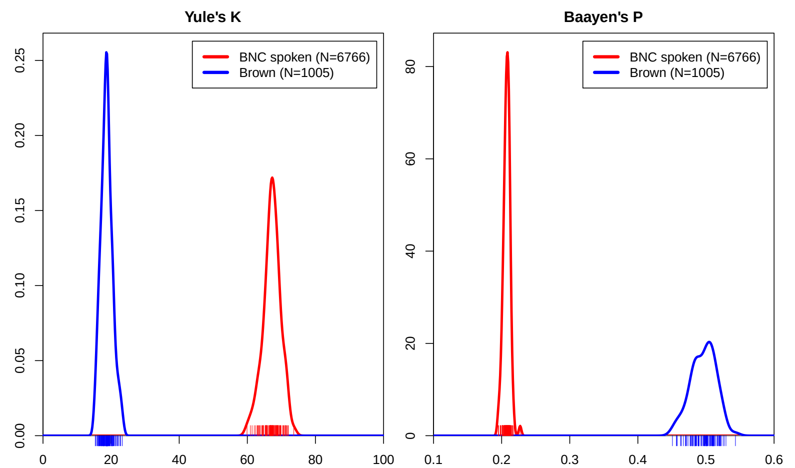
Sample size matters!

Brown corpus is too small for reliable LNRE parameter estimation



Sample size matters!

other productivity measures seem to be more robust



Thank you!

- ☞ There's much work to be done, of course!
- ☞ Talk about interpretation of measures in the coffee break?

References I

Baayen, R. Harald (1992). Quantitative aspects of morphological productivity. In G. Booij and J. van Marle (eds.), *Yearbook of Morphology 1991*, pages 109 – 150. Foris, Dordrecht.

Baayen, R. Harald (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Baroni, Marco and Evert, Stefan (2005). Testing the extrapolation quality of word frequency models. In P. Danielsson and M. Wagenmakers (eds.), *Proceedings of Corpus Linguistics 2005*, volume 1 of *The Corpus Linguistics Conference Series*. ISSN 1747-9398.

Baroni, Marco and Evert, Stefan (2007). Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 904–911, Prague, Czech Republic.

Efron, Bradley and Thisted, Ronald (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**(3), 435–447.

Evert, Stefan (2004a). A simple LNRE model for random character sequences. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2004)*, pages 411–422, Louvain-la-Neuve, Belgium.

References II

Evert, Stefan (2004b). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from `http://www.collocations.de/phd.html`.

Evert, Stefan and Baroni, Marco (2007). *zipfR*: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, pages 29–32, Prague, Czech Republic.

Evert, Stefan and Lüdeling, Anke (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*, pages 167–175, Lancaster. UCREL.

Evert, Stefan and Pipa, Gordon (2010). Probability estimation of rare events in linguistics and computational neuroscience. Presentation at the KogWis 2010 Conference, Potsdam, Germany.

Garrard, Peter; Maloney, Lisa M.; Hodges, John R.; Patterson, Karalyn (2005). The effects of very early alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, **128**(2), 250–260.

References III

Juola, Patrick (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3), 233–334.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events. Technical Report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.

Kiss, Tibor (2007). Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. *Zeitschrift für Sprachwissenschaft*, **26**(2), 317–345.

Mandelbrot, Benoit (1962). On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson (ed.), *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI.

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren*, pages 110 – 125.