

TEXT, SPEECH
AND LANGUAGE
TECHNOLOGY

WORD FREQUENCY DISTRIBUTIONS

R. Harald Baayen

CD-ROM included

Kluwer Academic Publishers

Word Frequency Distributions

Text, Speech and Language Technology

VOLUME 18

Series Editors

Nancy Ide, *Vassar College, New York*
Jean Véronis, *Université de Provence and CNRS, France*

Editorial Board

Harald Baayen, *Max Planck Institute for Psycholinguistics, The Netherlands*
Kenneth W. Church, *AT & T Bell Labs, New Jersey, USA*
Judith Klavans, *Columbia University, New York, USA*
David T. Barnard, *University of Regina, Canada*
Dan Tufis, *Romanian Academy of Sciences, Romania*
Joaquim Llisterri, *Universitat Autonoma de Barcelona, Spain*
Stig Johansson, *University of Oslo, Norway*
Joseph Mariani, *LIMSI-CNRS, France*

The titles published in this series are listed at the end of this volume.

Word Frequency Distributions

By

R. Harald Baayen

*University of Nijmegen
The Netherlands*



KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 0-7923-7017-1

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

All Rights Reserved

© 2001 Kluwer Academic Publishers

No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the copyright owner.

Printed in the Netherlands

to the memory of Rezo Chitashvili

Contents

List of Figures	ix
List of Tables	xix
Introduction	xxi
1 Word Frequencies	1
1.1 Introduction	2
1.2 The frequency spectrum	8
1.3 Zipf	13
1.4 The quest for characteristic constants	24
1.5 The lognormal distribution	32
1.6 Discussion	34
1.7 Bibliographical Comments	35
1.8 Questions	35
2 Non-parametric models	39
2.1 Basic concepts	39
2.2 The Urn model	42
2.3 The Structural Type Distribution	47
2.4 The LNRE zone	51
2.5 Good-Turing estimates	57
2.6 Interpolation and Extrapolation	63
2.6.1 Interpolation	64
2.6.2 Extrapolation	69
2.7 Discussion	76
2.8 Bibliographical Comments	76
2.9 Questions	77
3 Parametric models	79
3.1 Introduction	79
3.2 LNRE models	82
3.2.1 The Lognormal Structural Type Distribution	82
3.2.2 The Generalized Inverse Gauss-Poisson Structural Type Distribution	89
3.2.3 The Zipfian Family of LNRE Models	93
3.3 Evaluating Goodness of Fit	118
3.4 Parameter estimation	122
3.5 A comparative study	124
3.6 Comparing Lexical Measures Across Texts	132
3.7 Discussion	132
3.8 Bibliographical Comments	133

3.9	Questions	133
4	Mixture distributions	135
4.1	Introduction	135
4.2	Expectations, variances, and covariances	139
4.3	Examples of mixture distributions	142
4.3.1	A text-level mixture model	142
4.3.2	Morphological mixtures	145
4.4	Morphological Productivity	154
4.5	Discussion	158
4.6	Bibliographical Comments	160
4.7	Questions	160
5	The Randomness Assumption	161
5.1	The Randomness Assumption	161
5.1.1	Non-randomness and lexical specialization	162
5.1.2	Consequences of non-randomness	167
5.2	Adjusted LNRE models	173
5.2.1	Partition-based adjustment	174
5.2.2	Parameter-based adjustment	179
5.3	Discussion	192
5.4	Bibliographical Comments	193
6	Examples of Applications	195
6.1	Distributional properties of the lexicon	195
6.1.1	Word length and sample size	195
6.1.2	Matching reliability across corpora	199
6.2	Morphological productivity	203
6.2.1	Global analyses	203
6.2.2	Productivity and register	208
6.3	Authorship and Style	211
6.4	Beyond word frequency distributions	214
6.4.1	Counts of filarial worms on mites on rats	214
6.4.2	Year references	215
6.4.3	CV-structures	218
6.4.4	Word pairs	221
6.4.5	Discussion	221
6.5	Some practical guidelines	223
A	List of Symbols	237
B	Solutions to the exercises	241
C	Software	251
D	Data sets	289
Bibliography		321
Index		329

List of Figures

1.1	Vocabulary size $V(N)$ (panel A) and mean word frequency $N/V(N)$ (panel B) as a function of sample size N in Alice in Wonderland, measured at 20 equally-spaced intervals.	4
1.2	The sample relative frequency of the article a $p(a, N)$ (panel A) and the sample relative frequency of the article the $p(\text{the}, N)$ (panel B) as a function of the sample size N in Alice in Wonderland, measured at 20 equally-spaced intervals. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on a total of 5000 permutation runs.	6
1.3	The frequency spectrum of Alice in Wonderland (m : frequency class; $V(m, N)$: number of types with frequency m).	10
1.4	The empirical structural type distribution of Alice in Wonderland (m : frequency class; $g(m, N)$: number of types occurring m or more times).	11
1.5	Word frequency $f_z(z, N)$ as a function of Zipf rank z in the double logarithmic plane for $N = 13250$ (panel A) and for $N = 26505$ (panel B).	14
1.6	The rank-frequency distribution as a step function in the double logarithmic plane and the relation with the elements of the frequency spectrum: $V(m, N) = z_2 - z_1$.	17
1.7	The dependency of the two parameters of Zipf's zeta distribution, the intercept (panel A) and the slope (panel B) on the sample size N , plotted at 20 equally-spaced intervals for Alice in Wonderland (compare 1.5).	18
1.8	The frequency spectrum of Alice's Adventures in Wonderland in the double logarithmic plane. Left panel: the integer-valued spectrum with linear fits for $m = 1 \dots 15$ (dashed line) and for the complete range of m (dotted line). Right panel: the same spectrum elements after transformation into fractional values for higher-frequency ranks, with linear fits for $m = 1 \dots 15$ (dashed line) and for the complete range of m (dotted line).	19
1.9	The real-valued approximate frequency spectrum of Alice's Adventures in Wonderland in the double logarithmic plane with a simple Zipfian fit (left panel) and a Naranan-Balasubrahmanyam Zipfian fit (right panel).	22

1.10	<i>The dependency of the three parameters of the Naranan-Balasubrahmanyam Zipfian model as a function of the sample size N in Max Havelaar by Multatuli. The upper right panel plots C as a function of N, the lower left panel μ as a function of N, and the lower right panel γ as a function of N. The upper left panel plots $V(N)$ (upper circles), $V(1, N)$ (central circles), and $V(2, N)$ (lower circles) as a function of N at 20 equally-spaced intervals.</i>	23
1.11	<i>The spectrum ratios $V2/V1$ ($R(2, 1)$, circles) and $V3/V2$ ($R(3, 2)$, triangles) for Alice's Adventures in Wonderland (left panel) and Max Havelaar (right panel) for 20 equally spaced sample sizes N. The solid lines represent the corresponding expected values.</i>	24
1.12	<i>The characteristic constants K (panel A) and D (panel B) as a function of the sample size N in Alice in Wonderland. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.</i>	26
1.13	<i>The text characteristics R (panel A) and W (panel B) as a function of the sample size N in Alice in Wonderland. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.</i>	27
1.14	<i>V, N/V, R, and W as a function of N in a random sample of 10000 tokens from a population with 50 equiprobable types.</i>	28
1.15	<i>The text characteristics H (panel C) and S (panel D) as a function of the sample size N in Alice in Wonderland. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs. Panels A and B show the roughly linear dependency of the relative number of hapax legomena on $\log N$ that underlies H.</i>	29
1.16	<i>Herdan's 'law' applied to Alice in Wonderland. Panel A plots $V(N)$ as a function of N in the double logarithmic plane. Panel B plots Herdan's constant C as a function of the sample size N in Alice in Wonderland. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.</i>	30
1.17	<i>Yule's K as a function of N for Alice in Wonderland (dotted line) and Through the looking-glass (solid line).</i>	31
1.18	<i>The lognormal hypothesis. Panel A shows the estimated probability density function for log frequency in Alice in Wonderland (dashed and dotted line) and the estimated density function of a lognormal random variable with the same mean and standard deviation. Panel B plots the corresponding quantiles of the standard normal distribution.</i>	32
1.19	<i>The parameters of the lognormal model as a function of the sample size N for Alice in Wonderland.</i>	33

1.20	<i>The sample relative frequency of the article a $p(a, N)$ (panel A) and the sample relative frequency of the article the $p(\text{the}, N)$ (panel B) as a function of sample size N in Through the looking-glass, measured at 20 equally-spaced intervals. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on a total of 5000 permutation runs.</i>	36
1.21	<i>The frequency spectra of Through the looking-glass at the sample sizes $N = 14514$ and $N = 29028$ (m: frequency class; $V(m, N)$: number of types occurring m times).</i>	37
1.22	<i>The error function of Zipf's zeta function for Alice in Wonderland at $N = 13250$.</i>	38
2.1	<i>The structural type distribution. $\Delta G(\pi_j) = G(\pi_j) - G(\pi_{j+1})$ is the number of words in the population with probability π_j.</i>	48
2.2	<i>The growth rate of the vocabulary as the first derivative of the growth curve $E[V(N)]$: $\Delta y / \Delta x = E[V(1, N)]/N$.</i>	50
2.3	<i>$E[V(N)]$ and $E[V(m, N)]$ for $m = 1, 2, \dots, 5$ for 100 throws with a fair die.</i>	52
2.4	<i>The empirical vocabulary size $V(N)$ (large dots) and the first 5 empirical spectral elements $V(m, N)$ (small dots) for Alice in Wonderland, sampled at 20 equal-spaced intervals.</i>	53
2.5	<i>The expected vocabulary size $E[V(N)]$ (large dots) and the first 5 expected spectral elements $E[V(m, N)]$ (small dots) for the written texts in the British National Corpus, sampled at 20 equal-spaced intervals.</i>	54
2.6	<i>Estimating m^* for a corpus of the British newspaper The Independent. Left panel: The real-valued approximate spectrum elements $V_r(m, N)$ in the double logarithmic plane with the Naranan-Balasubrahmanyam Zipfian fit (solid line) for a subcorpus of 1 million words; Central panel: m^* as a function of m for the observations from eight separate newspaper corpora of 1 million words from The Independent, with the Good-Turing 95% confidence interval and the smoothed values derived from the Naranan-Balasubrahmanyam Zipfian fit shown in the left panel; Right panel: m^{**} as a function of m for the observations from eight separate newspaper corpora of 1 million words from The Independent, the Good-Turing 95% confidence interval, and the smoothed values derived from the Naranan-Balasubrahmanyam Zipfian fit shown in the left panel.</i>	62
2.7	<i>$E[V(N)]$ as a function of N using binomial and hypergeometric interpolation. The left panel shows the hypergeometric values and their 95% confidence intervals, using Monte Carlo simulations, for Alice in Wonderland. The dots plot the binomial expectations. The right panel illustrates the slight overestimation bias of the binomial model.</i>	68
2.8	<i>The empirical vocabulary growth curve $V(N)$ (solid line), its expectation using binomial interpolation (circles), and its expectation using the Poisson approximation (triangles), for Alice in Wonderland, at 20 equally spaced intervals.</i>	74

2.9	<i>Smoothing the first 15 spectrum elements of Alice in Wonderland using Zipf's law at the sample size $N = 10,602$.</i>	76
3.1	<i>Zipf's law as optimal for $N = 10602$. The plot shows the relative Mean Squared Error (MSE), calculated on the basis of the first 15 spectrum elements for Alice in Wonderland.</i>	80
3.2	<i>Quantile-quantile plot of log frequency for Alice in Wonderland.</i>	83
3.3	<i>Quantile-quantile plots for the empirical token (bottom line) and type (top line) structural distributions for Alice in Wonderland.</i>	87
3.4	<i>The empirical vocabulary size $V(N)$ (solid line) as a function of the text size N, its expectation using binomial interpolation (open circles), and its expectation using the lognormal structural type distribution (triangles), for Alice in Wonderland.</i>	88
3.5	<i>Observed (circles) and expected (solid line) relative spectrum elements for Alice in Wonderland, using Sichel's model with $\gamma = -0.5$, $b = 0.0236$, and $Z = 105.78$.</i>	93
3.6	<i>The empirical and expected growth curves of the vocabulary for Alice in Wonderland. The circles represent the observed vocabulary size $V(N)$, the triangles the expected vocabulary size $E[V(N)]$ using (3.42) for $Z = 10602$.</i>	99
3.7	<i>The relative frequency spectrum elements $\alpha(2, N)$ as a function of the sample size N for Alice in Wonderland. The dotted line represent the observed values, the solid line the expected values based on 100 permutation runs, and the dashed line the expected values for Zipf's law with $Z = 12225$.</i>	101
3.8	<i>Rank-frequency curves for Alice's adventures in Wonderland (upper left panel), for a sample of 1 million monomorphemic Dutch words (upper right panel), for a word frequency distribution generated by a first order Markov model using the phoneme transition frequencies of monomorphemic Dutch words (lower left panel), and for a word frequency distribution generated by a re-use and birth stochastic process (lower right panel).</i>	102
3.9	<i>The relation between log frequency and word length (top row) and log frequency and lexical density (bottom row) for a sample of 1 million monomorphemic Dutch words (left column), a Markov approximation (middle column), and a Re-use and birth model (right column).</i>	106
3.10	<i>Selection probabilities by frequency class (left panel) for $\Pr(m)$ (solid line) and $\Pr_{(e)}(m)$ (dashed line), and the corresponding contributions to the average amount of information (right panel, dashed and dotted lines, respectively).</i>	111
3.11	<i>Observed and expected frequency spectrum for $m = 1, 2, \dots, 15$ of the 'A'-texts of the British National Corpus. The dots represent the observed frequency spectrum, the dotted line the corresponding expectations using the extended Zipf's law ($Z = 191229.6$), and the solid line the corresponding expectations using the Yule-Simon model ($Z = 67494.14$, $\beta = 0.72978$).</i>	114

- 3.12 The first 15 spectrum elements on a double logarithmic scale (left) and the growth curve of the vocabulary, again using a double logarithmic scale, for Alice's adventures in wonderland. Open dots denote the observed spectrum elements, the solid line in the spectrum plot represents the Yule-Simon fit, and the dotted line the generalized inverse Gauss-Poisson fit. In the right panel, the triangles represent binomial interpolation, the solid line again represents the Yule-Simon fit, and the dotted line represents the generalized inverse Gauss-Poisson fit. 127
- 3.13 The first 15 spectrum elements on a double logarithmic scale (left) and the growth curve of the vocabulary, again using a double logarithmic scale, for the subcorpus of context-governed spoken English in the BNC. Open dots denote the observed spectrum elements, the solid line in the spectrum plot represents the Yule-Simon fit, and the dotted line the generalized inverse Gauss-Poisson fit. In the right panel, the triangles represent binomial interpolation, the solid line again represents the Yule-Simon fit, and the dotted line represents the generalized inverse Gauss-Poisson fit. 130
- 3.14 The first 15 spectrum elements on a double logarithmic scale for Tamm-saare's Truth and Justice (data from Tuldava, 1996). Observed spectrum elements are represented by circles, the generalized inverse Gauss-Poisson fit is represented by a dashed line, and the lognormal fit by a dotted line. 131
- 4.1 The first fifteen spectrum elements for simplex nouns with a Yule-Simon fit ($\hat{Z} = 7650$, $\hat{\beta} = 1.4178$, $\hat{V}_Z = 1654$, upper left), the complete spectrum (upper right), the estimated density (lower left), and the developmental profiles of $E[V(N)]$ and $E[V(m, N)]$, $m = 1, 2$ using binomial interpolation (lower right). The dashed line represents the asymptote $\lim_{N \rightarrow \infty} E[V(N)] = \hat{S} = 5042$ 136
- 4.2 The first fifteen spectrum elements for complex nouns with the suffix -heid with Yule-Simon fit ($\hat{Z} = 299.4$, $\hat{\beta} = 1.61$, $\hat{V}_Z = 150$, upper left), the complete spectrum (upper right), the estimated density (lower left), and the developmental profiles of the vocabulary size and the first two spectrum elements using binomial interpolation (lower right). 138
- 4.3 Illustrations of the equality $E[V(m, pN)|\{Z, \dots\}] = E[V(m, N)|\{\frac{Z}{p}, \dots\}]$ for a GIGP distribution with parameters $Z = 30$, $b = 0.001$, $\gamma = -0.4$: $E[V(N/3)|\{Z, \dots\}] = \frac{1}{3}E[V(N)|\{3Z, \dots\}]$ (upper left panel), $E[V(\bar{N}/3, m)|\{Z, \dots\}] = \frac{1}{3}E[V(\bar{N}, m)|\{3Z, \dots\}]$ for $m = 1$ (upper right panel), $m = 2$ (lower left panel), and $m = 3$ (lower right panel). The solid lines represent the expectations for the distribution with $Z = 10.0$, the dashed lines represent the expectations for the distribution with $Z' = 3Z = 30.0$ 141

- 4.4 Left panel: The first fifteen observed spectrum elements (circles), their expectations for a lognormal-GIGP model (solid line), and the corresponding expectations given a simple GIGP model (dashed line) for a Turkish text on archeology. Right panel: The mixture model for $m = 4, \dots, 15$ (solid line), its lognormal component (circles) and its GIGP component (asterisks). 143
- 4.5 Vocabulary growth curves for the Turkish text on archeology analyzed in Figure 4.4 (dashed line) as well as for a spoken English text from the British National Corpus (F71, short dashes) and a written English text (the first part of the Hound of the Baskervilles, long dashes). 144
- 4.6 The spectrum elements for complex nouns with the suffix -heid (circles), the Yule-Simon fit (dashed line), and a Lognormal-Yule-Simon mixture fit (solid line): $m = 1, \dots, 15$ (upper left), $m = 5, \dots, 100$ (upper right). The lower left panel shows the mixture components for $m = 2, \dots, 15$, using asterisks for the lognormal component and triangles for the Yule-Simon component and a solid line for the mixture model itself. The lower right panel shows the congruence of binomial interpolation (circles) and interpolation for the mixture model (solid line) and the simple Yule-Simon fit (dashed line). 148
- 4.7 The spectrum elements for complex nouns with the suffix -iteit (circles), the GIGP fit (dashed line), and a lognormal-GIGP mixture fit (solid line), for $m = 1, \dots, 15$ (upper left). The upper right panel shows the estimated density function for the relative frequencies using a logarithmic scale. The lower left panel shows the mixture components for $m = 2, \dots, 15$, using asterisks for the lognormal component and triangles for the GIGP component and a solid line for the mixture model itself. The lower right panel shows the congruence of binomial interpolation (circles) and interpolation for the mixture model (solid line) and the simple Yule-Simon fit (dashed line). 150
- 4.8 The first 15 spectrum elements for complex nouns with the suffix -ness in the written subcorpus of the British National Corpus (upper left) and in the context-governed spoken subcorpus (upper right). Observed values are represented by circles, the LNRE-based expectations (GIGP for the written and Yule-Simon for the spoken subcorpus) by solid lines. The lower left panel shows the LNRE-interpolated growth curves of the vocabulary for the written (upper curve) and spoken (lower curve) subcorpus. The dotted lines mark the 95% confidence interval for the written subcorpus. The lower right panel plots the mixture of the two distributions (circles), the GIGP fit to the mixture (solid line), the GIGP-Yule-Simon mixture fit (+), and the observed spectrum when a pure string-based type definition is used. 152
- 4.9 The contribution of singular nouns (left panel) and plural nouns (right panel) to the growth rate of nouns in Innes Appleby, \mathcal{P}^* , as a function of the sample size N 159

5.1	<i>The overestimation bias $E[V(N)] - V(N)$ in Alice in Wonderland (solid line) and in a sentence-randomized version of Alice in Wonderland (dotted line). For the first half of the text, significant estimation errors are highlighted.</i>	162
5.2	<i>Empirical (dotted line) and expected (solid line) growth curves for the concatenated texts of L. F. Baum's The Wonderful Wizard of Oz, election speeches by and interviews with B. Clinton, J. M. Barrie's Peter Pan, and L. Carroll's Alice in Wonderland, for 80 measurement points. The dotted vertical lines indicate the transition points between the texts.</i>	164
5.3	<i>Empirical (dashed line) and expected (solid line) growth curves of the non-underdispersed words in Alice in Wonderland.</i>	166
5.4	<i>Number of underdispersed word types (left panel, $VU(k)$) and tokens (right panel, $NU(k)$) for the partition of $k = 1, \dots, 40$ equally-sized chronologically ordered text chunks of Alice in Wonderland. The dotted lines represent least squares regression lines, the solid lines a non-parametric time series smoother.</i>	168
5.5	<i>Interpolation and extrapolation accuracy for Alice in Wonderland, conditioning on the frequency spectrum at $N = 13253$. The upper left panel shows the fit of Sichel's model to the frequency spectrum ($b = 0.0194$, $Z = 92.01$). The upper left panel plots the observed (dotted line) and expected vocabulary size given the Sichel fit. The bottom panel plots the estimation errors $E[V(N)] - V(N)$.</i>	169
5.6	<i>Left panel: part of the real-valued approximate frequency spectrum of the 1 million sample of The Independent of 1989 in the double logarithmic plane. The upper solid line represents the least-squares linear regression line to the frequency range to the left of the first dashed vertical line, the lower solid line is a least squares regression line to the right of the second vertical dashed line. Right panel: the corresponding plot for m^*. The upper line is fit to the spectrum elements to the right of the second dashed line, the lower line is a fit to the spectrum elements to the left of the first dashed line.</i>	171
5.7	<i>Spectrum and developmental profile of $V(N)$ and the first five spectrum elements for Alice in Wonderland. The dots represent the empirical values, the solid lines the corresponding expected values for the lognormal model.</i>	174
5.8	<i>Partition-based interpolation for Carroll's Alice in Wonderland (left) and Wells' The war of the worlds (right). The upper panels plot the observed vocabulary size and the spectrum elements $m = 1, 2, \dots, 5$ as a function of N, using solid lines for partition-based interpolation, dotted lines for standard binomial interpolation, and dots for the observed values. The lower panels show the dependency of the partition parameter p on N, using dots for the observed values and a dashed line for a non-parametric smoother using running medians.</i>	177

- 5.9 Partition-based adjusted Zipfian interpolation and extrapolation for Carroll's Alice in Wonderland (left, Yule-Simon fit) and Wells' The war of the worlds (right, Zipf fit). The observed vocabulary size and the spectrum elements $m = 1, 2, \dots, 5$ are represented by dots, their expectations based on the extended Zipf's law are shown by means of dotted lines, and their partition-based adjustments by means of solid lines. 178
- 5.10 A generalized inverse Gauss-Poisson fit to the combined samples of The Independent (left panel) and the individual empirical spectra of the 8 individual samples of 1 million words (right panel). The solid line represents the expected spectrum as interpolated from the combined spectrum shown in the left panel. 180
- 5.11 The three parameters of the generalized inverse Gauss-Poisson model for the eight one million word samples of The Independent (labeled 891 to 964) together with the parameters for the combined sample of 8 million words (labeled 0). 181
- 5.12 The dependence of LNRE parameters on the text length N illustrated for Alice in Wonderland. The top panels plot the developmental profiles of the parameters b and c of the generalized inverse Gauss-Poisson model, the bottom panel shows the profile of the single parameter Z of the extended Zipf's law. Observed profiles are represented by dots, their Monte-Carle expectations are represented by solid lines. 182
- 5.13 Diagnostic plots for Alice in Wonderland using a power link function to enhance the extended Zipf's law. 184
- 5.14 Diagnostic plots for the parameter-adjusted extended Zipf's law applied to Alice in Wonderland, with hand-tuned power fit. 186
- 5.15 Diagnostic plots for the parameter-adjusted generalized inverse Gauss-Poisson model applied to the 'A'-texts of the British National Corpus. 187
- 5.16 Non-homogeneity in the developmental profile: the case of the Hound of the Baskervilles. The upper left panel plots the growth curve of the vocabulary, and the upper right panel the growth curve of the hapax legomena. The bottom panels plot the developmental profiles of Z and c 188
- 5.17 Extrapolation accuracy of the parameter-adjusted extended Zipf's law for Alice in Wonderland (left panel) and Well's The war of the worlds (right panel), fitted at $N = N_0/2$. The dots plot the observed values, the dotted lines unadjusted interpolation and extrapolation, and the solid lines adjusted interpolation and extrapolation. 189
- 5.18 Three estimates of the growth rate of the vocabulary: the one-token growth rate $E[V(N+1)] - E[V(N)]$ for the parameter-adjusted Zipf smoothing (solid line), the unsmoothed sample estimates $V(1, N)/N$ (dashed line), as well as the (unadjusted) Zipf-smoothed estimates $E[V(1, N)]/N$ (dotted line), for Alice in Wonderland. 191

- 6.1 *The distribution of word length and sample size ($N = 1 \dots 8$ million word tokens). Upper left: number of types as a function of word length in letters; upper right: mean word frequency as a function of word length; lower left: proportion of types as a function of word length, the curve labelled A represents the corresponding graph for Alice's Adventures in Wonderland ($N = 26505$); lower right: proportion of types as a function of corpus size for word lengths 5 and 15.* 197
- 6.2 *Proportion of words (length 4, left panel; length 10, right panel) with frequency less than 10 (solid line), less than 20 (dotted line), and less than 100 (dashed line) as a function of the sample size N (in million word tokens from The Independent).* 198
- 6.3 *Boxplots showing the matching for log Surface Frequency and the contrast in log Base Frequency for the words in -heid in Experiment 3 of Bertram, Schreuder, and Baayen (1999) for the CELEX lexical database (based on a 42 million word corpus, upper row) and the Dutch Trouw corpus (4.2 million words, bottom row).* 200
- 6.4 *The difference between m and m^* for $m = 1, \dots, 50$ for the 8 million word sample of The Independent using a Narayan-Balasubramanian-Zipfian smoother for the calculation of the Good-Turing estimates.* 202
- 6.5 *Productivity statistics $P(N)$, $V(1, N)$, and $V(N)$ for selected Dutch affixes.* 205
- 6.6 *Observed and expected vocabulary size as a function of sample size using binomial interpolation for all words in the 8 million word sample of The Independent (upper left panel), for the 9449 words in -ness in the same corpus (lower left panel), and for the first 9449 words of Alice's Adventures in Wonderland (lower right panel).* 207
- 6.7 *Vocabulary growth curves for the German suffixes -bar (solid line), -sam (dashed line), and ös (dotted line) using raw data based on string matches (left panel) and manually corrected data (right panel).* 208
- 6.8 *Vocabulary growth curves $E[V(N)]$ using binomial interpolation for the English suffix -ness in the written subcorpus (W), the context-governed spoken subcorpus (C), and the demographic subcorpus (D) of the British National Corpus. The left panel plots $E[V(N)]$ as a function of the number of tokens in -ness in each subcorpus, the right panel renormalizes the horizontal axis to display the number of arbitrary tokens in a subcorpus.* 209
- 6.9 *Guiraud's R as a function of N for selected English texts. For the key to the abbreviations, see Table 6.1.* 212
- 6.10 *Developmental profiles in the plane spanned by $Z(N)$ and $K(N)$ for selected English texts. For the key to the abbreviations, see Table 6.1.* 213
- 6.11 *Fit of the generalized inverse Gauss-Poisson model (GIGP) and the Yule-Simon model to the data on filarial worms on mites on rats (Heller, 1997).* 216

6.12 References to years in the newspaper issues of the Frankfurter Allgemeine Zeitung that appeared in 1994. The upper left panel plots the frequency spectrum, the upper right panel the smoothed frequency spectrum, the lower left panel shows the rank-frequency distribution, and the lower right panel the distance-frequency distribution.	217
6.13 Fit of the generalized inverse Gauss-Poisson model (GIGP, solid line), the lognormal model (dashed line), and the Yule-Simon model (dotted line) to the first 15 ranks of the frequency spectrum of Gale and Sampson's (1995) counts of consonant-vowel patterns in English.	219
6.14 Fit of the generalized inverse Gauss-Poisson model (GIGP, solid line), the Naranan-Balasubrahmanyam Zipfian model (dashed line), and a least-squares regression (dotted line) for the frequency spectrum of Gale and Sampson's (1995) counts of consonant-vowel patterns in English.	220
6.15 Fit of the generalized inverse Gauss-Poisson model (GIGP, solid line) and the Naranan-Balasubrahmanyam Zipfian model (dashed line) to the first 20 spectrum elements of the word pairs in Multatuli's Max Havelaar. The left panel plots absolute values, the right panel uses a logarithmic scale for both axes.	222
6.16 Expected vocabulary growth curves for Alice in Wonderland and Through the looking glass using binomial interpolation.	226
6.17 Observed (circles and triangles) and expected spectrum (solid and dashed lines) for Alice in Wonderland and Through the looking glass respectively.	227
6.18 Main window of lexstats.	230
6.19 Parameter specification window of lexstats.	232
6.20 Parameter specification window of lexstats for mixture models. .	233
6.21 Plot window of lexstats for the frequency spectrum.	234
6.22 Plot window of lexstats for the vocabulary and spectral growth curves.	235

List of Tables

1.1	<i>Part of the word frequency list for Alice in Wonderland. i: arbitrary index for the word types; ω_i: the i-th word type; $f(i, 26505)$: the frequency of the i-th word type in the full text of 26505 word tokens.</i>	3
1.2	<i>Sample size N, vocabulary size $V(N)$, mean $N/V(N)$, standard deviation ($stdev$) and median word frequency for Through the looking-glass and Alice in Wonderland, as well as for the first 26505 words in Through the looking-glass.</i>	5
1.3	<i>The frequency spectrum $V(m, N)$ of Alice in Wonderland.</i>	9
1.4	<i>The empirical structural type distribution $g(m, N)$ of Alice in Wonderland.</i>	12
1.5	<i>The twenty most frequent words in Alice in Wonderland ordered according to decreasing frequency.</i>	13
1.6	<i>The relation between the Zipf rank, the empirical structural type distribution, and the spectrum elements, illustrated for $m = 1, 2, 3$ in Alice in Wonderland ($N = 26505$).</i>	18
3.1	<i>Observed and expected spectrum counts for Alice in Wonderland, with $r = 15$, using the inverse Gauss-Poisson model (Sichel) with $\hat{b} = 0.0236$, $\gamma = -0.5$, and $\hat{Z} = 105.78$, the lognormal model (Carroll) with $\hat{\mu} = -5.76$ and $\hat{\sigma} = 2.69$, and the extended Zipf model (Zipf) with $\hat{Z} = 12222$.</i>	119
3.2	<i>Overview of the statistics required for calculating $R(m, k, 15)$ for Alice in Wonderland using Sichel's model with γ fixed at -0.5 a priori.</i>	122
3.3	<i>Parameters and goodness-of-fit statistics for the lognormal model, the inverse Gauss-Poisson model, the extended Zipf's law, and the Yule-Simon model for selected texts using cost functions $C_1(3)$ and $C_1(2)$ (Alice: Alice in Wonderland; Through: Through the looking-glass; Wells: The War of the Worlds; Conan Doyle: Hound of the Baskervilles; BNC: the context-governed subcorpus of the British National Corpus).</i>	126
3.4	<i>Parameters and goodness-of-fit statistics for the lognormal model, the inverse Gauss-Poisson model, the extended Zipf's law, and the Yule-Simon model for selected texts using cost function $C_2(15)$ (Alice: Alice in Wonderland; Through: Through the looking-glass; Wells: The War of the Worlds; Conan Doyle: Hound of the Baskervilles; BNC: the context-governed subcorpus of the British National Corpus).</i>	128

4.1	<i>The twenty most frequent formations with -heid (left) and twenty randomly chosen examples of hapax legomena with -heid (right).</i>	146
4.2	<i>The twenty most frequent formations with -iteit (left) and twenty randomly selected examples of hapax legomena with -iteit (right).</i>	151
5.1	<i>Dispersion d_i, expected dispersion $E[d_i]$, frequency $f(i, N)$, and Monte Carlo probability of the dispersion $\Pr(d \leq d_i)$ for selected words in Alice in Wonderland for $K = 40$ equally-sized text chunks.</i>	165
5.2	<i>Comparison of frequency estimates: m^* is the Good-Turing estimate, m° is the sample-size adjusted population frequency, and \bar{m} is the average of m^* and m. The expectations $E[V(m, N/2)]$ are based on the extended Zipf's law.</i>	170
5.3	<i>Number of content words among the most frequent Zipf ranks in groups of 50. The Frequency and Log Frequency column pertain to the highest Zipf rank in each group.</i>	172
5.4	<i>Parameters and goodness-of-fit statistics for the extended Zipf's law and the generalized inverse Gauss-Poisson model and their parameter-adjusted variants for selected texts (Alice: Alice in Wonderland; Through: Through the looking-glass; Wells: The War of the Worlds; Conan Doyle: Hound of the Baskervilles.</i>	190
6.1	<i>Legend for the texts analyzed in Figures 6.9 and 6.10.</i>	211
6.2	<i>Observed and expected values for the bigram counts in Multatuli's Max Havelaar.</i>	221

Introduction

This book is an introduction to the statistical analysis of word frequency distributions, intended for linguists, psycholinguistics, and researchers working in the field of quantitative stylistics and anyone interested in quantitative aspects of lexical structure. Word frequency distributions are characterized by very large numbers of rare words. This property leads to strange statistical phenomena such as mean frequencies that systematically keep changing as the number of observations is increased, relative frequencies that even in large samples are not fully reliable estimators of population probabilities, and model parameters that emerge as functions of the text size.

Special statistical techniques for the analysis of distributions with large numbers of rare events can be found in various technical journals. The aim of this book is to make these techniques more accessible for non-specialists. Chapter 1 introduces some basic concepts and notation. Chapter 2 describes non-parametric methods for the analysis of word frequency distributions. The next chapter describes in detail three parametric models, the lognormal model, the Yule-Simon Zipfian model, and the generalized inverse Gauss-Poisson model. Chapter 4 introduces the concept of mixture distributions. Chapter 5 explores the effect of non-randomness in word use on the accuracy of the non-parametric and parametric models, all of which are based on the assumption that words occur independently and randomly in texts. Chapter 6 presents examples of applications.

Throughout the book, concepts of probability theory and statistics necessary to understand the analysis of word frequency distributions are carefully introduced. However, as this is not an introductory textbook to statistics and probability theory, readers with little background knowledge in these fields will find it useful to consult introductory textbooks such as Ross (1988) and Rice (1988). In order to make the text generally accessible, non-technical summaries precede the more technical sections, while the mathematical derivations have been kept simple by going through the proofs with small steps. This leads to the paradoxical situation that some pages may look very scary while being quite easy to read. A great many figures illustrate key concepts and results. Chapters 1 and 6 are relatively non-technical and should be generally accessible. Chapters 2–5 require some knowledge of mostly elementary calculus. As sections 2.5, 2.6.2, 3.2–3.4, and 4.2 are fairly technical, some readers may want to restrict themselves to the non-technical summaries preceding these sections.

Four appendices are included. Appendix A is a list of symbols. Appendix B gives solutions to the exercises found at the end of the first four chapters. Appendix C provides the documentation to the programs on the CDrom that comes with this book. As there is at present no generally available software available for carrying out the kind of statistical analyses described in this

book, I am making LEXSTATS available under the GNU General Public License. LEXSTATS is a suite of programs written in C including a graphical user interface written in Tcl/Tk. Updates can be obtained from the author by e-mail at baayen@mpi.nl. LEXSTATS is supported for LINUX only. It should run without problems on UNIX platforms, and the individual C-programs will probably run on other platforms as well. All C-programs require input and produce output that is in the *data frame* format of R and Splus, so that the user is not limited to the functionality provided by the graphical user interface. Finally, Appendix D summarizes the frequency distributions of the main data sets analyzed in this book.

I am indebted to Stefan Evert, Estate Khmaladze, Anke Luedeling, Richard Sproat, Arjuna Tuzzi, and especially Kyo Kageura for their careful reading of the manuscript and their detailed comments and suggestions for improvement. Most of all, I am indebted to Fiona Tweedie, with whom I have had the opportunity to collaborate on various issues discussed in this book. Without this wonderful collaboration, the chapter on mixture distributions would not exist. I remember with gratitude my friendship with Rezo Chitashvili, to whom this book is dedicated. It is a pleasure to be able to write that the Yule-Simon model, which he developed, emerges from the present study as an excellent model for word frequency distributions. The idea of writing a book along the present lines was born in the year before his untimely death. Thanks are also due to Antoinette Renouf, who kindly provided the data sets from her large longitudinal corpus of British newspapers, to Stephen Tweedie, who introduced me to the LINUX operating system, and to Jorn Baayen, who has been an excellent LINUX system administrator. And thanks to Tineke for making it all worthwhile.

Chapter 1

Word Frequencies

This chapter introduces two fundamental issues in lexical statistics. The first issue concerns the role of the sample size, the number of words in a text or corpus. The sample size crucially determines a great many measures that have been proposed as characteristic text constants. However, the values of these measures change systematically as a function of the sample size. Similarly, the parameters of many models for word frequency distribution are highly dependent on the sample size. This property sets lexical statistics apart from most other areas in statistics, where an increase in the sample size leads to enhanced accuracy and not to systematic changes in basic measures and parameters.

The second issue concerns the theoretical assumption underlying all theoretical models and tests used in lexical statistics, namely that words occur randomly in texts. This assumption is an obvious simplification that, however, offers the possibility of deriving useful formulae for text characteristics. The crucial question, however, is to what extent this simplifying assumption affects the reliability of these formulae when applied to actual texts and corpora.

Section 1.1 illustrates these two issues by means of an exploratory investigation of word frequencies in Lewis Carroll's *Alice's Adventures in Wonderland*, henceforth *Alice in Wonderland*. Although this is a small book with only 26505 words, it is large enough to reveal the kind of phenomena that emerge, often more strongly, in larger novels and text corpora. Section 1.2 introduces the fundamental concept of the frequency spectrum. Sections 1.3–1.5 review Zipf's rank-frequency model and the lognormal model, as well as a series of statistics that have been proposed as characteristic size-invariant textual constants.

A first objective of this chapter is to show that these statistics and many model parameters are seriously affected by changes in sample size as well as by the non-random organization of discourse. Another equally important objective is to familiarize the reader with some fundamental concepts and notational conventions.

1.1 Introduction

Consider the first sentence of *Alice in Wonderland*:

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

This sentence contains 57 instances of words, some of which are used more than once. The function words *of*, *or* and *the* occur three times, *Alice* and nine other words occur twice, the remaining 28 words occur only once. In all, this sentence with 57 word instances or **word tokens** contains 41 distinct letter strings or **word types**.¹

Obviously, these frequency counts are strictly bound to this particular sentence. In larger fragments of *Alice in Wonderland*, the frequency counts for these words are quite different. For instance, by the end of the book, Alice has been mentioned 398 times. While *sister* occurs twice in the first sentence just as *Alice* does, it appears only eight times in the complete text. Conversely, the determiner *the*, which appears three times in the first sentence, has a frequency count of 1631 in the novel as a whole. These counts illustrate a simple fact: word frequency depends on sample size. Denoting the **sample size**, the number of word tokens in the sample, by N ,

Definition 1.1 N : sample size in word tokens,

I will make the dependency on the sample size of the frequency of the i -th word (ω_i) in a list of word types explicit in my notation:

Definition 1.2 $f(i, N)$: frequency of ω_i in a sample of N tokens.

frequency is not relative frequency

Table 1.1 presents part of the **word frequency list** of the complete text of *Alice in Wonderland*. For each word ω_i , the frequency $f(i, N)$ is specified.

When we increase the size of our sample, for instance, from the 57 tokens of the first sentence to the 26505 tokens in the complete text of *Alice in Wonderland*, we not only find that the frequencies of the words we have already seen increase, we also encounter new types. The number of different types $V(N)$ we count in a sample of N tokens, the **vocabulary size**, is a non-decreasing function of N .

Definition 1.3 $V(N)$: number of types in a sample of N tokens.

The solid line in panel A of Figure 1.1 plots the development of the vocabulary size $V(N)$ in *Alice in Wonderland* as a function of the sample size N , measured at twenty equally-spaced intervals.

Clearly, the growth curve of the vocabulary size $V(N)$ is not a linear function of N . Initially, the vocabulary size increases quickly, but the rate at which

¹This is a string-based definition of types and tokens. Alternatively, inflectional variants such as *conversation* and *conversations* can be classified as two tokens of the same type, instead of treating them as tokens of two different types.

Table 1.1: Part of the word frequency list for Alice in Wonderland. i : arbitrary index for the word types; ω_i : the i -th word type; $f(i, 26505)$: the frequency of the i -th word type in the full text of 26505 word tokens.

i	ω_i	$f(i, 26505)$	i	ω_i	$f(i, 26505)$
1	a	629	23	of	510
2	alice	386	24	on	194
3	alice's	12	25	once	34
4	and	866	26	one	102
5	bank	3	27	or	77
6	beginning	14	28	peeped	3
7	book	7	29	pictures	4
8	but	170	30	reading	3
9	by	57	31	she	540
10	conversation	10	32	sister	8
11	conversations	1	33	sitting	10
12	do	81	34	the	1631
13	get	46	35	thought	74
14	had	177	36	tired	7
15	having	10	37	to	726
16	her	247	38	twice	5
17	in	365	39	use	18
18	into	67	40	very	144
19	is	108	41	was	356
20	it	528	42	what	136
21	no	90	43	without	26
22	nothing	34	44	...	

the vocabulary size increases as we proceed through the text decreases. By the end of the novel the vocabulary growth curve has not flattened out to a horizontal line. A horizontal line would have implied that no new words are added as N increases, which would have indicated that the full set of words judged by Carroll to be appropriate for this kind of story had been used. Instead, it is clear that if the story had continued, more new words would have appeared. Although we can regard *Alice in Wonderland* as the statistical population when we focus on this story as a literary unit, we can equally well view *Alice in Wonderland* as a sample of Carroll's language use. From the latter perspective, the shape of the growth curve $V(N)$ reveals that we have only just begun to sample Carroll's vocabulary.

Suppose that we want to compare *Alice in Wonderland* with *Through the looking-glass and what Alice found there*, henceforth *Through the looking-glass*. The latter is Carroll's second story about Alice. We might hypothesize that Carroll benefited from his experience in writing *Alice in Wonderland*, and that his greater experience as a writer might have lead to a more abundant use of the lexical resources of English. In other words, *Through the looking-glass* might be characterized by the greater vocabulary richness. Comparing $V(N)$ for the two books, we find that *Through the looking-glass* ($V(29028) = 2877$) has

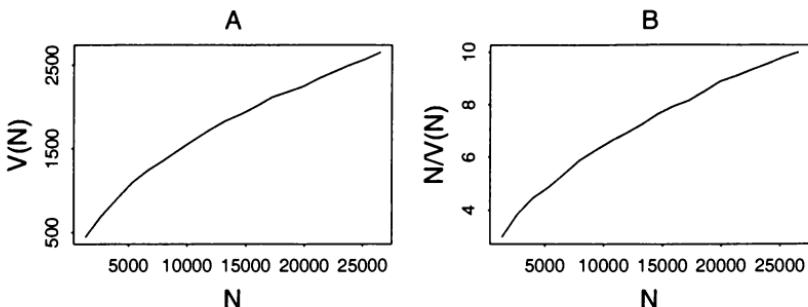


Figure 1.1: Vocabulary size $V(N)$ (panel A) and mean word frequency $N/V(N)$ (panel B) as a function of sample size N in *Alice in Wonderland*, measured at 20 equally-spaced intervals.

226 more types than *Alice in Wonderland* ($V(26505) = 2651$). However, since *Alice in Wonderland* is shorter than *Through the looking-glass*, we cannot simply compare their respective vocabulary sizes. If *Alice in Wonderland* had been longer, it would have contained more different words, possibly even more than *Through the looking-glass*.

$$\text{sum}_i (f(i, N) / V(N)) = N / V(N)$$

To adjust for the difference in text size, it seems reasonable at first sight to consider the mean token frequency of the word types, henceforth the **mean word frequency**. A greater vocabulary richness, one would think, should be reflected in a lower mean word frequency $N/V(N)$. For *Alice in Wonderland*, the mean word frequency is 10.00, for *Through the looking-glass*, the mean frequency equals 10.09. These numbers suggest that our hypothesis is wrong, and that Carroll did not exploit the lexical resources of English more fully in his second book.

Interestingly, this conclusion is unjustified. To see this, consider Panel B of Figure 1.1, which plots the mean word frequency $N/V(N)$ for *Alice in Wonderland* at twenty equally-spaced measurement points. The solid line shows the development of the mean through sampling time. Instead of randomly fluctuating around some fixed value, the mean word frequency increases non-linearly with the sample size in a similar way as the vocabulary size $V(N)$. Not only the mean, but also the median changes. For the first six measurement points, the median frequency equals 1, for the remaining measurement points, it equals 2. Apparently, simple statistics such as mean and median do not converge to their population values within the sample. This observation holds not only for a small book such as *Alice in Wonderland*, it generalizes to large novels and even to text corpora with tens of millions of words. Normally, individual sample means fluctuate randomly around the theoretical mean, with larger deviations for smaller samples and smaller deviations for larger sam-

ples. Apart from the magnitude of the deviations, there usually is no systematic pattern to the changes in the sample means as the sample size is increased. In the domain of lexical statistics, however, mean frequencies behave surprisingly differently, revealing a non-linear increase with N . In fact, panel B of Figure 1.1 seems a flat contradiction of the fact that the sample mean should become an increasingly accurate estimate of the population mean as the sample size increases. As we shall see, this is due to two factors. One factor is that we are dealing with counts of types instead of with properties of types. A second factor is the large numbers of extremely low-probability words that are present in lexical frequency distributions, distributions that belong to the class of **Large Number of Rare Events (LNRE)** distributions.

Table 1.2: Sample size N , vocabulary size $V(N)$, mean $N/V(N)$, standard deviation (stdev) and median word frequency for Through the looking-glass and Alice in Wonderland, as well as for the first 26505 words in Through the looking-glass.

	N	$V(N)$	$N/V(N)$	stdev	median
<i>Through the looking-glass</i>	29028	2877	10.09	50.91	2
	26505	2731	9.71	47.31	2
<i>Alice in Wonderland</i>	26505	2651	10.00	51.14	2

The dependency of the sample mean on the sample size implies that we have to correct for the difference in sample size before comparing the sample means. Table 1.2 shows that when we compare an equal number of tokens of the two texts, for instance, by selecting the first 26505 words of *Through the looking-glass*, we find that mean frequency for this text has decreased from 10.09 to 9.71, a change in the direction of our original hypothesis. Normally, fluctuations in the value of a mean are due to sampling error and should not be assigned significance. But we have seen that for our lexical data the sample mean increases as a function of the sample size. In our example, the smaller mean frequency observed for $N = 26505$ is not due to sampling error, it is a systematic change in the expected direction. From this point of view, we underestimate the difference in vocabulary richness between the two texts when we use the means of the complete texts. But if we adjust for sample size, we might as well compare the vocabulary sizes directly, instead of focusing on mean frequency. Carroll's second book contains 80 more types among its first 26505 tokens than his first, 3% of the vocabulary size of *Alice in Wonderland*. This suggests that *Through the looking-glass* displays the greater vocabulary richness. In section 3.6, I will introduce a technique for testing whether this difference in vocabulary size is statistically significant.

A second fundamental issue in lexical statistics concerns the non-random use of words in actual texts. However, to construct probabilistic models for word frequency distributions, models that, for instance, yield expressions for $V(N)$ as a function of N , it is convenient to assume that words occur randomly in texts. This is an obvious simplification. In a random rearrangement of all the words of *Alice in Wonderland*, the first 57 words are:

More find likely a somebody a you're lost again was you invent waited a on to time passion so partner about and with panting back-somersault queen as was were the open obliged ask the Alice much a do your as on if face come crab best not rapped gryphon I affair I to it see unlocking low.

Unlike the first sentence of *Alice in Wonderland*, this sequence of words is not semantically coherent. Moreover, sequences of words occur that are ruled out by the rules of syntax (*the Alice, a you're, was were, I to it see*). However, the methodological point at issue here is not whether the randomness assumption is wrong, but to what extent the simplifying assumption of random word use affects the accuracy of theoretical models. Are the effects of non-randomness visible at higher levels of abstraction? Do they introduce significant deviation between theoretically predicted and empirically observed values for statistics such as the vocabulary size $V(N)$? Are effects of non-randomness visibly present in the frequencies of individual words?

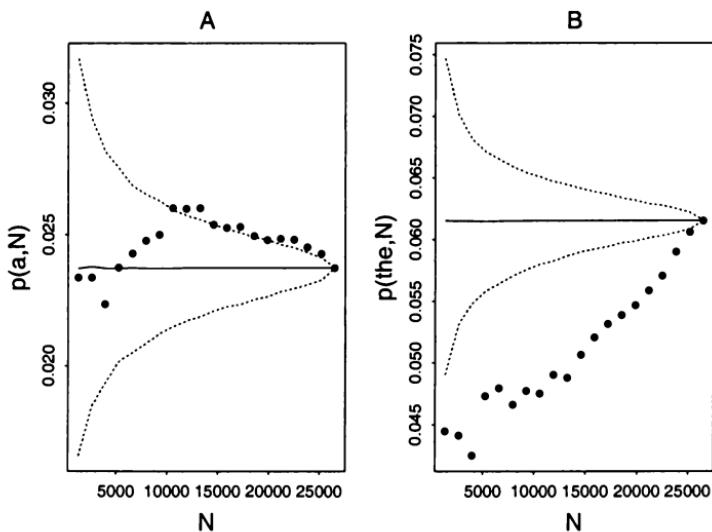


Figure 1.2: The sample relative frequency of the article a ($p(a, N)$) (panel A) and the sample relative frequency of the article the ($p(\text{the}, N)$) (panel B) as a function of the sample size N in *Alice in Wonderland*, measured at 20 equally-spaced intervals. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on a total of 5000 permutation runs.

Figure 1.2 plots the **sample relative frequencies** $p(i, N)$ of the definite and indefinite article in *Alice in Wonderland* using large dots.

Definition 1.4 $p(i, N) = \frac{f(i, N)}{N}$: sample relative frequency of ω_i .

If the articles are used randomly throughout the text, their sample relative frequencies should be approximately the same for any sample size N , i.e., their sample relative frequencies should show up as a horizontal line in a graph of $p(i, N)$ as a function of N . Figure 1.2, however, reveals that the observed sample relative frequencies do not show up as horizontal lines. Instead, they reveal non-random developmental profiles. The indefinite article *a* (panel A) is used more intensively in the central sections of the book than in the beginning or end. The definite article *the* (panel B) shows a more or less linear increase in relative sample frequency.

These developmental profiles might be due to chance. What is the probability that a random re-ordering of the words of *Alice in Wonderland* would lead to a similar pattern? We can approach this question by calculating the mean sample relative frequencies at twenty equally-spaced intervals for a large number of random permutations of the order of the words in *Alice in Wonderland*. As we shall see in more detail in Chapter 5, in a random permutation of the text the effects of cohesion in word use at the levels of sentence and discourse is eliminated. If the values actually observed for the text are more extreme than those observed for 95% of the permutation runs, then we know that the probability that the observed pattern arose by chance is less than 0.05.

Panels A and B of Figure 1.2 show the result of this randomization test for a total of 5000 permutations runs. For both articles, the average proportion or Monte Carlo mean, represented by a solid line, is constant, exactly what we expect when the tokens of a word are equally spread out over the text. The dotted lines mark the two-tailed 95% Monte Carlo confidence interval. Panel A shows that for measurement points 8–10, 13–14, and 16–19 the observed values for $p(a, N)$ fall outside this confidence interval. Apparently, *a* tends to be slightly overrepresented in the second half of the text. Turning to panel B, we find that with the exception of the final measurement point (the full text), all observed relative sample frequencies of *the* are well below the lower 95% confidence limit. Again we may conclude that the observed pattern for *the* is far from random.

The empirical developmental profiles of the articles are possibly linked with narrative development in *Alice in Wonderland* in the following way. In the initial sections of the book, new participants and new scenes are introduced, but as one continuous reading, Alice re-encounters participants and revisits places where she had been before. Since the indefinite article typically introduces new information and the definite article given information, the increase in the use of *the* and the decrease in the use of *a* in the second half of the text might be the consequence of thematic development in the narrative. Note, however, that this leaves the increase in the relative frequency of *a* in the first half of the text unexplained.

Summing up, in statistical analyses of textual data it is important to realize that the values of simple statistics such as means and proportions are heavily influenced by the sample size, for two reasons. First, the law of large numbers cannot be relied on when dealing with words and their frequencies of use. Second, authors do not use words at random. Word usage reflects lexical

cohesion both at the level of the sentence and at the level of discourse. These two factors should always be kept in mind when comparing the quantitative properties of textual materials.

1.2 The frequency spectrum

We have seen that statistics such as the sample mean and median increase when the sample size is increased. The variability of the sample mean has severe consequences for the comparison of texts. As illustrated for *Alice in Wonderland* and *Through the looking-glass*, it is possible to adjust for differences in sample size by considering a subset of the word tokens in the larger set. Such a procedure, however, has some obvious drawbacks. First, data in the larger sample have to be discarded, which implies a loss of information. Second, there are no good criteria available for deciding which tokens in the larger text should be discarded. Especially for novels and cohesive texts in general, the removal of any part of the text is completely arbitrary. Not surprisingly, considerable effort has been spent on the development of quantitative measures that characterize textual properties independently of sample size. This chapter reviews a series of such statistics. Unfortunately, the main thrust of the argument is a negative one: Almost all 'constants' proposed in the literature reveal specific developmental profiles in sampling time just as the sample mean and median. Before discussing a number of measures that have been put forward as text constants, I should first introduce the concept of the **grouped frequency distribution** or **frequency spectrum**.

Word frequencies in *Alice in Wonderland* range from 1 to 1631. The most frequent word is *the*, and this is the only word with this particular token frequency. Conversely, the lowest token frequency, 1, is represented by 1176 different words. The words which occur once only in a text are known as **hapax legomena**, from Greek *hapax*, 'once', and *legomenon*, 'read'. Typically, 1 is the frequency that is represented by the greatest number of words. The number of words that occur twice in *Alice in Wonderland*, 402, the so-called **dis legomena**, is substantially smaller, but in its turn almost twice the number of words that occur three times, 233. I will use the index m to denote these frequency classes. The number of word types in a given frequency class for a sample of size N will be denoted by $V(m, N)$. Thus, $V(1, N)$ denotes the number of hapax legomena, $V(2, N)$ the number of dis legomena, etc.

Definition 1.5 m : index for frequency class.

Definition 1.6 $V(m, N) = \sum_{i=1}^{V(N)} I_{[f(i, N)=m]}$: the number of types with frequency m in a sample of N tokens.

The identity operator $I_{[\alpha]}$ that appears in the definition of $V(m, N)$ yields the value 1 if the expression α is true, and zero otherwise. Note that N and $V(N)$ can be expressed in terms of m and $V(m, N)$:

$$N = \sum_m m V(m, N) \tag{1.1}$$

Table 1.3: *The frequency spectrum $V(m, N)$ of Alice in Wonderland.*

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
1	1176	31	3	62	1	144	1
2	402	32	4	63	1	145	1
3	233	33	4	67	2	148	1
4	154	34	3	68	4	151	1
5	99	35	4	73	1	153	1
6	57	37	1	74	1	170	1
7	65	38	4	75	1	177	1
8	52	39	4	77	2	179	1
9	32	40	4	79	1	182	1
10	36	41	2	80	1	194	1
11	23	42	2	81	1	211	1
12	20	43	2	82	2	247	1
13	34	44	1	83	2	263	1
14	20	45	4	85	1	280	1
15	12	46	1	87	1	356	1
16	9	47	1	88	2	364	1
17	9	48	1	90	1	365	1
18	10	49	4	93	1	386	1
19	8	50	2	94	1	410	1
20	5	51	4	96	2	460	1
21	6	52	3	98	1	510	1
22	3	53	1	102	2	528	1
23	3	54	3	108	1	540	1
24	6	55	3	113	1	629	1
25	9	56	1	114	1	726	1
26	4	57	2	121	1	866	1
27	6	58	2	128	1	1631	1
28	3	59	1	131	1		
29	6	60	2	133	1		
30	6	61	3	136	1		

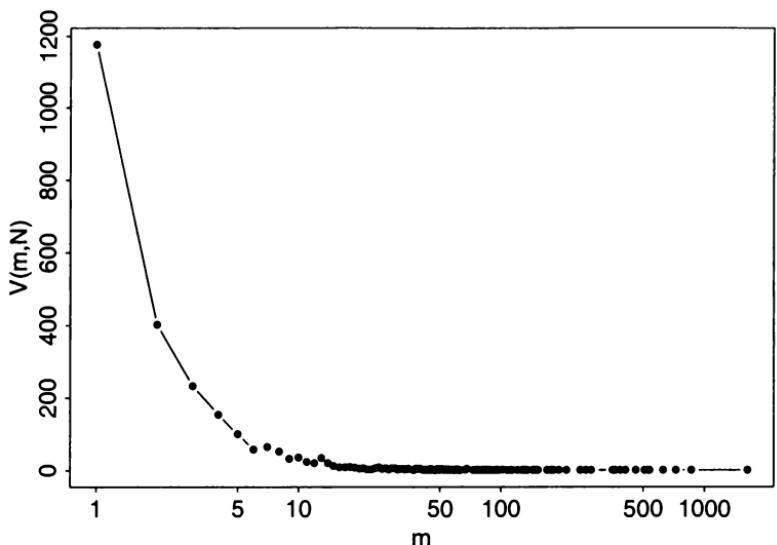


Figure 1.3: The frequency spectrum of Alice in Wonderland (m : frequency class; $V(m, N)$: number of types with frequency m).

$$V(N) = \sum_m V(m, N). \quad (1.2)$$

Table 1.3 lists $V(m, N)$ for the $N = 26505$ tokens of *Alice in Wonderland*, and Figure 1.3 visualizes this grouped frequency distribution or **frequency spectrum**. The horizontal axis plots the frequency classes m using a logarithmic scale. The vertical axis plots $V(m, N)$. Note that $V(m, N)$ is a rapidly decreasing function of m with a long tail of high frequencies m that are instantiated by very few types.

The curve of the frequency spectrum is smooth enough to suggest that in theory $V(m, N)$ might be a monotonically decreasing function of m for which the inequality

$$V(m, N) > V(m + 1, N) \quad (1.3)$$

holds. The irregularities observed from $m = 6$ onwards (see also Table 1.3) would then be due to sampling error. Highly skewed frequency spectra of this kind are typical in lexical statistics.

Closely related to the grouped frequency distribution $V(m, N)$ is the so-called **empirical structural type distribution** $g(m, N)$, which specifies the number of different word types which occur m or more times in a sample of N tokens.

Definition 1.7 $g(m, N) = \sum_{i=1}^{V(N)} I_{\{f(i, N) \geq m\}}$: the number of types with a frequency m or more in a sample of N tokens.

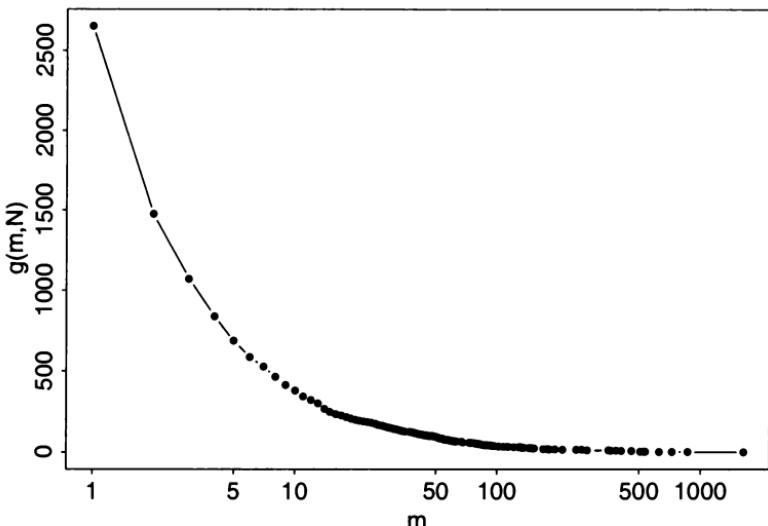


Figure 1.4: The empirical structural type distribution of *Alice in Wonderland* (m : frequency class; $g(m, N)$: number of types occurring m or more times).

Table 1.4 presents the empirical structural type distribution of *Alice in Wonderland*, and Figure 1.4 plots the corresponding empirical structural type distribution, again using a logarithmic scale for the horizontal axis. Note that $g(1, N)$ is equal to $V(N)$, and that for *the*, the highest-frequency word in the text, $g(1631, N) = 1$.

The grouped frequency distribution and the empirical structural type distribution are related by the following expressions:

$$V(m, N) = g(m, N) - g(m + 1, N), \quad (1.4)$$

$$g(m, N) = \sum_{w \geq m} V(w, N). \quad (1.5)$$

Table 1.4: *The empirical structural type distribution $g(m, N)$ of Alice in Wonderland.*

m	$g(m, N)$	m	$g(m, N)$	m	$g(m, N)$	m	$g(m, N)$
1	2651	31	143	62	67	144	27
2	1475	32	140	63	66	145	26
3	1073	33	136	67	65	148	25
4	840	34	132	68	63	151	24
5	686	35	129	73	59	153	23
6	587	37	125	74	58	170	22
7	530	38	124	75	57	177	21
8	465	39	120	77	56	179	20
9	413	40	116	79	54	182	19
10	381	41	112	80	53	194	18
11	345	42	110	81	52	211	17
12	322	43	108	82	51	247	16
13	302	44	106	83	49	263	15
14	268	45	105	85	47	280	14
15	248	46	101	87	46	356	13
16	236	47	100	88	45	364	12
17	227	48	99	90	43	365	11
18	218	49	98	93	42	386	10
19	208	50	94	94	41	410	9
20	200	51	92	96	40	460	8
21	195	52	88	98	38	510	7
22	189	53	85	102	37	528	6
23	186	54	84	108	35	540	5
24	183	55	81	113	34	629	4
25	177	56	78	114	33	726	3
26	168	57	77	121	32	866	2
27	164	58	75	128	31	1631	1
28	158	59	73	131	30		
29	155	60	72	133	29		
30	149	61	70	136	28		

1.3 Zipf

Among the earliest studies on word frequency distributions the work by Zipf (1935, 1949) figures prominently. Zipf ordered the words in his texts by decreasing frequency, and considered the relation between rank order and frequency. Consider Table 1.5, which lists the twenty most frequent words in

Table 1.5: *The twenty most frequent words in Alice in Wonderland ordered according to decreasing frequency.*

z	$f_z(z, N)$	word	z	$f_z(z, N)$	word
1	1631	<i>the</i>	11	365	<i>in</i>
2	866	<i>and</i>	12	364	<i>you</i>
3	726	<i>to</i>	13	356	<i>was</i>
4	629	<i>a</i>	14	280	<i>that</i>
5	540	<i>she</i>	15	263	<i>as</i>
6	528	<i>it</i>	16	247	<i>her</i>
7	510	<i>of</i>	17	211	<i>at</i>
8	460	<i>said</i>	18	194	<i>on</i>
9	410	<i>I</i>	19	182	<i>all</i>
10	386	<i>Alice</i>	20	179	<i>with</i>

Alice in Wonderland in their Zipfian rank order. The most frequent word, *the*, is assigned the **Zipf rank** $z = 1$, the next most frequent word, *and*, is assigned rank $z = 2$, and so on. Words with the same frequency are arranged in some arbitrary order and they receive successively larger Zipf ranks. For instance, the 1176 hapax legomena in *Alice in Wonderland* are assigned the Zipf ranks 1476, 1477, 1478, ..., 2651. (This implies that the actual Zipf rank of a hapax legomenon is not of interest, but rather the ranks at which the first and last hapax legomenon are observed.) I will use the notation $f_z(z, N)$ for the frequency of a word with Zipf rank z . Thus $f_z(1, N)$ is the frequency of the word with Zipf rank 1, the subscript indicating that the frequency is to be understood as with respect to a Zipfian ranking.

Definition 1.8 z : Zipf rank in a word list ordered by decreasing frequency.

Definition 1.9 $f_z(z, N)$: frequency in a sample of N tokens of a word with Zipf rank z .

The **Zipfian rank-frequency distribution** is the inverse of the empirical structural type distribution:

$$g(m, N) = z \Leftrightarrow f_z(z, N) = m. \quad (1.6)$$

For instance, for the highest-frequency word in *Alice in Wonderland*, *the*, we have

$$g(1631, N) = 1,$$

but at the same time

$$f_z(1, N) = 1631.$$

Similarly, at the low-frequency end of the frequency spectrum we have:

$$\begin{aligned} g(1, N) &= 2651, \\ f_z(2651, N) &= 1. \end{aligned}$$

In general, if a word with frequency m has Zipf rank z , then the frequency ordering underlying the Zipf ranking implies that there are z words with at least frequency m , which in turn implies that $g(m, N) = z$.

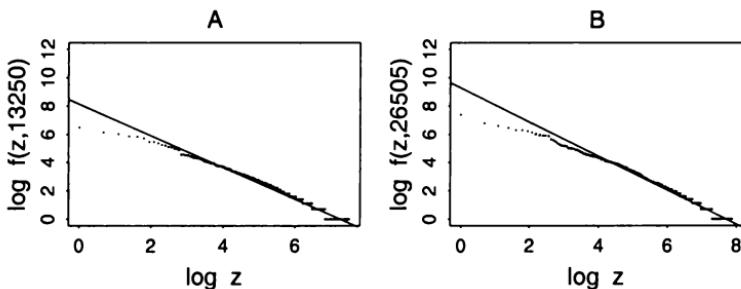


Figure 1.5: Word frequency $f_z(z, N)$ as a function of Zipf rank z in the double logarithmic plane for $N = 13250$ (panel A) and for $N = 26505$ (panel B).

Panel A of Figure 1.5 is a points plot of $\log f_z(z, N)$ against $\log z$ for the first 13250 tokens of *Alice in Wonderland*. The solid line is the corresponding least squares regression line. Note that the highest frequencies appear as individual points at the left hand edge of the plot, and that the large numbers of hapax legomena and dis legomena appear as horizontal line segments at the right-hand edge of the plot. The corresponding plot in panel B for the complete text reveals a highly similar pattern. Zipf observed such roughly linear plots for many different kinds of texts. This led him to formulate the following relation between $f_z(z, N)$ and z :

$$f_z(z, N) = \frac{C}{z^a}. \quad (1.7)$$

In (1.7), a is the free parameter of the model that determines the slope of the regression lines in Figure 1.5, C is a normalizing constant. Taking logarithms at both sides, the linear relation between $\log f_z(z, N)$ and $\log z$ follows immediately:

$$\begin{aligned} \log f_z(z, N) &= \log \frac{C}{z^a} \\ &= \log C - a \log z. \end{aligned}$$

This inverse relation between log Zipf rank and log frequency is known as **Zipf's law**.

Thus far, we have considered Zipf's law in terms of the absolute sample frequencies of words and their Zipf ranks. Absolute sample frequencies, however, are subject to sampling error, and will therefore diverge slightly from Zipf's law. Theoretically, the frequency of a word is expected to be $N\pi_i$, with π_i the population probability of ω_i (see section 2.2). Underlying the observed frequencies, there is a distribution of probabilities for which Zipf's law should also be valid. Let's therefore reformulate Zipf's law in terms of the probabilities of words. Instead of focusing on $f_z(z, N)$, we first consider the corresponding **relative sample frequencies** $f_z(z, N)/N$. Assume furthermore, for the sake of the argument, that these sample relative frequencies are reliable estimates of the population probabilities.

Definition 1.10 π_z : probability of the word ω_z with Zipf rank z .

Definition 1.11 $\hat{\pi}_z$: the probability of word ω_z estimated from its sample relative frequency: $\hat{\pi}_z = f_z(z, N)/N$.

We can now reformulate Zipf's law as

$$\pi_z = \frac{C}{z^a}. \quad (1.8)$$

In both (1.7) and (1.8), C is a normalizing constant. Its function in (1.8) is to ensure that the probabilities sum up to unity:

$$\sum_z \pi_z = 1.$$

In (1.7), it ensures that the frequencies $f_z(z, N)$ sum up to N :

$$\sum_z f_z(z, N) = N.$$

(Note that this implies that the value of the normalizing constant in (1.7) is N times that in (1.8). Going from frequencies to estimated probabilities therefore implies a leftward shift by $\log N$ in the double logarithmic plane.)

The probability distribution (1.8) is known as the **zeta distribution**, which owes its name to the Riemann Zeta function (after the German mathematician G.F.B. Riemann)

$$\zeta(a) = 1 + \left(\frac{1}{2}\right)^a + \left(\frac{1}{3}\right)^a + \dots + \left(\frac{1}{n}\right)^a + \dots$$

Apart from the (normalizing) constant C , the successive terms of the Riemann Zeta function spell out the probabilities of the words with Zipf rank $z = 1, 2, \dots$. In other words, the probability of a word with the n -th rank is given by the n -th term in the expansion of $\zeta(a)$. Thus we can restate the zeta

function in terms of the Zipf probabilities (1.8):

$$\begin{aligned}\zeta(a) &= 1 + \left(\frac{1}{2}\right)^a + \left(\frac{1}{3}\right)^a + \dots + \left(\frac{1}{n}\right)^a + \dots \\ &= (\pi_1 + \pi_2 + \pi_3 + \dots + \pi_n + \dots) \frac{1}{C}.\end{aligned}$$

Zipf often took a to equal unity, in which case the zeta function reduces to the so-called harmonic series

$$\sum_z \frac{1}{z} \quad (z = 1, 2, 3, \dots).$$

Zipf (1935) refers to the corresponding probability distribution

$$\pi_z = \frac{C}{z}$$

as the **standard harmonic distribution**.²

Given the standard harmonic distribution, $V(m, N)$ can be expressed as a function of m :

$$V(m, N) \propto \frac{1}{m(m+1)}. \quad (1.9)$$

To see this, consider two Zipf ranks z_1 and z_2 such that $f_z(z_1, N) = m+1$ and $f_z(z_2, N) = m$, with $m > 0$, where we choose z_1 such that there is no $z > z_1$ for which $f_z(z, N) = m+1$, and similarly z_2 such that there is no $z > z_2$ for which $f_z(z, N) = m$. In other words, we focus on the jumps in the rank-frequency step function, as illustrated graphically in Figure 1.6, and numerically for $m = 1, 2, 3$ in *Alice in Wonderland* in Table 1.6. Crucially, we can write

$$V(m, N) = z_2 - z_1.$$

Since $f_z(z_2, N) = \frac{C}{z_2} = m$ we have $z_2 = \frac{C}{m}$. Likewise, $z_1 = \frac{C}{m+1}$, and hence

$$V(m, N) = z_2 - z_1 = \frac{C}{m} - \frac{C}{m+1} = \frac{C}{m(m+1)}.$$

²The harmonic distribution does not converge. Because the probabilities π_z do not sum up to unity, it is not a proper probability distribution. However, since

$$\sum_{i=1}^V \frac{1}{i} - \log(V) \cong \gamma$$

for $V \rightarrow \infty$, with $\gamma = 0.57723$ (Euler's constant), we have that

$$\frac{1}{\log(V) + \gamma} \sum_{i=1}^V \frac{1}{i} = 1,$$

which allows us to use the harmonic distribution as an approximation for a probability distribu-

Since the number of hapax legomena tends to be roughly half the vocabulary size, the normalizing constant C is often taken to be $V(N)$:

$$V(m, N) = \frac{V(N)}{m(m+1)}. \quad (1.10)$$

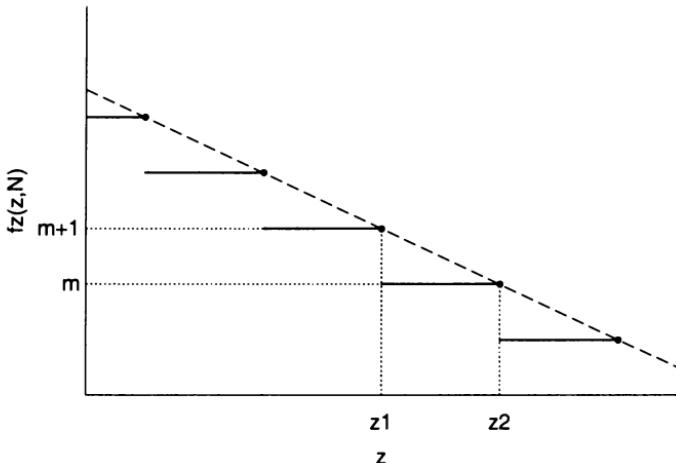


Figure 1.6: The rank-frequency distribution as a step function in the double logarithmic plane and the relation with the elements of the frequency spectrum: $V(m, N) = z_2 - z_1$.

Zipf (1935:47) hoped that the standard harmonic distribution would provide

... Dynamic Philology with a *standard curve of distribution* in reference to which the frequency distribution of any other language can be described. If the curve of the frequency distribution of a given language conforms at any point with the standard harmonic curve or if it deviates at any point either slightly or seriously above or below, these facts may shed welcome light on significant factors in the structure of language.

Zipf's hopes have been partially fulfilled. It is clear that in panel A of Figure 1.5 the frequencies of the lowest ranks deviate substantially from their theoretical values according to (1.7). Subsequent research by Mandelbrot (1953) has suggested that this deviation can be captured by introducing a second free parameter in the model, a proposal to which we discuss in more detail in section 3.2.3. Finally, the words occurring with the highest Zipf ranks are typically function words such as *the*, *a*, *for*, and *she* that have properties that differ fundamentally from content words such as *sister*, *white*, and *rabbit*.

The usefulness of Zipf's model, however, is severely limited because its parameters are highly dependent on the sample size N . To see this, consider again the graphs shown in Figure 1.5. Panel B plots the Zipfian curve

Table 1.6: The relation between the Zipf rank, the empirical structural type distribution, and the spectrum elements, illustrated for $m = 1, 2, 3$ in Alice in Wonderland ($N = 26505$).

Zipf rank	Number of spectrum elements
841	
:	
1073	$= g(3, N)$
1074	
:	
1475	$= g(2, N)$
1476	
:	
2651	$= g(1, N)$

233 tris legomena $= g(3, N) - g(4, N)$
 402 dis legomena $= g(2, N) - g(3, N)$
 1176 hapax legomena $= g(1, N) - g(2, N)$

for $N = 26505$, twice the sample size of panel A. The general shapes of the two curves are highly similar, although the divergence from linearity at the left hand side seems to have increased somewhat for the full text (panel B). More disconcerting is the observation that the slope increases from -1.119 for $N = 13250$ to -1.205 for $N = 26505$, and that the changes in the parameters of the model as a function of N are systematic, as shown in Figure 1.7. Panel A plots the intercept as a function of the sample size N , and panel B the slope. The intercept is an increasing function of N , the slope is a decreasing function of the sample size. Clearly, the parameters of the zeta distribution are subject to the same kind of systematic variation as the sample mean frequency. A way to take this variability into account in a principled way will be presented in Chapter 3.

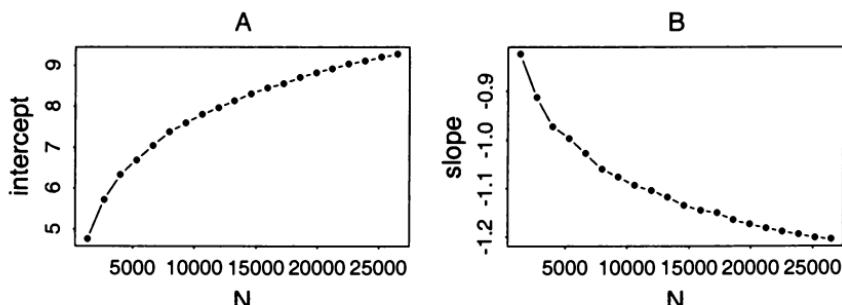


Figure 1.7: The dependency of the two parameters of Zipf's zeta distribution, the intercept (panel A) and the slope (panel B) on the sample size N , plotted at 20 equally-spaced intervals for Alice in Wonderland (compare 1.5).

Thus far, we have considered Zipf's law as formulated for the rank-frequency distribution. Zipf also proposed to analyze the frequency spectrum itself in terms of the zeta distribution. When we plot $V(m, N)$ against m in the double logarithmic plane, as shown for *Alice in Wonderland* in the left panel of Figure 1.8, we again find a roughly linear relation. Zipf (1935:40–44) points out that a model of the form

$$V(m, N) \propto \frac{1}{m^a} \quad (1.11)$$

is accurate primarily for the smaller values of m . This point is highlighted by the dashed line, which represents a linear fit $\log(V(m, N)) = a + b \log(m)$ based on the first 15 ranks. For these first 15 ranks, the fit seems quite reasonable, but it clearly does not capture the pattern among the higher-frequency ranks, which tend to have higher values than expected. When we base a linear fit on the full spectrum, the regression line, represented by a dotted line, deviates considerably from the observed lowest-frequency ranks.

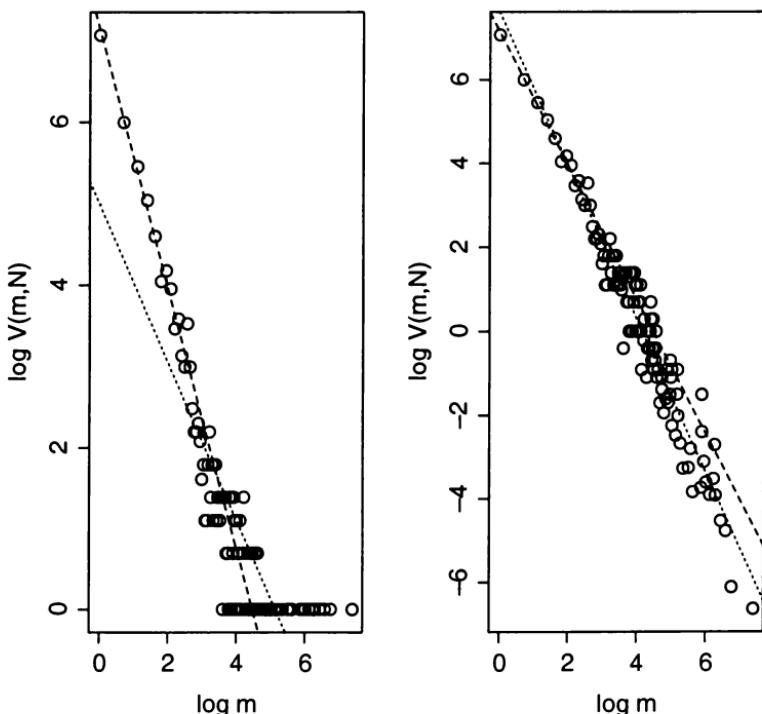


Figure 1.8: The frequency spectrum of Alice's Adventures in Wonderland in the double logarithmic plane. Left panel: the integer-valued spectrum with linear fits for $m = 1 \dots 15$ (dashed line) and for the complete range of m (dotted line). Right panel: the same spectrum elements after transformation into fractional values for higher-frequency ranks, with linear fits for $m = 1 \dots 15$ (dashed line) and for the complete range of m (dotted line).

In part, the problem that we are dealing with is a discretization problem. Word frequencies are integers, yet the power model

$$V(m, N) = am^b$$

which yields the straight line $\log(V(m, N)) = \log(a) + b \log(m)$ in the bi-logarithmic plane, expects words to have real-valued frequencies. Instead of having sparsely populated high-frequency ranks m with words occurring with integer frequencies, the model assumes that all ranks are populated with fractional numbers of types that become smaller as m increases. By itself, this need not be a problem, as long as there is a way to transform an integer-valued distribution into a real-valued distribution.

Church and Gale (1991) and Gale and Sampson (1995) propose the following technique to obtain fractional spectrum elements for the sparsely populated higher frequency ranks m . Let m_p and m_f denote the ranks for which $V(m, N) > 0$ that immediately precede and follow rank m . If $V(m-1, N) > 0$, we have that $m_p = m-1$, otherwise, $m_p < m-1$. Likewise, if $V(m+1, N) > 0$, then $m_f = m+1$, otherwise, it will be greater than $m+1$ due to intervening zero ranks. We can now define the real-valued $V_r(m, N)$ as follows:

$$V_r(m, N) = \begin{cases} V(1, N) & \text{if } m = 1 \\ \frac{2V(m, N)}{m_f - m_p} & \text{if } 1 < m < \max(m), \\ \frac{2V(m, N)}{2m - m_p} & \text{if } m = \max(m). \end{cases} \quad (1.12)$$

When zero ranks intervene between m and m_p or m_f , the difference $m_f - m_p$ will be greater than 2, so that $V_r(m, N) < V(m, N)$, otherwise $V_r(m, N)$ will be the same as $V(m, N)$. The right-hand panel of Figure 1.8 plots the resulting $V_r(m, N)$ for *Alice's Adventures in Wonderland* in the bi-logarithmic plane using circles, which now approximately follow a straight line. The downward curvature in the observed values for the lowest frequency ranks is not eliminated, however, as shown by the dashed line, a least squares regression line to the first 15 ranks. In contrast to the dotted line in the left panel, the dotted line in the right panel, the least-squares regression line using all observations, seems reasonable for all but the first two and last two ranks.

Definition (1.12) of $V_r(m, N)$ sets $V_r(m, N)$ to zero whenever $V(m, N) = 0$. Thus, following Church and Gale (1991), the right-hand panel of Figure 1.8 plots exactly the same number of points as the left panel of Figure 1.8. But because some integer-valued spectrum elements have been replaced by smaller real-valued spectrum elements, we have that

$$\sum_m V_r(m, N) < \sum_m V(m, N) = V,$$

$$\sum_m m V_r(m, N) < \sum_m m V(m, N) = N.$$

The discrepancy for $V(N)$, however, is easily solved by defining $V_r(m, N)$ for zero ranks as well. Let m_0 denote a rank for which $V(m, N) = 0$, and let p_m

denote the greatest nonzero rank smaller than m_0 and m_f the smallest nonzero rank greater than m_0 . Then

$$V_r(m_0, N) = \frac{V_r(m_p, N) + V_r(m_f, N)}{m_f - m_p}, \quad (1.13)$$

the average of the nonzero spectrum elements surrounding m_0 . With the addition of $V_r(m_0, N)$, the discrepancy between $V(N)$ and $\sum_m V_r(m, N)$ is removed. To see why this is so, consider a spectrum element m with $k = m_f - m_p - 2$ surrounding zero spectrum elements. The $V(m, N)$ types of rank m are reset to $2V(m, N)/(k + 2)$, leaving $kV(m, N)/(k + 2)$ fractional types which we divide equally among each of the k empty ranks. In this way, all $V(m, N)$ types are still present in the distribution, but now divided over $k + 1$ ranks instead of being concentrated in one rank m only. An empty rank m_0 therefore receives $V(m_p, N)/(k + 2)$ fractional types from its nearest left nonzero rank, and likewise $V(m_f, N)/(k + 2)$ fractional types from its nearest right nonzero rank, which immediately leads to (1.13). There is no guarantee, however, that $\sum_m m V_r(m, N)$ will equal N . For our present example of *Alice's Adventures in Wonderland*, $\sum_m m V_r(m, N) = 26893.61$ instead of 26505, even though $\sum_m V_r(m, N)$ is now identical to $V(N)$. This implies that the values of $V_r(m, N)$ are slightly too high.

Figure 1.8 illustrates that fitting a straight line to the real-valued approximate spectrum elements may not do justice to the slight downward curvature at the head of the spectrum. The left panel of Figure 1.9 shows that this curvature is handled in a more principled way when we smooth the spectrum using (1.10), with $V(26505) = 2651$:

$$V_z(m, N) = \frac{2651}{m(m+1)}.$$

Evaluating the goodness of fit in terms of the mean squared error (MSE) for the first 40 ranks,

$$\text{MSE}_{(40)} = \frac{\sum_{i=1}^{40} (V(m_i, N) - V_z(m_i, N))^2}{40},$$

we find that using (1.10) instead of a simple linear fit reduces the MSE from 30641.48 to 648.23. Although qualitatively and quantitatively a substantial improvement, the MSE remains quite high, and the left panel of Figure 1.9 shows that the curve of $V_z(m, N)$ tends to be too low for the higher ranks ($m > 20$).

A more flexible Zipfian smoother has been proposed by Naranan and Balasubrahmanyam (1998),

$$V_{nb}(m, N) = \frac{Ce^{-\mu/m}}{m^\gamma}, \quad (1.14)$$

which shares the term $1/m^\gamma$ with the simple Zipfian power function (1.11), but adds the term $e^{-\mu/m}$ to handle the downward curvature at the head of the frequency spectrum. The right panel of Figure 1.9 plots the fit of this model

to the data. Evaluated in terms of the MSE for the first 40 ranks, 77.37, we observe a substantial improvement in goodness of fit. (For details on how the parameters of (1.14) can be estimated, see section 3.4 in Chapter 3.)

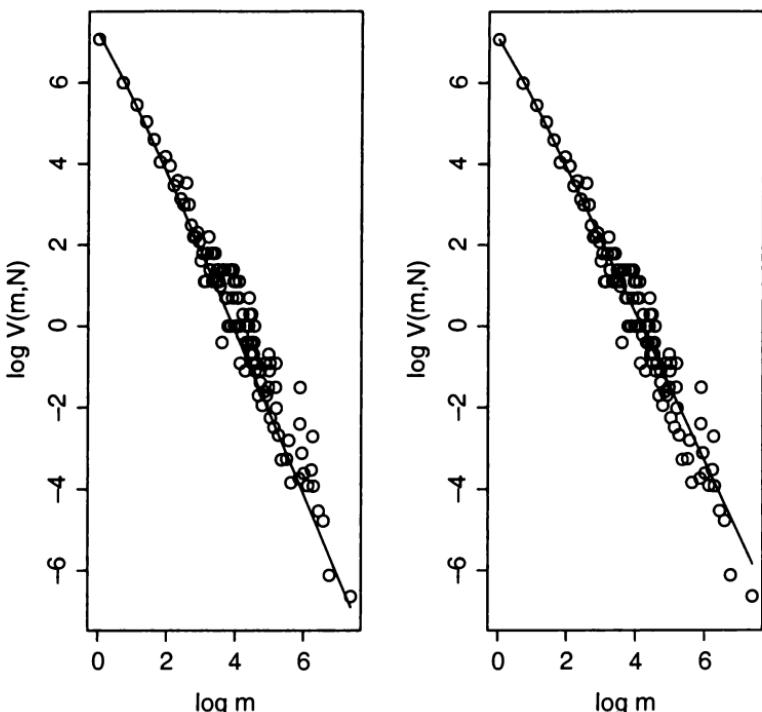


Figure 1.9: The real-valued approximate frequency spectrum of Alice’s Adventures in Wonderland in the double logarithmic plane with a simple Zipfian fit (left panel) and a Naranan-Balasubrahmanyam Zipfian fit (right panel).

Although the Naranan-Balasubrahmanyam Zipfian model generally provides very good fits to empirical spectra, it suffers from the same problem as observed in Figure 1.7 for the parameters of Zipf’s rank-frequency distribution, namely, systematic changes as a function of the sample size N . Figure 1.10 illustrates this systematic variability for a Dutch text, *Max Havelaar*, by Multatuli, the pseudonym of Eduard Douwes Dekker (1820-1887). The words of this text were re-arranged in a random order to eliminate possible effects of non-randomness in word use (see Chapter 5 for detailed discussion of the randomness assumption). The upper left panel plots $V(N)$ (upper circles), $V(1, N)$ (central circles), and $V(2, N)$ (bottom circles) as a function of N at 20 equally-spaced intervals. The remaining panels plot C (upper right), μ (lower left) and γ (lower right) as a function of N . Throughout the text, C appears to increase with N . During the first 8 measurement points, μ increases, after which it becomes more or less stable. From measurement point 9, γ emerges as a decreasing function of N . Thus, at least two parameters have to be adjusted to accommodate a change in sample size.

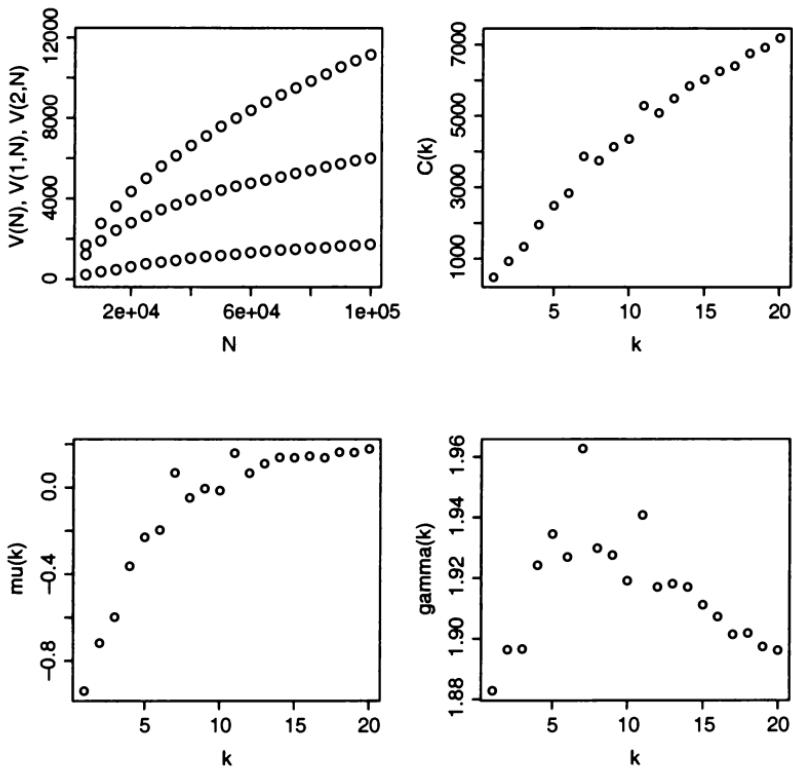


Figure 1.10: The dependency of the three parameters of the Naranan-Balasubrahmanyam Zipfian model as a function of the sample size N in Max Havelaar by Multatuli. The upper right panel plots C as a function of N , the lower left panel μ as a function of N , and the lower right panel γ as a function of N . The upper left panel plots $V(N)$ (upper circles), $V(1,N)$ (central circles), and $V(2,N)$ (lower circles) as a function of N at 20 equally-spaced intervals.

To see why the parameters of Naranan-Balasubrahmanyam Zipfian model change systematically as a function of N , consider the ratio of any two spectrum elements,

$$R(m, n) = V(m, N)/V(n, N).$$

Given Naranan and Balasubrahmanyam's model, we have that

$$\begin{aligned} R_{nb}(m, n) &= \frac{Ce^{-\mu/m}m^{-\gamma}}{Ce^{-\mu/n}n^{-\gamma}} \\ &= e^{-\frac{\mu(n-m)}{mn}} \left(\frac{m}{n}\right)^{-\gamma}. \end{aligned}$$

Note that C disappears in R_{nb} , which leaves us with two parameters to account for the kind of changes in the empirical values of these ratios illustrated

for *Alice's Adventures in Wonderland* and *Max Havelaar* in Figure 1.11. The circles represent $R(2, 1)$, the triangles $R(3, 2)$, and the solid lines the corresponding expected values (using binomial interpolation, see section 2.6). Because the empirical ratios change systematically with N , the parameters μ and γ have to be adjusted continuously.

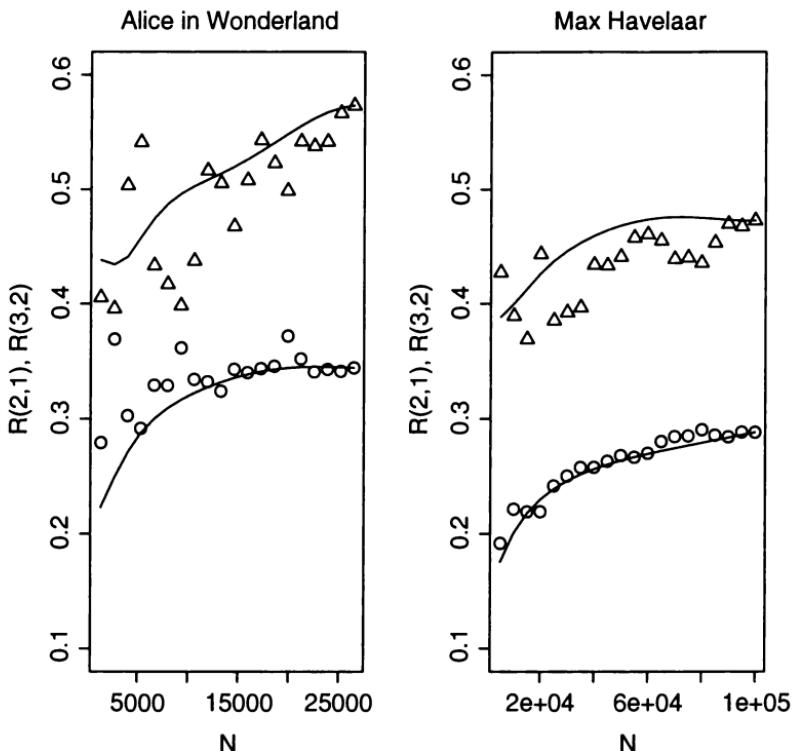


Figure 1.11: The spectrum ratios V_2/V_1 ($R(2, 1)$, circles) and V_3/V_2 ($R(3, 2)$, triangles) for *Alice's Adventures in Wonderland* (left panel) and *Max Havelaar* (right panel) for 20 equally spaced sample sizes N . The solid lines represent the corresponding expected values.

1.4 The quest for characteristic constants

We have seen that lexical measures such as mean frequency and vocabulary size, as well as the parameters of the zeta distribution and the Naranan-Ballasubrahmanyam Zipfian model all depend on the sample size. This state of affairs leads to severe problems when texts of different lengths have to be compared. Not surprisingly, a great many measures have been proposed as text constants, measures that do not vary with the size of the sample.

The oldest of these measures was developed by Yule (1944), in his sem-

inal book, *The Statistical Study of Literary Vocabulary*. In this comprehensive study of the frequency distributions of a great number of different texts, all compiled by hand on individual slips of paper, Yule proposed a quantitative textual measure that, apart from sampling fluctuations, should be independent of sample size. His so-called characteristic constant K ,

$$K = 10000 \frac{\sum_m m^2 V(m, N) - N}{N^2}, \quad (1.15)$$

is a measure of the rate at which words are repeated in a text. To see this, it is convenient to write K as follows:

$$K = 10000 \left\{ \sum_m [V(m, N) \frac{m}{N} \frac{m}{N}] - \frac{1}{N} \right\}.$$

The factor 10000 is a scale factor, introduced only to avoid overly small values of K that might otherwise be difficult to read. The proportion m/N is the sample estimate of the probability of sampling a word with token frequency m . Hence, m^2/N^2 is the probability of sampling such a word twice in a row, assuming that the word probabilities are constant (sampling with replacement).

A closely related measure has been proposed by Simpson (1949):

$$D = \sum_m V(m, N) \frac{m}{N} \cdot \frac{m-1}{N-1}. \quad (1.16)$$

Consider a word ω_i with frequency m in a sampling situation without replacement. The probability that the very first word sampled is precisely ω_i equals m/N . The probability that the next token sampled represents this same type is given by $(m-1)/(N-1)$: the number of remaining tokens of ω_i divided by the total number of remaining tokens. Thus $\frac{m}{N} \frac{m-1}{N-1}$ estimates the likelihood that two tokens of the same type are sampled in succession. The value of D is obtained by summation of this likelihood for all $V(N)$ types.

Both K and D are measures of the repeat rate. In Chapter 2, we shall see that they can also be viewed as weighted average probabilities. Both are heavily influenced by the highest frequency words, for which the probability of repeated use is greatest. The large dots in panels A and B of Figure 1.12 show that in *Alice in Wonderland* K and D reveal highly similar developmental patterns. Apparently, the repeat rate in this text is high in the initial sections, decreases first, and then slowly increases again.

Doubts concerning the stability of K and D for varying sample sizes have led to the formulation of other text constants. Recall that the mean frequency $N/V(N)$ varies with the sample size N . Would it be possible to eliminate this variation by considering simple functions of N and $V(N)$? Guiraud (1954) proposed a measure in which the square root of the sample size replaces the sample size in what is known as the type-token ratio, $V(N)/N$, the inverse of the mean type frequency, as follows:

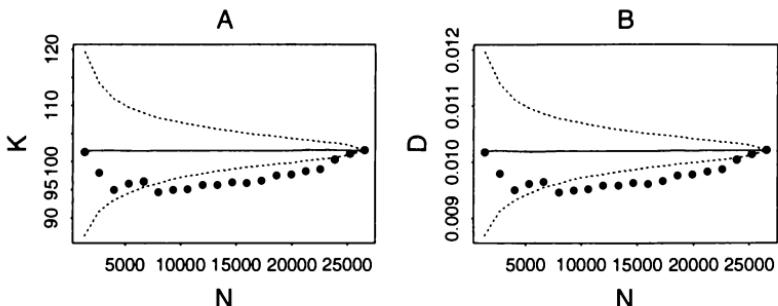


Figure 1.12: The characteristic constants K (panel A) and D (panel B) as a function of the sample size N in *Alice in Wonderland*. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.

and Brunet (1978) suggested a power relation between N and $V(N)$,

$$W = N^{V(N)^{-a}}. \quad (1.18)$$

If R and W are truly constants, then $V(N)$ reduces to very simple functions of N :

$$\begin{aligned} V(N) &= R\sqrt{N}, \\ V(N) &= \left(\frac{\log(W)}{\log(N)} \right)^{-\frac{1}{a}} = (\log_N W)^{-\frac{1}{a}}. \end{aligned}$$

But are R and W truly constant and independent of N ? The large dots in Figure 1.13, which plot the observed values of R and W for 20 equally-spaced values of N in *Alice in Wonderland*, show that this is not the case. In this novel, R and W are increasing functions of the sample size N . Unlike K and D , R and W have no probabilistic interpretation. The value of the parameter a in the expression for W , for instance, has no sensible interpretation and is usually fixed at 0.17, a heuristic value that has been found to produce the desired result of producing a roughly constant relation between N and $V(N)$.

It is easy to see why R and W change with N when we consider a population with a fixed number of types, for instance, a distribution of 10000 tokens sampled from a population with 50 equiprobable types, as shown in Figure 1.14. Each panel displays 40 measurement points that are 250 tokens apart. The upper left panel shows that all 50 types already appear in the sample once 250 tokens have been sampled. The upper right panel shows the linear increase of the average token frequency in the interval $N = [250, 10000]$. The lower left panel shows that R decreases as N is increased. Because V is fixed, the plot effectively shows the function $y = 50x^{-1/2}$. The lower right

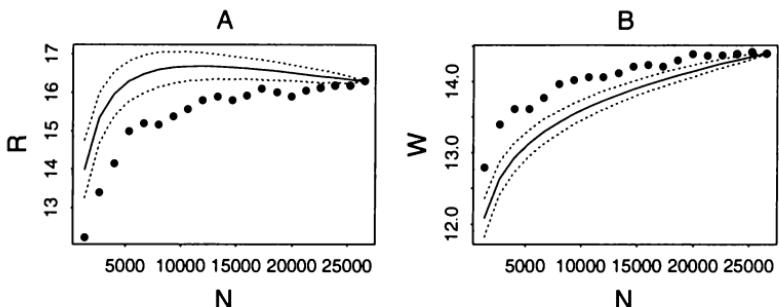


Figure 1.13: The text characteristics R (panel A) and W (panel B) as a function of the sample size N in Alice in Wonderland. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.

panel shows that W increases as a power function, $y = N^{0.514}$, a straight line in the double logarithmic plane. This example shows that once all types have been sampled, the constants systematically change when the sample size is increased. Because actual texts do not use all types that are available in the language, the changing magnitude of V affects the values of R and W . This is most clearly visible in the case of R . The left panel of Figure 1.13 shows that in *Alice's Adventures in Wonderland* R reveals the same downward curvature visible in Figure 1.14, but only after the first quarter of the text has been seen. Even for larger sample sizes, the appearance of new types slows down the rate at which R decreases, and guarantees that its value stays well above 1.0 throughout the text.

The next two measures focus on the low-frequency words in the frequency spectrum. Sichel (1975) observed that the proportion S of hapax legomena $V(2, N)$ of the vocabulary size $V(N)$ is more or less constant:

$$S = V(2, N)/V(N). \quad (1.19)$$

The proportion of hapax legomena on the vocabulary $V(1, N)/V(N)$ plays a key role in a measure proposed by Honoré (1979):

$$H = 100 \frac{\log N}{1 - \frac{V(1, N)}{V(N)}}. \quad (1.20)$$

The idea underlying (1.20) is that $V(1, N)/V(N)$ is a linear function of $\log N$:

$$\frac{V(1, N)}{V(N)} = a - b \log N$$

(see panels A and B of Figure 1.15). Since for $N = 1$ the number of hapaxes is equal to the number of types, a must be equal to 1 and hence $b = 100/H$. It

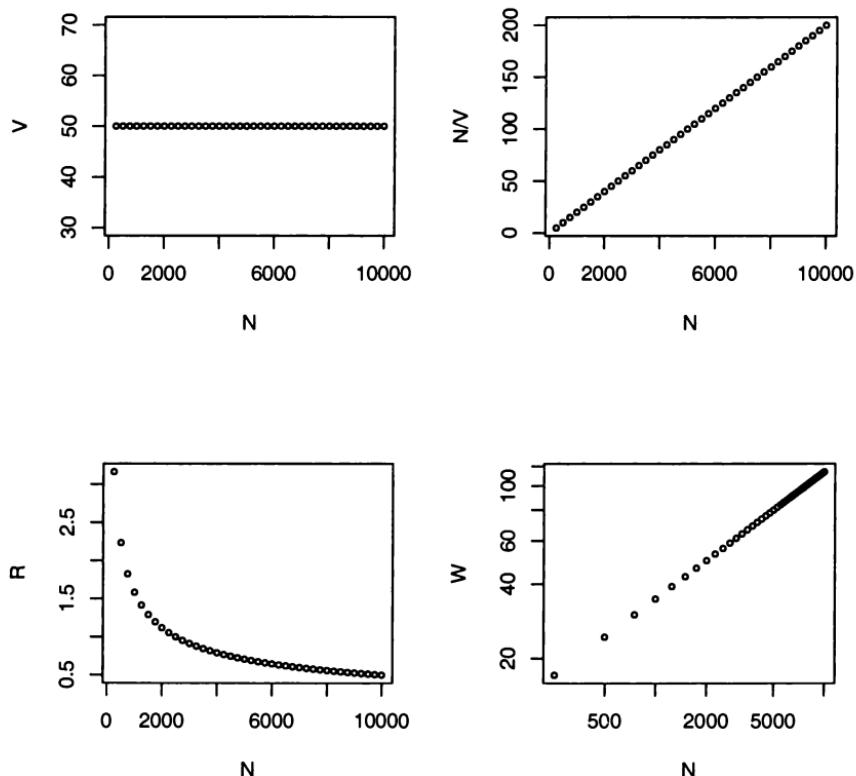


Figure 1.14: V , N/V , R , and W as a function of N in a random sample of 10000 tokens from a population with 50 equiprobable types.

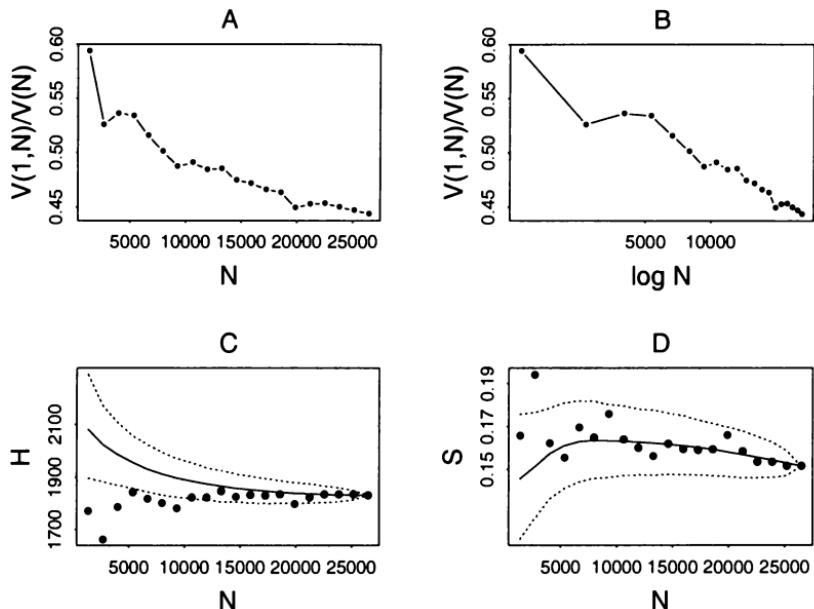


Figure 1.15: The text characteristics H (panel C) and S (panel D) as a function of the sample size N in *Alice in Wonderland*. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs. Panels A and B show the roughly linear dependency of the relative number of hapax legomena on $\log N$ that underlies H .

follows immediately that $H = 100/b$. Honoré does not consider the proportion $V(1, N)/V(N)$ by itself because this proportion generally decreases with increasing sample size (see panel B of Figure 1.15). The large dots in panel C of Figure 1.15 suggests that H is more or less constant for $N > 5000$ and converges rapidly to its final value of 1850. With respect to Sichel's constant S , the large dots in panel D reveal a slightly decreasing pattern with relatively large local fluctuations.

Herdan (1964) proposed a constant that is based on the observation that the growth curve of the vocabulary appears as roughly a straight line in the double logarithmic plane, as shown in panel A of Figure 1.16 for *Alice in Wonderland* by means of large dots. This suggests that a power model for $V(N)$ would be appropriate:

$$V(N) = aN^C.$$

Applying logarithms to both sides, we again obtain the equation for a straight line

$$\log V(N) = \log a + C \log N$$

with $\log a$ as intercept and C as slope. For sample size $N = 1$, $V(N)$ must also equal unity, and since $\log(1) = 0$, we have that $a = 1$. The remaining

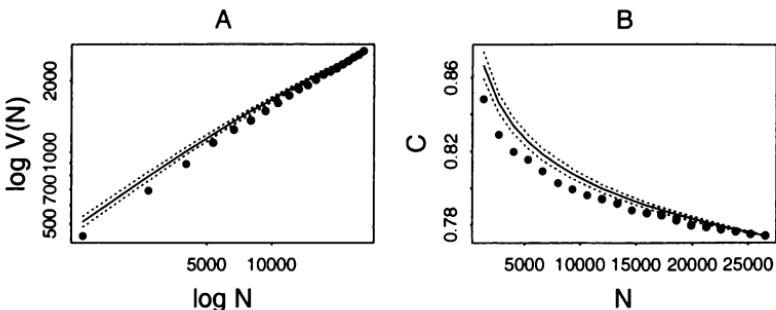


Figure 1.16: Herdan’s ‘law’ applied to *Alice in Wonderland*. Panel A plots $V(N)$ as a function of N in the double logarithmic plane. Panel B plots Herdan’s constant C as a function of the sample size N in *Alice in Wonderland*. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.

parameter C is therefore fully determined by N and $V(N)$:

$$C = \frac{\log V(N)}{\log N}. \quad (1.21)$$

Equation (1.21) is known as Herdan’s law. Herdan’s law is attractive in its simplicity, but on closer inspection C again emerges as highly dependent on the sample size N . The large dots in panel B of Figure 1.16 plot the empirical values of C for 20 equally-spaced intervals. Clearly, C is a decreasing function of N , and cannot be relied on as a characteristic constant that is independent of the sample size.

Of all the measures summarized in Figures 1.12–1.16, Honoré’s H (see Figure 1.15) appears to be least affected by changes in sample size for not too small N . All other measures change substantially as a function of N . The changes we observe in these figures have three possible sources. First, a given measure may be an inherently increasing or decreasing function of N . Second, a measure may in theory be a constant, but discourse structure might lead to a significantly different developmental profile through text time. Finally, all patterns observed in these figures might be due to random variability.

We can evaluate these explanations by considering the values of these measures for random permutations of the order of the words of *Alice in Wonderland*. The solid lines in Figures 1.12–1.16 show the Monte Carlo means of the ‘constants’ for the same 20 measurement points in text time, averaged over 5000 permutation runs. The dotted lines represent the corresponding 95% confidence interval. Figure 1.12 shows that in theory K and D are indeed independent of the sample size. Apart from fluctuations due to random sampling, their expected value does not change systematically with N . At the same time, we find that the empirical values of K and D in *Alice in Wonderland* observed for $N > 8000$ almost up to the end of the story are much lower

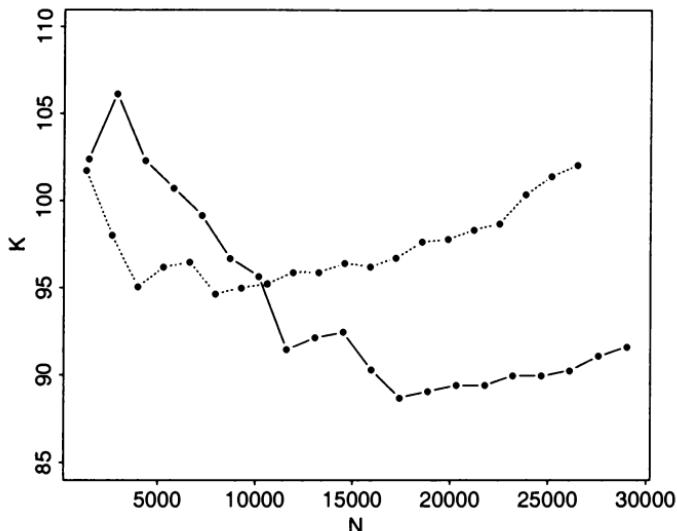


Figure 1.17: Yule's K as a function of N for *Alice in Wonderland* (dotted line) and *Through the looking-glass* (solid line).

than expected under chance conditions ($p < .05$). Random variability by itself does not explain the empirical curves of K and D : We are probably observing an effect of narrative structure.

Unlike K and D , R and W appear to be inherently dependent on N . Figure 1.13 suggests that R is a convex function with a maximum around $N = 10000$, and that W is a monotonically increasing function of N . The empirical values of R are significantly lower than those expected under chance conditions, while for W the observed values are nearly all higher than expected. Figure 1.15 indicates that for *Alice in Wonderland*, H is a decreasing function of N that, compared to R and W , converges relatively early to the value it assumes in the complete text. For $N < 10000$, the observed values of H seriously underestimate the expected values. Turning to S , we find that all but one of the empirical values of S are within the 95% confidence interval of the theoretical values. Nevertheless, S itself is not a constant, but, like R , a convex function of N . Finally, Figure 1.16 shows that C is a decreasing function of N , and its observed values are consistently lower than the theoretical ones.

We have seen that there is substantial variability in the values of all of the measures that have been proposed as characteristic constants. This variability can be inherent to a given measure (e.g., R , W), it can be due to the discourse organization of the text (e.g., K , D), or to a combination of these two factors. This variability imposes severe limits on the usefulness of these 'constants' for comparing texts of different size. Moreover, it is important to evaluate differences **between** two texts against the background of the differences **within** these texts. Figure 1.17 illustrates this point for *Alice in Wonderland* and *Through the looking-glass*. In *Alice in Wonderland*, the values of K range between 94.66 and 106.12, in *Through the looking-glass*, the range of K is from 88.68 to 102.05. Although the value of K for the complete text is lowest for *Through the*

looking-glass, it is in this very same text that K displays its greatest variability. In addition, the two texts reveal reasonably similar values for K for the first ten thousand tokens. Although *Through the looking-glass* has the lower overall repeat rate — note that this ties in nicely with its slightly higher lexical richness — the variability in the repeat rate itself within *Through the looking-glass* is so large that it becomes difficult to argue on the basis of K alone that *Through the looking-glass* differs from *Alice in Wonderland*. Texts are complex entities, and by using simple summary statistics one runs the risk of opting for too coarse a measure to identify similarities and differences between texts.

1.5 The lognormal distribution

An important model for word frequency distributions is the lognormal model. A random variable X is lognormally distributed if $Y = \log(X)$ follows a normal distribution. The lognormal model is sometimes used for skewed distributions with slowly decreasing right tails. Word frequency distributions are heavily skewed in this sense. Herdan (1960) and Carroll (1967) have therefore considered the possibility that word frequencies are lognormally distributed. The lognormal hypothesis is of special interest because many statistical tests presuppose normality. Word frequency distributions are decidedly non-normal. However, if they can be transformed into normal distributions by considering log frequency instead of absolute frequency, then these statistical tests can nevertheless be used after applying a simple logarithmic transformation.

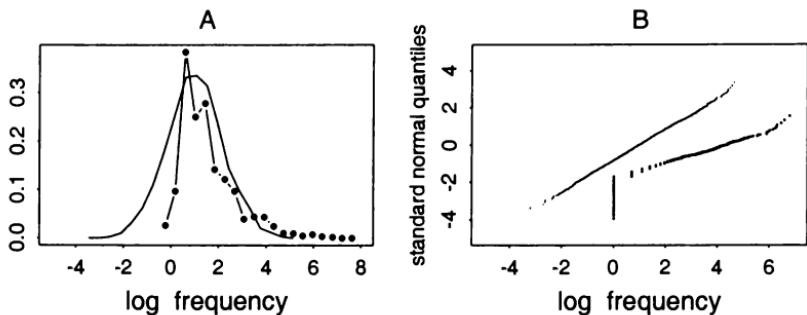


Figure 1.18: The lognormal hypothesis. Panel A shows the estimated probability density function for log frequency in *Alice in Wonderland* (dashed and dotted line) and the estimated density function of a lognormal random variable with the same mean and standard deviation. Panel B plots the corresponding quantiles of the standard normal distribution.

Unfortunately, it is not generally true that logarithmically transformed word frequencies are normally distributed. Figure 1.18 illustrates this point

for *Alice in Wonderland*. Panel A compares the distribution of log frequency in *Alice in Wonderland* (mean log frequency 0.974, variance 1.212, for 2651 word types) with 2651 random numbers from a lognormal distribution with the same logarithmic mean and variance by means of the estimated probability density functions. For the simulated lognormal distribution, represented by a solid line, we find a curve resembling the familiar bell-shaped curve of the normal distribution. The distribution of log frequency in *Alice in Wonderland*, represented by the dashed and dotted line, remains a skewed distribution.

Panel B is the corresponding Normal quantile-quantile plot. The horizontal axis plots log frequency sorted according to increasing frequencies. The vertical axis plots the quantiles of the standard normal, i.e., the values of a standard normal distribution corresponding to a given percentage of the sorted data. (For instance, the quantile for 2.5% of the data has the value of -1.96.) Normally distributed random variables show up as straight lines in Normal quantile-quantile plots, deviations from normality emerge as nonlinearities. In panel B, the upper line represents the simulated lognormal distribution, which, as expected, is a straight line. The log frequencies of *Alice in Wonderland* reveal a different pattern. Word frequencies are integer-valued, hence no log frequencies less than zero are attested. The vertical line at log frequency equals zero represents the hapax legomena, which jointly account for some 4.5% of all tokens. For the dis legomena and higher frequency ranks, the cumulated number of tokens reveals a more linear development.

Except for the lowest frequencies in the spectrum, the lognormal hypothesis seems to be reasonably well supported. This suggests that the problem that we are dealing with is at least in part one of discretization: Word frequency distributions are discrete. Normality, by contrast, presupposes a continuous random variable. Low-probability words either occur, in which case they are very likely to appear among the hapax legomena, or they do not occur. In the latter case, they do not appear in counts with fractional frequencies. Instead, they belong to the $V(0, N)$ unknown words with a frequency of zero.

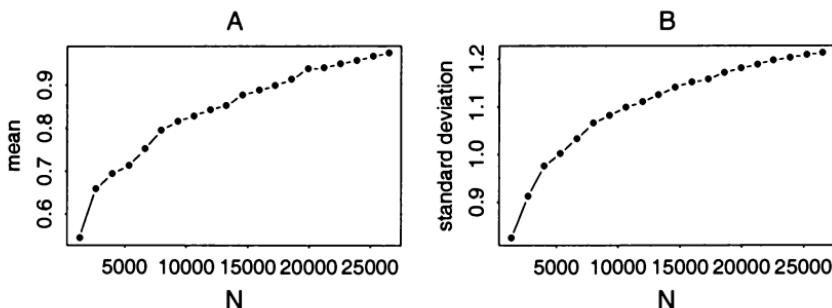


Figure 1.19: The parameters of the lognormal model as a function of the sample size N for *Alice in Wonderland*.

Even if we accept the lognormal model as a continuous approximation for

a discrete distribution, we again run into the problem that the parameters of the lognormal model change when the sample size is changed. Figure 1.19 illustrates this by now familiar phenomenon for *Alice in Wonderland*. Panel A plots the mean log frequency for twenty equally-spaced intervals, and panel B the corresponding standard deviations. Both the mean and the standard deviation appear as increasing functions of N . In chapter 3, we will show how the hypothesis of lognormality can be adjusted to avoid these problems.

1.6 Discussion

The main thrust of this chapter has been to show that for word frequency distributions the sample mean frequency and many other summary measures change in a highly systematic way as a function of the sample size. The parameters of the zeta (Zipf) and lognormal distributions are subject to exactly the same kind of systematic dependency on the sample size.

An additional complicating factor is that words are not randomly distributed in texts. Randomization tests show that the articles *a* and *the* are not uniformly distributed in *Alice in Wonderland*. Non-randomness in word usage similarly affects the measures that have been proposed as invariant with respect to sample size. Consequently, measures such as K and D , which in theory are true constants, nevertheless may reveal significantly non-random developmental profiles.

The prominent role of the sample size in shaping word frequency distributions, combined with the non-random way in which authors use their words in discourse, raises two important issues. First, when comparing texts or corpora, the characteristic constants that have been proposed as independent of the sample size should be interpreted with caution. They may reveal differences between authors, genre, or register, but when substantial differences in sample size are involved, the extent to which their values vary with sample size should be carefully considered. In addition, to gauge the importance of a difference in the value of a text characteristic for two or more texts, one should weigh the intertextual differences with respect to the intratextual variability of the text characteristic. It is only when the intertextual differences are larger than the intratextual differences that one may have some confidence that the differences are reliable.

Second, the law of large numbers does not appear to hold for word frequency distributions. Is the non-randomness in word use illustrated for the articles *the* and *a* in *Alice in Wonderland* to be held responsible? Or are lexical samples, even when encompassing tens of thousands or even millions of words, too small to allow the theoretical probabilities of words to be estimated from their sample relative frequencies? These issues are addressed in detail in the next chapters.

1.7 Bibliographical Comments

An elementary introduction to lexical statistics is Muller (1977). Muller (1979b) provides a useful collection of papers in the French tradition of lexical statistics. Other important studies in this tradition are Guiraud (1954), Brunet (1978), Honoré (1979), and Menard (1983). A more recent textbook is Lebart and Salem (1994).

In the Anglo-Saxon tradition, classic studies are Zipf (1935), Zipf (1949), Yule (1944), Carroll (1967), and Herdan (1960), Herdan (1964), Herdan (1966). Important technical papers are Good (1953), Good and Toulmin (1956), and Efron and Thisted (1976).

In the Eastern-European tradition, Orlov (1983a) and Orlov (1983b) are accessible studies. Guter and Arapov (1983) is a useful collection of studies on Zipf's law. The concept of structural distributions is developed in Khmaladze and Chitashvili (1989) (in Russian), part of which has appeared in English in Khmaladze (1987). A review article in English is Chitashvili and Baayen (1993).

For Monte Carlo methods, see Hammersley and Handscomb (1964) and Meyer (1956). A review of lexical constants is Tweedie and Baayen (1998), for non-randomness in the use of function words, see Damerau (1975).

Important journals are *Computers and the Humanities*, *Literary and Linguistic Computing*, *Journal of Quantitative Linguistics*, *Computational Linguistics*, and the book series *Glottometrika*. A number of important technical papers can be found in *Biometrika*.

1.8 Questions

1. Show, using definition 1.2, that the ratio $N/V(N)$ represents the mean token frequency.
2. How might non-randomness in word usage affect the accuracy of theoretical estimates for $V(N)$?
3. Figure 1.20 plots the relative sample frequencies of *a* and *the* in *Through the looking-glass*. Offer an explanation for the developmental profiles.
4. Interpret the expression $V(0, N)$. What does Figure 1.3 suggest about its magnitude, when we view *Alice in Wonderland* as a sample of Carroll's word use?
5. Figure 1.21 plots the frequency spectrum of *Through the looking-glass* for $N = 14514$ and $N = 29028$. Does the curve with the greater number of hapax legomena ($V(1, N)$) represent the larger sample or the smaller one?
6. Rewrite

$$g(m, N) > 2g(m + 1, N) - g(m + 2, N)$$

in terms of the frequency spectrum $V(m, N)$.

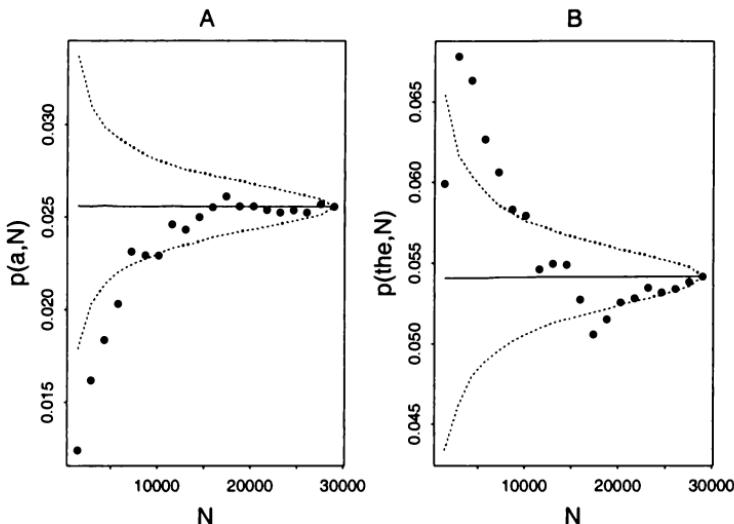


Figure 1.20: The sample relative frequency of the article a $p(a, N)$ (panel A) and the sample relative frequency of the article the , $p(\text{the}, N)$ (panel B) as a function of sample size N . Through the looking-glass, measured at 20 equally-spaced intervals. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on a total of 5000 permutation runs.

7. Figure 1.22 shows the error function of Zipf's zeta function for *Alice in Wonderland* at $N = 13250$, i.e., for each Zipf rank z it plots the difference between the observed frequency $f_z(z, 13250)$ and the expected frequency given by (1.7). What is the source of the striations at the right-hand side of the plot? Comment on the error pattern and its relevance for interpreting the high correlation ($r = -0.986, t_{(1827)} = -250.02, p = 0$) between Zipf rank and frequency.

8. Mandelbrot (1953) enriched Zipf's zeta distribution with a second free parameter b to enhance the model's accuracy for the high-frequency ranks:

$$f_z(z, N) = \frac{C}{(z + b)^a} .$$

Express $V(m, N)$ as a function of m and the parameters of the Zipf-Mandelbrot distribution.

9. Figure 1.7 shows that the intercept is an increasing function of N and that the slope is a decreasing function of N . Offer an explanation for this pattern.
10. Rewrite K (1.15) in terms of the word frequencies $f(i, N)$ instead of in terms of the frequency spectrum $V(m, N)$.
11. The randomized version of *Alice in Wonderland* reveals higher values for C than the original non-randomized version. Is the direction of the dif-

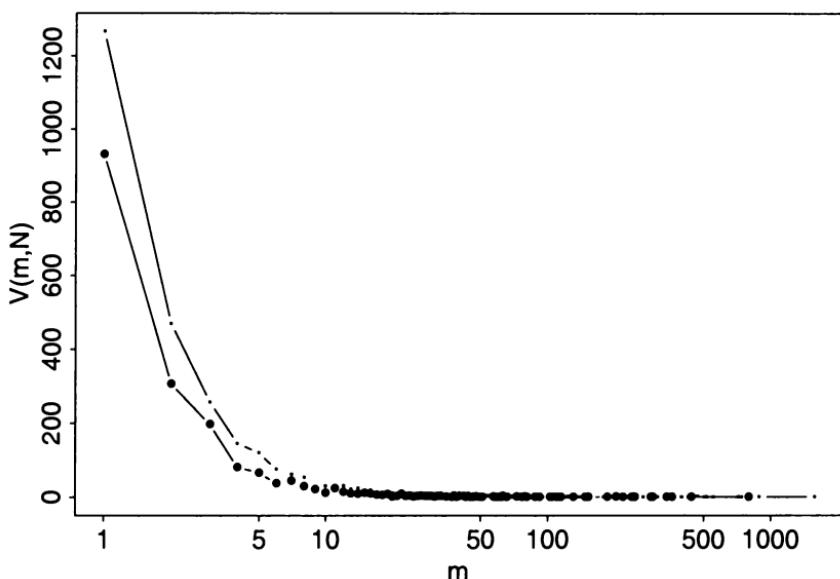


Figure 1.21: *The frequency spectra of Through the looking-glass at the sample sizes $N = 14514$ and $N = 29028$ (m : frequency class; $V(m, N)$: number of types occurring m times).*

ference between the two curves, higher values for the randomized version, what you would expect, or might the randomized version just as well have revealed lower values for C ?

12. Figure 1.19 shows that the rate at which mean and standard deviation change as N is increased decreases. Under what circumstances do you expect the mean and the standard deviation to have a limit for $N \rightarrow \infty$?
13. Discuss the usefulness of the type-token ratio V/N for the evaluation of the lexical richness of texts of different lengths.

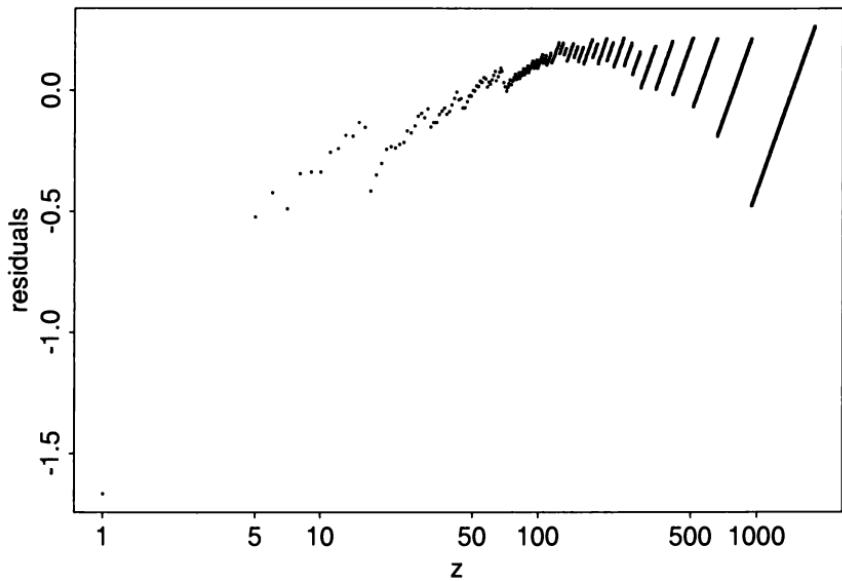


Figure 1.22: *The error function of Zipf's zeta function for Alice in Wonderland at $N = 13250$.*

Chapter 2

Non-parametric models

This chapter presents a range of statistical techniques that are available for the analysis of word frequency distributions. Section 2.1 introduces some basic probabilistic concepts. Section 2.2 discusses the urn model, according to which word use is viewed as random selection from a population with fixed probabilities for words to occur. The binomial model and the Poisson approximation to the binomial model are defined here. Section 2.3 is concerned with the structural type distribution, which allows us to restate the Poisson model in integral form. Section 2.4 introduces the concept of the LNRE zone, the range of sample sizes where the sample relative frequencies are not good estimates of the corresponding population probabilities. The next section (2.5) focuses on the Good-Turing estimates, which adjust sample relative frequencies for the non-negligible frequency weight of the unseen words. Methods for calculating the frequency spectrum for any sample size given the frequency spectrum for a given sample size are presented in sections 2.6 and 3.2.

2.1 Basic concepts

SUMMARY This section briefly presents the definitions of the expectation, variance, and covariance of a random variable, in combination with a summary of some basic properties of these operators. The Bernouilli distribution is also introduced, and the distinction between permutations and combinations is outlined.

In this section a summary review is presented of the main definitions and basic properties that we will need in the course of this chapter. For detailed discussion and proofs, the reader is referred to textbooks such as, e.g., Ross (1988).

Definition 2.1 The **expectation** $E[X]$ of a random variable X is given by

$$E[X] = \sum_x x \Pr(X = x).$$

The expectation of X , $E[X]$, is the sum of the values that X can assume, weighted by their probabilities. If X is the outcome of a toss with a fair die, then each outcome $1, 2, \dots, 6$ is equally likely with probability $1/6$, hence $E[X] = 21/6 = 3.5$. If the die is loaded such that the outcome 6 is twice as likely as any of the other outcomes, then $E[X] = \frac{1}{7}15 + \frac{2}{7}6 = 3.86$. The expectation of X can be viewed as its probability-weighted mean. A fundamental property of the expectation operator is that it can be interchanged with the summation operator:

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i]. \quad (2.1)$$

In other words, the expectation of a sum is the sum of the expectations. Other useful properties are

$$E[aX + b] = aE[X] + b, \quad (2.2)$$

and for independent random variable X and Y ,

$$E[XY] = E[X]E[Y]. \quad (2.3)$$

Definition 2.2 The variance $\text{VAR}[X]$ of a random variable X with expectation $E[X]$, is defined as

$$\begin{aligned} \text{VAR}[X] &= E[(X - E[X])^2] \\ &= E[X^2] - (E[X])^2. \end{aligned}$$

The variance $\text{VAR}[X]$ of X is a measure of variation, or spread, of the values around the mean. Important properties of the variance are

$$\text{VAR}[aX + b] = a^2\text{VAR}[X] \quad (2.4)$$

and, for independent random variables,

$$\text{VAR}\left[\sum_i^n X_i\right] = \sum_{i=1}^n \text{VAR}[X_i]. \quad (2.5)$$

Definition 2.3 The covariance $\text{COV}[X, Y]$ of the random variable X and Y with expectations $E[X]$ and $E[Y]$ is defined by

$$\begin{aligned} \text{COV}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

It is a measure for the extent to which X and Y are independent. When X and Y are independent, $\text{COV}[XY] = 0$. In general,

$$\text{COV}[a + bX, c + dY] = bd\text{COV}[X, Y], \quad (2.6)$$

and

$$\text{COV}\left[\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right] = \sum_{i=1}^n \sum_{j=1}^m \text{COV}[X_i, Y_j]. \quad (2.7)$$

A random variable X is said to be **Bernoulli-distributed**, when it assumes only two values, success ($X = 1$) and failure ($X = 0$):

Definition 2.4 A random variable X is Bernouilli-distributed when

$$\Pr(X = 1) = p$$

and

$$\Pr(X = 0) = 1 - p$$

The expectation of a Bernouilli random variable is equal to the probability of success: $E[X] = 1 \cdot \Pr(X = 1) + 0 \cdot \Pr(X = 0) = \Pr(X = 1)$. The variance of X is also easily obtained:

$$\begin{aligned}\text{VAR}[X] &= E[X^2] - (E[X])^2 \\ &= E[X] - (E[X])^2 \\ &= p - p^2 \\ &= p(1 - p).\end{aligned}$$

It is often convenient to write Bernouilli random variables by means of the indicator operator I :

$$\begin{aligned}E[I_{[X=x]}] &= 1 \cdot \Pr(X = x) + 0 \cdot \Pr(X \neq x) \\ &= \Pr(X = x).\end{aligned}$$

Finally, we need to distinguish between permutations and combinations. The notion **permutation** pertains to the number of different orders in which n objects can be arranged.

Definition 2.5 The number of different orders for n objects is written as $n!$ and is given by

$$n! = n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1.$$

The value of $0!$ equals unity.

The number of different ways to select k objects out of n objects is referred to as the number of **combinations**, and is defined as

Definition 2.6 The number of possible combinations of n objects taken k at a time, ' n choose k ', for $k \leq n$, equals

$$\binom{n}{k} = \frac{n!}{(n - k)!k!}.$$

For instance, consider the number of different groups of 3 objects out of 7 objects, labeled A, B, C, ..., G. The first object can be chosen in 7 ways, the second in 6 ways, and the third in 5. Hence we have a total of $7 \cdot 6 \cdot 5 = 210$ different groups. However, many of these groups are identical except for the order of the elements. There are 6 groups with the objects A, B, and C, for instance (ABC, ACB, BAC, BCA, CAB, CBA), and we do not want to count these 6 groups as different. To obtain the number of groups that differ only

with respect to the elements in the group without considering their order, we divide the total number of groups by the number of (irrelevant) orders for each group of 3 objects. Hence, the number of combinations of 3 out of 7, $\binom{7}{3}$, is

$$\frac{7!}{4! \cdot 3!} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35.$$

2.2 The Urn model

okay, an urn of words, rather than bag

SUMMARY *The urn model takes the use of words to be similar to the random selection of marbles from an urn. We derive two important and surprisingly powerful models from the urn model: the binomial model and the Poisson model. The Poisson model is an approximation of the binomial model that is especially useful because it is mathematically and computationally more tractable than the binomial model. Both models provide us with expressions for the expected vocabulary size $E[V(N)]$ and the expected frequency spectrum $E[V(m, N)]$.*

The urn model for word frequency distributions compares the use of words to the sampling of marbles from an urn. Consider an urn containing marbles of various colors. Each color corresponds with a marble type. A particular color may be extremely rare, or it may be represented by a great many individual marbles, the marble tokens. We randomly draw N marbles from the urn, assuming that the outcome of a given trial is completely independent from the outcome of any other trial. We then inspect our N marbles, and count the number of different colors represented in our sample, $V(N)$, or the number of colors that appear m times in the sample, $V(m, N)$. It will be clear that the assumptions underlying this simple probabilistic experiment, randomness and independence, are violated for language. We do not choose our words randomly from our vocabulary, and the kind of word used at one point influences the kind of word used next. In Chapter 5 we will study in some detail how violations of these assumptions affect the accuracy of the predictions based on the urn model. In this chapter, we will investigate how much progress can be made in understanding the dynamics of word frequency distributions on the basis of the very simple assumptions of the urn model.

In what follows, I will assume that the urn contains S different word types $\omega_i, i = 1, 2, 3, \dots, S$. With each word ω_i we associate a population probability $\pi_i, i = 1, 2, 3, \dots, S$ of being sampled. These probabilities can be thought of as the number of marbles with color i in the urn, divided by the total number of marbles in the urn. Sampling a word consists of randomly selecting a word token from the urn, inspecting its color, and returning it to the urn. Because we sample with replacement, the probabilities of the words do not change over time.

What is the probability $\Pr(f(i, N) = m)$ that word ω_i appears m times in a sample of N tokens? We can consider our sample of N tokens as a sequence of N trials with m successes (ω_i was drawn) and $N - m$ failures (ω_i was not drawn). The probability of a particular sequence of m successes and $N - m$ failures equals $\pi_i^m(1 - \pi_i)^{N-m}$. How many such sequences are there? This

question can be rephrased as: In how many ways can we select m trials from N trials to be labeled as a success? The number of ways in which we can select m objects from N objects is the number of combinations of N objects taken m at a time: $\binom{N}{m}$. Hence $\Pr(f(i, N) = m)$ equals

$$\Pr(f(i, N) = m) = \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m}. \quad (2.8)$$

For instance, suppose we have 5 trials, 3 success, and 2 failures, then there are $5!/(3!2!) = 10$ ways of labeling 3 out of the 5 trials as a success:

$$\begin{array}{lll} \text{s s s f f} & \text{s f s s f} & \text{f s s s f} \\ \text{s s f s f} & \text{s f s f s} & \text{f s s f s} \\ \text{s s f f s} & \text{s f f s s} & \text{f s f s s} \\ & & \text{f f s s s} \end{array}$$

Each labeling has probability $\pi^3(1 - \pi)^{5-3}$, hence

$$\Pr(f(i, 5) = 3) = 10\pi_i^3(1 - \pi_i)^2.$$

In general, we have the following definition of a binomially distributed random variable:

Definition 2.7 A random variable X is **binomially** (n, p) -distributed when

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}. \quad (2.9)$$

Given the urn model, the frequency of a word ω_i with probability π_i in a sample of N tokens is binomially (N, π_i) distributed.

What is the expected frequency of ω_i in the sample?

$$\begin{aligned} E[f(i, N)] &= \sum_{m=0}^N m \Pr(f(i, N) = m) \\ &= \sum_{m=0}^N m \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m} \\ &= \sum_{m=1}^N \frac{mN!}{(N-m)!m!} \pi_i^m (1 - \pi_i)^{N-m} \\ &= N\pi_i \sum_{m=1}^N \frac{(N-1)!}{(N-m)!(m-1)!} \pi_i^{m-1} (1 - \pi_i)^{N-m} \\ &\stackrel{k=m-1}{=} N\pi_i \sum_{k=0}^{N-1} \binom{N-1}{k} \pi_i^k (1 - \pi_i)^{N-1-k} \\ &= N\pi_i [\pi_i + (1 - \pi_i)]^{N-1} \\ &= N\pi_i. \end{aligned} \quad (2.10)$$

In the second but last step, we let $k = m - 1$. In the one but last step, we use the **binomial theorem**

Definition 2.8 Binomial Theorem:

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$$

(for a proof, see Ross, p. 9). Note that the binomial theorem has a probabilistic interpretation:

$$(\pi + (1 - \pi))^N = 1 = \sum_{m=0}^N \binom{N}{m} \pi^m (1 - \pi)^{N-m}$$

is the sum of the probabilities for all possible number of successes (ranging from no successes to only success), which sum up to unity.

Recall that we have defined $p(i, N)$ as the sample relative frequency of ω_i . Corresponding to the sample frequency and the sample probability, we have the expected frequency and the population probability. Thus, the following relations should be carefully distinguished:

$$\begin{aligned} E[f(i, N)] &= N\pi_i \\ f(i, N) &= Np(i, N). \end{aligned}$$

We are now in the position to obtain expressions for the spectrum elements, the expected number of word types with frequency m in a sample of N ,

$$\begin{aligned} E[V(m, N)] &= E\left[\sum_{i=1}^S I_{[f(i, N)=m]}\right] \\ &= \sum_{i=1}^S E[I_{[f(i, N)=m]}] \\ &= \sum_{i=1}^S \Pr(f(i, N) = m) \\ &= \sum_{i=1}^S \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m}, \end{aligned} \tag{2.11}$$

as well as for the number of different types in the sample:

$$\begin{aligned} E[V(N)] &= E\left[\sum_{m=1}^N V(m, N)\right] \\ &= \sum_{m=1}^N E[V(m, N)] \\ &= \sum_{m=1}^N \sum_{i=1}^S \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m}. \end{aligned} \tag{2.12}$$

This expression can be simplified by changing perspective. Instead of focusing on the words that appear in the sample, we have a closer look at the words that do not appear in the sample. Consider the case that word ω_i does not occur at all in our sample of N tokens. The probability that this happens equals $(1 - \pi_i)^N$, the probability of N failures. The complement of this probability, the probability that ω_i occurs at least once in the sample, is

$$1 - (1 - \pi_i)^N.$$

The number of types in the sample is the number of types that occur at least once. Hence

$$\begin{aligned} E[V(N)] &= E\left[\sum_{i=1}^S I_{[f(i,N) > 0]}\right] \\ &= \sum_{i=1}^S E[I_{[f(i,N) > 0]}] \\ &= \sum_{i=1}^S \Pr(I_{[f(i,N) > 0]}) \\ &= \sum_{i=1}^S (1 - (1 - \pi_i)^N) \\ &= S - \sum_{i=1}^S (1 - \pi_i)^N. \end{aligned} \tag{2.13}$$

The penultimate step gives the final step of (2.12), an equality which we can also state as the difference between the number of types in the population S and the expected number of types that do not occur in the sample.

The expressions for $E[V(N)]$ and $E[V(m, N)]$ can be considerably simplified by making use of the **Poisson** approximation to the binomial distribution. For large N and small p , a binomially (N, p) -distributed random variable is well approximated by a Poisson (λ) -distributed random variable with $\lambda = Np$:

Definition 2.9 A random variable X is Poisson- λ distributed when

$$\Pr(X = k) = \frac{(\lambda)^k}{k!} e^{-\lambda}. \tag{2.14}$$

Since text samples are generally large and word probabilities very small, the Poisson approximation is quite good. Given the approximation

$$\binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m} \approx \frac{(N\pi_i)^m}{m!} e^{-N\pi_i},$$

we can rewrite $E[V(N)]$ and $E[V(m, N)]$ as follows:

$$E[V(m, N)] = \sum_{i=1}^S \frac{(N\pi_i)^m}{m!} e^{-N\pi_i} \tag{2.15}$$

$$\mathbb{E}[V(N)] = \sum_{i=1}^S (1 - e^{-N\pi_i}). \quad (2.16)$$

Note that $e^{-N\pi_i}$ is the Poisson probability that word ω_i does not appear in a sample of N tokens.

Recall that in Chapter 1, the lexical constants D (Simpson) and K (Yule) were described as weighted average probabilities. We are now in the position to see how this interpretation arises. We focus on the interpretation of D , leaving the interpretation of K as an exercise. Using the fact that

$$\binom{N}{m} = \frac{N(N-1)}{m(m-1)} \binom{N-2}{m-2},$$

it is easy to see that $\mathbb{E}[D] = \sum_i \pi_i^2$:

$$\begin{aligned} E \left[\sum_{m=1}^N V(m, N) \frac{m}{N} \frac{m-1}{N-1} \right] &= \sum_{m=0}^N E[V(m, N)] \frac{m}{N} \frac{m-1}{N-1} \\ &= \sum_{m=0}^N \sum_{i=1}^S \left\{ \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m} \right\} \frac{m}{N} \frac{m-1}{N-1} \\ &= \sum_{i=1}^S \sum_{m=0}^N \frac{m}{N} \frac{m-1}{N-1} \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m} \\ &= \sum_{i=1}^S \sum_{m=2}^N \binom{N-2}{m-2} \pi_i^{m-2} (1 - \pi_i)^{N-m} \pi_i^2 \\ &= \sum_{i=1}^S \sum_{m'=0}^{N'} \left\{ \binom{N'}{m'} \pi_i^{m'} (1 - \pi_i)^{N'-m'} \right\} \pi_i^2 \\ &= \sum_{i=1}^S \pi_i^2 \end{aligned} \quad (2.17)$$

Thus, D is an unbiased estimate of the summed squared probabilities. The summed squared probabilities can be viewed as the expected probability of a random variable P assuming the values $\pi_i, i = 1, 2, \dots, S$ with probabilities $\pi_i, i = 1, 2, \dots, S$:

$$\mathbb{E}[P] = \sum_{i=1}^S \pi_i \pi_i.$$

In this interpretation, the random variable P describes a property of word tokens, namely, the fact that they represent a type with a probability π_i . Comparing word tokens with marbles in an urn, this amounts to the situation in which marbles have probabilities inscribed on them. The probability of sampling a token having probability π_i is itself equal to π_i , the proportion of the tokens representing type i in the urn.

2.3 The Structural Type Distribution

SUMMARY The structural type distribution allows us to rewrite the Poisson model in integral form. This leads to mathematically more convenient expressions. For instance, we can now easily obtain an expression for the rate at which the vocabulary size increases. The structural type distribution will play a key role in the discussion of the LNRE models in Chapter 3.

In section 1.2, the empirical structural type distribution $g(m, N)$ was introduced:

$$g(m, N) = \sum_{i=1}^{V(N)} I_{[f(i, N) \geq m]}.$$

Analogous to the empirical structural type distribution we have the **structural type distribution** $G(\pi)$, which pertains to the population probabilities $\pi_i, i = 1, 2, \dots, S$.

Definition 2.10 The structural type distribution $G(\pi) = \sum_{i=1}^S I_{[\pi_i \geq \pi]}$: the number of types in the population with probability greater or equal to π .

The structural type distribution is a step function with jumps at those points that correspond to probabilities which characterize at least one word type. Denoting the number of types with probability π by $V(\pi)$,

$$V(\pi) = \sum_{i=1}^S I_{[\pi_i = \pi]},$$

we can index those probabilities π for which $V(\pi) > 0$, such that $\pi_j < \pi_{j+1}$. The jumps at the probabilities $\pi_j, j = 1, 2, \dots, \kappa, \kappa \leq S$ are given by

$$\Delta G(\pi_j) = G(\pi_j) - G(\pi_{j+1}),$$

as illustrated in Figure 2.1. In other words, $\Delta G(\pi_j)$ denotes the number of words in the population with probability π_j .

We can now restate the expressions for the expected frequency spectrum and the expected vocabulary size in integral form:

$$\begin{aligned} E[V(m, N)] &= \sum_{i=1}^S \frac{(N\pi_i)^m}{m!} e^{-N\pi_i} \\ &= \sum_{j=1}^{\kappa} \frac{(N\pi_j)^m}{m!} e^{-N\pi_j} \Delta G(\pi_j) \\ &= \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} dG(\pi), \\ E[V(N)] &= \sum_i^S (1 - e^{-N\pi_i}) \end{aligned} \tag{2.18}$$

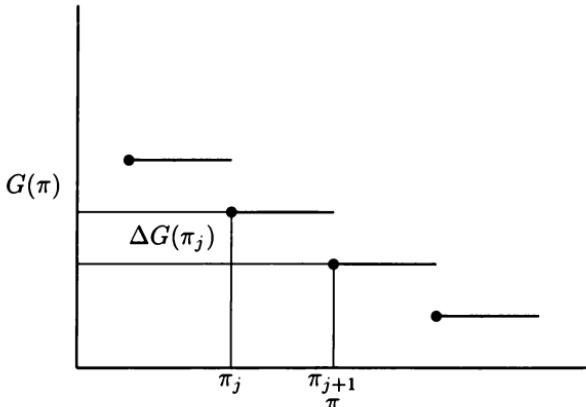


Figure 2.1: *The structural type distribution.* $\Delta G(\pi_j) = G(\pi_j) - G(\pi_{j+1})$ is the number of words in the population with probability π_j .

$$\begin{aligned} &= \sum_{j=1}^{\kappa} (1 - e^{-N\pi_j}) \Delta G(\pi_j) \\ &= \int_0^\infty (1 - e^{-N\pi}) dG(\pi). \end{aligned} \quad (2.19)$$

Note that $dG(\pi) = 0$ except for the intervals $dG(\pi)$ around the π_j , where $dG(\pi) = \Delta G(\pi)$. Also note that we have changed from sums to integrals. Formally, this transition from sums to integrals can be motivated in terms of Stieltjes integrals. Without going into the technical details, the idea is that, because $G(\pi)$ is a (non-increasing) step function defined on the interval $[0, 1]$ with jumps at points $(\pi_1, \pi_2, \pi_3, \dots, \pi_\kappa)$,

$$\Delta(G\pi_i) = G(\pi_i) - G(\pi_i+),$$

with $G(\pi_i+) = \lim_{\pi \downarrow \pi_i} G(\pi)$, we can write sums of the form

$$s = \sum_{i=1}^{\kappa} h(\pi_i) \Delta G(\pi_i),$$

with h some function of p , in the form

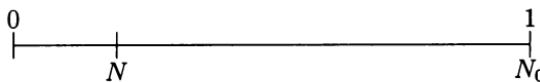
$$\begin{aligned} s &= \sum_{\pi} h(\pi) \Delta G(\pi) \\ &= \int_0^1 h(p) dG(p). \end{aligned}$$

In the case of the expression for $E[V(N)]$, for instance, $h(\pi) = 1 - e^{-N\pi}$.

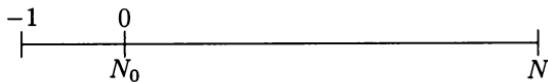
The reader may have observed that the integration intervals range from zero to infinity, even though probabilities are normalized for the interval $[0, 1]$.

Formally, integration over $[0, 1]$ leads to identical results. I have opted for the current notation with the larger integration interval to maintain the parallelism with a slightly different notational convention that is often found in the literature in which integration over $[0, \infty)$ is essential. Since the parameter of a Poisson- λ distributed random variable can be interpreted as the rate at which a particular event occurs, the question arises what unit of time is appropriate for measuring this rate. For expressions such as (2.18) and (2.19), a single word token is the basic unit of 'text time'. However, it is equally well possible to choose N_0 tokens as the basic unit of measurement. Defining $\lambda_i = N_0\pi_i$ is equivalent to saying that $N_0\pi_i$ occurrences of ω_i appear in a time interval of N_0 tokens. Conversely, defining $\lambda_i = \pi_i$ amounts to saying that π_i tokens of ω_i appear on average in a time unit of one token.

If we take N_0 tokens as our basic unit of measurement, the 'text time' t equals 1 for $N = N_0$. For $N < N_0$, $t = N/N_0 < 1$:



For $N > N_0$, it is convenient to map the interval $[0, N_0]$ onto the interval $[-1, 0]$,



and again we write $t = N/N_0$. The expressions for the expected vocabulary size and the spectrum elements can now be reformulated as follows:

$$\begin{aligned} E[V(m, t)] &= \int_0^\infty \frac{(t\lambda)^m}{m!} e^{-t\lambda} dG(\lambda) \\ E[V(t)] &= \int_0^\infty (1 - e^{-t\lambda}) dG(\lambda) \end{aligned}$$

In this form, the integration interval $[0, \infty)$ is crucial as $\lambda = N_0\pi$ now denotes a frequency: the frequency with which a word occurs in a sample of N_0 tokens. However, since a single word token is the most obvious unit of measurement, I will formulate theoretical expressions in terms of N and π rather than in terms of t and λ .

The first derivative of $E[V(N)]$ expresses the rate at which the vocabulary increases after sampling N tokens (see Figure 2.2):

$$\begin{aligned} \frac{d}{dN} E[V(N)] &= \frac{d}{dN} \int_0^\infty (1 - e^{-N\pi}) dG(\pi) \\ &= \int_0^\infty -\pi \cdot -e^{-N\pi} dG(\pi) \\ &= \frac{1}{N} \int_0^\infty N\pi e^{-N\pi} dG(\pi) \\ &= \frac{E[V(1, N)]}{N}. \end{aligned} \tag{2.20}$$

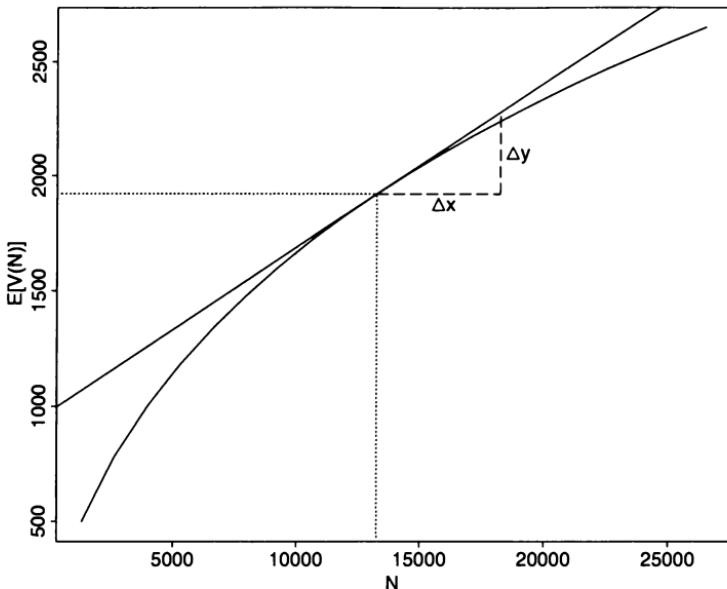


Figure 2.2: *The growth rate of the vocabulary as the first derivative of the growth curve $E[V(N)]$: $\Delta y/\Delta x = E[V(1, N)]/N$.*

For a given sample, the growth rate of the vocabulary can be estimated by the ratio of hapax legomena to the number of tokens, $V(1, N)/N$. This leads to the following definition:

Definition 2.11 The growth rate $\mathcal{P}(N)$ of the vocabulary, the rate at which the vocabulary size increases with increasing sample size, is defined by

$$\mathcal{P}(N) = \frac{d}{dN} E[V(N)] = \frac{E[V(1, N)]}{N}.$$

The growth rates of the individual spectral elements $E[V(m, N)]$ can likewise be obtained by differentiating in the variable N .

$$\begin{aligned}
 \frac{d}{dN} E[V(m, N)] &= \\
 &= \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} dG(\pi) \frac{d}{dN} \\
 &= \int_0^\infty \left(mN^{m-1} \frac{\pi^m}{m!} e^{-N\pi} - \frac{(N\pi)^m}{m!} \pi e^{-N\pi} \right) dG(\pi) \\
 &= \frac{1}{N} \int_0^\infty \left(m \frac{(N\pi)^m}{m!} e^{-N\pi} - (m+1) \frac{(N\pi)^{m+1}}{(m+1)!} e^{-N\pi} \right) dG(\pi) \\
 &= \frac{1}{N} (mE[V(m, N)] - (m+1)E[V(m+1, N)]). \tag{2.21}
 \end{aligned}$$

Denoting the sample sizes at which the spectral elements $E[V(m, N)]$ reach their maximum by N_m^* , we observe the following relations. At the moment

in sampling time that the number of hapax legomena reaches its maximum ($N = N_1^*$), the number of hapax legomena is exactly twice the number of dis legomena. The number of hapax legomena reaches its maximum when

$$\frac{d}{dN} E[V(1, N)] = 0,$$

which, by (2.21), implies that

$$\frac{E[V(1, N)] - 2E[V(2, N)]}{N} = 0, \quad (2.22)$$

hence $E[V(1, N)] = 2E[V(2, N)]$ at N_1^* . The same line of reasoning leads to the general relation

$$E[V(m, N_m^*)] = \frac{m+1}{m} E[V(m+1, N_m^*)]. \quad (2.23)$$

2.4 The LNRE zone

SUMMARY This section introduces the fundamental concept of the LNRE zone. Word frequency distributions differ from more familiar distributions with respect to the number of low-probability, rare, word types (events). Whereas there are relatively few low-probability events in familiar distributions such as the normal distribution, word frequency distributions are characterized by very Large Numbers of Rare Events. The general shape of the curves of the vocabulary size and the spectrum elements differ markedly from the corresponding curves for non-LNRE type count distributions. For non-LNRE distributions, the spectrum elements reach their maximum value very quickly, for LNRE distributions, the spectrum elements generally do not do so in normal sample size ranges. The LNRE zone can be described as the range of sample sizes for which it is clear from the shape of the spectral curves that we have only just begun to sample the types available in the population. Even large corpora with tens of millions of words are located in the LNRE zone.

The main part of this section is relatively non-technical, and should be generally accessible.

In Chapter 1, we have seen that the sample mean and a great many other text characteristics vary systematically with the sample size. How can this property of word frequency distributions be understood against the background of the urn model?

According to the law of large numbers, we have that for any probability distribution

$$(\pi_i, 1 \leq i \leq S)$$

with a finite vocabulary size S the relative sample frequencies will converge to the population probabilities

$$\lim_{N \rightarrow \infty} p(i, N) = \pi_i.$$

in probability as $N \rightarrow \infty$. As a simple consequence,

$$\lim_{N \rightarrow \infty} V(m, N) = 0$$

for all m . As we continue sampling, we reach a point where all types in the population have been sampled at least once: $V(0, N) = 0$. As we continue sampling tokens, we reach a point where each word type has been sampled at least twice: $V(1, N) = 0$. This process can be continued indefinitely. Step by step, the successive spectral frequencies $m = 3, 4, \dots$ will no longer be represented by any type in the sample. Figure 2.3 illustrates this phenomenon

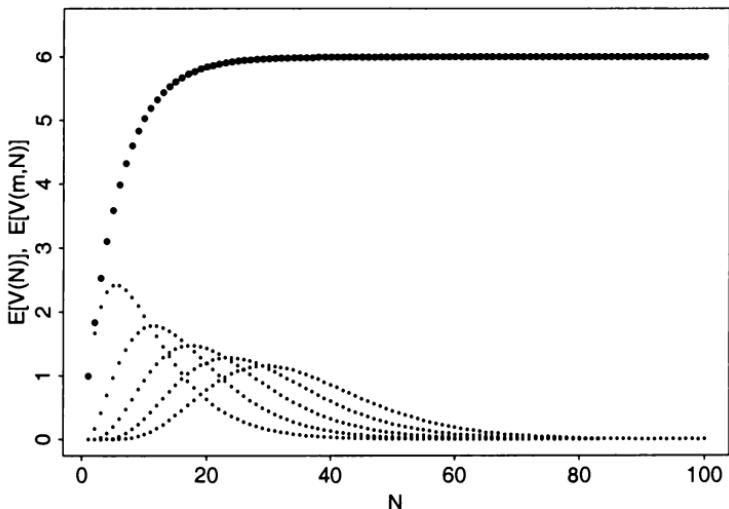


Figure 2.3: $E[V(N)]$ and $E[V(m, N)]$ for $m = 1, 2, \dots, 5$ for 100 throws with a fair die.

for an experiment with a fair die, in which we regard the 6 possible outcomes as 6 different word types. The large dots plot the expected vocabulary size $E[V(N)]$ as a function of N , the number of throws. After some 40 throws, all possible outcomes of throwing a die are expected to have appeared at least once. Of course, there is always a minute probability that one of the outcomes has not yet appeared: $E[V(N)]$ never reaches its horizontal asymptote. But after a small number of throws, $E[V(N)]$ is already very close to its asymptotic value, 6.

The curves represented by small dots in Figure 2.3 represent the spectrum elements $E[V(m, N)]$ for $m = 1, 2, \dots, 5$. The curve that is the first to reach

its maximum (at five throws of the die) is that of $E[V(1, N)]$. The next curve to reach its maximum is $E[V(2, N)]$, and the last one to do so is the curve of $E[V(5, N)]$. Note that after 40 throws, $E[V(1, 40)]$ is approximately zero. This is the point at which $E[V(N)]$ is virtually identical to its asymptotic value.

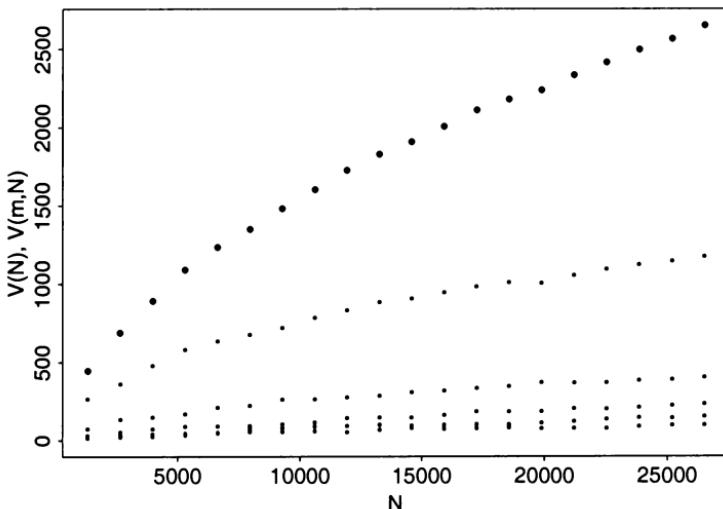


Figure 2.4: The empirical vocabulary size $V(N)$ (large dots) and the first 5 empirical spectral elements $V(m, N)$ (small dots) for *Alice in Wonderland*, sampled at 20 equal-spaced intervals.

Figure 2.3 is a simple illustration of the kind of pattern that is typically observed for a great many random variables. But for word frequency distributions, the usual pattern is qualitatively different. Figure 2.4 plots $V(N)$ and $V(m, N)$ ($m = 1, 2, \dots, 5$) for *Alice in Wonderland*. The large dots represent the empirical growth curve of the vocabulary, the small dots the individual spectrum elements. Figure 2.4 does not reveal a horizontal asymptote for $E[V(N)]$. The rate at which the vocabulary increases is clearly decreasing, but it is unclear whether the curve will ultimately reach a horizontal asymptote or an asymptote with a positive slope. A horizontal asymptote implies that the number of types in the population (S) is finite, a non-horizontal asymptote implies that S is infinite. Figure 2.4 also illustrates that the spectral elements $E[V(m, N)]$ do not reach their maximum value within the text itself. For $N \leq 26505$, they all emerge as increasing functions of N .

For all values of N , the number of hapax legomena exceeds the number of dis legomena, the number of dis legomena exceeds that of the tris legomena,

etc. If we assume, for the time being, that the number of types in the population S is finite, then the pattern for *Alice in Wonderland* in Figure 2.4 resembles the very beginning of the pattern in Figure 2.3. For the first 5 tokens of Figure 2.3, $E[V(N)]$ is increasing, and none of the spectral elements $E[V(m, N)]$ has yet reached its maximum. From this point of view, *Alice in Wonderland* is a sample of Carroll's word use that is much too small to allow us to see the asymptotic behavior of the vocabulary size and the spectrum elements.

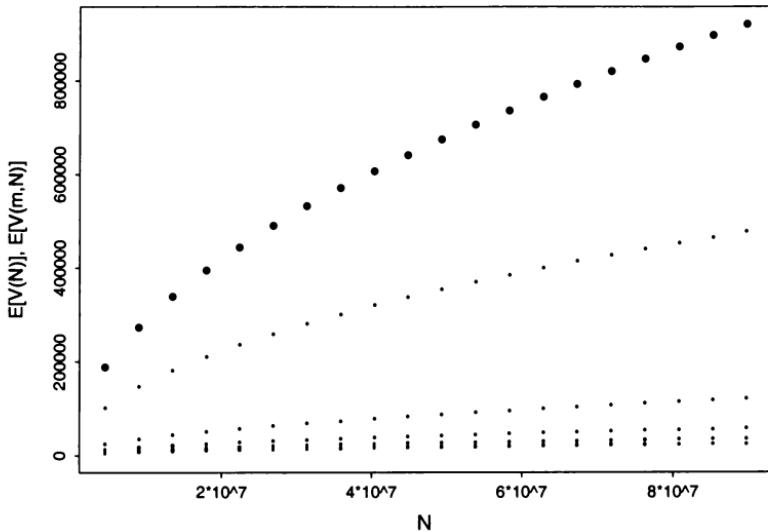


Figure 2.5: The expected vocabulary size $E[V(N)]$ (large dots) and the first 5 expected spectral elements $E[V(m, N)]$ (small dots) for the written texts in the British National Corpus, sampled at 20 equal-spaced intervals.

Figure 2.5 shows that a very similar pattern emerges when we exchange *Alice in Wonderland* ($N = 26,505$) for the written texts in the British National Corpus ($N = 89,739,914$).¹ Even after sampling nearly 90 million words, we are still in the sampling zone where the vocabulary size and the spectrum elements are increasing.

Word frequency distributions are **Large Number of Rare Event (LNRE) distributions**, distributions characterized by the presence of large numbers of

¹Figure 2.5 plots the expected values, as there is no natural order to the texts in a corpus. The technique used to obtain these expected values will be introduced in section 2.6. The definition of word type underlying these counts is the CLAWS word, disambiguated for part of speech.

words with very low probabilities of occurrence. In the British National Corpus, for instance, more than half of all types have a sample relative frequency of .00000001. Due to the large numbers of rare words, the sample size N has to be extremely large for the asymptotic properties of the distribution to emerge. In practice, almost all samples of words are located in the **LNRE zone**, the range of sample sizes where the vocabulary size is still increasing, and where the numbers of hapax legomena, dis legomena, etc., are non-negligible.

The concept of LNRE distributions was developed by Khmaladze (1987). He introduces two formal definitions for the LNRE concept. Let

$$\mathcal{V}(N) = (f(1, N), f(2, N), \dots, f(S, N))$$

denote the vector of frequencies of the word types as realized in a sample of N tokens. By increasing N , we obtain a sequence of such vectors,

$$\{\mathcal{V}(N)\}, \quad N = 1, 2, 3, \dots$$

Initially, most word frequencies will be zero, but as we increase the sample size, more and more words will appear with non-zero frequencies. The first formal definition states that

Definition 2.12 A sequence $\{\mathcal{V}(N)\}$ is a sequence with a large number of rare events if

$$\lim_{N \rightarrow \infty} \frac{E[V(1, N)]}{N} > 0.$$

According to this definition, we have an LNRE distribution in case the growth rate of the vocabulary remains greater than zero even when N is increased indefinitely. It is easy to see that for a fixed finite number of types S and fixed probabilities $\pi_i, i = 1, 2, \dots, S$,

$$\lim_{N \rightarrow \infty} f(i, N) = \infty,$$

hence

$$\lim_{N \rightarrow \infty} E[V(1, N)] = 0$$

and

$$\lim_{N \rightarrow \infty} E[V(N)] = S.$$

In these circumstances, the vocabulary growth rate will become zero. However, even for distributions with infinite S , a non-zero growth rate for $N \rightarrow \infty$ is not guaranteed, as N may grow too rapidly with respect to the expected number of hapax legomena. The second definition developed by Khmaladze is more lenient:

Definition 2.13 A sequence $\{\mathcal{V}(N)\}$ is a sequence with a large number of rare events if

$$\lim_{N \rightarrow \infty} \frac{E[V(1, N)]}{E[V(N)]} > 0 \text{ and } \lim_{N \rightarrow \infty} E[V(N)] = \infty.$$

This definition explicitly requires S to be infinite while requiring that the hapax legomena constitute a non-negligible proportion of the vocabulary. Note that a distribution satisfying the first definition will also satisfy the second definition, but that the reverse does not hold. For the statistical justification of these definitions and their theoretical implications, the reader is referred to Khmaladze (1987).

It is useful to distinguish between a central and a late LNRE zone (Chitashvili and Baayen, 1993). The **central LNRE zone** is the interval $(0, N_1^*)$, the range of sample sizes where the expected number of hapax legomena is increasing. The **late LNRE zone** is the interval (N_1^*, N_1^ϵ) , with N_1^ϵ the sample size at which $E[V(1, N)]$ differs from zero by less than some small number ϵ . The central LNRE zone in Figure 2.3 is the interval $(0, 5)$. After 5 throws of a fair die, the expected number of values that have appeared only once in the sample no longer increases. The late LNRE zone is the interval $(6, 48)$ for $\epsilon = 0.01$. Following the LNRE zone we have the interval (N_1^ϵ, ∞) in which, for practical purposes, the vocabulary size has reached its asymptotic limit. Both *Alice in Wonderland* and the set of written texts in the British National Corpus are situated in the central LNRE zone.

Another perspective on LNRE distributions is to ask under what circumstances we can convince ourselves that the empirical relative sample frequencies $p(i, N)$ are close enough to the theoretical probabilities π_i to allow us to replace theoretical expectations by empirical ones. If the sample is located outside of the LNRE zone, we can approximate the expected vocabulary size and the expected spectrum elements by the expressions

$$\begin{aligned}\hat{E}[V(m, N)] &= \sum_{i=1}^{V(N)} \frac{(p(i, N)N)^m}{m!} e^{-p(i, N)N} \\ &= \sum_{i=1}^{V(N)} \frac{f(i, N)^m}{m!} e^{-f(i, N)}\end{aligned}\tag{2.24}$$

$$\begin{aligned}\hat{E}[V(N)] &= \sum_{i=1}^{V(N)} (1 - e^{-p(i, N)N}) \\ &= \sum_{i=1}^{V(N)} (1 - e^{-f(i, N)})\end{aligned}\tag{2.25}$$

instead of using (2.15) and (2.16). To gauge whether the approximation is reliable, we may use the **coefficient of loss** C_L , the proportion of types that we lose by using the sample relative frequencies instead of the population probabilities:

$$\begin{aligned}C_L &= \frac{V(N) - \hat{E}[V(N)]}{V(N)} \\ &= \frac{V(N) - \sum_{i=1}^{V(N)} (1 - e^{-f(i, N)})}{V(N)}\end{aligned}$$

$$= \frac{\sum_{m \geq 1} V(m, N) e^{-m}}{V(N)}. \quad (2.26)$$

For our example of the ‘vocabulary’ of outcomes of throwing a die, the first value of N outside the late LNRE zone is 49. For this sample size we find, using the binomial model instead of the Poisson approximation, that $C_L = 0.0005$. By contrast, for the complete sample of *Alice in Wonderland* ($N = 26,505$), $C_L = 0.19$, and for the British National Corpus ($N = 89,739,914$) we find that $C_L = 0.21$. For both samples, the observed vocabulary is underestimated by roughly 20% when the population probabilities are replaced by the sample relative frequencies. Crucially, this underestimation is not due to the use of the Poisson approximation to the binomial probabilities. If we use the expression

$$C_L = \frac{\sum_{m \geq 1} V(m, N) (1 - \frac{m}{N})^N}{V(N)}$$

instead of (2.26) to obtain the coefficient of loss for *Alice in Wonderland*, the result is identical up to four decimal digits (0.189492 for the Poisson model, 0.189486 for the binomial model).

Summing up, word frequency distributions, even when obtained for samples of millions of words, are generally located in the LNRE zone, the sampling zone where sample relative frequencies cannot be used to obtain the expected values of the vocabulary size and the spectrum elements. There are methods that avoid using the relative sample frequencies as estimates of the population probabilities when calculating these expected values, but before introducing these, we first consider a technique for improving the sample relative frequencies as estimates of probabilities.

2.5 Good-Turing estimates

SUMMARY Word frequency distributions generally have a large growth rate even at the full sample size, implying that there are many more types to be sampled if more word tokens are added to the sample. This renders the use of sample relative frequencies as estimates of population probabilities problematic. Sample relative frequencies add up to unity. As estimates of population probabilities, they do not leave probability mass for unseen word types. However, the LNRE property of word frequency distributions shows that there are many unseen types. The probability mass of these unseen types is non-negligible. This section introduces the Good-Turing estimates, which adjust the sample relative frequencies, freeing probability mass for the unseen types. The probability mass of the unseen types for a given sample size will be shown to equal the growth rate of the vocabulary at that sample size.

If the growth rate of the vocabulary

$$\mathcal{P}(N) = \frac{d}{dN} E[V(N)] = \frac{E[V(1, N)]}{N}$$

as derived in (2.20) is greater than zero, the number of types in the sample does not exhaust the number of different types in the population. Consequently, the

relative sample frequencies $f(i, N)/N$ overestimate the population probabilities π_i . The sample relative frequencies sum up to unity,

$$\sum_{i=1}^{V(N)} \frac{f(i, N)}{N} = 1,$$

and do not leave any probability for the types that have not appeared in the sample but which, when $P(N) > 0$, nevertheless exist. In this section, we consider a technique that allows us to adjust sample relative frequencies for the joint probability of the unseen types.

This technique was developed by Good (1953), who in turn attributes it to Turing. Hence, the adjusted estimates are commonly referred to as Good-Turing estimates. If $f(i, N) = m$, then the Good-Turing estimate $f^*(i, N) = m^*$ of $f(i, N)$ is

$$f^*(i, N) = m^* \approx (m + 1) \frac{E[V(m + 1, N)]}{E[V(m, N)]}. \quad (2.27)$$

The following proof (presented here informally) is due to Church, Gale, and Kruskal (Church and Gale 1991).

Consider the situation in which we choose one particular word type from the S words in the population, and assume that each word type is equally probable to be selected from the list of word types. If we let W denote a random variable ranging over the indices $1, 2, \dots, S$, then

$$\Pr(W = w) = \frac{1}{S}. \quad (2.28)$$

In addition, independently and simultaneously, we randomly select two samples of N word tokens each. Let $f_1(W, N)$ denote the frequency of the chosen word ω_W in the first sample, and let $f_2(W, N)$ be its frequency in the second sample. If ω_W has occurred m times in the first sample, then its expected frequency in the second sample of the same size is given by

$$m^* = \frac{m + 1}{1 + 1/N} \frac{E[V(m + 1, N + 1)]}{E[V(m, N)]}. \quad (2.29)$$

Note that (2.27) follows from (2.29) when we allow ourselves to assume that $1/N$ is negligible, and that $E[V(m + 1, N + 1)] \approx E[V(m + 1, N)]$.

In order to prove (2.29), we need to introduce the concept of conditional expectations (see, e.g., Ross, 1976, 283-290). Let $p_X(x) = \Pr(X = x)$, let $p_Y(y) = \Pr(Y = y)$, and let $p(x, y) = \Pr(X = x, Y = y)$, the probability that $X = x$ and $Y = y$ simultaneously. The probability of X assuming the value x given that Y has the value y is the conditional probability

$$\Pr(X = x|Y = y) = \frac{p(x, y)}{p_Y(y)},$$

and the corresponding expectation is defined as

$$E[X|Y = y] = \sum_x x \Pr(X = x|Y = y).$$

The conditional expectation given $Y = y$ can be conceptualized as being an ordinary expectation with the set of possible outcomes reduced to those for which $Y = y$. Let $E[X|Y]$ denote that function of the random variable Y that assumes the value $E[X|Y = y]$ when $Y = y$. Since $E[X|Y]$ is itself a random variable, we can consider its expectation, which turns out to be identical to that of X :

$$\begin{aligned}
 E[E[X|Y]] &= \sum_y E[X|Y = y] \Pr(Y = y) \\
 &= \sum_y \sum_x x \Pr(X = x|Y = y) \Pr(Y = y) \\
 &= \sum_y \sum_x x \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} \Pr(Y = y) \\
 &= \sum_y \sum_x x \Pr(X = x, Y = y) \\
 &= \sum_x x \sum_y \Pr(X = x, Y = y) \\
 &= \sum_x x \Pr(X = x) \\
 &= E[X].
 \end{aligned} \tag{2.30}$$

Given (2.30), we can express the expected frequency $E[f(W, N)]$ of word type ω_W in the following form:

$$\begin{aligned}
 E[f(W, N)] &= E[E[f(W, N)|W = w]] \\
 &= E[E[f(w, N)]] \\
 &= E[N\pi_w] \\
 &= NE[\pi_w] \\
 &= N \sum_{i=1}^S \pi_i \Pr(W = i).
 \end{aligned} \tag{2.31}$$

This result carries over to expectations conditioned on some event Y :

$$E[f(W, N)|Y] = N \sum_{i=1}^S \pi_i \Pr(W = i|Y).$$

Next, consider the probability that ω_W has index w given that it occurs m times in the first sample:

$$\Pr(W = w | f_1(W, N) = m) = \frac{\Pr(W = w, f_1(W, N) = m)}{\Pr(f_1(W, N) = m)}. \tag{2.32}$$

Since the choice of w and the choice of the first sample are independent, we can rewrite the numerator as follows:

$$\Pr(W = w, f_1(W, N) = m) = \Pr(W = w, f_1(w, N) = m)$$

$$\begin{aligned}
&= \Pr(W = w) \Pr(f_1(w, N) = m) \\
&= \frac{1}{S} \binom{N}{m} \pi_w^m (1 - \pi_w)^{N-m}.
\end{aligned} \tag{2.33}$$

The probability in the denominator of (2.32) concerns the event that the frequency of our chosen word happens to be m . This situation can in principle arise for any of our word types:

$$\{f_1(W, N) = m\} = \bigcup_{i=1}^S \{W = i, f_1(W, N) = m\}.$$

Hence,

$$\begin{aligned}
\Pr(f_1(W, N) = m) &= \sum_{i=1}^S \Pr(W = i, f_1(W, N) = m) \\
&= \sum_{i=1}^S \Pr(W = i) \Pr(f_1(i, N) = m) \\
&= \sum_{i=1}^S \frac{1}{S} \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m},
\end{aligned}$$

and consequently,

$$\begin{aligned}
\Pr(W = w | f_1(W, N) = m) &= \frac{\frac{1}{S} \binom{N}{m} \pi_w^m (1 - \pi_w)^{N-m}}{\sum_{i=1}^S \frac{1}{S} \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m}} \\
&= \frac{\pi_w^m (1 - \pi_w)^{N-m}}{\sum_{i=1}^S \pi_i^m (1 - \pi_i)^{N-m}}.
\end{aligned} \tag{2.34}$$

We are now in the position to prove (2.29):

$$\begin{aligned}
m^* &= \mathbb{E}[f_2(W, N) | f_1(W, N) = m] \\
&= N \sum_{i=1}^S \pi_i \Pr(W = i | f_1(W, N) = m) \quad \text{by (2.31)} \\
&= N \sum_{i=1}^S \pi_i \frac{\pi_i^m (1 - \pi_i)^{N-m}}{\sum_{j=1}^S \pi_j^m (1 - \pi_j)^{N-m}} \quad \text{by (2.34)} \\
&= N \frac{\sum_{i=1}^S \pi_i^{m+1} (1 - \pi_i)^{N-m}}{\sum_{j=1}^S \pi_j^m (1 - \pi_j)^{N-m}} \\
&= N \frac{\sum_{i=1}^S \binom{N+1}{m+1} \pi_i^{m+1} (1 - \pi_i)^{N-m} / \binom{N+1}{m+1}}{\sum_{j=1}^S \binom{N}{m} \pi_j^m (1 - \pi_j)^{N-m} / \binom{N}{m}} \\
&= N \frac{\binom{N}{m} \mathbb{E}[V(m+1, N+1)]}{\binom{N+1}{m+1} \mathbb{E}[V(m, N)]} \\
&= \frac{m+1}{1 + 1/N} \frac{\mathbb{E}[V(m+1, N+1)]}{\mathbb{E}[V(m, N)]}, \quad \text{(2.35)}
\end{aligned}$$

which was to be shown.

Given the Good-Turing estimate m^* , we can obtain adjusted estimates of the population probability of a word ω_i with frequency $f(i, N) = m$ by replacing its relative sample frequency $p_{i,N} = f(i, N)/N$ by the Good-Turing probability $p^*(i, N)$:

$$p^*(i, N) = \frac{f^*(i, N)}{N} = \frac{m^*}{N} \approx \frac{(m+1)\text{E}[V(m+1, N)]}{N\text{E}[V(m, N)]} \quad (2.36)$$

$$\approx \frac{(m+1)V(m+1, N)}{NV(m, N)}. \quad (2.37)$$

Ideally, the expected values of the spectrum elements are to be used, but for small m and smoothly decreasing spectrum elements the empirical values $V(m, N)$ might also be used.

Recall that $\sum_m mV(m, N) = N$. Interestingly,

$$\begin{aligned} \sum_{m=1} m^* \text{E}[V(m, N)] &= \sum_{m=1} (m+1)\text{E}[V(m+1, N)] \\ &= \sum_{n=1} \{n\text{E}[V(n, N)]\} - \text{E}[V(1, N)] \\ &= N - \text{E}[V(1, N)], \end{aligned} \quad (2.38)$$

which implies that the total probability of all types in the sample is equal to

$$1 - \frac{\text{E}[V(1, N)]}{N}.$$

In other words, the unseen types jointly have a probability of $\text{E}[V(1, N)]/N$, which is exactly the growth rate of the vocabulary $\mathcal{P}(N)$ (2.20).

We will illustrate the behavior of m^* using the newspaper corpus of *The Independent* collected by Renouf and her colleagues (see, e.g., Renouf, 1993). Eight independent samples of one million words each were extracted from this corpus, ranging over newspapers from 1989 to newspapers from 1996. The left panel of Figure 2.6 plots the frequency ranks m and the real-valued approximate spectrum elements $V_r(m, N)$ for the sample from 1989, together with the Naranan-Balasubrahmanyam Zipfian spectrum fit. Note that the variance of the spectrum elements is greatest for the intermediate ranks m .

Using the Naranan-Balasubrahmanyam Zipfian smoothed spectrum elements as point of departure for calculating m^* , we obtain the solid line shown in the central panel of Figure 2.6. In addition to the expected value of m^* , this panel also shows the 95% confidence interval for $m^*, \hat{m}^* \pm 1.96\hat{\sigma}_{m^*}$, with

$$\hat{\sigma}_{m^*} = \sqrt{(m+1)^2 \frac{V(m+1, N)}{V(m, N)^2} \left(1 + \frac{V(m+1, N)}{V(m, N)}\right)} \quad (2.39)$$

(see Church and Gale, 1991). The points in this panel, as well as those in the right panel, represent the joint observations from 8 disjunct subcorpora of 1 million words each from *The Independent*. In terms of log units, the largest adjustments when going from m to m^* occur for the smallest values of m . As for

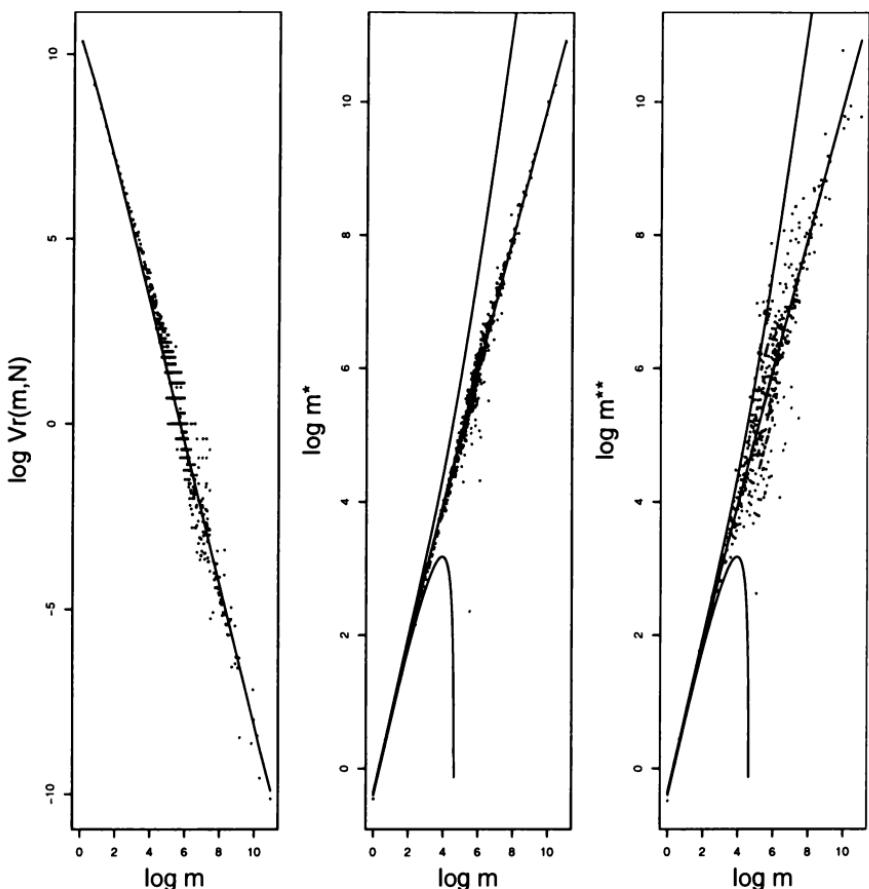


Figure 2.6: Estimating m^* for a corpus of the British newspaper The Independent. Left panel: The real-valued approximate spectrum elements $V_r(m, N)$ in the double logarithmic plane with the Naranan-Balasubrahmanyam Zipfian fit (solid line) for a subcorpus of 1 million words; Central panel: m^* as a function of m for the observations from eight separate newspaper corpora of 1 million words from The Independent, with the Good-Turing 95% confidence interval and the smoothed values derived from the Naranan-Balasubrahmanyam Zipfian fit shown in the left panel; Right panel: m^{**} as a function of m for the observations from eight separate newspaper corpora of 1 million words from The Independent, the Good-Turing 95% confidence interval, and the smoothed values derived from the Naranan-Balasubrahmanyam Zipfian fit shown in the left panel.

the preceding panel, we observe the greatest variance for the intermediate frequency ranges. Note that with only 7 additional 1 million word samples, the variation in the value of m^* is much smaller than the 95% confidence interval, as expected.

The expected values of m^* for the highest-frequency ranks are quite well predicted from the Naranan-Balasubrahmanyam Zipfian fit. It should be kept in mind, however, that for the highest-frequency ranks, the difference between $\log(m)$ and $\log(m^*)$ becomes vanishingly small. For instance, the highest-frequency rank, $m = 56038$, is adjusted to $m^* = 56037.098$, a difference of 0.000016 in log units. By contrast, the first rank, $m = 1$, is adjusted to $m^* = 0.685$, in log units a difference of 0.378.

The third panel of Figure 2.6 illustrates a curious property of an estimate of m^* that is not based on smoothed values such as provided by the Naranan-Balasubrahmanyam Zipfian model but on the real-valued approximate spectrum elements including the values that can be derived for the ranks for which $m = 0$, using (1.13), the average of its surrounding nonzero spectrum elements (see section 1.3):

$$V_r(m_0, N) = \frac{V_r(m_p, N) + V_r(m_f, N)}{m_f - m_p}.$$

Let m^{**} denote the Good-Turing estimate obtained by applying (2.27) to the real-valued spectrum $V_r(m, N)$ without smoothing. What panel three of Figure 2.6 shows is that the values of m^{**} have an upper bound that is reasonably well described by the upper 95% confidence interval derived from the smoothed fit. This suggests that the irregularities in the frequency spectrum may shed light on the variability of m^* . If this turns out to be a general pattern, we may use the lower bound of the m^{**} values as a perhaps more realistic replacement for the lower 95% confidence interval derived from the smoothed values, which is unrestricted for $m > 50$.

2.6 Interpolation and Extrapolation

SUMMARY *The general expressions for the vocabulary size and the spectrum elements require knowledge of the population probabilities, which are unknown. The Good-Turing estimates cannot be used to estimate these probabilities, because they themselves require knowledge of the spectrum elements. To avoid this mutual dependency, it is useful to condition on a given sample size. Given the vocabulary size and the spectrum elements at this sample size, we can work backwards to smaller sample sizes (interpolation) and forwards to larger sample sizes (extrapolation). Unfortunately, extrapolation to substantially larger sample sizes breaks down for technical reasons.*

In the previous sections, we have seen that population probabilities should not be replaced by sample relative frequencies when calculating the expectations of the number of types $E[V(N)]$ or the spectrum elements $E[V(m, N)]$.

This leaves us with the problem how to apply theoretical expressions such as

$$E[V(m, N)] = \sum_{i=1}^S \frac{(N\pi_i)^m}{m!} e^{-N\pi_i}$$

in practice. We cannot use the Good-Turing estimates to enhance our estimates of the population probabilities π_i , as these estimates themselves require prior knowledge of $E[V(m, N)]$. Moreover, a given sample provides information only for the $V(N)$ observed types, whereas we also need to know the probabilities of the $S - V(N)$ unseen types.

In this section, I will show how these problems can be avoided by conditioning on the frequency spectrum at a given sample size, say N_0 , and to work from there to other sample sizes N greater or smaller than N_0 . I discuss interpolation to sample sizes $N < N_0$ in section 2.6.1. In section 2.6.2, I show that the expressions for interpolation generalize for extrapolation to $N > N_0$.

2.6.1 Interpolation

Assume that the $f(i, N_0)$ tokens of word type ω_i are randomly distributed over the N_0 tokens of a given text. We divide this text into two parts, P_1 with $N < N_0$ tokens, and P_2 with $N_0 - N$ tokens. The probability that a particular token of ω_i occurs in P_1 is $\frac{N}{N_0}$. The number of times that ω_i occurs in P_1 is a binomially distributed random variable with parameters $f(i, N_0)$ and $\frac{N}{N_0}$. To obtain m tokens of ω_i in a sample of N tokens, we must draw m tokens of the $f(i, N_0)$ tokens in the complete text of N_0 tokens. The probability of m successes and $f(i, N_0) - m$ failures is

$$\left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{f(i, N_0) - m}$$

As there are $\binom{f(i, N_0)}{m}$ ways of choosing m tokens from $f(i, N_0)$ tokens, we have that

$$\Pr(f_{N_0}(i, N) = m) = \binom{f(i, N_0)}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{f(i, N_0) - m} \quad (2.40)$$

We write $f_{N_0}(i, N)$ instead of $f(i, N)$ to make explicit that we are conditioning on the frequency of ω_i in the larger sample of size N_0 .

Expressions for $E[V_{N_0}(N)]$ and $E[V_{N_0}(m, N)]$, the conditional vocabulary size and the conditional spectrum elements for a sample of size N given the frequency spectrum for N_0 tokens, can be derived from (2.40):

$$\begin{aligned} E[V_{N_0}(m, N)] &= E\left[\sum_{i=1}^{V(N_0)} I_{[f_{N_0}(i, N)=m]}\right] \\ &= \sum_{i=1}^{V(N_0)} E[I_{[f_{N_0}(i, N)=m]}] \end{aligned} \quad (2.41)$$

$$\begin{aligned}
&= \sum_{i=1}^{V(N_0)} \Pr(f_{N_0}(i, N) = m) \\
&= \sum_{i=1}^{V(N_0)} \binom{f(i, N_0)}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{f(i, N_0) - m} \\
&= \sum_{k \geq m} V(k, N_0) \binom{k}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{k-m},
\end{aligned}$$

and similarly,

$$\begin{aligned}
E[V_{N_0}(N)] &= E\left[\sum_{i=1}^{V(N_0)} I_{[f_{N_0}(i, N) > 0]}\right] \\
&= \sum_{i=1}^{V(N_0)} E[I_{[f_{N_0}(i, N) > 0]}] \\
&= \sum_{i=1}^{V(N_0)} \Pr(f_{N_0}(i, N) > 0) \\
&= \sum_{i=1}^{V(N_0)} \left(1 - \left(1 - \frac{N}{N_0}\right)^{f(i, N_0)}\right)
\end{aligned} \tag{2.42}$$

$$\begin{aligned}
&= \sum_{m=1}^{N_0} V(m, N_0) \left(1 - \left(1 - \frac{N}{N_0}\right)^m\right) \\
&= V(N_0) - \sum_{m=1}^{N_0} V(m, N_0) \left(1 - \frac{N}{N_0}\right)^m \\
&= V(N_0) + \sum_{m=1}^{N_0} (-1)^{m-1} V(m, N_0) \left(\frac{N}{N_0} - 1\right)^m \\
&= V(N_0) + \left(\frac{N}{N_0} - 1\right) V(1, N_0) - \left(\frac{N}{N_0} - 1\right)^2 V(2, N_0) \\
&\quad + \left(\frac{N}{N_0} - 1\right)^3 V(3, N_0) - \left(\frac{N}{N_0} - 1\right)^4 V(4, N_0) + \dots .
\end{aligned} \tag{2.43}$$

In the last two steps we write $E[V_{N_0}(N)]$ as an alternating sum of the successive spectrum elements. Note that

$$-\sum_{m=1}^{N_0} (-1)^{m-1} \left(\frac{N}{N_0} - 1\right)^m V(m, N_0)$$

is equal to $V(N_0) - E[V_{N_0}(N)]$, the number of new types observed in the sampling interval (N, N_0) .

The expressions for the vocabulary size and the spectrum elements that we have obtained hinge on the assumption that $f_{N_0}(i, N)$ is a binomially distributed random variable. This assumption implies an urn model in which we sample with replacement. However, sampling with replacement assigns nonzero probability to a hapax legomenon in the full text of N_0 tokens appearing twice in a smaller subset of N tokens. For tracing the theoretical development of the vocabulary within a given text, the binomial model therefore is an approximation. To gauge the accuracy of this approximation, we consider the exact probability for sampling without replacement. The frequency in a sample of N tokens of a word ω_i with $f(i, N_0)$ tokens in the complete text is a hypergeometric random variable with parameters N , $N_0 - N$, and $f(i, N_0)$:

Definition 2.14 A random variable X has a $(N, N_0 - N, f(i, N_0))$ hypergeometric distribution when

$$\Pr(X = m) = \frac{\binom{N}{m} \binom{N_0 - N}{f(i, N_0) - m}}{\binom{N_0}{f(i, N_0)}}.$$

To see this, first observe that there are $\binom{N_0}{f(i, N_0)}$ possible arrangements of the $f(i, N_0)$ tokens of ω_i in the text. For instance, for $N_0 = 5$ and $f(i, N_0) = 2$, we have 10 possible arrangements:

$$\begin{array}{ccccccccc} \omega_i & \omega_i & \cdot & & & & & & \\ \omega_i & \cdot & \omega_i & \cdot & \cdot & \cdot & & & \\ \omega_i & & \cdot & \omega_i & & \cdot & & & \\ \omega_i & \cdot & \cdot & \cdot & \cdot & \omega_i & & & \\ \cdot & \omega_i & \omega_i & \cdot & \cdot & & & & \\ \omega_i & \cdot & & \omega_i & \cdot & & & & \\ \omega_i & \cdot & \cdot & \cdot & \omega_i & & & & \\ \cdot & \omega_i & \omega_i & \cdot & & & & & \\ \omega_i & \cdot & & \omega_i & & & & & \\ \cdot & \omega_i & \omega_i & \cdot & & & & & \end{array}$$

There are $\binom{N}{m}$ possible arrangements for m tokens of ω_i in the first N tokens of the text. For instance, if $N = 3$ and $m = 2$, there are 3 arrangements of the two tokens of ω_i . Similarly, there are $\binom{N_0 - N}{f(i, N_0) - m}$ different arrangements of the $f(i, N_0) - m$ remaining tokens of ω_i in the second part of the text. In our present example, there is only one such arrangement ($\cdot \cdot$). The total number of possible arrangements with m tokens of ω_i in the first part and $f(i, N_0) - m$ tokens in the second part therefore is

$$\binom{N}{m} \binom{N_0 - N}{f(i, N_0) - m}.$$

The hypergeometric probability is obtained by dividing this number arrangements by the total number of possible arrangements of the $f(i, N_0)$ tokens of ω_i in the text, $\binom{N_0}{f(i, N_0)}$. For the present example, this probability is 3/10.

The expectation of a hypergeometric random variable with parameters $(N, N_0 - N, f(i, N_0))$ is fN/N_0 : Writing f for $f(i, N_0)$, we have:

$$\begin{aligned}
 E[X] &= \sum_{m=0}^f m \frac{\binom{N}{m} \binom{N_0-N}{f-m}}{\binom{N_0}{f}} \\
 &= \sum_{m=1}^f N \frac{\binom{N-1}{m-1} \binom{N_0-N}{f-m}}{\binom{N_0}{f}} \\
 &= \sum_{m=1}^f N \frac{f}{N_0} \frac{\binom{N-1}{m-1} \binom{N_0-N}{f-m}}{\binom{N_0-1}{f-1}} \\
 &= f \frac{N}{N_0} \sum_{m=0}^f \frac{\binom{N-1}{m} \binom{N_0-N}{f-1-m}}{\binom{N_0-1}{f-1}} \\
 &= fN/N_0.
 \end{aligned} \tag{2.44}$$

Note that this expectation is identical to the expectation of a binomially distributed random variable with parameters $(f(i, N_0), N/N_0)$. The variance of the hypergeometric random variable,

$$\text{VAR}[X] = f \frac{N}{N_0} \left(1 - \frac{N}{N_0}\right) \frac{N_0 - f}{N_0 - 1}, \tag{2.45}$$

is smaller, however, than the binomial variance $f \frac{N}{N_0} \left(1 - \frac{N}{N_0}\right)$ (see questions 3 and 6).

The expectation for the spectrum elements given the hypergeometric model can now be written as

$$\begin{aligned}
 E[V_{N_0}(m, N)] &= \sum_{i=1}^{V(N_0)} \frac{\binom{N}{m} \binom{N_0-N}{f(i, N_0)-m}}{\binom{N_0}{f(i, N_0)}} \\
 &= \sum_{k \geq m} V(k, N_0) \frac{\binom{N}{m} \binom{N_0-N}{k-m}}{\binom{N_0}{k}},
 \end{aligned}$$

and the expected vocabulary size as

$$\begin{aligned}
 E[V_{N_0}(N)] &= \sum_{m=1}^N E[V_{N_0}(m, N)] \\
 &= \sum_{m=1}^N \sum_{k \geq m} V(k, N_0) \frac{\binom{N}{m} \binom{N_0-N}{k-m}}{\binom{N_0}{k}}.
 \end{aligned}$$

These expressions are rather unwieldy compared to the binomial expressions. Fortunately, the binomial expressions are accurate approximations of the hypergeometric expressions for lexical data. For large N_0 , $N_0 \gg f(i, N)$, and $N \gg m$, the hypergeometric probability is very similar to the binomial probability. To see this, we first observe that for $a \gg b$, a large and b small,

$$\binom{a}{b} \approx \frac{a^b}{b!},$$

which is easily understood by writing

$$\begin{aligned} \binom{a}{b} &= \frac{a(a-1)(a-2)\dots(a-b+1)}{b!} \\ &= \frac{a^b}{b! \prod_{i=0}^{b-1} (1 - \frac{i}{a})}. \end{aligned}$$

Applying this approximation to each of the three terms of the hypergeometric probability immediately leads to the familiar binomial probability:

$$\begin{aligned} \frac{\binom{N}{m} \binom{N_0 - N}{f(i, N_0) - m}}{\binom{N_0}{f(i, N_0)}} &\approx N^m \frac{(N_0 - N)^{f(i, N_0) - m}}{N_0^{f(i, N_0)}} \frac{f(i, N_0)!}{m!(f(i, N_0) - m)!} \\ &= \binom{f(i, N_0)}{m} \left(\frac{N}{N_0}\right)^m \left(\frac{N_0 - N}{N_0}\right)^{f(i, N_0) - m} \end{aligned}$$

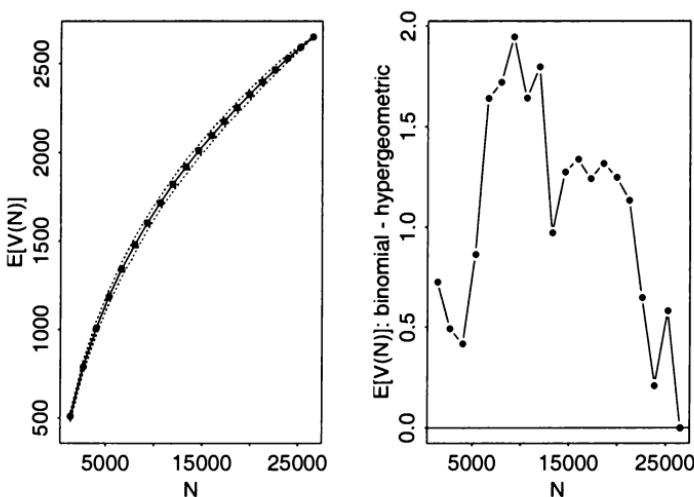


Figure 2.7: $E[V(N)]$ as a function of N using binomial and hypergeometric interpolation. The left panel shows the hypergeometric values and their 95% confidence intervals, using Monte Carlo simulations, for Alice in Wonderland. The dots plot the binomial expectations. The right panel illustrates the slight overestimation bias of the binomial model.

Figure 2.7 provides a graphical illustration of how good the binomial approximation is for *Alice in Wonderland*. The large dots in the left hand panel

represent $E[V_{N_0}(N)]$ using the binomial model. Their position is, at least to the eye, undistinguishable from the predictions of the hypergeometric model, plotted as a solid line. It is only upon close inspection that we find a slight overestimation bias for the binomial model of maximally 2 types, as shown in the right hand panel of Figure 2.7. This overestimation bias is well within the 95% confidence interval of the hypergeometric expectations. This example illustrates that the binomial model is an acceptable approximation for textual data.

2.6.2 Extrapolation

The expressions for $E[V_{N_0}(m, N)]$ and $E[V_{N_0}(N)]$ developed in the preceding section hold for $N < N_0$. In this section, we shall see that they are also valid for $N > N_0$. I will first present a proof for $E[V_{N_0}(N)]$ using the same line of reasoning that underlies the Good-Turing estimates. I will then outline a proof for $E[V_{N_0}(m, N)]$ which is based on writing $E[V_{N_0}(N)]$ in the form of a Taylor series. The Taylor series expansion of a function $f(x)$ in the point $x = a$ is

$$f(a) + f'(a) \cdot (x - a) + \frac{f''(a)}{2!} (x - a)^2 + \frac{f'''(a)}{3!} (x - a)^3 + \dots + \frac{f^{(n)}(a)}{n!} (x - a)^n + \dots, \quad (2.46)$$

where $f^{(n)}(a)$ denotes the value of the n th derivative of the function $f(x)$ at $x = a$. Consider, for instance, the function

$$f(x) = 3x^2 + 2x + 4$$

and its derivatives in $x = 2$:

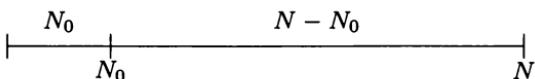
$$\begin{aligned} f'(x) &= 6x + 2 \\ f''(x) &= 6x \\ f'''(x) &= 0. \end{aligned}$$

It is easy to verify that

$$3x^2 + 2x + 4 = f(2) + f'(2) \cdot (x - 2) + f''(2)/6 + 0.$$

In this simple example, the use of the Taylor series expansion does not simplify the calculation of $f(x)$. However, for computing the values of more complex functions such as $\log(x)$, e^x , $\sin(x)$, and so on, the Taylor series expansion is very important. Often, the values of these functions can be approximated with a high degree of accuracy by computing a small number of derivatives $f^{(n)}(x)$. We shall see that in the Taylor series expansion of $E[V_{N_0}(N)]$, the given sample size N_0 plays the pivotal role of a in (2.46).

The path we shall follow to arrive at expressions for extrapolation starts with the idea to write $E[V_{N_0}(N)]$ as the sum of $E[V(N_0)]$ and the number of new types $E[V_{0, N_0}(N - N_0)]$ in an additional $N - N_0$ tokens.



Given independence in word use, we can view these $N - N_0$ tokens as a second, independent sample that we add to the N_0 tokens that we already have to form a new text of N tokens. What we need, then, is an expression for $E[V_{0,N_0}(N - N_0)]$, the number of types with zero frequency in a first sample of N_0 tokens and with nonzero frequency in a second independent sample of $N - N_0$ tokens. To obtain such an expression, let

$$i(W, N) = \begin{cases} 1 & \text{if } f(\omega_W, N) > 0 \\ 0 & \text{if } f(\omega_W, N) = 0 \end{cases}$$

denote a random Bernoulli variable that equals unity if the word ω_W indexed by the random variable W occurs in a sample of N tokens, and zero otherwise. The expectation of $i(W, N)$ is

$$\begin{aligned} E[i(W, N)] &= E[E[i(W, N)|W = w]] \\ &= E[E[i(w, N)]] \\ &= E[1 - e^{-N\pi_w}] \\ &= \sum_{i=1}^S (1 - e^{-N\pi_i}) \Pr(W = i). \end{aligned}$$

In section 2.5, we already derived an expression for the probability that W happens to index a word ω_w given that ω_w occurred m times in the sample of size N :

$$\begin{aligned} \Pr(W = w|f(W, N) = m) &= \frac{\binom{N}{m} \pi_w^m (1 - \pi_w)^{N-m}}{\sum_{i=1}^S \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m}} \\ &= \frac{\frac{(N\pi_w)^m}{m!} e^{-N\pi_w}}{E[V(m, N)]}. \end{aligned}$$

With these expressions we can calculate $V_{m,N_0}(N)$, the number of types appearing in a second sample of N tokens given that they also occur m times in our first sample of N_0 tokens:

$$V_{m,N_0}(N) = \sum_{i=1}^S \{i(W, N)|f(W, N_0) = m\}.$$

Since there are exactly $V(m, N_0)$ types for which $f(i, N_0) = m$, we can simplify this expression:

$$V_{m,N_0}(N) = V(m, N_0) \{i(W, N)|f(W, N_0) = m\}.$$

Before taking expectations on both sides, we first calculate

$$\begin{aligned} E[i(W, N)|f(W, N_0) = m] &= \sum_{i=1}^S (1 - e^{-N\pi_i}) \Pr(W = i|f_1(W, N_0) = m) \\ &= \sum_{i=1}^S (1 - e^{-N\pi_i}) \frac{\frac{(N_0\pi_i)^m}{m!} e^{-N_0\pi_i}}{E[V(m, N_0)]}. \end{aligned}$$

We can now write

$$\begin{aligned} E[V_{m,N_0}(N)] &= E[V(m, N_0) \{i(W, N) | f(W, N_0) = m\}] \\ &= E[V(m, N_0)] E[i(W, N) | f(W, N_0) = m] \\ &= \sum_{i=1}^S (1 - e^{-N\pi_i}) \frac{(N_0\pi_i)^m}{m!} e^{-N_0\pi_i}. \end{aligned}$$

This expression is simplified by making use of the series expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$:

$$1 - e^{-N\pi} = N\pi - \frac{(N\pi)^2}{2!} + \frac{(N\pi)^3}{3!} - \dots$$

We reformulate $(1 - e^{-N\pi}) \frac{(N_0\pi)^m}{m!} e^{-N_0\pi}$ as a series with as j^{th} term

$$\begin{aligned} (-1)^{j+1} \frac{(N_0\pi)^m}{m!} e^{-N_0\pi} \frac{(N\pi)^j}{j!} &= (-1)^{j+1} \frac{(N_0\pi)^m}{m!} e^{-N_0\pi} \left(\frac{N}{N_0}\right)^j \frac{(N_0\pi)^j}{j!} \\ &= (-1)^{j+1} \left(\frac{N}{N_0}\right)^j \frac{(N_0\pi)^{m+j}}{m! j!} e^{-N_0\pi} \\ &= (-1)^{j+1} \left(\frac{N}{N_0}\right)^j \binom{m+j}{m} \frac{(N_0\pi)^{m+j}}{(m+j)!} e^{-N_0\pi}, \end{aligned}$$

which leads directly to

$$\begin{aligned} E[V_{m,N_0}(N)] &= \sum_{i=1}^S (1 - e^{-N\pi_i}) \frac{(N_0\pi_i)^m}{m!} e^{-N_0\pi_i} \\ &= \sum_{i=1}^S \sum_{j=1}^{\infty} (-1)^{j+1} \left(\frac{N}{N_0}\right)^j \binom{m+j}{m} \frac{(N_0\pi)^{m+j}}{(m+j)!} e^{-N_0\pi} \\ &= \sum_{j=1}^{\infty} (-1)^{j+1} \left(\frac{N}{N_0}\right)^j \binom{m+j}{m} E[V(m+j, N_0)]. \quad (2.47) \end{aligned}$$

For $m = 0$, this expression simplifies to

$$E[V_{0,N_0}(N)] = \sum_{j=1}^{\infty} (-1)^{j+1} \left(\frac{N}{N_0}\right)^j E[V(j, N_0)], \quad (2.48)$$

the number of new types observed in a second sample of size N given a sample of size N_0 . We immediately have that

$$\begin{aligned} E[V_{N_0}(N)] &= E[V(N_0)] + E[V_{0,N_0}(N - N_0)] \\ &= E[V(N_0)] + \sum_{m=1}^{\infty} (-1)^{m+1} \left(\frac{N - N_0}{N_0}\right)^m E[V(m, N_0)], \end{aligned}$$

which becomes identical to (2.43) when we replace the expectations by the corresponding observed counts $V(N_0)$ and $V(m, N_0)$.

To obtain an expression for the expected spectrum elements for $N > N_0$, we make use of the Taylor series expansion of $E[V_{N_0}(N)]$. We rewrite (2.46), replacing a by x_0 ,

$$f(x) = f(x_0) + \sum_{n=1}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n,$$

and state $E[V_{N_0}(N)]$ in the same form:

$$E[V_{N_0}(N)] = E[V(N_0)] + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{E[V(n, N_0)]}{N_0^n} (N - N_0)^n. \quad (2.49)$$

We now have an expression that already starts to resemble a Taylor series, but for this resemblance to hold, (2.49) should be equivalent to

$$E[V_{N_0}(N)] = E[V(N_0)] + \sum_{n=1}^{\infty} \frac{E[V^{(n)}(N_0)]}{n!} (N - N_0)^n. \quad (2.50)$$

To check whether the successive terms of (2.49) and (2.50) are identical, we need an expression for $E[V^{(n)}(N)]$. Using (2.21) and (2.20), we can calculate the successive derivatives of $E[V(N)]$:

$$\begin{aligned} E[V^{(1)}(N)] &= \frac{E[V(1, N)]}{N} \\ E[V^{(2)}(N)] &= \frac{d}{dN} \frac{E[V(1, N)]}{N} \\ &= \frac{N \frac{E[V(1, N)] - 2E[V(2, N)]}{N} - E[V(1, N)]}{N^2} \\ &= -\frac{2E[V(2, N)]}{N^2} \\ E[V^{(n)}(N)] &= \frac{d}{dN} \frac{(n-1)!E[V(n-1, N)](-1)^n}{N^{n-1}} \\ &= (-1)^{n+1} \frac{n!E[V(n, N)]}{N^n} \end{aligned}$$

Substitution of $E[V^{(n)}(N)]$ in (2.50) immediately leads to (2.49), proving the equivalence of these two expressions for $E[V_{N_0}(N)]$. Importantly, we can now express $E[V(n, N)]$ in terms of the n^{th} derivative of $E[V(N)]$:

$$E[V(n, N)] = (-1)^{n+1} E[V^{(n)}(N)] \frac{N^n}{n!} \quad (2.51)$$

When we differentiate $E[V_{N_0}(N)]$ n times in N , this time using (2.49) as our point of departure, we obtain

$$E[V_{N_0}^{(n)}(N)] = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{E[V(m, N_0)]}{N_0^m} \frac{m!}{(m-n)!} (N - N_0)^{m-n}.$$

Substituting in (2.51), we obtain the expression we set out to prove:

$$\mathbb{E}[V_{N_0}(n, N)]$$

$$\begin{aligned}
&= (-1)^{n+1} \mathbb{E}[V_{N_0}^{(n)}(N)] \frac{N^n}{n!} \\
&= (-1)^{n+1} \sum_{m=1}^{\infty} (-1)^{m+1} \frac{\mathbb{E}[V(m, N_0)]}{N_0^m} \frac{m!}{(m-n)!} (N - N_0)^{m-n} \frac{N^n}{n!} \\
&= \sum_{m=1}^{\infty} \mathbb{E}[V(m, N_0)] \binom{m}{n} \left(\frac{N}{N_0}\right)^n \left(\frac{N_0 - N}{N_0}\right)^{m-n} (-1)^{2m+2} \\
&= \sum_{m \geq n}^{\infty} \mathbb{E}[V(m, N_0)] \binom{m}{n} \left(\frac{N}{N_0}\right)^n \left(\frac{N_0 - N}{N_0}\right)^{m-n}
\end{aligned} \tag{2.52}$$

It is important to realize that the Poisson model cannot be used for interpolation. Even though the Poisson approximation to the binomial probabilities leads to an expression for $\mathbb{E}[V_{N_0}(N)]$,

$$\begin{aligned}
\mathbb{E}[V_{N_0}(N)] &= \sum_{i=1}^{V(N_0)} \left(1 - \left(1 - \frac{N}{N_0}\right)^{f(i, N_0)}\right) \\
&\approx \sum_{i=1}^{V(N_0)} \left(1 - e^{-\frac{N}{N_0} f(i, N_0)}\right) \\
&= \sum_{i=1}^{V(N_0)} \left(1 - e^{-\frac{f(i, N_0)}{N_0} N}\right) \\
&= \sum_{i=1}^{V(N_0)} \left(1 - e^{-N p(i, N_0)}\right),
\end{aligned} \tag{2.53}$$

that is very similar to the general, unconditional, expression for $\mathbb{E}[V(N)]$,

$$\mathbb{E}[V(N)] = \sum_{i=1}^S \left(1 - e^{-N \pi_i}\right),$$

differing only in the use of the sample frequencies $p(i, N_0)$ instead of the population probabilities π_i , the coefficient of loss C_L (2.26) shows that for samples in the LNRE zone the use of the relative sample frequencies leads to serious underestimation of the vocabulary size. In fact, the approximation of $\left(1 - \frac{N}{N_0}\right)^{f(i, N_0)}$ by $e^{-\frac{N}{N_0} f(i, N_0)}$ is accurate only when $\frac{N}{N_0}$ is small compared to $f(i, N_0)$. For larger values of N and small values of $f(i, N_0)$,

$$e^{-\frac{N}{N_0} f(i, N_0)} \gg \left(1 - \frac{N}{N_0}\right)^{f(i, N_0)},$$

hence, the expected vocabulary size is underestimated.

How serious the underestimation is when we use the Poisson approximation is shown in Figure 2.8 for *Alice in Wonderland*. The solid line represents the observed development of the vocabulary size $V(N)$ through token time,

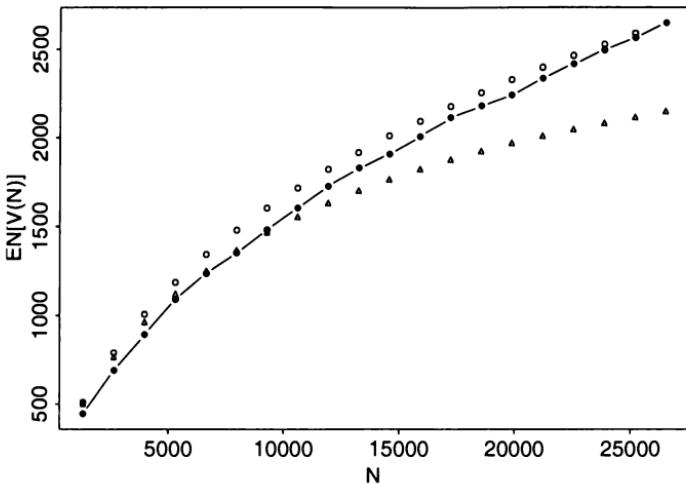


Figure 2.8: The empirical vocabulary growth curve $V(N)$ (solid line), its expectation using binomial interpolation (circles), and its expectation using the Poisson approximation (triangles), for Alice in Wonderland, at 20 equally spaced intervals.

the circles show its expectation using exact binomial interpolation (2.42) conditional on the frequency spectrum of the full text. The triangles represent the corresponding values using the Poisson approximation (2.53). The binomial expectations are slightly too high. As will become clear in section 5.1, this slight overestimation arises due to the nonrandom use of words. The Poisson approximation clearly results in a severe underestimation of the vocabulary size for the larger values of N : conditioning on the spectrum at N_0 as the population, the Poisson approximation to the binomial probabilities is too inaccurate to be of practical use.

Of course, it is possible to use the Poisson model to answer the slightly different question of how many types that occur in a (larger) text of N_0 tokens would appear in an independent, smaller text with N tokens. Since we are now dealing with two independent samples from some population, we can use the Poisson model to obtain the number of shared types $E[V_{\cdot, N_0}(N)]$:

$$E[V_{\cdot, N_0}(N)] = \sum_{i=1}^S (1 - e^{-N\pi_i})(1 - e^{-N_0\pi_i}).$$

The expected number of types in the smaller sample is simply

$$E[V(N)] = \sum_{i=1}^S (1 - e^{-N\pi_i}) > E[V_{< N_0}(N)].$$

Clearly, the smaller sample now contains types that do not occur in the larger sample. An independent smaller sample from the same population as the larger one is unconstrained with respect to the vocabulary observed for the larger sample. I have therefore limited the use of the term 'interpolation' to denoting binomial or hypergeometric estimations for a part of a given text, using the notation $E[V_{N_0}(N)]$ for the vocabulary size and $E[V_{N_0}(m, N)]$ for the spectrum elements. By necessity, all the words that occur in a part of size N also occur in the whole of N_0 tokens. For the cross-comparison of two independent samples from the same population, the smaller sample may contain words that do not occur in the larger one. For cross-comparisons, I use the notation $E[V_{< N_0}(N)]$ for the number of types in the sample with N tokens that also occur in the sample of N_0 tokens, $E[V_{< N_0}(m, N)]$ for the spectrum elements at N given that the types counted also occur in N_0 , $E[V_{n, N_0}(N)]$ for the number of types in N for which $f(i, N_0) = n$, and $E[V_{n, N_0}(m, N)]$ for the number of types with frequency n in N_0 and frequency m in N tokens.

Two serious problems remain concerning the practical application of (2.51) for sample sizes that are much larger than the observed text or corpus size N_0 . The first problem is that in actual textual data the spectrum elements $V(m, N)$ often reveal small irregularities, such as that $V(m, N)$ is slightly smaller instead of greater than $V(m + 1, N)$ for particular values of m . Such irregularities wreak havoc with the convergence of the alternating sum in (2.51). The curve for the frequency spectrum requires smoothing before (2.51) can be applied. Smoothing the observed frequency spectrum $V(m, N)$ amounts to replacing $V(m, N)$ by $E[V(m, N)]$, where the expected values are supplied by some smooth function such as, for instance, Zipf's harmonic distribution,

$$E[V(m, N)] = \frac{E[V(N)]}{m(m + 1)},$$

as illustrated for *Alice in Wonderland* in Figure 2.9 for $N = 10,602$.

A second problem is that for $N \gg N_0$ the alternating sum tends to diverge so rapidly that techniques for forcing convergence break down (see Good and Toulmin, 1956; Efron and Thisted, 1976). While it is no problem to interpolate to sample sizes $N < N_0$, and while extrapolation is possible in principle, in practice extrapolation using (2.52) is feasible only for $N < 2N_0$. To obtain more useful expressions for extrapolation, we need to solve the two problems of how to smooth the frequency spectrum and of how to avoid the divergence of the alternating sum. These questions are addressed in detail in the next chapter.

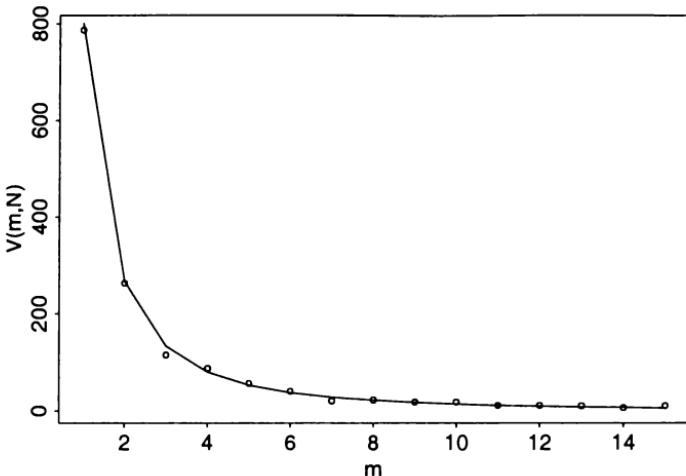


Figure 2.9: Smoothing the first 15 spectrum elements of Alice in Wonderland using Zipf's law at the sample size $N = 10,602$.

2.7 Discussion

In this chapter, we have explored the possibilities of non-parametric models, models that do not make any assumptions about the structural distributions. We have seen that non-parametric models provide the means for estimating the vocabulary size and the spectrum elements for sample sizes smaller than the observed sample size. For extrapolation beyond the observed sample size to larger (unobserved) sample sizes, these models fail for technical reasons. The non-parametric models also do not provide estimates of the population number of types. How to overcome these shortcomings is the topic of the next chapter.

2.8 Bibliographical Comments

Useful introductions to probability theory are Ross (1988) and the first chapters of Rice (1988). Urn models and their applications are analyzed in Johnson and Kotz (1977). The concept of LNRE distributions is introduced in Khmaladze (1987), the distinction between early and late LNRE zones is developed in Chitashvili and Baayen (1993). The Good-Turing estimates are introduced in Good (1953), further details can be found in Church and Gale (1991), Nádas (1985), Huang and Lo (1994), and Gale and Sampson (1995). Interpolation and extrapolation are discussed in a wide range of publications. Primary sources are Good (1953), Good and Toulmin (1956), Kalinin (1965), Efron and Thisted (1976), Muller (1977), Muller (1979a), and Muller (1979c).

2.9 Questions

1. Show that $E[K] = 10000 \sum_i \pi^2$ for $N \rightarrow \infty$.
2. The entropy of the frequency spectrum is

$$\mathcal{E} = - \sum_{i=1}^S \pi_i \log(\pi_i).$$

Explain, given that $-\log(\pi_i)$ is a measure of the amount of information, why \mathcal{E} can be viewed as the average amount of information.

3. Derive $E[f(i, N)] = N\pi_i$ using (2.1) and regarding a binomial random variable as a sum of Bernoulli trials. Also show that $\text{VAR}[f(i, N)]$ is $N\pi_i(1 - \pi_i)$.
 4. Assume that the probability of the determiner *the* in English is $\pi_{\text{the}} = 0.07$. What is the probability that *the* occurs exactly 7 times in a text of 100 words?
 5. What is the mean token frequency for the 6 types on a fair die? (Consider $\lim_{N \rightarrow \infty} \frac{N}{E[V(N)]}$.)
 6. Show that
- $$\text{VAR}[f(i, N)] = f \frac{N}{N_0} \left(1 - \frac{N}{N_0}\right) \frac{N_0 - f}{N_0 - 1}$$
- when $f(i, N)$ is an $(N, N_0 - N, f(i, N_0))$ -distributed hypergeometric random variable.
7. Which words are most affected by the Good-Turing adjustment for the unseen types: the low-frequency words, or the high-frequency words?
 8. Use Table 1.3 to calculate $(m + 1) * V(m + 1, N) / V(m, N)$ for $m = 6$ and use the result to explain why it is necessary to use the expected spectrum elements when the curve of $V(m, N)$ is not a smooth decreasing function of m .
 9. Show that the number of types in common to two independent samples of size N and N_0 tokens is given by

$$S - E[V(0, N)] - E[V(0, N_0)] + E[V(0, N + N_0)].$$

10. Show that the number of types that occur with a frequency of m in a sample of N tokens given that they occur with a frequency of k in a sample of N_0 tokens is

$$\sum_{i=1}^S E[V(k + m, N + N_0)] \frac{N^m N_0^k}{(N + N_0)^{k+m}} \binom{k + m}{m}.$$

11. Thisted and Efron (1987: 447) introduce $G(\lambda)$ as "The empirical cumulative distribution function" for the Poisson rates $\lambda_i, i = 1, \dots, S$ with which types appear. They give the following expression for $E[V(x, N_0)]$:

$$S \int_0^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} dG(\lambda).$$

Relate Thisted and Efron's $G(\lambda)$ to the structural type distribution.

12. Carroll (1970) proposes a standard frequency index $SFI = 10^{(10 \log p(i, N) + 10)}$. One of the studies Carroll refers to as justification for the logarithmic transformation is Shapiro (1969), who presents evidence that human perception is sensitive to log frequency rather than to absolute frequency. Revise the SFI using the Good-Turing frequency adjustment.

Chapter 3

Parametric models

This chapter introduces three families of LNRE models: Carroll's (1967) log-normal model, Sichel's (1975) generalized inverse Gauss-Poisson model, and Orlov and Chitashvili's (1983a,b) extended Zipf's law. By enriching the non-parametric expressions of the preceding chapter with parametric assumptions about the shape of the structural distribution, we can avoid the technical problems associated with the non-parametric methods. Section 3.2 describes the three LNRE models. Sections 3.3 and 3.6 discuss techniques for evaluating the goodness of fit of the LNRE models and for testing whether word frequency distributions are different by means of statistics such as $V(N)$ and $V(1, N)$.

3.1 Introduction

First consider the question of how to obtain a smoothed curve for the frequency spectrum. One very simple model that we might want to use as a smoother for the spectrum is Zipf's law in the form

$$V(m, N) = \frac{V(N)}{m(m+1)}. \quad (3.1)$$

Recall, however, that the parameters of Zipf's law change with the sample size, as shown in Chapter 1. This implies that (3.1), the Zeta distribution with the parameter a fixed at unity, would have to apply to one particular sample size only. Is there a sample size for *Alice in Wonderland* for which the harmonic distribution provides a reasonable fit? To study the appropriateness of the harmonic distribution, we need a measure for gauging how well its predictions for, say, the first 15 spectrum elements match the observed values. An informal measure is the relative mean squared error $\text{MSE}_r(m')$, which quantifies the extent to which the predicted values diverge from the observed ones relative to the vocabulary size for the spectrum elements $m = 1, 2, \dots, m'$:

$$\text{MSE}_r(m') = \frac{1}{m'} \sum_{m=1}^{m'} \left(\frac{\text{E}[V(m, N)] - V(m, N)}{V(N)} \right)^2. \quad (3.2)$$

Figure 3.1 shows the MSE_r scores for the harmonic distribution for 20 equal-spaced sample sizes of *Alice in Wonderland*. The u-shaped curve suggests that the harmonic distribution is inappropriate for the smallest sample sizes, that it is reasonable for samples of roughly 10,000 words, and that it becomes increasingly inappropriate again for larger values of N . To my knowledge, Orlov was the first to call attention to this kind of dependence of the goodness of fit on the sample size (see, e.g., Orlov (1983a) and references cited there).

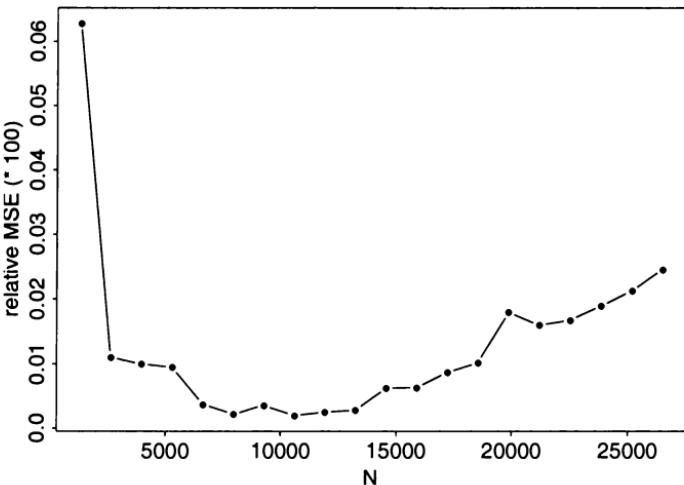


Figure 3.1: Zipf's law as optimal for $N = 10602$. The plot shows the relative Mean Squared Error (MSE), calculated on the basis of the first 15 spectrum elements for *Alice in Wonderland*.

The smoothing of the first 15 spectrum elements of *Alice in Wonderland* shown in Figure 2.9 shows that for the sample size $N = 10602$, the sample size for which the relative mean squared error is smallest for the given 20 measurement points, the harmonic distribution

$$E[V(m, 10602)] = \frac{V(10602)}{m(m+1)},$$

with $V(10602)$ the number of types observed for $N = 10602$ tokens, provides a fit that, at least to the eye, is quite reasonable. Interestingly, the sample size $N = 10602$ now emerges as a characteristic of *Alice in Wonderland*: it is the unique sample size at which Zipf's harmonic distribution is valid. We can regard this sample size, to which I will henceforth refer as the **Zipf size Z**, as a characteristic constant of this novel. In fact, we can define the sample size Z formally as the sample size at which the frequency spectrum of *Alice in Wonderland* converges to Zipf's harmonic spectrum law.

As we now have a smooth function for the spectrum of *Alice in Wonderland* at $N = Z$, we can replace the observed $V(m, Z)$ by the expected values $E[V(m, Z)]$ that follow from Zipf's harmonic distribution. This solves the smoothing problem, as we can now use the spectrum elements $E[V(m, Z)]$ for interpolation and extrapolation to other sample sizes. In other words, Z is the pivotal sample size N_0 on which we condition when considering larger or smaller sample sizes.

An analogous situation obtains when we try to approximate a Poisson random variable Y by a binomial variable X (the triangle scheme of experiment). For large n and small p , our binomially (n, p) -distributed random variable X is approximately Poisson(np)-distributed. When we consider particular values of n and p , say $n' = 100000$ and $p' = 0.0001$, and setting $X = 20$, we have that the binomial probability $\Pr(X = 20) = 0.001865$. Roughly the same probability is obtained using the Poisson approximation for $Y = 20$ with $\lambda = n'p' = 10$ ($\Pr(Y = 20) = 0.001866$). When we change the number of trials from $n' = 100000$ to $n = 50000$ while keeping the same p' , the approximation is no longer good: $\Pr(X = 20)$ becomes nearly zero (0.0000003). By introducing a new parameter $t = n'/n$ that takes into account the divergence from the original number of trials n' , we obtain a new binomial distribution with parameters tn and p' that again provides a good approximation. In the case of Zipf's law, Z plays a similar role as n' . When a sample is not at the Zipf size, we have to adjust for this fact by means of an extra parameter t . In this way we can further justify the use of continuous structural distributions in the modeling of discrete word frequency distributions: for a particular sample size Z , the discrete distribution asymptotically approximates the expected continuous distribution.

We have not yet considered the other problem, namely how to avoid divergence for the alternating sum (2.43) when extrapolating to larger sample sizes. The solution to this problem is to enrich the urn model with an explicit parametric expression for the structural type distribution. What we need is an hypothesis concerning the form of $G(\pi)$ that allows us, in the case of Zipf's law, to express $E[V(m, N)]$ in the form $V(N)/[m(m+1)]$, with $N = Z = 10602$ for *Alice in Wonderland*. In other words, we need to define the structural type distribution such that the following equations hold:

$$\begin{aligned} E[V(m, Z)] &= \int_0^\infty \frac{(Z\pi)^m}{m!} e^{-Z\pi} dG(\pi) \\ &= \frac{1}{Z} \int_0^\infty \frac{f(z)^m}{m!} e^{-f(z)} dG(f(z)) \\ &= \frac{1}{Z} \int_0^\infty \frac{\lambda^m}{m!} e^{-\lambda} dG(\lambda), \end{aligned}$$

writing $\lambda = f(z) = Z\pi$ for the rate (frequency) of a given type in a time unit of Z tokens. (For formal justification of writing $G(\pi)$ in integral form, see section 2.3.) Given $G(\pi)$, we can then extrapolate and interpolate to other

sample sizes. With $t = N/Z$, we then have

$$\mathbb{E}[V(m, N)] = \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} dG(\pi) \quad (3.3)$$

$$= \int_0^\infty \frac{\left(\frac{N}{Z}(Z\pi)\right)^m}{m!} e^{-\frac{N}{Z}(Z\pi)} dG(\pi) \quad (3.4)$$

$$= \frac{1}{Z} \int_0^\infty \frac{(t\lambda)^m}{m!} e^{-t\lambda} dG(\lambda).$$

A crucial point here is that an analytic expression for $V(m, N)$ must contain a parameter specifying Z as the time unit with respect to which the rate λ is defined.

The introduction of a parametric expression for $G(\pi)$ solves the problem of how to take the unseen types into account, as any hypothesis concerning the form of $G(\pi)$ entails an hypothesis concerning the distributional properties of the unseen types.

In what follows, I will discuss two proposals for $G(\pi)$, in section 3.2.1 the lognormal structural type distribution, and in section 3.2.2 the inverse Gauss-Poisson structural type distribution. It turns out that there is no explicit expression available for the structural type distribution for Zipf's law. In section 3.2.3, I will outline how an LNRE model can nevertheless be formulated by enriching the urn model with a parametric expression for the ratio $\mathbb{E}[V(m, N)]/\mathbb{E}[V(N)]$.

3.2 LNRE models

SUMMARY *To solve the extrapolation problem, we enrich the urn model with a hypothesis concerning the form of the structural type distribution. Given a parametric expression for the structural type distribution, we can obtain general expressions for the vocabulary size and the spectrum elements that work equally well for interpolation and extrapolation. The resulting LNRE models are parameterized in sample size, that is, they have a free parameter, Z , that specifies the pivotal theoretical sample size from which the model interpolates and extrapolates. Three LNRE models are at present available: Sichel's inverse Gauss-Poisson model, Carroll's lognormal model, and Orlov and Chitashvili's extended Zipf's law.*

3.2.1 The Lognormal Structural Type Distribution

In Chapter 1, the lognormal hypothesis was introduced for the frequencies of the individual words ω_i . In this simple form, the lognormal hypothesis has two obvious drawbacks. First, as we increase the sample size, the parameters of the model likewise change. Second, in a quantile-quantile plot, the lowest frequencies show up as vertical bars instead of being aligned with the other words, as shown in right panel of Figure 1.18, repeated here for convenience as Figure 3.2. The first problem can be avoided by taking the number of words

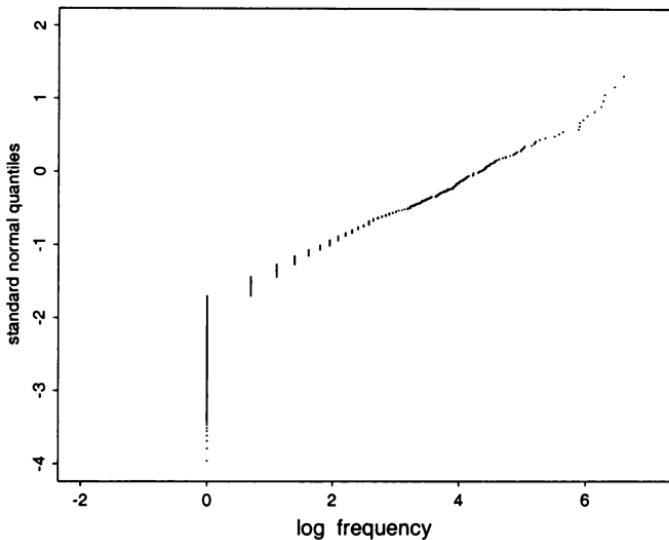


Figure 3.2: Quantile-quantile plot of log frequency for Alice in Wonderland.

that do not appear in the sample into account using the LNRE approach. With respect to the second problem, observe that the upper ends of the initial vertical bars in Figure 3.2 are nicely aligned with the remainder of the distribution. This suggests that it might be useful to reformulate the lognormal hypothesis in some cumulative form.

To this end, we may consider the **structural token distribution**, the probability mass of all words ω_i with probability $\pi_i \geq \pi$:

$$F(\pi) = \sum_{i=1}^S \pi_i I_{[\pi_i \geq \pi]}. \quad (3.5)$$

In the same way as for the structural type distribution (see 2.3), we index the ranked probabilities such that $\pi_j < \pi_{j+1}$, and define $V(\pi_j)$ as the number of types with probability π_j . Let there be κ distinct probability ranks, i.e., κ denotes the type count of different probabilities that the S types have ($\kappa < S$). The structural token distribution $F(\pi)$ is a step function with jumps $\Delta F(\pi_j)$ at those probabilities π_j for which $V(\pi_j) > 0$. Now

$$\begin{aligned} \Delta F(\pi_j) &= F(\pi_j) - F(\pi_{j+1}) \\ &= \sum_{s=j}^{\kappa} \pi_s V(\pi_s) - \sum_{s=j+1}^{\kappa} \pi_s V(\pi_s) \\ &= \pi_j V(\pi_j) \end{aligned}$$

$$\begin{aligned} &= \pi_j [G(\pi_j) - G(\pi_{j+1})] \\ &= \pi_j \Delta G(\pi_j), \end{aligned}$$

so that we can express the jumps of the structural type distribution in terms of the jumps of the structural token distribution: $\Delta G(\pi_j) = \Delta F(\pi_j)/\pi_j$.

Under Carroll's (1967, 1969) lognormal hypothesis, the structural token distribution has the following form:¹

$$F(\pi) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\pi}^{\infty} \frac{1}{x} e^{-\frac{1}{2\sigma^2}[\log(x)-\mu]^2} dx.$$

The corresponding probability density function $f(x)$,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2\sigma^2}[\log(x)-\mu]^2}, \quad (3.6)$$

is the density function of a lognormally distributed random variable, with mean $E[X] = e^{\mu + \frac{1}{2}\sigma^2}$.

A random variable X is said to follow a lognormal distribution if $\log X$ has a normal distribution. The central limit theorem provides a rationale for the use of the lognormal distribution when a random variable can be viewed as the product of independent other random variables having the same distribution:

$$Y_n = X_n X_{n-1} \cdots X_2 X_1.$$

Since

$$\log Y_n = \sum_{i=1}^n \log X_i$$

is the sum of independent identically distributed random variables, $\log Y_n$ will tend to normality as n increases. Carroll (1967, 1969) applies this rationale for the lognormal model as applied to word frequency distributions by viewing the use of a word as resulting from a selection process through a binary-branching decision tree. Decision probabilities are assigned to each branching point in the decision tree. The individual words of the vocabulary appear at the leaf nodes. Each word is associated with a unique path through the tree. This path specifies the sequence of choices that have to be made to reach the word starting from the root of the tree. The probability of selecting a word is the product of the decision probabilities of its path. Hence, the logarithm of this probability is the sum of the choice probabilities. Given the central limit theorem combined with the appropriate assumptions about the probability distribution of the choice probabilities and the length of the decision paths through the tree, it can be shown that the lognormal model follows (Chitashvili and Baayen, 1993).

The density function (3.6) belongs to a family of j -th moment-weighted functions $f_j(x)$:

$$f_j(x) = \frac{1}{\sigma\sqrt{2\pi}} x^j e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2}$$

¹Note that the constant $\pi = 3.14\dots$ is printed in bold to distinguish it from the probability π .

$$\begin{aligned}
&= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\log(x)^2 - 2\mu\log(x) + \mu^2 - 2\sigma^2 j \log(x)]} \\
&= \frac{1}{\sigma\sqrt{2\pi}} e^{j\mu + \frac{j^2}{2}\sigma^2} e^{-\frac{1}{2\sigma^2}[\log(x)^2 - 2\mu\log(x) + \mu^2 - 2\sigma^2 j \log(x) + j\mu2\sigma^2 + j^2\sigma^4]} \\
&= \frac{e^{j\mu + \frac{j^2}{2}\sigma^2}}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\log(x) - (\mu + j\sigma^2)]^2}.
\end{aligned}$$

Hence we can write the probability density function (3.6) in the form

$$f_{-1}(x) = e^{\frac{1}{2}\sigma^2 - \mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\log(x) - (\mu - \sigma^2)]^2}. \quad (3.7)$$

Since $\Delta G(\pi) = \Delta F(\pi)/\pi$, the integrand for $G(\pi)$ is $f_{-2}(x)$:

$$f_{-2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x^2} e^{-\frac{1}{2\sigma^2}[\log(x) - \mu]^2} \quad (3.8)$$

$$= e^{\frac{1}{2}\sigma^2 - \mu} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2\sigma^2}[\log(x) - (\mu - \sigma^2)]^2} \right\}. \quad (3.9)$$

The expression enclosed in curly brackets represents the density function of a lognormal random variable with mean $\mu - \sigma^2$ and standard deviation σ . The number of types in the population, S , follows immediately:

$$\begin{aligned}
S &= \int_0^\infty dG(\pi) \\
&= e^{\frac{1}{2}\sigma^2 - \mu} \int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2\sigma^2}[\log(x) - (\mu - \sigma^2)]^2} dx \\
&= e^{\frac{1}{2}\sigma^2 - \mu}.
\end{aligned} \quad (3.10)$$

Note that as S is finite, a lognormal distribution does not satisfy either of the two definitions for LNRE distributions (2.12) and (2.13) developed by Khmaladze (1987). However, a lognormal distribution will better approximate a strict LNRE distribution when S is large and when $E[V(1, N)]/E[V(N)]$ remains greater than zero for a large range of values of N .

In addition to the probability density function for the token distribution, we therefore have a second probability density function for the corresponding type distribution:

$$g(x) = \frac{e^{\mu - \frac{1}{2}\sigma^2}}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2\sigma^2}[\log(x) - (\mu - \sigma^2)]^2}. \quad (3.11)$$

This is again a lognormal probability density function, but now with mean $\mu - \sigma^2$ and standard deviation σ .

The lognormal structural type distribution $G(\pi)$ can now be defined as

$$\begin{aligned}
G(\pi) &= \frac{1}{\sigma\sqrt{2\pi}} \int_\pi^\infty \frac{1}{x^2} e^{-\frac{1}{2\sigma^2}[\log(x) - \mu]^2} dx \\
&= \frac{Z}{\sigma\sqrt{2\pi}} \int_{Z\pi}^\infty \frac{1}{f_z^2} e^{-\frac{1}{2\sigma^2}[\log(f_z)]^2} df_z,
\end{aligned} \quad (3.12)$$

with $\mu = \log(1/Z)$ and $f_{(z)} = Zx$. Equivalently, we have

$$G(\lambda) = \frac{Z}{\sigma\sqrt{2\pi}} \int_{\lambda}^{\infty} \frac{1}{f_{(z)}^2} e^{-\frac{1}{2\sigma^2}[\log(f_{(z)})]^2} df_{(z)},$$

with $\lambda = Z\pi$. Thus the parameter $\mu = \log(1/Z)$ emerges as the parameter defining the Zipf size $Z = e^{-\mu}$, the unit of measurement in text time with respect to which the rate λ at which words appear is defined. For *Alice in Wonderland*, Z is estimated at approximately 320, hence, the full text is far beyond the sample size at which relative sample frequencies can be used to estimate population probabilities for the structural type distribution.

When (3.12) is substituted into (3.4), the following expressions for the vocabulary size and the spectrum elements can be obtained:

$$\mathbb{E}[V(m, N)] = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\infty} \frac{(xN)^m}{x^2 m!} e^{-xN - \frac{1}{2\sigma^2}[\log(xZ)]^2} dx \quad (3.13)$$

$$\mathbb{E}[V(N)] = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\infty} (1 - e^{-xN}) \frac{1}{x^2} e^{-\frac{1}{2\sigma^2}[\log(xZ)]^2} dx. \quad (3.14)$$

For numerical integration, it is convenient to let $y = xN$ and to write

$$V(m, N) = \frac{N}{\sigma\sqrt{2\pi}} \int_0^{\infty} \frac{y^m}{y^2 m!} e^{-y - \frac{1}{2\sigma^2}[\log(yZ/N)]^2} dy,$$

$$V(N) = \frac{N}{\sigma\sqrt{2\pi}} \int_0^{\infty} (1 - e^{-y}) \frac{1}{y^2} e^{-\frac{1}{2\sigma^2}[\log(yZ/N)]^2} dy.$$

Recall that, as illustrated in Figure 3.2 for *Alice in Wonderland*, a quantile-quantile plot of the individual log word frequencies revealed vertical line segments for the lowest frequencies. By using the structural token and type distributions, these vertical lines no longer appear. The transition from individual word probabilities to the number of words with a given probability (by means of the structural type distribution) or to the probability mass of the words with a given probability (by means of the structural token distribution) solves this problem. This is illustrated for the same text in Figure 3.3, a quantile-quantile plot with on the horizontal axis the logarithm of the empirical relative frequencies m/N , ordered from low to high, and on the vertical axis the quantiles of the standard normal distribution. The lower dotted line represents the token distribution, the upper dotted line the type distribution.

The population mean and standard deviation were estimated by requiring $\mathbb{E}[V(N)] = V(N)$ and $\mathbb{E}[V(1, N)] = V(1, N)$, which leads to the estimates $\hat{\mu} = -5.768$ and $\hat{\sigma} = 2.689$ (see section 3.4 for further details on parameter estimation). The two slanted lines in Figure 3.3 represent the estimated theoretical token and type distributions. The two lines have the same slope ($1/\hat{\sigma}$), and they are $\hat{\sigma}$ Y-units apart. The mean of the token distribution is represented by a solid vertical line. It is located at the point where the token distribution intersects the X-axis. The mean of the type distribution is represented by a dashed line. It is found $\hat{\sigma}^2$ X-units to the left of μ , where the type distribution intersects the X-axis. Finally, the dotted line represents the sample mean.

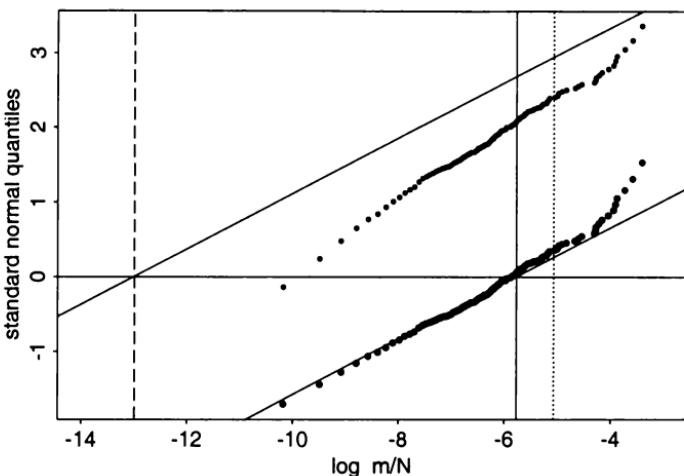


Figure 3.3: Quantile-quantile plots for the empirical token (bottom line) and type (top line) structural distributions for *Alice in Wonderland*.

Figure 3.3 illustrates a number of important properties of the lognormal token and type distributions. First, the sample mean (-5.074) overestimates the population mean (-5.768). Since

$$S = e^{\frac{1}{2}\sigma^2 - \mu},$$

it is clear that the model assumes the population number of types to be larger than inspection of the sample would suggest. Second, the theoretical standard deviation $\hat{\sigma} = 2.689$ is substantially larger than the sample standard deviation $s = 0.743$. In Figure 3.3, the slanted lines therefore have a somewhat smaller slope ($1/\hat{\sigma}$) than the empirical curves. The same difference is reflected in the distance between the type and token curves. Recall that the distance between the theoretical distributions is $\hat{\sigma}$ Y-units. Since $\hat{\sigma} > s$, the theoretical curves are further apart than their empirical analogs. This is a direct consequence of the LNRE assumption, according to which only a small subset of the words in the population actually appears in the sample. For *Alice in Wonderland*, we can estimate the population number of types at 11889 using (3.10). Of the 11889 words that Carroll might have used for writing *Alice in Wonderland*, only 22% (2651) actually appear, however. In other words, the population contains more low-probability types, its left-hand tail is larger than the sample suggests, and hence $\sigma > s$.

Third, Figure 3.3 reveals a slight deviation from normality at the right-hand side of the plot. For *Alice in Wonderland*, the highest-frequency words (e.g., *the, and, to, a, she, it, of, said, I, Alice, in, you,*) do not fall in line with

the rest of the distribution. Their misalignment is probably due to the specific properties of the highest-frequency words. Note that the highest-frequency words would have followed the lognormal pattern if their probabilities had been much higher. In other words, the highest-frequency words are simply not frequent enough. Interestingly, we have observed a similar pattern for Zipf's rank-frequency distribution. Here too, the highest-frequency words deviate from the general pattern, as illustrated for *Alice in Wonderland* in Figure 1.5. Mandelbrot (1953) captured this deviance by enriching the zeta distribution with an additional parameter. For the lognormal model, however, this is not possible.

Fortunately, the lognormal structural token distribution is quite robust with respect to the deviation from normality of the highest-frequency words. For instance, when we use (3.14) to estimate the vocabulary size at 20 equally spaced intervals for *Alice in Wonderland*, the result is very similar to that obtained by interpolation from the frequency spectrum using (2.42), as shown in Figure 3.4. As we have observed before, the theoretical values overestimate

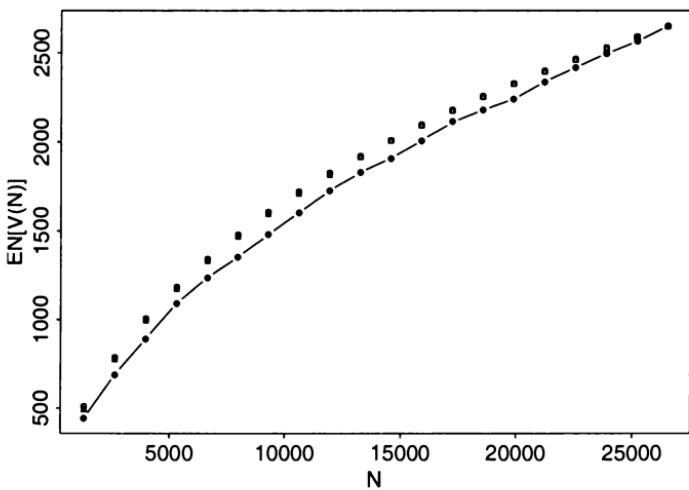


Figure 3.4: The empirical vocabulary size $V(N)$ (solid line) as a function of the text size N , its expectation using binomial interpolation (open circles), and its expectation using the lognormal structural type distribution (triangles), for *Alice in Wonderland*.

the empirical values. Crucially, however, binomial interpolation (open circles) and lognormal interpolation (triangles) lead to highly similar expected values, as expected for techniques that share the assumptions of the urn model. But whereas binomial interpolation proceeds from the complete text and regards it as the population, lognormal interpolation proceeds from the assumptions

that the text is a sample of a larger population.

3.2.2 The Generalized Inverse Gauss-Poisson Structural Type Distribution

Sichel (1975, 1986) developed a model that makes use of a parametric expression for the probability density function for the probabilities π_i of the words ω_i , the so-called generalized inverse Gauss-Poisson distribution, GIGP in short:

$$\psi(\pi) = \frac{(2/bc)^\gamma}{2K_\gamma(b)} \pi^{\gamma-1} e^{-\frac{\pi}{c} - \frac{b^2 c}{4\pi}}. \quad (3.15)$$

In (3.15), $K_\gamma(b)$ denotes the modified Bessel function of the second kind of order γ and argument b .² The generalized inverse Gauss-Poisson distribution has three free parameters, $-1 < \gamma < 0$, $b \geq 0$, and $c \geq 0$. $\psi(\pi)d\pi$ specifies the population proportion of words ω_i having a probability π_i in the interval $(\pi - \epsilon, \pi + \epsilon)$ (for small ϵ). Hence the structural type distribution for Sichel's model is

$$\begin{aligned} G(\pi) &= S \int_{\pi}^{\infty} \psi(x) dx \\ &= S \int_{\pi}^{\infty} \frac{(2/bc)^\gamma}{2K_\gamma(b)} x^{\gamma-1} e^{-\frac{x}{c} - \frac{b^2 c}{4x}} dx \\ &= S \frac{(2/b)^\gamma Z}{2K_\gamma(b)} \int_{\pi}^{\infty} (xZ)^{\gamma-1} e^{-xZ - \frac{b^2}{4xZ}} dx \\ &= S \frac{(2/b)^\gamma}{2K_\gamma(b)} \int_{\pi Z}^{\infty} f_{(z)}^{\gamma-1} e^{-f_{(z)} - \frac{b^2}{4f_{(z)}}} df_{(z)}. \end{aligned} \quad (3.16)$$

The last expression for $G(\pi)$ is obtained by replacing c with $1/Z$, and by writing $f_{(z)} = xZ$. Note that the parameter c defines the unit of measurement $Z = 1/c$ in sampling time for the rate λ at which a given word type appears:

$$G(\lambda) = S \frac{(2/b)^\gamma}{2K_\gamma(b)} \int_{\lambda}^{\infty} f_{(z)}^{\gamma-1} e^{-f_{(z)} - \frac{b^2}{4f_{(z)}}} df_{(z)}.$$

To complete Sichel's model, we need an expression for the population number of types S . We assume, for convenience, that the $V(\pi_i)$ words with probability π_i all have unique probabilities which differ from π_i by some varying small number ϵ . Under this assumption, all individual probabilities are

²The modified Bessel function of the second kind of order v is

$$K_v(z) = \frac{\pi}{2} \frac{I_{-v}(z) - I_v(z)}{\sin(v\pi)},$$

with

$$I_v(z) = \sum_{n=0}^{\infty} \frac{(z/2)^{v+2n}}{n! \Gamma(v+n+1)}.$$

Two important recurrence formulas are $K_{-v}(z) = K_v(z)$ and $K_{v-1}(z) - K_{v+1}(z) = -(2v/z)K_v(z)$. For $v = -0.5$, $K_{-0.5}(z) = (\pi/2z)^{1/2}e^{-z}$.

equiprobable, so that

$$\mathbb{E}[\pi] \sim \frac{1}{S} \sum_{i=1}^S \pi_i = \frac{1}{S}.$$

The mean of the distribution $\psi(\pi)$ is known to be

$$\mathbb{E}[\pi] = \frac{bc}{2} \frac{K_{\gamma+1}(b)}{K_\gamma(b)},$$

hence

$$S = \frac{2}{bc} \frac{K_\gamma(b)}{K_{\gamma+1}(b)}. \quad (3.17)$$

Because S is finite, Khmaladze (1987)'s formal definitions of LNRE distributions are not strictly satisfied, although for very large S Khmaladze's second definition (2.4) may be approximated quite well.

Combining (3.17), (3.16), and (2.18), writing $c = 1/Z$, and solving the integral³

$$\mathbb{E}[V(m, N)] = \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} \left(\frac{2}{b}\right)^{\gamma+1} \frac{Z^2}{2K_{\gamma+1}(b)} (Z\pi)^{\gamma-1} e^{-Z\pi - \frac{b^2}{4Z\pi}} d\pi,$$

the following expression for the spectrum elements is obtained:

$$\mathbb{E}[V(m, N)] = \frac{2Z}{bK_{\gamma+1}(b)(1+N/Z)^{\gamma/2}} \frac{\left(\frac{bN}{2Z\sqrt{1+N/Z}}\right)^m}{m!} K_{m+\gamma}(b\sqrt{1+N/Z}). \quad (3.18)$$

For $m = 0$ we have

$$\mathbb{E}[V(0, N)] = \frac{2Z}{bK_{\gamma+1}(b)(1+N/Z)^{\gamma/2}} K_\gamma(b\sqrt{1+N/Z}),$$

hence

$$\begin{aligned} \mathbb{E}[V(N)] &= \mathbb{E}[S - V(0, N)] \\ &= S - \mathbb{E}[V(0, N)] \\ &= \frac{2Z}{b} \frac{K_\gamma(b)}{K_{\gamma+1}(b)} - \frac{2Z}{bK_{\gamma+1}(b)(1+N/Z)^{\gamma/2}} K_\gamma(b\sqrt{1+N/Z}) \\ &= \frac{2Z}{b} \frac{K_\gamma(b)}{K_{\gamma+1}(b)} \left[1 - \frac{K_\gamma(b\sqrt{1+N/Z})}{(1+N/Z)^{\gamma/2} K_\gamma(b)} \right]. \end{aligned} \quad (3.19)$$

The remainder of this section considers the method suggested by Sichel to estimate the parameters of the GIGP model. For computational reasons, it turns out to be convenient to consider the relative spectrum elements

$$\alpha(m, N) = \frac{\mathbb{E}[V(m, N)]}{\mathbb{E}[V(N)]}$$

³See Luke (1962) for details on integrals of Bessel functions.

$$\begin{aligned}
&= \frac{\mathbb{E}[V(m, N)]}{S - \mathbb{E}[V(0, N)]} \\
&= \frac{1}{c_1 m!} \frac{c_2^m}{K_{\gamma+m}(b\sqrt{1+N/Z})}, \tag{3.20}
\end{aligned}$$

with

$$\begin{aligned}
c_1 &= (1 + N/Z)^{\gamma/2} K_\gamma(b) - K_\gamma(b\sqrt{1+N/Z}), \quad \text{and} \\
c_2 &= (bN)/(2Z\sqrt{1+N/Z}).
\end{aligned}$$

We can obtain an expression of $\alpha(m, N)$ in terms of $\alpha(m-1, N)$ and $\alpha(m-2, N)$ by making use of the recurrence relation

$$K_{\gamma+m}(q) = \frac{2(m+\gamma-1)}{q} K_{\gamma+m-1}(q) + K_{\gamma+m-2}(q).$$

Writing $q = b\sqrt{1+N/Z}$, we have

$$\begin{aligned}
\alpha(m, N) &= \frac{1}{c_1 m!} \frac{c_2^m}{K_{m+\gamma}(q)} \\
&= \frac{1}{c_1} \left[\frac{c_2}{m} \frac{c_2^{m-1}}{(m-1)!} \frac{2(\gamma+m-1)}{q} K_{\gamma+m-1}(q) \right. \\
&\quad \left. + \frac{c_2^2}{m(m-1)} \frac{c_2^{m-2}}{(m-2)!} K_{\gamma+m-2}(q) \right] \\
&= \frac{c_2}{m} \frac{2(\gamma+m-1)}{q} \alpha(m-1, N) + \frac{c_2^2}{m(m-1)} \alpha(2, N) \\
&= \frac{\gamma+m-1}{m} \frac{1}{Z/N+1} \alpha(m-1, N) \\
&\quad + \frac{\left(\frac{bN}{2Z\sqrt{1+N/Z}} \right)^2}{m(m-1)} \alpha(m-2, N). \tag{3.21}
\end{aligned}$$

For $m = 1$ and $\gamma = -0.5$, (3.20) reduces to

$$\alpha(1, N) = \frac{bN}{2Z\sqrt{1+N/Z}} \frac{1}{e^{b(\sqrt{1+N/Z}-1)} - 1}, \tag{3.22}$$

where we make use of the equality $K_{-1/2}(z) = K_{1/2}(z) = \sqrt{\pi/(2z)}e^{-z}$. To obtain $\alpha(2, N)$, we consider the ratio

$$\begin{aligned}
\frac{\alpha(2, N)}{\alpha(1, N)} &= \frac{\frac{1}{c_1} \frac{c_2^2}{2} K_{\gamma+2}(q)}{\frac{1}{c_1} c_2 K_{\gamma+1}(q)} \\
&= \frac{c_2}{2K_{\gamma+1}(q)} K_{\gamma+2}(q) \\
&= \frac{c_2}{2K_{\gamma+1}(q)} \left[K_\gamma(q) + \frac{2(\gamma+1)}{q} K_{\gamma+1}(q) \right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{c_2}{q} \left[\frac{qK_\gamma(q)}{2K_{\gamma+1}(q)} + \gamma + 1 \right] \\
&= \frac{1}{2} \frac{1}{Z/N + 1} \left[\frac{b\sqrt{1+N/Z}K_\gamma(b\sqrt{1+N/Z})}{2K_{\gamma+1}(b\sqrt{1+N/Z})} + \gamma + 1 \right]. \quad (3.23)
\end{aligned}$$

Hence,

$$\alpha(2, N) = \alpha(1, N) \frac{1}{2} \frac{1}{Z/N + 1} \left[\frac{b\sqrt{1+N/Z}K_\gamma(b\sqrt{1+N/Z})}{2K_{\gamma+1}(b\sqrt{1+N/Z})} + \gamma + 1 \right]. \quad (3.24)$$

For $\gamma = -0.5$, we use the equality $K_{-\nu}(z) = K_\nu(z)$ to simplify this expression:

$$\alpha(2, N) = \alpha(1, N) \frac{1}{4} \frac{1}{Z/N + 1} [b\sqrt{1+N/Z} + 1]. \quad (3.25)$$

Note that by fixing $\gamma = -0.5$, the relative frequency spectrum is completely determined by expressions ((3.22) for the hapax legomena, (3.25) for the dis legomena, and the general recurrence relation (3.21)) from which the Bessel functions have been eliminated. Moreover, for the expected vocabulary size we can write, again using $K_{-0.5}(z) = (\pi/2z)^{1/2}e^{-z}$,

$$\begin{aligned}
E[V(N)] &= \frac{2Z}{b} \frac{K_{-1/2}(b)}{K_{1/2}(b)} \left[1 - \frac{K_{-1/2}(b\sqrt{1+N/Z})}{(1+N/Z)^{-1/4}K_{-1/2}(b)} \right] \\
&= \frac{2Z}{b} \left[1 - \frac{\left(\frac{\pi}{2b\sqrt{1+N/Z}}\right)^{1/2} e^{-b\sqrt{1+N/Z}}}{\sqrt{1+N/Z}^{-1/2} \left(\frac{\pi}{2b}\right)^{1/2} e^{-b}} \right] \\
&= \frac{2Z}{b} \left[1 - e^{b(1-\sqrt{1+N/Z})} \right]. \quad (3.26)
\end{aligned}$$

The expected frequency spectrum follows immediately, since

$$E[V(m, N)] = E[V(N)]\alpha(m, N).$$

Finally, the population number of types simplifies to $S = \frac{2Z}{b}$.

The parameters of Sichel's model are easily estimated for $\gamma = -0.5$ by requiring that $V(N) = E[V(N)]$ and $V(1, N) = E[V(1, N)]$. Let $g = (1+N/Z)^{-1/2}$ and let $h = b\sqrt{1+N/Z}$. We first observe that

$$\frac{N}{E[V(1, N)]} = \frac{N/E[V(N)]}{\alpha(1, N)} = \frac{1}{g} e^{h(1-g)}, \quad (3.27)$$

and

$$\frac{E[V(N)]}{E[V(1, N)]} = \frac{1}{\alpha(1, N)} = [e^{h(1-g)} - 1] \frac{2}{h(1-g^2)}.$$

Let $x = h(1-g) = \log((gN)/E[V(1, N)])$, and note that $x(1+g) = h(1-g^2)$. This leads to an expression in which g is the only unknown variable when we

use $V(1, N)$ as an estimate of $E[V(1, N)]$, and $V(N)$ as an estimate of $E[V(N)]$:

$$\begin{aligned} \frac{V(N)}{V(1, N)} &= [e^x - 1] \frac{2}{x(1+g)} \\ &= \left[\frac{gN}{V(1, N)} - 1 \right] \frac{2}{\log\left(\frac{gN}{V(1, N)}\right)(1+g)}. \end{aligned} \quad (3.28)$$

From (3.28) we can solve g by iteration, and hence Z and b .

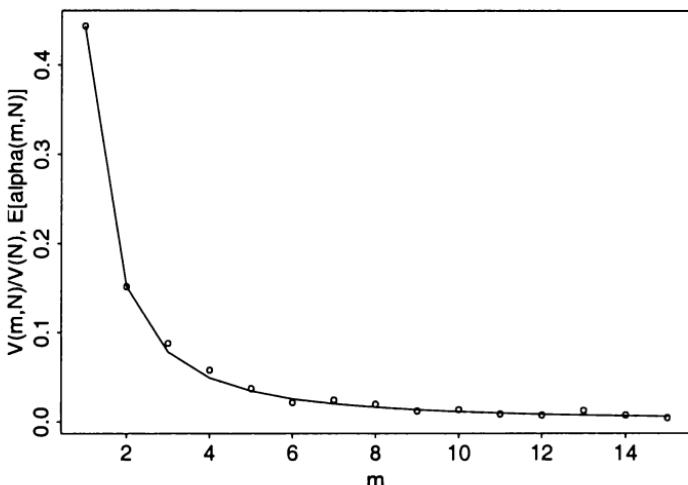


Figure 3.5: *Observed (circles) and expected (solid line) relative spectrum elements for Alice in Wonderland, using Sichel's model with $\gamma = -0.5$, $b = 0.0236$, and $Z = 105.78$.*

Figure 3.5 plots $V(m, N)/V(N)$ (using circles) and $\alpha(m, N)$ (using a solid line) for *Alice in Wonderland*, for the parameters $\gamma = -0.5$, $b = 0.0236$, and $Z = 105.78$. The fit is quite good for $m = 1$ and $m = 2$, but the model slightly underestimates the empirical values for $m = 3, 4, 5$. The number of types in the population estimated on the basis of this model equals 8948, which is somewhat lower than the estimate based on the lognormal model, 11889.

3.2.3 The Zipfian Family of LNRE Models

The LNRE models outlined in the preceding sections enriched the nonparametric expressions for the spectrum elements and the vocabulary size with parametric expressions for the structural type distribution $G(\pi)$. Unfortunately,

for Zipf's harmonic spectrum law

$$E[V(m, N)] = \frac{E[V(N)]}{m(m+1)}$$

and related models, no complete expression for the structural type distribution is available. Nevertheless, Orlov and Chitashvili (1983a,b) and Khmaladze and Chitashvili (1989) have shown that LNRE models can be defined by taking as one's point of departure the **relative spectrum elements** at the Zipf size Z :

$$\alpha(m, Z) = \frac{E[V(m, Z)]}{E[V(Z)]}.$$

Given a parametric expression for $\alpha(m, Z)$, we can express the expected spectrum elements at the Zipf size in the form

$$E[V(m, Z)] = \alpha(m, Z) E[V(Z)].$$

The expected spectrum elements for arbitrary sample size N can be obtained by interpolating and extrapolating from the Zipf size for which $\alpha(m, Z)$ holds, using (2.42) and (2.52) in the form

$$E[V(m, N)] = \sum_{k \geq m} E[V(k, Z)] \binom{k}{m} \left(\frac{N}{Z}\right)^m \left(1 - \frac{N}{Z}\right)^{k-m}$$

Given $E[V(m, N)]$, we derive $E[V(N)]$ by summation over m for $E[V(m, N)]$:

$$E[V(N)] = \sum_m E[V(m, N)].$$

The general expression for the relative spectrum elements introduced by Orlov and Chitashvili (1983a,b) for the Zipf size Z is

$$\alpha(m, Z, \alpha, \beta, \gamma) = \frac{\int_0^\infty \frac{(\log(1+x))^{\gamma-1} x^\alpha}{(1+x)^{m+1}(1+x)^\beta} dx}{\int_0^\infty \frac{(\log(1+x))^{\gamma-1} x^{\alpha-1}}{(1+x)^{\beta+1}} dx}. \quad (3.29)$$

This expression subsumes a family of models, most of which reduce to Zipf's law for specific choices of the parameters α, β , and γ , and defines the 'generalized Zipf's law':

1. Zipf (Zipf, 1935)

$$\alpha(m, Z, 1, 1, 1) = \frac{1}{m(m+1)} \quad (3.30)$$

2. Yule (Yule, 1924; Simon, 1955)

$$\alpha(m, Z, \beta, \beta, 1) = \frac{\Gamma(\beta+1)\Gamma(m)\beta}{\Gamma(m+\beta+1)} \quad (3.31)$$

3. Yule-Simon (Simon, 1960)

$$\alpha(m, Z, 1, \beta, 1) = \frac{\beta}{(m + \beta - 1)(m + \beta)}, \quad (\beta > 0), \quad (3.32)$$

4. Waring-Herdan-Muller (Herdan, 1960, 1964; Muller, 1979)

$$\alpha(m, Z, \alpha, \beta, 1) = \frac{\Gamma(\beta + 1)\alpha}{\Gamma(\beta + 1 - \alpha)} \cdot \frac{\Gamma(m + \beta - \alpha)}{\Gamma(m + \beta + 1)}, \quad (\alpha > 1, \beta \geq \alpha), \quad (3.33)$$

5. Karlin-Rouault (Rouault, 1978)

$$\alpha(m, Z, \alpha, 0, 1) = \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)}, \quad (0 < \alpha < 1), \quad (3.34)$$

6. Zipf-Mandelbrot (Mandelbrot, 1962)

$$\alpha(m, Z, 1, 1, \gamma) = \frac{1}{m^\gamma} - \frac{1}{(m + 1)^\gamma}, \quad (\gamma > 0). \quad (3.35)$$

Given (3.29) as a model for the relative spectrum elements at the Zipf size Z , the next question is how to obtain a more general expression for the relative spectrum elements for arbitrary sample size N . To answer this question, we combine the equality

$$E[V(m, Z)] = E[V(Z)]\alpha(m, Z, \alpha, \beta, \gamma)$$

with the expression for interpolation and extrapolation, (2.42), repeated here for convenience,

$$E[V_{N_0}(m, N)] = \sum_{k \geq m} V(k, N_0) \binom{k}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{k-m},$$

and we condition on the frequency spectrum at the Zipf size Z , substituting the expected spectrum values for the empirical ones:

$$\begin{aligned} E[V(m, N)] &= \sum_{k \geq m} E[V(k, Z)] \binom{k}{m} \left(\frac{N}{Z}\right)^m \left(1 - \frac{N}{Z}\right)^{k-m} \\ &= \sum_{k \geq m} E[V(Z)]\alpha(k, Z, \alpha, \beta, \gamma) \binom{k}{m} \left(\frac{N}{Z}\right)^m \left(1 - \frac{N}{Z}\right)^{k-m} \end{aligned}$$

For notational convenience, let

$$c = C(Z, \alpha, \beta, \gamma) = \frac{E[V(Z)]}{\int_0^\infty \frac{[\log(1+x)]^{\gamma-1} x^{\alpha-1}}{(1+x)^{\beta+1}} dx},$$

and write

$$E[V(m, N)] =$$

$$\begin{aligned}
&= \sum_{k \geq m} c \int_0^\infty \frac{(\log(1+x))^{\gamma-1} x^\alpha}{(1+x)^{k+1}(1+x)^\beta} dx \binom{k}{m} \left(\frac{N}{Z}\right)^m \left(1 - \frac{N}{Z}\right)^{k-m} \\
&= c \int_0^\infty \frac{(\log(1+x))^{\gamma-1} x^\alpha}{(1+x)^{\beta+1}} \sum_{k \geq m} \binom{k}{m} \left(\frac{N}{Z}\right)^m \left(1 - \frac{N}{Z}\right)^{k-m} \frac{1}{(1+x)^k} dx \\
&= c \left(\frac{N}{Z}\right)^m \int_0^\infty \frac{[\log(1+x)]^{\gamma-1} x^\alpha}{(N/Z + x)^{m+1}(1+x)^\beta} dx. \tag{3.36}
\end{aligned}$$

The last step makes use of the equality

$$\sum_{i=0}^{\infty} a^i \binom{m+i}{i} = \frac{1}{(1-a)^{m+1}},$$

as follows:

$$\begin{aligned}
&\sum_{k \geq m} \binom{k}{m} \left(\frac{N}{Z}\right)^m \left(1 - \frac{N}{Z}\right)^{k-m} \frac{1}{(1+x)^k} \\
&= \left(\frac{N}{Z}\right)^m \frac{1}{(1+x)^m} \sum_{k-m \geq 0} \frac{(1-N/Z)^{k-m}}{(1+x)^{k-m}} \binom{m+k-m}{k-m} \\
&= \left(\frac{N}{Z}\right)^m \frac{1}{(1+x)^m} \frac{1}{\left(1 - \frac{1-N/Z}{1+x}\right)^{m+1}} \\
&= \left(\frac{N}{Z}\right)^m \frac{(1+x)}{(N/Z + x)^{m+1}}.
\end{aligned}$$

Given (3.36), we can estimate the vocabulary size for arbitrary N :

$$\begin{aligned}
E[V(N)] &= \sum_{m=1}^{\infty} E[V(m, N)] \\
&= \sum_{m=1}^{\infty} C(Z, \alpha, \beta, \gamma) \left(\frac{N}{Z}\right)^m \int_0^\infty \frac{[\log(1+x)]^{\gamma-1} x^\alpha}{(N/Z + x)^{m+1}(1+x)^\beta} dx \\
&= C(Z, \alpha, \beta, \gamma) \left(\frac{N}{Z}\right) \int_0^\infty \frac{[\log(1+x)]^{\gamma-1} x^{\alpha-1}}{(N/Z + x)(1+x)^\beta} dx, \tag{3.37}
\end{aligned}$$

once again using the fact that $\sum_{i=0}^{\infty} a^i = 1/(1-a)$ and writing (for notational convenience) $t = N/Z$:

$$\begin{aligned}
\sum_{m=1}^{\infty} \frac{t^m}{(t+x)^{m+1}} &= \frac{t}{(t+x)^2} \sum_{m=1}^{\infty} \left(\frac{t}{(t+x)}\right)^{m-1} \\
&= \frac{t}{(t+x)^2} \sum_{i=0}^{\infty} \left(\frac{1}{1+x/t}\right)^i \\
&= \frac{t}{(t+x)x}.
\end{aligned}$$

To complete the exposition of the extended Zipfian family of LNRE models, we need a general expression for $E[V(Z)]$ for arbitrary α, β , and γ :

$$E[\widehat{V}(Z)] = Z \frac{\int_0^\infty \frac{[\log(1+x)]^{\gamma-1} x^{\alpha-1}}{(1+x)^{\beta+1}} dx}{\int_0^\infty \frac{[\log(1+x)]^{\gamma-1} x^{\alpha-2}}{(1+x)^{\beta+2p^*}} \left[(1+x)^{Zp^*} - 1 - \frac{Zp^*x}{1+x} \right] dx}. \quad (3.38)$$

Unfortunately, this general expression for $E[V(Z)]$ is computationally rather intractable. Moreover, estimating the parameters for the complete model is extremely difficult, due to the presence of many singularities in the parameter space. The main interest of this general model is theoretical. Fortunately, some of the more specific submodels of the general LNRE model are computationally tractable, and it is to these more specific models that we now turn.

Zipf

Zipf's harmonic spectrum law can be obtained from (3.29) by choosing $\alpha = \beta = \gamma = 1$, in which case, $\alpha(m, Z, \alpha, \beta, \gamma)$ reduces to the simplified expression

$$\alpha(m, Z, 1, 1, 1) = \frac{\int_0^\infty \frac{x}{(1+x)^{m+2}} dx}{\int_0^\infty \frac{1}{(1+x)^2} dx},$$

which can in turn be even further simplified by observing that the denominator equals unity,

$$\int_0^\infty \frac{1}{(1+x)^2} dx = \left[-\frac{1}{1+x} \right]_0^\infty = 1,$$

and by solving the numerator as follows:

$$\begin{aligned} \int_0^\infty \frac{x}{(1+x)^{m+2}} dx &= \int_0^\infty \frac{1+x}{(1+x)^{m+2}} dx - \int_0^\infty \frac{1}{(1+x)^{m+2}} dx \\ &= \left[-\frac{1}{m(1+x)^m} \right]_0^\infty + \left[-\frac{1}{(m+1)(1+x)^{m+1}} \right]_0^\infty \\ &= \frac{1}{m} + \frac{1}{m+1} \\ &= \frac{1}{m(m+1)}. \end{aligned}$$

Hence, for Zipf's law itself, $\alpha(m, Z, 1, 1, 1) = \frac{1}{m(m+1)}$.

In this form, Zipf's law describes the spectrum for one specific sample size Z . General expressions for the spectrum elements and the vocabulary size for arbitrary sample size N can be obtained as follows. Writing $t = N/Z$ and rewriting $x = y/t$, we have

$$\begin{aligned} E[V(m, N)] &= E[V(Z)]t^m \int_0^\infty \frac{y}{(t+y)^{m+1}(1+y)} dy \\ &= E[V(Z)]t^m \int_0^\infty \frac{tx}{(t+tx)^{m+1}(1+tx)} tdx \end{aligned}$$

$$= E[V(Z)]t \int_0^\infty \frac{x}{(1+x)^{m+1}(1+tx)} dx \quad (3.39)$$

$$\begin{aligned} E[V(N)] &= E[V(Z)]t \int_0^\infty \frac{1}{(t+y)(1+y)} dy \\ &= E[V(Z)]t \int_0^\infty \frac{1}{(t+tx)(1+tx)} tdx \\ &= E[V(Z)]t \int_0^\infty \frac{1}{(1+x)(1+tx)} dx \end{aligned} \quad (3.40)$$

(3.41)

This last integral can be simplified further. Writing $t = N/Z$, we have

$$\begin{aligned} &\int_0^\infty \frac{t}{(1+tx)(1+x)} dx \\ &= \int_0^\infty \left\{ \frac{t}{1+tx} - \frac{t}{t-1} \left[\frac{1}{1+x} - \frac{1}{1+tx} \right] \right\} dx \\ &= \int_0^\infty \frac{t}{1+tx} dx - \frac{t}{t-1} \left[\int_0^\infty \frac{1}{1+x} dx - \int_0^\infty \frac{1}{1+tx} dx \right] \\ &= \lim_{a \rightarrow \infty} \left[\log(1+tx) - \frac{t}{t-1} \log(1+x) + \frac{1}{t-1} \log(1+tx) \right]_0^a \\ &= \frac{t}{t-1} \lim_{a \rightarrow \infty} \left[\log \left(\frac{1+tx}{1+x} \right) \right]_0^a \\ &= \frac{t}{t-1} \lim_{a \rightarrow \infty} \left[\log \left(t + \frac{1-t}{1+x} \right) \right]_0^a \\ &= \frac{t}{t-1} \log(t). \end{aligned}$$

To complete the Zipfian expression for the vocabulary size, we need to know the value of $E[V(Z)]$. For the Zipf size Z , and writing m^* for the highest value of m realized for $N = Z$, we have

$$\begin{aligned} Z &= \sum_{m=1}^{m^*} m E[V(m, Z)] \\ &= \sum_{m=1}^{m^*} m \frac{E[V(Z)]}{m(m+1)} \\ &= E[V(Z)] \sum_{m=1}^{m^*} \frac{1}{m+1} \\ &\approx E[V(Z)] \log(m^*). \end{aligned}$$

If we allow ourselves to estimate m^* by Zp^* , with p^* the highest sample relative frequency at sample size N , that is, if we assume that the highest sample

relative frequency is constant and independent of N ,⁴ we obtain the following expression for $E[V_Z(N)]$:

$$E[V(N)] = \frac{Z}{\log(p^* Z)} \frac{N}{N - Z} \log(N/Z). \quad (3.42)$$

We can estimate the one free parameter of the extended Zipf's law Z by solving $E[V(N)] = V(N)$ by iteration. Note that (3.42) implies that for Zipf's harmonic distribution $S = \lim_{N \rightarrow \infty} E[V(N)]$ is infinite. I will refer to this LNRE modification of Zipf's law as the extended law of Zipf.

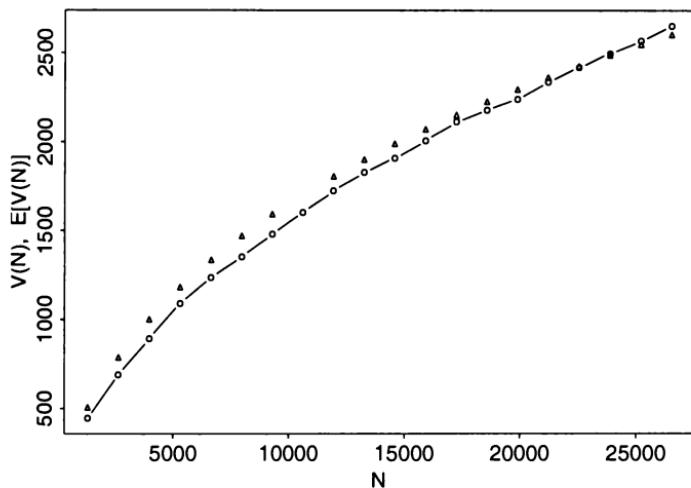


Figure 3.6: The empirical and expected growth curves of the vocabulary for Alice in Wonderland. The circles represent the observed vocabulary size $V(N)$, the triangles the expected vocabulary size $E[V(N)]$ using (3.42) for $Z = 10602$.

Figure 3.6 plots $E[V(N)]$ for *Alice in Wonderland* for $Z = 10602$, the sample size that appears as optimal in the plot of the relative mean squared error as a function of N (Figure 3.1). The pattern obtained is similar to that obtained for the lognormal and the inverse Gauss-Poisson models. The expected vocabulary size again overestimates the observed values for a wide range of sample sizes. Note, however, that for the highest values of N the theoretical vocabulary size slightly underestimates the empirical values. By choosing a slightly higher value for Z , 12225 — after all, the minimum relative mean squared error in Figure 3.6 appears in a shallow dip in the graph — the expected vo-

⁴As we have seen in Chapter 1, the highest-frequency words may show some systematic variation with N . As long as these fluctuations are small with respect to their overall magnitude, this variability need not be a serious problem for the estimation of p^* .

cabulary can be made to approximate the observed vocabulary more closely ($V(26505) = 2651$, $E[V_{12225}(26505)] = 2651.21$).

Recall that the LNRE zone was introduced as the range of sample sizes where the vocabulary size is still increasing, and where the values of the lowest spectrum elements are non-negligible. We have seen that *Alice in Wonderland* is located in the early LNRE zone: the numbers of hapax legomena, dis legomena, etc., are all increasing. How do the relative spectrum elements $\alpha(m, N)$ relate to the LNRE zone? And what is the relation of the Zipf size Z to the LNRE zone?

For Zipf's harmonic spectrum law, the Zipf size appears as the sample size for which the relative number of dis legomena $\alpha(2, N)$ reaches its maximum. To see this, we differentiate $\alpha(m, N)$ with respect to N :

$$\begin{aligned}\alpha(m, N) \frac{d}{dN} &= \frac{E[V(m, N)]}{E[V(N)]} \frac{d}{dN} \\ &= \frac{(E[V(N)] \frac{d}{dN}) E[V(m, N)] - (E[V(m, N)] \frac{d}{dN}) E[V(N)]}{E[V(N)]^2} \\ &= \frac{\{E[V(1, N)]/N\} E[V(m, N)]}{E[V(N)]^2} - \\ &\quad \frac{E[V(N)] \frac{1}{N} \{m E[V(m, N)] - (m+1) E[V(m+1, N)]\}}{E[V(N)]^2} \\ &= -\frac{1}{N} [\alpha(m, N)[m - \alpha(1, N)] - (m+1)\alpha(m+1)], \quad (3.43)\end{aligned}$$

using (2.20) and (2.21). Combining Zipf's law, $\alpha(m, Z) = \frac{1}{m(m+1)}$, with (3.43), it is easy to see that $\frac{d}{dN} \alpha(m, N) = 0$ for $m = 2$.

Figure 3.7 illustrates this relation between the Zipf size and $\alpha(2, N)$ for *Alice in Wonderland*. The dotted line represents the observed values for $\alpha(2, N)$. The solid line represents the value of $\alpha(2, N)$ averaged over 100 permutation runs of the text. The dashed line shows $\alpha(2, N)$ for $Z = 12225$ using (3.39). First observe that the theoretical curve of $\alpha(2, N)$ emerging from the randomization test has its maximum in the same range of sample sizes that also showed up in Figure 3.1 as the potential range for Z . The empirical values of $\alpha(2, N)$, however, depart from their expectation, notably so for the first two measurement points. This is probably due to the non-random use of words in the text.

The curve for Zipf's law yields expected values that are too large for the larger sample sizes. This is a consequence of the fact that the Zipfian expectations for $E[V(2, N)]$ are too high for large N . Even though Z was chosen to satisfy $E[V(26505)] = V(26505)$, the single parameter of this model (Z) does not allow us to simultaneously match for the number of hapax legomena ($V(1, 26505) = 1176$, $E[V(1, N)] = 1156.13$) and the number of dis legomena ($V(2, 26505) = 402$, $E[V(2, N)] = 433.23$). Below, it will become clear that enriching Zipf's harmonic law with the Yule-Simon parameter β allows the distributional structure of *Alice in Wonderland* to be captured more precisely.

The relative spectrum elements $\alpha(m, N)$ reach their maxima long before the absolute spectrum elements $E[V(m, N)]$. For samples outside the LNRE

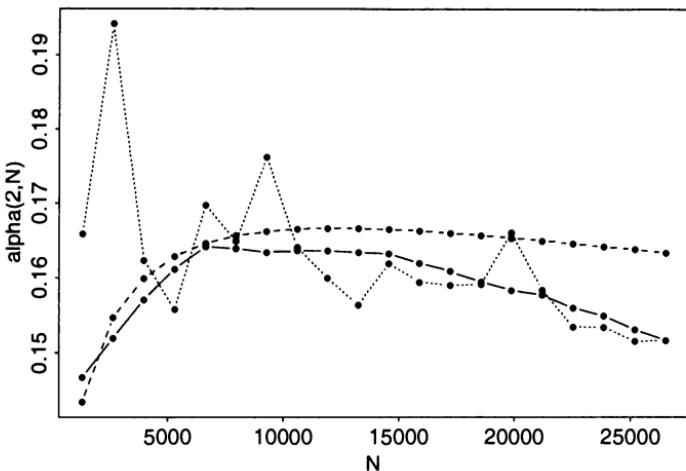


Figure 3.7: The relative frequency spectrum elements $\alpha(2, N)$ as a function of the sample size N for Alice in Wonderland. The dotted line represent the observed values, the solid line the expected values based on 100 permutation runs, and the dashed line the expected values for Zipf's law with $Z = 12225$.

zone, the graphs of $\alpha(m, N)$ appear as decreasing functions of N . Thus, the convex shape of the theoretical curves of $\alpha(2, N)$ for *Alice in Wonderland* show that this text is located in the LNRE zone. Texts for which the maximum of $\alpha(2, N)$ appears early on in sampling time are less centrally embedded in the early LNRE zone than texts for which Z is closer to or even greater than the text size itself. Since the vocabulary size increases with increasing Z , the Zipf size can be viewed both as a parameter of vocabulary richness, and as a parameter anchoring the text with respect to the LNRE zone.

Zipf-Mandelbrot

Zipf's harmonic distribution

$$\pi_z = \frac{C}{z}$$

has been enriched by Mandelbrot (1953, 1962) with two parameters,

$$\pi_z = \frac{C}{(z + b)^a}. \quad (3.44)$$

The parameter a modifies the slope of the rank-frequency graph in the double logarithmic plane. The parameter b is useful for introducing the downward

curvature observable in, for instance, the rank-frequency plot for *Alice's Adventures in Wonderland*, as shown in the upper left panel of Figure 3.8. When z is large, a (small) positive value of b hardly affects the predicted frequency. However, for small z , even a small value of b leads to lower predicted frequencies, leading to the kind of curvature shown by the dashed line.

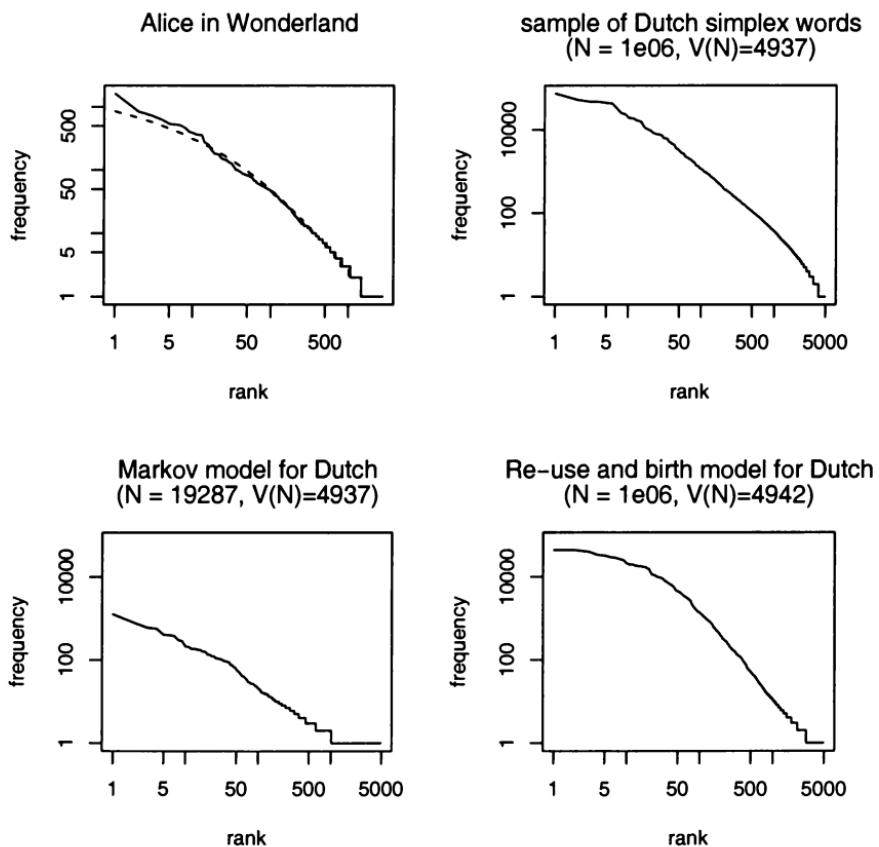


Figure 3.8: Rank-frequency curves for Alice's adventures in Wonderland (upper left panel), for a sample of 1 million monomorphemic Dutch words (upper right panel), for a word frequency distribution generated by a first order Markov model using the phoneme transition frequencies of monomorphemic Dutch words (lower left panel), and for a word frequency distribution generated by a re-use and birth stochastic process (lower right panel).

The Zipf-Mandelbrot rank-frequency relation (3.44) implies for the frequency spectrum that

$$V(m, N) = \sum_i I_{[f_i \geq m]} - \sum_i I_{[f_i \geq m+1]}$$

$$\begin{aligned}
 &= \left(\frac{C}{m} \right)^{\frac{1}{a}} - b - \left[\left(\frac{C}{(m+1)} \right)^{\frac{1}{a}} - b \right] \\
 &= C^{\frac{1}{a}} \left(\frac{1}{m^{1/a}} - \frac{1}{(m+1)^{1/a}} \right). \tag{3.45}
 \end{aligned}$$

Note that b no longer plays a role in the expression for the frequency spectrum. This is fine for the lower frequency ranks, where b has little to contribute, and where the rank-frequency relation is a step function rather than a continuous function. However, for the highest-frequency words, a continuous function seems more appropriate for the rank-frequency relation than a step function. This raises the question how accurate LNRE models might be with respect to the highest-frequency words. Firstly, models based on descriptions of the frequency spectrum such as the extended Zipf's law discussed in the previous section, and the extended Yule-Simon model to be discussed below, might be flawed because they build on expressions that do not take the highest-frequency words into account. Secondly, these models are typically fit to the head of the frequency spectrum, where we have small values of m and large values for $V(m)$, without regard for the shape of the spectrum for larger values of m .

Interestingly, the curvature observable in rank-frequency plots seems to arise, at least in part, as a result of a discretization problem. In the frequency spectrum, high-frequency words will often represent frequencies m that are attested only once ($V(m) = 1$). Generally, $E[V(m)]$ will be less than unity for such words. Here, the observed phenomenon is discrete and its theoretical expectation real-valued. Conversely, in the rank-frequency plane, a discrete theoretical notion, that of an integer-valued rank, is imposed on data for which it may not be appropriate. Evidence supporting this possibility is provided by the dashed line in the left panel of Figure 3.8. This line represents a GIGP fit to *Alice's Adventures in Wonderland*. The expected number of types was calculated for $m = 1, 2, \dots, 1624$, the frequency of the highest-frequency word in this book, *the*. The largest frequency m for which $E[V(m)] > 1.0$ is 70. Working backwards from $m = 1624$ to $m = 70$, the expected counts were cumulated and grouped into bins of 1.0. The first frequency m to yield a full bin was 851, the second 671, the third 570, etc. These bins were assigned the ranks 1, 2, 3, The resulting rank-frequency distribution, represented by the dashed line in Figure 3.8, shows that an LNRE model fitted to the counts of low-frequency types may nevertheless capture the main trend of the highest-frequency formations in the rank-frequency plane.

The appeal of the Zipf-Mandelbrot model lies in its rationale. Mandelbrot (1962) showed that the rank-frequency relation

$$f = \frac{K}{z^a} \tag{3.46}$$

can be understood as resulting from the optimization of the cost of coding on the one hand and maximization of the transmission of information. Consider the simple case in which there are M letters and a space character. Assume that the cost of coding the space is C_0 , and that the cost of coding a string with

n letters is n . There are M^n words with n letters, and $(M^n - 1)/(M - 1)$ words with less than n letters. When we rank the words according to increasing length, we have that the rank of the first word to have cost $C = n + C_0$ is

$$z = \frac{M^n - 1}{M - 1} + 1.$$

For large n ,

$$\begin{aligned}\log z &= \log \left(\frac{M^{n-1} - 1}{M - 1} + 1 \right) \\ &\approx \log(M^{C-C_0}) - \log(M - 1) \\ &\approx C \log M - C_0 \log M - \log(M - 1) \\ &\approx \log(M)C - L,\end{aligned}$$

implying that $\log z$ is approximately linear in the cost of coding C . Interestingly, the cost of coding is minimized when low-probability events have a high cost of coding and high-probability events have a low cost of coding. In information theory, the cost of coding is defined as $C = -2 \log \pi$, the number of bits required to encode a message with probability π . If we allow ourselves to use (3.46), we have that

$$\begin{aligned}C &= -\log \pi \\ &= -\log \left(\frac{K}{z^a} \right) \\ &= -\log K + a \log z,\end{aligned}$$

which again implies that $\log z$ is linear in the cost of coding C :

$$\log z = \frac{1}{a}C + \frac{\log K}{a}.$$

Since (3.46) implies that

$$\log z = -\frac{1}{a} \log f + \frac{\log K}{a},$$

we find that the slope a of the regression line in bi-logarithmic rank-frequency plots is inversely proportional to the string-productivity of a language. If a language has a large number of (equiprobable) letters, it will have many low-frequency words, other things being equal. Conversely, a language with a small alphabet will have a smaller number of possible words which are used more frequently. The latter type of language will have a steeper regression line than the former type of language.

Miller (1957) pointed out that the Zipf-Mandelbrot probabilities (3.44) are of the form predicted by Markov processes. Again, we only consider the proof for the simplest case in which all M letters are equiprobable. Let p denote the probability of a transition to the space character, and let $q = 1 - p$ denote the probability of a transition into any other letter. Since all M^n words of length

n are equiprobable, the probability of selecting a particular string i of length n equals

$$\begin{aligned} p_{(i,n)} &= \frac{pq^{n-1}}{M^n} \\ &= \frac{p}{q} e^{-n(\log M - \log q)}. \end{aligned}$$

Note that longer strings are less probable than shorter strings. As the number of words with a length not exceeding n equals

$$\sum_{i=1}^n M^i = \frac{M^{n+1} - M}{M - 1} \quad (M > 1),$$

and as the ranks of words of length n are numbered from $\frac{M^n - M}{M - 1} + 1$ to $\frac{M^{n+1} - M}{M - 1}$, the average rank $\bar{z}_{(n)}$ of a string of length n equals

$$\begin{aligned} \bar{z}_{(n)} &= \frac{\frac{M^n - M}{M - 1} + 1 + \frac{M^{n+1} - M}{M - 1}}{2} \\ &= M^n \frac{M + 1}{2(M - 1)} - \frac{M + 1}{2(M - 1)}, \end{aligned}$$

which is of the form $\bar{z}_{(n)} = M^n c + d$, hence $M^n = (\bar{z}_{(n)} - d)/c$. We now isolate M^n in $p_{(i,n)}$:

$$\begin{aligned} p_{(i,n)} &= \frac{p}{q} e^{-n(\log M - \log q)} \\ &= \frac{p}{q} e^{\log M^n[-(1-(\log q / \log M))]} \\ &= \frac{p}{q} \left(\frac{\bar{z}_{(n)} - d}{c} \right)^{-(1-(\log q / \log M))}, \end{aligned}$$

which, after reparameterization, emerges as being of the form

$$p_{(i,n)} = \frac{C}{(\bar{z}_{(n)} + b)^a}.$$

The right panel of Figure 3.8 shows the rank-frequency distribution generated by means of a first-order Markov model for Dutch. Instead of using equiprobable transitions, this model uses empirical phoneme transition probabilities calculated on the basis of the monomorphemic words (lemmas, excluding proper names) in the Dutch section of the CELEX lexical database. A random sample of 1 million monomorphemic Dutch words constitutes a baseline for comparison with a Markov-generated lexicon. The baseline sample, shown in the upper right panel of Figure 3.8, contains 4937 different word types. The first 4937 different types generated by the Markov process exhibit a roughly linear rank-frequency relation with a slight downward curvature for the lowest-valued ranks, as shown in the lower left panel.

A comparison of the lower left panel with the upper right panel shows that, the Markov process, while capturing the general shape of the rank-frequency relation, fails with respect to the type-token ratio. In the sample of Dutch, 4937 types account for 1 million tokens. In the Markovian approximation, the same number of types accounts for only 19287 tokens. The slope of the curve is smaller than the slope of the empirical curve, and the number of hapax legomena is far too large. Other lexical properties that correlate with word frequency, notably word length and lexical density, are captured only qualitatively but not quantitatively. This is illustrated in Figure 3.9.

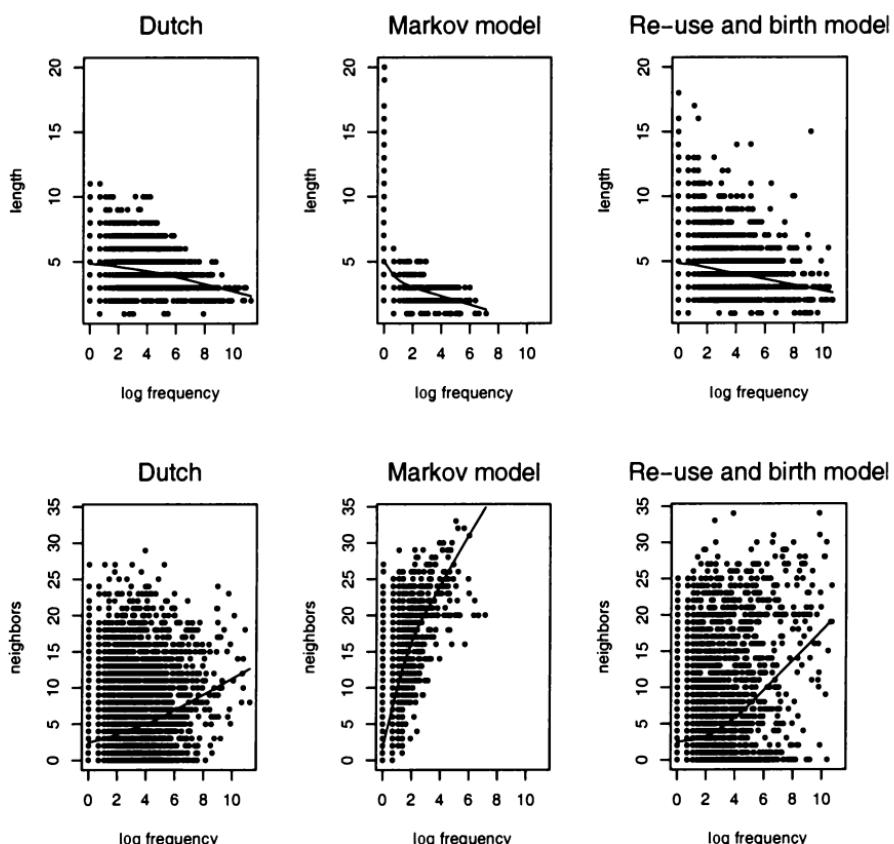


Figure 3.9: The relation between log frequency and word length (top row) and log frequency and lexical density (bottom row) for a sample of 1 million monomorphemic Dutch words (left column), a Markov approximation (middle column), and a Re-use and birth model (right column).

The upper panels of Figure 3.9 illustrate the relation between log frequency and word length. The Dutch sample of 1 million word tokens demonstrates

the negative correlation between word frequency and word length. The solid line is a non-parametric regression line. In the Markovian lexicon, we again observe the same correlation, but compared to the empirical data there is a shortage of high-frequency words and a large surplus of longer word lengths.

The lower panels of Figure 3.9 plot lexical density against log frequency. The lexical density of a word is the count of its lexical neighbors, the number of strings in the language at Hamming distance 1. For instance, *play*, *pram*, *tray*, and *prey* are lexical neighbors of *pray*. Lexical density has attracted considerable attention in psycholinguistics, as it appears to be a variable that affects the early stages of visual and auditory identification in language comprehension. The lower left panel of Figure 3.9 shows that the number of neighbors increases roughly linearly with log frequency. It also illustrates that the variance is huge. The next panel shows that the Markov approximation also captures the positive correlation of log frequency and density. This correlation is exaggerated, however, and the variance is too small compared to the empirical variance.

Summing up, a Markov process generates a lexicon that has the right kind of relations between frequency, word length, and lexical density. However, this lexicon is too sparsely populated: The token frequencies of the word types in the lexicon tend to be too low, and the word types are too similar to each other. Hence, there are no principled theoretical reasons to believe that the Zipf-Mandelbrot frequencies

$$f = \frac{C}{(z + b)^a} \quad (3.47)$$

must provide good fits to frequency distributions at a given Zipf size Z . Since the LNRE version of the Zipf-Mandelbrot law failed to provide accurate fits (Baayen, 1989), the details of the extended Zipf-Mandelbrot model need not concern us here. A more interesting model, both practically and theoretically, is the Yule-Simon model.

Yule-Simon

The Yule-Simon model replaces the expression for the relative spectrum elements at the Zipf size,

$$\alpha(m, Z, 1, 1, 1) = \frac{1}{m(m+1)}$$

with

$$\alpha(m, Z, 1, \beta, 1) = \frac{\beta}{(m + \beta - 1)(m + \beta)}. \quad (3.48)$$

This model for the relative frequency spectrum at the Zipf size Z instantiates a specific class of distributions of a more general model proposed in Simon (1960). Consider a stochastic process for generating words with the following properties. First, the probability of adding a new word to the vocabulary is constant throughout the generation process. Second, the probability that a

word with frequency m is re-used is proportional to $(m + c)V(m, N)$, with c specifying the 'birth rate'. Third, the rate d at which words drop out of use, the 'death rate', is proportional to $(m + d)V(m, N)$. Simon shows that these assumptions lead to a frequency spectrum of the form

$$\mathbb{E}[V(m, Z)] = A\lambda^m B(m + c, d - c + 1). \quad (3.49)$$

Here $B(z, w)$ denotes the Beta function

$$B(z, w) = \frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)},$$

with $\Gamma(x)$ denoting the Gamma function. The gamma function is related to the faculty function,

$$n! = \Gamma(n+1),$$

and it has the property that

$$\Gamma(x+1) = x\Gamma(x).$$

The expression of the spectrum (3.49) has the Yule-distribution

$$\mathbb{E}[V(m, Z)] = AB(m, \rho + 1)$$

as a special case ($\lambda = 1, c = 0, d = \rho$). The Yule-distribution in turn has Zipf's law as a special case for $\rho = 1$:

$$\begin{aligned} AB(m, 2) &= \frac{\Gamma(m)\Gamma(2)}{\Gamma(m+1)} \\ &= \frac{\Gamma(m)}{(m+1)\Gamma(m+1)} \\ &= \frac{\Gamma(m)}{(m+1)m\Gamma(m)} \\ &= \frac{A}{m(m+1)}, \end{aligned}$$

which is Zipf's law when $A = V(Z)$.

Returning to (3.49) and choosing $\lambda = 1$ and $d = c + 1$, we obtain

$$\begin{aligned} \mathbb{E}[V(m, N)] &= AB(m + d - 1, 2) \\ &= A \frac{\Gamma(m + d - 1)\Gamma(2)}{\Gamma(m + d + 1)} \\ &= A \frac{\Gamma(m + d - 1)}{(m + d)(m + d - 1)\Gamma(m + d - 1)} \\ &= \frac{A}{(m + d)(m + d - 1)}. \end{aligned}$$

Reparameterization ($\beta = d$, $\mathbb{E}[V(Z)]\beta = A$) immediately leads to (3.48), the form of the Yule-Simon model that in extended form has proved to be a useful LNRE model.

Before turning to the LNRE version of the Yule-Simon model, it is useful to consider in some more detail the rationale of this model. In the previous section, we saw that the Markovian rationale for the Zipf-Mandelbrot is unsatisfying with respect to magnitude of the correlations of frequency with length and lexical density, and with respect to the intensity with which words are used. A stochastic process along the lines described by Simon (1955, 1960), on the other hand, has nothing to say about the relation between frequency and length or frequency and lexical density. However, a stochastic process allows control of the birth rate, and hence offers the possibility of coming to grips with intensity of use.

Consider again Figure 3.8, but this time the lower right panel. The rank-frequency distribution in this panel is the result of a re-use and birth stochastic model. It is remarkably similar to the empirical distribution of Dutch monomorphemic words shown in the upper right panel. In contrast to the Markov-generated distribution in the lower left panel that we considered in the previous section, the numbers of tokens (1 million) and types (4937 and 4942 respectively) are well matched.

This distribution was generated by a stochastic process with the following properties. First, the probability of introducing a new type was set to 0.005. Second, the probability of re-using a word ω_i with frequency m , $\Pr(i, m)$, was a function of the joint probability $\Pr(m)$ of all $V(m)$ words with this frequency,

$$\Pr(m) = \frac{mV(m)}{\sum_m mV(m)}.$$

In Simon's original model, $\Pr(i, m) = \Pr(m)$. In the present model, however, it is set to

$$\Pr(i, m) = \Pr_{(e)}(m) = \frac{-\Pr(m) \log(\Pr(m))}{\sum_m -\Pr(m) \log(\Pr(m))}. \quad (3.50)$$

In other words, the probability of selecting a word belonging to frequency class m is proportional to the contribution of the tokens of this class to the by-class token entropy

$$H_{(m)} = \sum_m -\Pr(m) \log(\Pr(m)).$$

This rule of entropy-based proportionality avoids that the highest-frequency words receive excessively high token frequencies, and allows words in the intermediate frequency ranges a better chance of being re-used. It is a means to obtain a better trade-off between maximization of information transmission and optimization of the cost of coding information. In order to minimize the cost of coding of some word ω , $-\log(\Pr \omega)$ should be small. This is the case when ω is a high-frequency word. High-frequency words, however, have a low information load. It is the low-frequency words that have a high information load. To maximize information transmission, the lowest-frequency words should be re-used. Information transmission would in fact be maximal for a uniform distribution of the probabilities $\Pr(m)$. But then the entropy of the distribution would be maximal, implying the highest possible cost of coding.

The entropy-based selection rule (3.50) finds a balance between these opposing requirements. It does so by decreasing the likelihood of re-using high-frequency words and by slightly increasing this likelihood for low-frequency words. This is illustrated in the left panel of Figure 3.10 for the frequencies m in the sample of one million monomorphemic Dutch words underlying the upper right panel of Figure 3.8. The horizontal axis plots m , the vertical axis the selection probability. The solid line represents the selection probabilities $\text{Pr}(m)$. The dashed line represents the corresponding probabilities $\text{Pr}_{(e)}(m)$. We see that the highest-frequency words have a notably lower probability of being selected, and that the lower-frequency words have a slightly higher probability of being re-used.

Another way of looking at (3.50) is that it reduces the amount of information contributed by the higher-frequency words to the average information in the system. If we were to use selection by $\text{Pr}(m)$, the contribution of a frequency class m to the entropy would be $\text{Pr}_{(e)}(m)$, the dashed line in the right panel of Figure 3.10. When we use selection by $\text{Pr}_{(e)}(m)$, the contribution to the entropy becomes

$$\text{Pr}_{(e)(e)}(m) = \frac{-\text{Pr}_{(e)}(m) \log \text{Pr}_{(e)}(m)}{\sum_m -\text{Pr}_{(e)}(m) \log \text{Pr}_{(e)}(m)},$$

represented in the right panel of Figure 3.10 by a dotted line. For the higher frequencies, $\text{Pr}_{(e)(e)}(m) < \text{Pr}_{(e)}(m)$.

Interestingly, there is independent evidence supporting the lower information load of the higher-frequency words. It is well-known that higher-frequency words have more meanings or shades of meaning than is the case for lower-frequency words. A larger number of meanings implies increased dependency on context for interpretation. Hence, the amount of information contributed by such types out of context is reduced. This is implicitly modeled by means of the entropy-based selection rule. Note that this approach provides a semantic explanation for the downward curvature of the rank-frequency curve observable in the double logarithmic plane for the higher-frequency words, instead of the phonotactic explanation provided by a Markov process.

What a Simon-type rationale does not capture is the correlations of word frequency and word length and lexical density. A simple stochastic process has nothing to say about the shape of the words generated. Interestingly, the complementary advantages of the Mandelbrot and Simon approaches can be fruitfully combined by having a Markov model feed the re-use and birth process. Technically, this can be accomplished by having a Markov model generate a list of string types that preserves the order in which the types are generated. At each point in the re-use and birth process that a new word is called for, the next word from the list of Markov-generated types comes into play, strictly following the original order in the list. The right-hand panels of Figure 3.9 show that the frequency-length (upper panel) and frequency-density (lower panel) correlations bear a substantially improved resemblance to the empirical correlations in the leftmost column of this figure. We may conclude that the Yule-Simon model has a non-trivial rationale. Thanks to this ratio-

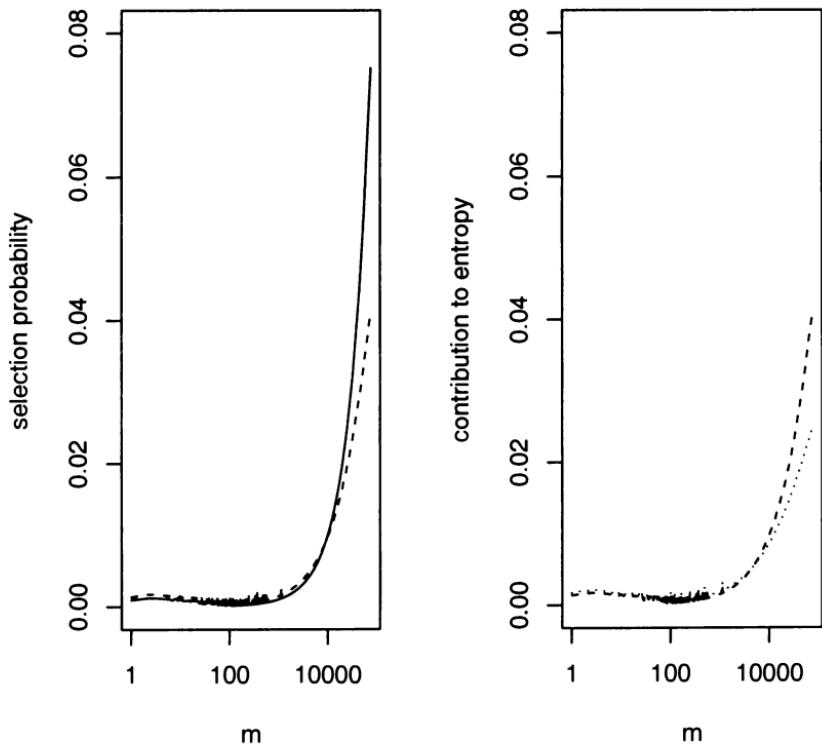


Figure 3.10: Selection probabilities by frequency class (left panel) for $\Pr(m)$ (solid line) and $\Pr_{(e)}(m)$ (dashed line), and the corresponding contributions to the average amount of information (right panel, dashed and dotted lines, respectively).

nale, its parameter β receives a direct interpretation as determining the type richness of the population.

In order for this model to become applicable to actual data sets, we need to consider its LNRE extension. The expression for the relative spectrum is obtained from the general definition of $\alpha(m, Z, \alpha, \beta, \gamma)$ presented above as (3.29) by replacing α and γ with 1,

$$\alpha(m, Z, 1, \beta, 1) = \frac{\int_0^\infty \frac{x}{(1+x)^{m+1+\beta}} dx}{\int_0^\infty \frac{1}{(1+x)^{\beta+1}} dx},$$

and by simplifying the denominator

$$\int_0^\infty \frac{1}{(1+x)^{\beta+1}} dx = \left[-\frac{1}{\beta} \cdot \frac{1}{(1+x)^\beta} \right]_0^\infty = \frac{1}{\beta}$$

as well as the numerator

$$\begin{aligned} \int_0^\infty \frac{x}{(1+x)^{m+1+\beta}} dx &= \int_0^\infty \frac{1+x}{(1+x)^{m+1+\beta}} - \int_0^\infty \frac{1}{(1+x)^{m+1+\beta}} \\ &= \left[-\frac{1}{m+\beta-1} \frac{1}{(1+x)^{m+\beta-1}} \right]_0^\infty - \left[-\frac{1}{m+\beta} \frac{1}{(1+x)^{m+\beta}} \right]_0^\infty \\ &= \frac{1}{m+\beta-1} - \frac{1}{m+\beta} \\ &= \frac{1}{(m+\beta-1)(m+\beta)}. \end{aligned}$$

We can estimate $E[V(Z)]$ using the general relations

$$E[V(m, Z)] = E[V(Z)]\alpha(m, Z)$$

and

$$N = \sum_m mV(m, N).$$

Writing m^* for the highest value of m realized at $N = Z$ we then have:

$$\begin{aligned} Z &= \sum_{m=1}^{m^*} mE[V(m, Z)] \\ &= \sum_{m=1}^{m^*} mE[V(Z)]\alpha(m, Z) \\ &= \sum_{m=1}^{m^*} E[V(Z)]\beta \sum_{m=1}^{m^*} \frac{m}{(m+\beta-1)(m+\beta)} \\ &= E[V(Z)]\beta \left(\sum_{m=1}^{m^*} \frac{\beta}{m+\beta} - \sum_{m=1}^{m^*} \frac{\beta-1}{m+\beta-1} \right) \\ &\approx E[V(Z)]\beta \log(m^*) \\ &\approx E[V(Z)]\beta \log(Zp^*), \end{aligned} \tag{3.51}$$

and hence,

$$E[V(Z)] = \frac{Z}{\beta \log(Zp^*)}. \quad (3.52)$$

When there are reasons to doubt the constancy of m^* , or when m^* is not available, $V(Z)$ or equivalently m^* can be introduced as a third free parameter of the model.

Combining these results with (3.36) and (3.37), we obtain the following expressions for the general vocabulary size $E[V(N)]$ and spectrum elements $E[V(m, N)]$:

$$E[V(N)] = E[V(Z)]\beta \frac{N}{Z} \int_0^\infty \frac{1}{(N/Z + x)(1+x)^\beta} dx, \quad (3.53)$$

$$E[V(m, N)] = E[V(Z)]\beta \left(\frac{N}{Z}\right)^m \int_0^\infty \frac{x}{(N/Z + x)^{m+1}(1+x)^\beta} dx. \quad (3.54)$$

Recall that for Zipf's law the population number of types S is infinite. For the Yule-Simon law, the magnitude of S depends on β , the parameter for the 'death rate' in Simon's stochastic model. The greater the death rate, the smaller the population number of types will be. Formally, with $t = N/Z$, we have:

$$\begin{aligned} S &= \lim_{N \rightarrow \infty} E[V(Z)]\beta \frac{N}{Z} \int_0^\infty \frac{1}{(N/Z + x)(1+x)^\beta} dx \\ &= \lim_{t \rightarrow \infty} E[V(Z)]\beta \int_0^\infty \frac{1}{(1+x)^\beta(1+x/t)} dx \\ &\approx \lim_{t \rightarrow \infty} E[V(Z)]\beta \int_0^\infty \frac{1}{(1+x)^\beta} dx \\ &= E[V(Z)]\beta \left[-\frac{1}{\beta-1} \frac{1}{(1+x)^{\beta-1}} \right]_0^\infty. \end{aligned}$$

For $\beta > 1$, this immediately leads to

$$S \approx E[V(Z)] \frac{\beta}{\beta-1}.$$

Note that for $\beta \downarrow 1$, the population number of types S approaches infinity. For $0 < \beta < 1$, we have

$$S \approx \left[\frac{(1+x)^{1-\beta}}{1-\beta} \right]_0^\infty = \infty.$$

In other words, for $\beta \gg 1$, the population size is finite, for β less than 1, the population size is infinite. In the latter case, Khmaladze's second definition (2.4) of an LNRE distribution is fully satisfied.

Figure 3.11 illustrates the advantage of adding the parameter β to the extended Zipf's law. The dots plot the first fifteen spectrum elements of the combined texts in the 'A' directory of the British National Corpus. The corresponding Zipfian expectations ($\hat{Z} = 191229.6$) are represented by a dotted

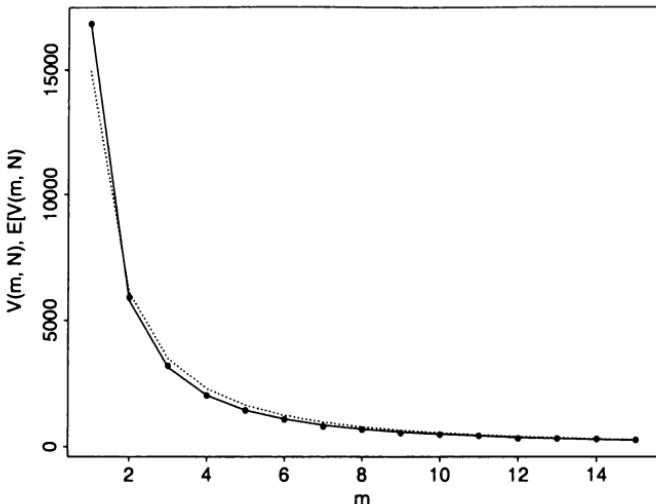


Figure 3.11: Observed and expected frequency spectrum for $m = 1, 2, \dots, 15$ of the 'A'-texts of the British National Corpus. The dots represent the observed frequency spectrum, the dotted line the corresponding expectations using the extended Zipf's law ($Z = 191229.6$), and the solid line the corresponding expectations using the Yule-Simon model ($Z = 67494.14$, $\beta = 0.72978$).

line. Zipf's law substantially underestimates the observed number of hapax legomena, and slightly overestimates the spectrum elements 3–7. With only one free parameter, Z , it cannot be adjusted without losing predictive accuracy with respect to the overall vocabulary size $V(N)$. A much improved fit is obtained with the Yule-Simon model for $\hat{Z} = 67494.14$ and $\hat{\beta} = 0.72978$.

Waring-Herdan-Muller

The Yule-Simon model is itself a special case of another distribution, the Waring distribution, that has been applied by Herdan (1964) and Muller (1977, 1979) to word frequency distributions. This distribution has the interesting property that it allows the population number of types to be estimated in a very simple way on the basis of Z , $V(Z)$, and $V(m, Z)$. Unfortunately, there is no computationally tractable LNRE version of the Waring-Herdan-Muller model available. Nevertheless, I discuss the model in some detail as it has enjoyed some popularity in the literature.

The Waring-Herdan-Muller model is based on the Waring distribution, a probability distribution that is derived from the series expansion of the ratio $1/(x - a)$. When we expand this ratio,

$$\begin{aligned}\frac{1}{x-a} &= \frac{1}{x} + \frac{a}{x} \cdot \frac{1}{(x+1)-(a+1)} \\ &= t_0 + R_0,\end{aligned}$$

the second fraction in the remainder R_0 is again of the form $1/(x - a)$. Expanding it leads to

$$\begin{aligned}\frac{1}{x-a} &= \frac{1}{x} + \frac{a}{x} \cdot \left(\frac{1}{x+1} + \frac{a+1}{x+1} \cdot \frac{1}{(x+2)-(a+2)} \right) \\ &= \frac{1}{x} + \frac{a}{x} \cdot \frac{1}{x+1} + \frac{a}{x} \cdot \frac{a+1}{x+1} \cdot \frac{1}{x-a} \\ &= t_0 + t_1 + R_1.\end{aligned}$$

Repeating the expansion of the rightmost fraction in the remainder results in the series

$$\frac{1}{x-a} = \frac{1}{x} + \sum_{i=1}^{\infty} \frac{1}{x} \prod_{j=1}^i \frac{a+j-1}{x+j} \quad (x-a > 1).$$

Multiplying both sides by $x - a$, we obtain a probability distribution:

$$\begin{aligned}\sum_{i=0}^{\infty} \pi_i &= \frac{x-a}{x} + \sum_{i=1}^{\infty} \frac{x-a}{x} \prod_{j=1}^i \frac{a+j-1}{x+j} \\ &= (x-a) \left\{ \frac{1}{x} + \frac{1}{x} \cdot \frac{a}{x+1} + \frac{1}{x} \cdot \frac{a}{x+1} \cdot \frac{a+1}{x+2} + \dots \right\} \\ &= (x-a)\{t_0 + t_1 + \dots + t_n + R_n\},\end{aligned}$$

with R_n the n -th remainder:

$$R_n = \frac{a}{x} \cdot \frac{a+1}{x+1} \cdot \frac{a+2}{x+2} \cdots \frac{a+n}{x+n} \cdot \frac{1}{x-a}.$$

Thus, the first two probabilities are $\pi_0 = \frac{x-a}{x}$ and $\pi_1 = \frac{x-a}{x} \frac{a}{x+1}$. Let M denote a random variable with values $m = 0, 1, 2, \dots$, and let the probability that M has the value m be defined by the Waring series, i.e.,

$$\Pr(M = m) = (x-a)t_m \quad (m = 0, 1, 2, \dots).$$

The expectation of M is

$$E[M] = \sum_{m=0}^{\infty} m\pi_m = (x-a) \sum_{m=0}^{\infty} mt_m,$$

and since

$$\begin{aligned}\sum_{m=0}^{\infty} mt_m &= t_1 + t_2 + t_3 + t_4 + \\ &\quad t_2 + t_3 + t_4 + \\ &\quad t_3 + t_4 + \\ &\quad t_4 + \\ &\quad +\end{aligned}$$

it follows that

$$\frac{E[M]}{x-a} = R_0 + R_1 + R_2 + \dots$$

$$\begin{aligned}
 &= \frac{1}{x-a} \left\{ \frac{a}{x} + \frac{a}{x} \cdot \frac{a+1}{x+1} + \frac{a}{x} \cdot \frac{a+1}{x+1} \cdot \frac{a+2}{x+2} + \dots \right\} \\
 &= \frac{a}{x-a} \left\{ \frac{1}{x} + \frac{1}{x} \cdot \frac{a+1}{x+1} + \frac{1}{x} \cdot \frac{a+1}{x+1} \cdot \frac{a+2}{x+2} + \dots \right\} \\
 &= \frac{a}{x-a} \cdot \frac{1}{x-(a+1)},
 \end{aligned}$$

so that

$$E[M] = \frac{a}{x-a-1}.$$

In the Waring-Herdan-Muller model, π_m is the probability of sampling a type that has the property of having frequency m , it being understood that all types are equiprobable irrespective of their frequency value. Hence, π_m can be defined as the proportion of types in the population that have frequency m :

$$\pi_m = \frac{E[V(m)]}{S}.$$

In order to estimate the parameters x and a , we consider the expected frequency

$$E[M] = \sum_m m\pi_m = \frac{N}{S} = \frac{a}{x-a-1}, \quad (3.55)$$

as well as its corresponding sample estimate

$$\hat{E}[M] = \sum_m mp_m = \frac{N}{E[V]},$$

with $p_m = E[V(m)]/E[V]$. Given these two expectations, we have that

$$\begin{aligned}
 E[M] &= \hat{E}[M] \frac{E[V]}{S} \\
 &= \hat{E}[M] \frac{S - E[V(0)]}{S} \\
 &= \hat{E}[M](1 - \pi_0) \\
 &= \hat{E}[M] \frac{a}{x},
 \end{aligned}$$

and hence,

$$S = E[V] \frac{x}{a}.$$

Let $q_i = 1 - p_i$, and consider p_1 :

$$\begin{aligned}
 p_1 &= \pi_1 \frac{S}{E[V]} \\
 &= \frac{x-a}{x} \cdot \frac{a}{x+1} \cdot \frac{x}{a} \\
 &= \frac{x-a}{x+1}.
 \end{aligned}$$

It follows that

$$x = \frac{a + p_1}{q_1}.$$

We use this to rewrite

$$\begin{aligned}\hat{E}[M] &= \frac{S}{E[V]} E[M] \\ &= \frac{x}{a} \cdot \frac{a}{x - a - 1} \\ &= \frac{a + p_1}{a + p_1 - q_1(a + 1)},\end{aligned}$$

from which we isolate a :

$$a = \frac{p_1 + m(q_1 - p_1)}{p_1 m - 1}.$$

Estimating p_1 and q_1 using N , V , and $V(1)$, we can then estimate a , x , and hence S . As

$$E[V(m)] = S\pi_m,$$

we obtain

$$E[V(m)] = E[V] \frac{x}{a} \cdot \frac{x - a}{x} \prod_{j=1}^m \frac{a + j - 1}{x + j}$$

as expression for the spectrum elements, and

$$\alpha(m) = \frac{E[V(m)]}{E[V]} = \frac{x}{a} \cdot \frac{x - a}{x} \prod_{j=1}^m \frac{a + j - 1}{x + j} \quad (3.56)$$

as expression of the relative spectrum elements. The expression for the relative spectrum elements given in (3.33),

$$\alpha(m, Z, \alpha, \beta, 1) = \frac{\Gamma(\beta + 1)\alpha}{\Gamma(\beta + 1 - \alpha)} \cdot \frac{\Gamma(m + \beta - \alpha)}{\Gamma(m + \beta + 1)}, \quad (\alpha > 1, \beta \geq \alpha),$$

is obtained by substituting $\beta = x$, $\beta - \alpha = a$, and by using the property of the Gamma function that $x\Gamma(x) = \Gamma(x + 1)$.

Note that according to the Waring-Herdan-Muller model the relative spectrum elements $\alpha(m)$ are independent of the sample size: N does not appear in (3.56). We know, however, that the relative spectrum changes with the sample size. The point in the derivation of the Waring-Herdan-Muller model where things go wrong from an LNRE perspective is (3.55), where the expected frequency $E[M]$ is first defined to equal N/S , where it is a function of N , and where it subsequently is defined to equal $a/(x - a - 1)$, which is independent of N . For (3.55) to be valid independently of the sample size, x , a , or both x and a should themselves be transformed into functions of the sample size.

Rouault

A final model that should be mentioned is interesting purely from a theoretical perspective. We have seen that the shape of the curve of the relative spectrum elements $\alpha(m, N) = V(m, N)/V(N)$ changes with N . Is this necessarily the case for all possible distributions? Rouault (1978) has shown that there is exactly one model for which $\alpha(m, N)$ does not change with N for $N \rightarrow \infty$. The limiting expression for $\alpha(m)$ according to this model is

$$\alpha(m) = \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)}.$$

Thus far, no word frequency distributions have been reported for which the relative spectrum has this particular shape.

3.3 Evaluating Goodness of Fit

SUMMARY This section outlines how the goodness of the fit of an LNRE model to an empirical distribution can be evaluated. Expressions for the variance of the vocabulary size and the spectrum elements are presented here.

Once the parameters of an LNRE model have been estimated, the question arises how well the model fits the data. Suppose we want know whether Sichel's model provides an adequate fit for *Alice in Wonderland*. Recall that for γ fixed at -0.5 we found that by choosing $b = 0.0236$ and $Z = 105.78$ we could satisfy the requirements $E[V(N)] = V(N)$ and $E[V(1, N)] = V(1, N)$. Figure 3.5 showed that, at least to the eye, the fit is quite good for $m = 2$ as well, but that the spectrum elements $m = 3$ and $m = 4$ are slightly underestimated. In order to evaluate how serious this underestimation is, we need some objective measure for the goodness of the fit.

Some authors (Sichel, 1975; Muller, 1979; Stein, Zucchini, and Juritz, 1987) evaluate the goodness of fit by means of a standard chi-squared test on the basis of the first r spectrum elements and their expectations, supplemented by a count of the cumulated spectrum elements $r + 1, r + 2, \dots$, henceforth denoted by $V(r+, N)$. Table 3.1 lists the relevant data with $r = 15$ for *Alice in Wonderland* using the inverse Gauss-Poisson, the lognormal, and the extended Zipf model. When we calculate the chi-square statistic

$$X^2 = \sum_{m=1}^{15} \frac{(V(m, N) - E[V(m, N)])^2}{E[V(m, N)]} + \frac{(V(15+, N) - E[V(15+, N)])^2}{E[V(15+, N)]}, \quad (3.57)$$

for Sichel's model (using columns 2 and 3 in Table 3.1), we find that $X_{(15)}^2 = 21.76$ and that $p = 0.1143$. This suggests that the fit provided by this model is acceptable, and that the somewhat lower expected values for $m = 3, 4$, for instance, are no source of worry.

Unfortunately, this use of the chi-square test is incorrect because the various spectrum elements have substantially different variances. The appropri-

Table 3.1: Observed and expected spectrum counts for Alice in Wonderland, with $r = 15$, using the inverse Gauss-Poisson model (Sichel) with $\hat{b} = 0.0236$, $\gamma = -0.5$, and $\hat{Z} = 105.78$, the lognormal model (Carroll) with $\hat{\mu} = -5.76$ and $\hat{\sigma} = 2.69$, and the extended Zipf model (Zipf) with $\hat{Z} = 12222$.

m	$V(m, N)$	E[V(m, N)]		
		Sichel	Carroll	Zipf
1	1176	1176.00	1173.12	1156.13
2	402	402.64	431.93	433.23
3	233	207.35	228.73	233.12
4	154	130.25	143.58	147.19
5	99	91.18	99.33	101.90
6	57	68.26	73.19	74.95
7	65	53.50	56.36	57.53
8	52	43.34	44.86	45.61
9	32	36.00	36.60	37.07
10	36	30.49	30.47	30.74
11	23	26.24	25.79	25.92
12	20	22.88	22.13	22.15
13	34	20.16	19.20	19.16
14	20	17.94	16.83	16.73
15	12	16.08	14.87	14.74
15+	236	308.70	234.01	235.06

ate chi-squared test takes these individual variances into account and proceeds as follows. We begin with constructing an error vector $e = o - \mu$,

$$e = o - \mu = \begin{pmatrix} V(N) \\ V(1, N) \\ V(2, N) \\ \vdots \\ V(r, N) \end{pmatrix} - \begin{pmatrix} E[V(N)] \\ E[V(1, N)] \\ E[V(2, N)] \\ \vdots \\ E[V(r, N)] \end{pmatrix},$$

taking into account both the vocabulary size $V(N)$ and the first r spectrum elements, 15 in our working example. Next, we construct a covariance matrix, $R(m, k, r)$, with the following elements:

$$R(m, k, r) = \begin{cases} \text{VAR}[V(N)] & \text{if } m = k = 0 \\ \text{COV}[V(N), V(k, N)] & \text{if } m = 0; k = 1, 2, \dots, r \\ \text{COV}[V(m, N), V(N)] & \text{if } m = 1, 2, \dots, r; k = 0 \\ \text{COV}[V(m, N), V(k, N)] & \text{if } m, k = 1, 2, \dots, r \end{cases} \quad (3.58)$$

If o is $\mathcal{N}(\mu, R)$ -distributed, then, given a free parameters,

$$X^2 = e^T R(m, k, r)^{-1} e \quad (3.59)$$

is $\chi^2_{(r+1-a)}$ -distributed (see, e.g., Morrison, 1971). For Sichel's model with $\gamma = -0.5$ fixed a priori, only b and Z are free to vary, hence $a = 2$ when we fit the inverse Gauss-Poisson model to the spectrum of *Alice in Wonderland*.

We calculate the elements of the covariance matrix in the following way. Our starting point is the equality

$$\text{COV}[X, Y] = E[XY] - E[X]E[Y],$$

and using

$$\text{COV}\left[\sum_i X_i, \sum_j Y_j\right] = \sum_i \sum_j \text{COV}[X_i, Y_j],$$

we can write

$$\begin{aligned} \text{COV}[V(m, N), V(k, N)] &= \text{COV}\left[\sum_{i=1}^S I_{\{f(i, N)=m\}}, \sum_{j=1}^S I_{\{f(j, N)=k\}}\right] \\ &= \sum_i \sum_j E[\{I_{\{f(i, N)=m\}}\} \cap \{I_{\{f(j, N)=k\}}\}] \\ &\quad - \sum_i \sum_j E[I_{\{f(i, N)=m\}}]E[I_{\{f(j, N)=k\}}]. \end{aligned} \quad (3.60)$$

Given (2.11), the second term of (3.60) is simply the product of $E[V(m, N)]$ and $E[V(k, N)]$. The first term concerns the expectation of the intersection of two events, the event that the frequency of ω_i equals m and the event that the frequency of ω_j equals k . First consider the case that $i = j$, the case that we are dealing with one word ω_i . Since one word can have only one frequency for a given sample size, the simultaneous occurrence of $f(i, N) = m$ and $f(j, N) = k$ for $i = j$ is possible only when $m = k$, and hence

$$\begin{aligned} E[\{I_{\{f(i, N)=m\}}\} \cap \{I_{\{f(j, N)=k\}}\}] &\stackrel{i=j}{=} I_{\{m=k\}} \sum_i \Pr(f(i, N) = m) \\ &= I_{\{m=k\}} E[V(m, N)]. \end{aligned}$$

Next consider the case that $i \neq j$, the situation in which we are dealing with two different words. The joint probability of words i and j having frequencies m and k is multinomially distributed:

$$\Pr(f(i, N) = m, f(j, N) = k) = \binom{N}{m+k} \binom{m+k}{m} \pi_i^m \pi_j^k (1 - \pi_i - \pi_j)^{N-m-k}.$$

It turns out that a simple expression for the covariance can nevertheless be obtained. Good and Toulmin (1956) present the approximation

$$\text{COV}[V(m, N), V(k, N)] = I_{\{m=k\}} E[V(m, N)] - \binom{m+k}{m} \frac{1}{2^{m+k}} E[V(m+k, 2N)], \quad (3.61)$$

which implies that we can effectively treat the events $\{f(i, N) = m\}$ and $\{f(j, N) = k\}$ as if they were independent:

$$\begin{aligned} \binom{m+k}{m} \frac{1}{2^{m+k}} \text{E}[V(m+k, 2N)] &= \sum_i \frac{(\pi_i 2N)^{m+k}}{(m+k)!} \binom{m+k}{m} \frac{1}{2^{m+k}} e^{-\pi_i 2N} \\ &= \text{E}[V(m, N)] \text{E}[V(k, N)] - \\ &\quad \sum_{i \neq j} \sum \frac{(\pi_i N)^m}{m!} \frac{(\pi_j N)^k}{k!} e^{-\pi_i N - \pi_j N} \end{aligned}$$

Since $\text{COV}[X, X] = \text{VAR}[X]$, (3.61) also specifies the variance of the spectrum elements. The variance of the vocabulary size is obtained on the basis of the relation between S , $V(0, N)$ and $V(N)$:

$$S = V(0, N) + V(N).$$

Since $\text{VAR}[aX + b] = a^2 \text{VAR}[X]$, we have that

$$\begin{aligned} \text{VAR}[V(N)] &= \text{VAR}[S - V(0, N)] \\ &= \text{VAR}[V(0, N)] \\ &= \text{E}[V(0, N)] - \text{E}[V(0, 2N)] \\ &= S - \text{E}[V(N)] - \{S - \text{E}[V(2N)]\} \\ &= \text{E}[V(2N)] - \text{E}[V(N)]. \end{aligned} \tag{3.62}$$

We use the equality

$$\text{COV}[aX + b, cY + d] = ac\text{COV}[X, Y]$$

to obtain the covariance of $V(N)$ and $V(m, N)$:

$$\begin{aligned} \text{COV}[V(m, N), V(N)] &= \text{COV}[V(m, N), S - V(0, N)] \\ &= -\text{COV}[V(m, N), V(0, N)] \\ &= \frac{1}{2^m} \text{E}[V(m, 2N)]. \end{aligned} \tag{3.63}$$

According to In 't Veld (1984:22–25), the approximations of Good and Toulmin (1956) are imprecise for small values of m and k . He proposes the following adjustments:

$$\begin{aligned} \text{VAR}[V(N)] &= \text{E}[V(2N)] - \text{E}[V(N)] - \frac{\text{E}[V(1, N)]^2}{N} \\ \text{VAR}[V(m, N)] &= \text{E}[V(m, N)] - \binom{2m}{m} \frac{1}{2^{2m}} \text{E}[V(2m, 2N)] - \\ &\quad \frac{1}{N} (m\text{E}[V(m, N)] - (m+1)\text{E}[V(m+1, N)])^2 \\ \text{COV}[V(m, N), V(N)] &= \frac{1}{2^m} \text{E}[V(m, 2N)] - \\ &\quad \frac{1}{N} \text{E}[V(1, N)] (\text{mE}[V(m, N)] - (m+1)\text{E}[V(m+1, N)]). \end{aligned}$$

Table 3.2: Overview of the statistics required for calculating $\mathbf{R}(m, k, 15)$ for Alice in Wonderland using Sichel's model with γ fixed at -0.5 a priori.

m	$V(m, N)$	$E[V(m, N)]$	m	$E[V(m, 2N)]$	m	$E[V(m, 2N)]$
1	1176	1176.00	1	1426.02	16	21.12
2	402	402.64	2	544.29	17	19.23
3	233	207.35	3	288.21	18	17.59
4	154	130.25	4	182.94	19	16.18
5	99	91.18	5	128.81	20	14.94
6	57	68.26	6	96.84	21	13.85
7	65	53.50	7	76.15	22	12.88
8	52	43.34	8	61.87	23	12.02
9	32	36.00	9	51.53	24	11.24
10	36	30.49	10	43.76	25	10.55
11	23	26.24	11	37.75	26	9.92
12	20	22.88	12	32.99	27	9.35
13	34	20.16	13	29.14	28	8.84
14	20	17.94	14	25.98	29	8.36
15	12	16.08	15	23.35	30	7.93

$$V(N) = 2651; E[V(N)] = 2651.0; E[V(2N)] = 3554.1.$$

In this book, we use (3.61), keeping in mind that this is an approximation that may slightly overestimate the variances and covariances.

Table 3.2 lists the data we need to calculate the covariance matrix

$$\hat{\mathbf{R}} = \begin{pmatrix} 903.10 & 713.01 & 136.07 & 36.03 & 11.43 & 4.03 & \cdots \\ 713.01 & 903.86 & -108.08 & -45.74 & -20.13 & -9.08 & \cdots \\ 136.07 & -108.08 & 333.40 & -40.25 & -22.70 & -12.49 & \cdots \\ 36.03 & -45.74 & -40.25 & 202.74 & -20.82 & -13.53 & \cdots \\ 11.43 & -20.13 & -22.70 & -20.82 & 137.08 & -12.68 & \cdots \\ 4.03 & -9.08 & -12.49 & -13.53 & -12.68 & 88.23 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

for Sichel's model with γ fixed at -0.5 a priori applied to *Alice in Wonderland* when, as before, we take the first 15 spectrum elements into account. Note that $\mathbf{R}(0, 0, 15) = \text{VAR}[V(N)]$, and that the other diagonal elements specify $\text{VAR}[V(1, N)]$, $\text{VAR}[V(2, N)]$, etc. Also note that the covariances of $V(N)$ and $V(m, N)$ are positive, and that the covariances of $V(m, N)$ and $V(k, N)$ are negative. Using (3.59), we find that $X_{(14)}^2 = 262.98$, $p = 0.0000$. Clearly, the fit is much worse than suggested by the inappropriate chi-square test (3.57).

3.4 Parameter estimation

SUMMARY This section addresses the question how to estimate the parameters of an

LNRE model for a given word frequency distribution.

There are no closed-form expressions for the parameters of the full generalized inverse Gauss-Poisson model with γ free, the Yule-Simon model, nor the lognormal model. Technically, this need not be a problem. Given a suitable cost function for goodness of fit, techniques are available for finding the parameter values for which this cost function is minimized. A technique that is especially useful is the downhill simplex method due to Nelder and Mead (1965). This function requires only the evaluation of the cost function and no calculation of derivatives. Given the complexity of the expressions of $E[V(N)]$ and $E[V(m, N)]$, this is a real advantage, even though the downhill simplex method is not the most efficient technique in terms of function calls. A disadvantage of the downhill simplex method is that it may find a local minimum instead of a global minimum depending on the initial starting values for the parameters with which it is initialized. Hence it may be necessary to run the procedure with different starting points in order to ascertain whether it will converge to the same, assumedly global, minimum.

Application of the downhill simplex method requires the formulation of a cost function. One possible approach is to require that the most remarkable characteristics of the spectrum, $N, V(N), V(1, N)$ and $V(2, N)$ should coincide with their expected values $E[V(N)], E[V(1, N)],$ and $E[V(2, N)]$. As we have seen, this method is used by Sichel (1986) for obtaining the parameters b and c of the generalized inverse Gauss-Poisson model when γ is fixed at -0.5 a priori, and it is also a convenient way of estimating the single parameter of the extended Zipf's law. In this approach, we define a cost function C_1 for a three-parameter model to be

$$C_1(2) = |E[V(N)] - V(N)| + |E[V(1, N)] - V(1, N)| + |E[V(2, N)] - V(2, N)|, \quad (3.64)$$

and for a two-parameter model to be

$$C_1(1) = |E[V(N)] - V(N)| + |E[V(1, N)] - V(1, N)|. \quad (3.65)$$

The advantages of this cost function are, first, the small number of expected values that have to be calculated, and, second, that the expected vocabulary size and the expected number of hapax legomena will be as similar as possible to the observed values. This is especially important when frequency distributions of novels are analyzed, for which the observed vocabulary size can arguably be regarded as a property of the text that has to be matched as closely as possible by an LNRE model. By requiring $E[V(N)]$ to equal $V(N)$, for instance, we can avoid a discontinuity between the observed growth curve of the vocabulary and the extrapolated growth curve.

When it is appropriate to regard a spectrum as being derived from a random sample, we can relax the strict requirements that come with cost function C_1 . Instead, we can consider a more general cost function that optimizes the fit using a larger part of the spectrum. Theoretically, one might consider a cost function that evaluates the fit in terms of the X^2 value for the first r ranks. Unfortunately, calculating X^2 is computationally very intensive as it requires

the computation of the first $2r$ expected spectrum elements at sample size $2N$. For practical applications, the use of X^2 in the cost function would render the simplex method too time costly.

A far less time-costly alternative is to formulate the cost function in terms of the mean squared error (MSE) for the first r ranks and including in the evaluation the match to $V(N)$ as well as the match to the sum of the remaining ranks, $V(r+, N) = V - \sum_{m=1}^r V(m, N)$:

$$\begin{aligned}\mathcal{C}_2(r) = & \frac{1}{r+2}(\{V(N) - E[V(N)]\}^2 \\ & + \sum_{m=1}^r \{V(m, N) - E[V(m, N)]\}^2 \\ & + \{V(r+, N) - E[V(r+, N)]\}^2).\end{aligned}\quad (3.66)$$

This cost function is a reasonably fast tool for optimizing the fit of an LNRE model.⁵

When estimating the parameters of Narayan and Balasubrahmanyam's implicit Zipfian model for the spectrum, special care is required, as the choice of parameters is not intrinsically constrained by the requirements

$$\begin{aligned}N &= \sum_m mE[V(m, N)] \\ V(N) &= \sum_m E[V(N)].\end{aligned}$$

The cost function $\mathcal{C}_2(r)$ should therefore be augmented with an extra term to keep the number of tokens balanced:

$$\begin{aligned}\mathcal{C}_3(r) = & \frac{1}{r+2}(\{V(N) - E[V(N)]\}^2 \\ & + \sum_{m=1}^r \{V(m, N) - E[V(m, N)]\}^2 \\ & + \{V(r+, N) - E[V(r+, N)]\}^2 + \\ & + \{N - \sum_m mE[V(m, N)]\}^2).\end{aligned}\quad (3.67)$$

3.5 A comparative study

SUMMARY In this section, the goodness of fit of the LNRE models is compared for various word frequency distributions. We shall see that the optimal model varies from text to text.

In this section we apply the LNRE models to the word frequency distributions of Carroll's *Alice's Adventures in Wonderland, Through the Looking-Glass*

⁵For the generalized inverse Gauss-Poisson model, expressions and algorithms for parameter estimation based on the log-likelihood function are available (see Stein, Zucchini, and Juritz, 1987, Heller, 1997, and Burrell and Fenton, 1993).

and what Alice found there, to Wells' *War of the Worlds*, to Conan-Doyle's *Hound of the Baskervilles*, and to the context-governed sub-corpus of the British National Corpus. Table 3.3 summarizes the estimated parameters as well as the goodness-of-fit statistics using cost function $C_1(r)$ with $r = 2$ for the three-parameter models, and $r = 1$ for the two-parameter models. Table 3.4 presents the corresponding results when the cost function $C_2(15)$ is used instead. It also tabulates the parameters and MSE for the Naranan-Balasubrahmanyam Zipfian spectrum fit, using cost function $C_2(15)$ (the fit labeled (1)) as well as cost function $C_3(15)$ (the fit labeled (2)).

First consider Table 3.3. In terms of the X^2 statistic and the associated p -values, the best fits for *Alice's Adventures in Wonderland* are obtained for the extended Zipf's law and the Yule-Simon model. Figure 3.12 plots the first 15 spectrum elements for the Yule-Simon fit and the generalized inverse Gauss-Poisson fit (left panel), as well as the interpolated growth curves of the vocabulary for these models together with the interpolated values obtained by means of binomial interpolation. Note that in both panels the solid line, representing the Yule-Simon fit, emerges as superior.

For *Through the looking-glass*, the extended Zipf's law provides a good fit, as does the lognormal model. In terms of the MSE, the latter provides an even better fit. Note that its chi-squared value is also slightly lower, but due to its greater number of parameters, its associated p -value is lower. For Wells' *War of the worlds*, the Yule-Simon model provides an excellent fit. No fully satisfactory fit emerges for Conan-Doyle's *Hound of the Baskervilles*. The Yule-Simon model and the lognormal model provide reasonable approximations. For the subcorpus of the British National Corpus ($N = 6154206$, $V(N) = 79883$), the chi-squared values as well as the MSE suggest that the fits are abysmal. Upon visual inspection, however, the fits emerge as not unreasonable. Consider Figure 3.13. The left panel shows the observed frequency spectrum for $m = 1 \dots 15$ using circles, as well as the Yule-Simon and generalized inverse Gauss-Poisson fits. The two fits are superimposed and indistinguishable in the plot, and provide a reasonable line through the observed data points. The right panel shows that the two fits both closely follow the interpolated values obtained by binomial interpolation (represented by triangles) as well. Given the rather large values of the spectrum elements involved, small divergences between the fit and the data easily give rise to high chi-squared values. This is a well-known problem of the chi-squared test (Grotjahn and Altmann, 1993): For large counts, even small errors easily lead to the rejection of reasonable models.

Thus far, we have fitted our models by requiring that the vocabulary size and the most salient spectrum elements be approximated as closely as possible by their expected values. Table 3.4 shows the effects of using cost function $C_2(15)$, which optimizes the parameters for a larger part of the spectrum. The model for which the use of this cost function leads to a considerable improvement is the generalized inverse Gauss-Poisson model. For all five word frequency distributions, both the MSE as well as the chi-squared value are substantially reduced. For Carroll's *Through the looking-glass*, this model now provides an acceptable fit. For the lognormal and the Yule-Simon models, using

Table 3.3: *Parameters and goodness-of-fit statistics for the lognormal model, the inverse Gauss-Poisson model, the extended Zipf's law, and the Yule-Simon model for selected texts using cost functions $C_1(3)$ and $C_1(2)$ (Alice: Alice in Wonderland; Through: Through the looking-glass; Wells: The War of the Worlds; Conan Doyle: Hound of the Baskervilles; BNC: the context-governed subcorpus of the British National Corpus).*

	Alice	Through	Wells	Conan Doyle	BNC
lognormal model					
\bar{Z}	317.34	347.74	724.05	431.29	337.84
$\hat{\mu}$	-5.76	-5.85	-6.58	-6.07	-5.82
$\hat{\sigma}$	2.69	2.68	2.99	3.067	3.890
X^2	34.89	21.58	50.39	33.84	4424.22
df	14	14	14	14	14
p	0.0015	0.0877	5.3E-6	0.0022	0
MSE	100.43	53.62	148.94	106.05	52023.2
inverse Gauss-Poisson, $\gamma = -0.5$					
\bar{Z}	105.78	115.33	265.00	176.11	402.79
\hat{b}	0.02364	0.02459	0.00682	0.00829	0.00359
X^2	262.98	302.65	1763.65	1322.82	567.96
df	14	14	14	14	14
p	0	0	0	0	0
MSE	421.42	447.87	5672.58	3341.28	35770.77
inverse Gauss-Poisson, γ free					
\bar{Z}	105.34	92.90	99.22	176.16	466.77
\hat{b}	0.0242	0.02183	0.0198	0.0074	0.00349
$\hat{\gamma}$	-0.501	-0.54	-0.6675	-0.5003	-0.4863
X^2	262.06	155.96	364.38	1317.06	1407.62
df	13	13	13	13	13
p	0	0	0	0.0000	0
MSE	416.52	243.93	1108.5	3366.21	73032.73
extended Zipf					
\bar{Z}	12222.19	12919.86	60741.1	31711.54	210933.5
X^2	29.05	22.22	19.74	227.84	121917.9
df	15	15	15	15	15
p	0.0158	0.1021	0.1822	0	0
MSE	126.22	75.27	363.47	3925.04	11059532
Yule-Simon					
\bar{Z}	8738.38	10765.20	38816.97	16271.02	7677.38
$\hat{\beta}$	0.89529	0.8097	0.85070	0.7335	0.5141
V_Z	1547.89	1856.07	5666.9	2880.67	2365.96
X^2	26.60	56.92	11.98	28.58	173.29
df	13	13	13	13	13
p	0.0141	0.0000	0.53	0.0075	0.0000
MSE	77.69	121.23	66.66	89.06	18034.76

$\mathcal{C}_2(r)$ instead of \mathcal{C}_1 sometimes leads to a small improvement (e.g., in the case of the lognormal model for *Alice's Adventures in Wonderland*), but it may lead to inferior fits as well (e.g., in the case of the lognormal model for *Through the Looking-glass*).

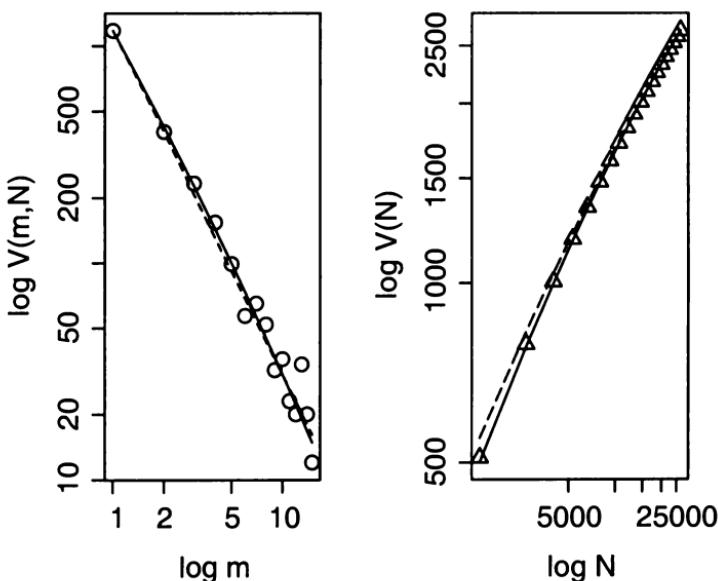


Figure 3.12: The first 15 spectrum elements on a double logarithmic scale (left) and the growth curve of the vocabulary, again using a double logarithmic scale, for Alice's adventures in wonderland. Open dots denote the observed spectrum elements, the solid line in the spectrum plot represents the Yule-Simon fit, and the dotted line the generalized inverse Gauss-Poisson fit. In the right panel, the triangles represent bimodal interpolation, the solid line again represents the Yule-Simon fit, and the dotted line represents the generalized inverse Gauss-Poisson fit.

Selecting the best fit by either of the two cost functions, we find that the Yule-Simon model is the best choice for Carroll's *Alice's Adventures in Wonderland* and Wells' *War of the Worlds*, that the extended Zipf's law outperforms the other models for Carroll's *Through the Looking-glass*, that the lognormal model is superior for Conan-Doyle's *Hound of the Baskervilles*, and that the general-

Table 3.4: Parameters and goodness-of-fit statistics for the lognormal model, the inverse Gauss-Poisson model, the extended Zipf's law, and the Yule-Simon model for selected texts using cost function $C_2(15)$ (Alice: *Alice in Wonderland*; Through: *Through the looking-glass*; Wells: *The War of the Worlds*; Conan Doyle: *Hound of the Baskervilles*; BNC: the context-governed subcorpus of the British National Corpus).

	Alice	Through	Wells	Conan Doyle	BNC
lognormal model					
\hat{Z}	320.14	304.01	722.20	425.20	365.57
$\hat{\sigma}$	2.69	2.92	2.99	3.09	3.852
X^2	32.38	83.30	48.97	26.67	2071.45
df	14	14	14	14	14
p	0.0035	0.0000	0.0000	0.0085	0
MSE	95.31	148.62	148.49	109.00	42494.78
inverse Gauss-Poisson, γ free					
\hat{Z}	59.18	52.31	47.48	40.33	325.90
\hat{b}	0.0294	0.02121	0.018	0.0155	0.00357
$\hat{\gamma}$	-0.609	-0.63	-0.75	-0.7062	-0.5179
X^2	77.56	24.61	78.51	92.39	201.78
df	13	13	13	13	13
p	0	0.026	0.0000	0.0000	0
MSE	128.98	129.54	574.47	519.22	16750.23
Yule-Simon					
\hat{Z}	8628.02	9664.41	32644.37	4475.89	6672.59
$\hat{\beta}$	0.8975	0.794	0.8113	0.6056	0.5087
\hat{V}_Z	1537.18	1751.67	5154.37	1373.80	2192.42
X^2	25.76	39.07	16.40	460.93	115.43
df	13	13	13	13	13
p	0.0183	0.00019	0.23	0	0.0000
MSE	74.14	97.41	76.38	1169.45	37325.44
Naranan-Balasubrahmanyam (1)					
\hat{C}	2276.57	2193.82	5773.24	3149.91	38494.83
$\hat{\mu}$	0.66	0.38	0.47	0.05	0.05
$\hat{\gamma}$	1.88	1.85	1.95	1.82	1.69
MSE	172.46	166.10	214.86	2526.28	2796649.60
Naranan-Balasubrahmanyam (2)					
\hat{C}	1978.47	1835.29	5361.92	3886.49	36251.93
$\hat{\mu}$	0.52	0.21	0.39	0.31	0.12
$\hat{\gamma}$	1.82	1.78	1.92	1.85	1.57
MSE	124.33	72.34	151.43	192.93	2453.72
N%	15.75	20.08	9.76	4.79	192.34

ized inverse Gauss-Poisson model is preferable for the subcorpus of the British National Corpus. It is surprising that the extended Zipf's law and the lognormal models perform so well, even though they have only one and two free parameters respectively.

Table 3.4 also summarizes the Naranan-Balasubrahmanyam Zipfian fits. The first set of fits, labeled (1), is obtained by imposing the restriction that $N = \sum_m mE[V(m, N)]$ (cost function \mathcal{C}_3). When we relax this restriction and use cost function \mathcal{C}_2 instead, we obtain the results tabulated under (2). The MSE values in Table 3.4 show that the fits with \mathcal{C}_3 are not as good as the fits with \mathcal{C}_2 . The latter fits are much superior in terms of the MSE compared to any of the other models for all texts, but, unfortunately, this is accomplished at the price of a substantial discrepancy between N and $N = \sum_m mE[V(m, N)]$. The row labeled $N\%$ lists the percentage of spurious tokens required to fit the spectrum:

$$N\% = \frac{\sum_m mE[V(m, N)] - N}{N}.$$

This percentage varies from roughly 5% to nearly 200%, which suggests that the goodness-of-fit for the low-frequency part of the spectrum is counterbalanced by a lack of goodness-of-fit for the higher-frequency ranks.

As a final example, consider the word frequency distribution of the Estonian novel *Truth and Justice* by A.H.Tammsaare discussed in Tuldava (1996). Tuldava claims, using visual inspection of a quantile-quantile plot such as presented in Figure 3.3 for *Alice in Wonderland* that graphical inspection 'has shown indisputable accordance of the empirical and the lognormal distribution.' (Tuldava, 1996: 48). When we analyze this word frequency distribution more precisely, the model that provides the optimal fit both in terms of the MSE as well as in terms of the chi-squared test is the generalized inverse Gauss-Poisson model ($MSE = 350.85$, $X^2_{(13)} = 40.12$, $p = 0.00013$, using cost function $\mathcal{C}_2(20)$). The lognormal model is much less satisfactory with $X^2(14)X2(14) = 420.46$, $p < 0.1e-12$ and $MSE = 1766.24$. Although the lognormal model is, as we have seen, a reasonable choice for some word frequency distributions, it is not the optimal choice for this Estonian text. The Sichel distribution provides a fit that, although not perfect, appears to be a more reliable choice. It should be kept in mind, however, that there is considerable variation in the development of the initial spectrum elements in this text, as shown in Figure 3.14. Given this not inconsiderable variation, the lognormal model does not appear to be completely unreasonable, but using objective criteria such as the MSE and the chi-squared test, the generalized inverse Gauss-Poisson model is to be preferred.

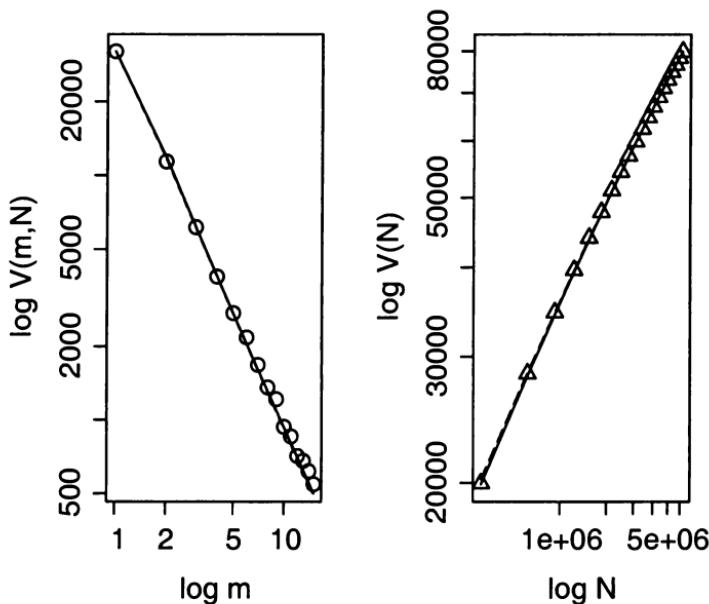


Figure 3.13: The first 15 spectrum elements on a double logarithmic scale (left) and the growth curve of the vocabulary, again using a double logarithmic scale, for the subcorpus of context-governed spoken English in the BNC. Open dots denote the observed spectrum elements, the solid line in the spectrum plot represents the Yule-Simon fit, and the dotted line the generalized inverse Gauss-Poisson fit. In the right panel, the triangles represent binomial interpolation, the solid line again represents the Yule-Simon fit, and the dotted line represents the generalized inverse Gauss-Poisson fit.

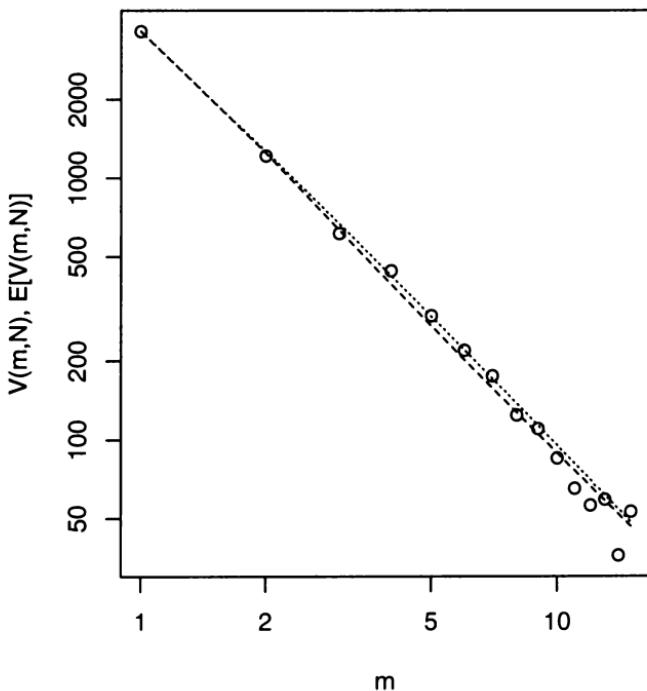


Figure 3.14: The first 15 spectrum elements on a double logarithmic scale for Tamm-saare's Truth and Justice (data from Tuldava, 1996). Observed spectrum elements are represented by circles, the generalized inverse Gauss-Poisson fit is represented by a dashed line, and the lognormal fit by a dotted line.

3.6 Comparing Lexical Measures Across Texts

SUMMARY Tests for comparing word frequency distributions with respect to vocabulary size and growth rate are presented in this section.

In Chapter 1, we compared *Alice in Wonderland* and *Through the looking-glass* and observed that for $N = 26505$ the vocabulary size of the latter exceeded that of the former by some 80 word types. Given the expressions for expectations and variances that we now have, we can test whether this difference is significant.

A general test that is useful here concerns the comparison of two normal random variables, X with mean μ_1 and variance σ_1^2 and Y with mean μ_2 and variance σ_2^2 . Under the Null-hypothesis that the two means do not differ significantly, the test statistic

$$Z = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad (3.68)$$

is $\mathcal{N}(0, 1)$ distributed. In order to apply this test to the comparison of vocabulary richness in *Alice in Wonderland* and *Through the looking-glass*, we first have to fit an LNRE model to the two texts in order to obtain estimates of the variance of $V(N)$. Using Sichel' model, we find that $\text{VAR}[V(N)] = 903.10$ for *Alice in Wonderland*. When we fit Sichel's model to the first 26505 words of *Through the looking-glass* ($X_{(9)}^2 = 45.83, p = 0.0000006$, again a less than optimal fit), we find that for this text $\text{VAR}[V(N)] = 956.60$. Assuming that $V(N)$ is a normally distributed random variable, we have that

$$\begin{aligned} Z &= \frac{E[V_1(N)] - E[V_2(N)]}{\sqrt{\text{VAR}[V_1(N)] + \text{VAR}[V_2(N)]}} \\ &= \frac{2731 - 2651}{\sqrt{956.60 + 903.10}} = 1.8551. \end{aligned} \quad (3.69)$$

Since $\Pr(|Z| > 1.8551) = 0.0636$, we conclude that the observed difference in vocabulary size is marginally significant.

The comparison of spectrum elements can likewise proceed on the basis of (3.68). However, special care is required for the growth rate of the vocabulary

$$\mathcal{P}(N) = \frac{E[V(1, N)]}{N}$$

has to be compared across texts of different lengths. For $N_1 \neq N_2$, the appropriate statistic is

$$Z = \frac{\frac{1}{N_1} E[V_1(1, N_1)] - \frac{1}{N_2} E[V_2(1, N_2)]}{\sqrt{\frac{1}{N_1^2} \text{VAR}[V_1(1, N_1)] + \frac{1}{N_2^2} \text{VAR}[V_2(1, N_2)]}}. \quad (3.70)$$

3.7 Discussion

This chapter discusses three LNRE models for word frequency distributions, Carroll's lognormal model, Sichel's generalized inverse Gauss-Poisson model,

and Orlov and Chitashvili's extended generalized Zipf's law. Sichel's model is computationally relatively simple, and requires little execution time. It is not always the best choice, however. Although computationally much more intensive, the lognormal model and the Yule-Simon model often provide excellent fits as well.

For the evaluation of the goodness of fit of an LNRE model, a multivariate chi-squared test can be carried out. It is generally useful to consider the χ^2 value in conjunction with the mean squared error. A good fit to the frequency spectrum, moreover, will also provide interpolated values for the vocabulary size and the spectrum elements that are very similar to those of binomial interpolation. When binomial interpolation and LNRE interpolation diverge, this is a sure sign that the fit is not optimal. For large samples, comparison with binomial interpolation is a means for checking whether the model is reasonable even when the chi-squared test suggests the model should be rejected.

3.8 Bibliographical Comments

For the lognormal model, see Carroll (1967, 1969) and Herdan (1960, 1964). The LNRE extension of Zipf's law is developed in Khmaladze and Chitashvili (1989). For the Yule-Simon model and its rationale, see Simon (1955, 1960, 1961) and also Lánský and Radil-Weiss (1980) and Chen and Leimkuhler (1989). Rationales for Zipf's law with special attention for the high-frequency tail of the frequency spectrum can be found in Mandelbrot (1953, 1959) and Miller (1957). Baayen (1991) describes the Mandelbrot-Simon model using a modified second-order Markov model. The relation between number of meanings and word frequency is studied in Reder, Anderson, and Bjork (1974), Paivio, Yuille, and Madigan (1968), Koehler (1986), and Hay (2000). The generalized inverse Gauss-Poisson model is described and applied to various kinds of data sets in Sichel (1975, 1986, 1997), Heller (1997), Stein, Zucchini, and Juritz (1987), Burrell and Fenton (1993), Price (1997), and Atkinson and Yeh (1982). Reviews of word frequency models are Baayen (1993) and Chitashvili and Baayen (1993). A general review article on estimating the number of species is Bunge and Fitzpatrick (1993). Further technical details on the statistics for the multivariate chi-square measure used in section 3.3 can be found in Morrison (1971).

3.9 Questions

1. Show that for $N = Z$ the growth rate $E[V(1, Z)]/Z$ is twice the type-token ratio $E[V(Z)]/N$ for Zipf's law and $1 + \beta$ times the type-token ratio for the Yule-Simon model.
2. Show that the expectation of a lognormal random variable equals $e^{\mu + \frac{1}{2}\sigma^2}$.
3. The probability density function of a Gamma-(λ, s) distributed random

variable X is

$$f(x) = \frac{\lambda e^{-\lambda x} (\lambda x)^{s-1}}{\Gamma(s)},$$

and its expectation is $E[X] = \frac{s}{\lambda}$. For $0 < s < 1$, the Gamma distribution is highly skewed to the left, which suggests it might serve as structural type distribution for an LNRE model. Prove the following equations

$$\begin{aligned} E[V(N)] &= \frac{\lambda}{s} \left(1 - \left(\frac{\lambda}{\lambda + N}\right)^s\right) \\ E[V(m, N)] &= \frac{\lambda}{s} \frac{\Gamma(m+s)}{\Gamma(m+1)} \left(\frac{N}{N+\lambda}\right)^m \left(\frac{\lambda}{N+\lambda}\right)^s \\ N_1^* &= S \end{aligned}$$

and explain why this model is too restricted for modeling most word frequency distributions.

4. Show that the denominator of (3.29) is the sum with respect to m of the numerator.

Chapter 4

Mixture distributions

4.1 Introduction

There are many ways in which texts are non-homogeneous. In the next chapter, we will see that the topical development in discourse may require adjustments to the standard LNRE models. Non-randomness in the way texts unfold through ‘text time’ N introduces a kind of non-homogeneity that is at odds with the basic assumptions underlying the urn model on which LNRE models are based. However, there are other kinds of non-homogeneity in texts that we have thus far not considered at all. Corpora are collections of texts, generally from different authors and covering a wide range of topics. Differences in register and authorial structure may make it impossible to fit a simple LNRE model to corpus-derived word frequency distributions. Even texts that are non-composite with respect to authorship, style, and register are nevertheless non-homogeneous in the sense that they are composed of words that differ widely with respect to their internal structure and quantitative properties. Some words, e.g., *dog*, *tree*, *write*, have no internal structure at all. Most words have some kind of internal structure, e.g., *dogs* consists of the base word *dog* and the plural suffix *-s*, and *unwillingness* has a layered structure that starts with the verb *will* and successively adds the affixes *-ing*, *un-*, and *-ness*:

<i>will</i>	(verb)
<i>willing</i>	(gerund in <i>-ing</i>)
<i>unwilling</i>	(adjective in <i>un-</i>)
<i>unwillingness</i>	(abstract noun in <i>-ness</i>)

Figures 4.1 and 4.2 illustrate how different the quantitative properties of morphologically defined subsets of words can be. Both figures are based the frequency lists in the CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993), which for Dutch is based on a corpus of 42 million words. Figure 4.1 presents a series of summary plots for the monomorphemic nouns in this corpus, and Figure 4.2 plots the same kind of summary plots for the Dutch suffix *-heid*, which enjoys similar use as the English suffix *-ness*.

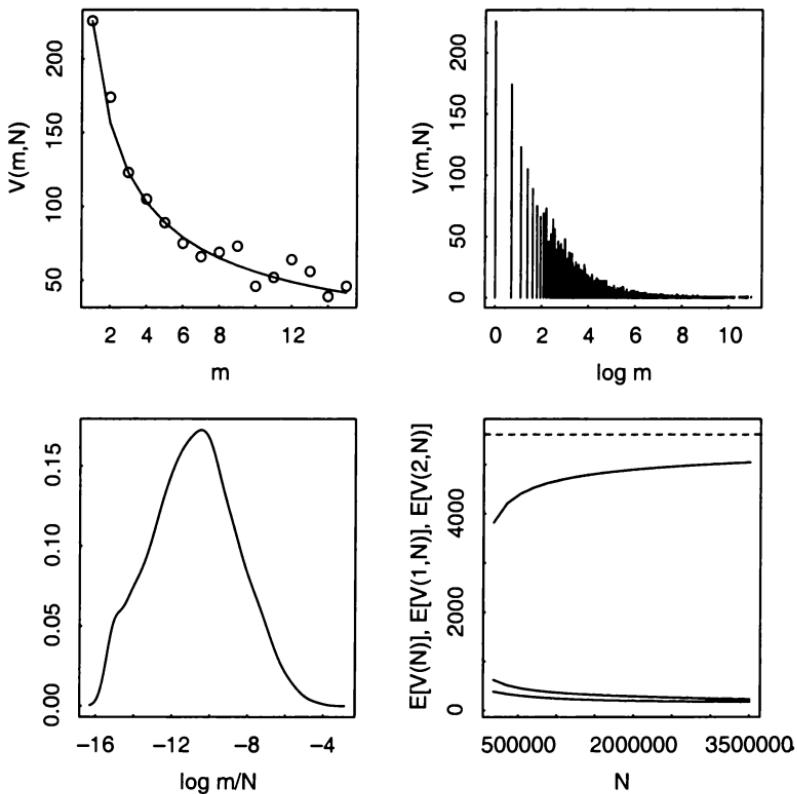


Figure 4.1: The first fifteen spectrum elements for simplex nouns with a Yule-Simon fit ($\hat{Z} = 7650$, $\hat{\beta} = 1.4178$, $\hat{V}_Z = 1654$, upper left), the complete spectrum (upper right), the estimated density (lower left), and the developmental profiles of $E[V(N)]$ and $E[V(m, N)]$, $m = 1, 2$ using binomial interpolation (lower right). The dashed line represents the asymptote $\lim_{N \rightarrow \infty} E[V(N)] = \hat{S} = 5042$.

First consider the upper left panels of both figures. The spectrum plot of the nouns has relatively few hapax legomena compared to the number of dis legomena, while the spectrum plot of the complex nouns with *-heid* has a great many hapax legomena and relatively few dis legomena. The upper right panels plot the full spectrum with a logarithmic scale for the X-axis, using high-density lines. For the nouns we observe that the central frequency ranges are much more densely populated than is the case for the words in *-heid*. The lower left panels show the estimated probability density functions for the logarithmically transformed relative frequencies. For the nouns, we see a curve that somewhat resembles the probability density function of a normal random variable. For the complex words in *-heid*, by contrast, we observe a high degree of skewness to the left. Finally, the bottom right panels show the interpolated growth curves of the vocabulary size and the first two spectrum elements. In the case of the nouns, we see that the vocabulary size increases slowly with an asymptote $Y = \hat{S} = 5042$.¹ The growth curves of the hapax and dis legomena emerge as decreasing functions of N , indicating that this sample is located in the late LNRE zone. Conversely, the *-heid* data are located in the central LNRE zone, and the Yule-Simon fit does not suggest that $E[V(N)]$ might have an asymptote.² These comparisons show that the distribution of simplex nouns and the distribution of complex words in *-heid* have quite different quantitative properties.

What this example shows is that a word frequency distribution of a text or corpus may be a composite entity consisting of parts with quite different quantitative properties. When there is reason to assume that a given word frequency distribution is not homogeneous distribution but rather a mixture of different distributions, we may partition the N observed tokens into t subsets with proportions p_1, \dots, p_t , $\sum_i^t = 1$, and regard $V(N)$ and $V(m, N)$ as mixtures of t components:

$$\begin{aligned} V(N) &= \sum_i V(p_i N), \\ V(m, N) &= \sum_i V(m, p_i N). \end{aligned}$$

In the next section, some formal aspects of analyzing word frequency distributions as mixtures are presented. For ease of exposition, I will sometimes refer to the generalized inverse Gauss-Poisson model by means of the abbreviation GIGP.

¹The fit of the Yule-Simon model to the noun data is not perfect, but given the irregularities in the observed spectrum, a MSE of 162.12 and the chi-squared test ($X_{(13)}^2 = 38.95, p = 0.0002$) are not unreasonable for a sample of $N = 3507305$ tokens and $V = 5042$ types.

²The Yule-Simon fit is very bad: MSE = 2718, 4, $X_{(13)}^2 = 1346, p < 0.00000000$. The estimate of S is therefore a very rough approximation at best. A much better fit will be discussed below.

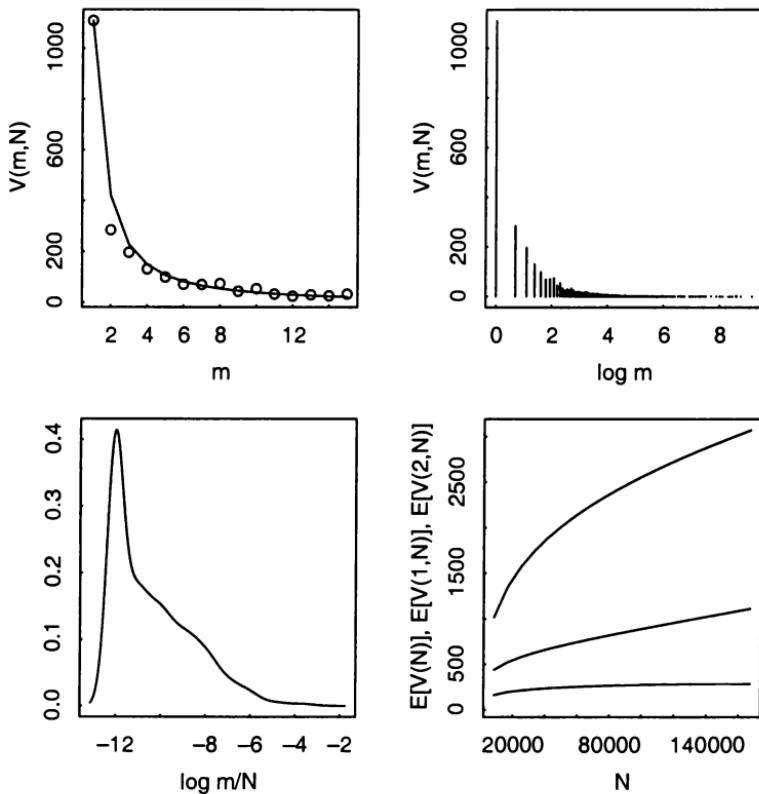


Figure 4.2: The first fifteen spectrum elements for complex nouns with the suffix -heid with Yule-Simon fit ($\hat{Z} = 299.4$, $\hat{\beta} = 1.61$, $\hat{V}_Z = 150$, upper left), the complete spectrum (upper right), the estimated density (lower left), and the developmental profiles of the vocabulary size and the first two spectrum elements using binomial interpolation (lower right).

4.2 Expectations, variances, and covariances

SUMMARY This section introduces expressions for the variances and covariances of LNRE mixture models. It is also shown that the expected spectrum is linear with respect to N and Z , which is indicative of a kind of proportional self-similarity.

We take the frequency spectrum as point of departure, and study the simple case in which $V(m, N)$ originates from just two distributions, one with parameters Z_1, a_1 , and b_1 , and one with parameters Z_2, a_2 , and b_2 . We assume that pN of our tokens have been sampled from the first distribution, and that $(1 - p)N$ tokens come from the second distribution. For the expectation of $V(m, N)$ we have

$$\begin{aligned} E[V(m, N)] &= E[V(m, pN)|\{Z_1, a_1, b_1\} + V(m, (1 - p)N)|\{Z_2, a_2, b_2\}] \\ &= E[V(m, pN)|\{Z_1, a_1, b_1\}] + E[V(m, (1 - p)N)|\{Z_2, a_2, b_2\}] \\ &= pE[V(m, N)|\{\frac{Z_1}{p}, a_1, b_1\}] + \\ &\quad (1 - p)E[V(m, N)|\{\frac{Z_2}{1 - p}, a_2, b_2\}]. \end{aligned} \tag{4.1}$$

The last step in (4.1) makes use of the following theorem:

Theorem A $E[V(m, pN)|\{Z, \dots\}] = pE[V(m, N)|\{\frac{Z}{p}, \dots\}]$

Equivalently, writing $Z' = Z/p$, we have that

$$E[V(m, pN)|\{pZ', \dots\}] = pE[V(m, N)|\{Z', \dots\}],$$

i.e., the expected spectrum is linear in p with respect to N and Z . This equality is illustrated in Figure 4.3. We first prove this theorem for the lognormal model:

$$\begin{aligned} E[V(m, pN)|\{Z, \sigma\}] &= \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty \frac{(xpN)^m}{x^2 m!} e^{-xpN - \frac{1}{2\sigma^2}[\log(xZ)]^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty p^2 \frac{(xpN)^m}{(xp)^2 m!} e^{-xpN - \frac{1}{2\sigma^2}[\log(xp\frac{Z}{p})]^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty p^2 \frac{((xpN)^m}{(xp)^2 m!} e^{-(xp)N - \frac{1}{2\sigma^2}[\log((xp)\frac{Z}{p})]^2} \frac{1}{p} d(xp) \\ &= p \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty \frac{(yN)^m}{y^2 m!} e^{-yN - \frac{1}{2\sigma^2}[\log(y\frac{Z}{p})]^2} dy \\ &= pE[V(m, N)|\{\frac{Z}{p}, \sigma\}]. \end{aligned}$$

For the Yule-Simon model, we similarly obtain that

$$\begin{aligned}
E[V(m, pN) | \{Z, \beta\}] &= \\
&= \frac{Z}{\beta \log(m^*)} \beta \left(\frac{pN}{Z} \right)^m \int_0^\infty \frac{x}{((pN)/Z + x)^{m+1} (1+x)^\beta} dx \\
&= p \frac{\frac{Z}{p}}{\beta \log(m^*)} \beta \left(\frac{N}{\frac{Z}{p}} \right)^m \int_0^\infty \frac{x}{(N/\frac{Z}{p} + x)^{m+1} (1+x)^\beta} dx \\
&= p E[V(m, N) | \{\frac{Z}{p}, \beta\}].
\end{aligned}$$

Note that when V_Z , which appears as $\frac{Z}{\beta \log(m^*)}$ in the above equations, is used as an independent third parameter (to avoid having to estimate m^* by means of Zp^*), we obtain a different result:

$$E[V(m, pN) | \{Z, \beta, V_Z\}] = E[V(m, N) | \{\frac{Z}{p}, \beta, V_Z\}]. \quad (4.2)$$

The proof for the generalized inverse Gauss-Poisson model is left as an exercise. Thus for each model, the change in sample size, pN instead of N , can be reformulated as a change in only one parameter, Z , the pivotal sample size of the LNRE model.

Likewise, we find for the expected vocabulary that

$$\begin{aligned}
E[V(N)] &= E[V(pN) | \{Z_1, a_1, b_1\} + V((1-p)N) | \{Z_2, a_2, b_2\}] \quad (4.3) \\
&= p E[V(N) | \{\frac{Z_1}{p}, a_1, b_1\}] + (1-p) E[V(N) | \{\frac{Z_2}{1-p}, a_2, b_2\}],
\end{aligned}$$

where we use the fact that the vocabulary size can be expressed as the sum of the spectrum elements:

$$\begin{aligned}
E[V(pN) | \{Z, a, b\}] &= E[\sum_m V(m, pN) | \{Z, a, b\}] \\
&= \sum_m E[V(m, pN) | \{Z, a, b\}] \\
&= \sum_m p E[V(m, N) | \{\frac{Z}{p}, a, b\}] \\
&= p E[\sum_m V(m, N) | \{\frac{Z}{p}, a, b\}] \\
&= p E[V(N) | \{\frac{Z}{p}, a, b\}].
\end{aligned}$$

Variances and covariances of the mixture model can be expressed as the sums of variances and covariances of the components. For the variance of the vocabulary size we have:

$$\begin{aligned}
\text{VAR}[V(N)] &= E[V(2N)] - E[V(N)] \\
&= E[V(2pN) | \{Z_1, a_1, b_1\}] + E[V(2(1-p)N) | \{Z_2, a_2, b_2\}] \\
&\quad - [E[V(pN) | \{Z_1, a_1, b_1\}] + E[V((1-p)N) | \{Z_2, a_2, b_2\}]] \\
&= \text{VAR}[V(pN)] + \text{VAR}[V((1-p)N)].
\end{aligned}$$

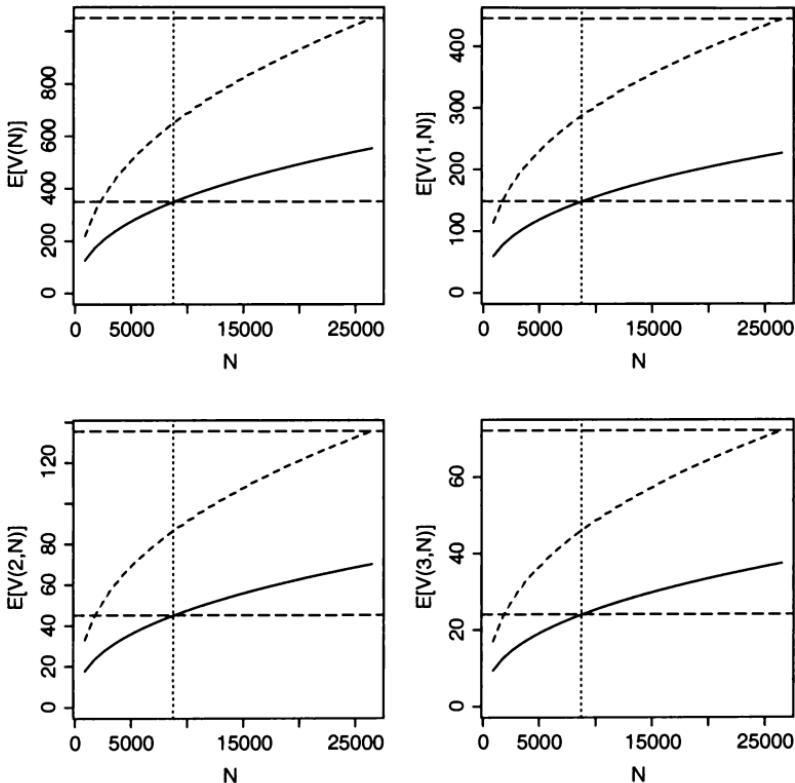


Figure 4.3: Illustrations of the equality $E[V(m, pN)|\{Z, \dots\}] = E[V(m, N)|\{\frac{Z}{p}, \dots\}]$ for a GIGP distribution with parameters $Z = 30, b = 0.001, \gamma = -0.4$: $E[V(N/3)|\{Z, \dots\}] = \frac{1}{3}E[V(N)|\{3Z, \dots\}]$ (upper left panel), $E[V(N/3, m)|\{Z, \dots\}] = \frac{1}{3}E[V(N, m)|\{3Z, \dots\}]$ for $m = 1$ (upper right panel), $m = 2$ (lower left panel), and $m = 3$ (lower right panel). The solid lines represent the expectations for the distribution with $Z = 10.0$, the dashed lines represent the expectations for the distribution with $Z' = 3Z = 30.0$.

For the covariances of the spectrum elements and the vocabulary size we obtain

$$\begin{aligned}\text{COV}[V(m, N), V(N)] &= \frac{1}{2^m} E[V(m, 2N)] \\ &= \frac{1}{2^m} [E[V(m, 2pN)] + E[V(m, 2(1-p)N)]] \\ &= \text{COV}[V(m, 2pN)] + \text{COV}[V(m, 2(1-p)N)],\end{aligned}$$

and along similar lines it can be shown that

$$\begin{aligned}\text{COV}[V(m, N), V(k, N)] &= \text{COV}[V(m, pN), V(k, pN)] + \\ &\quad \text{COV}[V(m, (1-p)N), V(k, (1-p)N)].\end{aligned}$$

These results generalize to mixtures with more than two components. The next section presents some examples of mixture analyses.

4.3 Examples of mixture distributions

SUMMARY This section discusses two kinds of mixture data, first an analysis of a Turkish text for which a mixture analysis seems appropriate, followed by three examples of mixture analyses of morphological data.

4.3.1 A text-level mixture model

A Turkish text on archeology ($N = 6939, V(N) = 3302, V(1, N) = 2326$) provides our first example for a text for which a mixture model is required. The left panel of Figure 4.4 plots $V(m, N)$ for $m = 1, \dots, 15$ using logarithmic scales for both axes with circles as plot symbol. The dashed line represents a simple GIGP fit ($\hat{Z} = 404.34, \hat{b} = 0.000000007149, \hat{\gamma} = -0.623$), with $\text{MSE} = 269.537$ and $X_{(13)}^2 = 96.84, p = 6.77e-16$. A much better fit is obtained with a mixture model with as base component a lognormal distribution ($\hat{Z} = 40.0, \hat{\sigma} = 1.5$) and as complement component a GIGP distribution ($\hat{Z} = 821.98, \hat{b} = 0.000000029947, \hat{\gamma} = -0.5717$), with mixing parameter $p = 0.2$ (i.e., 1388 tokens are attributed to the base component and 5551 tokens to the complement component): $\text{MSE} = 73.2474, X_{(15)}^2 = 33.44, p = 0.004$. This mixture fit is represented by a solid line in the left panel of Figure 4.4. The right hand panel shows the mixture fit (solid line), the lognormal base component (circles) and the GIGP complement component (asterisks) for frequency ranks $m = 4, \dots, 15$. Note that from rank 13 onwards the base curve accounts for more types than the complement curve, which dominates the first 11 frequency ranks.

A possible interpretation for this mixture model is that the base component represents the closed-class words of Turkish, while the complement component represents the nouns, verbs, and adjectives, most of which are morphologically complex. Compared to English, Turkish is a language with a very

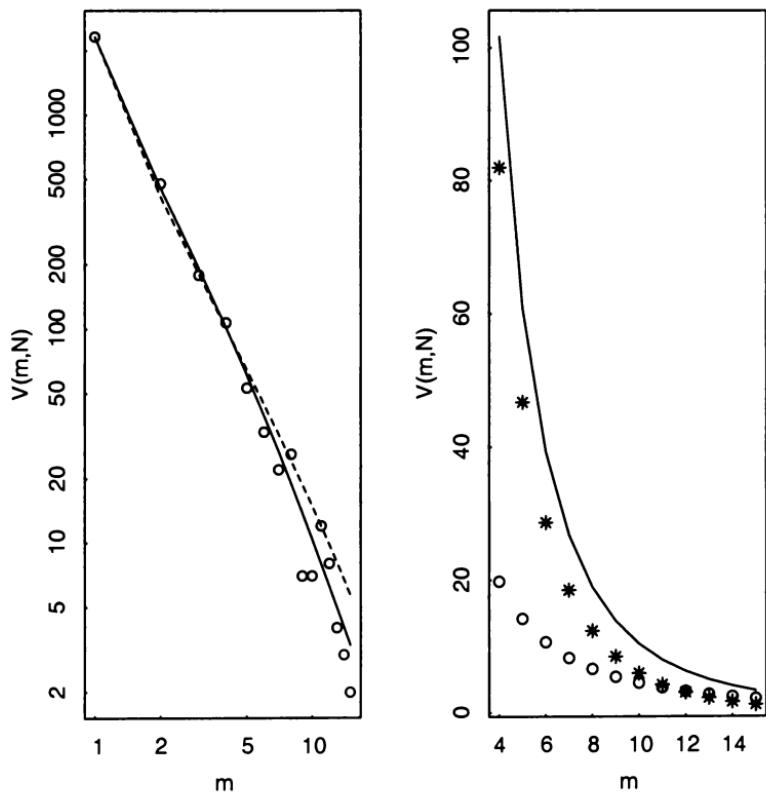


Figure 4.4: Left panel: The first fifteen observed spectrum elements (circles), their expectations for a lognormal-GIGP model (solid line), and the corresponding expectations given a simple GIGP model (dashed line) for a Turkish text on archeology. Right panel: The mixture model for $m = 4, \dots, 15$ (solid line), its lognormal component (circles) and its GIGP component (asterisks).

rich morphological system that makes it possible to create hundreds of complex words from a single root form. By way of illustration of the difference between the two languages, consider Figure 4.5, which plots the vocabulary growth curve for the Turkish archeology text (solid line), an English spoken text (a recorded meeting in an art gallery, text F71 from the British National Corpus, short dashes) of approximately the same length ($N = 6969$, $V(N) = 1078$, $V(1, N) = 477$), and a text consisting of the first 6939 words of Conan Doyle's 'Hound of the Baskervilles' (long dashes). Spoken language tends to be less type rich than written language, and this is what we see for the two kinds of English data: the curve of the spoken language is lower than that of the written language for the full range of N . But even the written English of Conan Doyle is much less type rich than the Turkish archeology text, the growth curve of which hardly shows any curvature at all.

According to the mixture model for the Turkish text, it is the GIGP component that drives the high rate at which the vocabulary size increases. This component accounts for 3199 of the 3302 types in the sample, and it estimates the population number of types at $\hat{S} = 547528241979$. Conversely, the lognormal base component accounts for only 103 out of 3302 types, and estimates its population size to be $\hat{S} = 123$.

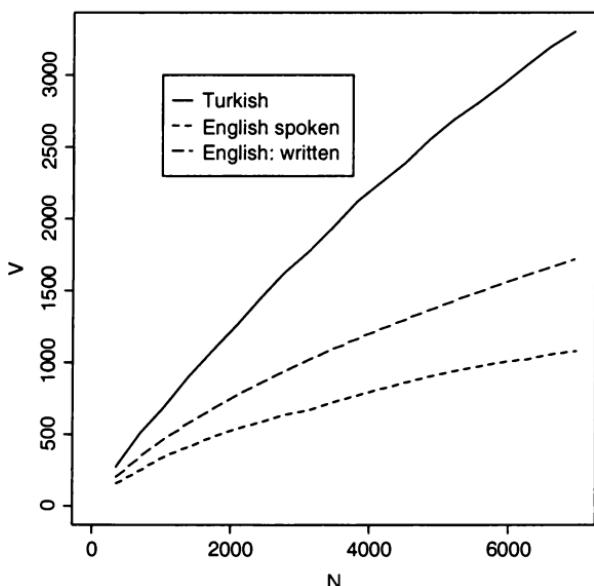


Figure 4.5: Vocabulary growth curves for the Turkish text on archeology analyzed in Figure 4.4 (dashed line) as well as for a spoken English text from the British National Corpus (F71, short dashes) and a written English text (the first part of the Hound of the Baskervilles, long dashes).

4.3.2 Morphological mixtures

As a first example of a morphological mixture distribution, consider the distribution of formations with the Dutch suffix *-heid* ('-ness') that we have already seen in passing in Figure 4.2. For *-heid*, the simple fit provided by the Yule-Simon model ($\hat{Z} = 299.4$, $\hat{\beta} = 1.61$, $\hat{V}_Z = 150$) is entirely unsatisfactory, as both the chi-squared and the mean squared error are large ($X^2_{(13)} = 1346$, $p < 0.000000$; $\text{mse} = 2718.4$). The source of this lack of goodness of fit is immediately apparent from the upper left panel of Figure 4.2: The model overestimates the number of dis legomena, thereby underestimating the steepness of the spectral curve for the smallest word frequencies m .

What is the source of this lack of goodness-of-fit? Before attempting to fit a mixture model, however, we first consider whether there are independent reasons for assuming that a mixture analysis is indeed called for. Interestingly, the suffix *-heid* serves two related but distinguishable purposes. On the one hand, it may express abstract concepts for measures, emotions, and properties. Typical examples are *snelheid*, 'speed', *waarheid*, 'truth', and *vrijheid*, 'freedom', words for well-established concepts. On the other hand, this suffix is also used in a more anaphoric sense, to refer to states of affairs. For instance, if John has just been described as being sad, this state can later be referred to as 'John's sadness'. The anaphoric function of *-heid* is more prominent for low-frequency words, the conceptual function is most clearly visible among the highest-frequency formations. Baayen and Neijt (1997) show that the highest-frequency *-heid* formations tend to occur relatively independently in discourse, while the lowest-frequency forms tend to be somewhat more tightly integrated in their context, exactly the kind of differentiation that one would expect for typical terms on the one hand, and contextually referring anaphoric formations on the other. This is a first indication that high-frequency formations in *-heid* may have different properties from low-frequency formations with this suffix.

A second indication that the high and low frequency forms are somewhat different can be obtained by studying the translation equivalents of high and low frequency words in *-heid*. Table 4.1 lists some randomly selected examples of hapax legomena with *-heid* in the right column and the twenty highest-frequency forms in the left column. Note that the hapax legomena invariably translate with *-ness*, but that the high-frequency formations have various translation equivalents: simplex words (*snelheid*, 'speed'), formations with obsolete suffixes such as *-th* (*waarheid*, 'truth') and *-dom* (*vrijheid*, 'freedom'), or with the marginally productive suffix *-ity* (*werkelijkheid*, 'work-ly-ness', 'reality'). In addition, the high-frequency formations tend to have more translation equivalents than the low-frequency formations. For instance, *enzaamheid* translates into 'solitude' or 'loneliness', and *mogelijkheid* into 'possibility', 'feasibility', 'opportunity', 'prospect', and 'eventuality'. Such a range of meanings is often already carried by the base word. However, the derived form need not express all meanings carried by the base. In the case of *mogelijkheid*, for instance, the meaning 'conceivability' is not in use, even though *mogelijk* carries the meaning 'conceivable'. (The translation equivalent of *conceivabil-*

ity is 'denkbaarheid'.) The fact that high-frequency formations in *-heid* denote meanings that in English are carried by words that have forms that are not supported by synchronously productive word formation patterns suggests that their meanings may exist relatively independent from the meanings of their base word, and that their interpretation need not crucially depend on their formal morphological structure.

A third indication comes from experimental work on lexical processing. Baayen, Schreuder, Bertram, and Tweedie (1999) report that high-frequency formations in *-heid* are processed to a greater degree on the basis of stored knowledge of these forms in the mental lexicon, while for low-frequency forms on-line computation of the meaning on the basis of the base adjective and the suffix plays a more prominent role. Possibly, comprehension through on-line parsing is more usual of *-heid* formations predominantly carrying the anaphoric function, while comprehension based on stored knowledge of the form of the formation itself is more usual for *-heid* formations predominantly expressing terms.

Table 4.1: *The twenty most frequent formations with -heid (left) and twenty randomly chosen examples of hapax legomena with -heid (right).*

veiligheid	1395	'safety'	bedaagdheid	'elderliness'
eenzaamheid	1485	'solitude'	bovenzinnelijkheid	'supersensoriness'
werkzaamheid	1524	'industry'	doofstomblindheid	'deaf-mute-blindness'
meerderheid	1572	'majority'	fijnbeschaafdheid	'well-culturedness'
snelheid	1694	'speed'	geliefdheid	'belovedness'
bevoegdheid	1815	'competence'	kinderloosheid	'childlessness'
gezondheid	1850	'health'	hoopvolheid	'hopefulness'
schoonheid	1856	'beauty'	klefheid	'stickiness'
persoonlijkheid	2227	'personality'	lichthoofdigheid	'lightheadedness'
verantwoordelijkheid	2659	'responsibility'	mysterieuosheid	'mysteriousness'
aanwezigheid	2683	'presence'	onbetekenendheid	'insignificantness'
eenheid	2984	'unity'	onkiesheid	'indelicateness'
zekerheid	3049	'certainty'	onvermeerderbaarheid	'unincreasableness'
moeilijkheid	4065	'difficulty'	overbelastheid	'overtaxedness'
gelegenheid	4435	'opportunity'	pronkievendheid	'ostentatiousness'
waarheid	4753	'truth'	schandaligheid	'outrageousness'
vrijheid	5298	'freedom'	sprookjesachtigheid	'fairy-tale-like-ness'
omstandigheid	5579	'circumstance'	uitgelezenheid	'exquisiteness'
werkelijkheid	6437	'reality'	vernuftigheid	'ingenuousness'
mogelijkheid	9623	possibility'	vreeswekkendheid	'terrifyingness'

Given that there are independent reasons for analyzing the word frequency distribution of *-heid* as a mixture of two distributions, we now have to ascertain the nature of the component distributions, to which I shall refer as the base component and the complement component. The base component can be a lognormal, a GIGP, or a Yule-Simon distribution, and the same holds for the complement component. Good fits are obtained when we select a lognormal distribution as base component for the terms, and either a GIGP or a Yule-Simon component for the anaphors. The lognormal-GIGP mixture model ($\hat{Z}_1 = 200$, $\hat{\sigma} = 2.05$, $\hat{Z}_2 = 83.6369$, $\hat{b} = 0.00000002715$, $\hat{\gamma} = -0.5643$) is superior in terms of the MSE (93.95) and inferior in terms of the chi-squared ($X^2_{(10)} = 18.21$, $p = 0.0515$), the lognormal-Yule-Simon mixture model ($\hat{Z}_1 =$

$200, \hat{\sigma} = 2.05, \hat{Z}_2 = 107.1620, \hat{\beta} = 0.3525, \hat{V}_Z = 107.16$) is superior in terms of the chi-squared ($X^2_{(10)} = 0.83, p = 0.9999$) and inferior in terms of the MSE (150.05). The mixing parameter for both models is set to 0.96: 160549 tokens are accounted for by the lognormal base component, and 6690 tokens are accounted for by the GIGP and the Yule-Simon component.

Figure 4.6 presents some diagnostic plots. The upper panels shows the observed frequency spectrum (circles) and the expected spectrum for the lognormal-Yule-Simon mixture (solid line) as well as the corresponding expectations derived from the simple Yule-Simon fit discussed above in relation to Figure 4.2 (dashed line). The upper left panel shows the first 15 spectrum elements. In the mixture analysis, the overestimation of $E[(2, N)]$ has been removed. The upper right panel plots the spectrum elements 5–100. The erratic line connects the observed data points, the solid line represents the expected values according to the mixture model, and the dashed line running below the solid line again represents the simple Yule-Simon fit. The simple model underestimates the spectrum elements for the higher frequencies m . This underestimation is corrected for in the mixture model. Thus, the mixture model is more precise both for the smallest ranks m and for the medium frequency ranges.

The lower left panel shows the contributions of the two component models. The lognormal component takes care of the majority of the spectrum counts $E[V(m, N)]$ for $m > 4$, the complement component is the dominant contributor to the spectrum counts for the smallest values of m . This division of labor is reflected in the composition of the estimated number of types in the population, $\hat{S} = 660193617889$, of which 1635.3 are contributed by the base component and 660193616253.97 by the complement component. The bottom right panel, finally, shows that the interpolation accuracy of the mixture model has improved considerably. Using binomial interpolation as the baseline (represented by circles), we observe that the interpolated values for the mixture model (solid lines) are virtually identical to those of the non-parametric model, while the interpolated values predicted by the simple Yule-Simon model (dashed lines) underestimate the binomial values for $V(N)$ and overestimate $V(1, N)$ and $V(2, N)$. We conclude that the mixture model is a non-trivial improvement with respect to the simple Yule-Simon model.

As a second example of a morphological mixture distribution, we turn to the distribution of the morphological category of words with the suffix *-iteit*, the Dutch equivalent of the English suffix *-ity*. The Dutch suffix *-iteit* is much less productive than *-ity* in English, in part because many English words in *-ity* have translation equivalents with *-heid* in Dutch (see Table 4.1 above).³ As in the case of the distribution of *-heid* formations, there is no simple LNRE model that provides an acceptable fit. The fit shown in the upper left-hand panel of Figure 4.7 by means of a dashed line is that of the GIGP model (MSE: 234.52; $X^2_{(13)} = 1230.79$). It is roughly as bad as that of the Yule-Simon model (MSE:

³This difference in productivity is clearly visible in the counts for $V(N)$. In the 18 million Cobuild corpus underlying the English frequency counts in the CELEX lexical database, 611 formations with *-ity* are found, whereas for *-iteit* the total number of types is only 362, using a much larger corpus of 42 million words.

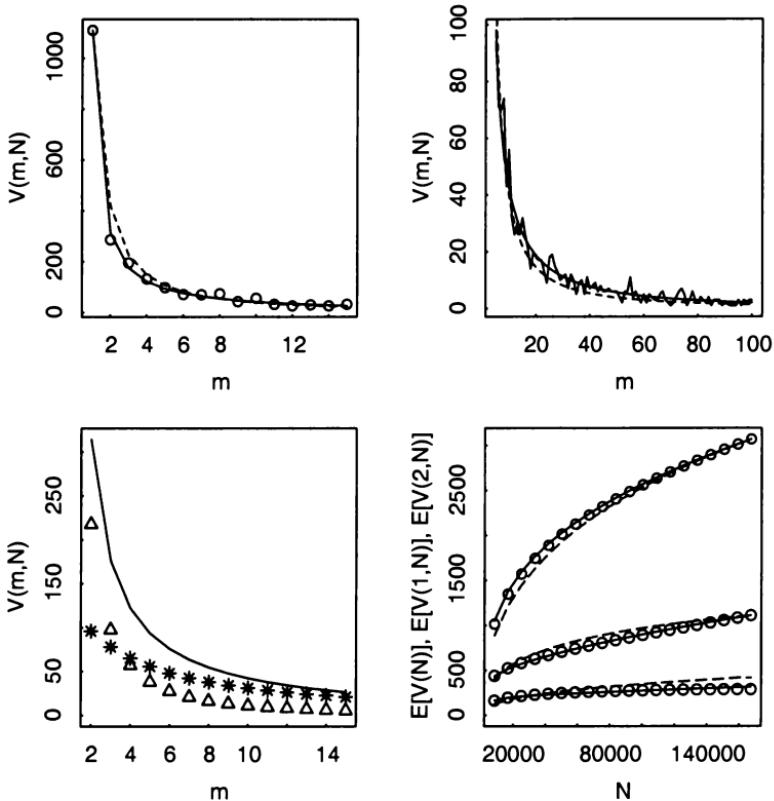


Figure 4.6: The spectrum elements for complex nouns with the suffix -heid (circles), the Yule-Simon fit (dashed line), and a Lognormal-Yule-Simon mixture fit (solid line): $m = 1, \dots, 15$ (upper left), $m = 5, \dots, 100$ (upper right). The lower left panel shows the mixture components for $m = 2, \dots, 15$, using asterisks for the lognormal component and triangles for the Yule-Simon component and a solid line for the mixture model itself. The lower right panel shows the congruence of binomial interpolation (circles) and interpolation for the mixture model (solid line) and the simple Yule-Simon fit (dashed line).

241.62 ; $X^2_{(13)} = 1156.73$). Just as we have seen for *-heid*, the steep slope of the head of the frequency spectrum is not captured adequately. Especially for the frequency ranks $m = 2, 3, 4, 5$, the simple LNRE models overestimate the spectrum values.

The upper right panel of Figure 4.7 plots the estimated probability density function for the sample relative frequencies. Contrary to what one would expect for an LNRE model, we observe a bi-modal function. This bimodality suggests that a mixture analysis is called for. Indeed, a much improved fit is obtained when we approach the distribution of *-iteit* formations as a mixture of a lognormal distribution with parameters $\hat{Z} = 40.0$ and $\hat{\sigma} = 1.8$ and a GIGP distribution with parameters $\hat{Z} = 0.6018$, $\hat{b} = 0.00000000021$, and $\hat{\gamma} = -0.7120$ ($MSE = 16.25$),⁴ with as mixing parameter $p = 0.95$. The improved fit is represented by a solid line in the left panel of Figure 4.7.

The lower left panel of Figure 4.7 shows the contributing mixture components for the frequency ranks $m = 2, \dots, 15$, using asterisks for the lognormal component and triangles for the GIGP component. The latter accounts for almost all hapax legomena and roughly half of the dis legomena, the former accounts for the majority of types with the higher frequency ranks m .

The lower right panel of Figure 4.7 shows that the interpolation accuracy of the mixture model (solid line) is much improved both qualitatively and quantitatively compared to the simple GIGP fit (dashed line), where we again use parameter-free binomial interpolation as baseline. Note that the simple model suggests that growth curves of the hapax and dis legomena are monotonically increasing functions of N (for the current 20 measurement points), while both binomial interpolation and interpolation on the basis of the mixture model suggest points of inflection for both curves.

In order to see how we might interpret the mixture model, consider Table 4.2, which lists the twenty highest-frequency formations with *-iteit* (left columns) as well as twenty randomly selected examples of low-frequency forms.

The words in the left-hand column are all ordinary words of the more formal, ‘learned’ registers of Dutch, with the exception of some commonly used words such as *electriciteit*, ‘electricity’, and *activiteit*, ‘activity’. The words listed in the right-hand column are in no way ordinary words, however. Words such as *commoditeit*, *municipaliteit*, *hospitaliteit*, and *chariteit* are unknown to me as Dutch words, even though they are readily interpretable given the English words *commodity*, *municipality*, *hospitality*, and *charity*. These words seem to be calques, highly marked alternatives for the more normal words *droogte*, *artikel*, *gemeente*, *gastvrijheid* and *liefdadigheidsinstelling*. Other words, such as *ariditeit*, *bilinealiteit*, and *hypertoniciteit* are technical terms used only in specific areas of scientific inquiry.

Thus it seems likely that the word frequency distribution of *-iteit* is a mixture of a set of well-known non-scientific words on the one hand, and a set of scientific terms (often calqued from English) on the other hand. According to the mixture fit, the sample nearly exhausts the set of common non-scientific

⁴For these data, a negative value was obtained for X^2 , which is probably due to irregularities in the observed spectrum from $m = 6$ onwards.

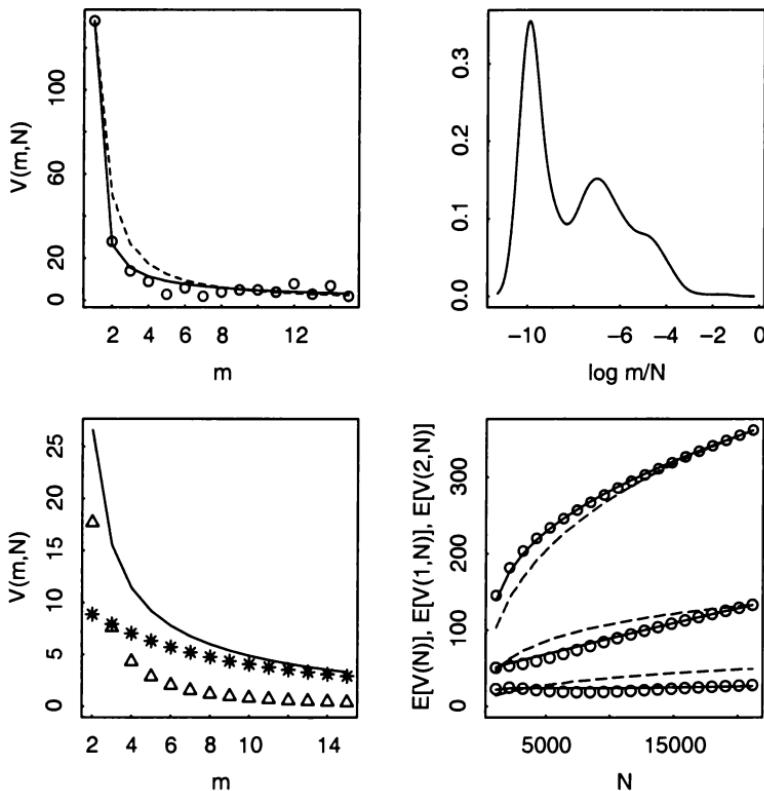


Figure 4.7: The spectrum elements for complex nouns with the suffix -iteit (circles), the GIGP fit (dashed line), and a lognormal-GIGP mixture fit (solid line), for $m = 1, \dots, 15$ (upper left). The upper right panel shows the estimated density function for the relative frequencies using a logarithmic scale. The lower left panel shows the mixture components for $m = 2, \dots, 15$, using asterisks for the lognormal component and triangles for the GIGP component and a solid line for the mixture model itself. The lower right panel shows the congruence of binomial interpolation (circles) and interpolation for the mixture model (solid line) and the simple Yule-Simon fit (dashed line).

Table 4.2: *The twenty most frequent formations with -iteit (left) and twenty randomly selected examples of hapax legomena with -iteit (right).*

<i>pluriformiteit</i>	237	'pluriformity'	<i>ariditeit</i>	'aridity'
<i>stabiliteit</i>	243	'stability'	<i>bilinealiteit</i>	'bilineality'
<i>vitaliteit</i>	250	'vitality'	<i>chariteit</i>	'charity'
<i>rationaliteit</i>	256	'rationality'	<i>commoditeit</i>	'commodity'
<i>nationaliteit</i>	285	'nationality'	<i>curieusiteit</i>	'curiosity'
<i>populariteit</i>	294	'popularity'	<i>diviniteit</i>	'divinity'
<i>specialiteit</i>	297	'speciality'	<i>essentialiteit</i>	'essentiality'
<i>objectiviteit</i>	317	'objectivity'	<i>habitualiteit</i>	'habituality'
<i>intimiteit</i>	363	'intimacy'	<i>hospitaliteit</i>	'hospitality'
<i>elektriciteit</i>	383	'electricity'	<i>hypertoniciteit</i>	'hypertonicity'
<i>totaliteit</i>	395	'totality'	<i>improductiviteit</i>	'unproductiveness'
<i>continuiteit</i>	425	'continuity'	<i>ingenuositeit</i>	'ingenuousness'
<i>mentaliteit</i>	462	'mentality'	<i>irritabiliteit</i>	'irritability'
<i>publiciteit</i>	468	'publicity'	<i>localiteit</i>	'locality'
<i>intensiteit</i>	476	'intensity'	<i>monoseksualiteit</i>	'monosexuality'
<i>creativiteit</i>	481	'creativity'	<i>municipaliteit</i>	'municipality'
<i>solidariteit</i>	585	'solidarity'	<i>notabiliteit</i>	'noteability'
<i>seksualiteit</i>	683	'sexuality'	<i>operabiliteit</i>	'operability'
<i>realiteit</i>	1487	'reality'	<i>paranormaliteit</i>	'paranormality'
<i>activiteit</i>	4500	'activity'	<i>posteriteit</i>	'posterity'

words (190 words in the sample out of an estimated total of 202), while the sample only shows the tip of the iceberg with respect to the set of scientific terms (172 in the sample out of an estimated 39972140866221 potential types).

As a third example of a morphological mixture distribution, we turn to the words in *-ness* in two subcorpora of the British National Corpus, the subcorpus of written English, and the subcorpus of context-governed spoken English. The extent to which affixes are used in the spoken registers of English differs substantially from their use in the written registers. The written subcorpus is much larger than the spoken subcorpus. Not surprisingly, we find many more *-ness* formations in the former ($N = 106957, V = 2466$) than in the latter ($N = 4037, V = 310$).

The upper left panel of Figure 4.8 plots a GIGP fit to the first fifteen spectrum elements of the *-ness* words in the subcorpus of written English ($\text{MSE} = 53.76, X_{(13)}^2 = 29.60, p = 0.0054; \hat{Z} = 24.58, \hat{b} = 0.00834, \hat{\gamma} = -0.5$), the upper right panel similarly plots a Yule-Simon fit for the subcorpus of spoken English ($\text{MSE} = 7.49, X_{(13)}^2 = 34.70, p = 0.0009; \hat{Z} = 80.57, \hat{\beta} = 0.5048, \hat{V}_Z = 32.23$). The lower left panel illustrates the difference between the two distributions by plotting the interpolated growth curves of the vocabulary size. The upper curve represents the first part of the growth curve for the written subcorpus, the lower curve the complete growth curve for the spoken subcorpus. The dotted lines mark the 95% confidence interval for the formations from the written subcorpus. The curve of the spoken subcorpus comes nowhere near this confidence interval. Clearly, we are dealing with sets of words with different underlying distributions.

The lower right panel plots the mixture spectrum,

$$V(m, N) = V(m, N_1) + V(m, N_2)$$

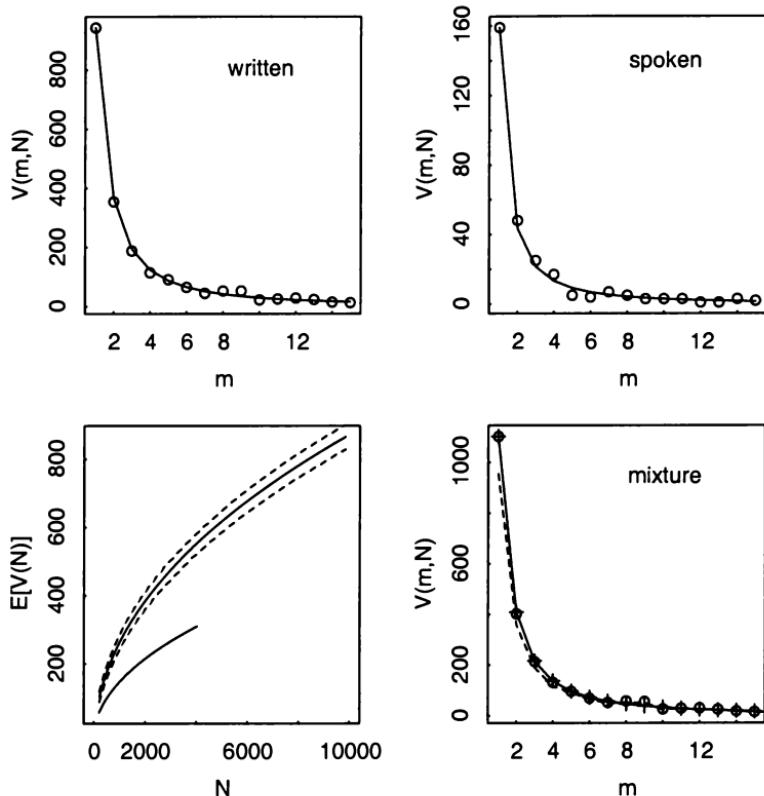


Figure 4.8: The first 15 spectrum elements for complex nouns with the suffix -ness in the written subcorpus of the British National Corpus (upper left) and in the context-governed spoken subcorpus (upper right). Observed values are represented by circles, the LNRE-based expectations (GIGP for the written and Yule-Simon for the spoken subcorpus) by solid lines. The lower left panel shows the LNRE-interpolated growth curves of the vocabulary for the written (upper curve) and spoken (lower curve) subcorpus. The dotted lines mark the 95% confidence interval for the written subcorpus. The lower right panel plots the mixture of the two distributions (circles), the GIGP fit to the mixture (solid line), the GIGP-Yule-Simon mixture fit (+), and the observed spectrum when a pure string-based type definition is used.

using circles for the observed values $V(m, N)$, a solid line for a simple GIGP fit ($\text{MSE} = 39.21$, $X^2_{(13)} = 18.55$, $p = 0.1377$; $\hat{Z} = 28.26$, $\hat{b} = 0.00769$, $\hat{\gamma} = -0.5$), and + symbols for the mixture model with the GIGP and Yule-Simon components as fitted previously for the individual subcorpora with mixing parameter $p = 0.0364$ ($\text{MSE} = 38.14$; $X^2_{(9)} = 18.69$, $p = 0.0280$). To the eye, the two fits are indistinguishable. The mixture model has a marginally smaller MSE and a smaller X^2 value. In the assessment of the X^2 value, however, we have to take the greater number of parameters into account in degrees of freedom, so that the simple GIGP fit emerges as the more probable model.

This example illustrates two important points. First, a good fit of a simple LNRE model to an observed frequency spectrum does not necessarily imply that the underlying distribution is homogeneous. Many of the simple models to texts such as *Alice's adventures in wonderland* provide reasonable analyses at a macroscopic level, even though they can be analyzed at the microscopic level of morphological structure as mixtures of various kinds of words.

Second, a *-ness* formation such as *willingness*, which occurs both in the written subcorpus and in the spoken subcorpus, contributes two types to the mixture distribution, a 'written' type occurring 1157 times, and a 'spoken' type occurring 35 times. The definition of what constitutes a type in this example is based on both the shape of the word (*willingness* and *illness* being different types) and the kind of language setting in which the word is used (spoken *willingness* being a different type than written *willingness*). When we use a pure string-based type definition, collapsing *willingness* in the written subcorpus and *willingness* in the spoken subcorpus into one type, we obtain a word frequency distribution with a slightly smaller number of types. Notably, the number of hapax legomena is smaller, as shown by the dashed line in the lower right panel of Figure 4.8. Note that it is not possible to estimate the component distributions in the written and spoken corpora from the collapsed distribution without further additional assumptions about how the probability of a given string type to occur in one component is related to its probability to occur in the other mixture component. This is an issue awaiting further research.

Since the present mixture model theory presupposes disjunct component distributions, we briefly reconsider to what extent our interpretation of the mixture distributions of words in *-iteit* and *-heid* is appropriate. Recall that the words in *-iteit* fall into two sets, general words (e.g., *electriciteit*, 'electricity') and specialized words (scientific terms, calques; e.g., *operabiliteit*, 'operability'). A given word cannot be a general word of the language community on the one hand, and at the same time a specialized word used only by people in a particular professional context on the other, as required by mixture theory.

Turning to the words in *-heid*, the motivation for the mixture model should be formulated with care, as the simple assumption that a given *-heid* formation is either a term or an anaphoric word is obviously wrong. Although a word such as *snelheid*, 'quick-ness', is the Dutch term for the notions of speed and velocity, it can be used in an anaphoric way as well to refer to previous situations in the discourse (*Langzaam jogde de atleet langs. Zijn snelheid ontlokte gefluit bij het publiek* 'Slowly the athlete jogged by. His speed elicited

whistling from the public'). In fact, a given occurrence of *snelheid* might even instantiate the anaphoric function and the term function simultaneously. For our mixture model to be valid, therefore, we have to define precisely in what sense the two kinds of *-heid* formations constitute disjunct sets. We can do so by taking as point of departure that all *-heid* words can serve the anaphoric function, while only a subset subserves the term function in the sense that they denote concepts that are part of the linguistic and cognitive knowledge of speakers of Dutch. The anaphoric words, then, can be defined negatively as those *-heid* formations that do not denote standard concepts and hence subserve the anaphoric function only. In this interpretation, we again have two disjunct sets, as required.

4.4 Morphological Productivity

SUMMARY *Word frequency distributions can be analyzed as mixtures of distributions of monomorphemic and polymorphemic words. The component distributions have quite different formal properties: Some are LNRE distributions, others are distributions outside the LNRE zone with hardly any rare types. The quantitative properties of mixture components tie in with the linguistic concept of morphological productivity, a pre-theoretical notion with various interpretations that each can be formalized statistically.*

In the previous section, we have seen that a word frequency distribution can be analyzed morphologically as a mixture distribution with as component distributions various kinds of simplex and complex words, e.g., function words (*the, a, of, to, that, ...*), monomorphemic verbs (*think, describe, hallucinate*), complex nouns in *-ness* (*gratefulness, goodness*), complex adjectives with *un-* (*unthinkable, unfriendly*), etc. It has long been noticed that some of the word formation patterns are very type-rich, while others comprise only a few types. An example of the latter is the English suffix *-th*, which occurs in only a handful of words such as *warmth, strength, length, and breadth*. In linguistics, the problem of vocabulary richness with respect to word formation patterns is known as the problem of morphological productivity. Word formation patterns can be 'productive' in two ways. It can be productive in the sense that many rather than few types have been observed. And it can be productive in the sense that the pattern can be extended easily to form new formations that have not been observed before. Sometimes the number of potential types, the number of types that could possibly be formed following a given word formation pattern, is also taken into consideration. Most linguistic studies of morphological productivity have attempted to capture these aspects of the phenomenon in terms of qualitative properties of word formation patterns. This approach, however, has not led to a consistent theory of the various aspects of lexical type richness. This section shows how the various aspects of morphological productivity can be formalized in terms of mixture model theory.

Let the lexicon \mathcal{L} be a partition of L sets of words according to their morphological structure, henceforth morphological categories \mathcal{M} :

$$\mathcal{L} = \bigcup_{i=1}^L \mathcal{M}_i. \quad (4.4)$$

A given word belongs to one morphological category only. Thus, *unfriendly* belongs to the morphological category of *un-*, *friendly* to the morphological category of *-ly*, and *friend* to the morphological category of simplex nouns. We define the mixture model for \mathcal{L} in terms of the component subvocabularies for a sample of N tokens:

$$\mathbb{E}[V(N)] = \sum_{i=1}^L \mathbb{E}[V(p_i N) | \{Z_i, \dots\}]. \quad (4.5)$$

Given (4.5), we can formalize the notion of productivity in the sense of what has been produced, henceforth **extent of use** (\mathcal{E}), in terms of the proportion of types contributed by morphological category \mathcal{M} :

$$\mathcal{E}_i = \frac{\mathbb{E}[V(p_i N) | \{Z_i, \dots\}]}{\sum_{j=i}^L \mathbb{E}[V(p_j N) | \{Z_j, \dots\}]} \quad (4.6)$$

When comparing morphological categories with respect to their extent of use, we can simply compare $\mathbb{E}[V(p_i N) | \{Z_i, \dots\}]$.

The estimated population number of types S_i for a given morphological category \mathcal{M} is the appropriate measure for gauging the **potentiaality** of a word formation pattern, the number of types it might possibly give rise to.

Theorem B $S = \sum_{i=1}^L S_i$.

Theorem B states that the mixture model allows us to view the potentiaality of the vocabulary as a whole as the sum of the potentiaality of its component distributions. To prove Theorem B, first observe that by Theorem A we can rewrite (4.5) as

$$\mathbb{E}[V(N)] = \sum_{i=1}^L p_i \mathbb{E}[V(N) | \{\frac{Z_i}{p_i}, \dots\}], \quad (4.7)$$

so that

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathbb{E}[V(N)] &= \lim_{N \rightarrow \infty} \sum_{i=1}^L p_i \mathbb{E}[V(N) | \{\frac{Z_i}{p_i}, \dots\}] \\ &= \sum_{i=1}^L p_i \lim_{N \rightarrow \infty} \mathbb{E}[V(N) | \{\frac{Z_i}{p_i}, \dots\}] \\ &= \sum_{i=1}^L p_i \frac{S_i}{p_i} \\ &= \sum_{i=1}^L S_i. \end{aligned}$$

In the last but one step, we make use of yet a third theorem:

Theorem C $S| \{Z, \dots\} = pS| \{pZ, \dots\}$.

First consider the validity of Theorem C for Sichel's generalized inverse Gauss-Poisson model. Since

$$S = \frac{2Z}{b} \frac{K_\gamma(b)}{K_{\gamma+1}(b)}, \quad (4.8)$$

we immediately have that changing Z to Z/p changes S into pS . For the Yule-Simon model, S is infinite whatever the value of Z may be when $\beta \leq 1$. But for $\beta > 1$, we have that

$$S = E[V(Z)] \frac{\beta}{\beta - 1} = Z \frac{1}{\log(m^*)(\beta - 1)}, \quad (4.9)$$

from which Theorem C follows immediately. The proof for the lognormal model is left as an exercise. To evaluate the potentiality of a given morphological category M_i , we can compare its estimated population richness S_i with the corresponding values for other morphological categories. Alternatively, one can investigate the extent to which S_i exceeds $V(p_i N)$ by means of the potentiality index

$$\mathcal{I} = \frac{\hat{S}_i}{V(p_i N)}. \quad (4.10)$$

A third statistic that is useful for making precise the notion of degree of productivity is the growth rate of the vocabulary

$$\mathcal{P} = \frac{E[V(1, N)]}{N}. \quad (4.11)$$

The vocabulary growth rate can be interpreted as a probability, the probability that a new type will be sampled after N tokens have been sampled. To see this, consider sampling without replacement from an urn with N tokens. The probability that the first token sampled represents a type that will never again be encountered in the sampling process equals $V(1, N)/N$. By symmetry, this is also the probability that the last token sampled represents a new type. At the sample size $N - 1$, $V(1, N)/N$ is the probability that the next token to be sampled represents a new type.

There are two ways in which we can use \mathcal{P} to formalize the notion 'degree of productivity' of the i -th component of a lexical mixture distribution. First, we can consider the growth rate of the i -th component itself, i.e.,

$$\mathcal{P}_i = \frac{E[V(1, p_i N)|\{Z_i, \dots\}]}{p_i N}. \quad (4.12)$$

This growth rate is the conditional probability that the next token to be sampled represents a new type, given that this token belongs to the i -th component. Let $\{A\}$ denote the event that the new token represents a new type,

and let $\{B\}$ denote the event that the new token belongs to the i -th mixture component. Then

$$\begin{aligned}\mathcal{P}_i &= \Pr(\{A\}|\{B\}) \\ &= \frac{\Pr(\{A\} \cap \Pr(\{B\}))}{\Pr(\{B\})} \\ &= \frac{\frac{\mathbb{E}[V(1, p_i N) | \{Z_i, \dots\}]}{N}}{\frac{p_i N}{N}} \\ &= \frac{\mathbb{E}[V(1, p_i N) | \{Z_i, \dots\}]}{p_i N}.\end{aligned}$$

Morphological categories can be productive in the sense that new formations can be coined with ease, and nevertheless appear in texts with small numbers of types. In such cases, \mathcal{P} is always much larger than for unproductive morphological categories appearing with similar numbers of types. I will refer to \mathcal{P} as the category-conditioned degree of productivity.

We can also consider the probability that the next token to be sampled represents a type belonging to the i -th component, given that this token represents a new type that has not been observed before for any of the mixture components.

$$\begin{aligned}\mathcal{P}_i^* &= \Pr(\{B\}|\{A\}) \\ &= \frac{\Pr(\{B\} \cap \Pr(\{A\}))}{\Pr(\{A\})} \\ &= \frac{\frac{\mathbb{E}[V(1, p_i N) | \{Z_i, \dots\}]}{N}}{\frac{\sum_{i=1}^L \mathbb{E}[V(1, p_i N) | \{Z_i, \dots\}]}{N}} \\ &= \frac{\mathbb{E}[V(1, p_i N) | \{Z_i, \dots\}]}{\mathbb{E}[V(1, N)]}.\end{aligned}$$

This statistic measures the contribution of the i -th mixture component to the growth rate of the mixture distribution itself. It is useful for gauging how often new formations will be encountered when reading through a text or corpus. Note that the ratio of two \mathcal{P}^* -values is equal to the ratio of the corresponding expected numbers of hapax legomena. I will refer to \mathcal{P}^{ast} as the hapax-conditioned degree of productivity.

A closely related measure is the probability of sampling a new type belonging to category \mathcal{M}_i , the unconditioned degree of productivity \mathcal{P}^{**} :

$$\mathcal{P}^{**} = \frac{\mathbb{E}[V(1, p_i N) | \{Z_i, \dots\}]}{N}. \quad (4.13)$$

As in the case of \mathcal{P}^* , comparisons of morphological categories in terms of \mathcal{P}^{**} amount to comparisons in terms of numbers of hapax legomena.

Figure 4.9 illustrates the complementary nature of \mathcal{P} and \mathcal{P}^* as measures capturing aspects of the intuitive notion of 'degree of productivity'. The data

for this example are the common nouns in Innes' novel *The bloody wood*, partitioned into singular nouns and plural nouns. Evaluated in terms of \mathcal{P} , the plurals ($\mathcal{P}(386) = 173/386 = 0.45$) are more productive than the singulars ($\mathcal{P}(2352) = 602/2352 = 0.26$). In terms of \mathcal{P}^* , however, the singulars ($\mathcal{P}^*(2738) = 602/770 = 0.78$) are more productive than the plurals ($\mathcal{P}^*(2738) = 173/770 = 0.22$). At first sight, the two measures seem to contradict each other. We can resolve this apparent contradiction by taking into account how \mathcal{P}^* develops over sampling time. The left panel of Figure 4.9 shows that in this simple mixture of singulars and plurals, $\mathcal{P}^*(N)$ is a decreasing function of N . The right panel shows that it is an increasing function of N for the plurals. As we proceed through the text, the relative contribution to the growth rate of the vocabulary increases for the plurals and decreases for the singulars. There are more singulars ($V(2352) = 985$) than plurals ($V(386) = 236$) in the complete novel, so that it comes as no surprise that there are more singular hapaxes than plural hapaxes. For the range of sample sizes observed, then, the probability that a new type is a singular is larger than that of being a plural. The opposite ways in which \mathcal{P}^* develops in sampling time for the two component distributions shows, however, that for larger sample sizes their relative contributions to the growth rate may change in favor of the plurals. This possibility is highlighted by inspection of the category-conditioned degree of productivity for singulars and plurals, which is larger for the plurals (0.45) than for the singulars (0.26). Estimates of S provide further support for the greater type richness of plurals.⁵ In sum, the category-conditioned degree of productivity provides insight into the long-run potentiality of a morphological category, while the hapax-conditioned degree of productivity measures its short-term immediate potentiality.

4.5 Discussion

Word frequency distributions have component distributions with greatly varying quantitative properties. Most word frequency distributions of affixes are prototypical LNRE distributions with large numbers of hapax legomena, while especially categories of monomorphemic nouns, verbs, and adjectives are typically located far outside the LNRE zone. The LNRE property of a word frequency distribution coincides with the linguistic property of being productive, which makes it possible to make the notion of morphological productivity more precise by defining the various ways in which this notion can be interpreted with formal probabilistic definitions.

In this chapter, we have seen that expressions for expectations, variances and covariances, and the population number of types can be obtained for a mixture model on the basis of the corresponding expressions of the component distributions. A challenge for future research is to find adequate mod-

⁵The GIGP model provided excellent fits for both the singulars ($\hat{Z} = 193.57, \hat{b} = 0.05, \hat{\gamma} = -0.53; X_{(13)}^2 = 6.02, p = 0.95$) and the plurals ($\hat{Z} = 75.96, \hat{b} = 0.0000000578, \hat{\gamma} = -0.555; X_{(13)}^2 = 3.41, p = 0.996$). For the singulars, S is estimated to be 8363, for the plurals, it is estimated to huge (14459718668).

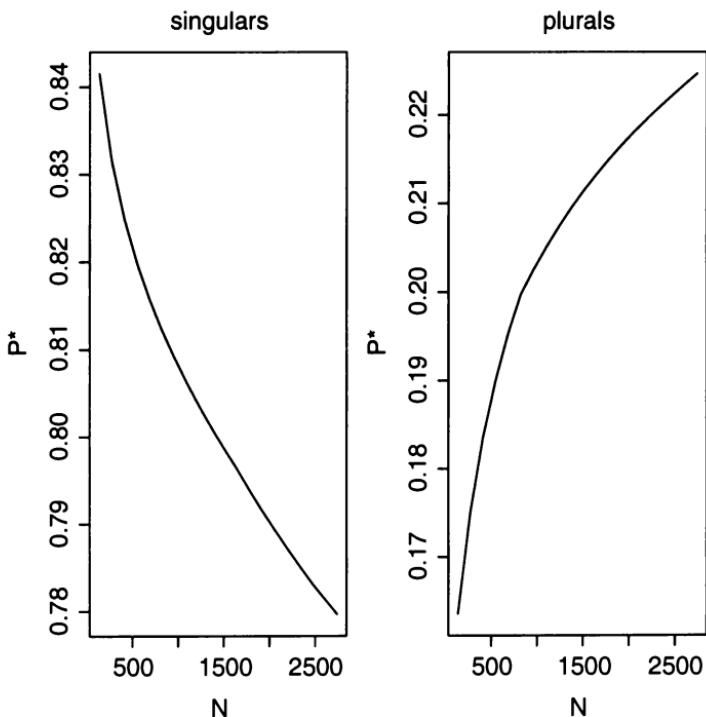


Figure 4.9: The contribution of singular nouns (left panel) and plural nouns (right panel) to the growth rate of nouns in Innes Appleby, P^* , as a function of the sample size N .

els for word frequency distributions that fall outside the LNRE zone, and for which, e.g., the mode has shifted from $m = 1$ to $m = 2, 3, \dots$. Without such models, no complete mixture analysis of a word frequency distribution can be completed.

An important aspect of modeling word frequency distributions as mixtures is that we can do justice in principle to the observation that even distributions that fall outside the LNRE zone still contain small numbers of even the lowest frequencies. There is no frequency cut-off point beyond which words belong to either an LNRE or a non-LNRE distribution. From this perspective, the proposal of Narayan and Balasubrahmanyam (1998) to model the head and the tail of the frequency spectrum with different functions is a first approximation at best.

4.6 Bibliographical Comments

Component distributions such as the nouns or the verbs in a given text have long been studied (see, e.g., Yule, 1944). A systematic study of how such component distributions combine and in what way they contribute to the properties of the overall mixing distribution has yet to be undertaken. The hypothesis that morphological categories are responsible for the LNRE property of word frequency distributions can be found in Chitashvili and Baayen (1993). Baayen and Lieber (1997) call attention the phenomenon that word frequencies may have bimodal densities, without developing a formal theory for LNRE mixtures. Tweedie and Baayen (1999) is a technical study of mixture distributions. A statistical textbook on mixture models is Titterington, Smith, and Makov (1985).

4.7 Questions

1. Show that

$$\begin{aligned}\text{COV}[V(m, N), V(k, N)] &= \text{COV}[V(m, pN), V(k, pN)] \\ &\quad + \text{COV}[V(m, (1-p)N), V(k, (1-p)N)].\end{aligned}$$

2. Discuss the bulge in the *-heid* density with respect to the mixture model.
3. Prove Theorem A for the generalized inverse Gauss-Poisson model.
4. Prove Theorem C for the lognormal model.
5. Provide an explanation for the large estimate of S for plural nouns, which is larger than that for singular nouns even though formally plurals are coined from singulars.

Chapter 5

The Randomness Assumption

The non-parametric and parametric models described in the preceding chapters all depart from the assumption that words appear randomly in texts. Obviously, this assumption is a simplification. Section 5.1 investigates the consequences of non-randomness in word use on the accuracy of the theoretical predictions of our models. Section 5.2 shows how the accuracy of these models can be increased by taking the effects of non-randomness in word use into account.

5.1 The Randomness Assumption

SUMMARY *The models for word frequency distributions developed thus far all build on the randomness assumption. Words, however, do not occur randomly in texts. This section explores how departure from randomness affects the accuracy of theoretical models. For interpolation, an overestimation bias is observed, and for extrapolation an underestimation bias. Discourse-level cohesion in word use is shown to be responsible for these biases.*

In the preceding sections, we have repeatedly seen that the interpolated expected vocabulary size tends to overestimate the observed vocabulary size for a wide range of sample sizes. Figures 2.8, 3.4, and 3.6 illustrate this overestimation bias for *Alice in Wonderland*, which we observe for binomial interpolation and LNRE models alike. In this section, we will first trace this overestimation bias to a particular kind of violation of the randomness assumption involving lexical specialization. Next, we will study how the accuracy of other measures such as the estimated population number of types S and the Good-Turing frequency estimates are affected.

5.1.1 Non-randomness and lexical specialization

A preliminary question that arises when we consider the overestimation bias is whether the difference between the observed and expected vocabulary size is in fact significant for a wide range of sample sizes. In other words, does the overestimation bias signal a significant discrepancy between model and data that requires further study, or can we safely ignore it?

This question is easily answered for sample sizes $M \leq N/2$, when we allow ourselves to estimate $\text{VAR}[V(N)] = E[V(2N)] - E[V(N)]$ by $V(2N) - V(N)$. Figure 5.1 plots the overestimation bias $E[V(N)] - V(N)$ for *Alice in Wonderland*, highlighting those measurement points for which $V(N)$ is significantly smaller than its expectation:

$$\frac{|V(N) - E[V(N)]|}{\sqrt{V(2N) - V(N)}} > 1.96. \quad (5.1)$$

For this text, the overestimation bias indeed appears to be significant throughout the first half of the text.

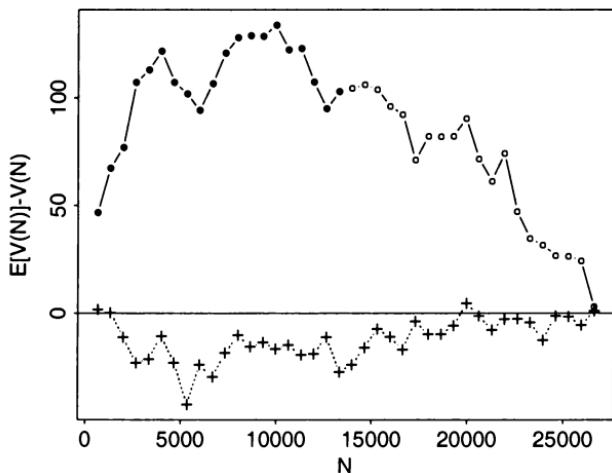


Figure 5.1: The overestimation bias $E[V(N)] - V(N)$ in *Alice in Wonderland* (solid line) and in a sentence-randomized version of *Alice in Wonderland* (dotted line). For the first half of the text, significant estimation errors are highlighted.

Given that the overestimation bias is significant, the question arises what kind of violations of the randomness assumption underlie its appearance. One possible explanation focuses on the role of sentential syntax in constraining word use. Syntactic rules impose many restrictions on the occurrence of words within the sentence. In normal written English, for instance, it is unlikely to

find two successive occurrences of the determiner *the*. The urn model, however, assigns the sequence *the the* a non-negligible probability of 0.0036 (the relative frequency of *the* is approximately 0.06), predicting that it should occur once on every other page. This is not what we find in normal texts. Clearly, the urn model has no predictive power for the occurrences of words in sentences. Does this imply that syntax is responsible for the overestimation bias?

To test this hypothesis, we may reason as follows. If syntax is truly responsible, the overestimation bias should remain present when we randomly permute the order of the sentences in a given text, without changing the order of the words within the individual sentences. Applied to *Alice in Wonderland*, sentence-randomization yields incoherent sequences of grammatical sentences such as:

For a minute or two she stood looking at the house and wondering what to do next when suddenly a footman in livery came running out of the wood. They all sat down at once in a large ring with the mouse in the middle. If you didn't sign it, said the king, that only makes the matter worse. The knave did so very carefully with one foot.

The dotted line in Figure 5.1 plots $E[V(N)] - V(N)$ for a sentence-randomized version of *Alice in Wonderland*. Interestingly, sentence randomization appears to remove the overestimation bias. At least for the first half of the text, the difference between the expected and observed vocabulary size never reaches significance, and the shape of the curve does not suggest that significant differences are to be expected for the second half of the text. We may conclude that sentential syntax is not responsible for the overestimation bias. This suggests that this bias must arise at higher levels of textual organization: paragraph structure, or even the discourse structure of the full text itself.

Discourse structure influences word use through shifts and changes in subject matter. Changes in topic bring along changes in vocabulary. Each topic domain invites its own specialized subvocabulary. Figure 5.2 illustrates this simple point by concatenating four completely different texts: L. F. Baum's *The Wonderful Wizard of Oz*, election speeches by and interviews with B. Clinton, J. M. Barrie's *Peter Pan*, and L. Carroll's *Alice in Wonderland*. The dotted line represents the vocabulary size $V(N)$ for this composite text, the solid line plots the corresponding expectations. The dotted vertical lines indicate the transitions between the texts. The transition between *The Wonderful Wizard of Oz* and the Clinton texts is marked by a sudden increase in the growth rate of the vocabulary, the other transitions are barely visible to the eye as slight irregularities in the growth curve at best. Clearly, the Clinton texts introduce large numbers of words in the domain of politics and economics than are not used by Baum. Conversely, given the combined texts of Baum and Clinton, the other two novels do not contribute large numbers of new words to the vocabulary that has already been established, even though the vocabulary size is still increasing substantially.

Crucially, the overestimation bias is huge for especially the first half of the composite text. This is due to the concentration of Clinton's political vocab-

ulary in just one part of the text. Instead of being spread out evenly over the complete text, the sudden influx of political and economic terms is concentrated in a limited part of the text, introducing an abrupt change in the growth curve. Instead of contributing from the start to $V(N)$, this subvocabulary starts to appear only after some 40000 tokens. Unsurprisingly, $E[V(N)]$, which is based on the assumption of a homogeneous use of words, overestimates the observed vocabulary size $V(N)$.

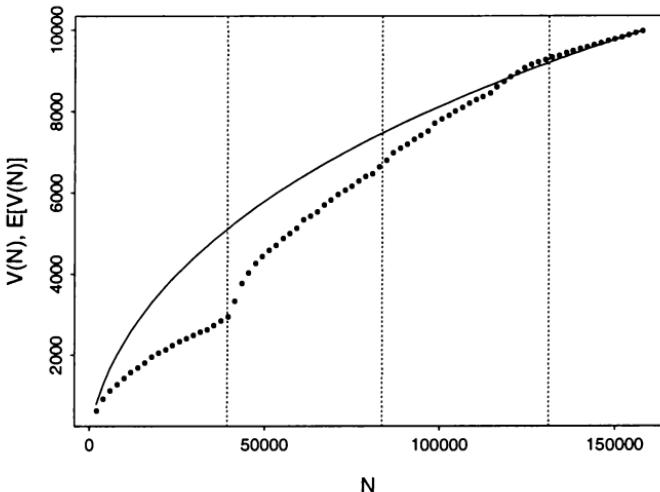


Figure 5.2: Empirical (dotted line) and expected (solid line) growth curves for the concatenated texts of L. F. Baum's *The Wonderful Wizard of Oz*, election speeches by and interviews with B. Clinton, J. M. Barrie's *Peter Pan*, and L. Carroll's *Alice in Wonderland*, for 80 measurement points. The dotted vertical lines indicate the transition points between the texts.

To make this line of reasoning more precise, we need to formalize the notion of lexical specialization. To do so, it is convenient to introduce the concept of underdispersion. When we partition a text into K text chunks, the dispersion d_i of a word ω_i is defined as the number of text chunks in which this word is found.

Definition 5.1 The dispersion d_i of a word ω_i is the number of different text chunks in which ω_i occurs.

A word is **underdispersed** if its observed dispersion is smaller than its expected dispersion. Since underdispersed words occur in fewer text chunks than expected under chance conditions, they are the words the tokens of which occur concentrated in particular parts of the text, instead of being spread out evenly throughout the text. Hence, the underdispersed words are the words

that, if our hypothesis is correct, should be responsible for the overestimation bias.

Table 5.1: Dispersion d_i , expected dispersion $E[d_i]$, frequency $f(i, N)$, and Monte Carlo probability of the dispersion $\Pr(d \leq d_i)$ for selected words in Alice in Wonderland for $K = 40$ equally-sized text chunks.

w_i	d_i	$E[d_i]$	$f(i, N)$	$\Pr(d_i)$	w_i	d_i	$E[d_i]$	$f(i, N)$	$\Pr(d_i)$
the	40	40.0	1629	1.000	then	34	36.3	94	0.132
and	40	40.0	864	1.000	no	35	35.9	90	0.401
a	40	40.0	628	1.000	know	33	35.7	88	0.091
she	40	40.0	540	1.000	them	32	35.7	88	0.029
said	39	40.0	461	0.000	queen	18	32.8	68	0.000
Alice	40	40.0	386	1.000	king	9	31.5	61	0.000
you	38	40.0	364	0.000	turtle	6	30.3	56	0.000
one	35	37.0	102	0.143	hatter	7	30.1	55	0.000
up	35	36.7	98	0.202	mock	7	30.1	55	0.000
his	29	36.5	96	0.000	gryphon	7	29.8	54	0.000

Table 5.1 lists the observed and expected dispersion of a selection of words in *Alice in Wonderland* along with their frequency and the probability $\Pr(d \leq d_i)$ of observing d_i or an even lower dispersion under chance conditions. Table 5.1 is based on a partition of *Alice in Wonderland* into 40 equally-sized text chunks ($K = 40$). The listed probabilities and expectations were calculated using Monte Carlo simulations. The text of *Alice in Wonderland* was randomized 1000 times. For each permutation run, the dispersion of each word in that particular permutation was obtained. The proportion of permutation runs for which the dispersion was smaller or equal to the empirical dispersion is listed in Table 5.1 as $\Pr(d_i)$. For instance, 29 of the 1000 simulation runs revealed a dispersion of 32 or less for *them*. Hence, *them* is significantly underdispersed at the 5% significance level.¹ High-frequency words, especially words for which $f(i, N) \gg K$, are likely to occur in all text chunks ($d_i = K$). This is what we find for, e.g., the determiners *the*, *a* and for *Alice*. Other words only occur in a subset of chunks. For very low-frequency words ($f(i, N) \ll K$), this is to be expected, but as the frequency of a word increases, its dispersion should likewise increase. Often, the key words of a text are underdispersed. Examples in Table 5.1 are *queen*, *king*, *turtle*, *hatter*, *mock*, and *gryphon*. *Queen*, for instance, occurs in 18 text chunks, but given a frequency of 68 it might be expected to occur in some 33 text chunks. Since none of the 1000 permutation runs revealed a dispersion less than or equal to 18, the probability that the observed dispersion $d = 18$ is due to chance is very small ($p < 0.001$).

If the underdispersed words are responsible for the overestimation bias, their removal from *Alice in Wonderland* should result in a text for which no overestimation bias is observed. Interestingly, after removal of the 5642 word tokens of the 345 words that are significantly underdispersed at the 5% sig-

¹We use a one-tailed test as it makes no sense to suppose that words would be overdispersed. For an analytic approach for evaluating underdispersion, see Baayen (1996b).

nificance level, we obtain an apocopated version of *Alice in Wonderland* for which the observed and expected vocabulary size indeed are no longer significantly different. This is shown in Figure 5.3. The solid line, which represents the expected growth curve, is only slightly higher than the curve of the observed vocabulary sizes (dashed line). Using (5.1), we find that the difference between the two curves is not significant for the first 20 measurement points, and visual inspection of Figure 5.3 suggests that the same probably holds for the second half of the text. In other words, we can indeed bring the text in line with the urn model by removing those words the dispersion of which reveals that they violate the randomness assumption of the urn model.

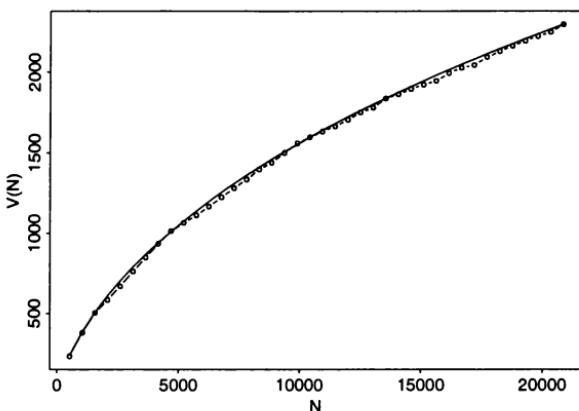


Figure 5.3: Empirical (dashed line) and expected (solid line) growth curves of the non-underdispersed words in *Alice in Wonderland*.

Given that the underdispersed words are responsible for the overestimation bias that we have observed for *Alice in Wonderland*, the question arises in what way the underdispersed words introduce this bias. To answer this question, it is useful to study in some more detail how the types and tokens of the underdispersed words are distributed over the 40 chunks into which we partitioned *Alice in Wonderland*. What we do not expect to find is a homogeneous, uniform distribution. A uniform distribution is predicted by the urn model, and would not explain how the underdispersed words cause the overestimation bias for $E[V(N)]$.

In order to study the distribution of underdispersed words over the $k = 1, 2, \dots, 40$ text chunks, let $f_K(i, k)$ denote the frequency of word ω_i in text chunk k , and let

$$d_{i,k} = \begin{cases} 1 & \text{iff } \omega_i \text{ underdispersed and } f_K(i, k) > 0 \\ 0 & \text{otherwise,} \end{cases} \quad (5.2)$$

so that the number of underdispersed types in text slice k , $VU(k)$, and the corresponding number of underdispersed tokens, $NU(k)$, can be defined as

$$VU(k) = \sum_i d_{i,k} \quad (5.3)$$

$$NU(k) = \sum_i d_{i,k} \cdot f_K(i, k). \quad (5.4)$$

Figure 5.4 plots $VU(k)$ (left panel) and $NU(k)$ (right panel) for *Alice in Wonderland*. The dotted lines, least squares regression lines, suggest that the numbers of underdispersed types and tokens tend to increase with increasing text length ($F(1, 38) = 3.295, p < 0.08$ for the types, $F(1, 38) = 27.09, p < 0.001$ for the tokens). The solid lines represent non-parametric smoothed curves that do not impose a linear structure on the distribution.² The non-parametric smoother supports the increase in token intensity for increasing k . For the underdispersed types, it suggests a slightly more complicated pattern, with a lack of underdispersed types for the very first chunks, and again around chunk 15. The overall pattern suggested by Figure 5.4 is that in *Alice in Wonderland* the underdispersed words are underrepresented most clearly in the initial parts of the novel.

Interestingly, the number of expected underdispersed words in the first text chunk, $E[VU(k)]$, calculated by applying (2.42) to the frequency spectrum of the set of underdispersed words, equals 78.14. The observed number of underdispersed words in this text chunk equals 39. Hence, we have an overestimation bias $E[VU(k)] - VU(k)$ of some 39 word types. The overestimation bias for the underdispersed words is of the same order of magnitude as the overestimation bias $E[V(N)] - V(N)$ observed for the text (including both normally dispersed and underdispersed words): 45. Clearly, this general overestimation bias of $k = 1$ is for the most part due to the overestimation of the subset of underdispersed words. For the next chunks, similar calculations reveal additional deficits in underdispersed words, which cumulate with the deficits in the preceding chunks to form an increasingly large overestimation bias for $E[V(N)]$. Conversely, the later text chunks are characterized by a surplus of underdispersed types and tokens, and therefore contribute to decreasing this deficit. By the end of the text, the deficit is completely cancelled out, and the overestimation bias has vanished: $E[V(N)] = V(N)$.

A note of caution, however, is required here. The consistent overestimation bias ($E[V(N)] > V(N)$) that we have observed for *Alice in Wonderland* is but one of the different error patterns that one finds when comparing the expected and observed vocabulary growth curves. For a detailed study of more complicated patterns, for which the precise way in which the underdispersed words effect the divergence between $E[V(N)]$ and $V(N)$ is often more difficult to discern, the reader is referred to Baayen (1996a).

5.1.2 Consequences of non-randomness

The overestimation bias that we have observed for *Alice in Wonderland* is not without consequences for other statistics such as the estimated number of types in the population S and the Good-Turing estimates. First consider how S is affected. The immediate consequence of the overestimation bias for the

²The solid lines were obtained by means of time series smoothing with running medians (Tukey 1977).

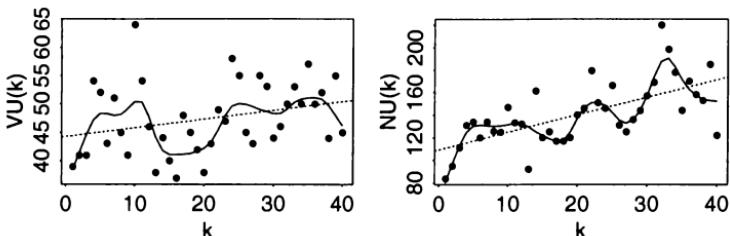


Figure 5.4: Number of underdispersed word types (left panel, $VU(k)$) and tokens (right panel, $NU(k)$) for the partition of $k = 1, \dots, 40$ equally-sized chronologically ordered text chunks of *Alice in Wonderland*. The dotted lines represent least squares regression lines, the solid lines a non-parametric time series smoother.

known ‘sampling time’ $M = 1, 2, \dots, N$ is an underestimation bias for $M > N$. This is illustrated for *Alice in Wonderland* in Figure 5.5. In order to compare the observed with the extrapolated values of $V(N)$, we choose as our point of departure the frequency spectrum of *Alice in Wonderland* at exactly halfway through the book at $N = 13253$. As LNRE model, we opt for a generalized inverse Gauss-Poisson fit to the first 10 spectrum elements, which, for $\hat{b} = 0.0194$, $\hat{Z} = 92.01$, and $\hat{\gamma} = -0.5$ seems reasonable ($X_9^2 = 26.48, p = 0.0017$). The upper left panel of Figure 5.5 visualizes this fit. Using (3.26), we calculate $E[V(N)]$ for 20 measurement points that jointly partition the full text of *Alice in Wonderland* into 20 equally large chunks. The upper right panel of Figure 5.5 shows the empirical growth curve (dotted line) and its expectation as calculated from the frequency spectrum at $N = 13253$. The bottom panel highlights the estimation errors by plotting $E[(V)] - V(N)$. As expected, we again find an overestimation bias for $M < 13253$, but for $M > 13253$, the error is now reversed and surfaces as an underestimation error. For instance, the model predicts a vocabulary size of 2531.03 for $N = 26505$, whereas in fact $V(26505) = 2651$, a difference of 120 word types. From this, we can infer that the population number of types $S = \lim_{N \rightarrow \infty} E[V(N)]$ is likewise underestimated.

The non-randomness in word use exhibited by the underdispersed words also affects the Good-Turing estimates. We again take the frequency spectrum of *Alice in Wonderland* at sample size $N/2$ as our point of departure, and use the above fit of Sichel’s model to calculate the expectations of $V(m, N/2)$. Given these smoothed values, we calculate the Good-Turing adjusted frequency estimates m^* using (2.27). How do these estimates compare with the known frequencies of the words involved in the complete text, our ‘population’? To answer this question, we first calculate the summed frequencies in the full text ($N = 26505$) of all words with frequency m at sample size $N/2$, $s(m)$:

$$s(m) = \sum_{i=1}^{V(N)} I_{\{f(i, N/2) = m\}} \cdot f(i, N).$$

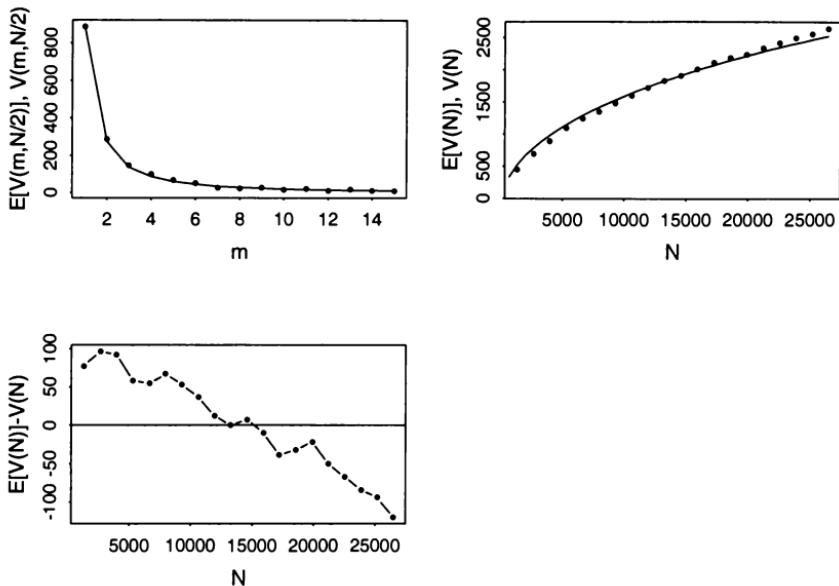


Figure 5.5: Interpolation and extrapolation accuracy for Alice in Wonderland, conditioning on the frequency spectrum at $N = 13253$. The upper left panel shows the fit of Sichel's model to the frequency spectrum ($b = 0.0194$, $Z = 92.01$). The upper left panel plots the observed (dotted line) and expected vocabulary size given the Sichel fit. The bottom panel plots the estimation errors $E[V(N)] - V(N)$.

The average frequency in the complete text of the words for which $f(i, N/2) = m$ equals

$$s(m)/V(m, N/2).$$

For sample size $N/2$, this average frequency has to be adjusted for the difference in sample size as we want to compare it with the frequency m at the sample size $N/2$. Assuming that the tokens of the words for which $f(i, N/2) = m$ are uniformly distributed over the full text, we have that the population-based frequency m^o can be defined as

$$m^o = \frac{1}{2} \frac{s(m)}{V(m, N/2)}. \quad (5.5)$$

Table 5.2 lists m , m^* , and m^o for the first 10 spectrum elements at $N = 13253$. Note that the observed frequencies m overestimate their population values m^o , whereas the first five Good-Turing estimates underestimate them. A heuristic measure that seems to work fairly well for the lowest values of m is to take the average of m and m^* , \bar{m} , as an estimate of a word's frequency that takes into account that texts violate the randomness assumption. As shown in Table 5.2, the resulting estimate \bar{m} is closer to the population value m^o than

Table 5.2: Comparison of frequency estimates: m^* is the Good-Turing estimate, m^o is the sample-size adjusted population frequency, and \bar{m} is the average of m^* and m . The expectations $E[V(m, N/2)]$ are based on the extended Zipf's law.

$V(N/2)$	$E[V(N/2)]$	m	m^*	m^o	\bar{m}
888	865.991	1	0.702	0.840	0.851
286	303.826	2	1.548	1.719	1.774
146	156.732	3	2.451	2.644	2.726
98	96.054	4	3.385	3.923	3.692
68	65.029	5	4.336	4.434	4.668
50	46.995	6	5.298	5.060	5.649
25	35.572	7	6.269	6.280	6.635
22	27.873	8	7.244	5.886	7.622
26	22.435	9	8.224	7.500	8.612
15	18.451	10	9.207	8.000	9.604

the Good-Turing estimates for $m < 5$. For other materials, the range of values for which \bar{m} seems reasonable is somewhat larger (see Baayen (1996a)).

Figure 5.6 illustrates how the effect of lexical specialization can even emerge in the plot of the real-valued approximate spectrum elements and the plot of the Good-Turing estimate m^* . Using the data from *The Independent* (see section 2.5), and focusing on the spectrum elements with $m > 55$, we observe a slight downward shift for $V_r(m, N)$ and a slight upward shift for m^* when we compare spectrum elements for which $m < 250$ (marked by the left dashed line) and the spectrum elements for which $m > 500$ (marked by the right dashed line). When we fit a straight line to the spectrum elements with $55 < m \leq 250$, we obtain the upper solid line in the left panel. The corresponding line in the right panel is the lower solid line. Similarly, the straight line fitted to the spectrum elements with $m > 500$ is the lower line in the left panel. The corresponding line in the right panel for m^* is the upper solid line.

The highest-frequency ranks in the distribution are populated predominantly by function words, words with primarily a grammatical function such as *the*, *a*, *this*, *over*, *my*, to be distinguished from content words such as *walk*, *green*, *house*, and *armchair*. As shown in Table 5.3, the content words are scarce among the highest Zipf ranks, while they predominate among the lower Zipf-ranks.³ Comparing the two regression lines, we see that there are more content words with high values of m than one would expect on the basis of the trend visible among the function words. Turning to the right panel, we see that Good-Turing estimates for the function words as a group are higher than the Good-Turing estimates for the content words. What we observe is the effect of lexical specialization affects the content words but not the function words. Function words are required in any text, and tend to have full dispersion. By contrast, content words tend to be topic-specific. They tend to oc-

³For studies of the information-theoretic difference between function words and content words see Baayen (1991) and Naranan and Balasubrahmanyam (1996, 1998),

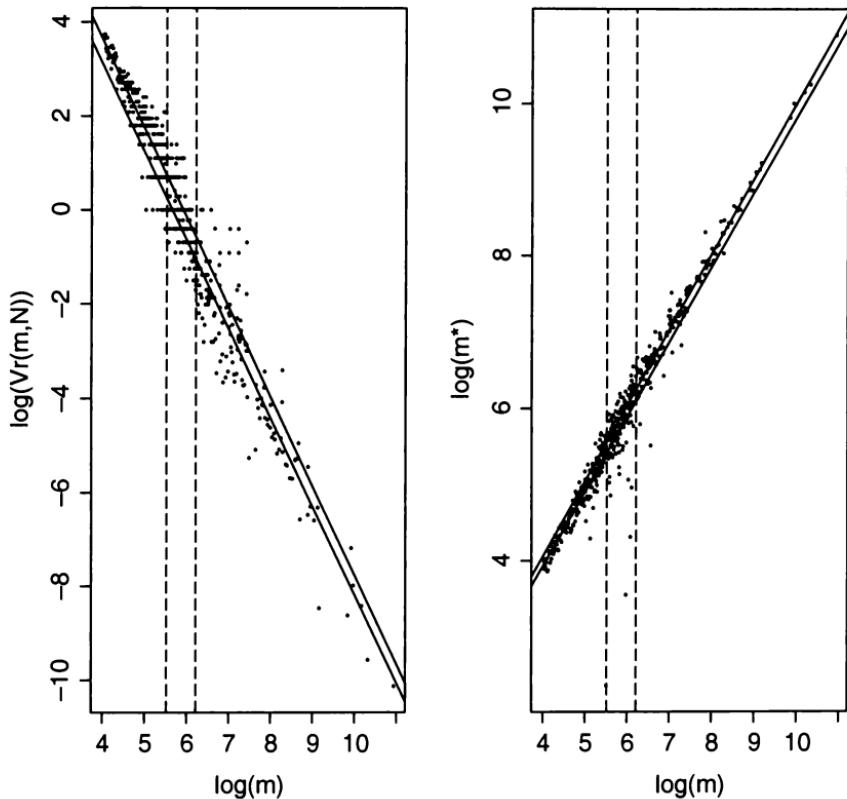


Figure 5.6: Left panel: part of the real-valued approximate frequency spectrum of the 1 million sample of The Independent of 1989 in the double logarithmic plane. The upper solid line represents the least-squares linear regression line to the frequency range to the left of the first dashed vertical line, the lower solid line is a least squares regression line to the right of the second vertical dashed line. Right panel: the corresponding plot for m^* . The upper line is fit to the spectrum elements to the right of the second dashed line, the lower line is a fit to the spectrum elements to the left of the first dashed line.

cur with slightly higher frequencies in the subcorpus of *The Independent* from 1989 due to the cohesive structure of normal texts. Hence, when we calculate the average frequencies of these words in other independent samples of the same newspaper, we observe that these frequencies tend to be somewhat lower compared to those of the function words. Thus, the effect of lexical specialization and non-randomness can be discerned in the $m-m^*$ plot and even in plot of the real-valued approximate frequency spectrum itself.

Table 5.3: *Number of content words among the most frequent Zipf ranks in groups of 50. The Frequency and Log Frequency column pertain to the highest Zipf rank in each group.*

Zipf Ranks	Content	Frequency	Log Frequency
1 – 50	1	56038	10.94
51 – 100	14	1657	7.41
101 – 150	18	813	6.70
151 – 200	31	509	6.23
201 – 250	28	398	5.99
250 – 300	34	349	5.86
301 – 350	38	308	5.73
351 – 400	42	275	5.62
401 – 450	44	241	5.48
451 – 500	38	224	5.41
501 – 550	45	209	5.34

Finally, consider the growth rate of the vocabulary which, as we have seen in section 2.5, can be viewed as an estimate of the probability mass of unseen word types given the assumptions of the urn model. For *Alice in Wonderland* at $N = 13253$, we find that $\mathcal{P}(13253) = 888/13253 = 0.067$, or, using the expected number of hapaxes, $865/13253 = 0.065$. In fact, however, the summed relative frequency of the unseen types in the complete text equals 0.063, a slightly lower value. Comparisons for other texts also suggest that $\mathcal{P}(N)$ tends to overestimate the probability mass of the unseen types.

It is easy to see why \mathcal{P} appears to be an upper bound for coherent text by focussing on its interpretation as the probability of the unseen types. Given the urn model, the probability that the very first token sampled represents a type that will not be represented by any other token is $E[V(1, N)]/N$. By symmetry, this probability is identical to the probability that the very last token sampled represents a new type. In turn, this probability approximates the probability that, when we increase the sample with one additional token, this token will represent a new type. However, real texts do not follow the urn model. Our analysis of underdispersion shows that, once used, many words are much more likely to be used again in the immediate context than expected under chance conditions. Hence, the probability that after sampling N tokens the $N + 1$ -st token will represent a new type is less than $E[V(1, N)]/N$. Due to the cohesive structure of texts, the words that have already been used have a slightly higher probability of being used again than predicted by the urn model.

Recall that according to the urn model, $\mathcal{P} = E[V(1, N)]/N$ simultaneously quantifies both the growth rate of the vocabulary and the probability mass of unseen types. Non-randomness in word use drives a wedge between these two interpretations of \mathcal{P} . The probability mass of unseen types is **smaller** than $E[V(1, N)]/N$. Conversely, the growth rate of the empirical vocabulary size is **larger** — as illustrated in Figure 5.5, the empirical curve grows more quickly and reaches higher values in extrapolation than predicted by the urn model. In other words, texts have even more rare words, and these words have even lower sample relative frequencies, than expected on the basis of the urn model.

5.2 Adjusted LNRE models

SUMMARY This section discusses two techniques for adjusting LNRE models for the non-randomness in word use. The first technique introduces a new parameter expressing the amount of lexical specialization. The second technique introduces link functions that takes the parameters of the LNRE models themselves to be functions of the sample size. Both techniques lead to a substantial improvements in interpolation and extrapolation accuracy.

We have seen that the non-random use of words in discourse affects the accuracy of the predictions of LNRE models with respect to interpolation and extrapolation. This leads to the paradoxical situation that an LNRE model may provide a reasonable fit for the frequency spectrum, while revealing salient errors for the developmental profiles of the vocabulary size and the spectrum elements $m = 1, 2, 3, \dots$. Consider, for instance, Figure 5.7, which shows two diagnostic plots for the lognormal model applied to *Alice in Wonderland*. The left panel shows the observed and expected frequency spectrum. The observed spectrum elements are represented by dots, the expected spectrum is represented by a solid line.

However, when we consider the right panel of Figure 5.7, a different picture emerges. This panel plots the empirical developmental profiles of $V(N)$ and $V(m, N)$ for $m = 1, 2, \dots, 5$ by means of dots, and the corresponding expected profiles by means of solid lines. Observe that the theoretical curves for the vocabulary size, the hapax legomena, and the dis legomena all overestimate their empirical analogs, even though $E[V(N)] \approx V(N)$ and $E[V(1, N)] \approx V(1, N)$ for the last measurement point. Clearly, LNRE models need to be adjusted to take the effect of the non-random use of words in discourse into account. Otherwise, either the developmental profile of the frequency spectrum and the vocabulary size, or the frequency spectrum at the final text size will be modeled adequately, but not both simultaneously.

There are two different ways in which this problem of adjusting for non-randomness can be approached. In the first approach we partition the vocabulary into lexically specialized words on the one hand, and normal, unspecialized words on the other, and formulate separate submodels for both sets. In the second approach, we enrich LNRE models with a special parameter that

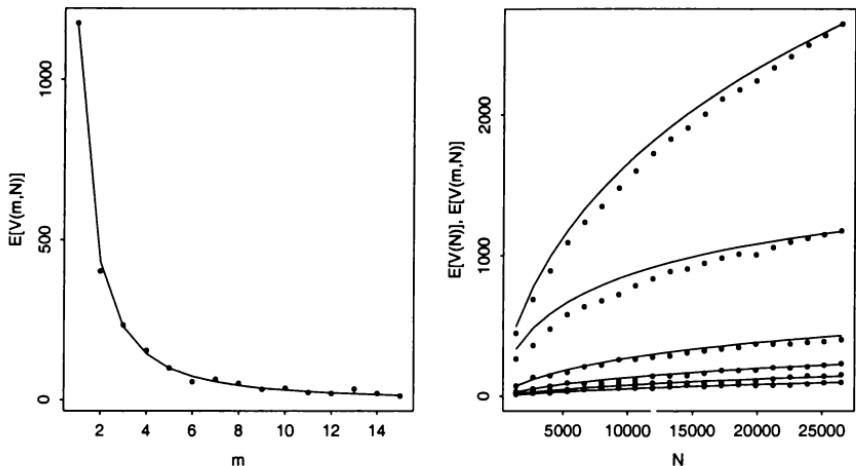


Figure 5.7: Spectrum and developmental profile of $V(N)$ and the first five spectrum elements for Alice in Wonderland. The dots represent the empirical values, the solid lines the corresponding expected values for the lognormal model.

models the effects of non-randomness on the growth curve of the vocabulary and the developmental profile of the spectrum elements. Section 5.2.1 explores partition-based adjustment, and section 5.2.2 parameter-based adjustment for the extended Zipf's law and the generalized inverse Gauss-Poisson model with γ fixed at -0.5 a priori.

5.2.1 Partition-based adjustment

For partition-based adjustment, we take as our point of departure the model proposed by Hubert and Labbe (1988) for adjusting the binomial interpolation model. In section 2.6.1, the following expressions for binomial interpolation of the vocabulary size and the spectrum elements were introduced:

$$E_{bin}[V(N)] = \sum_m V(m, N_0) \left(1 - \left(1 - \frac{N}{N_0}\right)^m\right) \quad (5.6)$$

$$E_{bin}[V(m, N)] = \sum_{k \geq m} V(k, N_0) \binom{k}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{k-m} \quad (5.7)$$

Hubert and Labbe (1988) proposed the following adjustment of these interpolation formulas:

$$\begin{aligned} E_{adj}[V(N)] &= p \frac{N}{N_0} V(N_0) + (1-p) \sum_m V(m, N_0) \left(1 - \left(1 - \frac{N}{N_0}\right)^m\right) \\ E_{adj}[V(m, N)] &= p \frac{N}{N_0} V(m, N_0) \\ &= +(1-p) \sum_{k \geq m} V(k, N_0) \binom{k}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{k-m} \end{aligned} \quad (5.8)$$

These adjusted expressions for interpolation, in which the parameter p specifies the proportion of specialized word types, can be derived as follows. First, we partition the $V(N_0)$ words in a given text into specialized and non-specialized words. Let \mathcal{S} denote the set of specialized words, comprising $pV(N_0)$ types. We assume that the $(1-p)V(N_0)$ non-specialized words are, as before, binomially $(f(i, N_0), N/N_0)$ -distributed. This leaves us with the specialized words, the distribution or distributions of which are unknown. Hubert and Labbe (1988) make three simplifying assumptions that allow them to regard the specialized words as having a uniform distribution. The first simplifying assumption is that the tokens of specialized words occur concentrated in a single text slice. The second simplifying assumption is that the text fragments in which specialized words appear are evenly spread out over the text. The third assumption is that each frequency rank m yields the same proportion p of specialized words, irrespective of whether we are dealing with hapax legomena or with words occurring a large number of times in the text. Jointly, these assumptions yield the simplest possible model for the distribution of specialized words: a uniform, random distribution in 'text time'.

Divide the text into two parts, P_1 with N tokens and P_2 with $N_0 - N$ tokens. Let

$$X_i = \begin{cases} 1 & \text{if } \omega_i \in \mathcal{S} \text{ and } \omega_i \text{ occurs in } P_1 \\ 0 & \text{otherwise,} \end{cases} \quad (5.9)$$

and let

$$Y_i = \begin{cases} 1 & \text{if } \omega_i \notin \mathcal{S} \text{ and } \omega_i \text{ occurs in } P_1 \\ 0 & \text{otherwise.} \end{cases} \quad (5.10)$$

The total number of types in P_1 is simply $\sum_i X_i + \sum_j Y_j$. If word $\omega_i \in \mathcal{S}$, then its $f(i, N_0)$ tokens will all appear in the same part of the text. The probability that they will appear in part P_1 is $\frac{N}{N_0}$. Hence

$$\begin{aligned} E[V(N)] &= E[\sum_i X_i + \sum_j Y_j] \\ &= \sum_{\omega_i \in \mathcal{S}} \Pr(X_i = 1) + \sum_{\omega_j \notin \mathcal{S}} \Pr(Y_j = 1) \\ &= \sum_{\omega_i \in \mathcal{S}} \frac{N}{N_0} + \sum_{\omega_j \notin \mathcal{S}} \left(1 - \left(1 - \frac{N}{N_0}\right)^{f(j, N_0)}\right) \end{aligned}$$

$$\begin{aligned}
 &= pV(N_0) \frac{N}{N_0} + (1-p) \sum_{\omega_j \notin S} \left(1 - \left(1 - \frac{N}{N_0}\right)^{f(j, N_0)}\right) \\
 &= pV(N_0) \frac{N}{N_0} + \sum_m (1-p)V(m, N_0) \left(1 - \left(1 - \frac{N}{N_0}\right)^m\right). \quad (5.11)
 \end{aligned}$$

The adjusted expression for the expected spectrum elements is obtained analogously, and is left as an exercise. The parameter p can be determined by minimizing the mean squared error for K measurement points:

$$\text{MSE} = \sum_{k=1}^K \frac{(V(N_k) - E_{bin}[V(N_k)])^2}{K}, \quad (5.12)$$

with E_{bin} denoting the binomial expectation.

The top panels of Figure 5.8 illustrate for *Alice in Wonderland* and Well's *The war of the worlds* that this simple adjustment of the binomial model may lead to a substantial improvement in the fit. Whereas the binomial model overestimates the vocabulary size and the spectrum elements, the adjusted model yields curves that closely follow the observed values.

This adjustment procedure has two disadvantages, however. First, some care is required with respect to the interpretation of p as a parameter of lexical specialization. For *Alice in Wonderland*, the optimal estimate of p equals 0.17. This value is somewhat higher than the proportion of words that are underdispersed at the 5% significance level, 0.13. As shown in Figure 5.3, the expected growth curve of the vocabulary for the remaining 87% of the words in *Alice in Wonderland* provides a good fit to the observed growth curve of the vocabulary. This suggests that the estimated proportion of 0.17 is too high. This overestimation would then be due to the simplifying assumptions of the adjusted model, at least one of which is clearly wrong for *Alice in Wonderland*. Whereas the adjusted model assumes that the specialized words are uniformly distributed in the text, specialized words are underrepresented in the initial chapters of *Alice in Wonderland*, as shown in Figure 5.4.

A second disadvantage is that the proportion p is itself subject to the dependency on the sample size that characterizes lexical measures in general. This dependency is illustrated in the bottom panels of Figure 5.8 for *Alice in Wonderland* and Wells' *The war of the worlds*. The dashed lines highlight the main trend in the development of p by means of a non-parametric smoother using running medians (Tukey, 1977). In *Alice in Wonderland*, p decreases up to $N = 20000$, after which point it slightly increases. In *The war of the worlds*, an initial decrease is followed by a steep increase, which levels off towards the end of the novel. This again illustrates that lexical specialization is not uniformly distributed through text time.

In spite of these shortcomings, partition-based adjustment seems a promising technique to incorporate into LNRE models in order to enhance not only interpolation but also extrapolation accuracy. As discussed in section 5.1.2, overestimation for interpolation goes hand in hand with underestimation when extrapolating (see, e.g., Figure 5.5). We therefore rewrite the partition-adjusted

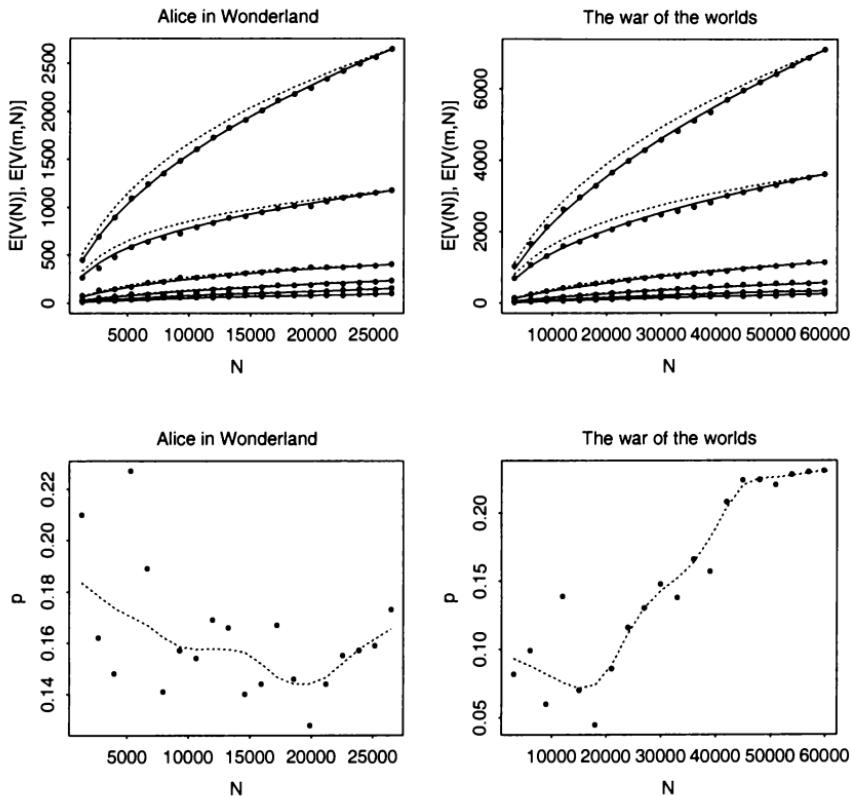


Figure 5.8: Partition-based interpolation for Carroll's Alice in Wonderland (left) and Wells' The war of the worlds (right). The upper panels plot the observed vocabulary size and the spectrum elements $m = 1, 2, \dots, 5$ as a function of N , using solid lines for partition-based interpolation, dotted lines for standard binomial interpolation, and dots for the observed values. The lower panels show the dependency of the partition parameter p on N , using dots for the observed values and a dashed line for a non-parametric smoother using running medians.

binomial expressions (5.9) in LNRE form,

$$\begin{aligned} E_{part}[V(N)] &= p \frac{N}{N_0} E_{LNRE}[V(N_0)] + (1-p) E_{LNRE}[V(N)], \\ E_{part}[V(m, N)] &= p \frac{N}{N_0} E_{LNRE}[V(m, N_0)] \\ &\quad + (1-p) E_{LNRE}[V(m, N)]. \end{aligned} \quad (5.13)$$

with E_{part} denoting the partition-adjusted expectation and E_{LNRE} the expectation given an LNRE model that adequately fits the frequency spectrum at a given sample size N_0 . Note that these expressions generalize over interpolation $N/N_0 < 1$ and extrapolation $N/N_0 > 1$.

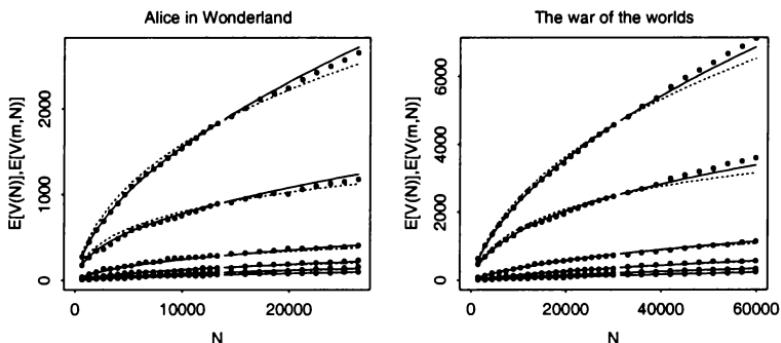


Figure 5.9: Partition-based adjusted Zipfian interpolation and extrapolation for Carroll's *Alice in Wonderland* (left, Yule-Simon fit) and Wells' *The war of the worlds* (right, Zipf fit). The observed vocabulary size and the spectrum elements $m = 1, 2, \dots, 5$ are represented by dots, their expectations based on the extended Zipf's law are shown by means of dotted lines, and their partition-based adjustments by means of solid lines.

Figure 5.9 shows the effect of partition-based adjustment on extrapolation accuracy for *Alice in Wonderland* and *The war of the worlds*. For both texts, N_0 was fixed at half the length of the complete novel. This allows us to use the second halves of these texts to check the accuracy of the model predictions. For both first halves, the specialization parameter p was determined by minimizing (5.12). The Yule-Simon model emerged as optimal for *Alice in Wonderland* ($X_{14}^{(2)} = 15.03, p = .38$, MSE=19.93). For *The war of the worlds*, the extended Zipf's law provided the best fit $X_{15}^{(2)} = 37.95, p = .001$, MSE=81.66). These models were subsequently used to provide the LNRE expectations in (5.13). As for partition-adjusted binomial interpolation, partition-adjusted LNRE interpolation leads to a fit that, at least to the eye, seems quite good. For extrapolation, we also observe an increase in accuracy, although here the ex-

pected extrapolation curves still deviate from the observed vocabulary size and spectrum elements. We observe slight overestimation for *Alice in Wonderland*, while for *The war of the worlds* we find more substantial underestimation. We conclude that the addition of the specialization parameter leads to a substantial improvement in accuracy, be it that the model's predictions for extrapolation remain slightly off-target.

The larger deviance observed for *The war of the worlds* is probably due to the larger difference between the values of the specialization proportion p . For *Alice in Wonderland*, the value of p at half the text size happens to be quite similar to that at the final text size (.166 versus .173). In the case of *The war of the worlds*, however, the value of p at the first half of the text is substantially smaller than that for the complete text (.148 versus .231). In other words, the assumption that lexical specialization is random, and that hence p is a sample-size invariant text characteristic, lies at the heart of the lack of extrapolation precision.

There are two ways in which we can proceed from here. One possibility is to parameterize the specialization proportion p as a function of the sample size N . Given a parametric function for $p(N)$ that captures fluctuations in lexical specialization in text time, better fits might be obtained. This approach has two disadvantages, however. First, the developmental profiles of p shown in Figure 5.8 are quite erratic, and suggest that extrapolation of $p(N)$ is hazardous. Second, parameterizing p adds at least two parameters to the underlying LNRE model. This raises the question whether it might be possible to achieve a similar gain in accuracy by directly adjusting a parameter of the LNRE model itself. This possibility is explored in the next section.

5.2.2 Parameter-based adjustment

In order to introduce parameter-based adjustment, we first return to a basic observation in Chapter 1, namely, that various measures of lexical richness and characteristic text constants vary systematically with the text length — in theory, in practice, due to non-random text structure, or both in theory and in practice. The parameter of lexical specialization p discussed in the preceding section again reveals such a dependence on N . In order to capture this theoretical dependence on N in a principled way, LNRE models have been developed. The parameters of these models are in principle invariant with respect to N , as long as the textual data to which these models are applied do not substantially violate the randomness assumption.

As an illustration of this theoretical sample-size invariance, consider again the data of the English newspaper *The Independent* discussed above in section 5.1.1. We have eight samples of one million words each. The first sample contains the first million words of the newspapers appearing in the last quarter of 1989, the next sample contains the first million words of the first quarter of 1990, the third sample contains the first million words of the second quarter of 1991, and so on. Thus, we have 8 relatively independent samples, which between them will have little discourse cohesion in common. Internally, each sample will be less cohesively structured than a novel, so that in all we may

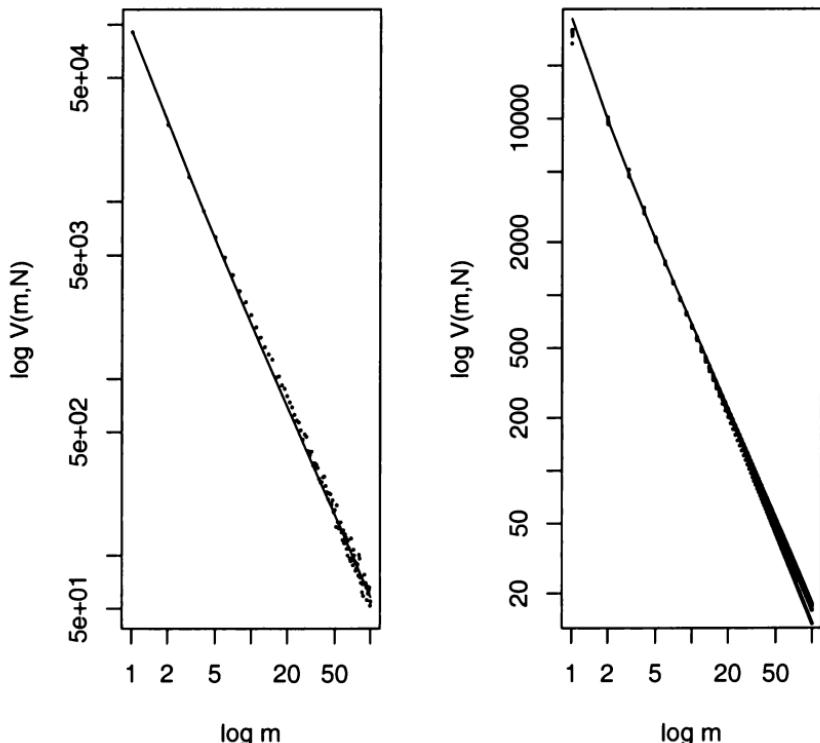


Figure 5.10: A generalized inverse Gauss-Poisson fit to the combined samples of The Independent (left panel) and the individual empirical spectra of the 8 individual samples of 1 million words (right panel). The solid line represents the expected spectrum as interpolated from the combined spectrum shown in the left panel.

expect that the randomness assumption is not too severely violated.

The left panel of Figure 5.10 displays the observed spectrum (dots) and expected spectrum (solid line) in the double logarithmic plane for the first 100 ranks of the combined word frequency distribution of all 8 samples. Parameters were estimated using cost function $C_2(100)$. The right panel shows the corresponding individual spectra of these samples (dots) and the expected spectrum as calculated on the basis of the fit shown in the left panel. Except for the hapax legomena, the number of which is clearly overestimated, the interpolated spectrum values fall within the range of the 8 empirical spectra of one million words.

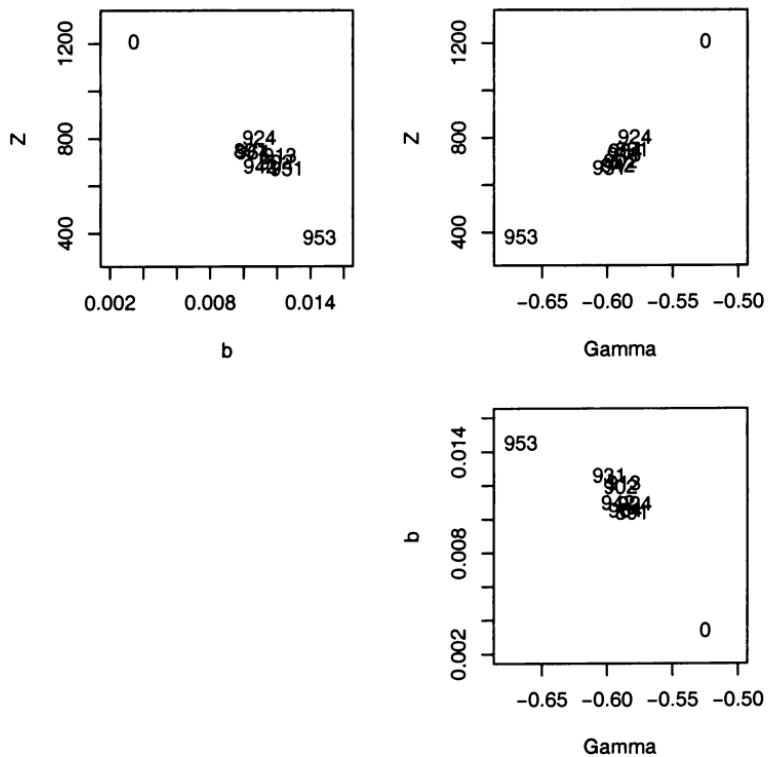


Figure 5.11: The three parameters of the generalized inverse Gauss-Poisson model for the eight one million word samples of The Independent (labeled 891 to 964) together with the parameters for the combined sample of 8 million words (labeled 0).

Figure 5.11 shows the scatter of the parameter values for the eight samples (labeled 891 to 964) and the combined sample of 8 million words (labeled 0).

Seven of the eight samples cluster closely together, the sample from the third quarter of 1995 (953) shows a high value for b and a low value for Z and γ . Conversely, the parameters of the combined data set are characterized by a low value of b and high values of Z and γ . Although the parameters of the combined word frequency distribution clearly assume outlier values, it is reassuring to find that there is one other outlier with nearly equally extreme values. Considered jointly, these results suggest that this kind of newspaper data does not violate the randomness assumption to such an extent that the generalized inverse Gauss-Poisson becomes inapplicable.

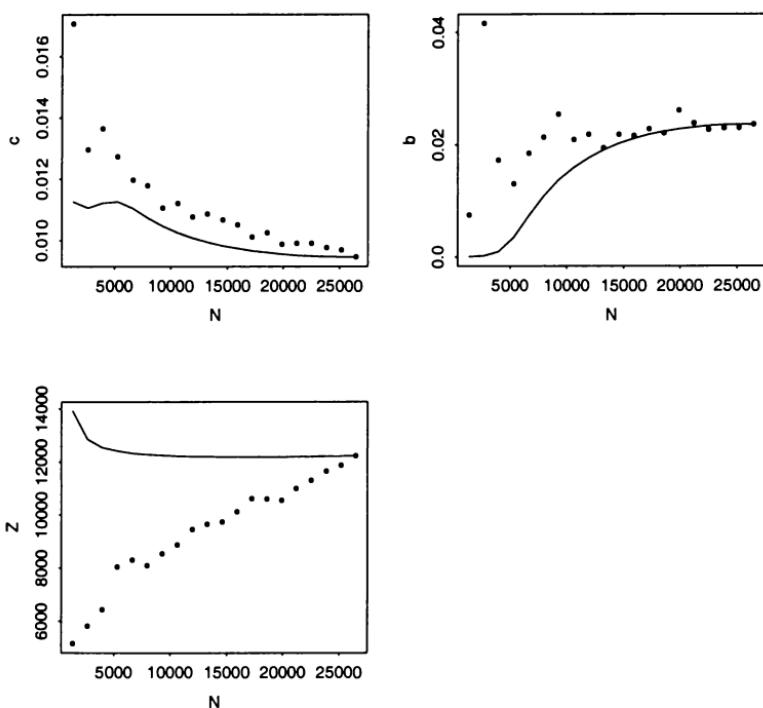


Figure 5.12: The dependence of LNRE parameters on the text length N illustrated for Alice in Wonderland. The top panels plot the developmental profiles of the parameters b and c of the generalized inverse Gauss-Poisson model, the bottom panel shows the profile of the single parameter Z of the extended Zipf's law. Observed profiles are represented by dots, their Monte-Carle expectations are represented by solid lines.

Let us now consider a text with a much more cohesive structure at the level of discourse, *Alice's Adventures in Wonderland*. How well do the parameters of LNRE models live up to the claim of sample size invariance when we study their developmental profiles in exactly the same way as we did for p and other textual constants?

Figure 5.12 suggests that LNRE parameters are just as dependent on the text length N as all other measures and constants that we have encountered thus far. Parameters were estimated using cost function C_1 in order to match $E[V(N)]$ as closely as possible with $V(N)$ itself. The upper panels show how the parameters b and c of the generalized inverse Gauss-Poisson model vary with N , the bottom panel shows the dependence of the Zipf size Z of the extended Zipf's law on N . The dots represent the parameter values as estimated for 20 equally-spaced measurement points. The solid lines show the corresponding Monte-Carlo means averaged over 5000 permutation runs. The observed value of c decreases with increasing text length, while the values of b and Z increase with N .

The theoretical curves for c and b reveal developmental profiles of a similar nature as the observed values.⁴ For the extended Zipf's law, we find a more stable pattern: following an initial sharp decrease, the value Z remains more or less fixed at 12000. Why do the Monte-Carlo developmental profiles of b , c , and Z show a functional dependence on N ? At first sight, it seems that this dependence points to some flaw in LNRE theory. This, however, is too strong a conclusion. In fact, the Monte-Carlo developmental profiles provide a test for goodness-of-fit. If an LNRE model provides a satisfactory fit, the developmental profiles of its parameters show up as horizontal lines. Functional dependence of the parameters on N arises when the model does not provide an adequate fit. For instance, the extended Zipf's law slightly underestimates the expected vocabulary size for small N compared to binomial interpolation. When we fit this model to these small text sizes, the parameter Z will be chosen such that it will optimally fit the vocabulary size at small N , hence, Z is assigned higher values. In other words, because the developmental profile is based on locally optimal fits, it brings to light how the parameters have to be adjusted in order to compensate for inadequacies in the fit as obtained on the basis of the spectrum of the complete text.

The empirical developmental profiles show how LNRE parameters have to be adjusted to locally optimize fits to texts governed by non-random word use. Again taking the extended Zipf's law as an example, we can see that overestimation of the expected vocabulary size for $N < N_0$ is compensated for by decreasing Z .

The empirical developmental profile of Z suggests a fairly simple functional dependence on N that might be captured in terms of a power model:

$$Z(N) = a_1 N^{a_2} \quad (5.14)$$

For *Alice in Wonderland*, least-squares parameter estimation suggests $a_1 = 625.4501$ and $a_2 = 0.2883$ as optimal ($r^2 = .980$). We can now replace Z in the theoretical expressions of the extended Zipf's law by the link function

$$Z(N) = 625.4501N^{0.2883}.$$

⁴Some care is required with respect to the interpretation of the theoretical curves for the first 5 measurement points. For these small text lengths, a substantial number of permutation runs failed to yield a fit satisfying $E[V(N)] = V(N)$ and $E[V(1, N)] = V(1, N)$. In other words, the first 5 Monte-Carlo means are conditional on the availability of a fit, and would probably have been lower for c and higher for b if the more extreme distributions could have been fitted.

For this link function, the expression for the expected vocabulary size (3.42) can now be written in the form

$$\begin{aligned} E[V(N)] &= \frac{Z(N)}{\log(p^*Z(N))} \frac{N}{N - Z(N)} \log(N/Z(N)) \\ &= \frac{a_1 N^{a_2+1}}{\log(p^*a_1 N^{a_2})} \frac{1}{1 - a_1 N^{a_2-1}} \log\left(\frac{N^{1-a_2}}{a_1}\right), \end{aligned} \quad (5.15)$$

effectively changing a one-parameter model into a two-parameter model. Of course, other choices for the link function will lead to different expressions for $E[V(N)]$.

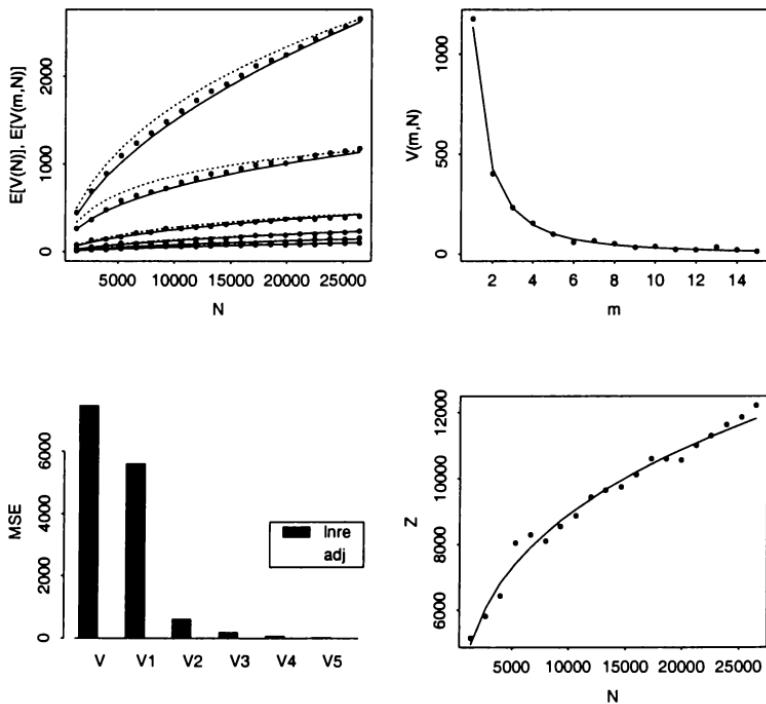


Figure 5.13: Diagnostic plots for Alice in Wonderland using a power link function to enhance the extended Zipf's law.

Figure 5.13 presents some diagnostic plots of the parameter-adjusted extended Zipf's law for *Alice in Wonderland*. The bottom right panel shows how Z increases as a function of N . The dots show the observed values, the solid line the power fit $Z(N) = 625.45N^{0.2883}$, which is just slightly better than a logarithmic fit. The power model for $Z(N)$ is not as good as one might wish: for large N , it appears to underestimate $Z(N)$. Nevertheless, adjusting Zipf's

law with a power link function leads to a substantial increase of goodness of fit. The upper left panel visualizes the growth curves of the vocabulary and the first five spectrum elements. The observed values are represented by dots, the values expected on the basis of the (unadjusted) extended Zipf's law by a dotted line, and the power-adjusted expectations by solid lines. The bottom left panel presents a bar chart of the mean squared error for the unadjusted (black bars) and the adjusted (grey bars) models for the vocabulary and the first five spectrum elements. Clearly, the power-adjustment leads to a much better fit. The upper right panel summarizes the observed (dots) and adjusted (solid line) spectrum elements at the full text size N_0 . The adjustment leads to a slight underestimation for the hapax legomena, but seems otherwise quite acceptable for the frequency ranks $m = 2, \dots, 15$. This impression is confirmed by the chi-square test, $X_{(14)}^2 = 27.77, p = 0.0153$, which suggests that this fit is roughly equivalent to the unadjusted fit ($X_{(15)}^2 = 29.05, p = 0.0158$). (It should be kept in mind, however, that this is due to the higher tolerance of the chi-square test for deviations of $E[V(N)]$ and $E[V(1, N)]$). By contrast, the mean squared error, 126.21 for the unadjusted spectrum, and 341.3 for the adjusted spectrum, more heavily penalizes the underestimation of the hapax legomena.)

Figure 5.14 presents a similar set of diagnostic plots for a hand-crafted power fit with parameters $a_1 = 617.68$ and $a_2 = 0.2898$. This choice of parameters leads to a real improvement of the fit ($X_{(14)}^2 = 21.69, p = 0.0852$, MSE = 165.51). With respect to the developmental profiles of $V(N)$ and $V(1, N)$, note that the slight underestimation observed for $E[V(N)]$ and $E[V(1, N)]$ in Figure 5.13 is substantially reduced. It is only for $E[V(2, N)]$ that the fit is slightly less good.

We can adjust the generalized inverse Gauss-Poisson model with γ fixed at -0.5 a priori in a similar way, although the introduction of a link function is somewhat less straightforward here due to the presence of two parameters instead of one. Of the parameters b and c , the latter has an interpretation similar to the parameter Z of the extended Zipf's law as locator in the LNRE zone. Since $1/c$ and Z also have similar developmental profiles, and since c tends to reveal smoother profiles than b (see, e.g., Figure 5.12), it is convenient to select c as the parameter for which to introduce a link function $c(N)$. In the unadjusted model, we can calculate b once we know c . From (3.27), and writing $g = (1 + Nc)^{-1/2}$ as before, we have that

$$b = \frac{\log(gN/E[V(1, N)])}{(1 - g)\sqrt{1 + Nc}} \quad (5.16)$$

When fitting the model to a complete text ($N = N_0$), we can allow ourselves to estimate $E[V(1, N_0)]$ by $V(1, N_0)$. Unfortunately, we do not know for arbitrary N greater or smaller than N_0 how to estimate $E[V(1, N)]$ from the frequency spectrum $\{V(m, N)\}$. Given a value for the adjusted parameter $c(N)$, we cannot straightforwardly apply (5.16) to obtain b . However, given $c(N)$, we can choose $b(N)$ such that at N_0 the expected number of hapax legomena will equal the observed number of hapax legomena at N_0 , in line with the use of cost function C_1 . In other words, writing $g(N_0) = (1 + N_0c(N))^{-1/2}$, we

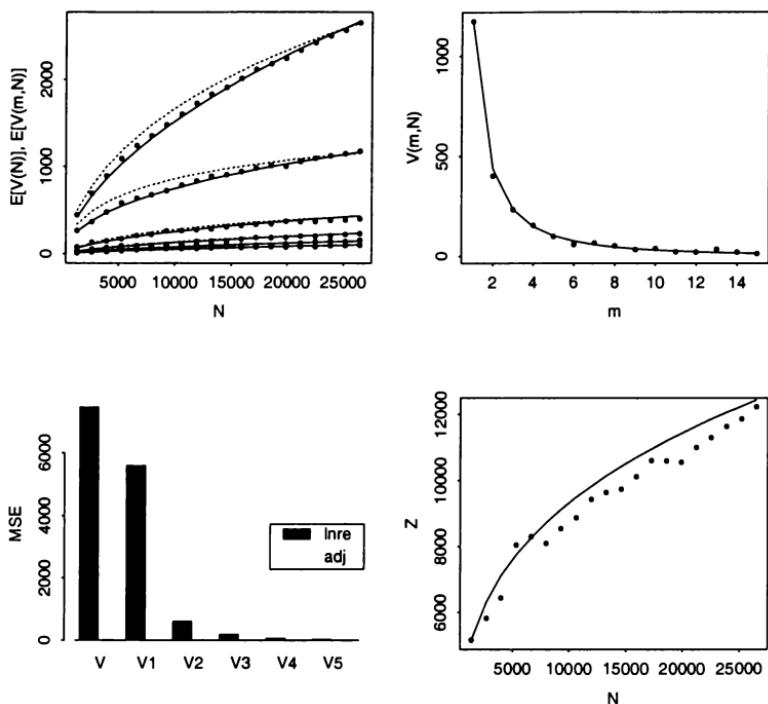


Figure 5.14: Diagnostic plots for the parameter-adjusted extended Zipf's law applied to Alice in Wonderland, with hand-tuned power fit.

estimate $b(N)$ as follows:

$$b(N) = \frac{\log(g(N_0)N_0/E[V(1, N_0)])}{(1 - g(N_0))\sqrt{1 + N_0c(N)}}, \quad (5.17)$$

replacing $E[V(1, N_0)]$ by $V(1, N_0)$.

Figure 5.15 illustrates the application of the parameter-adjusted generalized inverse Gauss-Poisson model to the 'A'-texts of the British National corpus, which comprise some 800000 word tokens. For this collection of texts, the extended Zipf's law is inadequate, notably so with respect to $V(1, N_0)$, which is severely underestimated. The parameter-adjusted fit for the spectrum elements at N_0 of Sichel's model is shown in the upper right panel of Figure 5.15. To the eye, the fit is very good. Due to the very large values of $V(N_0)$ and $V(m, N_0)$, however, both the chi-square and the mean squared error are very high, 1877.20 and 53807.04, respectively. Fortunately, these high values are not a source of worry, as the chi-square is known to become very large for large numbers of observations, and will almost always lead to the rejection of the Null-Hypothesis (see, e.g., Grotjahn and Altmann, 1993, for discussion). The

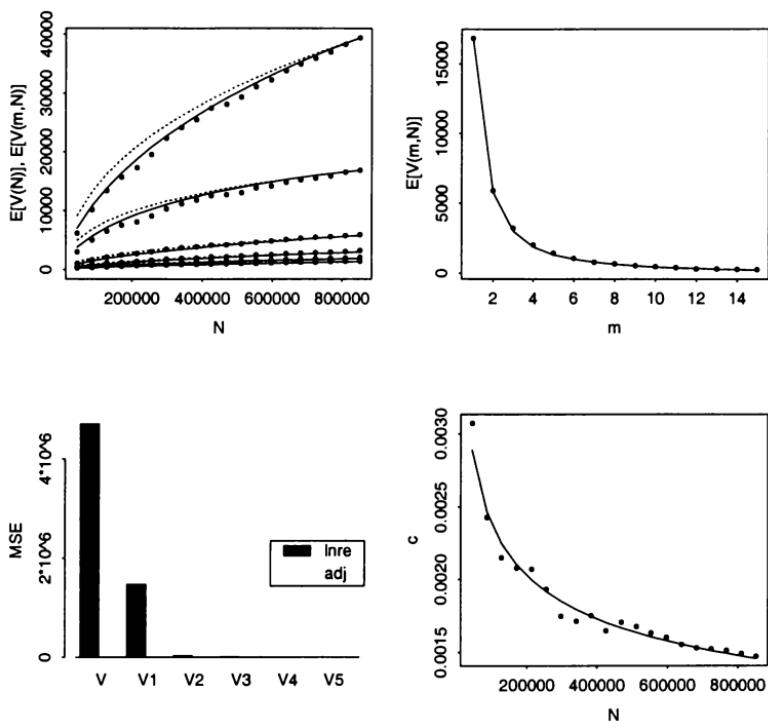


Figure 5.15: Diagnostic plots for the parameter-adjusted generalized inverse Gauss-Poisson model applied to the 'A'-texts of the British National Corpus.

left-hand panels of Figure 5.15 show that the use of a power link function for c ,

$$c(N) = 0.0334N^{-0.2295},$$

leads to a substantial increase in interpolation accuracy. This power fit is shown in the bottom right panel. It seems reasonable, except perhaps for the first measurement point. More accurate interpolation results can probably be obtained with a better link function.

Conan-Doyle's *Hound of the Baskervilles* is an example of a text to which LNRE models and their adjusted variants should probably not be applied. The top panels of Figure 5.16 strongly suggest a marked change in vocabulary structure around $N = 20000$. Both the growth curve of the vocabulary (left panel) and the growth curve of the hapax legomena (right panel) level off, to continue as more or less linear functions of N . Not surprisingly, this lexical structure gives rise to erratic developmental profiles for Z and c , as shown in the bottom panels. These unpredictable developmental profiles suggest that we are dealing with a non-homogeneous text, or at least with a text violat-

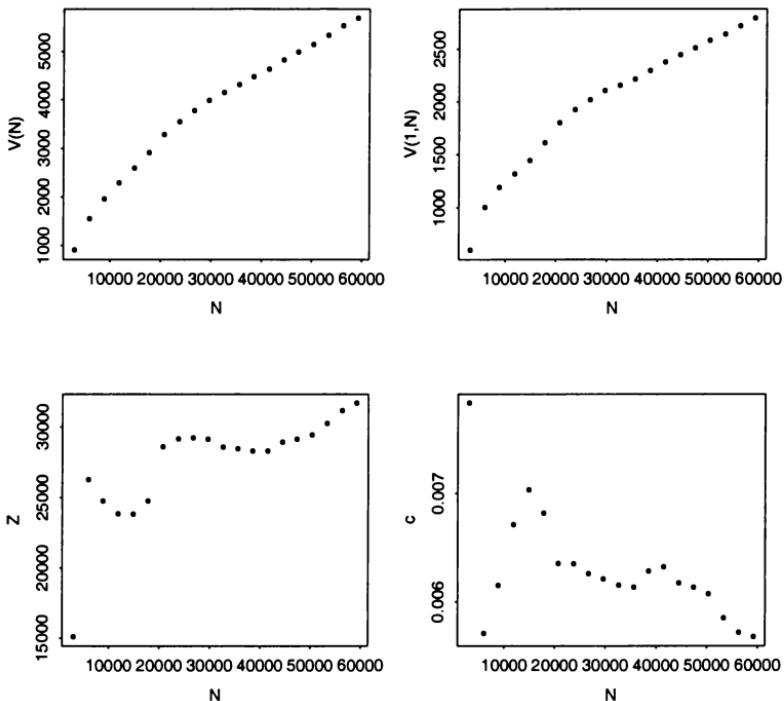


Figure 5.16: Non-homogeneity in the developmental profile: the case of the Hound of the Baskervilles. The upper left panel plots the growth curve of the vocabulary, and the upper right panel the growth curve of the hapax legomena. The bottom panels plot the developmental profiles of Z and c .

ing the randomness assumption in such a fundamental way that adjusting for non-randomness by means of link functions is no longer feasible.

We have not yet considered to what extent extrapolation accuracy is improved through parameter-based adjustment. Figure 5.17 shows the extrapolation results for the same texts considered earlier in the context of partition-based adjustment, *Alice in Wonderland* (left panel) and *The war of the worlds* (right panel). For both texts, both the extended Zipf's law (dotted line) and the parameter-adjusted extended Zipf's law (solid line) was fitted to the frequency spectrum and the developmental profile of Z at $N = N_0/2$. Subsequently, both models were used to extrapolate from $N_0/2$ to N_0 . Given the observed values (represented by dots), parameter-adjustment leads to a substantial increase in extrapolation accuracy. For *Alice in Wonderland*, we observe a slight overestimation for $E[V(N)]$ and $E[V(2, N)]$, which is probably largely due to a non-perfect power fit for $Z(N)$. In the case of *The war of the worlds*, however, the adjustment seems right on target. The underlying linear fit for

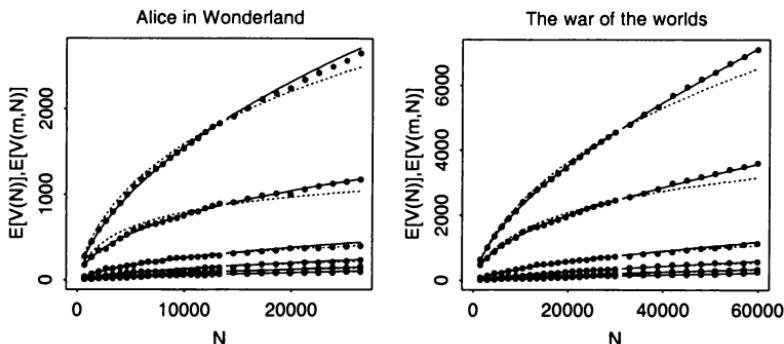


Figure 5.17: Extrapolation accuracy of the parameter-adjusted extended Zipf's law for Alice in Wonderland (left panel) and Wells' The war of the worlds (right panel), fitted at $N = N_0/2$. The dots plot the observed values, the dotted lines unadjusted interpolation and extrapolation, and the solid lines adjusted interpolation and extrapolation.

$Z(N)$ apparently leads to quite satisfactory extrapolation accuracy up to at least $2N_0$.

Table 5.4 summarizes some details on the fits obtained for the extended Zipf's law, the generalized inverse Gauss-Poisson model with γ fixed at -0.5 , and their parameter-adjusted variants, for five different texts: *Alice in Wonderland*, *Through the looking-glass and what Alice found there*, Wells' *The war of the worlds*, and Conan Doyle's *Hound of the Baskervilles*. For each fit, Table 5.4 provides details on the mean squared error and the chi-squared value for the spectrum at the full text size. Note that adjusting Zipf's law tends to lead to a slight decrease in the value of X^2 that, however, is paired with an increase in the MSE. For the generalized inverse Gauss-Poisson model, the fits also tend to be slightly better in terms of the chi-square and slightly worse in terms of the MSE. As the fit to the spectrum of the full text itself is more or less equivalent for the unadjusted and adjusted models, the advantage of the adjusted model lies primarily in its increased interpolation and extrapolation accuracy, without this advantage being seriously offset by a deteriorating spectrum fit.

Various link functions appear as reasonable models for the developmental profiles of the parameters c and Z . For Z , a power model

$$Z(N) = a_1 N^{a_2}$$

and a logarithmic model

$$Z(N) = a_1 + a_2 \log(N)$$

tend to yield roughly equivalent fits. However, for *The war of the worlds*, a

Table 5.4: Parameters and goodness-of-fit statistics for the extended Zipf's law and the generalized inverse Gauss-Poisson model and their parameter-adjusted variants for selected texts (Alice: Alice in Wonderland; Through: Through the looking-glass; Wells: The War of the Worlds; Conan Doyle: Hound of the Baskervilles).

	Alice	Through	Wells	Conan Doyle
extended Zipf's law				
\hat{Z}	12222.2	12919.9	60741.1	31711.5
X^2	29.05	22.22	19.74	227.84
df	15	15	15	15
p	0.0158	0.1021	0.1822	0.0000
MSE	126.22	75.27	363.47	3925.0
adjusted extended Zipf's law				
\hat{Z}	12443.5	12443.8	61782.0	30958.1
X^2	21.69	22.26	13.34	352.28
df	14	14	14	14
p	0.0852	0.0735	0.5001	0.0000
MSE	165.51	415.81	534.0	5338.1
link function	power	power	linear	logarithmic
\hat{a}_1	617.68	1628.07	35937.3	-12914.44
\hat{a}_2	0.2898	0.1979	0.4312	3992.42
generalized inverse Gauss-Poisson				
\hat{b}	0.0236	0.0246	0.0068	0.0083
\hat{Z}	105.78	115.33	265.00	176.11
X^2	262.98	302.65	1763.65	1322.82
df	14	14	14	14
p	0.0000	0.0000	0.0000	0.0000
MSE	421.43	447.87	5672.58	3341.28
adjusted generalized inverse Gauss-Poisson				
\hat{b}	0.0234	0.0232	0.0067	0.0068
\hat{c}	0.0095	0.0089	0.0038	0.0060
X^2	245.38	236.36	-	963.12
df	13	13	13	13
p	0.0000	0.0000	-	0.0000
MSE	421.49	511.53	5692.89	3844.67
link function	power	power	exponential	logarithmic
\hat{a}_1	0.0567	0.0548	0.0056	0.0103
\hat{a}_2	-0.1754	-0.1765	0.0000	-0.0004

linear fit

$$Z(N) = a_1 + a_2 N$$

is clearly optimal. For c , only the logarithmic and power models were found to yield acceptable fits. The exponential fit listed for *The war of the worlds*,

$$c(N) = e^{a_1 + a_2 N}$$

was optimal in the least-squares sense, but upon visual inspection completely failed to capture the developmental trend in the data.

The extrapolation accuracy of adjusted LNRE models depends on the accuracy of the fit for $Z(N)$ or $c(N)$. Simple models such as the power model and the logarithmic model appear to capture the main trends with enough precision to enhance interpolation accuracy. However, fits with these models tend to remain approximations that fail to capture the exact developmental profile. Clearly, more research is required with respect to the appropriate link functions for further enhancement of extrapolation accuracy. In addition, the development of parameter-adjusted variants of the more powerful three-parameter models may be expected to lead to improved accuracy.

Finally, recall that non-randomness in word use causes $\mathcal{P} = E[V(1, N)]/N$ as a measure of the growth rate of the vocabulary to be too small. The underestimation bias of \mathcal{P} is illustrated in Figure 5.18 for *Alice in Wonderland*.

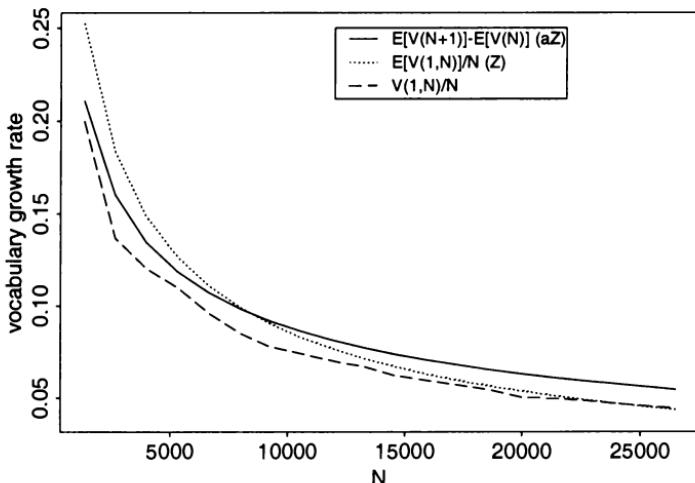


Figure 5.18: Three estimates of the growth rate of the vocabulary: the one-token growth rate $E[V(N + 1)] - E[V(N)]$ for the parameter-adjusted Zipf smoothing (solid line), the unsmoothed sample estimates $V(1, N)/N$ (dashed line), as well as the (unadjusted) Zipf-smoothed estimates $E[V(1, N)]/N$ (dotted line), for Alice in Wonderland.

The dashed line plots the growth rate of the vocabulary using the unsmoothed sample estimates $V(1, N)/N$. The dotted line plots the smoothed estimates $E[V(1, N)]/N$ using the (unadjusted) extended Zipf's law. The two curves converge to approximately the same final value. The overestimation bias for small N will clearly be reversed when extrapolating to values of N greater than the empirical text size. In order to obtain a better approximation of the empirical growth rate of the vocabulary we can use the by-token vocabulary growth rate

$$GV(N) = E[V(N + 1)] - E[V(N)], \quad (5.18)$$

the expected average increase in the vocabulary size obtained by adding one token to the sample, represented by a solid line in Figure 5.18. Note that the adjusted growth rate (solid line) and the unsmoothed growth rate (dashed line) follow the same trajectory, except for a translation along the vertical axis. We can correct for this by defining

$$\varepsilon_{GV} = GV(N_0) - V(1, N_0)/N_0,$$

with N_0 the complete text size, and by using $GV'(N)$,

$$GV' = GV(N) - \varepsilon_{GV}$$

instead of $GV(N)$.

5.3 Discussion

In the preceding chapters we have seen that by combining the urn model with specific assumptions concerning the structural distributions, it is possible to formulate theoretical models for the frequency spectrum and related statistics. The urn model, however, is an obvious simplification that does not do justice to the many intricate patterns of lexical dependency in texts. An analysis of lexical specialization reveals that the accuracy of the theoretical models is primarily affected by the locally concentrated, underdispersed use of key words. Unlike in experimental statistics, where it is often possible to neutralize known dependencies by counterbalancing, texts cannot be brought in line with the randomness assumption without destroying the 'texture' of the text itself. Hence, we have studied in some detail in what way the values of various statistics are affected by textual non-randomness. While it is important to keep the limitations the urn model in mind, it is only fair to point out in its defence that, in the light of the enormous linguistic complexity of texts, it is rather surprising that LNRE models capture the main trends in the data as well as they do. In order to adjust LNRE models for the effects of non-randomness, two adjustment techniques are available, partition-based adjustment and parameter-based adjustment. Both techniques lead to considerable improvement both with respect to interpolation accuracy as with respect to extrapolation accuracy. Further research is required, however, into the appropriate link functions required for these adjustment techniques.

5.4 Bibliographical Comments

The randomness assumption and the consequences of non-randomness in word use have received relatively little attention in the literature. An early study is In 't Veld (1984). For partition-based adjustment, see Hubert and Labbe (1988, 1988a) and Labbe and Hubert (1994). Section 5.1 is based on Baayen (1996a). For section 5.2, see Baayen & Tweedie (1998) and Baayen & Tweedie (1999).

Chapter 6

Examples of Applications

This chapter discusses examples of domains of inquiry where LNRE models or the theory of LNRE distributions can be applied. The first examples are drawn from the study of distributional properties of the lexicon. The second set of examples concerns the study of morphological productivity. A brief discussion of the pros and cons of using measures based on word frequency distributions in authorship studies is followed by a set of examples of applications outside the domain of word frequency distributions. The final section offers some practical guidelines, and provides a detailed example of how the program library for running LNRE analyses described in Appendix C can be used.

6.1 Distributional properties of the lexicon

6.1.1 Word length and sample size

Word frequency is a lexical variable that is known to correlate with a great many other variables such as word length (Best and Zhu, 1994; Wimmer, Koehler, Grotjahn, and Altmann, 1994), number of meanings, (Paivio, Yuille, and Madigan, 1968; Reder, Anderson, and Bjork, 1974; Koehler, 1986), number and summed frequency of morphologically related words (Schreuder and Baayen, 1997), and the number and average frequency of phonologically or orthographically similar words (Landauer and Streeter, 1973; Baayen, 1991; Frauenfelder, Schreuder, Hellwig, and Baayen, 1993). A question that has not received principled discussion in the literature is to what extent the LNRE property of word frequency distributions is reflected in the variables correlating with word frequency. In this section, I will illustrate this problem for word frequency and word length.

Researchers studying the distribution of word length are not unaware of the potential problem that the dependency of the frequency spectrum on the sample size might cause for the modeling of word length distributions. However, Best and Zhu (1994), following Hammerl (1990), claim that texts with more than 2000 words tend to become inhomogeneous with respect to topic

and style. By studying short texts of roughly 1000 words, it is hoped to avoid the problems of text-internal inhomogeneity. Wimmer, Koehler, Grotjahn, and Altmann (1994) sketch the possibility of modeling the effect of factors such as register, topic domain, and authorship on the shape of word length distributions once the characteristics of homogeneous texts have been firmly established.

Although this approach to the modelling of word length distributions is a sensible and practical solution to the problem of textual non-homogeneity, it does not offer a principled answer to the question in what way the LNRE property of word frequency distributions might affect one's conclusions concerning the appropriate model for the associated word length distribution.

Figure 6.1 illustrates the problem for the eight one-million word samples of *The Independent* that we have studied before in chapters 1 and 2. The upper right panel plots the number of types as a function of word length for increasing corpus size from 1 to 8 million word tokens. Not surprisingly, the number of types with say length 5 increases with N . As we increase N , we allow more and more population types with length 5 into the sample. The upper right panel shows that the average frequency of the words with a given length also increases with N . The bottom left panel graphs the proportion of types for the various word lengths. The solid line represents the distribution for 1 million words, the dotted line the distribution for 8 million words, and the dashed line the distribution for *Alice's Adventures in Wonderland*, which has as sample size of approximately 25,000 words. Note that the curve shifts to the higher word lengths as we increase the sample size. This property of word length distributions is highlighted for word lengths 5 and 15 in the bottom right panel of Figure 6.1. As we increase the corpus size N , shown on the horizontal axis, the proportion of types with word length 5 slowly decreases, while the proportion of types with word length 15 slowly increases.

It does not make sense to attribute this systematic change in the word length distribution to factors variations in authorship, register, or style. Internally, the samples of the independent are non-homogeneous with respect to authorship, topic domain, and style. But for our 8 samples of 1 million words from the same newspaper sampled at different periods through a period of more than 6 years, the sample-internal effects of non-homogeneity are probably comparable for each individual sample, so that we may consider the 8 samples as comparable and, at a higher level of granularity, homogeneous. This example shows that models for word length distributions have to be parameterized for sample size, just as models for word frequency distributions.

The fact that distributions that are interrelated with word frequency distributions inherit the LNRE problem has serious consequences when distributional properties of words in the lexicon are compared using corpora of different sample size. Consider Figure 6.2, which plots the number of words with frequency lower than 10 (solid line), lower than 20 (dotted line), and lower than 100 (dashed line) as a function of corpus size using the data of *The Independent* for words of length 4 (left panel) and words of length 10 (right panel). In a corpus of 1 million word tokens, the proportion of words of length 5 with frequency ≤ 10 is much smaller than the proportion of words of length 5 in a

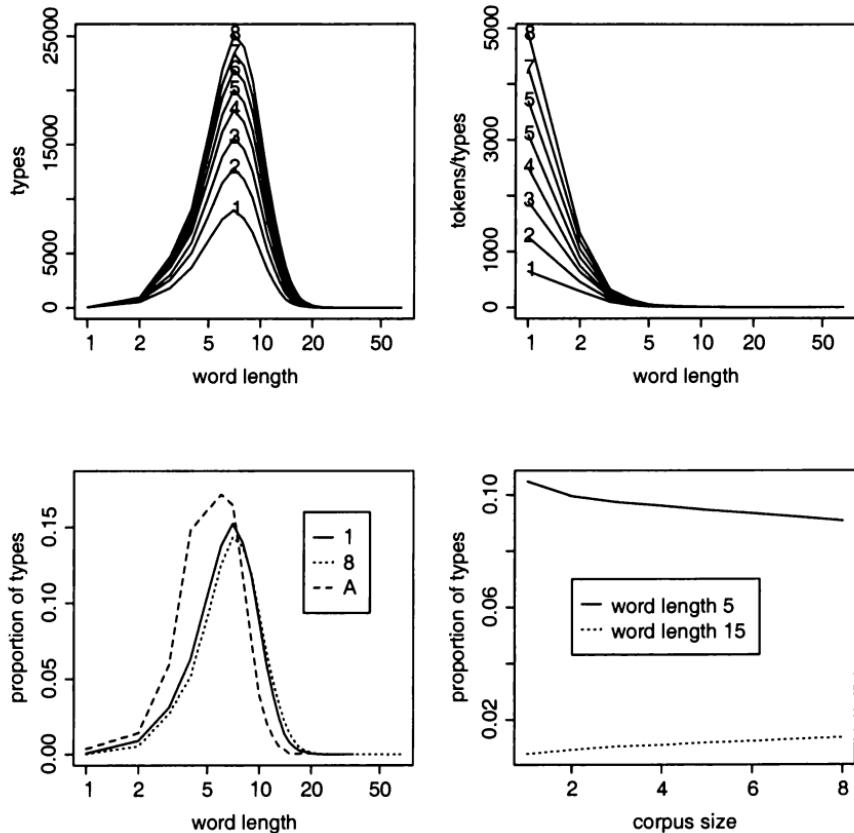


Figure 6.1: The distribution of word length and sample size ($N = 1 \dots 8$ million word tokens). Upper left: number of types as a function of word length in letters; upper right: mean word frequency as a function of word length; lower left: proportion of types as a function of word length, the curve labelled A represents the corresponding graph for Alice's Adventures in Wonderland ($N = 26505$); lower right: proportion of types as a function of corpus size for word lengths 5 and 15.

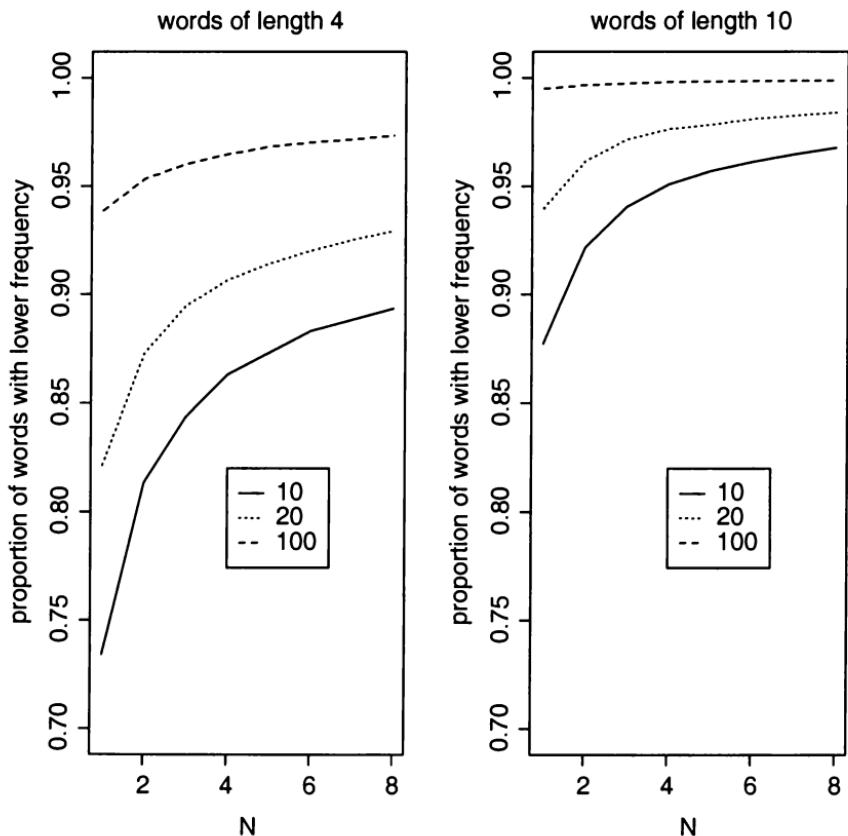


Figure 6.2: Proportion of words (length 4, left panel; length 10, right panel) with frequency less than 10 (solid line), less than 20 (dotted line), and less than 100 (dashed line) as a function of the sample size N (in million word tokens from The Independent).

corpus with 8 million words. At first blush, this suggests that the two corpora sample radically different populations, while in fact, as is the case in this example, they stem from exactly the same population. Conversely, the words of length 5 at sample size 8 reveal a proportion of types with frequency ≤ 10 that is of the same order of magnitude as the corresponding proportion for words of length 10 in a 1 million sample. When sample size is ignored, this leads to the incorrect conclusion that the two word lengths appear with roughly equal proportions of words in the frequency range 1...10.

When distributional properties of words have to be compared on the basis of frequency lists compiled from corpora of different size, say N_1 and N_2 , $N_1 < N_2$, it is advisable to randomly sample N_1 tokens from the larger sample proportional the relative frequencies of the words in the larger corpus. Although the new sample of N_1 tokens will not be completely compatible with an independent sample of N_1 words — non-randomness in word use is not taken into account, and the $V(0, N_2)$ words that do not appear in the larger sample have not had the chance to appear in the new sample — the distributional properties of the two corpora can now be compared with a substantially reduced risk of distortion due to the difference in sample size. An example of a distributional study in which this resampling technique has been used is Frauenfelder, Schreuder, Hellwig, and Baayen (1993).

6.1.2 Matching reliability across corpora

An issue that occasionally arises in psycholinguistic studies of lexical processing is whether experimental designs are feasible in which words are matched pairwise for Surface Frequency but contrasted with respect to Base Frequency. The Surface Frequency of a word such as *gallop* is the frequency of occurrence of exactly the orthographic string *gallop*. The Base Frequency of *gallop* is the summed frequency of all its inflectional variants (*gallop*, *gallops*, *galloping*, *galloped*).

The top panels of Figure 6.3 plot the matching for the words in *-heid* used in Experiment 3 of Bertram, Schreuder, and Baayen (1999). This Experiment investigates whether words with a high log Base Frequency (the H set of words) are processed faster than words with a low log Base Frequency (the L set of words) when log Surface Frequency is held constant between the two conditions. The upper left panel presents boxplots for the matching for Surface Frequency, and the right panel the corresponding boxplots for the contrast in Base Frequency, using the word frequencies as available in the CELEX lexical database. These frequency counts are based on a corpus of 42 million words.

The question now is whether the words with a high Base Frequency and a low Surface Frequency are in fact outliers due to sampling error. If a word has a high Base Frequency, then one would expect it also to have a high Surface Frequency. If instead it occurs with a low Base Frequency, it might be the case that this low Surface Frequency does not reflect the true population frequency. In other words, the experimental design would capitalize on random sampling error, which would imply that the matching would be invalid. It would then not be possible to trace a difference between the H set and the

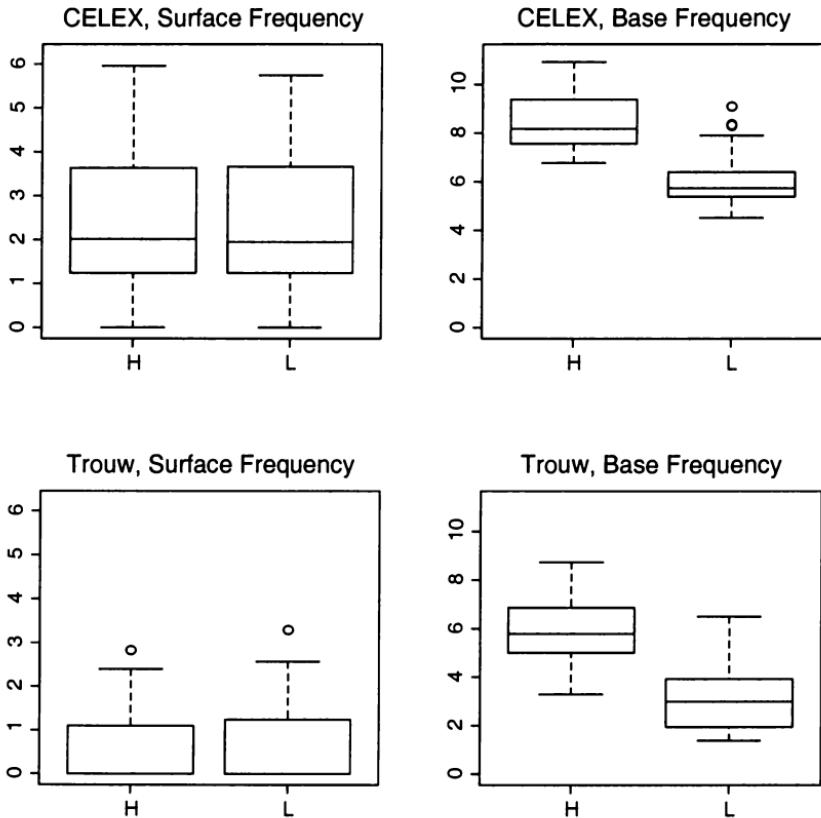


Figure 6.3: Boxplots showing the matching for log Surface Frequency and the contrast in log Base Frequency for the words in -heid in Experiment 3 of Bertram, Schreuder, and Baayen (1999) for the CELEX lexical database (based on a 42 million word corpus, upper row) and the Dutch Trouw corpus (4.2 million words, bottom row).

L set in the average experimental response latencies unambiguously to the difference in Base Frequency. The effect might just as well arise due to a hidden difference in Surface Frequency. This problem is discussed both in Taft (1979) and in Bertram, Schreuder, and Baayen (1999), probably in response to the same anonymous reviewer, who claims the design is flawed by the problem of regression towards the mean: the words with a low Surface Frequency would be extreme outliers and their real frequencies would be closer to the average Surface Frequency of words with a high Base Frequency.

In order to make sure that the issue of sampling error does not arise, one can check whether the matching still holds using the frequencies of occurrence of the same words in a second corpus. Taft (1979) found that his matching was retained in such a second corpus, and the bottom panels of Figure 6.3 show, using a 4.2 million word corpus of the Dutch newspaper *Trouw*, that the matching is unaffected by changing from one corpus to another for the materials in *-heid* of Bertram et al. (1999) as well. The question to be addressed here is whether this is a coincidence or not.

The first to realize is that the term 'regression towards the mean' is somewhat misleading in this context. The classical example of regression towards the mean is that the sons of fathers of more than average size tended to be smaller than their fathers, while sons of fathers with less than average size tended to be larger than their fathers. Given the general positive correlation between Surface Frequency and Base Frequency, it is indeed true that words with a high Base Frequency and a low Surface frequency are exceptional, just as large fathers with even larger sons are exceptional. But this is not the issue, as the experimental design selects very specific words without any claims as to how well they represent the population of words.

This leaves us with the problem of sampling error. How likely is it that a word with a high Base Frequency and a low Surface Frequency is in fact a word with a high Base Frequency and a medium Surface Frequency? How unlikely is such a combination? Consider the case of nouns. Some nouns denote objects that typically occur in pairs or groups (*eyes, legs, sheep*), others denote objects that tend to occur by themselves (*nose, moon, table*). The singulars of the former type of nouns will have a high Base Frequency (the summed frequencies of *eye* and *eyes*) and a low Surface Frequency (the frequency of *eye*) without being in any sense strange or exceptional. The singular and plural frequencies reflects the frequencies with which objects appear singly or in pairs or groups in the real world. Thus, there is no a-priori linguistic reason to assume that there is something wrong with the sampled frequencies.

There is also no a-priori statistical reason to assume that the combination of a high Base Frequency in combination with a low Surface Frequency is suspect. Given the Good-Turing estimate m^* of a word ω occurring m times in a corpus with N_1 words,

$$m^* = \frac{(m+1)\text{E}[V(m+1, N_1)]}{\text{E}[V(m, N_1)]},$$

we can estimate the frequency $f(\omega, N_2)$ of this word in a corpus of size N_2 to

be

$$f^*(\omega, N_2) = \frac{N_2}{N_1} \frac{(m+1)\mathbb{E}[V(m+1, N_1)]}{\mathbb{E}[V(m, N_1)]}.$$

When $N_2 < N_1$, the Good-Turing frequency $f^*(\omega, N_2)$ will be slightly smaller than $\frac{N_2}{N_1}m$. For large m , the difference between m and $f^*(\omega, N_2)$ will be somewhat larger than for small m .

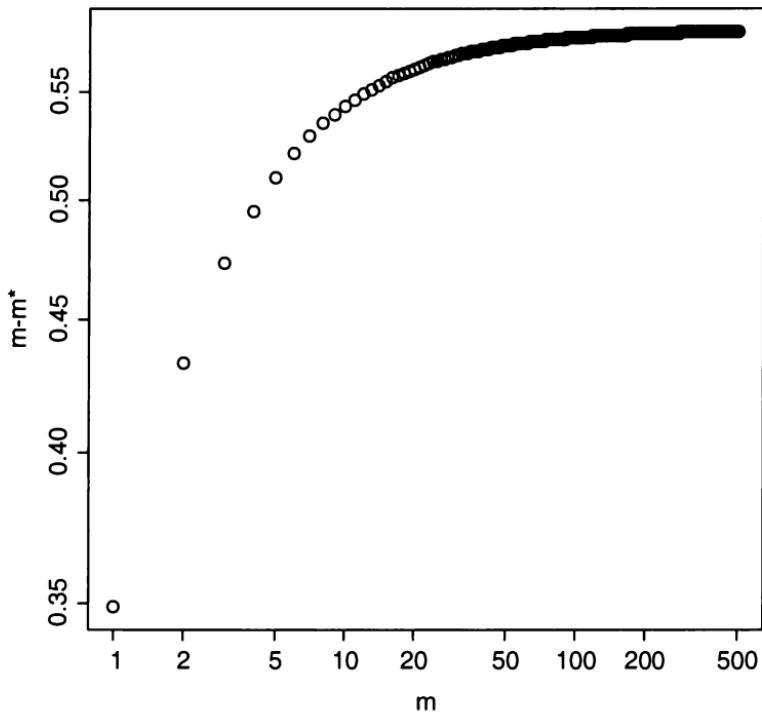


Figure 6.4: The difference between m and m^* for $m = 1, \dots, 50$ for the 8 million word sample of The Independent using a Naranan-Balasubrahmanyam Zipfian smoother for the calculation of the Good-Turing estimates.

Figure 6.4 illustrates that the difference between m and the expected Good-Turing frequency m^* increases with m but nevertheless remains quite small. Therefore, words with a high Base Frequency will still have a high Base Frequency in another corpus, and words with a low Surface Frequency are unlikely to show up with a substantially higher Surface Frequency in another corpus, irrespective of whether their Base Frequency is high or low. Even the fact that the variance of m^* increases with m , illustrated in the second panel

of Figure 2.6 in Chapter 2, does not cause a problem because, although the variance will be greater for the high Base Frequency set and the low Base Frequency set than for the high and low Surface Frequency sets, the average Base Frequency contrast between the high and low sets will remain in tact. This is exactly what we observe for the Dutch experimental data graphically summarized in Figure 6.3. The variance in the high and low log Base Frequency sets is 1.27 and 1.63 respectively for the CELEX counts, while in the much smaller Dutch newspaper corpus, the variances are 2.02 and 2.11 respectively.

We conclude that, in the absence of well-motivated doubts concerning the balance of Surface and Base frequency (as when, for instance, the word *harpsichord* is found to occur 300 times in the plural and only once in the singular), there is no reason to assume that the present experimental design is flawed by inevitable ‘regression towards the mean’.

6.2 Morphological productivity

Texts may differ with respect to a wide range of factors, such as authorship, register or genre, style, topic domain, and intended audience. Any of these factors may substantially influence the degree of productivity of a word formation pattern. In this section, we first present a global analysis in which the productivity of selected Dutch affixes is measured on the basis of a large corpus, without paying attention to the various kinds of texts contained in this corpus. To the extent that a corpus reflects in some way the use of a given language in the language community, this approach provides insight into the global productivity of word formation patterns. In the second subsection, we then proceed to show how severely one factor, register, may affect the degree of productivity of an affix, using data from English.

6.2.1 Global analyses

In section 4.4, a number of measures for quantifying the pre-theoretical notion of morphological productivity were introduced. In this section, we first show how the simplest productivity measures $V(N)$, $V(1, N)$, and $\mathcal{P}(N)$ can be used to gauge the degree to which a morphological category is productive. We will then discuss to what extent the results might be distorted by the fact that words are not used randomly in texts. Finally, we will briefly discuss the evaluation of morphological productivity in terms of the population number of types S .

Figure 6.5 visualizes the three key productivity statistics, $V(N)$, $V(1, N)$, and $\mathcal{P}(N)$ for the following morphological categories:

N	monomorphemic nouns
A	monomorphemic adjectives
V	monomorphemic verbs
-in	female agent or occupation nouns (<i>boerin</i> , ‘female farmer’)
-ster	female agent or occupation nouns

	(<i>denkster</i> , 'female thinker')
-se	female agent or occupation nouns (<i>dominese</i> , 'female preacher')
-e	female agent or occupation nouns (<i>gidse</i> , 'female guide')
-es	female agent or occupation nouns (<i>prinses</i> , 'princess')
-erd	pejorative nouns (<i>viezerd</i> , 'dirty person')
-her	iterative verbs (<i>heronderzoeken</i> , 're-investigate')
-elijk	descriptive adjectives (<i>werkelijk</i> , 'work-like', i.e., 'really')
-iteit	abstract nouns (<i>complexiteit</i> , 'complexity')
-heid	abstract nouns (<i>groenheid</i> , 'greenness')
-baar	adjectives in -able (<i>werkbaar</i> , 'workable')
-ver	forms verbs (<i>verhuizen</i> , 'to move house')

The three sets of monomorphemic words are all unproductive. Their values for $V(N)$ are large, but their category-internal growth rates (\mathcal{P}) are the lowest of all categories described here. Their position in the upper left corner of the bottom right scatterplot of Figure 6.5 shows that it is unlikely that additional tokens of simplex nouns, verbs, or adjectives will represent new types. In the case of the simplex verbs, even the shape of the word frequency distribution, which has its mode at $m = 2$ instead of at $m = 1$, reveals that at this sample size this category has moved out of the LNRE zone altogether.

At the opposite corner of this plot we find the unproductive suffix *-se*. Even though *-se* displays a high \mathcal{P} -value, its associated extremely low value of $V(N)$ shows that we are dealing with a suffix that is represented by just a few hapax legomena only. The contribution of this suffix to the growth rate of the vocabulary as a whole, evaluated in terms of $V(1, N)$ (i.e., in terms of \mathcal{P}^* , but without normalizing by the joint number of hapax legomena of all morphological categories), is likewise minimal, see, e.g., the upper right panel.

Truly productive suffixes occupy the upper right corner of all three panels of Figure 6.5, a position indicating they are strong in both $V(N)$, $V(1, N)$, and \mathcal{P} . The suffix *-heid* is prototypically productive in this sense. The suffixes *-baar*, *-ster*, and *-e* are strong in $\mathcal{P}(N)$ and $V(1, N)$, but they are weaker in $V(N)$, which suggests that these categories are productive in principle, in the sense that many new formations may be expected for larger samples, but that at the same time the category as a whole is not as useful as that of *-heid*, which is represented by roughly 5 times as many types in the corpus.

Note that inspection of $V(1, N)$ and $\mathcal{P}(N)$ allows us to analyze with more precision than a simple type count the degree of productivity of morphologi-

cal categories that appear with approximately the same number of types. Consider, e.g., *-eljk* (generally judged to be unproductive), *-iteit* (productive), *-ster* (productive), *-e* (productive), *-baar* (productive), and the simplex adjectives (unproductive). It is only through inspection of \mathcal{P} that we can differentiate between the unproductive categories, which have lower \mathcal{P} -values and lower numbers of hapax legomena, and the more productive categories, which have higher \mathcal{P} -values and higher numbers of hapax legomena. In order to come to grips with the quantitative aspects of morphological productivity, it is necessary to inspect both the number of types realized in the sample and the likelihood that hitherto unobserved types will appear when the sample is increased.

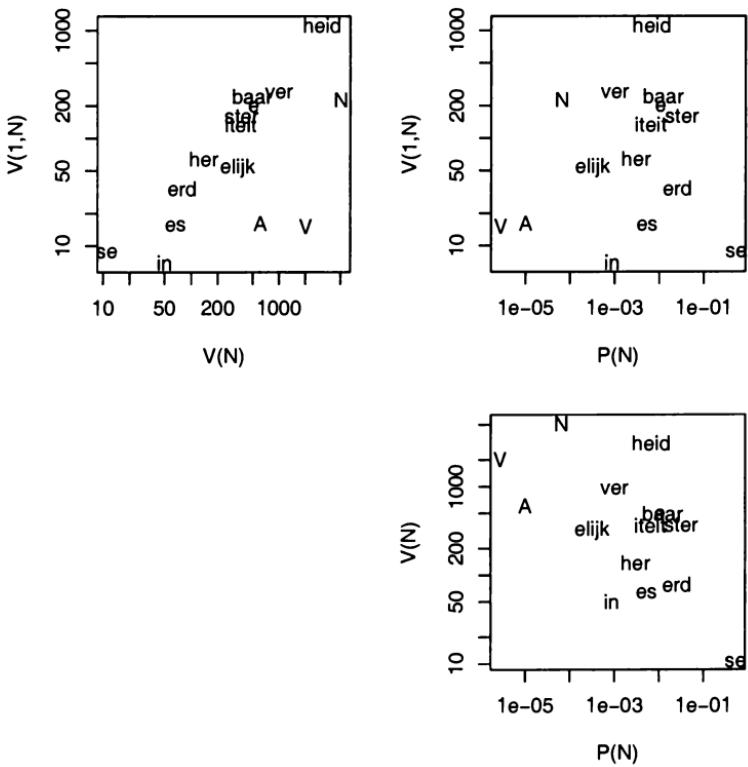


Figure 6.5: Productivity statistics $\mathcal{P}(N)$, $V(1, N)$, and $V(N)$ for selected Dutch affixes.

Given the importance of measures relating to the growth rate of the vocabulary size, we should ask ourselves to what extent these estimates might be affected by the non-randomness in word use that we have studied in Chap-

ter 5. Recall that the growth curve of the vocabulary tends to have an overestimation bias, as shown for the 8 million word corpus of *The Independent* in the upper left panel of Figure 6.6. The values expected on the basis of binomial interpolation, represented by a solid line, are higher than the observed values, represented by circles. This implies that the theoretical growth rate of the vocabulary will be too small. The question now is whether the effect of non-randomness remains when we sample words with a particular structure, e.g., all words ending in *-ness*, instead of sampling all words. The sampling of subsets of words implies that the effects of topic continuity in discourse will be reduced substantially. The bottom left panel shows that, as expected, the overestimation bias is substantially reduced for the 9449 word tokens in *-ness* in *The Independent*, compared to an equal number of word tokens in *Alice's Adventures in Wonderland* (bottom right panel). A slight overestimation bias remains, however, as most expected values are slightly higher than the observed values. This is probably due to the slight tendency for low-frequency complex forms with a particular morphological structure to occur in other's company. For instance, words in *-ness* tend to be used as near-synonyms in paratactic constructions such as *dejectedness* and *dumbfoundedness* (see Baayen and Neijt, 1997, for detailed data on the Dutch equivalent of *-ness*, *-heid*). This kind of underdispersion contributes to the overestimation bias just as the underdispersion caused by topic continuity at the discourse level. For subsets of morphologically complex words, however, the bias is quite small, which allows us to have some confidence in the accuracy of productivity measures based on the growth rate of the vocabulary.

For a validation study of \mathcal{P} and \mathcal{P}^* as measures of productivity, see Baayen (1994) and Baayen and Renouf (1996).

This analysis of productivity can be complemented by estimating the population number of types S . For instance, for the unproductive suffix *-in*, a lognormal fit ($\hat{Z} = 3.86$, $\hat{\sigma} = 2.35$) suggests that $\hat{S} = 61.1$ is only slightly larger than the observed number of types, 50. This is in line with its very low degree of productivity, $\mathcal{P} = 7/7985 = 0.00088$. For the productive suffix *-ster*, the observed number of types is fairly small (370) compared to *-heid* (3070), but the population number of types estimated on the basis of the Yule-Simon model ($\hat{Z} = 63.23$, $\hat{\beta} = 0.60$, $\hat{V}(Z) = 37.94$) is infinite given that $\hat{\beta} < 1$. Estimating S may require the use of a mixture model, however, as illustrated for *-heid* and *-iteit* in section 4.3.2. In the case of *-ster*, for instance, a mixture model may be more appropriate given the mediocre goodness of fit of the simple Yule-Simon model ($X_{(13)}^2 = 52.20$, $p = 0.000001$; MSE = 20.02).

When applying these methods, it is important to make sure that the data are reliable. Figure 6.7 illustrates this point for three German suffixes, *-bar* (solid line), *-sam* (dashed line), and *ös* (dotted line) using data made available to me by Anke Lüdeling and Stefan Evert. The left panel shows the vocabulary growth curves for these three suffixes using counts based on simple string matches. The curves suggest substantial differences in productivity. Given that all three curves are increasing, it would seem that neither *ös* nor *-sam* are truly unproductive. However, after manual correction of the raw data, it becomes clear that *ös* and *-sam* are probably completely unproductive: Their

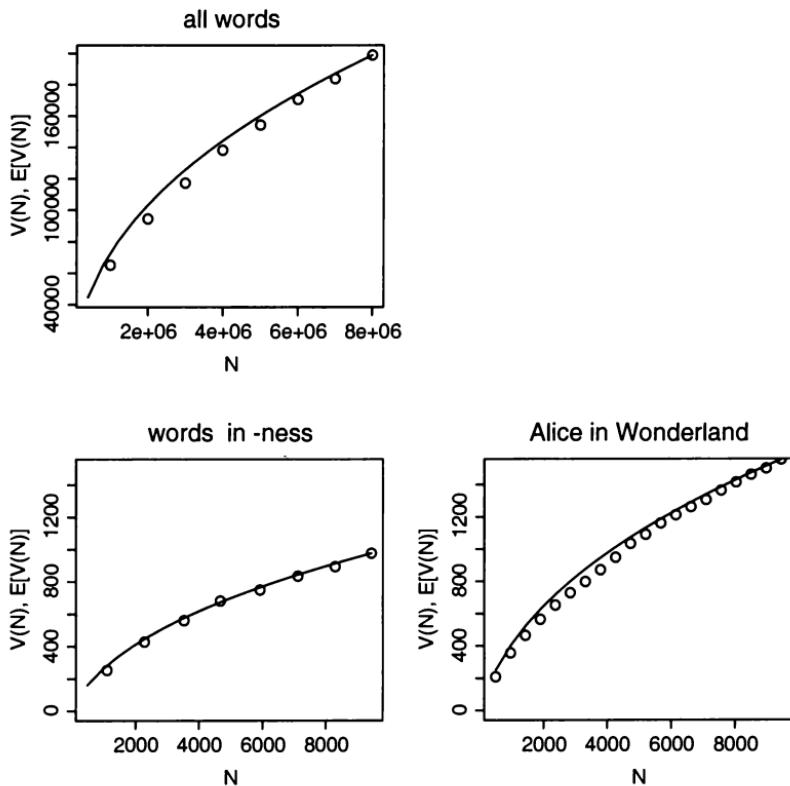


Figure 6.6: Observed and expected vocabulary size as a function of sample size using binomial interpolation for all words in the 8 million word sample of *The Independent* (upper left panel), for the 9449 words in *-ness* in the same corpus (lower left panel), and for the first 9449 words of Alice's Adventures in Wonderland (lower right panel).

vocabulary growth curves now emerge as nearly having reached their asymptotic value.

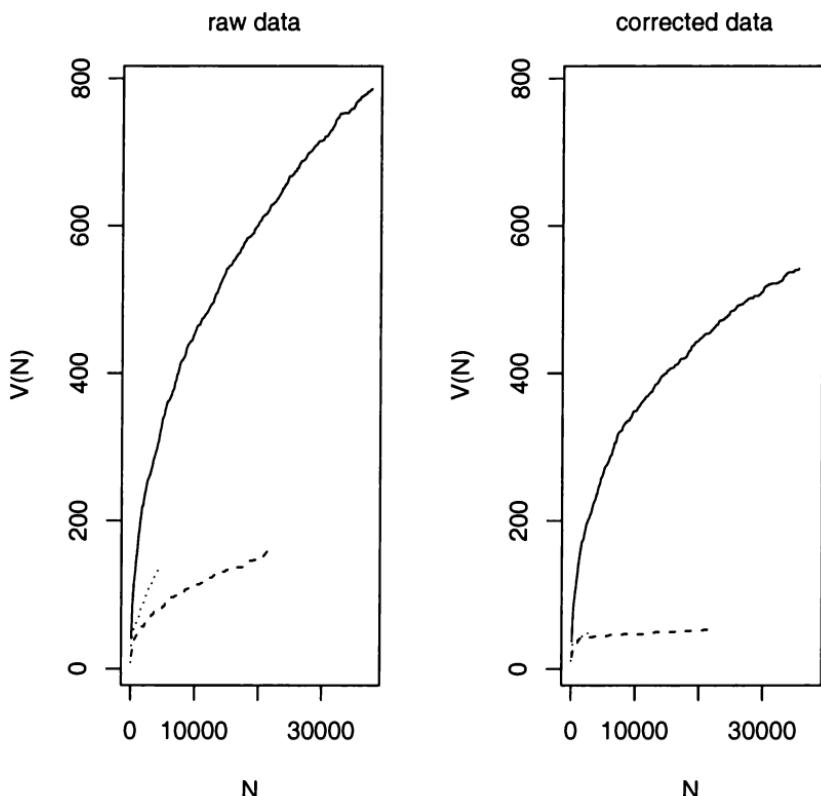


Figure 6.7: Vocabulary growth curves for the German suffixes -bar (solid line), -sam (dashed line), and ös (dotted line) using raw data based on string matches (left panel) and manually corrected data (right panel).

6.2.2 Productivity and register

The productivity of morphological categories may vary substantially on dimensions such as authorship, style, register, topic domain, and intended readership. An analysis of the way in which morphological categories cluster together on these dimensions is outside the scope of this book. For an exploratory study, the reader is referred to Baayen (1995). The aim of the present section is to illustrate the extent to which the degree of productivity is co-determined by these factors by comparing the use of the English suffix *-ness* in the three main subcorpora of the British National Corpus: the subcorpus

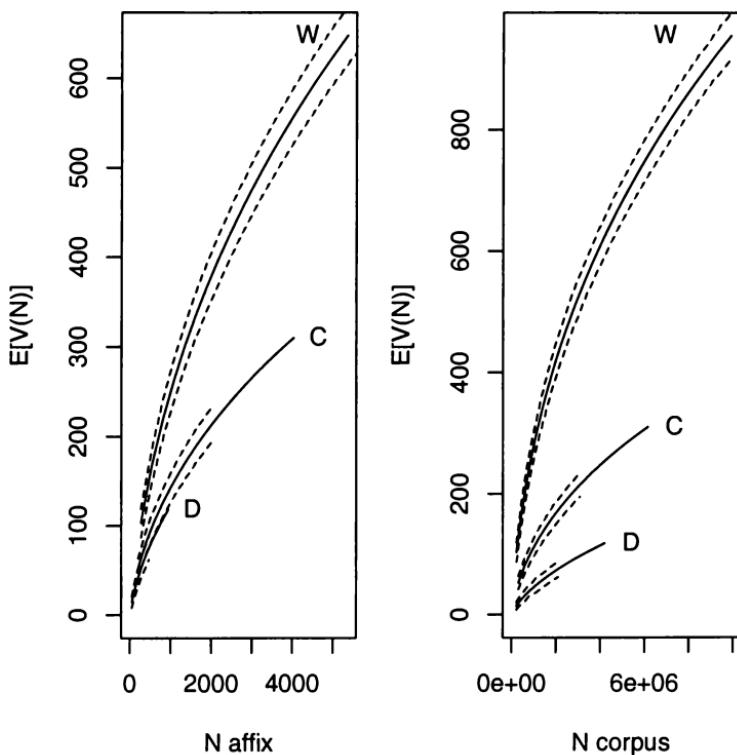


Figure 6.8: Vocabulary growth curves $E[V(N)]$ using binomial interpolation for the English suffix -ness in the written subcorpus (W), the context-governed spoken subcorpus (C), and the demographic subcorpus (D) of the British National Corpus. The left panel plots $E[V(N)]$ as a function of the number of tokens in -ness in each subcorpus, the right panel renormalizes the horizontal axis to display the number of arbitrary tokens in a subcorpus.

of written English (henceforth W), the subcorpus of spontaneous English conversations (the so-called demographic subcorpus, henceforth D), and the subcorpus of spoken English in more controlled contexts (the so-called context-governed subcorpus, henceforth C, which contains interviews, speeches, meetings, etc.). We have already cursorily investigated part of this data set in the context of mixture distributions (see section 4.3.2).

The left panel of Figure 6.8 presents the growth curves of the vocabulary size for the words in *-ness* in each of these three subcorpora. Thus, we find roughly 4000 word tokens in *-ness* in subcorpus C, which jointly represent roughly 300 types. Subcorpus D contains only some 1000 word tokens with *-ness*, in all slightly more than 100 types. Conversely, subcorpus W contains the largest number of tokens and types in *-ness*. Only the first 5000 tokens in *-ness* from this subcorpus have been graphed.

The differences in the numbers of tokens and types observed are by themselves not very informative because the subcorpora in which they have been counted differ in size. The subcorpus of written language W contains in all 89740544 tokens, the subcorpus of spontaneous conversations D contains in all 4211216 tokens, and the context-governed subcorpus C 6154248 tokens. It is more interesting to compare the number of types observed for a fixed number of tokens from each subcorpus using binomial interpolation. The resulting curves are represented by solid lines in the left panel of Figure 6.8, which also shows the 95% confidence intervals using dashed lines for the first $N/2$ tokens, using

$$\text{VAR}[V(N)] = V(2N) - V(N)$$

to estimate the variance of $V(N)$ given by (3.62). The three subcorpora do not differentiate for very small N . For the complete sample size of subcorpus D, however, all three curves are already significantly different.

Instead of inspecting the number of types observed for a given number of word tokens in *-ness* for the three subcorpora, we can also ask ourselves how many types in *-ness* are to be expected for a given number of arbitrary tokens from the subcorpora, including the many word tokens representing types that do not end in *-ness*. Renormalization of the X-axis, under the assumption that the N_{affix} tokens in *-ness* are uniformly distributed over the N_{corpus} tokens of a given subcorpus, leads to the right panel of Figure 6.8. The dissociation of the three registers now emerges even more clearly. The position of subcorpus C in between W and D is in line with the relative complexity of the kind of English in the three subcorpora. The subcorpus of spontaneous conversations (D) contains unprepared spoken materials, the context-governed subcorpus contains spoken materials such as lectures and interviews, materials that have been prepared to some degree, while a high degree of preparation is characteristic for written texts. Clearly, the likelihood of using abstract nouns in *-ness* increases for the more complex registers of English.

For a more detailed study of variations in productivity as a function of register in English, the reader is referred to Plag, Dalton-Puffer, and Baayen (1999). Kageura (1999) is a study of the productivity of term formation in different kinds of texts across the lexical strata of native and loan words.

6.3 Authorship and Style

Various researchers have used one or a small set of 'characteristic constants' and sometimes even the type-token ratio to study issues of authorship and literary style (e.g., Yule, 1944; Guiraud, 1954; Muller, 1977; Brunet, 1978; Sichel, 1986; Holmes, 1992, 1994; Holmes and Forsyth, 1995; Martindale and McKenzie, 1995; Whissell, 1996; Gani, 1997). In Chapter 1, we have already seen that most proposed constants are not truly constant in practice, and that in theory only Yule's K , Simpson's D , and the parameters of LNRE models (conditional on having obtained a reliable fit) are truly independent of the sample size. A question was left unanswered was to what extent the variation in the values of a constant poses a serious problem for the study of authorship and style. If the variation within a text is small compared to the variation between texts and especially between texts in different styles by different authors, then there is no serious problem. However, when the variation is such that the intra-textual variation is as big as the inter-textual variation, the usefulness of summary textual measures is considerably reduced.

Table 6.1: *Legend for the texts analyzed in Figures 6.9 and 6.10.*

Author	Title	N	Legend
L.F.Baum	The Wonderful Wizard of Oz	39282	b1
	Tip Manufactures a Pumpkinhead	41571	b2
E.Brontë	Wuthering Heights	116534	b2
L.Carroll	Alice's Adventures in Wonderland	26505	a1
	Through the Looking-glass and what Alice Found there	29053	a2
A.Conan Doyle	The Sign of Four	43125	c1
	The Hound of the Baskervilles	43125	c1
	The Valley of Fear	57746	c3
H.James	Confidence	76512	j1
	The Europeans	59800	j2
St Luke	Gospel according to St Luke (KJV)	25939	L1
	The Acts of the Apostles (KJV)	24246	L2
J.London	The Sea Wolf	105925	l1
	The Call of the Wild	31891	l2
H.G.Wells	The War of the Worlds	60187	w1
	The Invisible Man	48599	w2

The kind of pattern one typically finds when studying the inter-textual and intra-textual variation of these constants jointly is exemplified for Guiraud's R in Figure 6.9. The horizontal axis plots the sample size N , the vertical axis plots Guiraud's (1954) R as a function of the sample size. For each of the texts listed in Table 6.1 the development of R at twenty equally-spaced intervals through the text size N of that particular text is shown. Some texts by the same author show a remarkably similar pattern, as, e.g., the two texts by Carroll (a1 and a2). Other texts by the same author show very different developmental profiles, such as the texts by Baum (b1 and b2), which are separated

from each other by the texts by Carroll and St Luke. The non-constancy of R makes it difficult to compare texts directly. For instance, R is on the increase throughout *The War of the Worlds* (w1), while after an initial small rise it is continuously decreasing for *The Sea Wolf* (l1). What interpretation should we assign to the circumstantial fact that both texts end up with a roughly similar value of R ?

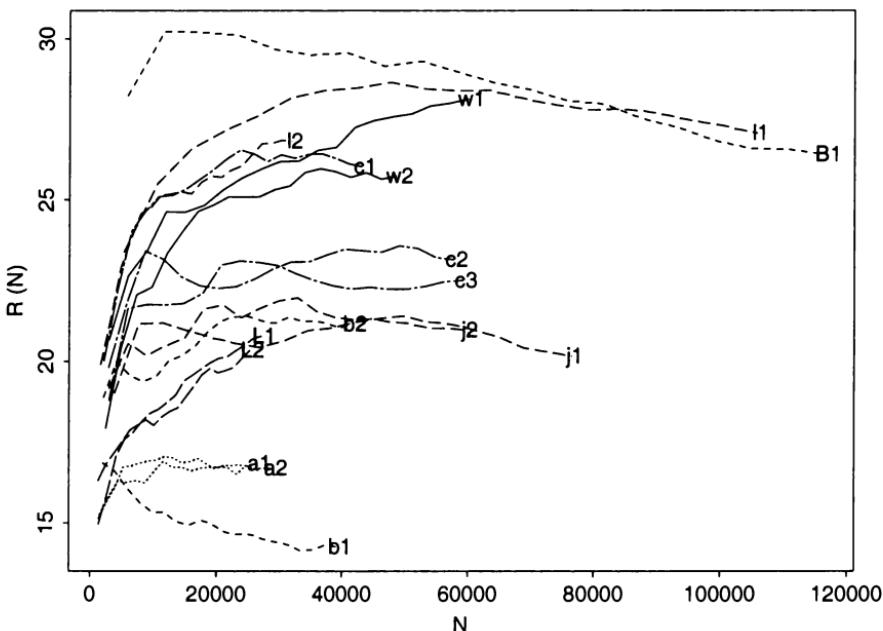


Figure 6.9: Guiraud's R as a function of N for selected English texts. For the key to the abbreviations, see Table 6.1.

Although R is too variable to allow reliable authorial separation, it is nevertheless clear that the developmental profiles of $R(N)$ tend to be similar for texts by the same author. A detailed survey of lexical constants and their variability by Tweedie and Baayen (1998) suggests that the constants partition into two sets with as most stable representatives K and Z . Yule's K captures the repetitiveness of the text, while Z focuses on its lexical richness. With respect to Z , it should be kept in mind that in this study Z was estimated for the simple extended Zipf's law by requiring that $E[V(N)] = V(N)$. This means that accuracy is guaranteed (to the extent that this is possible without adjusting for non-randomness in word use) only for the growth curve of the vocabulary. In many cases, an adequate characterization of the complete frequency spectrum will require the Yule-Simon model or one of the other LNRE models. As a measure of lexical richness, however, Z by itself emerged as the most constant measure of lexical richness with good inter-textual discriminatory power.

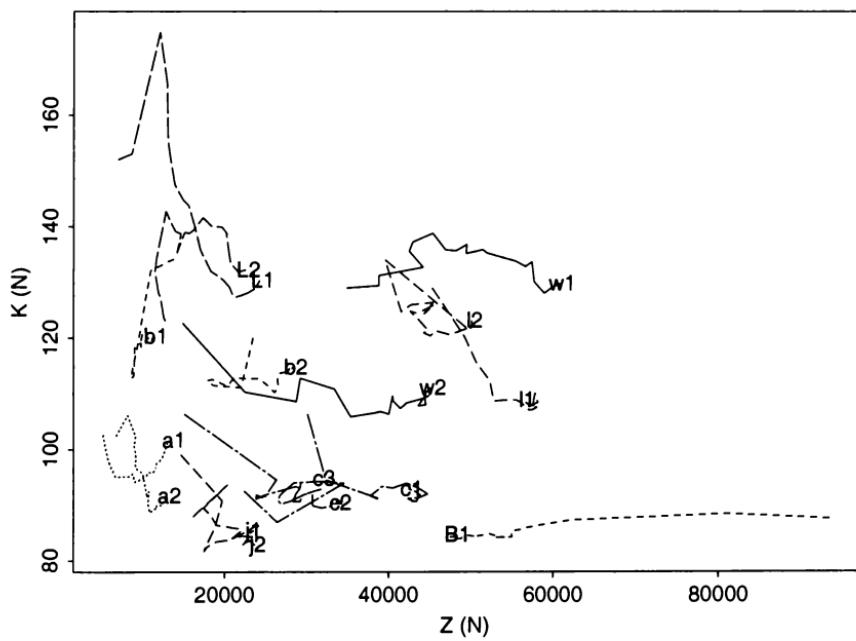


Figure 6.10: Developmental profiles in the plane spanned by $Z(N)$ and $K(N)$ for selected English texts. For the key to the abbreviations, see Table 6.1.

Figure 6.10 plots the developmental profiles of K and Z for the same texts as listed in Table 6.1, for 20 equally-spaced intervals through text time. The legend of each text is placed at the final measurement point, the full text size. For Brontë's *Wuthering Heights*, for instance, the value of K remains relatively constant throughout the text, while Z decreases from more than 90000 to less than 50000. The developmental profiles in the Z - K plane reveal reasonable but not complete authorial separation. The texts by St Luke (L1, L2), Carroll (a1, a2), James (j1, j2), Conan-Doyle (c1, c2, c3), and london (l1m l2) occupy nearly distinct regions. The texts by Baum (b1, b2) and Wells (w1, w2) show that intra-author variability may be greater than inter-author variability.

It will be clear that questions of authorship and style cannot be resolved in terms of characteristic summary measures of the frequency spectrum only. Much better results can be obtained by combining Z and K with other diagnostic features such as the relative frequencies of the highest-frequency function words (Mosteller and Wallace, 1964; Burrows, 1992), the relative frequencies of syntactic constructions (Baayen, Van Halteren, and Tweedie, 1996), and the productivity P of a wide range of morphological categories (Baayen, 1994). At the same time, the claim advanced by Naranan and Balasubrahmanyam (1998:38) that '*a word frequency analysis of a text can reveal nothing about its characteristics* (author, language, style, type of literature, etc.)' is also unfounded. The frequency spectrum summarized in terms of K and Z , although not a complete and infallible guide on its own, captures important aspects of authorial structure.

6.4 Beyond word frequency distributions

This section presents four examples of applications of LNRE models outside the domain of word frequency distributions. The first example comes from biology, the second example comes from history, the third and fourth examples are again from the domain of linguistics.

6.4.1 Counts of filarial worms on mites on rats

An example of an LNRE distribution in biology that has been studied in detail with respect to the generalized inverse Gauss-Poisson model concerns counts of filarial worms on mites on rats. A total of 2600 mites on rats infected with Cotton Rat Filariasis were examined microscopically after feeding. The number of filarial worms on each mite was recorded. The number of mites with no filarial worms was 1155, the number of mites with one filarial worm was 553, and the total number of filarial worms observed was 1445.

Heller (1997) discusses the maximum likelihood estimates for the parameters of the generalized inverse Gauss-Poisson model for the data, using the frequency ranks $m = 1, 2, \dots$. Figure 6.11 plots the observed counts by means of circles, as well as the generalized inverse Gauss-Poisson fit (sc gigp, solid line) and the Yule-Simon fit (dashed line). The Yule-Simon fit does not capture the pattern in the data. By contrast, the GIGP model provides an excellent fit

with a MSE of 13.11 and $X^2(13) = 13.0, p = 0.4476$ for $Z = 265.664, b = 0.1784$, and $\gamma = -0.4468$. The population number of types is estimated at 2645. Given the observed 1445 types, the estimated number of types with zero frequency is 1200, which is of the same order of magnitude as the actual number of types with zero frequency, 1155.

The parameters for this fit were estimated using cost function $C_2(10)$, i.e., the mean squared error was minimized for the first 10 frequency ranks. The resulting parameter values are very similar to those obtained by Heller (1997) using maximum likelihood estimates: $Z = 268.0575, b = 0.1747$, and $\gamma = -0.4457$.¹ The number of types with zero frequency is estimated to be 1193. All these values are reassuringly close to the values obtained using the simplex downhill estimation method with the MSE cost function.

6.4.2 Year references

Polman and Baayen (2000) studied the references to years in the time period 1200–1993 appearing in the 1994 issues of three newspapers. Aspects of the frequency distribution of year references of the Frankfurter *Algemeine Zeitung* are summarized in Figure 6.12. The upper left panel plots the frequency spectrum using logarithmic transformations for both m and $V(m)$. The marked downward curvature at the left side of the graph shows that this is a distribution with relatively few rare events, and this is confirmed by the low value of the coefficient of loss: $C_L = 0.011$. The scarcity of low-frequency year references is due to the (practical) decision not to include the occasional references to years before 1200 in the study. At the right side of the plot, we observe the familiar scatter of very high values of m for which $V(m) = 1$. The structure of the spectrum is better brought to light by plotting the real-valued $V_r(m)$ against m using the smoothing technique described in section 1.3. Note that there seems to be some discontinuity around $\log m = 6$.

The lower left panel plots the rank-frequency distribution. Interestingly, we again observe a discontinuity around $\log \text{rank} = 4$, which corresponds to $\log \text{frequency} = 6$. The reason for this discontinuity becomes apparent when the lower right panel is considered. This panel plots log distance in time from 1994 on the horizontal axis, and log frequency on the vertical axis. Thus, 1993 is at distance 1, 1992 at distance 2, etc. Distance in time supplies an external ranking on the frequencies of year references that we can compare to the internal ranking supplied by the frequencies themselves. A comparison of the left and right panels suggests that the rank-frequency distribution captures the main pattern of the distance-frequency distribution and ignores the noise. The breakpoint that is clearly visible even in the distance-frequency distribution can be located around 1935 and is statistically reliable (see Polman and Baayen, 2000, for further details). Other newspapers also show a breakpoint just before or during the second world war, suggesting that this is a pivotal period for present-day historical consciousness.

¹Heller (1979) makes use of a re-parameterization in which b and $Z = 1/c$ are replaced by $\alpha = b\sqrt{1 + cN}$ and $\xi = (bcN)/2$. Conversely, $b = \sqrt{\alpha^2 + \xi^2} - \xi$, and $c = \frac{2\xi}{N(\sqrt{\alpha^2 + \xi^2} - \xi)}$.

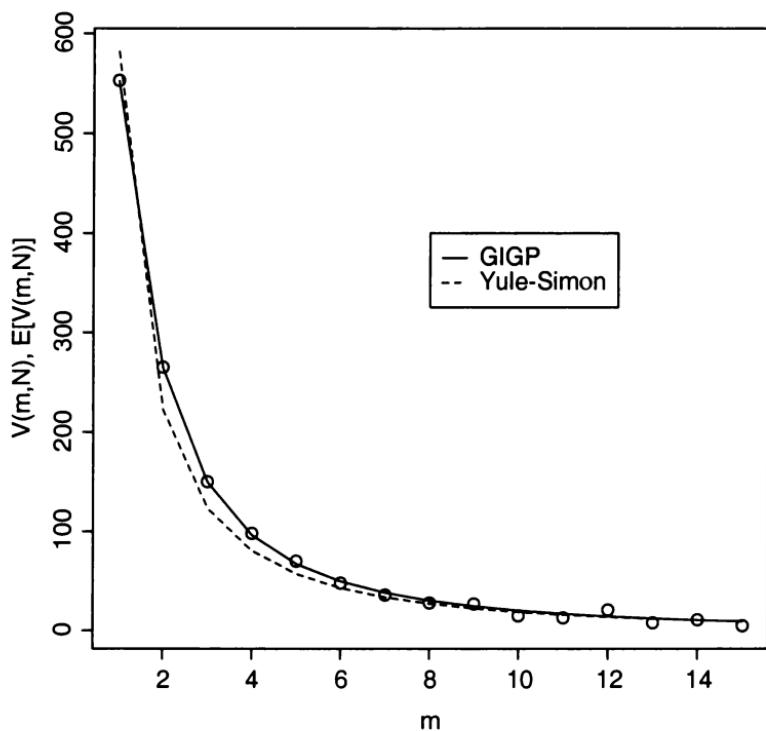


Figure 6.11: Fit of the generalized inverse Gauss-Poisson model (GIGP) and the Yule-Simon model to the data on filarial worms on mites on rats (Heller, 1997).

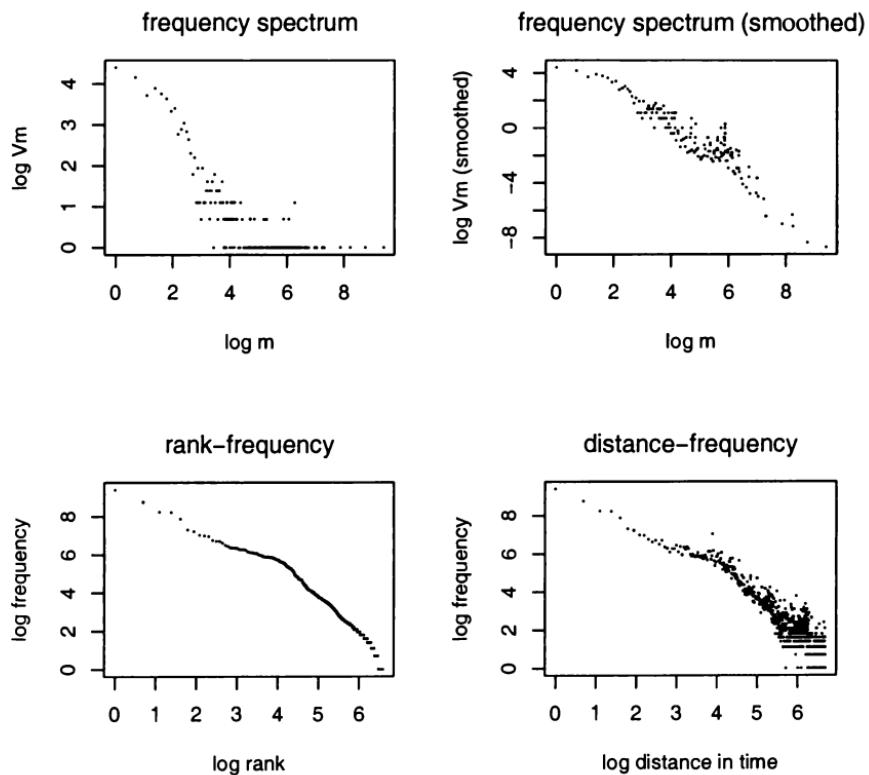


Figure 6.12: References to years in the newspaper issues of the Frankfurter Allgemeine Zeitung that appeared in 1994. The upper left panel plots the frequency spectrum, the upper right panel the smoothed frequency spectrum, the lower left panel shows the rank-frequency distribution, and the lower right panel the distance-frequency distribution.

Breakpoints such as observed here fall outside the scope of LNRE analyses. In the case of the year references in the Frankfurter Allgemeine Zeitung, the GIGP nevertheless provides a reasonable fit for $\hat{Z} = 8.6792$, $\hat{b} = 0.0318$, and $\hat{\gamma} = -0.5752$ ($X^2_{(13)} = 21.63$, $p = 0.0613$). Interestingly, the estimated number of year references in the population equals $\hat{S} = 782$. The actually observed number of year-references is 731, while the maximum number of different year-references in the period 1200–1993 is 793. The similarity of the logically possible number of reference types and the estimated number of possible reference types suggests that the GIGP succeeds quite well in capturing the overall statistical structure of these data.

6.4.3 CV-structures

Gale and Sampson (1995) discuss the calculation of Good-Turing estimates for a count of consonant-vowel-reduced vowel patterns in the TIMIT speech database. Instead of applying a power model to the estimated real-valued spectrum elements and correcting where necessary for counts that diverge too much (see section 1.3 for details), we will investigate whether an LNRE model can provide an acceptable fit.

Figure 6.13 plots the head of the frequency spectrum using circles, the generalized inverse Gauss-Poisson fit using a solid line, the lognormal fit using a dashed line, and the Yule-Simon fit using a dotted line. Again, the generalized inverse Gauss-Poisson model provides a very good fit ($MSE = 5.70$, $x^2_{(13)} = 17.91$, $p = 0.161$ for $\hat{Z} = 2.9473$, $\hat{b} = 0.00233$, and $\hat{\gamma} = -0.4199$, using cost function $C_2(10)$). Of the other two LNRE models, the lognormal model also provides a reasonable fit, at least in terms of the MSE, 14.62, for $\hat{Z} = 1.535$ and $\hat{\sigma} = 3.746$ (no positive chi-squared value is available for this fit). The Yule-Simon model does not provide an accurate fit ($MSE = 32.713$, $X^2_{(13)} = 45.83$, $p = 0.00002$).

Figure 6.14 plots the complete spectrum using the approximated real-valued spectrum elements to facilitate comparison with Gale and Sampson (1995). The generalized inverse Gauss-Poisson fit is represented by a solid line. A Naranan-Balasubrahmanyam fit (using cost function $C_3(20)$, $\hat{C} = 85.944$, $\hat{\mu} = -0.3269$, $\hat{\gamma} = 1.389$, $MSE = 5.42$) is shown using a dashed line. The dotted line represents a least-squares regression line.

First, observe that the least-squares regression line does not do justice to the lowest-frequency ranks. Second, note that the Naranan-Balasubrahmanyam fit and the least-squares fit are indistinguishable for the highest-frequency ranks. Third, we see that the Naranan-Balasubrahmanyam fit and the Sichel fit make very similar predictions for the low-frequency ranks. The reason that the Sichel fit predicts lower values for the highest-frequency ranks might be a matter of accumulating error in the recursive calculation of $\alpha(m, N)$ (see equation (3.21) in chapter 2). This error is also visible when we compare the empirical sample size $N = 30934$ with $N' = \sum_m m E[V(m, N)]$. In theory, the two should be identical, but in the case of the Sichel fit $N - N' = 8121.34$, the order of magnitude of the highest-frequency type (7846), while in the case of the Naranan-Balasubrahmanyam fit, $N - N' = -2749.48$.

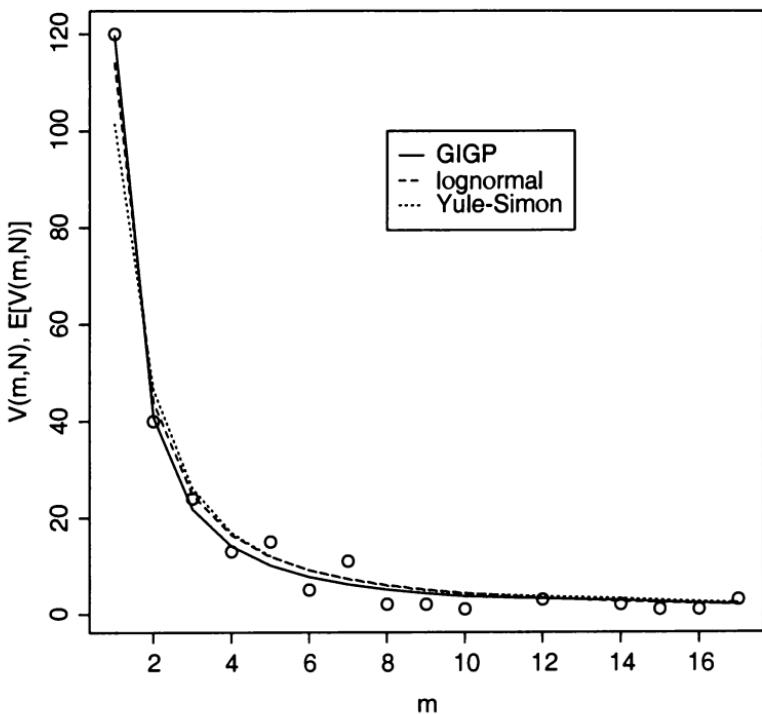


Figure 6.13: Fit of the generalized inverse Gauss-Poisson model (GIGP, solid line), the lognormal model (dashed line), and the Yule-Simon model (dotted line) to the first 15 ranks of the frequency spectrum of Gale and Sampson's (1995) counts of consonant-vowel patterns in English.

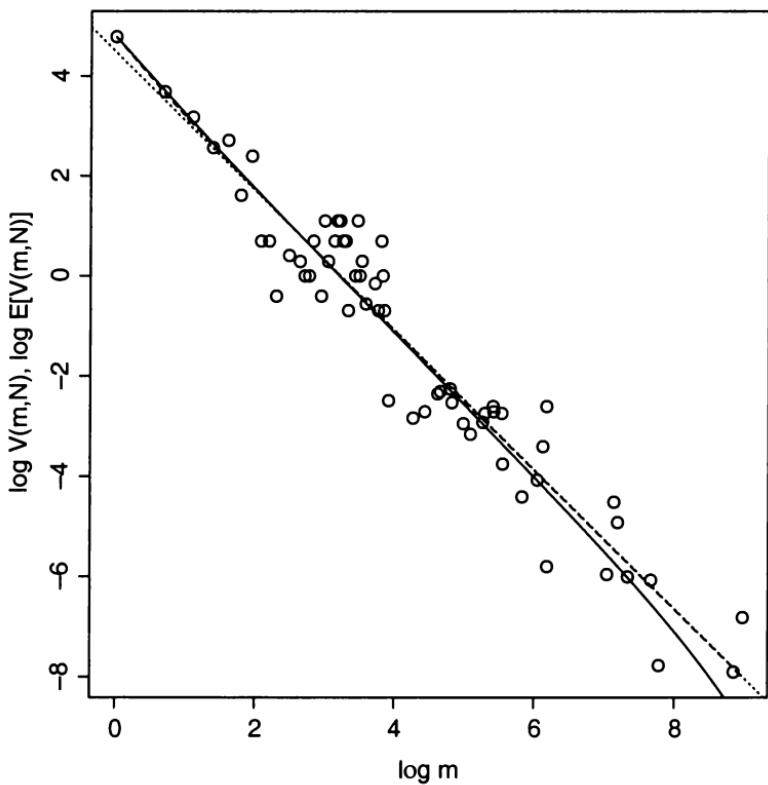


Figure 6.14: Fit of the generalized inverse Gauss-Poisson model (GIGP, solid line), the Naranan-Balasubrahmanyam Zipfian model (dashed line), and a least-squares regression (dotted line) for the frequency spectrum of Gale and Sampson's (1995) counts of consonant-vowel patterns in English.

This suggests that the Naranan-Balasubrahmanyam model is a good choice when smoothed values are required across the full frequency spectrum. When the purpose of modeling is to gain insight in the dynamic aspects of the distribution (interpolation, extrapolation, estimation of the number of types with zero frequency, in this example $\hat{S} - E[V(N)] = 1173 - 310 = 863$), the generalized inverse Gauss-Poisson model is the obvious choice.

6.4.4 Word pairs

With the last example we enter the domain of *very* large number of rare event distributions. Instead of counting words, we now count word pairs, or, in the terminology of computational linguistics, bigrams. The left panel of Figure 6.15 plots the first 20 spectrum elements for the bigrams in Multatuli's *Max Havelaar* using circles, the generalized inverse Gauss-Poisson fit ($\hat{Z} = 1378.72$, $\hat{b} = 0.0186$, $\hat{\gamma} = -0.9999999815$, $MSE = 630.39$, $X^2_{(13)} = 85.68$) using a solid line, and the Naranan-Balasubrahmanyam fit ($\hat{C} = 18185.82$, $\hat{\mu} = -0.9713$, $\hat{\gamma} = 2.311$, $MSE = 1834.68$) using a dashed line. The left panel highlights the extremely large number of hapax legomena that is typical for bigram data. The right panel uses a logarithmic scale for the Y-axis to make the difference between the two fits for the larger values of m visible to the eye.

Table 6.2 tabulates N , $V(N)$, and the first two spectrum elements and their expected values. Note that the Naranan-Balasubrahmanyam fit is less good than the LNRE fit, to some extent for $V(N)$, and markedly so for $V(1, N)$ and $V(2, N)$.

Table 6.2: *Observed and expected values for the bigram counts in Multatuli's Max Havelaar.*

	observed	Sichel	Naranan-Balasubrahmanyam
N	99766	99765	99770
$V(N)$	59156	59168	59088
$V(1, N)$	47974	47952	48037
$V(2, N)$	5825	5813	5957

The number of tokens in this example, 99766, is quite small compared to the huge bigram counts that are currently analyzed in the context of language engineering, and it remains to be seen whether the generalized inverse Gauss-Poisson model is flexible enough to handle bigram data calculated for corpora with tens of even hundreds of millions of words.

6.4.5 Discussion

Of the three LNRE models, the one that appears to be the most useful outside the domain of word frequency distributions is the generalized inverse Gauss-Poisson model. When a fit to the spectrum suffices, the Naranan-Balasubrahmanyam model is a good alternative. The Yule-Simon and lognormal models, which when applied to word frequency distributions often emerged as

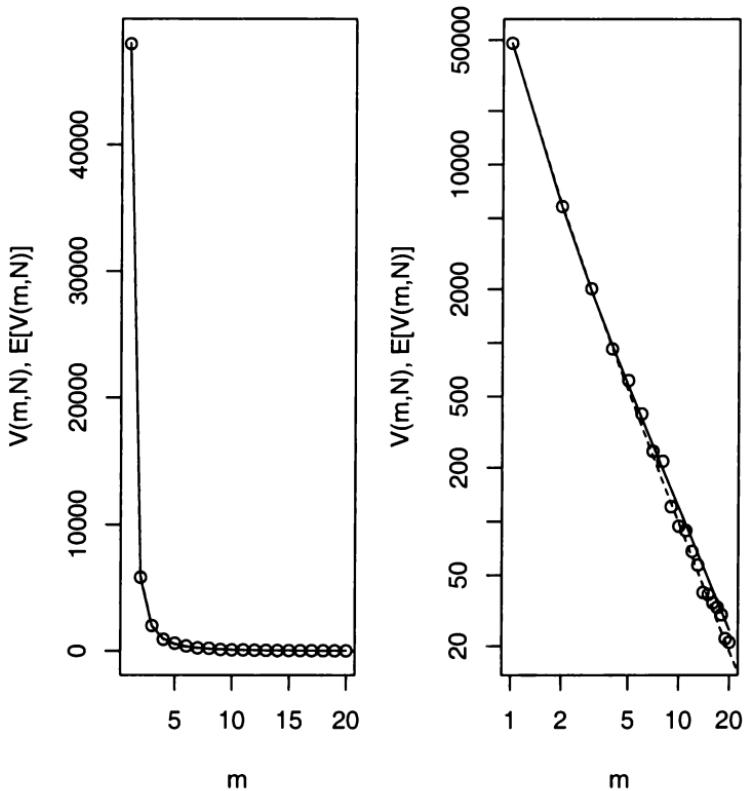


Figure 6.15: Fit of the generalized inverse Gauss-Poisson model (GIGP, solid line) and the Naranan-Balasubrahmanyam Zipfian model (dashed line) to the first 20 spectrum elements of the word pairs in Multatuli's Max Havelaar. The left panel plots absolute values, the right panel uses a logarithmic scale for both axes.

superior to the generalized inverse Gauss-Poisson model, did not provide acceptable fits. It is at present unclear why the models show this almost complementary distribution of applicability, although it is interesting to keep in mind that specific stochastic rationales are available for the lognormal model (Carroll, 1969, Chitashvili and Baayen, 1993) and the Yule-Simon model (Simon, 1955) as models for word frequency distributions, while no such rationale is known for the generalized inverse Gauss-Poisson model.

6.5 Some practical guidelines

This section offers some practical guidelines for analyzing word frequency distributions. By working through a simple example, the comparison of *Alice in wonderland* and *Through the looking-glass*, we illustrate both a simple but useful general approach as well as a series of practical commands: commands at the level of the (LINUX/UNIX) command shell, commands for basic statistical processing using the statistical programming environment of R² (or Splus), and commands in the graphical user interface of the LEXSTATS library documented in Appendix C for carrying out specific LNRE modelling.

The very first stage of an analysis will often consist of preprocessing. The LEXSTATS library contains a utility, spectrum, that takes a text file as input and produces a variety of output files. By default, the input file is assumed to have the extension .txt, and, again by default, any SGML markup is removed. Given the availability of the text file alice.txt in the current directory, we execute the following commands in the shell:

```
% spectrum alice.txt
spectrum: completed first scan of alice.txt: N = 26505
spectrum: calculating developmental profile of alice.txt
[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17]
[18] [19] [20]
spectrum: completed second scan of alice.txt:
N = 26505 V = 2651
printed developmental profile
printed word frequency list
printed frequency spectrum
printed summary statistics
sorting word frequency list
scanning text and printing in zvec format
% head alice.spc # show first 10 lines of alice.spc
m Vm
1 1176
2 402
3 233
4 154
5 99
6 57
```

²For detailed information about the R project, also known as GNU S, see the web page at <http://www.r-project.org>.

```
7 65
8 52
9 32
```

Spectrum outputs `alice.spc`, the first 10 lines of which are shown above. Note that columns are labelled. As we shall see below, this makes it very easy to load these files into R or Splus. The most important other files produced by spectrum are `alice.obs`, which tabulates the values of various lexical constants as described in section 1.4, for by default 20 equally-spaced intervals, `alice.sum`, which summarizes the values of various lexical measures for the full text size, and `alice.wf1`, which is a word frequency list.

A useful second step in many analyses is to inspect the vocabulary growth curve. The empirical growth curves are available in `alice.obs` with column labels N and V for N and $V(N)$. The expected growth curves can be obtained using `binomint`, which applies binomial interpolation as defined by (2.42). After having applied spectrum to `through.txt` to obtain `through.spc`, we use `binomint` to obtain the files `alice.int` and `through.int`:

```
% binomint alice.spc
binomint: loaded spectrum, N = 26505, V = 2651
% binomint through.spc
binomint: loaded spectrum, N = 28767, V = 3085
% head alice.spc
N EV EV1 EV2 EV3 EV4 EV5
1326.00 509.49 333.29 74.30 32.56 18.45 11.51
2652.00 789.59 481.65 119.51 51.92 30.09 20.51
3978.00 1005.62 585.73 158.20 69.80 38.85 25.61
5304.00 1185.47 665.33 190.90 87.46 48.52 30.72
6630.00 1341.07 729.67 218.30 103.61 58.78 36.82
7955.00 1478.92 783.75 241.61 117.70 68.63 43.62
9280.00 1603.26 830.52 261.96 129.92 77.43 50.49
10605.00 1716.85 871.78 280.13 140.73 84.99 56.82
11930.00 1821.65 908.69 296.58 150.61 91.47 62.29
```

By default, `binomint` carries out binomial interpolation for 20 equally spaced intervals for $E[V(N)]$ and the first 5 spectrum elements. We can plot the expected growth curves of *Alice in Wonderland* and *Through the looking glass* using R by typing the following commands to the R prompt:

```
> alice.bin = read.table("alice.bin", T)
> alice.bin[1:4, 1:3]
N EV EV1
1 1326 509.49 333.29
2 2652 789.59 481.65
3 3978 1005.62 585.73
4 5304 1185.47 665.33
> through.bin = read.table("through.bin", T)
> plot(alice.bin$N, alice.bin$EV,
+ xlab="N", ylab="E[V(N)]",
+ xlim=range(0, alice.bin$N, through.bin$N),
```

```
+ ylim = range(0, alice.bin$EV, through.bin$EV),
+ type="p", pch=1)
> points(through.bin$N, through.bin$EV, pch=2)
> lines(alice.bin$N[1:10], alice.bin$EV[1:10] + 1.96*
+ sqrt(alice.bin$EV[seq(2,20,2)] - alice.bin$EV[1:10]),
+ lty=1)
> lines(alice.bin$N[1:10], alice.bin$EV[1:10] - 1.96*
+ sqrt(alice.bin$EV[seq(2,20,2)] - alice.bin$EV[1:10]),
+ lty=1)
> lines(through.bin$N[1:10], through.bin$EV[1:10] + 1.96*
+ sqrt(through.bin$EV[seq(2,20,2)] - through.bin$EV[1:10]),
+ lty=2)
> lines(through.bin$N[1:10], through.bin$EV[1:10] - 1.96*
+ sqrt(through.bin$EV[seq(2,20,2)] - through.bin$EV[1:10]),
+ lty=2)
```

The command `read.table("alice.spc", T)` loads the file `alice.spc` into R. The second argument, T, specifies that the columns of filename are preceded by a header line with their names. The result of `read.table` is assigned (by the assignment operator `->`) to the data frame `alice.bin`. By imposing restrictions on its rows and columns, we can select subsections of the data frame for display. In the above example, the first four rows and first three columns were selected for display. Note that R expands an expression such as `1:4` into the vector `(1, 2, 3, 4)`. Columns of a data frame are specified by means of the \$ operator followed by the appropriate column name. For example, `alice.bin$EV` is the vector of the expected values of the vocabulary size. We plot the growth curve of *Alice in Wonderland* using the `plot` function, where we take care (by specifying `xlim` and `ylim` explicitly) to set up the tick marks on the horizontal and vertical axes in such a way that the full range of values of N and $E[V(N)]$ in both texts can be accommodated. We then add the data points from *Through the looking-glass* using the `points` function. Finally, we add the 95% confidence intervals for the first halves of both texts, which are easily calculated since $\text{VAR}[V(N)] = E[V(2N)] - E[V(N)]$ by (3.62). For *Alice in Wonderland*, we use solid lines (`lty=1`), for *Through the looking-glass*, we use dashed lines (`lty=2`). The result is shown in Figure 6.16. Note that by the end of the first halves of both texts, the confidence intervals no longer overlap. This indicates that *Through the looking-glass* has a reliably higher vocabulary richness.

By inspecting the frequency spectra, we can gain further insight into this difference in type richness. We read both spectra into R, and plot the first 15 spectrum elements using circles for *Alice in Wonderland* (`pch=1`) and triangles for *Through the looking-glass* (`pch=2`). We also add the corresponding expected values according to Sichel's GIGP model. These expectations can be obtained by typing `lnreSgam alice.spc` at the shell prompt. This program will interactively ask the user to specify on the command line how to fit the model. The graphical user interface of LEXSTATS discussed below makes it much easier to use and inspect LNRE models. The reader is referred to Appendix C for further technical details concerning the use of `lnreSgam` by itself. The expected spectra of *Alice in Wonderland* and *Through the looking-glass*

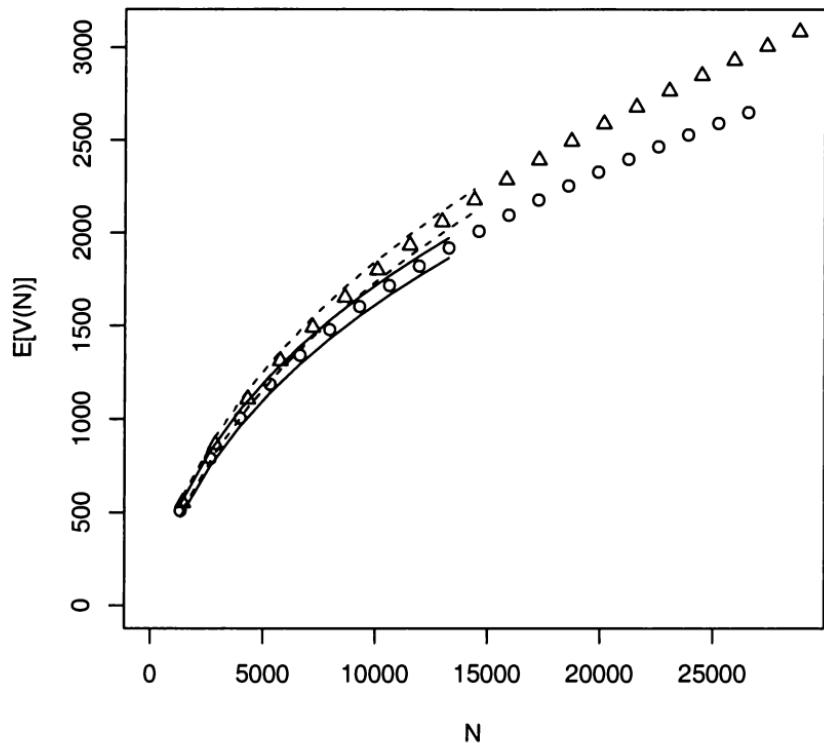


Figure 6.16: *Expected vocabulary growth curves for Alice in Wonderland and Through the looking glass using binomial interpolation.*

as calculated from `alice.spc` and `through.spc` become available in the files `alice_G.spc` and `through_G.spc` respectively. We read these tables into R and plot the expectations using a solid line for *Alice in wonderland* and a dashed line for *Through the looking-glass*. We plot the head of the spectrum by restricting the vectors `$m` and `$Vm` to the first 15 elements. The result is shown in Figure 6.17.

```
> alice.spc = read.table("alice.spc", T)
> through.spc = read.table("through.spc", T)
plot(alice.spc$m[1:15], alice.spc$Vm[1:15],
+ xlab="m", ylab="V(m,N), E[V(m,N)]",
+ ylim = range(0, alice.spc$Vm[1:15], through.spc$Vm[1:15]),
+ type="p", pch=1)
> points(through.spc$m[1:15], through.spc$Vm[1:15], pch=2)
```

```
> alice.G.spc = read.table("alice_G.spc", T)
> through.G.spc = read.table("through_G.spc", T)
> lines(alice.G.spc$m[1:15], alice.G.spc$EVm, lty=1)
> lines(through.G.spc$m[1:15], through.G.spc$EVm, lty=2)
```

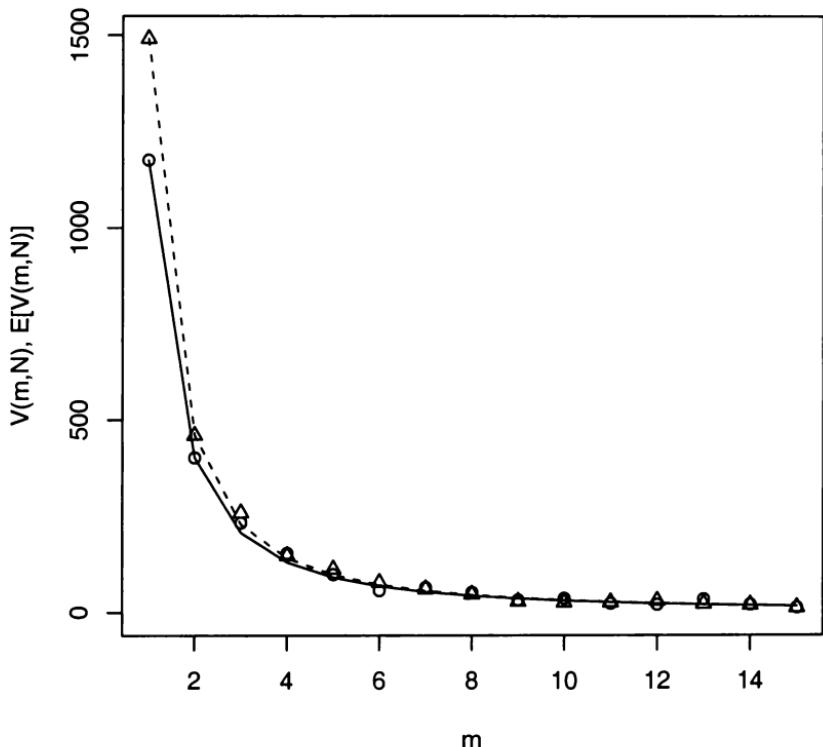


Figure 6.17: Observed (circles and triangles) and expected spectrum (solid and dashed lines) for Alice in Wonderland and Through the looking glass respectively.

Note that the spectra of *Alice in Wonderland* and *Through the looking-glass* differ most substantially for the lowest values of m . We may ask ourselves for what values of m these differences are statistically reliable. To answer this question, we need the covariance matrices (3.58) for both texts. Because calculation of the covariance matrix requires knowledge of various expectations at twice the observed sample size, a fitted LNRE model is required. In this example, we have the GIGP fits available. The following command typed at the shell prompt,

```
% lnreChi2 alice.G.spc
x2(13) = 267.32, p = 0.000000, MSE = 428.02, rMSE = 0.000061
```

carries out the multivariate chi-squared test of goodness of fit (3.59). For *Alice in Wonderland*, a straightforward GIGP model clearly does not provide an appropriate fit, even though the fits are quite reasonable to the eye, as can be seen in Figure 6.17. As explained in (3.5), a much better fit can be obtained by using a different cost function (C_2 , defined in (3.66)). The `lnreChi2` program produces a summary of the evaluation of goodness of fit in `alice.G.chi`, and it calculates the covariance matrix which is saved in the file `alice.G.cov`. We load both covariance matrices into R, keeping in mind that the variances of the spectrum elements $m = 1, 2, \dots$ (see equation 3.61) are on the main diagonal at positions `alice.G.cov[m+1,m+1]` since `alice.G.cov[1,1]` contains the variance of the vocabulary size. We define a function in R that takes two spectra and the corresponding covariance matrices as input and that calculates, for a specified frequency m , the probability that the expected values of $E[V(m, N)]$ are the same using (3.68). Applied to the first three values of m , we see that only the numbers of hapax and dis legomena differ reliably. We also define a second function that tests whether the vocabulary growth rates of the two texts at their full text sizes are reliably different, using (3.70). The sequence of commands in R is:

```
> through.G.cov - read.table("through.G.cov")
> alice.G.cov - read.table("alice.G.cov")
> spectrum.Z.fnc - function(spc1, spc2, cov1, cov2, m) {
+ return(2*(1-pnorm(abs((spc1$EVm[m] - spc2$EVm[m])/
+ sqrt(cov1[m+1,m+1]+cov2[m+1,m+1])))))}
> spectrum.Z.fnc(alice.G.spc, through.G.spc, alice.G.cov,
+ through.G.cov, 1)
[1] 4.487077e-12
> spectrum.Z.fnc(alice.G.spc, through.G.spc, alice.G.cov,
+ through.G.cov, 2)
[1] 0.03029123
> spectrum.Z.fnc(alice.G.spc, through.G.spc, alice.G.cov,
+ through.G.cov, 3)
[1] 0.2610231
> growthRate.Z.fnc - function(spc1, spc2, cov1, cov2) {
+ N1 - sum(spc1$m*spc1$Vm)
+ N2 - sum(spc2$m*spc2$Vm)
+ return(2*(1-pnorm(abs((spc1$EVm[1]/N1 - spc2$EVm[1]/N2)/
+ sqrt(cov1[2,2]/(N1*N1)+cov2[2,2]/(N2*N2))))))
> growthRate.Z.fnc(alice.G.spc, through.G.spc, alice.G.cov,
+ through.G.cov)
[1] 3.647066e-07
```

The main ideas underlying the present example are the following. As a first step, inspect the growth curves, applying simple binomial interpolation and estimating the variance of the vocabulary size for the first half of the text. Inspection of the spectrum is useful for tracing how many spectrum elements contribute to observed differences in vocabulary size. If a simple fit to the

frequency spectrum is required, the Naranan-Balasubrahmanyam model is a reliable and fast choice. The program `spectfit` provides this fit, together with Good-Turing frequency estimates (see Appendix C for further details).

For more detailed analyses, LNRE models are required. Of the three LNRE models, the generalized inverse Gauss-Poisson model is computationally convenient, while the Yule-Simon model and the lognormal model are quite computationally intensive. The generalized inverse Gauss-Poisson model is a very useful tool to start with. If it does not provide a good fit, it is worth considering using another cost function. The Yule-Simon model often is an excellent albeit computationally cumbersome alternative. For distributions in the late LNRE zone, the lognormal model may be a good first choice. Sometimes, mixture models are required. Unfortunately, fitting mixture models to the data is as yet not an automatic process and hence time-consuming.

The choice between the LNRE models should also be guided by one's a priori assumptions with respect to whether the population number of types is finite or possibly infinite. In the example of filarial worms on mites on rats in section 6.4.1, for instance, it is clear from the start that S must be finite. This makes the Yule-Simon model less attractive in case its fit is based on parameter values that imply an infinite population size.

The multivariate test of goodness of fit (3.59) is quite severe. Often, it rejects fits that seem quite adequate to the eye when plotted. Two things should be kept in mind here. First, it may be that there is some systematic way in which the observed and expected counts differ. This may indicate that, for instance, a mixture model is called for instead of a simple LNRE model. Second, non-systematic irregularities in the shape of the spectrum often lead to high X^2 values. Such irregularities are probably due to substantial non-randomness in word use. In the absence of systematic differences between the observed and expected spectrum, all that the goodness of fit test tells us in this case is that the text is not the result of a random process.

Finally, we turn to a brief overview of the LEXSTATS graphical user interface to the LNRE program library. The user interface is started by typing `lexstats` to the shell prompt. The results of the analyses carried out by `lexstats` are saved in a directory named `.lexstats`. If such a directory does not exist in the working directory where `lexstats` is run, data will be stored in a `.lexstats` in your home directory. As in R and Splus, this makes it possible to archive result files by project. Unlike in R or Splus, all files in the `.lexstats` are simple ASCII files. In fact, other programs such as R can be run without problem in the `.lexstats` directory. Any files created independently in this directory should not make use of the LNRE file extensions listed in Appendix C.

The main window that appears when `lexstats` is called is shown in Figure 6.18. Most labels and buttons come with a balloon help, a small window that pops up near the label or button when the mouse pointer is moved onto it and that explains its function. The main window has four panels, the first of which lists the available analyses. In the example shown in Figure 6.18, there is one such analysis with spectrum file `alice.txt`, which has 117 different spectrum values (`spc`), the first 15 of which come with an expected

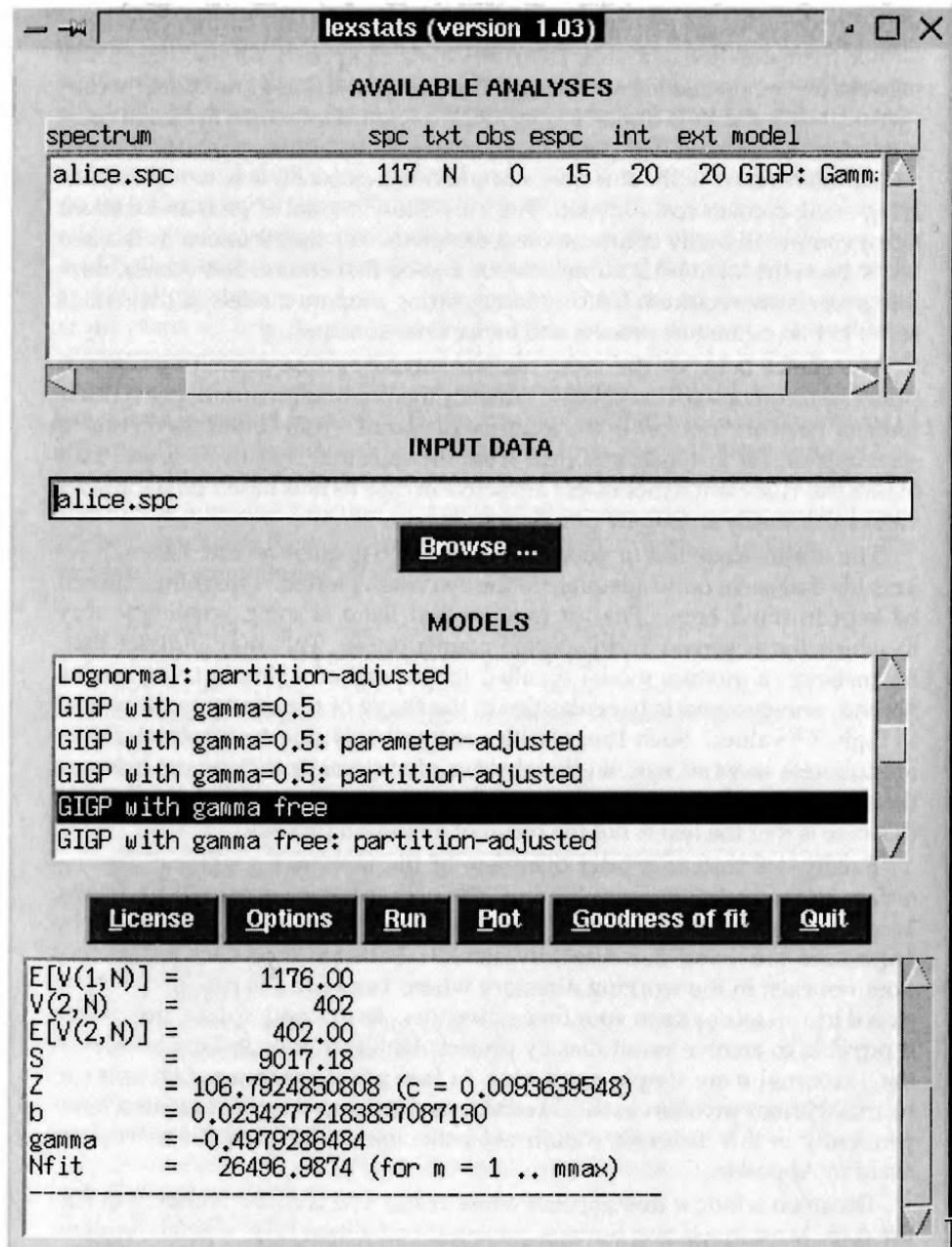


Figure 6.18: Main window of lexstats.

value (espc) according to the GIGP model. There are 20 equally-spaced 20 chunks with interpolated spectrum elements (irt), and 20 such chunks with extrapolated values (ext). In the current .lexstats directory, no observed profile alice.obs is available (obs = 0), nor a file of the form alice.txt with the original text (txt = N). An analysis is selected by a single mouse click with the left button. A following middle mouse click loads the analysis into memory, a following right mouse click deletes the analysis by removing all files pertaining to this analysis from the .lexstats directory.

The panel labelled INPUT DATA is the place to specify the input file for a new analysis. Normally, the input will be a spectrum file such as alice.spc. However, a text file such as alice.txt can also be specified, in which case it will be automatically preprocessed by the spectrum program provided that the text file has the extension .txt. Finally, files in the format of a word frequency list (alice.wf1) are also acceptable as input. The Browse button allows the user to browse the file system for the appropriate input file.

The third panel specifies which models can be applied, LNRE models, including adjusted models and mixture models, as well as the Naranan-Balasubrahmanyam Zipfian spectrum smoother. A model is selected by a single click of the left mouse button. The bottom panel lists the output of the analyses. The buttons above this panel can be used to inspect the GNU general public license under which LEXSTATS is distributed, to exit the program, to specify various options including the range of spectrum elements to be included and the kind of cost function to use for parameter estimation, to inspect the goodness of fit, to plot the results, and, last but not least, to run an analysis.

When the Run button is selected, the window shown in Figure 6.19 pops up. This window presents initial parameter values for running the simplex minimization routine. The button labelled Test Run allows the user to test whether the initial parameters indeed provide a reasonable starting point. Instead of running simplex minimization, it is also possible to run analyses with parameters of one's own choice.

In the case that a mixture model is selected, the window that pops up once Run has been selected in the main window is slightly different, as illustrated in Figure 6.20. The user has to specify the mixture parameter p , or, equivalently, the number of tokens of the mixture to be assigned to the base model. After specification of the parameters of the base model, the button labelled Test Base should be selected to test whether the base model is feasible. A wrong choice of parameter values may lead to a base model that does not leave enough tokens or types for the complement model. Once a base model has been selected, the button labelled Test Complement allows the user to check whether a reasonable complement model is available. If so, the Run button initiates the calculation of the interpolated and extrapolated spectrum values. By selecting Run Parameters, the model will be calculated from the parameters specified by the user for both the base and the complement models.

Once an analysis has been completed, it is possible to visually inspect the fit by selecting the Plot button in the main window. The most important kind of plots available are the spectrum plot, illustrated in Figure 6.21, and

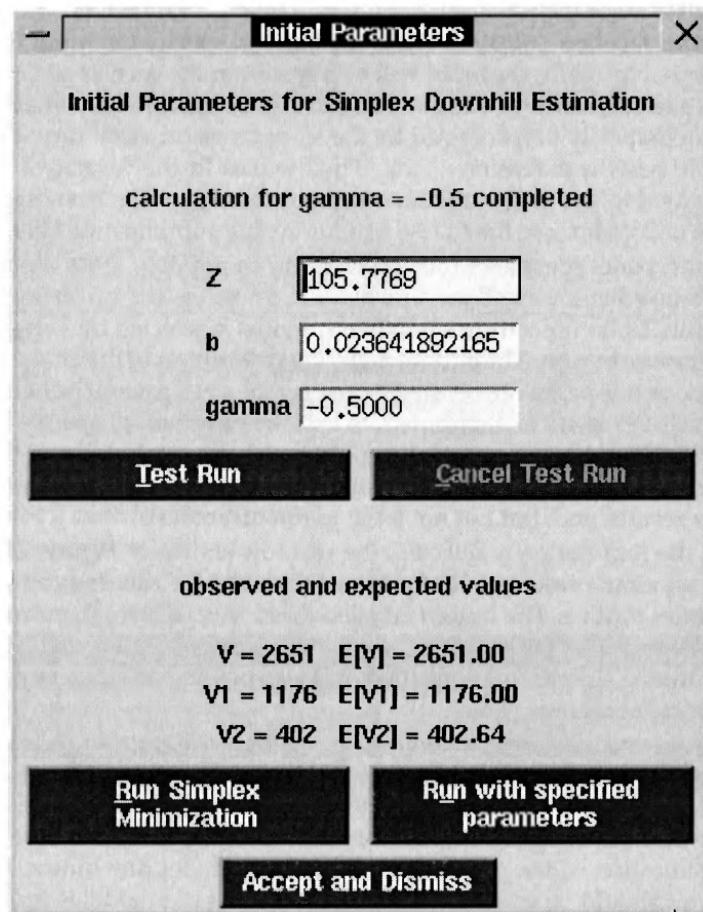


Figure 6.19: Parameter specification window of lexstats.

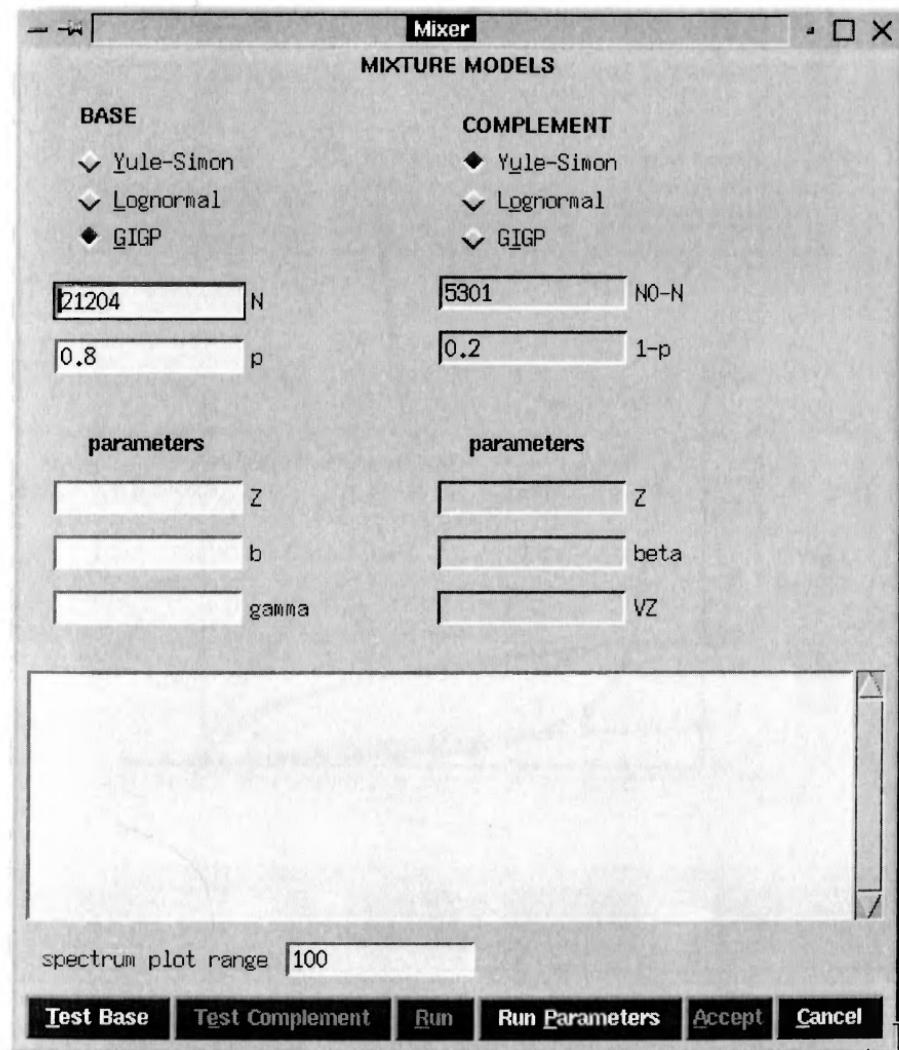


Figure 6.20: Parameter specification window of lexstats for mixture models.

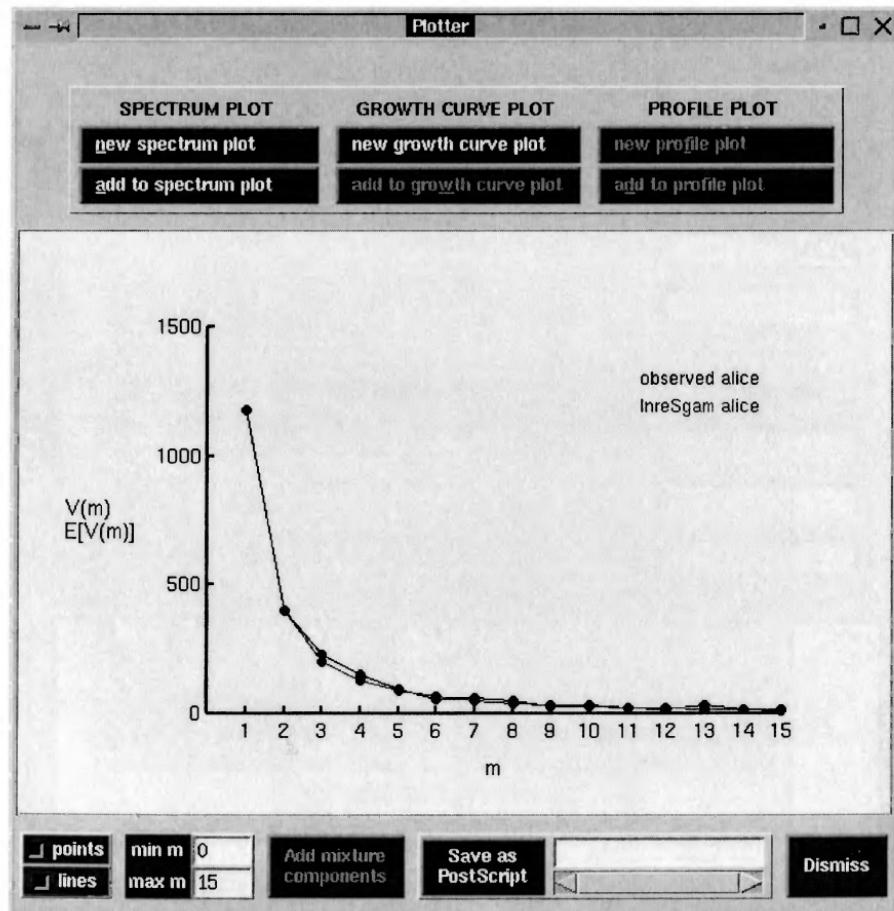


Figure 6.21: Plot window of flexstats for the frequency spectrum.

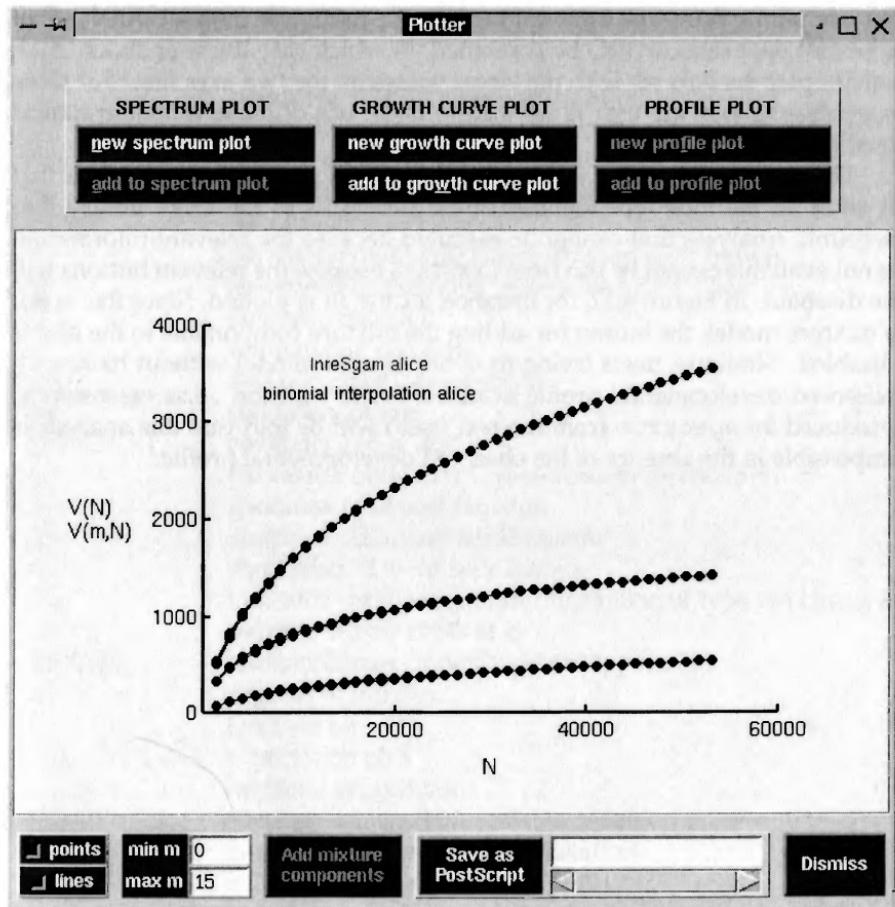


Figure 6.22: Plot window of lexstats for the vocabulary and spectral growth curves.

the growth curve plot, illustrated in Figure 6.22. When running a spectrum plot, it is possible to add the base and complement components in the case of a mixture model. When running a growth curve plot, the user is given the choice to add the binomial interpolation curves and the extrapolated curves as calculated by the selected model. If the observed profile exists in the form of a file such as `alice.obs` in the `.lexstats` directory, the observed growth profile will also be shown. The presence of an observed profile is also required for running a profile plot, a graph in which the developmental profiles of two lexical measures (various characteristic constants, but also N and $V(N)$) can be traced. It is possible to add a new analysis to an existing plot. This is useful for comparing how well different models fit a particular data set. Analyses of different data sets can also be combined, in which case the user should take care to plot the data set with the larger values on the two axes first. For more specialized plots, the user is advised to make use of the excellent graphical facilities of R.

The main advantage of using the LEXSTATS graphical user interface is that it takes all the interdependencies of the programs in the LNRE library into account. Analyses that cannot be executed because the relevant information is not available cannot be run from LEXSTATS because the relevant buttons will be disabled. In Figure 6.22, for instance, a GIGP fit is plotted. Since this is not a mixture model, the button for adding the mixture components to the plot is disabled. Similarly, users trying to fit an adjusted model without having an observed developmental profile available (a file with the `.obs` extension as produced by spectrum from the text itself) will be told that this analysis is impossible in the absence of the observed developmental profile.

Appendix A

List of Symbols

a	parameter of Zipf's zeta distribution
$\alpha(m, N)$	parameter of the Waring-Herdan-Muller distribution
α	relative spectrum $E[V(m, N)]/E[V(N)]$
$B(\cdot, \cdot)$	parameter of extended generalized Zipf's law
b	Beta function
β	parameter of inverse Gauss-Poisson distribution
C	parameter of extended generalized Zipf's law
C_L	Herdan's constant
$\text{COV}[X, Y]$	normalizing constant
c	cost of coding (Mandelbrot)
C	coefficient of loss
D	covariance of X and Y
d_i	parameter of inverse Gauss-Poisson distribution
$d_{i,k}$	goodness of fit cost function
$\Delta V(k)$	Simpson's characteristic constant
$\Delta VU(k)$	dispersion of word type ω_i
df	indicator variable for underdispersion of type i in chunk k
e	number of new types at k
$E[X]$	number of new underdispersed types at k
$E_{bin}[\cdot]$	degrees of freedom
$E_{adj}[\cdot]$	error vector
$E_{part}[\cdot]$	expectation of X
$E_{LNRE}[\cdot]$	binomial expectation
$\hat{E}[X]$	partition-adjusted binomial expectation
ε_{GV}	partition-adjusted LNRE expectation
$F(\pi)$	expectation based on LNRE model
$f(i, N)$	expectation of X estimated from the sample
$f(\omega_W, N)$	estimation error by-token growth rate
$f^*(i, N)$	probability mass of types with $\pi_i \geq \pi$
	frequency of word ω_i in sample of N tokens
	frequency of word ω_W in sample of N tokens
	Good-Turing estimate of $f(i, N)$

$f_{N_0}(i, N)$	$f(i, N)$ conditional on sample of N_0 tokens
$f_z(z, N)$	frequency of word with Zipf rank z in sample of N tokens
$f_{i,k}$	token frequency of i -th word in the k -th text slice
$f_j(x)$	j -th moment-weighted function
$\text{GV}(N)$	by-token vocabulary growth rate $E[V(N+1)] - E[V(N)]$
$G(\pi)$	number of types ω_i with $\pi_i \geq \pi$
$\Gamma(\cdot)$	Gamma function
$g(m, N)$	number of types with $f(i, N) \geq m$
γ	parameter of Sichel's inverse Gauss-Poisson model
	parameter of extended generalized Zipf's law
	parameter of the Naranan-Balasubrahmanyam Zipfian model
γ	Euler's constant
H	Honoré's constant
$H_{(m)}$	by-class token entropy
$I_{[\cdot]}$	indicator operator
i	index for word types $1, \dots, S$
$i(W, N)$	indicator operator for $f(\omega_W, N)$
k	index for text chunks $1, \dots, K$
K	number of (equal-sized) text chunks
	Yule's characteristic constant
	normalizing constant
$K_\gamma(\cdot)$	Bessel function of the second kind of order γ
λ	parameter of Poisson distribution
M	number of letters (Mandelbrot/Miller)
MSE	mean squared error
MSE_r	relative mean squared error
m	frequency rank (spectrum element)
m^*	Good-Turing estimate of m
m'	highest frequency m in the sample
m^o	highest frequency rank for MSE
μ	expected population-based frequency
N	mean
N_0	sample size in word tokens
N_m^*	pivotal sample size for interpolation or extrapolation
$NU(k)$	sample size for which $V(m, N)$ reaches maximum
ω_i	number of underdispersed tokens in chunk k
\mathcal{P}	the i -th word type
\mathcal{P}^*	vocabulary growth rate $E[V(1, N)]/N$
\mathcal{P}^{**}	category-conditioned degree of productivity
\Pr	hapax-conditioned degree of productivity
$\Pr(m)$	unconditioned degree of productivity
p	probability
$p(i, N)$	selection probability of word with frequency m
p^*	Hubert-Labbe coefficient of vocabulary partition
$p^*(i, N)$	sample relative frequency of word ω_i
	maximum sample relative frequency
	Good-Turing estimate of $p(i, N)$

π_i	probability of ω_i
π_z	probability of word type with Zipf rank z
$\hat{\pi}_z$	$f_z(z, N)/N$
$\psi(\pi)$	pdf of inverse Gauss-Poisson model
R	Guiraud's constant
\mathbf{R}	covariance matrix
$R(m, n)$	$V(m, N)/V(n, N)$
R_{nb}	$R(m, n)$ for the Naranan-Balasubrahmanyam smoother
r	number of frequency ranks for covariance matrix
S	Sichel's constant
	population number of types
σ	standard deviation
t	N/N_0
$\text{VAR}[X]$	variance of X
$V(N)$	vocabulary size (number of types among N tokens)
$V_r(m, N)$	real-valued approximate spectrum elements
$V^{(n)}(N)$	n -th derivative of $V(N)$
$V_{N_0}(N)$	$V(N)$ conditional on sample of N_0 tokens
$V_{\cdot, N_0}(N)$	number of types in N with $f(i, N_0) \Rightarrow 0$
$V_{n, N_0}(N)$	number of types in N with $f(i, N_0) = n$
$V(N_k)$	number of types in the first $\frac{kN}{K}$ tokens
$V(\pi)$	number of types with probability π
$V(m, N)$	number of types with frequency m in a sample of N tokens
$V_{N_0}(m, N)$	$V(m, N)$ conditional on sample of N_0 tokens
$V_{\cdot, N_0}(m, N)$	number of types with $f(i, N) = m$ that also occur in N_0 with $f(i, N_0) = n$
$V_{n, N_0}(m, N)$	number of types in N that occur in N_0 with $f(i, N_0) = n$
$VU(k)$	number of underdispersed types in chunk k
W	Brunet's constant
x	random variable ranging over indices $i = 1, 2, \dots, S$
Z	parameter of the Waring-Herdan-Muller distribution
z	Zipf size
	Z-score
	Zipf rank
$\zeta(\cdot)$	Riemann zeta function

Appendix B

Solutions to the exercises

CHAPTER 1

1. Since $\sum_{i=1}^{V(N)} f(i, N) = N$, and given $V(N)$ different words,

$$\frac{1}{V(N)} \sum_{i=1}^{V(N)} f(i, N) = \frac{N}{V(N)}.$$

2. See section 5.1.
3. A possible explanation is that we are observing an effect of sequelhood. In the beginning of *Through the looking-glass*, the definite article is overrepresented and the indefinite article is underrepresented. This suggests that, especially in the beginning of this sequel to *Alice in Wonderland*, Carroll tends to appeal to common knowledge, knowledge shared by him and his readers.
4. $V(0, N)$ denotes the number of words in the population that do not occur in the sample. As $V(1, N)$ is much larger than $V(2, N)$, and since in turn $V(2, N)$ is much larger than $V(3, N)$, etc., it is likely that $V(0, N)$ is much larger than $V(1, N)$.
5. The larger text, which follows from the fact that the curve with the highest number of hapax legomena also realizes the highest frequency m . As will become clear in section 2.4, in most textual data sets, the number of hapax legomena is an increasing function of the sample size.
6. This expression is equivalent to the inequality (1.3).
7. The discrete nature of word frequencies introduces the strange striations at the right-hand edge of the plot: all hapax legomena have frequency 1, all dis legomena frequency 2, etc., instead of decreasing frequencies with increasing rank z . The curvature of the error function at the left hand side highlights the failure of the zeta function to capture the details of the rank-frequency relation for the highest-frequency words. Note that

with the large number of observations, and hence the high number of degrees of freedom, a significant correlation is not necessarily an interesting and important correlation, especially if the error function reveals a non-random pattern.

8. See Figure 1.6:

$$V(m, N) = \frac{C'}{m^{1/a}} - b - \left(\frac{C'}{(m+1)^{1/a}} - b \right) = C' \left(\frac{1}{m^{1/a}} - \frac{1}{(m+1)^{1/a}} \right),$$

with $C' = C^{1/a}$.

9. The high-frequency words in the left-hand side of the rank-frequency plot have fairly stable relative frequencies. Increasing the sample size does not affect the Zipfian curve very much. On the other hand, as we increase the sample, large numbers of new types appear at the right-hand side of the plot, increasing the importance given by the least squares regression to the highest ranks. Consequently, the regression line is 'pulled downwards' by the lowest-frequency words, resulting in a greater (negative) value of a and in a larger intercept.
10. $K = 10000 \frac{\sum_i f(i, N)^2 - N}{N^2}$.
11. See chapter 5, section 5.1.
12. For a small finite population vocabulary size, we may expect the parameters of the lognormal model to become stable.
13. The type-token ratio is the inverse mean frequency. As such, it is equally dependent on the sample size as the mean frequency itself. It is useless for the comparison of texts of different lengths.

CHAPTER 2

1. Since

$$\lim_{N \rightarrow \infty} E[K] = \lim_{N \rightarrow \infty} \frac{10000}{N^2} \left[\sum_m m^2 E[V(m, N)] - N \right],$$

we focus on $\sum_{m=1}^{\infty} m^2 E[V(m, N)]$. First note that

$$m \binom{N}{m} = N \binom{N-1}{m-1}.$$

We use this relation twice to simplify this expression:

$$\begin{aligned}
& \sum_{m=1}^{\infty} m^2 \mathbb{E}[V(m, N)] = \\
&= \sum_{m=1}^{\infty} m^2 \sum_{i=1}^S \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m} \\
&\stackrel{n=m-1}{=} \sum_{i=1}^S N \pi_i \sum_{n=0}^{\infty} (1+n) \binom{N-1}{n} \pi_i^n (1 - \pi_i)^{N-1-n} \\
&= \sum_{i=1}^S N \pi_i \left[1 + \sum_{n=1}^{\infty} n \binom{N-1}{n} \pi_i^n (1 - \pi_i)^{N-1-n} \right] \\
&\stackrel{r=n-1}{=} \sum_{i=1}^S N \pi_i \left[1 + (N-1) \pi_i \sum_{r=0}^{\infty} \binom{N-2}{r} \pi_i^r (1 - \pi_i)^{N-2-r} \right] \\
&= N + (N-1)N \sum_{i=1}^S \pi_i^2.
\end{aligned}$$

Hence

$$\begin{aligned}
\lim_{N \rightarrow \infty} \mathbb{E}[K] &= \lim_{N \rightarrow \infty} \frac{10000}{N^2} N(N-1) \sum_{i=1}^S \pi_i^2 \\
&= 10000 \sum_{i=1}^S \pi_i^2,
\end{aligned}$$

which was to be proved.

2. $\mathcal{E} = \mathbb{E}[X]$, with X the information load $-\log(\pi)$.

3. Let

$$X_i = \begin{cases} 1 & \text{if a token of } \omega_i \text{ is sampled, and} \\ 0 & \text{otherwise.} \end{cases}$$

Since $\Pr(X_i = 1) = \pi_i$, we have

$$\begin{aligned}
\mathbb{E}[f(i, N)] &= \mathbb{E}\left[\sum_{j=1}^N X_j\right] \\
&= \sum_{j=1}^N \mathbb{E}[X_j] \\
&= N\pi_i.
\end{aligned}$$

Similarly, since $\mathbb{E}[X_i^2] = \mathbb{E}[X_i]$,

$$\begin{aligned}
\text{VAR}[f(i, N)] &= \text{VAR}\left[\sum_{j=1}^N X_j\right] \\
&= \sum_{j=1}^N \text{VAR}[X_j]
\end{aligned}$$

$$\begin{aligned}
 &= \sum_{j=1}^N \{\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2\} \\
 &= \sum_{j=1}^N (p - p^2) \\
 &= Np(1 - p).
 \end{aligned}$$

4. The frequency of *the* is a Binomially $(100, 0.07)$ -distributed random variable. Using the Poisson approximation, we have

$$\Pr(f(\text{the}, 100) = 7) = \frac{7^7}{7!} e^{-7} = 0.149.$$

The exact value using the binomial distribution itself is 0.154. In R or Splus, this exact value can be obtained by `pbinom(7, 100, 0.07) - pbinom(6, 100, 0.07)`.

5. Since $\lim_{N \rightarrow \infty} \mathbb{E}[V(N)] = 6$, the mean token frequency is $N/6$.
6. Let $X = f(i, N)$ denote an $(N, N_0 - N, f(i, N_0))$ -distributed hypergeometric random variable. Since $\text{VAR}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and since $\mathbb{E}[X] = f(i, N_0)N/N_0$, we first consider $\mathbb{E}[X^2]$. Let x denote the frequency of ω_i in a given sample of N words, and let f denote its frequency in the text with N_0 words. We then have:

$$\begin{aligned}
 \mathbb{E}[X^2] &= \sum_{x=0}^f x^2 \frac{\binom{N}{x} \binom{N_0-N}{f-x}}{\binom{N_0}{f}} \\
 &= \sum_{x=0}^f x(x-1) \frac{N(N-1)(N-2)!}{(N-x)!x(x-1)(x-2)!} \frac{\binom{N_0-N}{f-x}}{\binom{N_0}{f}} \\
 &\quad + \sum_{x=0}^f x \frac{\binom{N}{x} \binom{N_0-N}{f-x}}{\binom{N}{f}} \\
 &= \sum_{x=2}^f N(N-1) \frac{\binom{N-2}{x-2} \binom{N_0-N}{f-x}}{\frac{N_0(N_0-1)}{f(f-1)} \binom{N_0-2}{f-2}} + \mathbb{E}[X] \\
 &= \frac{f(f-1)N(N-1)}{N_0(N_0-1)} \sum_{x=0}^f \frac{\binom{N-2}{x} \binom{N_0-N}{f-2-x}}{\binom{N_0-2}{f-2}} + \mathbb{E}[X] \\
 &= \frac{f(f-1)N(N-1)}{N_0(N_0-1)} + \mathbb{E}[X].
 \end{aligned}$$

Hence

$$\begin{aligned}
 \text{VAR}[X] &= \frac{f(f-1)N(N-1)}{N_0(N_0-1)} + f \frac{N}{N_0} - \left(f \frac{N}{N_0}\right)^2 \\
 &= \frac{fN}{N_0(N_0-1)} [(f-1)(N-1) + (1 - fN/N_0)(N_0-1)]
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{fN}{N_0(N_0 - 1)}(1 - N/N_0)(N_0 - f) \\
 &= f \frac{N}{N_0} \left(1 - \frac{N}{N_0}\right) \frac{N_0 - f}{N_0 - 1}.
 \end{aligned}$$

7. The relative adjustment $(m - m^*)/m$ is greatest for the low-frequency words. For instance, consider a sample for which Zipf's law is valid at the sample size Z . We then have

$$\begin{aligned}
 m^* &= \frac{(m+1)\mathbb{E}[V(m+1, Z)]}{m\mathbb{E}[V(m, Z)]} \\
 &= \frac{(m+1)\mathbb{E}[V(Z)]m(m+1)}{(m+1)(m+2)\mathbb{E}[V(Z)]} \\
 &= \frac{m(m+1)}{m+2}.
 \end{aligned}$$

Thus, the adjusted frequency of a hapax legomenon at Z is 0.667, hence the relative adjustment is 0.333. For a higher-frequency word with, say, $f(i, Z) = 1000$, the relative adjustment is $(1000 - 999.002)/1000 = 0.001$.

8. Without smoothing, m^* can become negative.

9. Using the method and notation of 2.6, we have

$$\begin{aligned}
 \mathbb{E}[V_{\cdot, N_0}(N)] &= \int_0^\infty (1 - e^{-N_0\pi})(1 - e^{N\pi})dG(\pi) \\
 &= \int_0^\infty 1 - e^{-N_0\pi} - e^{-N\pi} + e^{-(N_0+N)\pi}dG(\pi) \\
 &= S - \mathbb{E}[V(0, N_0)] - \mathbb{E}[V(0, N)] + \mathbb{E}[V(0, N_0 + N)].
 \end{aligned}$$

10.

$$\begin{aligned}
 \mathbb{E}[V_{k, N_0}(m, N)] &= \int_0^\infty \frac{(N_0\pi)^k}{k!} e^{-N_0\pi} \frac{(N\pi)^m}{m!} e^{-N\pi} dG(\pi) \\
 &= \int_0^\infty \pi^{k+m} e^{-(N+N_0)\pi} \frac{N_0^k N^m}{(k+m)!} \binom{k+m}{m} dG(\pi) \\
 &= \mathbb{E}[V(k+m, N+N_0)] \binom{k+m}{m} \frac{N_0^k N^m}{(N+N_0)^{k+m}}.
 \end{aligned}$$

11. Let $G_{TE}(\lambda)$ denote Thisted and Efron's empirical cumulative distribution function. Then the structural type distribution is

$$G(\lambda) = \frac{G_{TE}(\lambda)}{S}.$$

12. If we standardize with respect to a fixed sample size N_0 , we can revised Carroll's SFI as follows:

$$\text{SFI}' = m^*/N_0,$$

the expected Good-Turing adjusted relative frequency at the standard sample size N_0 , which can be logarithmically transformed if so desired.

CHAPTER 3

1. For the Yule-Simon model, $\alpha(1, Z) = \frac{1}{1+\beta}$, consequently $E[V(Z)] = (1 + \beta)E[V(1, Z)]$ and hence

$$\frac{E[V(Z)]}{Z} = (1 + \beta) \frac{E[V(1, Z)]}{Z},$$

which was to be proved. For Zipf's law, $\beta = 1$, hence the type-token ratio for this model is twice the growth rate at the sample size Z .

2. The expectation of a lognormally distributed random variable X is

$$\begin{aligned} E[X] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{x} e^{-\frac{1}{2\sigma^2}[\log(x)-\mu]^2} x dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[\log(x)-\mu]^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[y-\mu]^2} e^y dy \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[(y-(\mu+\sigma^2))^2 - (2\mu\sigma^2 + \sigma^4)]} dy \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[y-(\mu+\sigma^2)]^2} e^{\mu+\frac{1}{2}\sigma^2} dy \\ &= e^{\mu+\frac{1}{2}\sigma^2} \end{aligned}$$

3. Given the structural type distribution

$$G(\pi) = S \int_{\pi}^{\infty} \frac{\lambda e^{-\lambda x} (\lambda x)^{s-1}}{\Gamma(s)} dx,$$

we first observe that

$$S = \frac{1}{E[X]} = \frac{\lambda}{s}.$$

Therefore,

$$\begin{aligned} E[V(N)] &= \int_0^{\infty} (1 - e^{-N\pi}) dG(\pi) \\ &= \int_0^{\infty} (1 - e^{-N\pi}) \frac{\lambda}{s} \frac{\lambda e^{-\lambda\pi} (\lambda\pi)^{s-1}}{\Gamma(s)} d\pi \\ &= \frac{\lambda}{s} \left(1 - \int_0^{\infty} \frac{\lambda e^{-(\lambda+N)\pi} (\lambda\pi)^{s-1}}{\Gamma(s)} d\pi \right) \\ &= \frac{\lambda}{s} \left[\left(\frac{\lambda}{\lambda+N} \right)^s \int_0^{\infty} \frac{(\lambda+N)e^{-\lambda+N\pi} (\lambda+N)^{s-1} \pi^{s-1}}{\Gamma(s)} d\pi \right] \\ &= \frac{\lambda}{s} \left(1 - \left(\frac{\lambda}{\lambda+N} \right)^s \right). \end{aligned}$$

Similarly,

$$\begin{aligned}
 E[V(m, N)] &= \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} \frac{\lambda}{s} \frac{\lambda e^{-\lambda\pi} (\lambda\pi)^{s-1}}{\Gamma(s)} d\pi \\
 &= \frac{\lambda}{s} \int_0^\infty \frac{N^m}{m!} \frac{\lambda e^{-(N+\lambda)\pi} \lambda^{s-1} \pi^{m+s-1}}{\Gamma(m+s)} d\pi \\
 &= \frac{\lambda \Gamma(m+s) N^m \lambda^s}{s (N+\lambda)^{m+s} m!} \cdot \\
 &\quad \int_0^\infty \frac{(N+\lambda)e^{-(N+\lambda)\pi} (N+\lambda)^{m+s-1} \pi^{m+s-1}}{\Gamma(m+s)} d\pi \\
 &= \frac{\lambda \Gamma(m+s)}{s \Gamma(m+1)} \left(\frac{N}{N+\lambda} \right)^m \left(\frac{\lambda}{N+\lambda} \right)^s.
 \end{aligned}$$

At N_1^* , the sample size where $E[V(1, N)]$ reaches its maximum, we have

$$\frac{1}{N} E[V(1, N)] - \frac{2}{N} E[V(2, N)] = 0,$$

hence

$$\frac{\lambda \Gamma(1+s)}{s \Gamma(2)} \left(\frac{N}{N+\lambda} \right) \left(\frac{\lambda}{N+\lambda} \right)^s = \frac{\lambda \Gamma(2+s)}{s \Gamma(3+1)} \left(\frac{N}{N+\lambda} \right)^2 \left(\frac{\lambda}{N+\lambda} \right)^s,$$

which, after some algebraic manipulation, leads to

$$N_1^* = \frac{\lambda}{s} = S.$$

To see the severe restrictions this last equality imposes, it is useful to compare it with Sichel's generalized inverse Gauss-Poisson model, which estimates \hat{S} at 8998 types for *Alice's adventures in wonderland*, for which $N = 26505$. But the present Gamma model would have to assume, for the same S , that the number of hapaxes reaches its maximum for $N = 8998$, contrary to fact. Thus, this model is appropriate only for distributions in the very late LNRE zone.

4.

$$\begin{aligned}
 \sum_{m=1}^{\infty} \int_0^{\infty} \frac{[\log(1+x)]^{\gamma-1} x^{\alpha}}{(1+x)^{m+1} (1+x)^{\beta}} dx &= \\
 &= \int_0^{\infty} \frac{[\log(1+x)]^{\gamma-1} x^{\alpha}}{(1+x)^{\beta+2}} \sum_{m=1}^{\infty} \frac{1}{(1+x)^{m-1}} dx \\
 &= \int_0^{\infty} \frac{[\log(1+x)]^{\gamma-1} x^{\alpha-1}}{(1+x)^{\beta+1}} dx.
 \end{aligned}$$

For the last step, we have made use of the equality $\sum_{i=0}^{\infty} a^i = 1/(1-a)$:

$$\begin{aligned}\sum_{m=1}^{\infty} \frac{1}{(1+x)^{m-1}} &= \sum_{n=0}^{\infty} \left(\frac{1}{1+x}\right)^n \\ &= \frac{1}{1 - \frac{1}{1+x}} \\ &= \frac{1+x}{x}.\end{aligned}$$

CHAPTER 5

1. The covariance of the spectrum elements in a mixture model can be expressed in terms of the sum of the covariances of the mixture elements:

$$\text{COV}[V(m, N), V(k, N)] =$$

$$\begin{aligned}&= I_{[m=k]} \{E[V(m, pN)|\{Z_1, a_1, b_1\}] + E[V(m, (1-p)N)|\{Z_2, a_2, b_2\}]\} \\&\quad - \binom{m+k}{m} \frac{1}{2^{m+k}} \cdot \\&\quad \{E[V(m+k, 2pN)|\{Z_1, a_1, b_1\}] + E[V(m+k, 2(1-p)N)|\{Z_2, a_2, b_2\}]\} \\&= I_{[m=k]} E[V(m, pN)|\{Z_1, a_1, b_1\}] \\&\quad - \binom{m+k}{m} \frac{1}{2^{m+k}} E[V(m+k, 2pN)|\{Z_1, a_1, b_1\}] \\&\quad + I_{[m=k]} E[V(m, (1-p)N)|\{Z_2, a_2, b_2\}] \\&\quad - \binom{m+k}{m} \frac{1}{2^{m+k}} E[V(m+k, 2(1-p)N)|\{Z_2, a_2, b_2\}] \\&= \text{COV}[V(m, pN), V(k, pN)|\{Z_1, a_1, b_1\}] + \\&\quad \text{COV}[V(m, (1-p)N), V(k, (1-p)N)|\{Z_2, a_2, b_2\}].\end{aligned}$$

2. A proof of Theorem A for the generalized inverse Gauss-Poisson model:

$$\begin{aligned}E[V(m, pN)|\{Z, b, \gamma\}] &= \frac{2Z}{bK_{\gamma+1}(b)(1+(pN)/Z)^{\gamma/2}} \frac{\left(\frac{b(pN)}{2Z\sqrt{1+(pN)/Z}}\right)^m}{m!} \cdot \\&\quad \cdot K_{m+\gamma}(b\sqrt{1+(pN)/Z}) \\&= p \frac{2\frac{Z}{p}}{bK_{\gamma+1}(b)(1+N/\frac{Z}{p})^{\gamma/2}} \frac{\left(\frac{bN}{2\frac{Z}{p}\sqrt{1+N/\frac{Z}{p}}}\right)^m}{m!} \cdot \\&\quad \cdot K_{m+\gamma}(b\sqrt{1+N/\frac{Z}{p}}) \\&= p E[V(m, N)|\{\frac{Z}{p}, b, \gamma\}].\end{aligned}$$

-
- 3. For the lognormal model, $S = e^{\frac{1}{2}\sigma^2 - \mu}$. As $\mu = \log 1/Z$, we have that $S = Ze^{\frac{1}{2}\sigma^2}$, from which theorem C follows immediately.
 - 4. Singulars are not always conceptually prior to plurals. There are many nouns for which the plural form denotes the natural number of occurrence (e.g., *eyes*, *feet*, *leaves*, contrast nouns such as *noses*, *mouths*, *dairy*), for which the singular denotes the natural number of occurrence, see Tiersma (1982) and Baayen, Dijkstra, and Schreuder (1997). Some nouns even occur without a corresponding singular (*scissors*, *trousers*). It may be necessary, therefore, to model the distribution of plural forms itself as a mixture distribution of two different kinds of plural forms: conceptually prior plural forms (*scissors*, *eyes*), and conceptually secondary plural forms (*noses*, *dairy*).

Appendix C

Software

This appendix summarizes a series of programs for analyzing frequency spectra using LNRE models. These programs are available under the GNU public license in the hope that they will be useful but *without any warranty*; without even the implied warranty of *merchantability* or *fitness for a particular purpose*.

The programs documented below are written in C and can be run from the command line in LINUX and UNIX systems. Preprocessing is done by `spectrum`, which takes a text as input, and outputs various files including a word frequency list, a file with the developmental profiles of the text, and a frequency spectrum. These files, in various combinations, form the input for the LNRE analyses. In case only a word frequency list is available, `wlf2spc` can be used to derive the frequency spectrum that is the point of departure for all LNRE models. The adjusted models described in Chapter 5, however, cannot be applied in this case, as they crucially depend on the empirical developmental structure of the text.

All output files, except summary files, are in table format with column headers, so that they can be easily serve as input to the statistical programming environments R and Splus.

A graphical user interface to most C programs is provided by the tcl/tk program `lexstats`, designed for LINUX and LINUX-like systems such as UNIX. LEXSTATS facilitates parameter estimation, notably so for mixture models, for which as yet no reliable automated procedures are available. It provides the possibility to calculate the multivariate X^2 and the MSE, and it also allows the user to inspect the goodness of fit graphically by means of a plotting routine for the frequency spectrum and the developmental spectrum of the vocabulary size and the spectrum elements. Not accessible from LEXSTATS are the programs `mcprofile` and `mcdisp`, which implement Monte Carlo methods for the observed developmental profile of a text and its lexical dispersion characteristics as described in sections 1.4 and 5.1.1.

In what follows, three tables first summarize the programs by group, the output file types by group, and the file type extensions. Next, the input and output specifications of each program are described in alphabetical order.

INDEX OF PROGRAMS BY GROUP

BASIC ANALYSES

spectrum	calculate spectrum and word frequency list from text
wf12spc	build frequency spectrum from a word frequency list

MONTE CARLO METHODS

mcdisp	dispersion analysis
mcprofile	calculate empirical and theoretical developmental profiles

NON-PARAMETRIC INTERPOLATION

binomint	binomial interpolation
labhub	partition-adjusted binomial interpolation (Labbe-Hubert)

SPECTRUM SMOOTHING

spectfit	Naranan-Balasubrahmanyam fit and Good-Turing estimation
----------	---------------------------------------------------------

LNRE MODELS

standard models

lnreZipf	Zipf
lnreYuSi	Yule-Simon
lnreCarr	lognormal

lnreSich	generalized inverse Gauss-Poisson ($\gamma = -0.5$)
lnreSgam	generalized inverse Gauss-Poisson (γ free)
<i>partition-adjusted models</i>	
ad2Zipf	partition-adjusted Zipf's law
ad2YuSi	partition-adjusted Yule-Simon model
ad2Carr	partition-adjusted lognormal model
ad2Sich	partition-adjusted inverse Gauss-Poisson with $\gamma = -0.5$
ad2Sgam	partition-adjusted inverse Gauss-Poisson with γ free
<i>parameter-adjusted models</i>	
adjZipf	parameter-adjusted Zipf's law
adjSich	parameter-adjusted inverse Gauss-Poisson with $\gamma = -0.5$
<i>mixture models</i>	
lexstats	any combination of two standard LNRE models
EVALUATING GOODNESS OF FIT	
lnreChi2	multivariate chi-squared test and mean squared error

INDEX OF OUTPUT FILE TYPES BY GROUP

BASIC ANALYSES

.obs	empirical developmental profile	spectrum
.spc	empirical frequency spectrum	spectrum
.sum	summary of statistics	wfl2spc
.wfl	empirical word frequency list	spectrum
.zrk	empirical rank-frequency list	spectrum
.zvc	text in vector form with Zipf ranks	spectrum

MONTE CARLO METHODS

.mch	upper 95% confidence interval	mcprofile
.mcl	lower 95% confidence interval	mcprofile
.mcm	mean profile	mcprofile
.mco	observed empirical profile	mcprofile
.mcd	observed and expected dispersion	mcdisp
.fik	dispersion frequency table	mcdisp

NON-PARAMETRIC INTERPOLATION

.bin	binomial interpolation profiles	binomint
.lhu	binomial interpolation profiles with partition-based adjustment	labhub

SPECTRUM SMOOTHING

N.spc	expected spectrum and Good-Turing estimates	spectfit
N.sum	summary statistics and MSE	spectfit

LNRE MODELS

_xX.spc	observed and expected spectrum	all models
_xX.fsp	expected spectrum for extended range	standard models
_xX.sp2	expected spectrum at $2N_0$	all models
_xX.ev2	$V(N_0)$, $E[V(N_0)]$, $E[V(2N_0)]$	all models
_xX.int	interpolated spectrum and $E[V(N)]$	all models
_xX.ext	extrapolated spectrum and $E[V(N)]$	all models
_xX.fit	fitted data points	adjZipf
_xX.sta	coefficients of the fit	adjSich
_xX.sum	summary statistics and parameters	adjZipf

SIGNIFICANCE TESTING

_xX.chi	goodness-of-fit statistics	lnreChi2
_xX.cov	covariance matrix for LNRE model	lnreChi2

FILE EXTENSIONS OF LNRE MODELS

SPECTRUM SMOOTHING:

N.xxx Naranan-Balasubrahmanyam

STANDARD LNRE MODELS:

Z.xxx Zipf

Y.xxx Yule-Simon

C.xxx lognormal

S.xxx generalized inverse Gauss-Poisson ($\gamma = -0.5$)

G.xxx generalized inverse Gauss-Poisson (γ free)

PARAMETER ADJUSTED LNRE MODELS:

aZ.xxx Zipf

aS.xxx generalized inverse Gauss-Poisson ($\gamma = -0.5$)

PARTITION ADJUSTED LNRE MODELS:

bZ.xxx Zipf

bY.xxx Yule-Simon

bC.xxx lognormal

bS.xxx generalized inverse Gauss-Poisson ($\gamma = -0.5$)

bG.xxx generalized inverse Gauss-Poisson (γ free)

MIXTURE MODELS:

XY.xxx X: base component

Y: complement component

ad2Carr	partition-adjusted lognormal
----------------	------------------------------

ad2Carr (-h -mW -kX -KY -EZ -H) *text.spc*

Partition-adjusted lognormal model. Estimates the partition parameter p by minimizing the mean squared error. Requires the availability of the standard lognormal summary file (output of *lnreCarr* run with cost function C_1) and the developmental profile file (output of *spectrum*) in the current directory. For the details of the output files, see *lnreCarr*.

INPUT

text.spc: the frequency spectrum

text.obs should be available in the working directory

text.C.sum should also be available in the working directory

OPTIONS

-h: display on-line help

-mW: number of frequency ranks m in the fit is set to W (default: 15)

-kX: number of chunks for interpolation is set to X (default: 20)

-KY: number of chunks for extrapolation is set to Y (default: 20)

-EZ: extrapolation sample size is set to Z (default: $2N_0$)

-H: input file has no header (default: with header)

OUTPUT

text.bC.spc: expected frequency spectrum

text.bC.sp2: expected frequency spectrum at $2N_0$

text.bC.ev2: $V(N_0)$, $E[V(N_0)]$ and $E[V(2N_0)]$

text.bC.int: interpolation statistics

text.bC.ext: extrapolation statistics

text.bC.sum: summary of main statistics

ad2Sgam

partition-adjusted generalized inverse Gauss-Poisson

```
ad2Sgam (-h -mW -kX -KY -EZ -H -L -v) text.spc
```

Partition-adjusted extended inverse Gauss-Poisson analysis with free γ . Estimates the partition parameter p by minimizing the mean squared error. Note that the observed developmental profile should be available in the working directory (output of spectrum). For a detailed description of the output files, see `lnreSgam`.

Since there are no closed-form expressions for estimating the parameters of the complete inverse Gauss-Poisson model with free γ , the program asks the user whether to attempt downhill simplex minimization provide interactive user-guided minimization. In the case downhill simplex minimization is selected, the program calculates $E[V(N)]$ and $E[V(1, N)]$ for the simpler model with $\gamma = -0.5$. The user is offered the choice between using these parameters as starting point for minimization, or to specify another starting point. The program uses cost function C_1 in the simplex subroutine.

In the case of interactive user-guided minimization, the program prompts for Z , b and γ , and calculates the expected vocabulary size and the expected number of hapax legomena and dis legomena for the specified parameters. The user is offered the option of adjusting the parameter estimates before proceeding to the main analyses.

INPUT

`text.spc`: the frequency spectrum

`text.obs` should be available in the working directory

OPTIONS

`-h`: display on-line help

`-mW`: number of frequency ranks m in the fit is set to W (default: 15)

`-kX`: number of chunks for interpolation is set to X (default: 20)

`-KY`: number of chunks for extrapolation is set to Y (default: 20)

`-EZ`: extrapolation sample size is set to Z (default: $2N_0$)

`-H`: input file has no header (default: with header)

OUTPUT

`text.bG.spc`: expected frequency spectrum

`text_bG.sp2`: expected frequency spectrum at $2N_0$
`text_bG.ev2`: $V(N_0)$, $E[V(N_0)]$ and $E[V(2N_0)]$
`text_bG.int`: interpolation statistics
`text_bG.ext`: extrapolation statistics
`text_bG.sum`: summary of main statistics

TECHNICAL DETAILS

See `lnreSgam`.

ad2Sich partition-adjusted generalized inverse Gauss-Poisson ($\gamma = -0.5$)

```
ad2Sich (-h -mW -kX -KY -EZ -H) text.spc
```

Partition-adjusted inverse Gauss-Poisson model with γ fixed at -0.5 a priori. Uses cost function C_1 and estimates the partition parameter p by minimizing the mean squared error. For a detailed description of the standard LNRE output files, see lnreSich.

INPUT

`text.spc`: the frequency spectrum

`text.obs` should be available in the working directory

OPTIONS

`-h`: display on-line help

`-mW`: number of frequency ranks m in the fit is set to W (default: 15)

`-kX`: number of chunks for interpolation is set to X (default: 20)

`-KY`: number of chunks for extrapolation is set to Y (default: 20)

`-EZ`: extrapolation sample size is set to Z (default: $2N_0$)

`-H`: input file has no header (default: with header)

OUTPUT

`text_bs.spc`: expected frequency spectrum

`text_bs.sp2`: expected frequency spectrum at $2N_0$

`text_bs.ev2`: $V(N_0)$, $E[V(N_0)]$ and $E[V(2N_0)]$

`text_bs.sum`: summary of main statistics, including the specialization parameter p and the corresponding MSE

`text_bs.int`: interpolation statistics

`text_bs.ext`: extrapolation statistics

ad2YuSi**partition-adjusted Yule-Simon**

ad2YuSi (-h -mW -kX -KY -EZ -H) text.spc

Partition-adjusted Yule-Simon model. Requires the developmental profile file (output of spectrum) and the standard Yule-Simon summary file (output of lnreYuSi run with cost function \mathcal{C}_1) to be available in the current directory. Estimates the partition parameter p by minimizing the mean squared error. For a detailed description of the standard LNRE output files, see lnreYuSi.

INPUT**text.spc:** the frequency spectrum**text.obs** should be available in the working directory**text_Y.sum** should also be available in the working directory**OPTIONS****-h:** display on-line help**-mW:** number of frequency ranks m in the fit is set to W (default: 15)**-kX:** number of chunks for interpolation is set to X (default: 20)**-KY:** number of chunks for extrapolation is set to Y (default: 20)**-EZ:** extrapolation sample size is set to Z (default: $2N_0$)**-H:** input file has no header (default: with header)**OUTPUT****text_bY.spc:** expected frequency spectrum**text_bY.sp2:** expected frequency spectrum at $2N_0$ **text_bY.ev2:** $V(N_0)$, $E[V(N_0)]$ and $E[V(2N_0)]$ **text_bY.int:** interpolation statistics**text_bY.ext:** extrapolation statistics**text_bY.sum:** summary of main statistics

ad2Zipf**partition-adjusted Zipf**

```
ad2Zipf (-h -mW -kX -KY -EZ -H) text.spc
```

Partition-adjusted extended Zipf's law. Estimates the partition parameter p by minimizing the mean squared error. Estimates Z by requiring that $V(N) = E[V(N)]$. For a detailed description of the standard LNRE output files, see `lnreZipf`.

INPUT

`text.spc`: the frequency spectrum

`text.obs` should be available in the working directory

OPTIONS

`-h`: display on-line help

`-mW`: number of frequency ranks m in the fit is set to W (default: 15)

`-kX`: number of chunks for interpolation is set to X (default: 20)

`-KY`: number of chunks for extrapolation is set to Y (default: 20)

`-EZ`: extrapolation sample size is set to Z (default: $2N_0$)

`-H`: input file has no header (default: with header)

OUTPUT

`text_bZ.spc`: expected frequency spectrum

`text_bZ.sp2`: expected frequency spectrum at $2N_0$

`text_bZ.ev2`: $V(N_0)$, $E[V(N_0)]$ and $E[V(2N_0)]$

`text_bZ.int`: interpolation statistics

`text_bZ.ext`: extrapolation statistics

`text_bZ.sum`: summary of main statistics

adjSich parameter-adjusted generalized inverse Gauss-Poisson ($\gamma = -0.5$)

```
adjSich (-h -mU -kV -KW -EX -H -fY -gZ) text.spc
```

Parameter-adjusted inverse Gauss-Poisson model with γ fixed at -0.5 a priori. Parameter estimation on the basis of cost function \mathcal{C}_1 . Considers four link functions (linear, logarithmic, exponential, and power) and selects the model that is optimal in the least squares sense for adjustment. Adjustment by hand of the selected model is possible by means of the parameters f and g . For a detailed description of the standard LNRE output files, see lnreSich.

INPUT

`text.spc`: the frequency spectrum;
`text.obs` should be available in the working directory.

OPTIONS

- `-h`: display on-line help
- `-mU`: number of frequency ranks m is set U (default: 15)
- `-kV`: number of chunks for interpolation is set to V (default: 20)
- `-KW`: number of chunks for extrapolation is set to W (default: 20)
- `-EX`: extrapolation sample size is set to X (default: $2N_0$)
- `-H`: input file has no header (default: with header)
- `-fY`: adjust intercept of link function by Y
- `-gZ`: adjust slope of link function by Z

OUTPUT

`text.aS.spc`: expected frequency spectrum
`text.aS.sp2`: expected frequency spectrum at $2N_0$
`text.aS.ev2`: $V(N_0)$, $E[V(N_0)]$ and $E[V(2N)]$
`text.aS.int`: interpolation statistics
`text.aS.ext`: extrapolation statistics
`text.aS.sum`: summary of main statistics
`text.aS.fit`: observed and expected developmental profiles of c

text_as_sta: statistics for the goodness-of-fit of the link function

t: the t-statistic

r: the Pearson correlation r

a: the intercept of the (transformed) linear fit

b: the slope of the (transformed) linear fit

obs: linear link $c(N) = a + bN$

log: logarithmic link $c(N) = a + b \log(N)$

exp: exponential link $c(N) = e^{a+bN}$

pow: power link $c(N) = e^{a+b \log(N)}$

adjZipf**parameter-adjusted Zipf**

adjZipf (-h -mU -kV -KW -EX -H -fY -gZ) text.spc

Parameter-adjusted extended Zipf's law. Considers four link functions (linear, logarithmic, exponential, and power) and selects the optimal model for adjustment. Adjustment by hand of the selected model parameters is possible by means of the parameters f and g . Z is estimated by requiring that $V(N) = E[V(N)]$. For a detailed description of the output files, see `lnreZipf`.

INPUT**text.spc:** the frequency spectrum**text.obs** should be available in the working directory.**OPTIONS****-h:** display on-line help**-mU:** number of frequency ranks m in the fit is set to U (default: 15)**-kV:** number of chunks for interpolation is set to V (default: 20)**-KW:** number of chunks for extrapolation is set to W (default: 20)**-EX:** extrapolation sample size is set to X (default: $2N_0$)**-H:** input file has no header (default: with header)**-fY:** adjust intercept of link function by Y **-gZ:** adjust slope of link function by Z **OUTPUT****text.aZ.spc:** expected frequency spectrum**text.aZ.sp2:** expected frequency spectrum at $2N_0$ **text.aZ.ev2:** $V(N_0)$, $E[V(N_0)]$ and $E[V(2N_0)]$ **text.aZ.int:** interpolation statistics**text.aZ.ext:** extrapolation statistics**text.aZ.sum:** summary of main statistics**text.aZ.fit:** observed and expected developmental profiles of Z

text_az.sta: statistics for the goodness-of-fit of the link function

t: the t-statistic t

r: the Pearson correlation r

a: the intercept of the (transformed) linear fit

b: the slope of the (transformed) linear fit

obs: linear link $Z(N) = a + bN$

log: logarithmic link $Z(N) = a + b \log(N)$

exp: exponential link $Z(N) = e^{a+bN}$

pow: power link $Z(N) = e^{a+b \log(N)}$

binomintbinomial interpolation

binomint (-h -mY -kZ -H) **text.spc**Binomial interpolation of $V(N)$ and $V(m, N)$ **INPUT****text.spc**: the frequency spectrum**OPTIONS****-h**: display on-line help**-mY**: number of frequency ranks m for interpolation is set to Y
(default: 5)**-kZ**: number of chunks for interpolation is set to Z (default: 20)**-H**: input file has no header (default: with header)**OUTPUT****text.bin**: interpolation data for $E[V(N)]$ (EV) and
 $E[V(m, N)]$ (EV1, EV2, EV3, ...).

labhub**partition-adjusted binomial interpolation**

```
labhub (-h -mX -kY -H -p -PZ) text.spc
```

Partition-adjusted binomial interpolation with estimation of the partition parameter p by means of minimization of the mean squared error (Hubert and Labb  , 1988).

INPUT

`text.spc`: the frequency spectrum

`text.obs` should be available in the working directory

OPTIONS

`-h`: display on-line help

`-mW`: number of frequency ranks m for interpolation is set to W
(default: 5)

`-kX`: number of chunks for interpolation is set to X (default: 20)

`-H`: input file has no header (default: with header)

`-pY`: required precision for estimating p is set to Y (default: 100, i.e. .01)

`-PZ`: fix p a priori to equal Z

OUTPUT

`text.lhu`: expected frequency spectrum

The optimal value for the specialization parameter p and the corresponding mean squared error are printed to standard output.

lnreCarr	lognormal
-----------------	-----------

lnreCarr (-h -mW -kX -KY -EZ -H -eR -ss -Nn -Vv) **text.spc**

The standard LNRE model based on the lognormal distribution. Since there are no closed-form expressions for estimating the parameters of the lognormal model, the program asks the user whether to use down-hill simplex minimization or to provide interactive user-guided minimization. In the case downhill simplex minimization is selected, the program calculates $E[V(N)]$ and $E[V(1, N)]$ for $Z = 200$ and $\sigma = 1.5$. The user is offered the choice between using these parameters as starting point for minimization, or to specify another starting point. By default, cost function C_1 is used. In order to use cost function $C_2(r)$, use the **-e** option.

INPUT

text.spc: frequency spectrum

OPTIONS

- h: display on-line help
- mW: number of ranks in fit is set to W (default: 15)
- kX: number of chunks for interpolation is set to X (default: 20)
- KY: number of chunks for extrapolation is set to Y (default: 20)
- EZ: extrapolation sample size is set to Z (default: $2N_0$)
- H: input files do not have a header (default: header is presupposed)
- eR: use cost function $C_2(r)$ with $r = R$
- ss: calculate only the expected spectrum for S ranks,
output on **text.C.fsp**
- Nn: force N to equal n (in case of a partial spectrum)
- Vv: force $V(N)$ to equal v (in case of a partial spectrum)

OUTPUT

text.C.spc: observed and expected frequency spectrum

m: m (frequency)

Vm: $V(m, N)$ (frequency at sample size N)

EVm: $E[V(m, N)]$ (expected frequency at sample size N)

`text_C.fsp`: expected frequency spectrum

`m`: m (frequency)

`EVm`: $E[V(m, N)]$ (expected frequency at sample size N)

`text_C.sp2`: expected frequency spectrum at $2N$

`m`: m (frequency)

`EVm2N`: $E[V(m, 2N)]$ (expected frequency at sample size $2N$)

`text_C.ev2`: vocabulary size statistics

`V`: $V(N)$ (observed vocabulary size at N)

`EV`: $E[V(N)]$ (expected vocabulary size at N)

`EV2N`: $E[V(2N)]$ (expected vocabulary size at $2N$)

`text_C.sum`: summary statistics and estimated parameters

`N`: N (number of tokens)

`V(N)`: $V(N)$ (observed number of types)

`E[V(N)]`: $E[V(N)]$ (expected number of types)

`V(1,N)`: $V(1, N)$ (observed number of hapax legomena)

`E[V(1,N)]`: $E[V(1, N)]$ (expected number of hapax legomena)

`Z`: $\hat{Z} = e^{-\hat{\mu}}$ (parameter)

`mean`: $\hat{\mu}$ (mean)

`stddev`: $\hat{\sigma}$ (standard deviation)

`S`: \hat{S} (population number of types)

`text_C.int`, `text_C.ext`: interpolation and extrapolation statistics

`N`: N (number of tokens)

`E[V(N)]`: $E[V(N)]$ (expected number of types)

`Alpha1`: $E[\alpha(1)] (E[V(1, N)]/E[V(N)])$

`EV1-5`, ... : $E[V(1 - 5, N)]$ (expected spectrum elements)

`GV`: $E[V(N + 1)] - E[V(N)]$ (token-unit growth rate)

TECHNICAL DETAILS

The integrals of the lognormal model are evaluated by means of Romberg integration for the interval [0.000001, 1000.0], using the subroutine `qromb` of Press et al. (1988). The downhill simplex minimization method is used for parameter estimation, using the subroutine `amoeba` of Press et al. (1988).

lnreChi2	multivariate χ^2 -test
-----------------	-----------------------------

lnreChi2 (-h -mZ) text_xx.spc

A multivariate chi-squared test for evaluating goodness of fit.

INPUT

text_xx.spc: expected frequency spectrum at N_0

text_xx.sp2: expected spectrum at $2N_0$

text_xx.ev2: expected vocabulary size at N_0 and $2N_0$

The *.sp2* and *.ev2* files are produced by any LNRE analysis. Only the *_xx.spc* input file requires specification on the command line, the other files should be available in the working directory.

OPTIONS

-h: display on-line help

-mZ: use the first *m* ranks only (default: use all *m* in *_X.spc*)

OUTPUT

text_xx.chi: goodness-of-fit statistics

text_xx.cov: the covariance matrix

lnreSgamgeneralized inverse Gauss-Poisson (γ free)

```
lnreSgam (-h -mW -kX -KY -EZ -H -eR -sS -Nn -Vv -S)
          text.spc
```

The standard LNRE model based on the generalized inverse Gauss-Poisson distribution. Since there are no closed-form expressions for estimating the parameters of the full generalized inverse Gauss-Poisson model with free γ , the program asks the user whether to use downhill simplex minimization or to provide interactive user-guided minimization. In the case downhill simplex minimization is selected, the program calculates $E[V(N)]$ and $E[V(1, N)]$ for the simpler model with $\gamma = -0.5$. The user is offered the choice between using these parameters as starting point for minimization, or to specify another starting point. By default, cost function C_1 is used, but cost function $C_2(r)$ can be selected as well using the $-e$ option.

INPUT

`text.spc`: frequency spectrum

OPTIONS

- `-h`: display on-line help
- `-mW`: number of ranks in fit is set to W (default: 15)
- `-kX`: number of chunks for interpolation is set to X (default: 20)
- `-KY`: number of chunks for extrapolation is set to Y (default: 20)
- `-EZ`: extrapolation sample size is set to Z (default: $2N_0$)
- `-H`: input files do not have a header (default: header is presupposed)
- `-eR`: use cost function $C_2(r)$ with $r = R$
- `-sS`: calculate only the expected spectrum for S ranks,
output on `text_G.fsp`
- `-Nn`: force N to equal n (in case of a partial spectrum)
- `-Vv`: force $V(N)$ to equal v (in case of a partial spectrum)
- `-S`: calculate Good-Turing estimates (output in `text_G.str`)

OUTPUT

`text_G.spc`: observed and expected frequency spectrum

m: m (frequency)

Vm: $V(m, N)$ (frequency at sample size N)

EVm: $E[V(m, N)]$ (expected frequency at sample size N)

text_G.fsp: expected frequency spectrum

m: m (frequency)

EVm: $E[V(m, N)]$ (expected frequency at sample size N)

text_G.sp2: expected frequency spectrum at $2N$

m: m (frequency)

EVm2N: $E[V(m, 2N)]$ (expected frequency at sample size $2N$)

text_G.ev2: vocabulary size statistics

V: $V(N)$ (observed vocabulary size at N)

EV: $E[V(N)]$ (expected vocabulary size at N)

EV2N: $E[V(2N)]$ (expected vocabulary size at $2N$)

text_G.int, text_G.ext: interpolation and extrapolation statistics

N: N (number of tokens)

E[V(N)]: $E[V(N)]$ (expected number of types)

Alpha1: $E[\alpha(1)]$ ($E[V(1, N)]/E[V(N)]$)

EV1-5, ...: $E[V(1 - 5, N)]$ (expected spectrum elements)

GV: $E[V(N + 1)] - E[V(N)]$ (token-unit growth rate)

text_G.sum: summary statistics and estimated parameters

N: N (number of tokens)

V(N): $V(N)$ (observed number of types)

E[V(N)]: $E[V(N)]$ (expected number of types)

V(1, N): $V(1, N)$ (observed number of hapax legomena)

E[V(1, N)]: $E[V(1, N)]$ (expected number of hapax legomena)

S: \hat{S} (population number of types)

b: \hat{b} (parameter)

c: \hat{c} (parameter)

z: $\hat{Z} = 1/\hat{c}$ (parameter)

gamma: $\hat{\gamma}$ (parameter)

text_G.str: Good-Turing estimates based on the GIGP fit

m: the frequency spectrum

mstar: the corresponding Good-Turing estimates

TECHNICAL DETAILS

The Bessel function $K_v(z)$ of real order v ,

$$K_v(z) = \frac{\pi}{2} \frac{I_{-v}(z) - I_v(z)}{\sin(v\pi)},$$

is itself defined in terms of the simpler function

$$I_v(z) = \sum_{n=0}^{\infty} \frac{(z/2)^{v+2n}}{n! \Gamma(v+n+1)},$$

which is calculated up to the point where two successive terms of the sum differ by less than 1.0e-9. The downhill simplex minimization method is used for parameter estimation, using the subroutine amoeba of Press et al. (1988).

lnreSichgeneralized inverse Gauss-Poisson ($\gamma = -0.5$)

lnreSich (-h -mW -kX -KY -EZ -H) **text.spc**

The standard LNRE model based on the generalized inverse Gauss-Poisson distribution with γ fixed at -0.5. Parameters are estimated using cost function C_1 as suggested in Sichel (1986).

INPUT**text.spc**: frequency spectrum**OPTIONS****-h**: display on-line help**-mW**: number of ranks in fit is set to W (default: 15)**-kX**: number of chunks for interpolation is set to X (default: 20)**-KY**: number of chunks for extrapolation is set to Y (default: 20)**-EZ**: extrapolation sample size is set to Z (default: $2N_0$)**-H**: input files do not have a header (default: with header)**OUTPUT****text_S.spc**: observed and expected frequency spectrum**m**: m (frequency)**Vm**: $V(m, N)$ (frequency at sample size N)**EVm**: $E[V(m, N)]$ (expected frequency at sample size N)**text_S.sp2**: expected frequency spectrum at $2N$ **m**: m (frequency)**EVm2N**: $E[V(m, 2N)]$ (expected frequency at sample size $2N$)**text_S.ev2**: vocabulary size statistics**V**: $V(N)$ (observed vocabulary size at N)**EV**: $E[V(N)]$ (expected vocabulary size at N)**EV2N**: $E[V(2N)]$ (expected vocabulary size at $2N$)**text_S.int**, **text_S.ext**: interpolation and extrapolation statistics**N**: N (number of tokens)**E[V(N)]**: $E[V(N)]$ (expected number of types)**Alpha1**: $E[\alpha(1)]$ ($E[V(1, N)]/E[V(N)]$)

EV1-5, . . . : $E[V(1 - 5, N)]$ (expected spectrum elements)
GV: $E[V(N + 1)] - E[V(N)]$ (token-unit growth rate)
text_S.sum: summary statistics and estimated parameters
N: N (number of tokens)
V(N): $V(N)$ (observed number of types)
 $E[V(N)]$: $E[V(N)]$ (expected number of types)
 $V(1, N)$: $V(1, N)$ (observed number of hapax legomena)
 $E[V(1, N)]$: $E[V(1, N)]$ (expected number of hapax legomena)
S: \hat{S} (population number of types)
b: \hat{b} (parameter)
c: \hat{c} (parameter)
z: $\hat{Z} = 1/\hat{c}$ (parameter)
gamma: $\hat{\gamma}$ (parameter, fixed a-priori at -0.5)

lnreYuSi**Yule-Simon**

```
lnreYuSi (-h -mW -kX -KY -EZ -H -eR -ss -Nn -Vn -P)
          text.spc
```

The standard Yule-Simon LNRE model. Since there are no closed-form expressions for estimating the parameters of the Yule-Simon model, the program asks the user whether to use downhill simplex minimization or to provide interactive user-guided minimization. In the case downhill simplex minimization is selected, the program asks for initial values for Z , β , and $V(Z)$. A reasonable starting point is (200,0.5,150). By default, cost function \mathcal{C}_1 is used, but cost function $\mathcal{C}_2(r)$ can be selected as well using the $-e$ option.

INPUT

`text.spc`: frequency spectrum

OPTIONS

`-h`: display on-line help

`-mW`: number of ranks in fit is set to W (default: 15)

`-kX`: number of chunks for interpolation is set to X (default: 20)

`-KY`: number of chunks for extrapolation is set to Y (default: 20)

`-EZ`: extrapolation sample size is set to Z (default: $2N_0$)

`-H`: input files do not have a header (default: with header)

`-eR`: use cost function $\mathcal{C}_2(r)$ with $r = R$

`-ss`: calculate only the expected spectrum for S ranks,

`output on text_Y.fsp`

`-Nn`: force N to equal n (in case of a partial spectrum)

`-Vv`: force $V(N)$ to equal v (in case of a partial spectrum)

`-P`: estimate $V(Z)$ from m^* instead of including $V(Z)$ as
third parameter

OUTPUT

`text_Y.spc`: observed and expected frequency spectrum

`m`: m (frequency)

`Vm`: $V(m, N)$ (frequency at sample size N)

EVm: $E[V(m, N)]$ (expected frequency at sample size N)
text_Y.fsp: expected frequency spectrum
 m: m (frequency)
 EVm: $E[V(m, N)]$ (expected frequency at sample size N)
text_Y.sp2: expected frequency spectrum at $2N$
 m: m (frequency)
 EVm2N: $E[V(m, 2N)]$ (expected frequency at sample size $2N$)
text_Y.ev2: vocabulary size statistics
 V: $V(N)$ (observed vocabulary size at N)
 EV: $E[V(N)]$ (expected vocabulary size at N)
 EV2N: $E[V(2N)]$ (expected vocabulary size at $2N$)
text_Y.int, text_Y.ext: interpolation and extrapolation statistics
 N: N (number of tokens)
 E[V(N)]: $E[V(N)]$ (expected number of types)
 Alpha1: $E[\alpha(1)]$ ($E[V(1, N)]/E[V(N)]$)
 EV1-5, . . . : $E[V(1 - 5, N)]$ (expected spectrum elements)
 GV: $E[V(N + 1)] - E[V(N)]$ (token-unit growth rate)
text_Y.sum: summary statistics and estimated parameters
 N: N (number of tokens)
 V(N): $V(N)$ (observed number of types)
 E[V(N)]: $E[V(N)]$ (expected number of types)
 V(1, N): $V(1, N)$ (observed number of hapax legomena)
 E[V(1, N)]: $E[V(1, N)]$ (expected number of hapax legomena)
 z: \hat{Z} (parameter)
 beta: $\hat{\beta}$ (parameter)
 vz: $\hat{V}(Z)$ (parameter)
 s: \hat{S} (population number of types)

TECHNICAL DETAILS

The integrals of the Yule-Simon model are evaluated by means of Romberg integration for the interval [0.000001, 10000.0], using the subroutine qromb of Press et al. (1988). The downhill simplex minimization method is used for parameter estimation, using the subroutine amoeba of Press et al. (1988).

lnreZipf

extended Zipf

```
lnreZipf (-h -mW -kX -KY -EZ -H -sS) text.spc
```

The standard basic Zipfian LNRE model. The parameter Z is estimated by requiring that $V(N) = E[V(N)]$.

INPUT

`text.spc`: frequency spectrum

OPTIONS

`-h`: display on-line help

`-mW`: number of ranks in fit is set to W (default: 15)

`-kX`: number of chunks for interpolation is set to X (default: 20)

`-KY`: number of chunks for extrapolation is set to Y (default: 20)

`-EZ`: extrapolation sample size is set to Z (default: $2N_0$)

`-H`: input files do not have a header (default: header is presupposed)

`-sS`: calculate only the expected spectrum for S ranks,

output on `text.Z.fsp`

OUTPUT

`text.Z.spc`: observed and expected frequency spectrum

`m`: m (frequency)

`Vm`: $V(m, N)$ (frequency at sample size N)

`EVm`: $E[V(m, N)]$ (expected frequency at sample size N)

`text.Z.fsp`: expected frequency spectrum

`m`: m (frequency)

`EVm`: $E[V(m, N)]$ (expected frequency at sample size N)

`text.Z.sp2`: expected frequency spectrum at $2N$

`m`: m (frequency)

`EVm2N`: $E[V(m, 2N)]$ (expected frequency at sample size $2N$)

`text.Z.ev2`: vocabulary size statistics

`v`: $V(N)$ (observed vocabulary size at N)

`EV`: $E[V(N)]$ (expected vocabulary size at N)

EV2N: $E[V(2N)]$ (expected vocabulary size at $2N$)
text_Z.int, text_Z.ext: interpolation and extrapolation statistics
N: N (number of tokens)
 $E[V(N)]$: $E[V(N)]$ (expected number of types)
Alpha1: $E[\alpha(1)]$ ($E[V(1, N)]/E[V(N)]$)
EV1-5, . . . : $E[V(1 - 5, N)]$ (expected spectrum elements)
GV: $E[V(N + 1)] - E[V(N)]$ (token-unit growth rate)
text_Z.sum: summary statistics and estimated parameters
N: N (number of tokens)
 $V(N)$: $V(N)$ (observed number of types)
 $E[V(N)]$: $E[V(N)]$ (expected number of types)
 $V(1, N)$: $V(1, N)$ (observed number of hapax legomena)
 $E[V(1, N)]$: $E[V(1, N)]$ (expected number of hapax legomena)
z: \hat{Z} (parameter)
s: \hat{S} (population number of types)

TECHNICAL DETAILS

The integrals of the extended Zipf model are evaluated by means of Romberg integration for the interval $[0.000001, 10000.0]$, using the subroutine `qromb` of Press et al. (1988). The downhill simplex minimization method is used for parameter estimation, using the subroutine `amoeba` of Press et al. (1988).

mcdisp

```
mcdisp (-kX -pY -sZ -H) text.zvc text.spc
```

Monte Carlo based dispersion analysis.

INPUT

text.zvc: text in vector format with Zipf ranks

text.spc: the corresponding frequency spectrum

OPTIONS

-kX: number of text chunks is set to X

-pY: number of permutation runs is set to Y

-sZ: seed for random generator is set to Z

-H: input files do not have the standard header

OUTPUT

text.mcd: list with for each word type:

z: the Zipf rank z

Frequency: $f(i, N)$

Obs: observed dispersion d_i

Exp: expected dispersion $E[d_i]$ using the binomial model

StDev: the corresponding standard deviation

Z: the corresponding Z -score

MCperc: proportion of simulation runs with dispersion $\leq d_i$.

text.fik: list of word types and their frequencies for each text chunk

TECHNICAL DETAILS

The maximum text length currently implemented equals 100000 word tokens, the maximum number of types 20000, and the maximum number of text chunks 100.

mprofilecalculate developmental profile statistics

```
mprofile (-kv -pw -H -ax -ty -cz -z) text.zvc
```

Monte Carlo based analysis of the developmental profile.

INPUT

`text.zvc`: text in vector format with Zipf ranks

OPTIONS

`-kv`: number of text chunks (default: 20)

`-pw`: number of permutation runs (default: 0)

`-ax`: Zipf rank of first word to be traced (default: 4)

`-ty`: Zipf rank of second word to be traced (default: 1)

`-cz`: confidence interval (default: 95)

`-z`: use local instead of global p^* for Zipf's law

`-H`: input files do not have a header

OUTPUT

`text.mcm`: table with Monte Carlo means

`text.mcl`: table with lower 95% Monte Carlo confidence interval

`text.mch`: table with upper 95% Monte Carlo confidence interval

`text.mco`: table with observed values

Each table lists the values of 19 measures (columns) for each of the specified text chunks (rows):

N: N (number of tokens)

K: K (Yule's characteristic constant)

D: D (Simpson's diversity index)

V: V (number of types)

V1: $V(1, N)$ (number of hapax legomena)

V2: $V(2, N)$ (number of dis legomena)

V3: $V(3, N)$ (number of tris legomena)

V4: $V(4, N)$ (number of types with frequency 4)

V5: $V(5, N)$ (number of types with frequency 5)

R: R (Guiraud's constant)
W: W (Brunet's constant)
S: S (Sichel's constant)
H: H (Honoré's constant)
C: C (Herdan's constant)
E: E (sample entropy)
1M: $\hat{\mu}$ (mean log frequency)
1St: $\hat{\sigma}$ (standard deviation of log frequency)
b: \hat{b} (parameter of Sichel's model)
c: \hat{c} (parameter of Sichel's model)
a1: $\alpha(1, N)$ (relative number of hapax legomena)
z: \hat{Z} (parameter of extended Zipf's law)
fa: frequency of first word with specified Zipf rank
fthe: frequency of second word with specified Zipf rank
sLmean: sample mean of lognormal model
sLstddev: sample standard deviation of lognormal model

TECHNICAL DETAILS

The maximum text length implemented equals 220000 word tokens, the maximum number of word types 20000, the maximum number of permutation runs 5000, and the maximum number of text chunks 40.

spectfit Naranan-Balasubrahmanyam smoother, Good-Turing estimation

```
spectfit (-h -mW -H -eR -n -v) text.spc
```

This program fits the Naranan-Balasubrahmanyam model and computes Good-Turing estimates.

INPUT

`text.spc`: frequency spectrum

OPTIONS

`-h`: display on-line help

`-mW`: number of ranks in fit is set to W (default: 15)

`-H`: input files do not have a header (default: with header)

`-eR`: use cost function $\mathcal{C}_2(r)$ with $r = R$

`-v`: include $|V(N) - E[V(N)]|$ in cost function

`-n`: include $|N - N_{fit}|$ in cost function

OUTPUT

`text_N.spc`: observed and expected frequency spectrum

`m`: m (frequency)

`Vm`: $V(m, N)$ (frequency at sample size N)

`SVm`: $V_r(m, N)$ (real-valued spectrum)

`EVm`: $E[V(m, N)]$ (expected frequency at sample size N)

`mStar`: m^* (Good-Turing estimate using $E[V(m, N)]$)

`mStarRaw`: m^* (Good-Turing estimate using $V_r(m, N)$)

`StdevMstar`: standard deviation of m^*

`text_N.sum`: summary statistics and estimated parameters

`N`: N (number of tokens)

`V(N)`: $V(N)$ (observed number of types)

`E[V(N)]`: $E[V(N)]$ (expected number of types)

`V(m, N)`: $V(m, N)$

`E[V(m, N)]`: $E[V(m, N)]$

`C`: \hat{C} (first parameter)

`mu`: $\hat{\mu}$ (second parameter)

γ : $\hat{\gamma}$ (third parameter)

MSE: mean squared error

Nfit: $\sum_m m E[V(m, N)]$

Nproxy: $\sum_{m=1}^{m\max} m V_r(m, N)$

Vproxy: $\sum_{m=1}^{m\max} V_r(m, N)$

spectrum construct frequency spectrum and profile data

spectrum (-kZ -e -s -n) *text.txt*

This program takes a text as input and produces a word frequency list, the frequency spectrum, and the observed developmental profile.

INPUT

text.txt: ASCII text file (SGML markup is ignored)

OPTIONS

- e: input text file does not have .txt extension
- s: short output (suppresses creation of *text.zvc*, *text.zrk*)
- n: do not attempt to remove Sgml code
- kZ: number of text chunks is set at *Z* (default: 20)

OUTPUT

text.obs: empirical developmental profile statistics

N: *N* (number of tokens)

K: *K* (Yule's characteristic constant)

D: *D* (Simpson's diversity index)

V: *V* (number of types)

V1: *V*(1, *N*) (number of hapax legomena)

V2: *V*(2, *N*) (number of dis legomena)

V3: *V*(3, *N*) (number of tris legomena)

V4: *V*(4, *N*) (number of types with frequency 4)

V5: *V*(5, *N*) (number of types with frequency 5)

R: *R* (Guiraud's constant)

W: *W* (Brunet's constant)

S: *S* (Sichel's constant)

H: *H* (Honoré's constant)

C: *C* (Herdan's constant)

E: *E* (sample entropy)

1M: $\hat{\mu}$ (mean log frequency)

1St: $\hat{\sigma}$ (standard deviation of log frequency)

b: b (parameter of Sichel's model)
c: c (parameter of Sichel's model)
a1: $\alpha(1, N)$ (relative number of hapax legomena)
Z: Z (parameter of extended Zipf's law)
fa: frequency of first word with specified Zipf rank
fthe: frequency of second word with specified Zipf rank
sLmean: sample mean of lognormal model
sLstddev: sample standard deviation of lognormal model
text.zvc: the text in Zipf-vector format
 Word: the word tokens
 z: the Zipf ranks of the corresponding word types
text.wfl: the word frequency list
 Word: the word types ω_i
 Frequency: the frequencies $f(i, N)$ of these word types
text.spc: the frequency spectrum
 m: the frequency rank m
 Vm: $V(m, N)$, the number of words with frequency m
text.zrk: the Zipfian rank-frequency list
 z: the Zipf rank z
 fz: $f(z, N)$, the frequency of the word with Zipf rank z
text.sum: summary statistics for complete text

TECHNICAL DETAILS

The maximum number of different word types equals 40000, the maximum number of text chunks 40.

wfl2spc**construct frequency spectrum from word frequency list**

wfl2spc (-e -m) text.txt

This program takes a word frequency list as input and outputs the frequency spectrum.

INPUT

text.wfl: a word frequency list with columns labeled WORD and FREQUENCY

OPTIONS

-e: input does not have .wfl extension

-m: input does not have a header

OUTPUT

text.sum: summary statistics:

N: N (number of tokens)

K: K (Yule's characteristic constant)

D: D (Simpson's diversity index)

V: V (number of types)

V1: $V(1, N)$ (number of hapax legomena)

V2: $V(2, N)$ (number of dis legomena)

V3: $V(3, N)$ (number of tris legomena)

V4: $V(4, N)$ (number of types with frequency 4)

V5: $V(5, N)$ (number of types with frequency 5)

R: R (Guiraud's constant)

W: W (Brunet's constant)

S: S (Sichel's constant)

H: H (Honoré's constant)

C: C (Herdan's constant)

E: E (sample entropy)

1M: $\hat{\mu}$ (mean log frequency)

1St: $\hat{\sigma}$ (standard deviation of log frequency)

b: b (parameter of Sichel's model)

c: c (parameter of Sichel's model)

a1: $\alpha(1, N)$ (relative number of hapax legomena)

z: Z (parameter of extended Zipf's law)

fa: not implemented (available in the input file)

fthe: not implemented (available in the input file)

sLmean: sample mean of lognormal model

sLstddev: sample standard deviation of lognormal model

text.spc: the frequency spectrum

m: the frequency rank m

Vm: $V(m, N)$, the number of words with frequency m

TECHNICAL DETAILS

The maximum number of different word types that can be accommodated equals 40000.

Appendix D

Data sets

INDEX OF DATA SETS BY GROUP

TEXTS

- L.Carroll, *Alice's adventures in wonderland*
- L.Carroll, *Through the Looking-glass and what Alice found there*
- H.G.Wells, *War of the worlds*
- A.Conan-Doyle, *Hound of the Baskervilles*
- E.Douwes Dekker, *Max Havelaar*
- Turkish archeology text
- A.H.Tammsaare, *Truth and justice*

CORPORA

- British national corpus: context-governed subcorpus
- Aviator corpus: 1 million word sample from 1989
- Aviator corpus: 8 million word sample from 1989–1996

MORPHOLOGICAL CATEGORIES

- The Dutch suffix *-heid* in a 42 million corpus
- The Dutch suffix *-iteit* in a 42 million corpus
- The Dutch suffix *-ster* in a 42 million corpus
- The Dutch suffix *-in* in a 42 million corpus
- Dutch simplex nouns in a 42 million corpus

English singulars in Innes' *The bloody wood*

English plurals in Innes' *The bloody wood*

English nouns in *-ness* in the written subcorpus of the BNC

English nouns in *-ness* in the context-governed subcorpus of the BNC

English nouns in *-ness* in the demographic subcorpus of the BNC

OTHER

Filarial worms on mites on rats (Stein, Zucchini, and Juritz, 1987)

CV-patterns (Gale and Sampson, 1995)

Word pairs (bigrams) in E.Douwes-Dekker's *Max Havelaar*

Year references (Polman and Baayen, in press)

alice.spcL. Carroll, *Alice's Adventures in Wonderland*

SOURCE: Oxford Text Archive

SUMMARY STATISTICS: $N = 26505$, $V(N) = 2651$, $\mathcal{P}(N) = 0.0444$

FREQUENCY SPECTRUM:

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
1	1176	2	402	3	233	4	154
5	99	6	57	7	65	8	52
9	32	10	36	11	23	12	20
13	34	14	20	15	12	16	9
17	9	18	10	19	8	20	5
21	6	22	3	23	3	24	6
25	9	26	4	27	6	28	3
29	6	30	6	31	3	32	4
33	4	34	3	35	4	37	1
38	4	39	4	40	4	41	2
42	2	43	2	44	1	45	4
46	1	47	1	48	1	49	4
50	2	51	4	52	3	53	1
54	3	55	3	56	1	57	2
58	2	59	1	60	2	61	3
62	1	63	1	67	2	68	4
73	1	74	1	75	1	77	2
79	1	80	1	81	1	82	2
83	2	85	1	87	1	88	2
90	1	93	1	94	1	96	2
98	1	102	2	108	1	113	1
114	1	121	1	128	1	131	1
133	1	136	1	144	1	145	1
148	1	151	1	153	1	170	1
177	1	179	1	182	1	194	1
211	1	247	1	263	1	280	1
356	1	364	1	365	1	386	1
410	1	460	1	510	1	528	1
540	1	629	1	726	1	866	1
1631	1						

through.spcL. Carroll, *Through the looking-glass and what Alice found there*

SOURCE: Oxford Text Archive

SUMMARY STATISTICS: $N = 28767$, $V(N) = 3085$, $\mathcal{P} = 0.0518$

FREQUENCY SPECTRUM:

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
1	1491	2	460	3	259	4	148
5	113	6	78	7	61	8	47
9	28	10	26	11	26	12	30
13	22	14	19	15	12	16	21
17	12	18	11	19	16	20	9
21	7	22	9	23	2	24	3
25	1	26	5	27	3	28	7
29	5	30	2	31	5	32	3
33	2	34	5	35	5	36	2
37	5	38	3	39	2	40	1
41	2	42	2	45	1	46	3
48	4	49	2	50	2	51	3
52	4	53	2	54	4	55	2
56	1	57	2	58	1	59	1
60	1	61	2	62	2	63	3
64	3	65	1	66	1	67	3
69	1	70	4	72	1	73	1
74	2	75	1	78	1	79	2
80	1	84	2	86	2	87	1
89	1	90	1	93	1	94	1
101	1	104	1	112	1	113	1
115	1	116	1	119	1	121	2
123	2	132	1	135	1	139	1
140	1	145	1	147	1	150	1
151	1	177	2	180	1	193	1
195	1	209	1	211	1	229	1
247	1	268	1	300	1	309	1
354	1	399	1	425	1	470	2
502	1	505	1	517	1	545	1
705	1	739	1	836	1	1555	1

war.spcH.G. Wells, *War of the worlds*

SOURCE: Oxford Text Archive

SUMMARY STATISTICS: $N = 59938$, $V(N) = 7112$, $\mathcal{P} = 0.0603$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	3613	2	1138	3	567	4	340
5	250	6	177	7	135	8	93
9	72	10	67	11	44	12	46
13	44	14	42	15	38	16	31
17	24	18	26	19	16	20	18
21	14	22	13	23	11	24	7
25	8	26	10	27	8	28	6
29	9	30	9	31	4	32	8
33	2	34	6	35	9	37	6
38	6	39	7	40	4	41	6
42	3	43	6	44	3	45	4
46	2	47	3	48	6	49	6
50	5	52	1	53	4	55	3
57	4	58	2	59	2	60	2
61	3	63	3	65	2	66	4
67	3	68	3	69	2	70	1
71	4	72	1	73	1	74	1
75	2	76	1	78	2	79	1
82	1	85	3	87	1	88	1
90	1	91	1	94	1	96	1
99	3	100	1	101	5	102	2
103	1	108	1	112	1	114	1
116	2	117	1	120	2	124	1
128	1	129	2	140	1	142	1
146	1	150	1	154	3	158	1
164	1	166	2	167	1	171	2
174	1	177	1	181	1	184	1
185	1	191	1	198	1	199	1
207	1	213	1	218	1	231	1

(continued)

m	$V(m, N)$						
243	1	247	1	248	1	250	1
254	1	266	1	292	1	320	1
327	1	343	1	378	1	379	1
420	1	441	1	446	1	447	1
469	1	579	1	647	1	766	1
850	1	991	1	1172	1	1257	1
1605	1	2297	1	2487	1	4775	1

hound.spcA.Conan-Doyle, *Hound of the Baskervilles*

SOURCE: Oxford Text Archive

SUMMARY STATISTICS: $N = 59241$, $V(N) = 5741$, $\mathcal{P} = 0.0479$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	2836	2	889	3	449	4	280
5	208	6	137	7	116	8	92
9	86	10	52	11	48	12	40
13	33	14	25	15	34	16	22
17	20	18	15	19	13	20	13
21	17	22	14	23	9	24	12
25	7	26	16	27	5	28	8
29	7	30	7	31	7	32	4
33	6	34	8	35	2	36	7
37	5	38	3	39	5	40	4
41	3	42	3	43	8	44	3
45	3	46	4	47	1	48	1
49	2	50	1	52	1	54	4
55	5	57	3	58	1	60	2
61	1	62	5	63	3	64	2
65	3	66	2	67	3	68	2
69	2	70	1	71	1	72	1
73	1	74	2	77	2	80	1
82	1	84	2	85	1	86	2
87	1	88	3	89	1	90	2
92	1	94	1	97	2	99	2
102	1	104	1	105	1	107	1
110	2	111	1	112	1	113	3
114	1	118	2	123	1	128	1
137	2	140	1	141	1	143	1
146	1	149	1	151	1	155	1
165	1	167	1	171	1	175	1
178	2	182	1	185	1	190	1
192	1	199	1	200	1	201	2
202	1	205	1	207	1	209	1
211	1	215	1	222	1	233	1

(continued)

m	$V(m, N)$						
240	1	242	1	244	1	264	1
286	1	298	1	314	1	315	1
326	1	329	1	337	1	350	1
364	1	374	1	400	1	405	1
416	1	419	1	441	1	453	1
479	1	506	1	541	1	624	1
689	1	803	1	827	1	911	1
914	1	980	1	1132	1	1305	1
1407	1	1465	1	1592	1	1627	1
3327	1						

havelaar.spc**E.Douwes Dekker, Max Havelaar**

SOURCE: Oxford Text Archive

SUMMARY STATISTICS: $N = 99767$, $V(N) = 11161$, $\mathcal{P} = 0.0602$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	6004	2	1731	3	819	4	491
5	368	6	258	7	168	8	137
9	123	10	108	11	80	12	52
13	60	14	57	15	39	16	34
17	37	18	33	19	19	20	33
21	21	22	19	23	20	24	18
25	14	26	13	27	9	28	9
29	13	30	13	31	9	32	9
33	10	34	9	35	6	36	5
37	10	38	7	39	9	40	6
41	8	42	8	43	8	44	5
45	3	46	6	47	4	48	4
49	2	50	6	51	3	52	4
53	1	54	3	55	5	56	4
57	3	58	4	59	8	61	8
62	2	63	2	64	4	65	2
66	5	67	5	68	2	69	3
70	3	71	1	72	2	73	1
74	1	75	3	76	3	78	4
79	2	80	2	81	1	82	2
83	4	86	1	87	2	88	2
90	1	92	1	93	2	96	3
98	2	101	3	102	1	105	1
106	1	107	1	109	3	110	1
111	1	113	2	114	1	115	1
116	1	120	1	121	1	122	1
123	1	125	1	126	3	127	1
128	2	135	1	139	1	145	2
147	3	151	1	154	1	156	1
161	1	162	1	165	1	169	1
170	1	171	1	177	1	184	1

(continued)

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
188	1	190		1	194	1	198
202	3	208		1	222	1	223
228	1	234		1	235	1	236
238	1	242		2	244	2	262
272	1	283		1	285	1	286
289	1	300		1	308	1	317
323	1	344		1	358	1	360
365	1	369		1	384	1	391
416	1	430		1	437	2	443
452	1	453		1	477	1	479
494	1	541		1	631	1	650
653	1	710		1	714	1	736
920	1	957		1	990	1	1159
1168	1	1267		1	1335	1	1423
1644	1	1686		1	1834	1	1894
1955	1	2032		1	2306	1	2782
4826	1						

turkish.spc**Turkish archeology text**

SOURCE: WWW

SUMMARY STATISTICS: $N = 6939, V(N) = 3302, \mathcal{P} = 0.3352$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	2326	2	477	3	178	4	107
5	53	6	33	7	22	8	26
9	7	10	7	11	12	12	8
13	4	14	3	15	2	16	7
17	4	18	2	20	1	21	4
22	1	23	1	24	2	28	2
32	2	34	1	36	1	38	1
43	1	44	1	51	1	56	1
68	1	69	1	193	1	222	1

estonian.spcA.H.Tammsaare, *Truth and Justice*

SOURCE: Tuldava (1996)

SUMMARY STATISTICS: $N = 160356$, $V(N) = 8228$, $\mathcal{P} = 0.0227$

FREQUENCY SPECTRUM (partial):

m	$V(m, N)$						
1	3637	2	1216	3	613	4	441
5	297	6	219	7	175	8	124
9	110	10	85	11	65	12	56
13	59	14	36	15	53	16	48
17	24	18	41	19	30	20	34
21	15	22	25	23	21	24	30
25	19	26	18	27	18	28	8
29	17	30	15	31	20	32	8
33	20	34	17	35	15	36	14
37	15	38	10	39	8	40	7
41	16	42	15	43	13	44	13
45	13	46	6	47	4	48	7
49	10	50	4	51	9	52	7
53	10	54	7	55	6	56	2
57	6	58	2	59	6	60	4
61	7	62	7	63	3	64	4
65	3	66	1	67	4	68	3
69	5	70	8	71	6	72	4
73	5	74	4	75	3	76	6
77	2	78	4	79	3	80	2
81	4	82	4	83	5	84	2
85	2	86	2	87	7	88	1
89	1	90	5	91	1	92	2
93	2	94	1	95	4	96	3
97	1	99	2	100	1	101	5
1759	1	2251	1	2268	1	2470	1
2492	1	2807	1	3715	1	4268	1
5210	1	7168	1				

bnc.spc

The context-governed subcorpus of the BNC

SOURCE: British National Corpus

SUMMARY STATISTICS: $N = 6154206$, $V(N) = 79883$, $\mathcal{P} = 0.0052$

FREQUENCY SPECTRUM (partial):

m	$V(m, N)$						
1	32042	2	11383	3	6153	4	3866
5	2739	6	2178	7	1679	8	1358
9	1214	10	937	11	856	12	712
13	680	14	617	15	546	16	457
17	415	18	401	19	351	20	359
21	327	22	293	23	291	24	262
25	228	26	199	27	229	28	195
29	190	30	171	31	162	32	163
33	159	34	138	35	139	36	131
37	138	38	128	39	125	40	125
41	123	42	112	43	105	44	98
45	120	46	89	47	97	48	79
49	90	50	94	51	79	52	87
53	89	54	78	55	82	56	71
57	68	58	66	59	59	60	65
61	65	62	61	63	61	64	50
65	60	66	60	67	52	68	49
69	57	70	46	71	38	72	50
73	40	74	35	75	47	76	66
77	44	78	33	79	42	80	31
81	42	82	41	83	34	84	27
85	42	86	31	87	31	88	27
89	43	90	25	91	31	92	32
93	31	94	24	95	24	96	27
97	30	98	27	99	38	100	20
101	19	102	25	103	26	104	21
70296	1	75237	1	75509	1	82272	1
105313	1	117140	1	117906	1	126064	1
134074	1	136692	1	170675	1	295636	1

in1.spcsample of 1 million words from *The Independent*

SOURCE: A.Renouf's AVIATOR corpus

SUMMARY STATISTICS: $N = 1000000$, $V(N) = 65135$, $\mathcal{P} = 0.0311$

FREQUENCY SPECTRUM (partial):

m	$V(m, N)$						
1	31114	2	9562	3	4987	4	3099
5	2094	6	1487	7	1237	8	1048
9	860	10	700	11	588	12	503
13	503	14	385	15	344	16	305
17	310	18	247	19	249	20	209
21	221	22	167	23	172	24	164
25	160	26	177	27	144	28	115
29	128	30	113	31	109	32	105
33	100	34	100	35	103	36	68
37	79	38	70	39	73	40	53
41	65	42	50	43	64	44	59
45	64	46	52	47	35	48	64
49	51	50	36	51	41	52	59
53	43	54	39	55	41	56	38
57	42	58	31	59	29	60	39
61	25	62	27	63	27	64	32
65	26	66	20	67	33	68	23
69	31	70	27	71	24	72	25
73	28	74	26	75	13	76	16
77	16	78	17	79	15	80	20
81	16	82	19	83	19	84	15
85	13	86	16	87	15	88	17
89	19	90	10	91	9	92	13
93	17	94	14	95	15	96	13
97	13	98	18	99	11	100	15
101	8	102	10	103	14	104	6
7301	1	7695	1	8387	1	9158	1
9499	1	18799	1	20362	1	21435	1
26187	1	30358	1	56038	1		

in8.spc

sample of 8 million words from *The Independent*

SOURCE: A.Renouf's AVIATOR corpus

SUMMARY STATISTICS: $N = 7789402$, $V(N) = 175146$, $\mathcal{P} = 0.0098$

FREQUENCY SPECTRUM (partial):

m	$V(m, N)$						
1	76181	2	23622	3	12306	4	8010
5	5780	6	4451	7	3573	8	2921
9	2507	10	2133	11	1830	12	1616
13	1434	14	1296	15	1204	16	952
17	957	18	872	19	826	20	748
21	708	22	651	23	591	24	552
25	541	26	489	27	429	28	466
29	449	30	368	31	356	32	361
33	304	34	317	35	310	36	303
37	260	38	248	39	248	40	271
41	272	42	232	43	220	44	197
45	213	46	214	47	186	48	194
49	162	50	173	51	139	52	180
53	146	54	139	55	143	56	143
57	126	58	117	59	131	60	113
61	129	62	126	63	103	64	96
65	119	66	106	67	91	68	89
69	119	70	93	71	81	72	103
73	96	74	90	75	79	76	82
77	85	78	87	79	77	80	97
81	93	82	71	83	64	84	72
85	56	86	66	87	58	88	65
89	76	90	63	91	68	92	66
93	54	94	66	95	58	96	59
97	62	98	63	99	51	100	53
101	50	102	50	103	58	104	44
56965	1	61840	1	71021	1	73801	1
79995	1	146618	1	174148	1	175374	1
205740	1	229765	1	437992	1		

heid.spcNouns in *-heid* in the CELEX lexical database

SOURCE: Baayen, Piepenbrock, and Van Rijn (1993)

SUMMARY STATISTICS: $N = 167239$, $V(N) = 3070$, $\mathcal{P} = 0.0066$

FREQUENCY SPECTRUM (partial):

m	$V(m, N)$						
1	1110	2	286	3	195	4	132
5	98	6	71	7	70	8	74
9	43	10	56	11	32	12	26
13	30	14	26	15	32	16	26
17	21	18	14	19	20	20	18
21	17	22	17	23	13	24	9
25	18	26	19	27	15	28	13
29	10	30	11	31	12	32	7
33	12	34	10	35	5	36	9
37	10	38	5	39	11	40	7
41	8	42	9	43	6	44	8
45	6	46	6	47	7	48	6
49	6	50	6	51	4	52	2
53	6	54	6	55	11	56	4
57	6	58	4	59	2	60	5
61	2	62	5	63	4	64	3
65	3	66	5	67	6	68	3
69	2	70	1	71	2	72	4
73	6	74	7	75	5	76	1
77	3	78	6	79	3	80	2
81	4	82	3	83	2	84	4
85	1	86	3	87	3	88	3
89	1	90	2	92	1	93	1
94	1	95	3	96	1	97	3
98	2	99	3	100	3	103	5
3352	1	4065	1	4435	1	4752	1
5298	1	5380	1	5579	1	6437	1
9623	1						

iteit.spc**Nouns in *-iteit* in the CELEX lexical database**

SOURCE: Baayen, Piepenbrock, and Van Rijn (1993)

SUMMARY STATISTICS: $N = 21156$, $V(N) = 362$, $\mathcal{P} = 0.0063$

FREQUENCY SPECTRUM:

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
1	133	2	28	3	14	4	9
5	3	6	6	7	2	8	4
9	5	10	5	11	4	12	8
13	3	14	7	15	2	16	2
17	3	18	3	19	1	20	3
21	5	22	3	23	2	24	5
25	3	26	2	27	1	28	2
29	2	30	1	31	1	32	5
33	2	37	1	38	2	39	1
40	1	41	1	42	3	43	1
44	1	48	1	49	1	50	1
53	1	54	1	55	2	59	3
62	1	63	3	64	1	69	1
71	1	72	1	73	1	75	1
79	1	86	1	88	2	90	1
95	1	102	1	104	1	105	1
120	1	128	3	131	2	140	1
153	1	156	1	158	1	162	1
163	1	169	1	178	1	180	1
181	1	183	1	206	1	218	1
219	1	220	1	221	1	224	1
237	1	243	1	250	1	256	1
285	1	293	1	297	1	317	1
363	1	383	1	395	1	425	1
462	1	468	1	476	1	481	1
585	1	683	1	1487	1	4500	1

ster.spc**Nouns in -ster in the CELEX lexical database**

SOURCE: Baayen, Piepenbrock, and Van Rijn (1993)

SUMMARY STATISTICS: $N = 5120$, $V(N) = 370$, $\mathcal{P} = 0.0314$

FREQUENCY SPECTRUM:

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
1	161	2	41	3	31	4	13
5	13	6	10	7	13	8	11
9	10	10	5	11	1	12	4
13	9	14	2	15	1	17	2
19	1	20	2	22	2	23	3
25	1	26	2	27	1	29	1
30	1	33	2	34	1	35	1
37	1	38	1	39	1	40	2
43	1	45	1	46	2	57	1
58	1	61	1	62	1	69	1
87	1	95	1	102	2	109	1
140	1	150	1	266	1	330	1
362	1	1141	1				

SOURCE: Baayen, Piepenbrock, and Van Rijn (1993)

SUMMARY STATISTICS: $N = 7985$, $V(N) = 50$, $\mathcal{P} = 0.0009$

FREQUENCY SPECTRUM:

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
1	7	2	4	3	2	4	1
6	2	7	1	10	1	11	2
13	1	15	1	18	4	19	1
24	2	26	1	30	1	33	1
34	1	40	1	45	1	58	1
59	1	79	1	85	1	104	1
112	1	123	1	199	1	216	1
299	1	334	1	478	1	593	1
1740	1	3055	1				

noun.spc**Simplex nouns in the CELEX lexical database**

SOURCE: Baayen, Piepenbrock, and Van Rijn (1993)

SUMMARY STATISTICS: $N = 3507305$, $V(N) = 5042$, $\mathcal{P} = 0.0001$

FREQUENCY SPECTRUM (partial):

m	$V(m, N)$						
1	226	2	174	3	123	4	105
5	89	6	75	7	66	8	69
9	73	10	46	11	52	12	64
13	56	14	39	15	46	16	42
17	44	18	39	19	33	20	48
21	24	22	32	23	32	24	37
25	36	26	24	27	36	28	28
29	21	30	22	31	25	32	29
33	23	34	21	35	23	36	27
37	22	38	23	39	21	40	23
41	16	42	21	43	16	44	22
45	12	46	27	47	24	48	15
49	16	50	14	51	19	52	15
53	14	54	18	55	18	56	15
57	5	58	15	59	15	60	16
61	11	62	14	63	17	64	19
65	14	66	10	67	8	68	11
69	11	70	12	71	10	72	14
73	10	74	7	75	13	76	14
77	6	78	15	79	16	80	14
81	10	82	9	83	15	84	11
85	9	86	12	87	3	88	12
89	13	90	10	91	14	92	10
93	9	94	6	95	4	96	6
97	8	98	7	99	14	100	7
101	6	102	8	103	14	104	9
24408	1	25099	1	25272	1	26719	1
27996	1	34753	1	38126	1	39630	1
40727	1	43573	1	45945	1	48452	1
50439	1	58078	1				

sing.spcsingular nouns in *The bloody wood* by M.Innes

SOURCE: TOSCA corpus

SUMMARY STATISTICS: $N = 2352$, $V(N) = 985$, $\mathcal{P} = 0.2560$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	602	2	165	3	76	4	39
5	25	6	12	7	14	8	11
9	6	10	7	11	6	13	4
15	2	17	2	18	1	19	2
20	3	21	2	25	1	26	1
29	1	31	1	36	1	42	1

plur.spcplural nouns in *The bloody wood* by M.Innes

SOURCE: TOSCA corpus (Keulen, 1986)

SUMMARY STATISTICS: $N = 386$, $V(N) = 236$, $\mathcal{P} = 0.4482$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	173	2	31	3	16	4	4
5	5	6	3	7	1	9	1
11	1	17	1				

nessw.spc

Nouns in *-ness* in BNC, written subcorpus

SOURCE: British National Corpus

SUMMARY STATISTICS: $N = 106957$, $V(N) = 2466$, $\mathcal{P} = 0.0088$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	943	2	354	3	188	4	114
5	91	6	65	7	45	8	53
9	53	10	23	11	26	12	29
13	24	14	15	15	13	16	15
17	9	18	10	19	14	20	10
21	7	22	12	23	7	24	11
25	11	26	4	27	11	28	6
29	9	30	7	31	7	32	7
33	5	34	6	35	2	36	10
37	1	38	5	39	4	40	6
41	4	42	2	43	6	44	4
45	6	46	7	47	2	48	1
49	2	50	3	51	2	52	2
53	5	54	3	56	1	57	1
58	3	59	5	60	1	61	2
62	3	63	2	65	3	67	2
68	2	69	2	71	4	72	2
73	1	74	1	77	3	79	2
81	2	82	2	83	1	84	2
85	2	86	2	88	1	89	1
90	2	91	1	94	1	96	2
97	1	99	1	100	2	101	1
102	1	103	1	104	1	105	2
106	2	107	1	109	1	110	1
111	1	116	2	117	1	119	1

(continued)

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
120	1	121	1	124	1	125	1
128	1	129	3	131	1	133	1
134	2	141	2	144	1	146	1
147	1	148	2	151	3	152	1
155	2	156	1	159	2	160	1
162	2	165	1	168	1	171	1
176	1	177	1	179	1	180	3
183	1	193	1	194	1	196	1
200	1	202	1	206	1	207	1
217	1	219	2	229	1	231	1
236	1	239	1	240	1	241	1
250	1	261	1	263	1	276	2
277	1	288	2	293	1	296	1
305	1	312	2	313	1	314	1
326	2	375	1	385	1	388	1
392	1	394	1	396	1	407	1
468	1	512	1	553	1	568	1
571	1	599	1	644	1	679	1
707	1	719	1	768	1	783	1
797	1	1105	1	1132	1	1157	1
1516	1	1663	1	2036	1	2560	1
2639	1	3366	1	3480	1	3594	1
36883	1						

nessscg.spc**Nouns in *-ness* in BNC, context-governed subcorpus**

SOURCE: British National Corpus

SUMMARY STATISTICS: $N = 4037$, $V(N) = 310$, $\mathcal{P} = 0.0394$

FREQUENCY SPECTRUM:

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
1	159	2	48	3	25	4	17
5	5	6	4	7	7	8	5
9	3	10	3	11	3	12	1
13	1	14	3	15	2	16	1
17	1	19	1	20	1	23	1
24	2	25	1	31	1	33	1
35	2	38	1	39	1	41	1
44	1	47	1	57	1	72	1
86	1	88	1	172	1	193	1
2135		1					

nessd.spc**Nouns in *-ness* in BNC, demographic subcorpus**

SOURCE: British National Corpus

SUMMARY STATISTICS: $N = 938$, $V(N) = 118$, $\mathcal{P} = 0.0810$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	76	2	13	3	11	4	3
5	3	6	1	7	2	8	2
9	0	10	0	11	0	12	1
13	1	14	0	15	1	16	1
24	1	226	1	434	1		

filarial.spc**filarial worms on mites on rats**

SOURCE: Stein, Zucchini, and Juritz (1987)

SUMMARY STATISTICS: $N = 6781$, $V(N) = 1445$, $\mathcal{P} = 0.0816$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	553	2	265	3	150	4	98
5	70	6	48	7	36	8	28
9	27	10	15	11	13	12	21
13	8	14	11	15	5	16	9
17	9	18	9	19	8	20	5
21	2	22	4	23	5	24	3
25	5	26	1	28	2	30	5
31	6	32	4	33	2	34	1
35	2	36	4	37	1	41	2
45	1	51	1	59	1	62	1
63	1	64	1	67	1	78	1

cv.spc**CV-patterns in the TIMIT speech database**

SOURCE: Gale and Sampson (1995)

SUMMARY STATISTICS: $N = 30934$, $V(N) = 310$, $\mathcal{P} = 0.0039$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	120	2	40	3	24	4	13
5	15	6	5	7	11	8	2
9	2	10	1	12	3	14	2
15	1	16	1	17	3	19	1
20	3	21	2	23	3	24	3
25	3	26	2	27	2	28	1
31	2	32	3	33	1	34	2
36	2	41	3	43	1	45	3
46	1	47	1	50	1	71	1
84	1	101	1	105	1	121	1
124	1	146	1	162	1	193	1
199	1	224	1	226	1	254	1
257	1	339	1	421	1	456	1
481	1	483	1	1140	1	1256	1
1322	1	1530	1	2131	1	2395	1
6925	1	7846	1				

pairs.spcWord pairs in *Max Havelaar* by E.Douwes Dekker

SOURCE: Oxford Text Archive

SUMMARY STATISTICS: $N = 99766$, $V(N) = 59156$, $\mathcal{P} = 0.4809$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	47974	2	5825	3	2013	4	925
5	615	6	400	7	247	8	217
9	121	10	94	11	89	12	68
13	57	14	40	15	39	16	35
17	33	18	30	19	22	20	21
21	25	22	16	23	21	24	13
25	14	26	14	27	14	28	7
29	15	30	12	31	5	32	6
33	3	34	8	35	2	36	8
37	3	38	3	39	6	40	5
41	1	42	2	43	3	44	4
45	2	46	5	47	3	48	4
49	2	51	2	52	4	53	4
56	3	57	2	58	2	60	1
61	2	64	1	65	1	66	2
68	4	69	2	70	1	72	2
73	1	75	1	77	2	78	1
81	2	82	1	83	1	87	1
88	1	92	1	93	1	94	1
96	1	103	1	105	1	107	2
120	1	125	1	140	1	159	1
170	1	174	1	199	1	204	1
222	1	241	1	242	1	428	1
493	1						

years.spcYear references in the *Frankfurter Allgemeine Zeitung*

SOURCE: Polman and Baayen (in press)

SUMMARY STATISTICS: $N = 72259$, $V(N) = 731$, $\mathcal{P} = 0.0011$

FREQUENCY SPECTRUM:

m	$V(m, N)$						
1	81	2	64	3	41	4	49
5	43	6	38	7	28	8	30
9	16	10	18	11	21	12	17
13	14	14	10	15	6	16	9
17	3	18	7	19	3	20	2
21	7	22	3	23	3	24	4
25	5	26	4	27	2	28	3
29	4	30	5	31	1	32	6
33	3	34	4	35	4	36	2
37	4	38	4	39	3	41	3
42	5	43	3	44	2	45	1
46	2	47	1	48	1	49	3
50	3	51	2	53	2	55	1
56	2	57	3	58	1	60	1
61	3	62	2	63	1	64	3
65	1	69	2	70	2	75	1
78	1	80	3	84	2	89	1
92	1	93	1	96	1	105	1
107	1	109	1	110	1	111	2
112	2	113	1	122	1	125	1
128	1	129	2	131	1	139	1
140	1	155	1	160	1	172	1
175	1	179	2	194	1	195	1
198	2	203	1	209	1	214	1
215	1	221	1	234	1	244	1
245	1	257	1	262	1	267	1
272	1	276	1	277	1	280	1
287	1	298	1	309	1	310	1
311	1	316	1	327	1	331	1
341	1	344	1	345	1	351	1
352	1	355	1	357	2	358	1

(continued)

m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$	m	$V(m, N)$
359	1	361	1	370	1	383	1
392	1	393	1	404	1	430	2
434	1	439	1	449	1	454	1
461	1	472	1	508	1	528	3
538	1	565	1	576	1	580	1
587	1	629	1	667	1	749	1
820	1	821	1	855	1	1061	1
1086	1	1138	1	1383	1	1496	1
2639	1	3713	1	3791	1	6338	1
12068	1						

Bibliography

- Atkinson, A. and Yeh, L.: 1982, Inference for Sichel's compound poisson distribution, *Journal of the American Statistical Association* **77**, 153–157.
- Baayen, R. H.: 1991, A stochastic process for word frequency distributions, *Proceedings of the 29th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Berkeley, pp. 271–278.
- Baayen, R. H.: 1993, Statistical models for word frequency distributions: A linguistic evaluation, *Computers and the Humanities* **26**, 347–363.
- Baayen, R. H.: 1994a, Derivational productivity and text typology, *Journal of Quantitative Linguistics* **1**, 16–34.
- Baayen, R. H.: 1994b, Productivity in production, *Language and Cognitive Processes* **9**, 447–469.
- Baayen, R. H.: 1996a, The effect of lexical specialization on the growth curve of the vocabulary, *Computational Linguistics* **22**, 455–480.
- Baayen, R. H.: 1996b, The randomness assumption in word frequency statistics, in G. Perissinotto (ed.), *Research in Humanities Computing 5*, Oxford University Press, Oxford, pp. 17–31.
- Baayen, R. H. and Lieber, R.: 1997, Word frequency distributions and lexical semantics, *Computers and the Humanities* **30**, 281–291.
- Baayen, R. H. and Neijt, A.: 1997, Productivity in context: a case study of a Dutch suffix, *Linguistics* **35**, 565–587.
- Baayen, R. H. and Renouf, A.: 1996, Chronicling The Times: Productive Lexical Innovations in an English Newspaper, *Language* **72**, 69–96.
- Baayen, R. H. and Tweedie, F. J.: 1998a, Enhancing LNRE models with partition-based adjustment, *Proceedings of JADT 1998*, Université Nice Sophia Antipolis, Nice, pp. 29–37.
- Baayen, R. H. and Tweedie, F. J.: 1998b, Sample-size invariance of LNRE model parameters: problems and opportunities, *Journal of Quantitative Linguistics* **5**, 145–154.

- Baayen, R. H., Dijkstra, T. and Schreuder, R.: 1997, Singulars and plurals in Dutch: Evidence for a parallel dual route model, *Journal of Memory and Language* **36**, 94–117.
- Baayen, R. H., Piepenbrock, R. and van Rijn, H.: 1993, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.
- Baayen, R. H., Schreuder, R., Bertram, R. and Tweedie, F.: 1999, The semantic functions of the Dutch suffix *-heid*: evidence from lexicography, lexical statistics, and psycholinguistics, in M. Nenonen and J. Jarvikivi (eds), *Languages, minds, and brains*, Joensuu University, Joensuu, pp. 1–29.
- Baayen, R. H., Van Halteren, H. and Tweedie, F. J.: 1996, Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, *Literary and Linguistic Computing* **11**, 121–131.
- Balasubrahmanyam, V. and Naranan, S.: 1996, Quantitative linguistics and complex system studies, *Journal of Quantitative Linguistics* **3**, 177–228.
- Bertram, R., Schreuder, R. and Baayen, R. H.: 2000, The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity, *Journal of Experimental Psychology: Memory, Learning, and Cognition* **26**, 419–511.
- Best, K.-H. and Zhu, J.: 1994, Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische, in U. Klenk (ed.), *Computatio Linguae II (ZDL Beiheft 83)*, Franz Steiner Verlag, Stuttgart, pp. 19–30.
- Brunet, E.: 1978, *Le Vocabulaire de Jean Giraudoux*, Vol. 1 of *TLQ*, Slatkine, Genève.
- Bunge, J. and Fitzpatrick, M.: 1993, Estimating the number of species: a review, *Journal of the American Statistical Association* **88**, 364–373.
- Burrell, Q. and Fenton, M.: 1993, Yes, the GIGP really does work – and is workable, *Journal of the American Society for Information Science* **44**, 61–69.
- Burrows, J. F.: 1992, Computers and the study of literature, in C. S. Butler (ed.), *Computers and Written Texts*, Blackwell, Oxford, pp. 167–204.
- Carroll, J. B.: 1967, On sampling from a lognormal model of word frequency distribution, in H. Kučera and W. N. Francis (eds), *Computational Analysis of Present-Day American English*, Brown University Press, Providence, pp. 406–424.
- Carroll, J. B.: 1969, A rationale for an asymptotic lognormal form of word frequency distributions, *Research Bulletin*, Educational Testing Service, Princeton.

- Carroll, J. B.: 1970, An alternative to Julland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (sfi), *Computer Studies in the Humanities and Verbal Behavior* 3, 61–65.
- Chen, Y. and Leimkuhler, F.: 1989, A type-token identity in the Simon-Yule model of text, *Journal of the American Society for Information Science* 40, 45–53.
- Chitashvili, R. J. and Baayen, R. H.: 1993, Word frequency distributions, in G. Altmann and L. Hřebíček (eds), *Quantitative Text Analysis*, Wissenschaftlicher Verlag Trier, Trier, pp. 54–135.
- Church, K. and Gale, W.: 1991, A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams, *Computer Speech and Language* 5, 19–54.
- Damerau, F.: 1975, The use of function word frequencies as indicators of style, *Computers and the Humanities* 9, 271–280.
- Efron, B. and Thisted, R.: 1976, Estimating the number of unseen species: How many words did Shakespeare know?, *Biometrika* 63, 435–447.
- Frauenfelder, U. H., Baayen, R. H., Hellwig, F. M. and Schreuder, R.: 1993, Neighborhood density and frequency across languages and modalities, *Journal of Memory and Language* 32, 781–804.
- Gale, W. A. and Sampson, G.: 1995, Good-Turing frequency estimation without tears, *Journal of Quantitative Linguistics* 2, 217–237.
- Gani, J.: 1997, Characterizing an author's vocabulary, *South African Statistical Journal* 31, 1–11.
- Good, I. J.: 1953, The population frequencies of species and the estimation of population parameters, *Biometrika* 40, 237–264.
- Good, I. J. and Toulmin, G. H.: 1956, The number of new species and the increase in population coverage, when a sample is increased, *Biometrika* 43, 45–63.
- Grotjahn, R. and Altmann, G.: 1993, Modeling the distribution of word length: some methodological problems, in R. Koehler and B.B. Rieger (eds), *Contributions to quantitative linguistics*, Kluwer, Dordrecht, pp. 141–153.
- Guiraud, H.: 1954, *Les Caractères Statistiques du Vocabulaire*, Presses Universitaires de France, Paris.
- Guiter, H. and Arapov, M. V. (eds): 1983, *Studies on Zipf's Law*, Brockmeyer, Bochum.
- Hammerl, R.: 1990, Untersuchungen zur Verteilung der Wortarten im Text, in L. Hřebíček (ed.), *Glottometrika* 11, Brockmeyer, Bochum, pp. 142–156.

- Hammersley, J. and Handscomb, D.: 1964, *Monte Carlo Methods*, Chapman and Hall, London.
- Heller, G.: 1997, Estimation of the number of classes, *South African Statistical Journal* 31, 65–90.
- Herdan, G.: 1960, *Type-Token Mathematics*, Mouton, The Hague.
- Herdan, G.: 1964, *Quantitative Linguistics*, Butterworths, London.
- Herdan, G.: 1966, *The advanced Theory of Language as Choice and Chance. (Kommunikation und Kybernetik in Einzeldarstellungen, Bd. 4)*, Springer, Berlin. Heidelberg: New York.
- Holmes, D. I.: 1992, A stylometric analysis of Mormon scripture and related texts, *Journal of the Royal Statistical Society Series A* 155, 91–120.
- Holmes, D. I.: 1994, Authorship attribution, *Computers and the Humanities* 28(2), 87–106.
- Holmes, D. I. and Forsyth, R.: 1995, The Federalist revisited: new directions in authorship attribution, *Literary & Linguistic Computing* 10, 111–127.
- Honoré, A.: 1979, Some simple measures of richness of vocabulary, *Association of Literary and Linguistic Computing Bulletin* pp. 172–179.
- Huang, B. H. and Lo, S. H.: 1994, On the bias of the Turnig-Good estimate of probabilities, *IEEE transactions on signal processing* 42, 496–498.
- Hubert, P. and Labbe, D.: 1988a, A model of vocabulary partition, *Literary and Linguistic Computing* 3, 223–225.
- Hubert, P. and Labbe, D.: 1988b, Un modèle de partition du vocabulaire, in D. Labbe, P. Thoiron and D. Serant (eds), *Etudes sur la richesse et les structures lexicales*, Slatkine-Champion, Paris, pp. 93–114.
- Johnson, N. L. and Kotz, S.: 1977, *Urn Models and Their Application. An Approach to Modern Discrete Probability Theory*, John Wiley & Sons, New York.
- Kageura, K.: 1998, A statistical analysis of morphemes in Japanese terminology, *Proceedings of Coling 1998*, Coling-ACL'98, Montreal, pp. 8–14.
- Kalinin, V. M.: 1965, Functionals related to the Poisson distribution, and statistical structure of a text, in J. V. Finnik (ed.), *Articles on Mathematical Statistics and the Theory of Probability*, Steklov Institute of Mathematics 79, American Mathematical Society, Providence, Rhode Island, pp. 202–220.
- Keulen, F.: 1986, The Dutch computer corpus pilot project, in J. Aarts and W. Meijs (eds), *Corpus linguistics II. New studies in the analysis and exploitation of computer corpora*, Rodopi, Amsterdam.
- Khmaladze, E. V.: 1987, The statistical analysis of large number of rare events, *Technical Report Report MS-R8804*, Dept. of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

- Khmaladze, E. V. and Chitashvili, R. J.: 1989, Statistical analysis of large number of rare events and related problems, *Transactions of the Tbilisi Mathematical Institute* 91, 196–245.
- Koehler, R.: 1986, *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*, Brockmeyer, Bochum.
- Labbe, D. and Hubert, P.: 1993, La richesse du vocabulaire (vocabulary richness), Centre de Recherche sur le Politique, l'Administration et le Territoire.
- Landauer, T. K. and Streeter, L. A.: 1973, Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition, *Journal of Learning and Verbal Behavior* 12, 119–131.
- Lánský, P. and Radil-Weiss, T.: 1980, A generalization of the Yule-Simon model, with special reference to word association tests and neural cell assembly formation, *Journal of Mathematical Psychology* 21, 53–65.
- Lebart, L. and Salem, A.: 1994, *Statistique Textuelle*, Dunod, Paris.
- Mandelbrot, B.: 1953, An information theory of the statistical structure of language, in W. E. Jackson (ed.), *Communication theory*, Academic Press, New York, pp. 503–512.
- Mandelbrot, B.: 1959, A note on a class of skew distribution functions: Analysis and critique of a paper by H.A. Simon, *Information and Control* 2, 90–99.
- Mandelbrot, B.: 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*, American Mathematical Society, Providence, Rhode Island, pp. 190–219.
- Martindale, C. and McKenzie, D.: 1995, On the utility of content analysis in author attribution: the Federalist, *Computers and the Humanities* 29, 259–270.
- Menard, N.: 1983, *Mesure de la Richesse Lexicale. Théorie et vérifications expérimentales. Etudes stylométriques et sociolinguistiques*, Slatkine-Champion, Genève.
- Meyer, H. (ed.): 1956, *Symposium on Monte Carlo methods*, Wiley, New York.
- Miller, G. A.: 1957, Some effects of intermittent silence, *The American Journal of Psychology* 52, 311–314.
- Morrison, D. F.: 1971, *Multivariate Statistical Methods*, McGraw-Hill Kokusisha, Tokyo.
- Mosteller, F. and Wallace, D. L.: 1964, *Applied Bayesian and Classical Inference. The case of the Federalist Papers*, Springer, New York.
- Muller, C.: 1977, *Principes et méthodes de statistique lexicale*, Hachette, Paris.

- Muller, C.: 1979a, Du nouveau sur les distributions lexicales: la formule de Waring-Herdan, in C. Muller (ed.), *Langue Française et Linguistique Quantitative*, Slatkine, Genève, pp. 177–195.
- Muller, C.: 1979b, *Langue Française et Linguistique Quantitative*, Slatkine, Genève.
- Muller, C.: 1979c, Peut-on estimer l'étendue d'un lexique?, *Langue Française et Linguistique Quantitative*, Slatkine, Genève, pp. 399–425.
- Nádas, A.: 1985, On Turing's formula for word probabilities, *IEEE Transactions on Acoustic Speech Signal Processing ASSP-33*, 1414–1416.
- Naranan, S. and Balasubrahmanyam, V.: 1998, Models for power law relations in linguistics and information science, *Journal of Quantitative Linguistics* 5, 35–61.
- Nelder, J. and Mead, R.: 1965, A simplex method for function minimization, *Computer Journal* 7, 308–313.
- Orlov, J. K.: 1983a, Dynamik der Häufigkeitsstrukturen, in H. Guiter and M. V. Arapov (eds), *Studies on Zipf's Law*, Brockmeyer, Bochum, pp. 116–153.
- Orlov, J. K.: 1983b, Ein Modell der Häufigkeitsstruktur des Vokabulars, in H. Guiter and M. V. Arapov (eds), *Studies on Zipf's Law*, Brockmeyer, Bochum, pp. 154–233.
- Orlov, J. K. and Chitashvili, R. Y.: 1982a, On some problems of statistical estimation in relatively small samples, *Bulletin of the Academy of Sciences, Georgia* 108, 513–516.
- Orlov, J. K. and Chitashvili, R. Y.: 1982b, On the distribution of frequency spectrum in small samples from populations with a large number of events, *Bulletin of the Academy of Sciences, Georgia* 108, 297–300.
- Orlov, J. K. and Chitashvili, R. Y.: 1983a, Generalized Z-distribution generating the well-known "rank-distributions", *Bulletin of the Academy of Sciences, Georgia* 110, 269–272.
- Orlov, J. K. and Chitashvili, R. Y.: 1983b, On the statistical interpretation of Zipf's law, *Bulletin of the Academy of Sciences, Georgia* 109, 505–508.
- Paivio, A., Yuille, J. C. and Madigan, S.: 1968, Concreteness, imagery, and meaningfulness values for 925 nouns, *Journal of Experimental Psychology Monograph*.
- Plag, I., Dalton-Puffer, C. and Baayen, R. H.: 1999, Productivity and register, *Journal of English Language and Linguistics*.
- Polman, T. and Baayen, R.: 2000, Computing historical consciousness. A quantitative inquiry into the presence of the past in newspaper texts, *Computers and the Humanities* (in press).

- Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T.: 1988, *Numerical Recipes in C. The art of Scientific Computing*, Cambridge University Press, Cambridge.
- Price, T.: 1997, Vocabulary growth functions, *South African Statistical Journal* **14**, 39–63.
- Reder, L. M., Anderson, J. R. and Bjork, R. A.: 1974, A semantic interpretation of encoding specificity, *Journal of Experimental Psychology* **102**, 648–656.
- Renouf, A.: 1993, A word in time: First findings from the investigation of dynamic text, in J. Aarts, P. De Haan and N. Oostdijk (eds), *English Language Corpora: Design, Analysis, and Exploitation*, Rodopi, Amsterdam.
- Rice, J. A.: 1988, *Mathematical Statistics and Data Analysis*, Wadsworth & Brooks, Pacific Grove.
- Ross, S. M.: 1988, *A First Course in Probability*, Macmillan Publishing Company, New York.
- Rouault, A.: 1978, Loi de Zipf et sources markoviennes, *Annales de l'Institute H. Poincaré* **14**, 169–188.
- Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **36**, 118–139.
- Shapiro, B. J.: 1969, The subjective estimation of word frequency, *Journal of verbal learning and verbal behavior* **8**, 248–251.
- Sichel, H. S.: 1975, On a distribution law for word frequencies, *Journal of the American Statistical Association* **70**, 542–547.
- Sichel, H. S.: 1986, Word frequency distributions and type-token characteristics, *Mathematical Scientist* **11**, 45–72.
- Sichel, H. S.: 1997, Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gauss-Poisson distribution, *South African Statistical Journal* **31**, 13–37.
- Simon, H. A.: 1955, On a class of skew distribution functions, *Biometrika* **42**, 435–440.
- Simon, H. A.: 1960, Some further notes on a class of skew distribution functions, *Information and Control* **3**, 80–88.
- Simon, H. A.: 1961, Reply to "final note" by Benoit Mandelbrot, *Information and Control* **4**, 217–223.
- Simpson, E. H.: 1949, Measurement of diversity, *Nature* **163**, 168.
- Stein, G. Z., Zucchini, W. and Juritz, J.: 1987, Parameter estimation for the Sichel distribution and its multivariate extension, *Journal of the American Statistical Association* **82**, 938–944.

- Taft, M.: 1979, Recognition of affixed words and the word frequency effect, *Memory and Cognition* 7, 263–272.
- Thisted, R. and Efron, B.: 1987, Did Shakespeare write a newly discovered poem?, *Biometrika* 74, 445–455.
- Tiersma, P. M.: 1982, Local and general markedness, *Language* 58, 832–849.
- Titterington, D. M., Smith, A. F. M. and Makov, U. E.: 1985, *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons, Chichester.
- Tukey, J. W.: 1977, *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.
- Tuldava, J.: 1996, The frequency spectrum of text and vocabulary, *Journal of Quantitative Linguistics* 3, 38–50.
- Tweedie, F. J. and Baayen, R. H.: 1998, How variable may a constant be? Measures of lexical richness in perspective, *Computers and the Humanities* 32, 323–352.
- Veld, R.: 1984, *Hoe willekeurig kiest een schrijver zijn woorden? Een urnmodel voor onderzoek naar de frequenties van woorden, munten, achternamen en vissen (How arbitrarily does a writer select his words? an urn model for researching the frequencies of words, coins, surnames, and fish)*, Master's thesis, University of Amsterdam.
- Whissell, C.: 1996, Traditional and emotional stylometric analysis of the songs of beatles Paul McCartney and John Lennon, *Computers and the Humanities* 30, 257–265.
- Wimmer, G., Koehler, R., Grotjahn, R. and Altmann, G.: 1994, Towards a theory of word length distribution, *Journal of Quantitative Linguistics* 1, 98–106.
- Yule, G. U.: 1924, A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S., *Philosophical Transactions of the Royal Society of London Ser. B* 213, 21–87.
- Yule, G. U.: 1944, *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge.
- Zipf, G. K.: 1935, *The Psycho-Biology of Language*, Houghton Mifflin, Boston.
- Zipf, G. K.: 1949, *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*, Hafner, New York.

Index

- adjusted LNRE models, 173
adjustment
 parameter-based, 179
 partition-based, 174
Altmann, 125, 186, 195, 196
Anderson, 195
Arapov, 35
Atkinson, 133

Baayen, 35, 56, 76, 133, 145, 146,
 160, 165, 167, 170, 193, 195,
 201, 206, 210, 223, 249
Balasubrahmanyam, 21, 61, 124, 129,
 160, 170, 214, 221, 229
Bertram, 146, 201
Bessel function, 89, 90
Best, 195
binomial
 distribution
 as approximation of hypergeometric distribution, 67
 interpolation, 64, 228
 probability, 43
 theorem, 43
Bjork, 195
Brunet, 25, 35, 211
Bunge, 133
Burrell, 124, 133
Burrows, 214

Carroll, 32, 35, 78, 84, 133, 223, 245
characteristic constants, 24, 34, 179,
 183, 236
C, 29, 31
D, 25, 30, 46
H, 27, 30, 31
K, 25, 30, 46, 242
R, 25, 31
S, 27, 31

W, 26, 31
Z, 80
Chen, 133
chi-squared test, 118, 119, 125, 228
Chitashvili, 35, 56, 76, 84, 94, 133,
 160, 223
Church, 20, 58, 61, 76
coefficient of loss, 56, 57, 73, 215
combination, 41
confidence interval, 61
cost function, 123, 183, 185
covariance
 definition of, 40, 120
 matrix, 119

Dalton-Puffer, 210
Damerau, 35
developmental profile, 7, 8, 30, 34,
 36, 173, 174, 176, 179, 182,
 183, 185
and link functions, 189
erratic, 187
goodness-of-fit testing, 183
Dijkstra, 249
dis legomena, 8, 14, 27, 33, 51, 53,
 55, 92, 100, 173, 228
discourse cohesion, 7
discretization, 20, 33, 103
dispersion, 164
distribution
 LNRE, 5, 54, 56, 57
 Bernoulli, 40
 binomial, 43, 45, 67
 distance-frequency, 215
 empirical structural type, 10, 11,
 13, 47
 grouped frequency, 8, 10, 11
 harmonic, 75, 79, 80, 99

- hypergeometric, 66, 67
inverse Gauss–Poisson, 89
lognormal, 32, 33
lognormal token, 87
lognormal type, 85, 87
normal, 32, 33
Poisson, 45
probability, 15, 16, 51
rank-frequency, 13, 215
standard harmonic, 16
structural, 35
structural token, 83, 84, 86
structural type, 47, 78, 81, 83,
 85, 86, 89, 94
uniform, 34, 166, 175, 176
Waring, 114
word frequency, 13, 33, 42, 51,
 53, 54
zeta, 15, 18, 37, 75, 88
Zipf–Mandelbrot, 37
downhill simplex method, 123, 231
- Efron, 35, 75, 76, 78, 245
error vector, 119
Euler's constant, 16
expectation, 39
extrapolation, 63, 69, 173
 parameter-adjusted LNRE, 188
 partition-adjusted LNRE, 176
- Fenton, 124, 133
Fitzpatrick, 133
Forsyth, 211
Frauenfelder, 195
frequency spectrum, 8, 10, 14, 44,
 47, 64, 77, 168, 173, 178,
 192
 and goodness-of-fit, 118, 119
 expected, 92
 relative, 90, 92, 94, 100
smoothing, 75, 79
variance of, 121
Yule-Simon, 113
Zipf, 95, 97
- Gale, 20, 58, 61, 76, 218
Gamma distribution, 134
Gani, 211
- generalized inverse Gauss-Poisson
 model
 parameter-adjusted, 185
Good, 35, 58, 75, 76, 120, 121
Good-Turing estimates, 57, 58, 201,
 218, 229
 affected by non-randomness, 168
goodness of fit, 21, 118, 123, 228,
 229, 231
Grotjahn, 125, 186, 195, 196
growth curve
 hapax legomena, 187
 spectrum elements, 185
 vocabulary, 2, 29, 53, 99, 163,
 166, 167, 174, 176, 185, 187,
 228
growth rate of vocabulary, 49, 132,
 172, 173, 191, 204, 205, 228
 by-token, 192
Guiraud, 25, 34, 211
Guitér, 35
- Hammerl, 195
Hammersley, 35
Handscomb, 35
hapax legomena, 8, 13, 14, 17, 27,
 33, 36, 50, 51, 53, 55, 65,
 92, 100, 172, 173, 175, 185,
 204, 205, 221, 228
- Heller, 124, 133, 214
Hellwig, 195
Herdan, 29, 32, 35, 95, 114, 133
Herdan's law, 30
Holmes, 211
Honoré, 27, 35
Hubert, 174, 175, 193, 267
hypergeometric
 distribution, 66
 probability, 66
- In 't Veld, 121, 193
indicator operator, 41
interpolation, 63, 64, 173
 binomial, 64, 65, 174
hypergeometric, 67
parameter-adjusted LNRE, 179
partition-adjusted LNRE, 176
Poisson, 73

- inverse Gauss-Poisson model, 89, 218
- Johnson, 76
- Juritz, 124, 133
- Kageura, 210
- Kalinin, 76
- key words, 165
- Khmaladze, 35, 55, 56, 76, 85, 90, 94, 113, 133
- Koehler, 195, 196
- Kotz, 76
- Kruskal, 58
- Lánský, 133
- Labbé, 267
- Labbe, 174, 175, 193
- Landauer, 195
- law of large numbers, 51
- Lebart, 35
- Leimkuhler, 133
- lexical specialization, 162, 164, 175, 176
- Lieber, 160
- link function, 183, 189
- LNRE, 55
- distribution, 5, 54, 57
 - models
 - adjusted, 173
- LNRE zone, 51, 55, 57, 100, 101
- central, 56
 - late, 56
- log-likelihood, 124
- lognormal model, 32, 37, 82, 84
- density function, 84
 - structural token distribution, 84, 87
 - structural type distribution, 85, 87
- Luke, 90
- Madigan, 195
- Makov, 160
- Mandelbrot, 17, 37, 88, 95, 133
- Martindale, 211
- McKenzie, 211
- Mead, 123
- mean
- comparison of, 132
 - lognormal, 85
 - estimation of, 86
 - token distribution, 86
 - type distribution, 86
 - log frequency, 33, 34
 - Monte Carlo, 6, 30
 - population, 87
 - probability, 90
 - probability weighted, 40
 - sample, 87
 - sample relative frequency, 7
 - word frequency, 4, 5, 34
- mean squared error, 21, 124, 176, 185
- relative, 79, 80
- mean type frequency, 25
- Menard, 35
- Meyer, 35
- Miller, 133
- mixture models, 135, 139
- model
- inverse Gauss-Poisson, 89
 - lognormal, 32, 82, 84
 - Waring-Herdan-Muller, 114
 - Yule-Simon, 107
 - Zipf, 93
- Monte Carlo methods, 7, 30, 35, 165, 183
- Morrison, 120, 133
- Mosteller, 214
- Muller, 34, 76, 95, 114, 118, 211
- Nadas, 76
- Naranan, 21, 61, 124, 129, 160, 170, 214, 221, 229
- Neijt, 145, 206
- Nelder, 123
- non-randomness, 6, 7, 34, 36, 168, 173, 179, 188, 191, 192
- consequences of, 167
- normal quantile-quantile plot, 33, 82
- Orlov, 35, 80, 94
- overestimation bias, 162
- and discourse, 163

- and syntax, 163
 and underdispersion, 165, 166
 for vocabulary size, 161
- Paivio, 195
 parameter estimation, 123, 231
 generalized inverse Gauss-Poisson,
 92
 lognormal, 86
 Zipf, 97, 99
 permutation, 41
 of sentence order, 163
 of word order, 7, 30
 run, 7, 30, 100, 165, 183
- Plag, 210
 Poisson
 approximation to Binomial, 45
 probability, 45
- Polman, 215
 population number of types, 42, 45,
 53, 58, 85, 89, 92, 99, 113,
 121, 147, 151, 167, 206
- population probabilities, 47, 51
 power model, 20
 Price, 133
 productivity, 145, 147, 154, 157, 158,
 203
- Radil-Weiss, 133
 randomization test, 7, 29, 30, 100,
 163, 183
 randomness assumption, 1, 6, 22,
 100, 161, 162, 166, 169, 188,
 192
- rank
 Zipf, 13
 rank-frequency relation, 13, 88
 rank-frequency step function, 16
 real-valued approximate spectrum,
 172
- real-valued spectrum, 20, 61
 Reder, 195
 Renouf, 61, 206
 Rice, 76
 Riemann Zeta function, 15
 Ross, 39
 Rouault, 95, 118
- Salem, 35
 sample size, 1, 2, 5, 17, 22, 25, 30,
 33, 34, 51, 55, 79–81, 86,
 161, 176
- Sampson, 20, 76, 218
 Schreuder, 146, 195, 201, 249
 Shapiro, 78
 Sichel, 27, 89, 118, 123, 133, 211
 Simon, 94, 95, 133, 223
 Simpson, 25, 46, 211
 Smith, 160
 smoothing, 21, 63, 75, 202, 215, 217,
 221, 231
- spectrum, see frequency spectrum
 standardized frequency index, 245
 Stein, 124, 133
 Streeter, 195
 structural distribution, 35, 86, 192
 structural token distribution, 83
 lognormal
 robustness of, 88
 structural type distribution, 47, 78,
 81, 94
 Gamma, 246
 inverse Gauss-Poisson, 82, 89
 lognormal, 82, 85
- Taft, 201
 Taylor series, 69, 72
 Thisted, 35, 75, 76, 78, 245
 Tiersma, 249
 Titterington, 160
 token, 2
 Toulmin, 35, 75, 76, 120, 121
 triangle scheme, 81
 Tukey, 167, 176
 Tuldava, 129
 Tweedie, 35, 146, 160, 193, 214
 type, 2
 type-token ratio, 25, 37, 133
- underdispersion, 164
 and key words, 165
 uniform distribution, 34, 175, 176
 urn model, 42, 43, 163, 166
- Van Halteren, 214
 variance, 40

vocabulary

growth rate, 50
size, 2, 4, 27, 47, 51, 52, 54, 55,
73, 96, 101, 121, 161, 166,
173, 179, 184
generalized inverse Gauss-Poisson,
92
lognormal, 86
Yule-Simon, 113
Zipf, 98

Wallace, 214

Waring, 114

Whissell, 211

Wimmer, 195, 196

word frequency list, 3

Yeh, 133

Yuille, 195

Yule, 24, 35, 46, 94, 95, 211

Z-score, 132

Zhu, 195

Zipf, 13, 16, 17, 35, 94

Zipf rank, 13, 16, 37

Zipf size, 80, 94, 183

Zipf's law, 15, 79–81, 94, 97, 100

extended, 99, 186

parameter-adjusted, 182–184, 188

partition-adjusted, 178

Zucchini, 124, 133

Text, Speech and Language Technology

1. H. Bunt and M. Tomita (eds.): *Recent Advances in Parsing Technology*. 1996 ISBN 0-7923-4152-X
2. S. Young and G. Bloothooft (eds.): *Corpus-Based Methods in Language and Speech Processing*. 1997 ISBN 0-7923-4463-4
3. T. Dutoit: *An Introduction to Text-to-Speech Synthesis*. 1997 ISBN 0-7923-4498-7
4. L. Lebart, A. Salem and L. Berry: *Exploring Textual Data*. 1998 ISBN 0-7923-4840-0
5. J. Carson-Berndsen, *Time Map Phonology*. 1998 ISBN 0-7923-4883-4
6. P. Saint-Dizier (ed.): *Predicative Forms in Natural Language and in Lexical Knowledge Bases*. 1999 ISBN 0-7923-5499-0
7. T. Strzalkowski (ed.): *Natural Language Information Retrieval*. 1999 ISBN 0-7923-5685-3
8. J. Harrington and S. Cassiday: *Techniques in Speech Acoustics*. 1999 ISBN 0-7923-5731-0
9. H. van Halteren (ed.): *Syntactic Wordclass Tagging*. 1999 ISBN 0-7923-5896-1
10. E. Viegas (ed.): *Breadth and Depth of Semantic Lexicons*. 1999 ISBN 0-7923-6039-7
11. S. Armstrong, K. Church, P. Isabelle, S. Nanzi, E. Tzoukermann and D. Yarowsky (eds.): *Natural Language Processing Using Very Large Corpora*. 1999 ISBN 0-7923-6055-9
12. F. Van Eynde and D. Gibbon (eds.): *Lexicon Development for Speech and Language Processing*. 2000 ISBN 0-7923-6368-X; Pb: 07923-6369-8
13. J. Véronis (ed.): *Parallel Text Processing*. Alignment and Use of Translation Corpora. 2000 ISBN 0-7923-6546-1
14. M. Horne (ed.): *Prosody: Theory and Experiment*. Studies Presented to Gösta Bruce. 2000 ISBN 0-7923-6579-8
15. A. Botinis (ed.): *Intonation*. Analysis, Modelling and Technology. 2000 ISBN 0-7923-6605-0
16. H. Bunt and A. Nijholt (eds.): *Advances in Probabilistic and Other Parsing Technologies*. 2000 ISBN 0-7923-6616-6
17. J.-C. Junqua and G. van Noord (eds.): *Robustness in Languages and Speech Technology*. 2001 ISBN 0-7923-6790-1
18. R.H. Baayen: *Word Frequency Distributions*. 2001 ISBN 0-7923-7017-1

TEXT, SPEECH AND LANGUAGE TECHNOLOGY
Series Editors: Nancy Ide and Jean Véronis

WORD FREQUENCY DISTRIBUTIONS

R. Harald Baayen

This book is a comprehensive introduction to the statistical analysis of word frequency distributions, intended for computational linguists, corpus linguists, psycholinguists, and researchers in the field of quantitative stylistics. Word frequency distributions are characterized by very large numbers of rare words. This property leads to strange phenomena such as mean frequencies that systematically change as the number of observations is increased, relative frequencies that even in large samples are not fully reliable estimators of population probabilities, and model parameters that vary with text or corpus size. Special statistical techniques for the analysis of distributions with large numbers of rare events can be found in various technical journals. The aim of this book is to make these techniques more accessible for non-specialists, both theoretically, by means of a careful introduction of the underlying probabilistic and statistical concepts, and practically, by providing a program library implementing the main models for word frequency distributions discussed.

