Journal of Quantitative Linguistics
1998, Vol. 5, No. 3, pp. 145–154

0929-6174/98/0503-0145$12.00

Sample-Size Invariance of LNRE Model Parameters: Problems and Opportunities *

R. Harald Baayen and Fiona J. Tweedie
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands and
University of Glasgow, United Kingdom

ABSTRACT

A well-known problem in the domain of lexical statistics concerns the dependence of measures of lexical richness on the text size. To avoid this dependence, Large Number of Rare Event (LNRE) models have been developed, the parameters of which are, in theory, sample-size invariant. In practice, however, the parameters of LNRE models may nevertheless reveal considerable dependence on the text size. We show that for LNRE models this dependence is a direct consequence of two factors: the non-random use of words in texts on the one hand, and a theoretical lack of goodness-of-fit on the other hand. We describe how we can use this dependence to our advantage to enhance the interpolation and extrapolation accuracy of LNRE models. In addition, we outline methods for carrying out cross-text comparisons using the empirical developmental profiles of LNRE parameters.

INTRODUCTION

The number of different word types $V(N)$ observed for N word tokens is an increasing function of N. This dependence of the vocabulary size $V(N)$ on the sample size N makes it impossible to use type counts as sample-size invariant characteristics of texts or corpora. Not surprisingly, various alternative measures have been put forward as sample-size invariant point statistics of lexical richness. Unfortunately, in practice all of these measures vary systematically with N (see Tweedie & Baayen, 1998, for a review).

It is well-known that the free parameter of Zipf's law in its original form (Zipf, 1935, 1949) is similarly subject to this dependence on the sample size, as shown by Orlov (1983a, 1983b). However, Orlov and Chitashvili (1982a, 1982b, 1983a, 1983b) developed an extension of Zipf's law in which this dependence is accounted for in a principled way by means of an additional parameter Z, the unique sample size for which Zipf's law in its simple form holds. The extended Zipf's law belongs to the class of LNRE models, models for distributions with Large Numbers of Rare Events. Other LNRE models are Carroll's lognormal model (Carroll, 1967) and Sichel's inverse Gauss-Poisson model (Sichel, 1986; see Chitashvili & Baayen, 1993, for a review).

The parameters of LNRE models are **in theory** invariant with respect to the sample size. The problem addressed in this paper is that **in practice** the parameters of LNRE models may nevertheless reveal substantial dependence on N. In the following section we focus on the sources of this systematic variation in the values of LNRE parameters, and we will outline how we can put this dependency to advantage to enhance the accuracy of the interpolation and extrapolation predictions of LNRE models.

We can also put the empirical dependence of LNRE parameters on the sample size to advantage for the comparison of texts of different lengths. Instead of comparing texts on the basis of a single value for a given parameter, we can now compare them on the basis of their devel-

*Address correspondence to: R. Harald Baayen, Max Planck Institute for Psycholinguistics, Wundtlaan 1, NL 6525 XD Nijmegen, The Netherlands. Tel.: +31-24-3521510. Fax: +31-24-3521213. E-mail: baayen@mpi.nl.

opmental profiles, which provide a much richer source of information. One method for doing so is outlined in the third section.

LNRE PARAMETERS AND SAMPLE SIZE

The upper panels of Figure 1 show that the parameter Z of the extended Zipf's law (Orlov, 1983a) and the parameter b of the inverse Gauß-Poisson law (Sichel, 1986) are no exception to the observation that most measures advanced as independent of the text length N tend to vary systematically with N. The horizontal axes of Figure 1, which is based on Carroll's *Alice's Adventures in Wonderland*, display the sample size N. The vertical axes of the upper panels display the values of Z (left) and b (right) as a function of N. The dots show the observed, empirical values of these parameters as estimated for the sequence of 20 equally-spaced text lengths $N = 1326, 2652, ..., 25180, 26505$ on the basis of the frequency spectra at these points in 'sample time'. For the extended Zipf's law, we find that the estimated value of Z increases with N. For the inverse Gauss-Poisson model, b also reveals considerable variation, especially for small N. The estimates of b, however, tend to converge relatively quickly to its final value as estimated for the complete text.

The solid lines in the upper panels of Figure 1 show the expected values of Z and b as calculated on the basis of a series of 5000 randomisations of the words in *Alice's Adventures in Wonderland*. Again, we observe a clear pattern of dependence on the sample size N. In the case of Z, we see an initial steep decline, after which Z stabilises, albeit with some very small concave curvature. In the case of b, we find a convex function that levels off by the end of the text.

Why do the empirical and theoretical developmental profiles of Z and b show this dependence on N? First consider the theoretical dependence as revealed by the Monte Carlo simulations. Does this dependence imply that LNRE models fail to eliminate the ubiquitous dependence on N which they were designed to overcome, not only in practice, but also in theo-

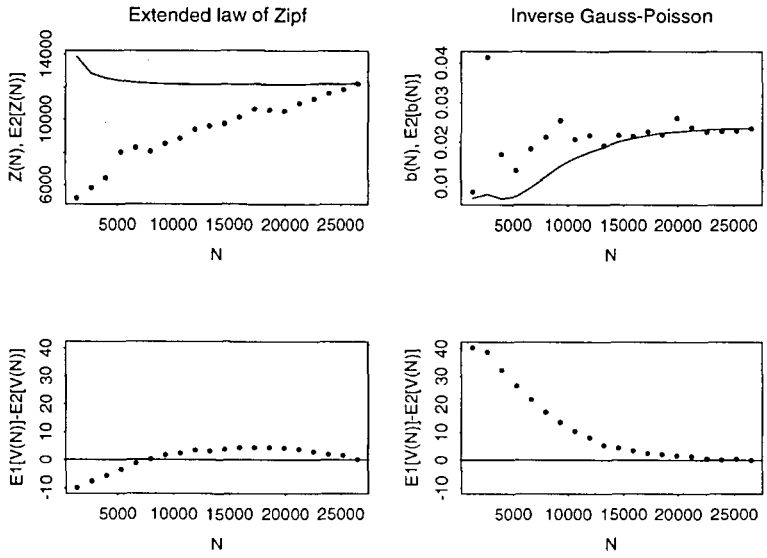

Fig. 1. Observed (dots) and Monte-Carlo-based expectations (solid lines) for Z (upper left) and b (upper right), and the error for the vocabulary size for the extended Zipf's law (bottom left) and the inverse Gauß-Poisson law (bottom right) as calculated for *Alice's Adventures in Wonderland*, measured at 20 equally-spaced intervals. $E1$: expectation based on the model for the full text ($E_{LNRE}[V(N)]$); E2: Monte Carlo expectation ($E_{MC}[V(N)]$).

ry? Fortunately, this is not the case. The changing values of Z and b are a direct consequence of imperfections in the fit of the LNRE models to the empirical frequency spectrum of *Alice's Adventures in Wonderland*. To see this, consider the bottom panels of Figure 1, which plot the difference between two theoretical growth curves, $E_{LNRE}[V(N)]$ and $E_{MC}[V(N)]$. The expected vocabulary growth curve $E_{LNRE}[V(N)]$ is obtained by estimating the parameters of the models for the complete text (N = 26505), followed by interpolation of the expected vocabulary size for 20 equally-spaced smaller sample sizes using

$$E_{LNRE}[V(N)] = \frac{Z}{\log(p^*Z)} \frac{N}{N-Z} \log(N/Z) \quad (1)$$

for the extended Zipf's law (with p^* the maximum relative frequency in the text), and using

$$E_{LNRE}[V(N)] = \frac{2}{bc} \left[1-e^{b(1-\sqrt{1+Nc})}\right] \quad (2)$$

for the inverse Gauss-Poisson law (with c the second parameter of the model, and fixing its third parameter, γ, at -0.5 a priori). The second expected vocabulary growth curve, $E_{MC}[V(N)]$, is obtained by calculating the average vocabulary size in a series of 5000 permutation runs, i.e. Monte Carlo expectations, for the same 20 measurement points in sample time. Since both expectations are based on the urn model, their values should be identical. Hence, their difference $E_{LNRE}[V(N)] - E_{MC}[V(N)]$ is a diagnostic for how well an LNRE model fits the basic urn model.

For the extended Zipf's law, we observe a convex curvature. For small N, the model underestimates $V(N)$, for medium N, it reveals a slight overestimation bias compared to the Monte Carlo expectations. Since increasing Z leads to an increase in $E_{LNRE}[V(N)]$ by (1), the underestimation bias of $E_{LNRE}[V(N)]$ observed for small N is compensated for by increasing Z when estimating this parameter for small sample sizes in the randomisations. For larger sample sizes, the mismatch between $E_{LNRE}[V(N)]$ and $E_{MC}[V(N)]$ is so small that the value of Z is hardly affected, and approaches constancy.

Turning to the inverse Gauss-Poisson law, we find an overestimation bias for $E_{MC}[V(N)]$ that decreases with increasing N. This model accommodates its overestimation bias by increasing c and by decreasing b as N becomes smaller. Compared to Z, the value of b becomes reasonably stable at a rather late moment in sampling time N. This is due to the larger bias of $E_{LNRE}[V(N)]$ for the inverse Gauss-Poisson model. Consequently, greater changes in the parameters are required to accommodate the model to the structure of the frequency spectra of the smaller sample sizes in the Monte Carlo simulations.

Since the observed dependence of LNRE parameters on the sample size directly reflects the accuracy of LNRE models, we can make use of this dependence to evaluate the goodness-of-fit of these models. Traditionally, the goodness-of-fit of LNRE models is evaluated by means of chi-square tests. Unfortunately, the appropriate chi-square test (using the covariance matrix of the spectrum elements) often leads to the rejection of theoretical models with p-values that may be as small as 10^{-8} (see also Grotjahn & Altmann, 1993), even when fits are obtained that are perfectly reasonable to the eye. Instead of using the chi-square test, the extent to which LNRE parameters change as a function of N can be used as a measure of goodness-of-fit: the less accommodation required, the better the fit of the model. As a practical measure, we propose to use the percentage of measurement points for which the absolute error

$|E_{LNRE}[V(N)] - E_{MC}[V(N)]|$

falls below a given tolerance threshold δ:

$$D(K, \delta) = \frac{1}{K} \sum_{i=1}^{K} I\left[\left|E_{LNRE}[V(N_k)] - E_{MC}[V(N_k)]\right| < \delta\right], \quad (3)$$

with K the number of measurement points (20 in our examples) and $V(N_k)$ the vocabulary size at the k^{th} measurement point, and with $I[\cdot]$ the indicator operator. We choose the model error δ as small as possible, but such that the proportions $D(K, \delta)$ for the two models differ significantly. For the extended Zipf's law and the in-

verse Gauss-Poisson law, the smallest significant difference is found for $\delta = 5$: $D(20,5) = 0.85$ for Z, and $D(20,5) = 0.5$ for b ($p < 0.05$). These calculations formalise the visual impression of Figure 1 that the extended Zipf's law provides the better fit to *Alice's Adventures in Wonderland*, even though it has only one parameter to vary instead of two.

The above test for goodness-of-fit pits the predictions of LNRE models against the predictions of the urn model (without replacement). However, words do not occur randomly in texts. As illustrated in the upper panels of Figure 1, the observed values of the LNRE parameters diverge from their Monte Carlo expectations for a wide range of measurement points. This is due to the non-random, underdispersed use of words in discourse, which, in the case of *Alice's Adventures in Wonderland*, causes the empirical vocabulary size to be substantially smaller than its theoretical expectation for all 20 measurement points (see Baayen, 1996, for detailed discussion). In the case of the extended Zipf's law, the overestimation bias of the theoretical estimates is compensated for when we estimate Z for smaller text lengths using (1). In order to match the expected and the observed vocabulary size, we have to lower $E_{LNRE}[V(N)]$ compared to what we would expect given the complete text, and hence Z has to be lowered too. In the case of the inverse Gauss-Poisson law, the parameters b and c are likewise adjusted to meet the requirement that for each measurement point the expected vocabulary size should be equal to its expectation given the frequency spectrum at that measurement point. Note that, in fact, the empirical developmental profile of, for instance, Z, locally optimises the model not only with respect the effects of non-randomness in word use, but also with respect to the slight misfit compared to the urn model itself.

Interestingly, the developmental profiles of Z and b themselves reveal a fairly regular dependence on the sample size N. In the case of Z, the functional dependence of Z on N might be captured by a power function

$$Z(N)=a_1 N^{a2}. \tag{4}$$

For our text, least squares estimation suggests $a_1 = 617.69$ and $a_2 = 0.290$ (with $R^2 = .981$). We can now replace Z in (1) by the link function $Z(N)$:

$$E[V(N)] = \frac{Z(N)}{\log(p^*Z(N))}\frac{N}{N-Z(N)}\log(N/Z(N)) = \frac{a_1 N^{a2+1}}{\log(p^*a_1 N^{a2})}\frac{1}{1-a_1 N^{a2-1}}\log\left(\frac{N^{1-a2}}{a_1}\right) \tag{5}$$

In this way we change a one-parameter model into a model with two free parameters. Another link function that we have found to be useful for some texts is the linear function $Z(N) = a_1 + a_2 N$.

Figure 2 shows the gain in interpolation and extrapolation accuracy obtained by adjusting Orlov and Chitashvili's Zipfian LNRE model with the link function $Z(N)$. The dots represent the observed values of the vocabulary size $V(N)$ and the first five spectrum elements $V(m, N)$ ($V(1, N)$ represents the number of hapax legomena; $V(2, N)$ the number of dis legomena, etc.). The dotted lines were obtained using (1) with Z estimated at exactly half the length of the complete text. Note that interpolation leads to a slight overestimation of the vocabulary size, whereas extrapolation leads to substantial underestimation. The solid lines in the left panel represent adjusted interpolation and extrapolation from the middle of the text using a power function for the link $Z(N)$. For *Alice's Adventures in Wonderland*, the adjusted LNRE clearly is considerably more accurate, although it reveals a slight overestimation bias for extrapolation that can be traced to a slight flaw in the goodness of fit of the power model to the developmental profile of Z. For Wells' *The War of the Worlds*, however, a linear link function for $Z(N)$ yields quite satisfactory precision for both interpolation and extrapolation (the solid lines in the right panel of Figure 2). We conclude that the adjustment of LNRE models by means of link functions seems promising as a means to obtain models that not only provide good fits to frequency spectra, but that are also accurate with respect to interpolation and extrapolation.

COMPARING DEVELOPMENTAL PROFILES OF LNRE PARAMETERS

We have seen that the parameters of LNRE models are not textual constants but that they vary systematically with the text length. We have used this variability to our advantage to enhance the accuracy of LNRE models. In this section, we will show how we can also employ this variability to enhance comparisons of texts based on LNRE parameters.

Consider Figure 3, which illustrates how variable the parameters Z (left panel) and b (right panel) are across a range of texts. The texts that we have investigated for this study are listed in the Appendix, they include children's books by Carroll ($a1$, $a2$) and Baum ($b1$, $b2$), novels by Wells ($w1$, $w2$), London ($l1$, $l2$), James ($j1$, $j2$), and Conan Doyle ($c1$, $c2$), as well as two books from the King James version of the Bible, the gospel according to St. Luke and the Acts of the Apostles, also held to be written by St. Luke ($L1$, $L2$). Note that there is some authorial structure in the plots. For instance, the texts by Carroll ($a1$, $a2$) have very similar developmental profiles for $Z(N)$, and the same holds to some extent for those by James ($j1$, $j2$).

At the same time, it will be clear that the within-author variation is by no means smaller than the between-author variation. For instance, the texts by Baum ($b1$, $b2$) are separated by the texts by Carroll, Luke, and James. There is also some evidence for genre differences: the children's books tend to have lower values for $Z(N)$ (they are less rich in vocabulary), and they show up with higher values for $b(N)$.

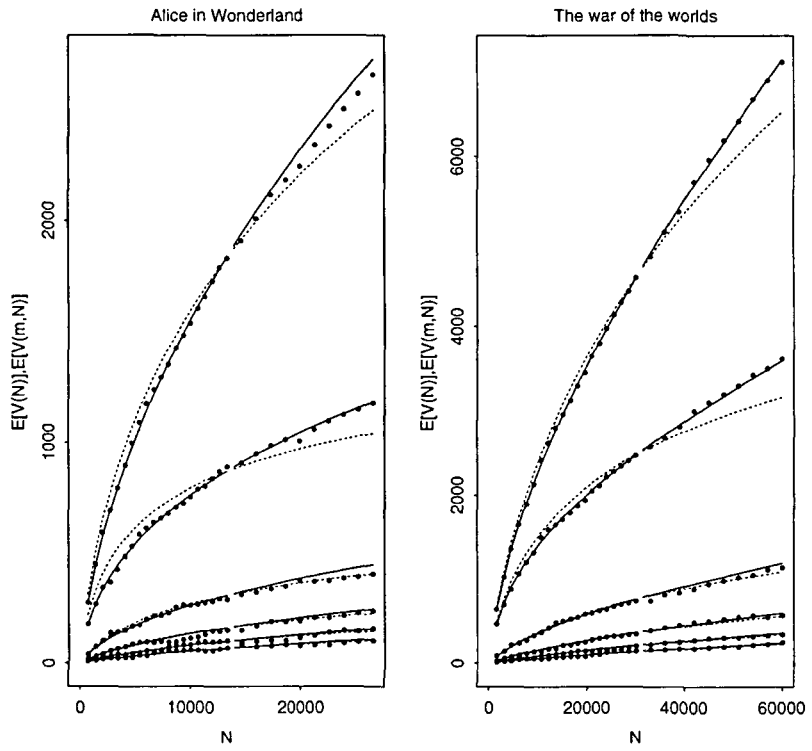

Fig. 2. Interpolation and extrapolation accuracy for Carroll's *Alice's Adventures in Wonderland* (left) and Wells' *The War of the Worlds* (right). The observed development of the vocabulary size $V(N)$ and that of the first 5 spectrum elements $V(m, N)$ are represented by dots. The dotted lines show unadjusted interpolation and extrapolation from the middle of the text, the solid lines show the corresponding adjusted curves using the link function $Z(N)$.

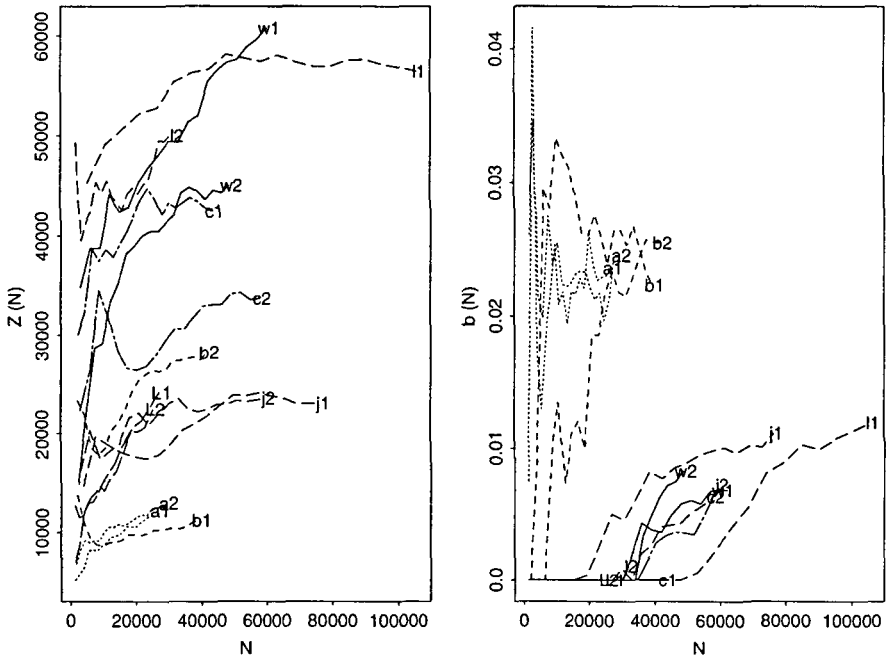

Fig. 3. Empirical values of Z and b as functions of N in selected texts. $a1$, $a2$: Carroll; $b1$, $b2$: Baum; $c1$, $c2$: Conan Doyle; $j1$, $j2$: James; $l1$, $l2$: London; $L1$, $L2$: Luke, Acts (*KJV*); $w1$, $w2$: Wells. See the Appendix for bibliographical details.

The question that arises when looking at these plots is which differences are statistically significant. For example, the texts by James ($j1, j2$) have similar final values for $Z(N)$ as the texts by Luke ($L1$, $L2$). Judged in terms of these final values, the texts by Luke and James cannot be distinguished. Their developmental profiles, however, are quite different, which suggests they might differ significantly in the way in which this final value is reached. In what follows, we will use analysis of variance techniques to investigate such similarities and differences between texts and authors, focusing on $Z(N)$.[1] The same methodology can be applied to other lexical measures as well, Yule's K (Yule, 1944) among others.

We first consider the question how to ascertain whether texts have significantly different developmental profiles. To do so, we use a model of the form

$$Z_{ik} = \mu + \alpha_1 + \gamma_k + \in_{ik}, \qquad (6)$$

where Z_{ik} represents the observed values of Z in text i, $i = a1, a2, ..., w2$ at measurement point k, $k = 1, ..., K$, here $K = 20$. The value of μ is roughly the grand mean, the mean of all observations of all texts jointly. The α_i terms can be thought of as text effects, they present the deviations from the grand mean for each individual text. The γ_k terms describe the changes as repeated measures are taken through the texts.[2] The $\in_{ik}$ terms are error terms, with the usual

[1] In the right panel of Figure 3, there are a fair number of cases with zero values for b. These result form the absence of a fit of the inverse Gauss-Poisson model to the text at the given length, in which case b is set to zero. Hence, we have not carried out any further analyses using these data.

[2] Technically, α_{a1} and γ_1 are constrained to equal zero, their actual values being absorbed into μ. Note that this model presupposes that the repeated measures develop in parallel for the various texts in our sample, as it does not include an interaction term. Hence, the present model is inappropriate for $b(N)$ in Figure 3, where the texts by Carroll and Baum do not develop in parallel with the texts by the other authors.

normality, independence and zero mean assumptions. Note that we are comparing the texts at each measurement point, i.e., the first recorded value of Z, $Z_{a1,1}$ is compared with other first values of $Z_{i,1}$, $Z_{a2,1}$, ..., $Z_{w2,1}$, regardless of the values of N_1 in each text.

Fitting model (6) to the values of Z_{ik} resulted in significant Text and Repeated Measures factors ($p << .0001$ for both, using the Most-Conservative Test; Geisser and Greenhouse, 1958). The Multiple R^2 value is 97.14%, indicating that a very good fit to the data has been achieved, with only 2.86% of the variation not being explained by (6). To determine which of the texts are significantly different from which other texts we construct t-tests and Bonferroni-corrected confidence intervals. Performing separate tests and obtaining 95% confidence intervals for each possible difference between texts would increase the overall Type I error, the probability that a significant difference is observed without it being actually present. For example, if we are interested in the differences between three texts, then there are three possible comparisons; A-B, A-C, and B-C. If we were to test these individually then, rather than having overall confidence intervals of 95%, we only have a confidence level of $0.95^3 = 0.857$. The Bonferroni adjustment takes the form of adjusting the probabilities in the usual t-statistic, by dividing the desired confidence level by the number of possible comparisons (1/2 (3*2)). For the three texts A, B and C, the desired probability level for two-tailed tests would be:

$$1 - \frac{\frac{1}{2}\ (1 - 0.95)}{\frac{1}{2}\ (3 * 2)} = 0.99167.$$

Using the Bonferroni adjustment ensures that the Type I error taken over all 14*13 textual comparisons for our data does not exceed 5%. The calculation of the confidence intervals is as follows:

$$\overline{Z_{i.}} - \overline{Z_{i.}'} \pm t\left((I-1)(K-1), 1 - \frac{\frac{1}{2}\ (1-c)}{\frac{1}{2}\ (I(I-1)}\right)\sqrt{\frac{2}{K} * MS_{RES}}, \quad (7)$$

where I is the number of texts under consideration, here 14, c is the desired confidence level, here 95%, and MS_{RES} the residual mean squared error, here 7630566. Hence,

$$\begin{aligned}\overline{Z_{i.}} - \overline{Z_{i.}'} &\pm t(247, 0.9997)\ \sqrt{2/20 * 7630566} \\ &\pm 3.47 * 873.53, \\ &\pm 3031.15.\end{aligned}$$

We present the results by positioning the texts in increasing order of their means, and by drawing lines under texts which do not have significantly different developmental profiles, as shown in Table 1. It can be seen that texts $b1$, $a1$ and $a2$ are not significantly different; neither are texts $L2$ and $L1$, $j1$, $j2$, and $b2$, or $w2$ and $c1$. These groups as well as the remaining texts are all significantly different from each other.

It is clear that some level of authorial structure is present. At the same time, two texts by a given author can be significantly different, e.g., $w1$ and $w2$, $b1$ and $b2$, or $l1$ and $l2$. Interestingly, the texts by James ($j1$, $j2$) are distinguished from those by Luke ($L1$, $L2$), even though the values of Z for the complete texts are quite similar. Apparently, the narrative organisation of the Luke texts differs sufficiently from that of the James texts to give rise to reliably different developmental profiles of Z.

Thus far we have considered a model in which each text is treated separately; the fact that some texts have the same author is not taken into account, and tracing authorial structure is left to the analyst. We can investigate explicitly to what extent authors produce texts that

Table 1. Results from fitting the model $Z_{ik} = \mu + \alpha_i + \gamma_k + \in_{ik}$.

$b1$	$a1$	$a2$	$L2$	$L1$	$j1$	$j2$	$b2$
10129	10242	10585	16644	17201	20918	21578	24263
$c2$	$w2$	$c1$	$l2$	$w1$	$l1$		
30433	37813	40625	45384	49506	55317		

Note. Horizontal lines group texts that are not significantly different.

differ with respect to vocabulary richness as measured by Z by introducing a new element to our model, $\beta_{j(i)}$:

$$Z_{ij(i)k} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \epsilon_{ij(i)k}. \quad (8)$$

Here $Z_{ij(i)k}$ represents the observed values of Z in text j, $j = 1,2$ by each author i, i = a, b, ..., w at measurement point k, $k = 1, ..., 20$. As before, the value of μ is a grand mean, and the γ_k terms describe the changes as repeated measures are taken through the text. In this model the α_i terms can be thought of as author effects while the $\beta_{j(i)}$ represent text effects nested within authors. When this model is fitted to the observed values of $Z_{ij(i)k}$, the Text factor in (6) is partitioned into an Author factor and a Text within Author factor (with 6 and 7 degrees of freedom respectively). Both of these factors are extremely significant ($p << .0001$). The Multiple R^2 value and the MS_{RES} remain the same as the nesting has not altered the amount of variation explained by the model, it has only partitioned it differently. The nested structure alters the degrees of freedom and the standard error in the Bonferroni-adjusted confidence intervals. We now calculate the confidence intervals as follows:

$$\overline{Z_{i..}} - \overline{Z_{i..}} \pm t\left(I(J-1), 1-\frac{\frac{1}{2}(1-c)}{\frac{1}{2}(I(I-1)}\right)\sqrt{\frac{2}{JK} * MS_{TwA}}$$

$$\pm t(7,0.9988)\sqrt{2/40*771313560}$$
$$\pm 4.63*6210.13$$
$$\pm 28752.90,$$

where I is now the number of authors, here 7, and MS_{TwA} is the Mean Square associated with the Text within Author factor. Table 2 groups together those authors whose texts are not significantly different.

While for some authors' texts the developmental profiles of Z are quite similar, e.g., for Carroll, Luke, and James, other authors have texts with quite different developmental profiles. The substantial variation introduced by the latter authors (e.g., Baum, Wells, and Conan Doyle) leads to a large sum of squares in the nested text factor. This results in a much wider confidence interval, which in turn leads to the observed weak discriminatory power of the present model for distinguishing between authors. Indeed, our seven authors fall into two main (overlapping) groups; Carroll, St Luke, Baum and James, and Conan Doyle, versus Conan Doyle, Wells and London. This analysis complements the previous one by making clear that the within-author variation is so large that its swamps most of the between-author variation. In order to tease authors apart on the basis of their texts, many more discriminant variables need to be included in the analysis, for instance, function words (Burrows, 1992) and syntactic patterns (Baayen, van Halteren, & Tweedie, 1996). Nevertheless, our by-text and by-author analyses show that, compared to analyses based on the final values of Z, the use of developmental profiles leads to improved inference.[3]

Table 2. Results from Fitting Model $Z_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \epsilon_{ij(i)k}$.

Carroll	Luke	Baum	James	Doyle	Wells	London
10413	16922	17196	21248	35529	43659	50351

Note. Horizontal lines group authors that are not significantly different.

CONCLUSIONS

We have called attention to the dependence of the parameters of *LNRE* models on the sample size. This dependence is observed for both the empirical development of a text, as well as for its theoretical development in Monte Carlo simulations. We have first focused on the new possibilities that this finding offers for goodness-of-fit testing, and for enhancing the interpolation and extrapolation accuracy of *LNRE* models by

[3] For a randomisation-based technique for comparing developmental profiles, see Tweedie & Baayen (1998).

means of link functions.These link functions allow us to take the non-random aspect of word use at the level of discourse organization in texts into account.

Secondly, we have shown that using developmental profiles of textual measures such as the parameters of *LNRE* models can lead to improved inference. We have outlined how to test for differences between individual texts, and how to investigate authorial structure. Both analyses reveal some author-related similarities, but it is evident that in our data within-author variability is at least as large as between-author variability. Indeed, texts by the same author may turn out to be significantly different once a model is fitted. For more reliable authorship discrimination, a greater range of variables (possibly with their developmental profiles) should be taken into account.

REFERENCES

Baayen, R.H. (1996). The effect of lexical specialisation on the growth curve of the vocabulary. *Computational Linguistics, 22*, 455–480.

Baayen, R.H., Van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*, 121–131.

Burrows, J.F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing, 7*, 91–109.

Carroll, J.B. (1967). On sampling from a lognormal model of word frequency distribution. In H. Kucera & W.N. Francis (Eds.), *Computational analysis of present-day American English* (pp. 406–424). Providence RI: Brown University Press.

Chitashvili, R.J., & Baayen, R.H. (1993). Word frequency distributions. In G. Altmann & L. Hřebiček (Eds.), *Quantitative Text Analysis* (pp. 54–135). Trier: Wissenschaftlicher Verlag Trier.

Geisser, S., & Greenhouse, S. (1958). Extension of Box's results on the use of the F-distribution in multivariate analysis. *Annals of Mathematical Statistics, 29*, 855–891.

Grotjahn, R., & Altmann, G. (1993). Modeling the distribution of word length: Some methodological problems. In R. Köhler & B.B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 141–153). Dordrecht: Kluwer.

Orlov, J.K. (1983a). Dynamik der Häufigkeitsstrukturen. In H. Guiter & M.V. Arapov (Eds.), *Studies on Zipf's Law* (pp. 116-153). Bochum: Brockmeyer.

Orlov, J.K. (1983b). Ein Model der Häufigskeitstruktur des Vokabulars. In H. Guiter & M.V. Arapov (Eds.), *Studies on Zipf's Law* (pp.154–233). Bochum: Brockmeyer.

Orlov, J.K., & Chitashvili, R.J. (1982a). On some problems of statistical estimation in relatively small samples. *Bulletin of the Academy of Sciences, Georgia, 108*, 513–516.

Orlov, J.K., & Chitashvili, R.J. (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., Chitashvili, R.J. (1983a). Generalized Z-distribution generating the well-known „rank-distributions''. *Bulletin of the Academy of Sciences, Georgia, 110*, 269–272.

Orlov, J.K., & Chitashvili, R.J. (1983b). On the statistical interpretation of Zipf's law. *Bulletin of the Academy of Sciences, Georgia, 109*, 505–508.

Sichel, H.S. (1986). Word frequency distibutions and type-token characteristics. *Mathematical Scientist, 11*, 45–72.

Tweedie, F.J., & Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. Submitted.

Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zipf, G.K. (1935). *The psycho-biology of language*. Boston: Houghton Mifflin.

Zipf, G.K. (1949). *Human behavior and the principle of the least effort. An introduction to human ecology*. New York: Hafner.

APPENDIX

Author	Title	Key
Baum, L.F.	*The Wonderful Wizard of Oz*	*b*1
	Tip Manufactures a Pumpkinhead	*b*2
Carroll, L.	*Alice's Adventures in Wonderland*	*a*1
	Through the Looking-glass and what Alice Found There	*a*2
Conan Doyle, A.	*The Sign of Four*	*c*1
	The Valley of Fear	*c*2
James	*Confidence*	*j*1
	The Europeans	*j*2
St Luke	*Gospel according to St Luke (KJV)*	*L*1
	Acts of the Apostles (KJV)	*L*2
London, J.	*The Sea Wolf*	*l*1
	The Call of the Wild	*l*2
Wells, H.G.	*The War of the Worlds*	*w*1
	The Invisible Man	*w*2