The impossibility of measuring productivity in small samples

Stefan Evert
FAU Erlangen-Nürnberg

MISC Workshop @ HU Berlin
18 May 2018

Tokens & types

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 15$: number of **tokens** = sample size
- $V = 7$: number of distinct **types** = **vocabulary size** (*recently*, *very*, *not*, *otherwise*, *much*, *merely*, *now*)

Tokens & types

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 15$: number of **tokens** = sample size
- $V = 7$: number of distinct **types** = **vocabulary size** (*recently*, *very*, *not*, *otherwise*, *much*, *merely*, *now*)

type-frequency list

w	f_w
recently	1
very	5
not	3
otherwise	1
much	2
merely	2
now	1

Zipf ranking

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 15$: number of **tokens** = sample size
- $V = 7$: number of distinct **types** = **vocabulary size** (*recently*, *very*, *not*, *otherwise*, *much*, *merely*, *now*)

Zipf ranking

w	r	f_r
very	1	5
not	2	3
merely	3	2
much	4	2
now	5	1
otherwise	6	1
recently	7	1

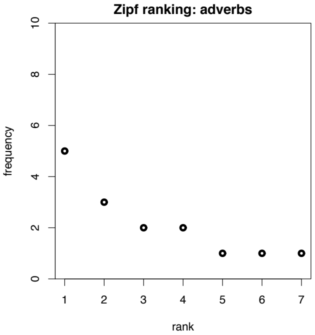
Zipf ranking

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 15$: number of **tokens** = sample size
- $V = 7$: number of distinct **types** = **vocabulary size** (*recently*, *very*, *not*, *otherwise*, *much*, *merely*, *now*)

Zipf ranking

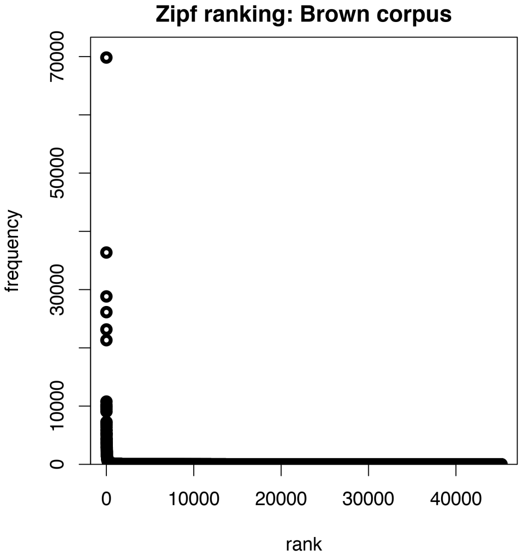
w	r	f_r
very	1	5
not	2	3
merely	3	2
much	4	2
now	5	1
otherwise	6	1
recently	7	1



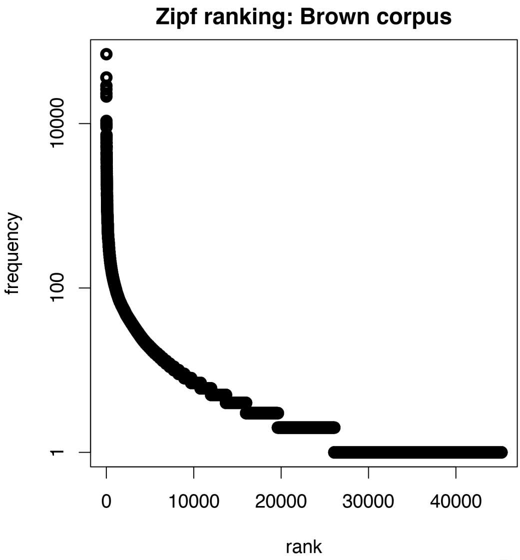
A realistic Zipf ranking: the Brown corpus

top frequencies			bottom frequencies		
r	**f**	**word**	**rank range**	**f**	**randomly selected examples**
1	69836	the	7731 – 8271	10	schedules, polynomials, bleak
2	36365	of	8272 – 8922	9	tolerance, shaved, hymn
3	28826	and	8923 – 9703	8	decreased, abolish, irresistible
4	26126	to	9704 – 10783	7	immunity, cruising, titan
5	23157	a	10784 – 11985	6	geographic, lauro, portrayed
6	21314	in	11986 – 13690	5	grigori, slashing, developer
7	10777	that	13691 – 15991	4	sheath, gaulle, ellipsoids
8	10182	is	15992 – 19627	3	mc, initials, abstracted
9	9968	was	19628 – 26085	2	thar, slackening, deluxe
10	9801	he	26086 – 45215	1	beck, encompasses, second-place

A realistic Zipf ranking: the Brown corpus



A realistic Zipf ranking: the Brown corpus



Frequency spectrum

- pool types with $f = 1$ (hapax legomena), types with $f = 2$ (dis legomena), ..., $f = m$, ...
- $V_1 = 3$: number of hapax legomena (*now, otherwise, recently*)
- $V_2 = 2$: number of dis legomena (*merely, much*)
- general definition: $V_m = |\{w \mid f_w = m\}|$

Zipf ranking

w	r	f_r
very	1	5
not	2	3
merely	3	2
much	4	2
now	5	1
otherwise	6	1
recently	7	1

frequency spectrum

m	V_m
1	3
2	2
3	1
5	1

Frequency spectrum

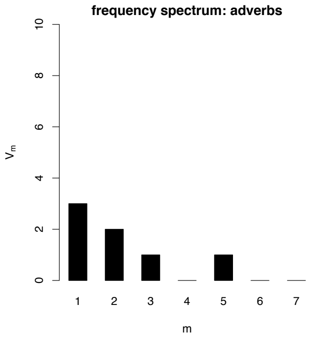
- pool types with $f = 1$ (hapax legomena), types with $f = 2$ (dis legomena), ..., $f = m$, ...
- $V_1 = 3$: number of hapax legomena (*now, otherwise, recently*)
- $V_2 = 2$: number of dis legomena (*merely, much*)
- general definition: $V_m = |\{w \mid f_w = m\}|$

Zipf ranking

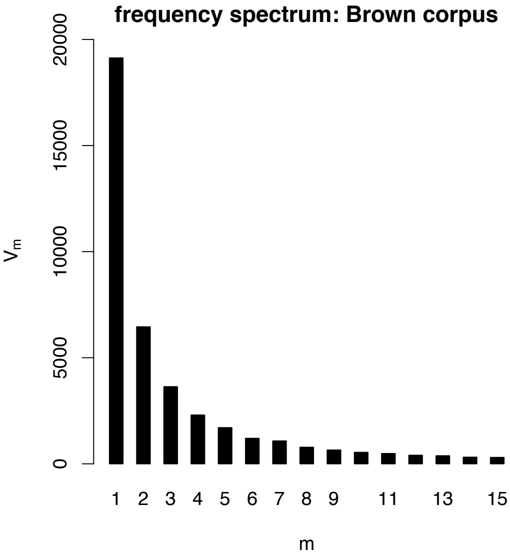
w	r	f_r
very	1	5
not	2	3
merely	3	2
much	4	2
now	5	1
otherwise	6	1
recently	7	1

frequency spectrum

m	V_m
1	3
2	2
3	1
5	1



A realistic frequency spectrum: the Brown corpus



Vocabulary growth curve

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 1,\ V(N) = 1,\ V_1(N) = 1$

Vocabulary growth curve

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 1,\ V(N) = 1,\ V_1(N) = 1$
- $N = 3,\ V(N) = 3,\ V_1(N) = 3$

Vocabulary growth curve

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 1,\ V(N) = 1,\ V_1(N) = 1$
- $N = 3,\ V(N) = 3,\ V_1(N) = 3$
- $N = 7,\ V(N) = 5,\ V_1(N) = 4$

Vocabulary growth curve

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 1, \ V(N) = 1, \ V_1(N) = 1$
- $N = 3, \ V(N) = 3, \ V_1(N) = 3$
- $N = 7, \ V(N) = 5, \ V_1(N) = 4$
- $N = 12, \ V(N) = 7, \ V_1(N) = 4$

Vocabulary growth curve

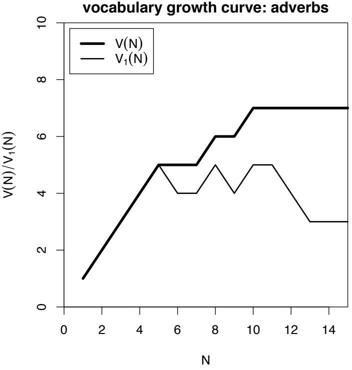
our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 1,\ V(N) = 1,\ V_1(N) = 1$
- $N = 3,\ V(N) = 3,\ V_1(N) = 3$
- $N = 7,\ V(N) = 5,\ V_1(N) = 4$
- $N = 12,\ V(N) = 7,\ V_1(N) = 4$
- $N = 15,\ V(N) = 7,\ V_1(N) = 3$

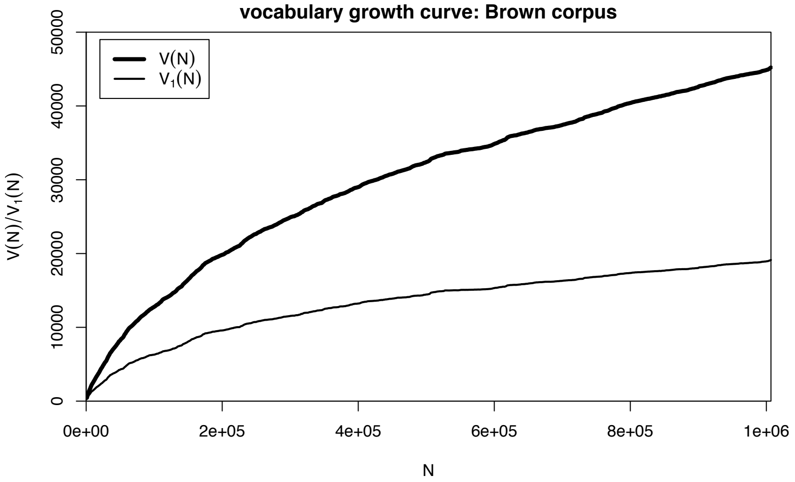
Vocabulary growth curve

our sample: *recently*, *very*, *not*, *otherwise*, *much*, *very*, *very*, *merely*, *not*, *now*, *very*, *much*, *merely*, *not*, *very*

- $N = 1,\ V(N) = 1,\ V_1(N) = 1$
- $N = 3,\ V(N) = 3,\ V_1(N) = 3$
- $N = 7,\ V(N) = 5,\ V_1(N) = 4$
- $N = 12,\ V(N) = 7,\ V_1(N) = 4$
- $N = 15,\ V(N) = 7,\ V_1(N) = 3$

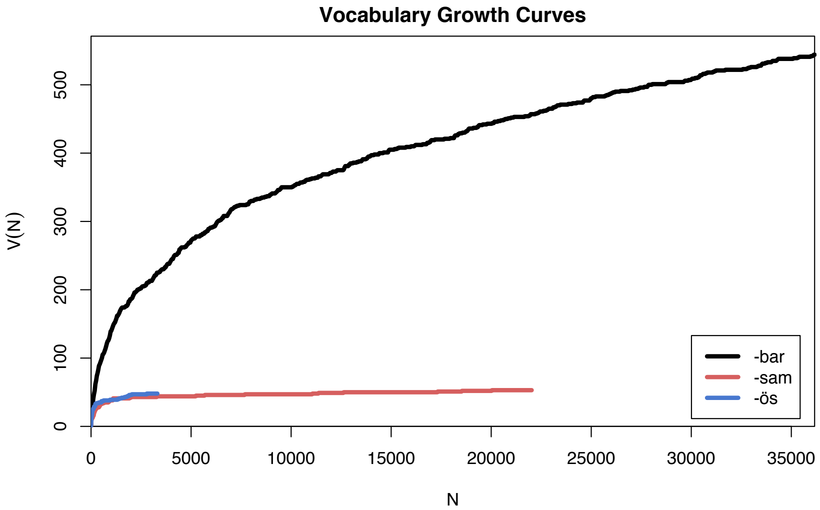


A realistic vocabulary growth curve: the Brown corpus



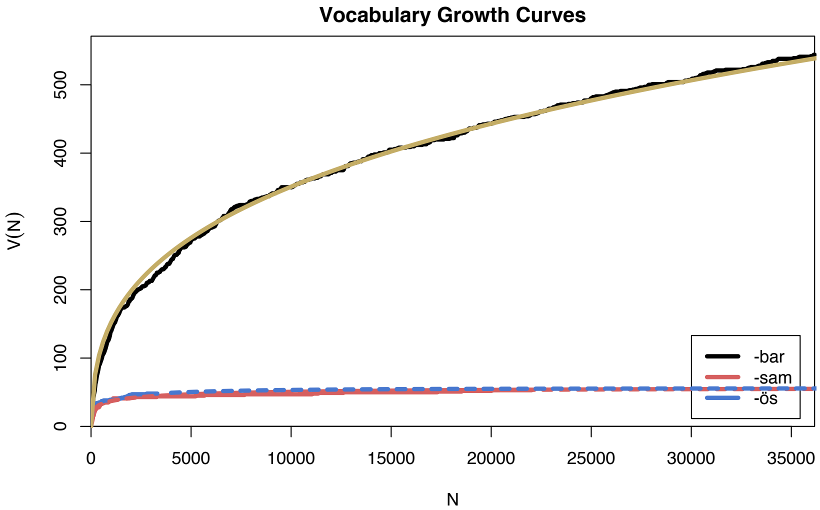
Measuring morphological productivity

example from Evert and Lüdeling (2001)



Measuring morphological productivity

example from Evert and Lüdeling (2001)



Quantitative measures of productivity

(Tweedie and Baayen 1998; Baayen 2001)

- Baayen's (1991) productivity index $\mathcal{P}$
  (slope of vocabulary growth curve)

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

- TTR = type-token ratio

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

- Population size

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

- Herdan's law (1964)

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

Quantitative measures of productivity

(Tweedie and Baayen 1998; Baayen 2001)

- Baayen's (1991) productivity index $\mathcal{P}$ (slope of vocabulary growth curve)

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

- TTR = type-token ratio

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

- Population size

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

- Herdan's law (1964)

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

- 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)

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

- Honoré (1979)

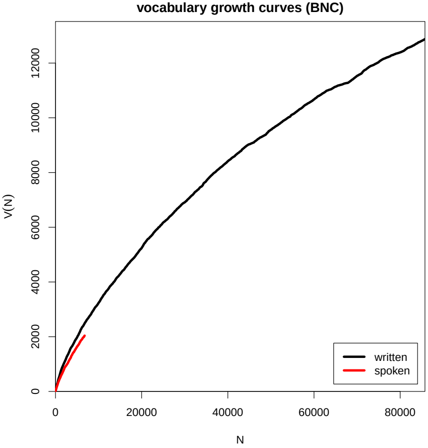
$$H = \frac{\log N}{1 - \frac{V_1}{V}}$$

Productivity measures for bare singulars in the BNC

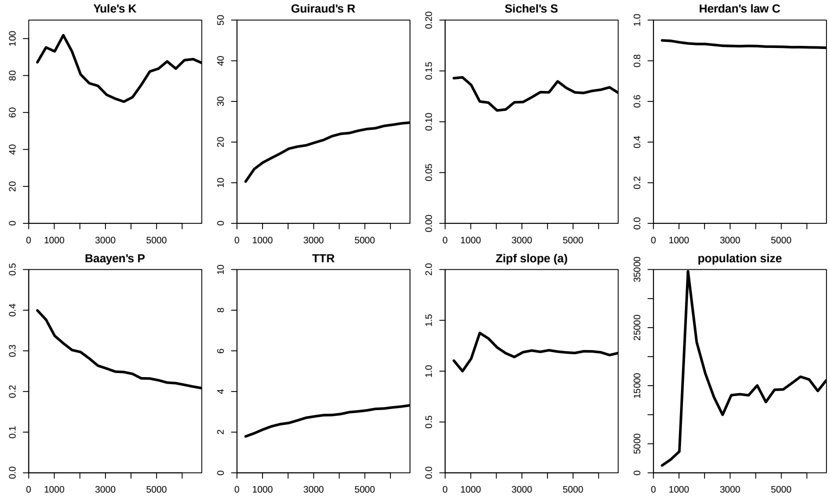
	spoken	written
V	2,039	12,876
N	6,766	85,750
K	86.84	28.57
R	24.79	43.97
S	0.13	0.15
C	0.86	0.83
$\mathcal{P}$	0.21	0.08
TTR	0.301	0.150
pop. S	15,958	36,874

Productivity measures for bare singulars in the BNC

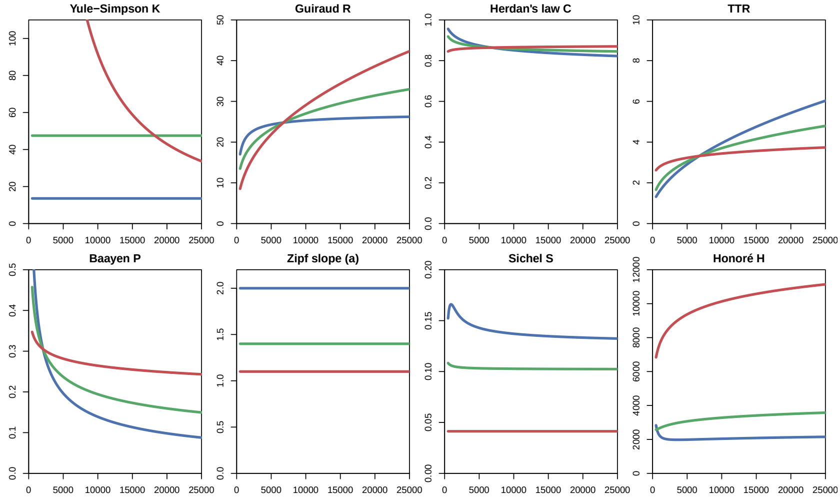
	spoken	written
V	2,039	12,876
N	6,766	85,750
K	86.84	28.57
R	24.79	43.97
S	0.13	0.15
C	0.86	0.83
$\mathcal{P}$	0.21	0.08
TTR	0.301	0.150
pop. S	15,958	36,874



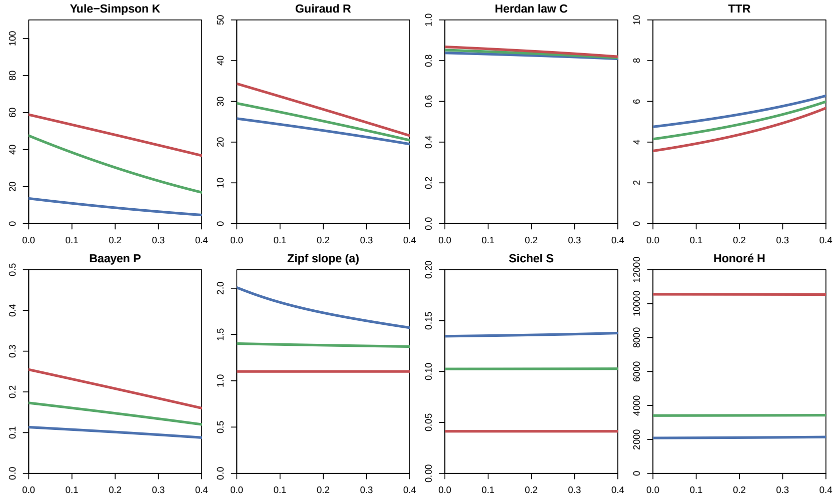
Are these "lexical constants" really constant?



Simulation: sample size



Simulation: frequent lexicalized types



LNRE models

- State-of-the-art approach to measuring productivity: **LNRE models** (Baayen 2001)
  - LNRE = Large Number of Rare Events
  - Baayen (2001) has 887 citations on Google Scholar
- Standard implementation: **zipfR** (Evert and Baroni 2007)
  - 76 citations on Google Scholar
  - only a few search results for Baayen's `lexstats` software
- LNRE uses various approximations and simplifications to obtain a tractable and elegant model
  - LNRE model usually minor component of complex procedure
  - often applied to very large samples ($N > 1$ M tokens)

The LNRE population

- Population: set of S types w_i with occurrence **probabilities** π_i
- $S =$ **population diversity** can be finite or infinite ($S = \infty$)
- Not interested in specific types ➜ arrange by decreasing probability: $\pi_1 \geq \pi_2 \geq \pi_3 \geq \cdots$
  - ☞ impossible to determine probabilities of all individual types
- Normalization: $\pi_1 + \pi_2 + \ldots + \pi_S = 1$

- **parametric** statistical **model** to describe full population (esp. for $S = \infty$), i.e. a function $i \mapsto \pi_i$
  - type probabilities π_i cannot be estimated reliably from a sample, but parameters of this function can
  - NB: population index $i \neq$ Zipf rank r

Zipf-Mandelbrot law as a population model

- Zipf-Mandelbrot law for type probabilities:

$$\pi_i := \frac{C}{(i+b)^a}$$

- Two free parameters: $a > 1$ and $b \geq 0$
  - ☞ C is not a parameter but a normalization constant, needed to ensure that $\sum_i \pi_i = 1$
- Third parameter: $S > 0$ or $S = \infty$

- This is the **Zipf-Mandelbrot** population model (Evert 2004)

Samples: type frequency list & spectrum

rank r	f_r	type i
1	37	6
2	36	1
3	33	3
4	31	7
5	31	10
6	30	5
7	28	12
8	27	2
9	24	4
10	24	16
11	23	8
12	22	14
⋮	⋮	⋮

m	V_m
1	83
2	22
3	20
4	12
5	10
6	5
7	5
8	3
9	3
10	3
⋮	⋮

sample #1

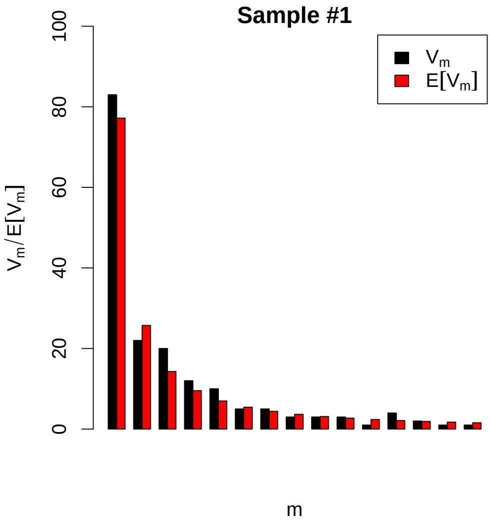
Samples: type frequency list & spectrum

rank r	f_r	type i
1	39	2
2	34	3
3	30	5
4	29	10
5	28	8
6	26	1
7	25	13
8	24	7
9	23	6
10	23	11
11	20	4
12	19	17
$\vdots$	$\vdots$	$\vdots$

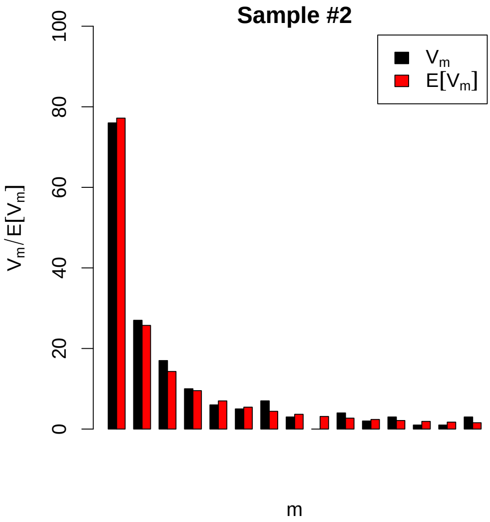
m	V_m
1	76
2	27
3	17
4	10
5	6
6	5
7	7
8	3
10	4
11	2
$\vdots$	$\vdots$

sample #2

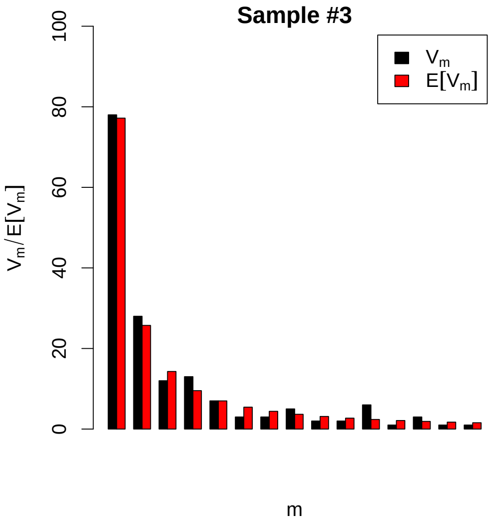
Expectation: frequency spectrum



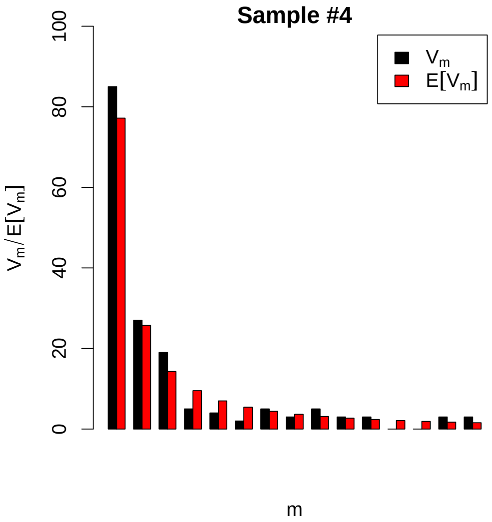
Expectation: frequency spectrum



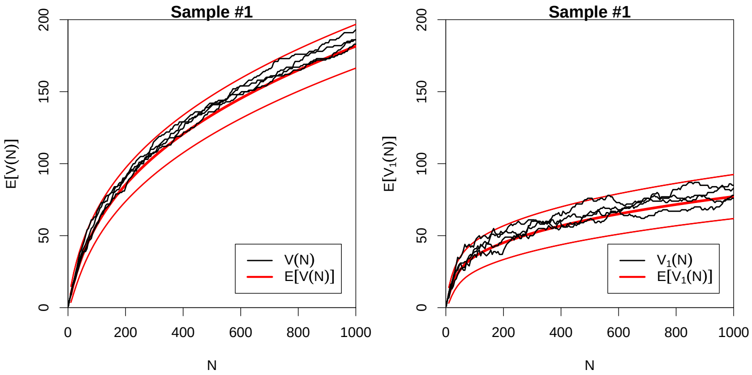
Expectation: frequency spectrum



Expectation: frequency spectrum



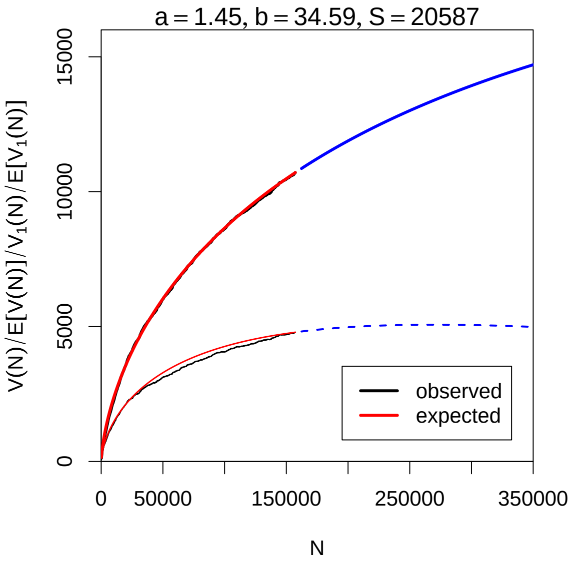
Expectation: vocabulary growth curve



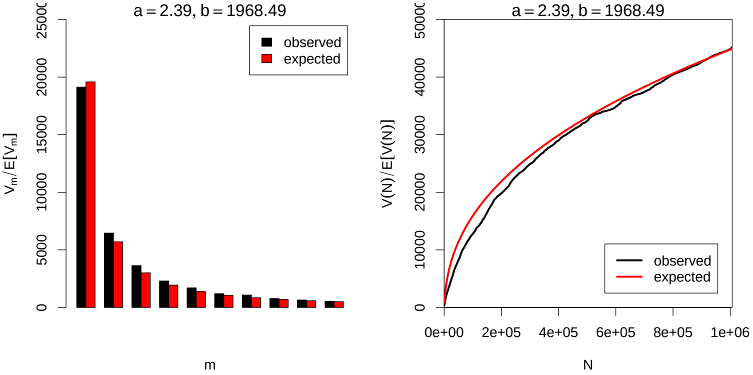
"Confidence intervals" indicate predicted sampling distribution:

☞ for 95% of samples generated by the LNRE model, VGC will fall within the range delimited by the thin red lines

Extrapolating vocabulary growth



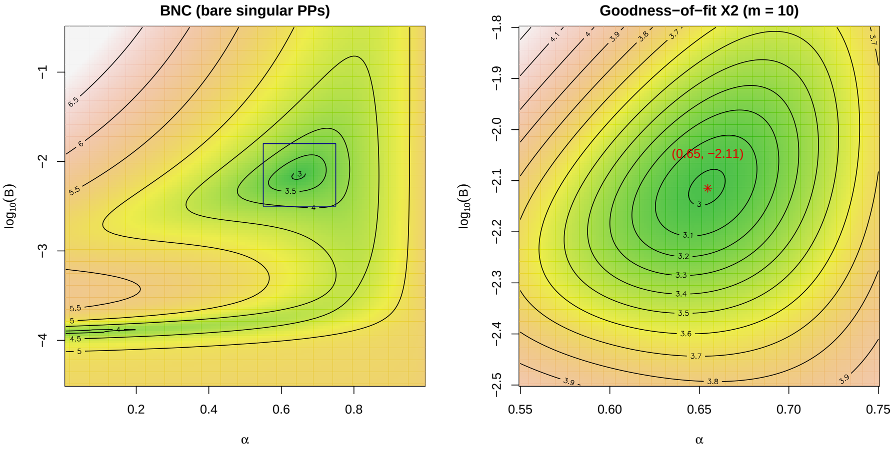
Parameter estimation



- By trial & error we found $a = 2.0$ and $b = 550$
- Automatic estimation procedure based on minimisation of suitable cost function: $a = 2.39$ and $b = 1968$

Parameter estimation

ZM model with free parameters $0 < \alpha < 1$ and $B > 0$



Problems of LNRE models

- Assumption: corpus data = random sample
  - ☞ holds reasonably well for morphological productivity
  - ☞ simple but effective ECHO correction (Baroni and Evert 2007)

Problems of LNRE models

- Assumption: corpus data = random sample
  - ☞ holds reasonably well for morphological productivity
  - ☞ simple but effective ECHO correction (Baroni and Evert 2007)
- Approximation: independent Poisson sampling
  - instead of correct multinomial sampling distribution

Problems of LNRE models

- Assumption: corpus data = random sample
  - ☞ holds reasonably well for morphological productivity
  - ☞ simple but effective ECHO correction (Baroni and Evert 2007)
- Approximation: independent Poisson sampling
  - instead of correct multinomial sampling distribution
- Approximation: multivariate normal distribution of V and V_m
  - true sampling distribution is completely intractable

Problems of LNRE models

- Assumption: corpus data = random sample
  - ☞ holds reasonably well for morphological productivity
  - ☞ simple but effective ECHO correction (Baroni and Evert 2007)
- Approximation: independent Poisson sampling
  - instead of correct multinomial sampling distribution
- Approximation: multivariate normal distribution of V and V_m
  - true sampling distribution is completely intractable
- Approximation: continuous type density function $g(\pi)$
  - instead of discrete type probabilities of Z-M law

Problems of LNRE models

- Assumption: corpus data = random sample
  - ☞ holds reasonably well for morphological productivity
  - ☞ simple but effective ECHO correction (Baroni and Evert 2007)
- Approximation: independent Poisson sampling
  - instead of correct multinomial sampling distribution
- Approximation: multivariate normal distribution of V and V_m
  - true sampling distribution is completely intractable
- Approximation: continuous type density function $g(\pi)$
  - instead of discrete type probabilities of Z-M law

➥ Wide-spread irresponsible application of LNRE models to small samples (e.g. Lüdeling and Evert 2005)

How reliable are the fitted models?

Three potential issues:

How reliable are the fitted models?

Three potential issues:

1. Model assumptions $\neq$ population
   (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - ☞ model cannot be adequate, regardless of parameter settings

How reliable are the fitted models?

Three potential issues:

1. Model assumptions $\neq$ population
   (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - ☞ model cannot be adequate, regardless of parameter settings

2. Parameter estimation unsuccessful
   (i.e. suboptimal goodness-of-fit to training data)
   - ☞ optimization algorithm trapped in local minimum
   - ☞ can result in highly inaccurate model

How reliable are the fitted models?

Three potential issues:

1. Model assumptions $\neq$ population
   (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - ☞ model cannot be adequate, regardless of parameter settings
2. Parameter estimation unsuccessful
   (i.e. suboptimal goodness-of-fit to training data)
   - ☞ optimization algorithm trapped in local minimum
   - ☞ can result in highly inaccurate model
3. Uncertainty due to sampling variation
   (i.e. training data differ from population distribution)
   - ☞ model fitted to training data, may not reflect true population
   - ☞ another training sample would have led to different parameters
   - ☞ especially critical for small samples ($N < 10{,}000$)

How reliable are the fitted models?

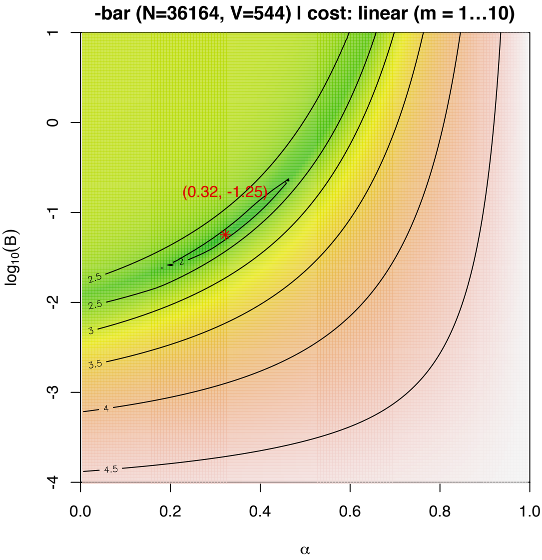
Three potential issues:

1. **Model assumptions $\neq$ population**
   (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - ☞ model cannot be adequate, regardless of parameter settings

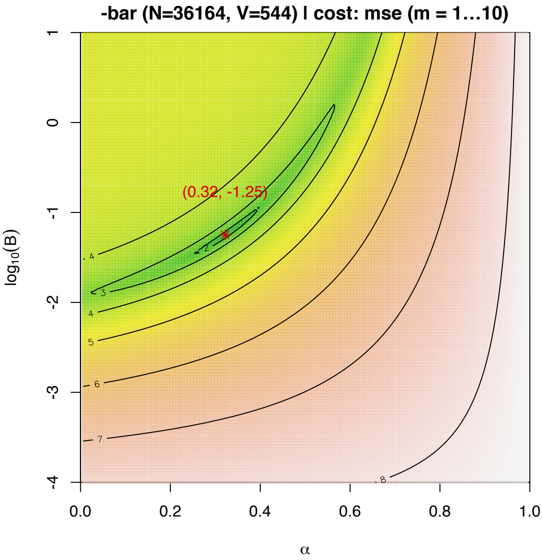
2. **Parameter estimation unsuccessful**
   (i.e. suboptimal goodness-of-fit to training data)
   - ☞ optimization algorithm trapped in local minimum
   - ☞ can result in highly inaccurate model

3. Uncertainty due to sampling variation
   (i.e. training data differ from population distribution)
   - ☞ model fitted to training data, may not reflect true population
   - ☞ another training sample would have led to different parameters
   - ☞ especially critical for small samples ($N < 10{,}000$)

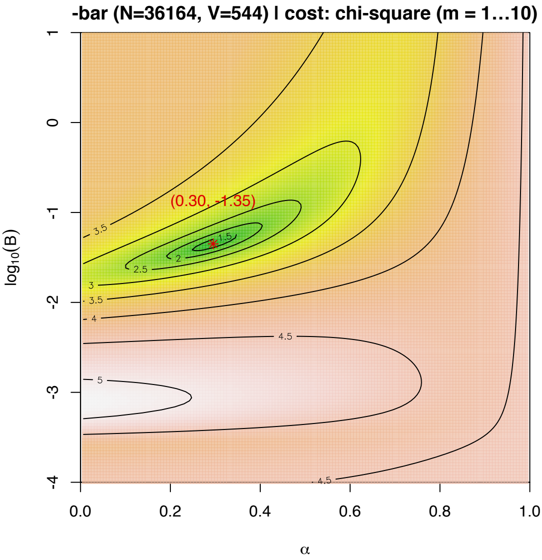
Cost functions for German word-formation affixes



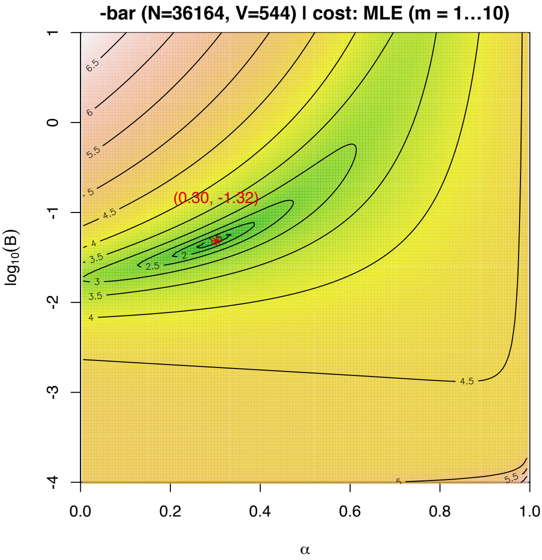
Cost functions for German word-formation affixes



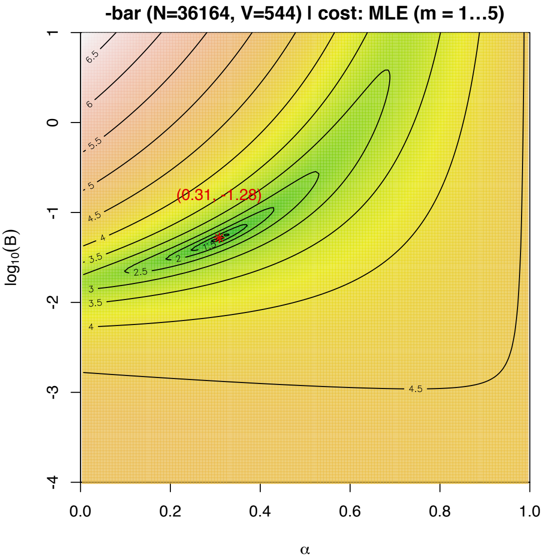
Cost functions for German word-formation affixes



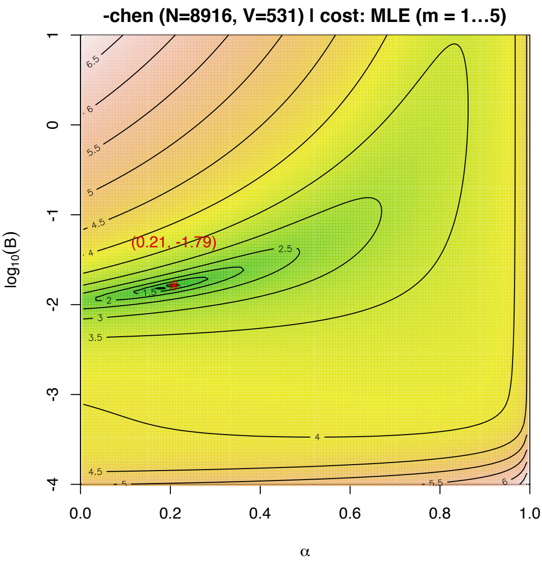
Cost functions for German word-formation affixes



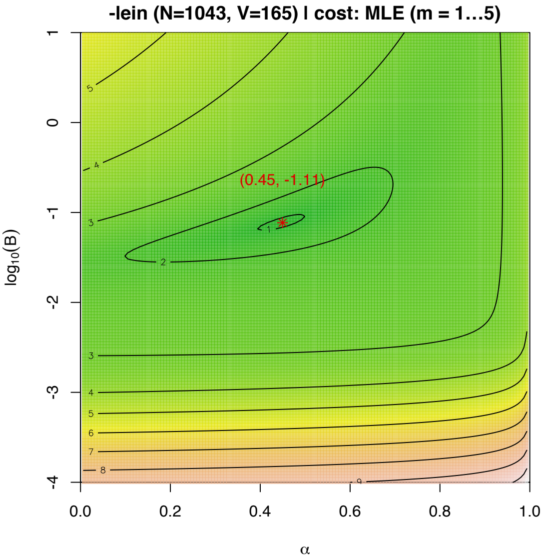
Cost functions for German word-formation affixes



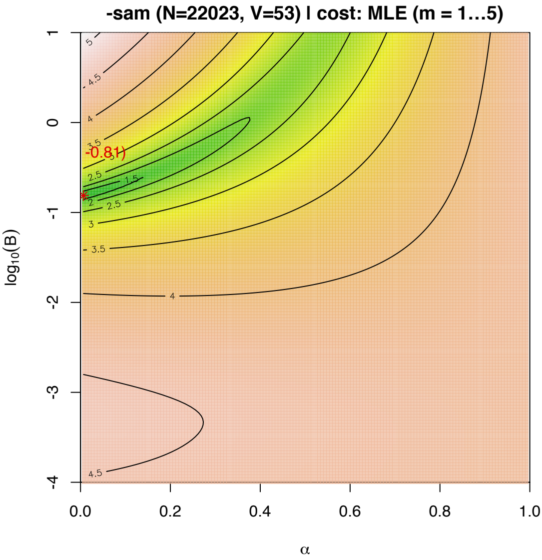
Cost functions for German word-formation affixes



Cost functions for German word-formation affixes



Cost functions for German word-formation affixes



Goodness-of-fit

(Baayen 2001, Sec. 3.3)

- Statistics: confidence intervals for population coefficients by inverting hypothesis tests (all $H_0 : \mu = x$ with $p > .05$)

Goodness-of-fit

(Baayen 2001, Sec. 3.3)

- Statistics: confidence intervals for population coefficients by inverting hypothesis tests (all $H_0 : \mu = x$ with $p > .05$)
- Multivariate normal approximation for $\mathbf{V} = (V, V_1, \ldots, V_k)$:

$$\Pr(\mathbf{V} = \mathbf{v}) \sim \frac{e^{-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{v}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^{k+1} \det \boldsymbol{\Sigma}}}$$

with $\boldsymbol{\mu} = (\mathrm{E}[V], \mathrm{E}[V_1], \mathrm{E}[V_2], \ldots)$ and $\boldsymbol{\Sigma}$ = covariance matrix

Goodness-of-fit

(Baayen 2001, Sec. 3.3)

- Statistics: confidence intervals for population coefficients by inverting hypothesis tests (all $H_0 : \mu = x$ with $p > .05$)
- Multivariate normal approximation for $\mathbf{V} = (V, V_1, \ldots, V_k)$:

$$\Pr(\mathbf{V} = \mathbf{v}) \sim \frac{e^{-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{v}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^{k+1} \det \boldsymbol{\Sigma}}}$$

with $\boldsymbol{\mu} = (\mathrm{E}[V], \mathrm{E}[V_1], \mathrm{E}[V_2], \ldots)$ and $\boldsymbol{\Sigma}$ = covariance matrix

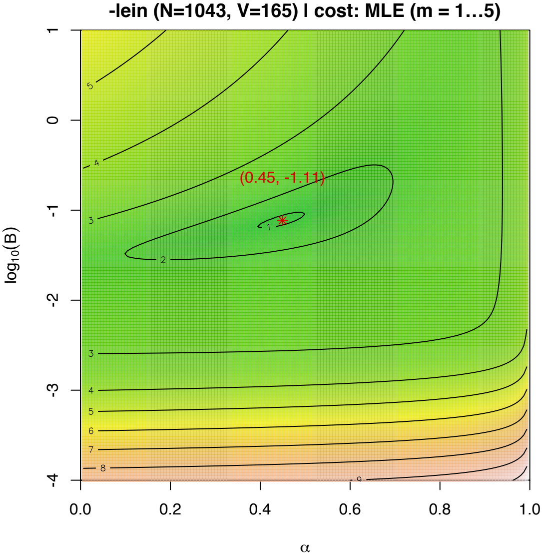
- Test statistic

$$X^2 = (\mathbf{V} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{V} - \boldsymbol{\mu}) \sim \chi^2_{k+1}$$

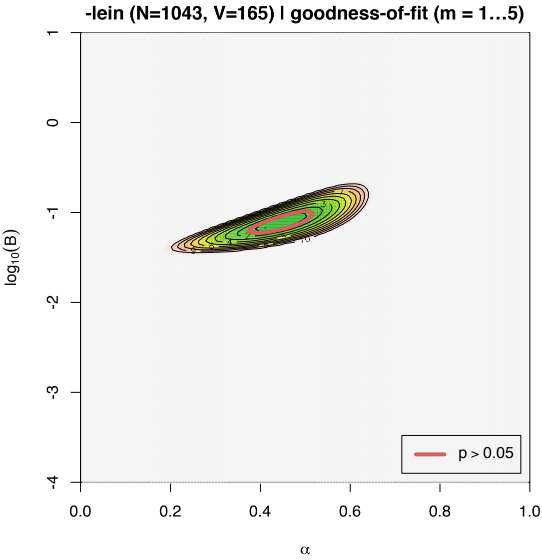
➥ Multivariate chi-squared test of **goodness-of-fit**

- ☞ significant rejection of the LNRE model for $p < .05$

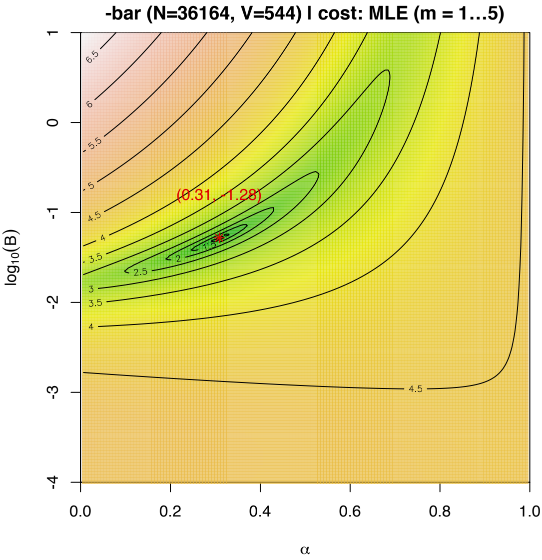
Confidence sets based on goodness-of-fit test?



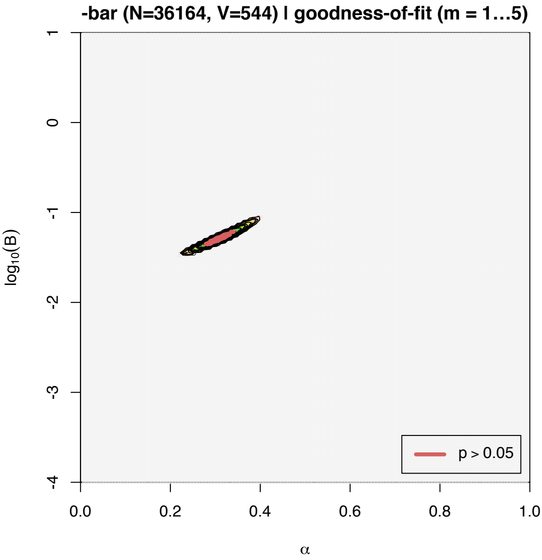
Confidence sets based on goodness-of-fit test?



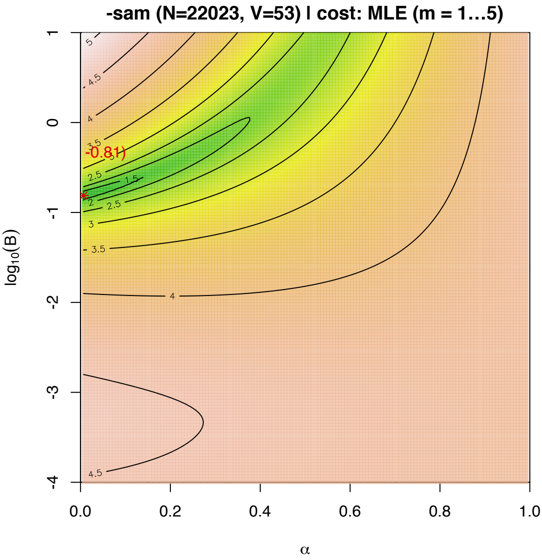
Confidence sets based on goodness-of-fit test?



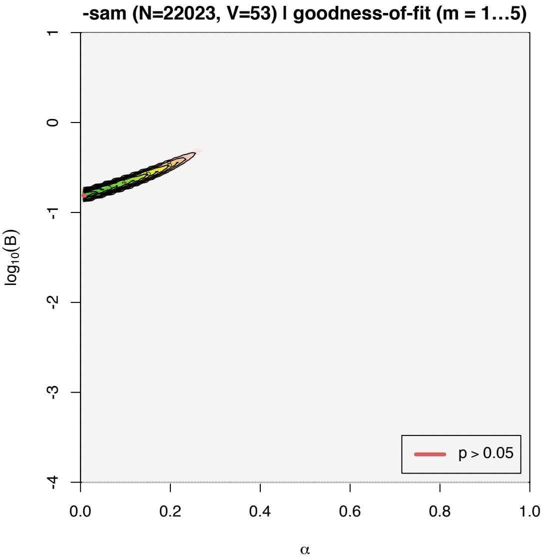
Confidence sets based on goodness-of-fit test?



Confidence sets based on goodness-of-fit test?

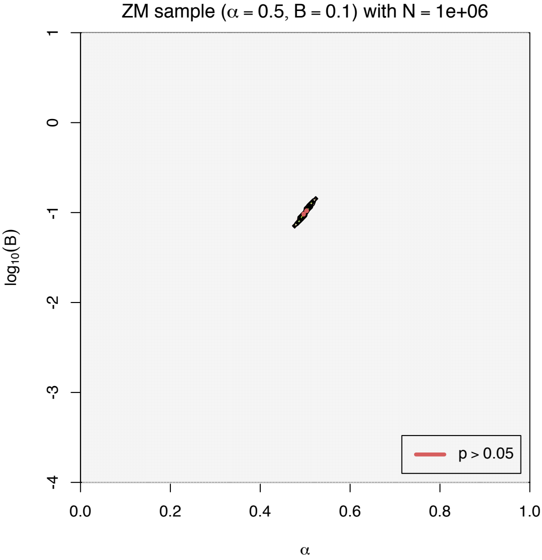


Confidence sets based on goodness-of-fit test?



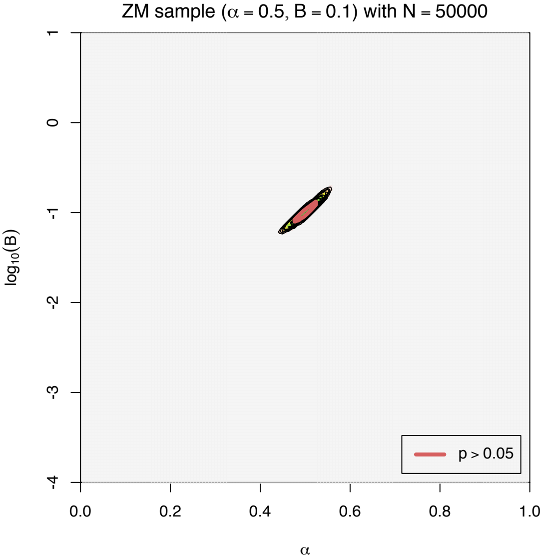
Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



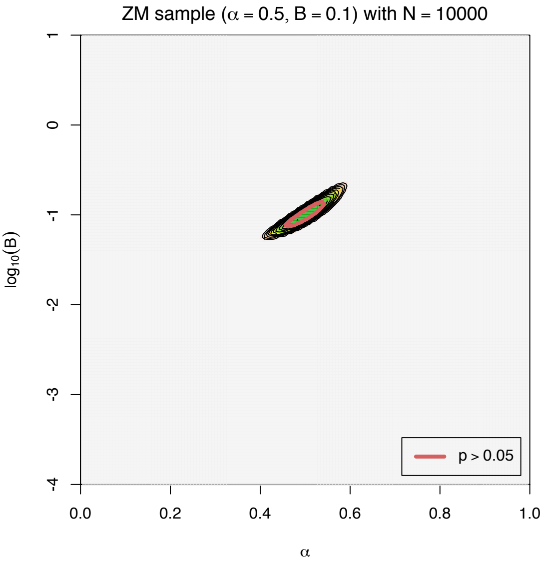
Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



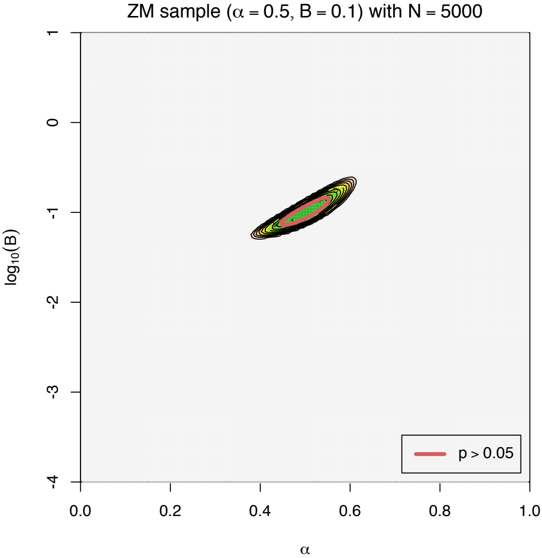
Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



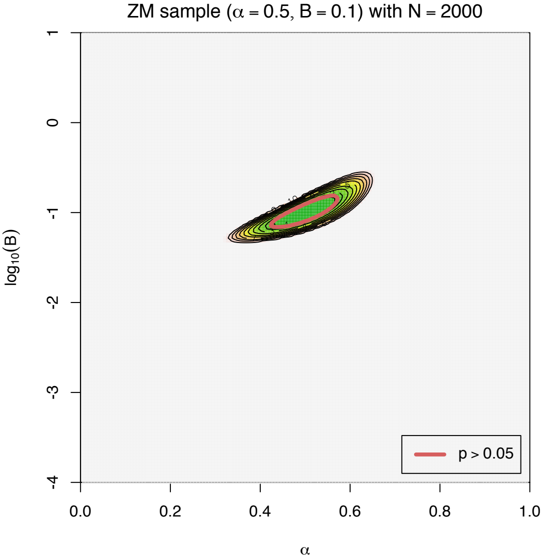
Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



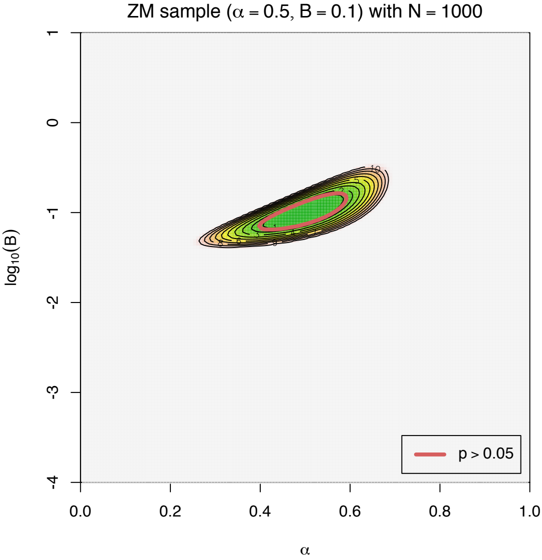
Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



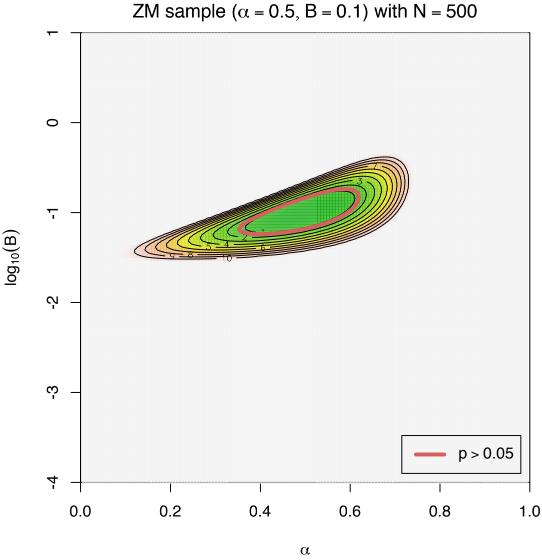
Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



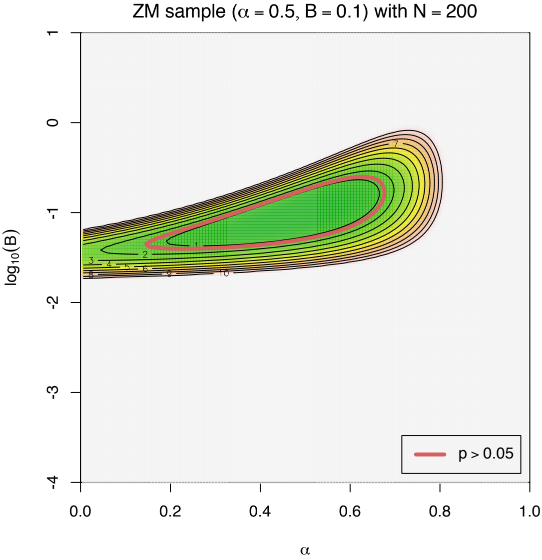
Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



Confidence sets for idealized samples from ZM population

➜ X^2 tests model parameters rather than goodness-of-fit



How reliable are the fitted models?

Three potential issues:

1. Model assumptions $\neq$ population
   (e.g. distribution does not follow a Zipf-Mandelbrot law)
   - ☞ model cannot be adequate, regardless of parameter settings
2. Parameter estimation unsuccessful
   (i.e. suboptimal goodness-of-fit to training data)
   - ☞ optimization algorithm trapped in local minimum
   - ☞ can result in highly inaccurate model
3. **Uncertainty due to sampling variation**
   (i.e. training data differ from population distribution)
   - ☞ model fitted to training data, may not reflect true population
   - ☞ another training sample would have led to different parameters
   - ☞ especially critical for small samples ($N < 10{,}000$)

Bootstrapping

- An empirical approach to sampling variation:
  - take many random samples from the same population
  - estimate LNRE model from each sample
  - analyse distribution of model parameters, goodness-of-fit, etc. (mean, median, s.d., boxplot, histogram, ...)
  - problem: how to obtain the additional samples?

Bootstrapping

- An empirical approach to sampling variation:
  - take many random samples from the same population
  - estimate LNRE model from each sample
  - analyse distribution of model parameters, goodness-of-fit, etc. (mean, median, s.d., boxplot, histogram, ...)
  - problem: how to obtain the additional samples?
- Bootstrapping (Efron 1979)
  - resample from observed data *with replacement*
  - this approach is not suitable for type-token distributions (resamples underestimate vocabulary size V!)

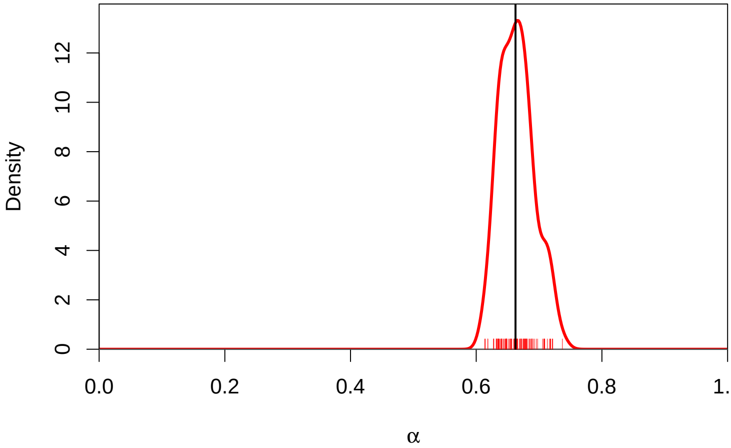
Bootstrapping

- An empirical approach to sampling variation:
  - take many random samples from the same population
  - estimate LNRE model from each sample
  - analyse distribution of model parameters, goodness-of-fit, etc. (mean, median, s.d., boxplot, histogram, . . .)
  - problem: how to obtain the additional samples?
- Bootstrapping (Efron 1979)
  - resample from observed data *with replacement*
  - this approach is not suitable for type-token distributions (resamples underestimate vocabulary size V!)
- Parametric bootstrapping
  - use fitted model to generate samples, i.e. sample from the population described by the model
  - advantage: "correct" parameter values are known

Bootstrapping

parametric bootstrapping with 100 replicates, fZM samples for $N = 3467$

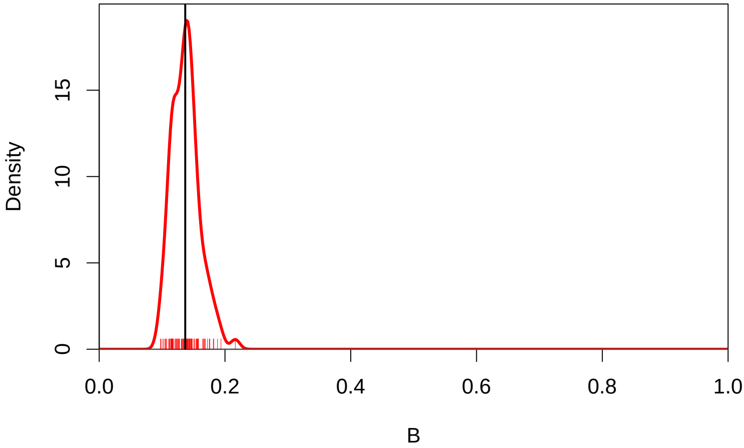
Zipfian slope $a = 1/\alpha$



Bootstrapping

parametric bootstrapping with 100 replicates, fZM samples for $N = 3467$

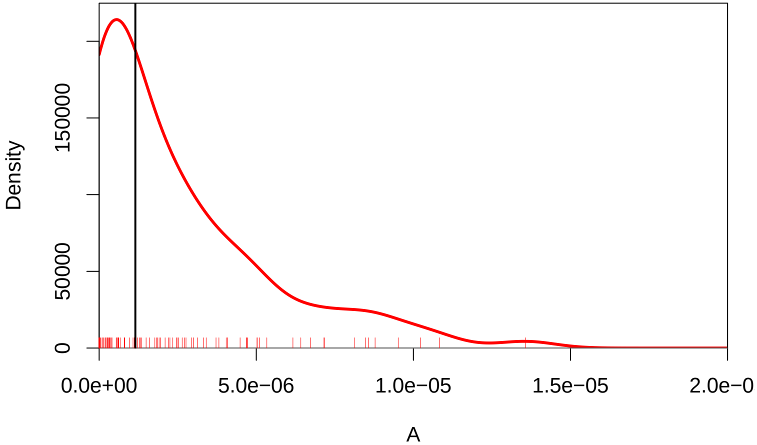
Offset $b = (1 - \alpha)/(B \cdot \alpha)$



Bootstrapping

parametric bootstrapping with 100 replicates, fZM samples for $N = 3467$

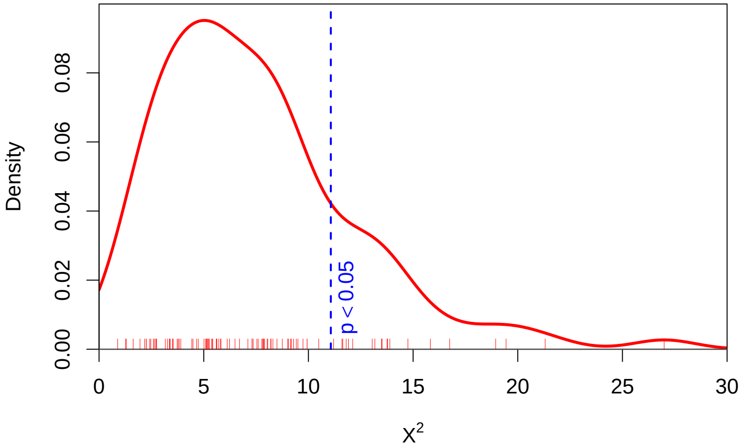
fZM probability cutoff $A = \pi_S$



Bootstrapping

parametric bootstrapping with 100 replicates, fZM samples for $N = 3467$

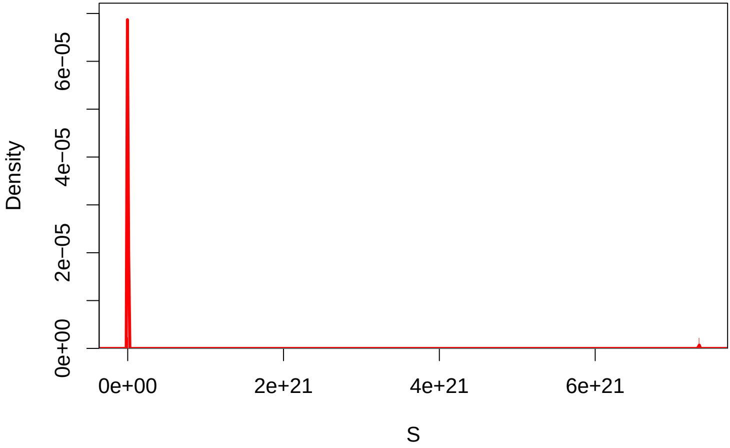
Goodness-of-fit statistic X^2 (model not plausible for $X^2 > 11$)



Bootstrapping

parametric bootstrapping with 100 replicates, fZM samples for $N = 3467$

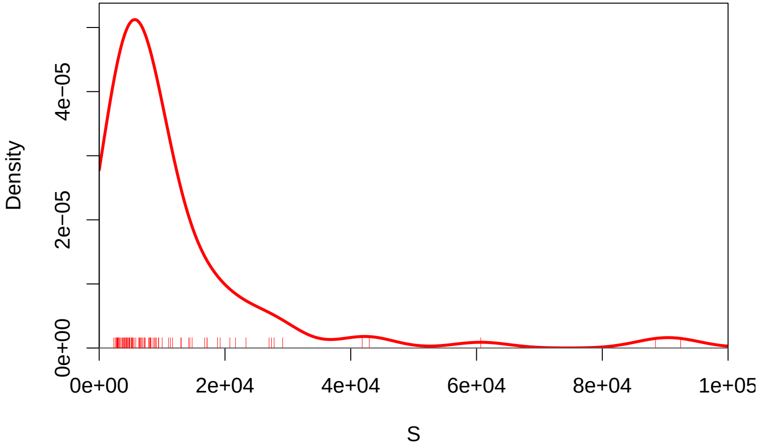
Population diversity S



Bootstrapping

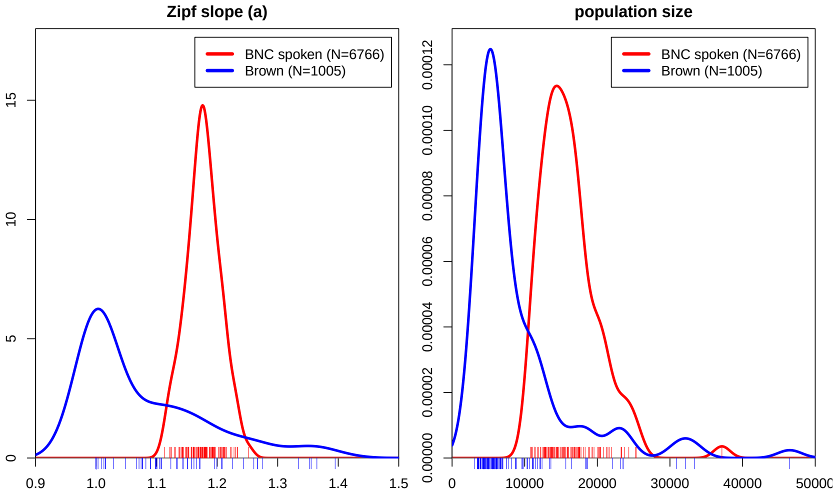
parametric bootstrapping with 100 replicates, fZM samples for $N = 3467$

Population diversity S



Sample size matters!

Brown corpus is too small for reliable LNRE parameter estimation (bare singulars)



Thank you!

References I

Baayen, Harald (1991). A stochastic process for word frequency distributions. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 271–278.

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

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 (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Evert, Stefan (2004). 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.

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.

References II

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.

Herdan, Gustav (1964). *Quantitative Linguistics*. Butterworths, London.

Lüdeling, Anke and Evert, Stefan (2005). The emergence of productive non-medical *-itis*. corpus evidence and qualitative analysis. In S. Kepser and M. Reis (eds.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*, pages 351–370. Mouton de Gruyter, Berlin.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542–547.

Tweedie, Fiona J. and Baayen, R. Harald (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, **32**, 323–352.

Yule, G. Udny (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge.