When word frequencies do not regress towards the mean

R. Harald Baayen, Fermín Moscoso del Prado Martín, Robert Schreuder and Lee Wurm

Ever since Gernsbacher (1984), it is widely believed that word frequency counts based on corpora are unreliable, particularly for the highest and lowest frequency words due to regression towards the mean. In this study, however, we show that word frequency counts across corpora are **not** subject to regression towards the mean, neither in theory nor in practice. Sampling error due to underdispersion, however, remains a serious concern.

1. Introduction

Several studies addressing frequency effects in morphological processing have made use of factorial designs contrasting high and low frequency words (e.g., Taft, 1979; Sereno and Jongman, 1997; Baayen, Dijkstra, and Schreuder, 1997). However, the use of such designs has been questioned on the grounds of the unreliability of both very low and very high word frequencies. The problem is described by Gernsbacher (1984) as follows:

> Acknowledging the potential unreliability of printed frequency, several have suggested that these probable confounds are due to regression towards the mean, that is, the statistical probability that with a different sample of an independent variable, the extreme points on a normal distribution will assume a "truer" value, one closer to the mean of that distribution ... Regression towards the mean is particularly probable when two highly correlated variables are factorially combined and when the measurement of either independent variable is noisy. Arranging groups of stimuli that are extremely high or low along one variable and simultaneously extremely high or low along its covariate variable, and vice versa, is often done by capitalizing on measurement error found in either variable. Thus, though it is believed that the values of each variable are well matched within either level of the opposite variable, it is possible that their "true" values are not. (Gernsbacher 1984: 276)

Gernsbacher's paper has led many researchers studying morphological processing to abandon the study of frequency effects, and to rely predominantly or exclusively on priming paradigms in which target words are their own controls. Recently, Ford, Davis, and Marslen-Wilson (this volume) have argued that regression designs in which the words with extreme frequencies are excluded are to be preferred above factorial designs contrasting extreme frequency classes. The claim that it is predominantly the extreme frequencies that are unreliable is motivated by the argument of regression towards the mean. In this study, however, we show that regression towards the mean does not take place for word frequency distributions.

In what follows, we first describe the phenomenon of regression towards the mean, which is part and parcel of bivariate normal distributions. We then introduce the bivariate Poisson-Lognormal distribution, which is appropriate for word frequency counts. Finally, we present some empirical bivariate word frequency distributions, which illustrate the absence of regression towards the mean and the presence of sampling error (due to underdispersion) notably in the medium frequency ranges.

2. Regression towards the mean

What is regression towards the mean? The term 'regression towards mediocrity' was introduced by Galton (1822-1911) for a dataset in which the heights of sons (Y) was plotted against the heights of their fathers (X). The resulting regression line had a positive slope smaller than 1, indicating that a very tall father was likely to have a son who was less tall. Conversely, a very short father was likely to have a somewhat taller son. The heights of the sons 'regressed' towards the mean, and this gave 'linear regression' its name. Fisher (1918) showed that, given the laws of genetics, the slope of the regression line has to be less than one when the heights of sons are plotted against those of their fathers.

It is useful to consider the properties of bivariate normal distributions in more detail to understand the phenomenon of regression towards the mean. A bivariate normal distribution (X, Y) has five parameters: the means of X and Y, the variances of X and Y, and the correlation ρ between X and Y. Let's assume for the moment that word frequencies are normally distributed, and for ease of exposition, let's also assume that the means of

X and Y are 0 and that their variances are 1, i.e., X and Y are standard normal random variables. How can we simulate a dataset with frequencies from two corpora, C_X and C_Y. What we need here are two distributions, a 'general' frequency distribution and a 'sample' frequency distribution. The general distribution is the distribution of the population usage rates of the different word types (their population mean token frequencies). The sample distribution of a given word type is the distribution of the token frequencies with which that particular word type, given its associated usage rate, appears across different samples (corpora).

We can generate a bivariate standard normal sample (simulating frequency data for two corpora) by first sampling a set of population usage rates from the general distribution, that should be normally distributed with mean 0 and variance ρ: $N(0, \rho)$. Why we need the variance to equal the correlation coefficient ρ will become clear below. Subsequently, we generate individual token frequencies for the word types by sampling from their corresponding sampling distributions. For a given word ω_i with usage rate μ_i this sampling distribution has mean μ_i and variance $1 - \rho$. In other words, word token frequencies are $N(\ N(0,\ \rho),\ 1 - \rho)$-distributed. By choosing the variances of the general and sampling distributions to be ρ and $1 - \rho$, we ensure that the marginal distributions X (the frequencies from corpus C_X) and Y (the frequencies from corpus C_Y) follow a standard normal distribution. The variance of the marginal distributions is the sum of the variances of the general distribution and the sampling distribution, and the correlation between X and Y is ρ. Because we have a standard normal bivariate distribution, there is a very simple relation between the slope of the regression line and ρ: the slope is equal to ρ.

The upper left panel of Figure 1 plots an example of 1000 points from a bivariate standard normal distribution with $\rho = 0.9$ and slope 0.9. The dashed line represents the line $Y = X$, the solid line is the regression line. The lower left panel of Figure 2 shows the corresponding density. Note that most of the observations are located around the center of the plot.

An important property of bivariate normal distributions is that the amount of regression towards the mean can be expressed as a function of ρ. When ρ equals 1, i.e., X and Y correlate perfectly, there is no regression towards the mean. When X and Y are totally uncorrelated, regression towards the mean is maximal. In general, the amount of regression towards the mean equals 1-ρ, the variance of the sampling distribution.

The slope of the regression line in Figure 1 is less than 1. However, when the mean of X is less than the mean of Y, the regression line may

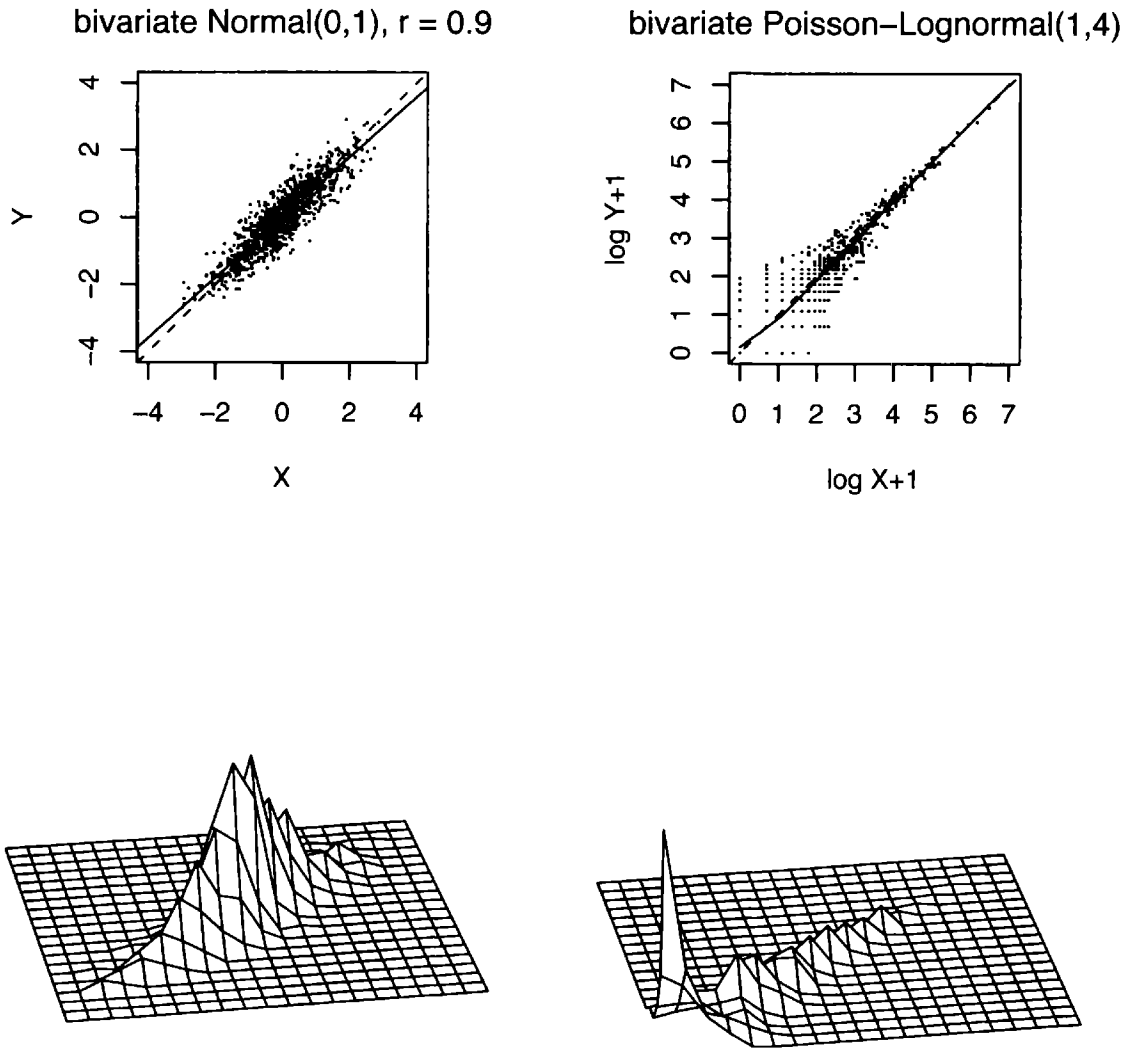

Figure 1. A bivariate standard normal density with $\rho = 0.9$ (left panels) and a bivariate Poisson-Lognormal(1, 4) density (right panels).

have a slope greater than 1. In this case, there would not be regression **towards** the mean, but regression **from** the mean. Thus, one should have an a-priori reason for expecting regression towards the mean rather than regression from the mean.

Summing up, word frequencies from two corpora will show regression towards the mean under the assumption that they are bivariate normal random variables.

3. Word frequency distributions

Word frequencies, however, are not $N(N(0, \rho), 1 - \rho)$-distributed, and it is an open question whether the phenomenon of regression towards the mean still arises. In fact, it turns out that both the general distribution and the sampling distribution of word frequencies are non-normal. There are several models for the general distribution of word frequencies (see, e.g., Baayen, 2001). One such model, that we will use here for reasons of simplicity, is the lognormal model (see, e.g., Carrol, 1967). The lognormal distribution has density

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

If X has a Lognormal(μ,σ^2) distribution, then $\log(X)$ follows a Normal(μ,σ^2) distribution. For given μ and σ, the lognormal model defines the usage rates with which individual words appear in text or speech.

To simulate a realistic data set of word frequencies from two corpora of the same size, we begin with sampling usage rates λ for the word types from a lognormal distribution with mean μ and variance σ^2.

Given these usage rates, we generate the token frequencies in a given corpus by sampling from a Poisson distribution. The Poisson distribution defines the probability that the frequency X of a word will be m given that the word has usage rate λ:

$$\Pr(X = m) = \frac{\lambda^m}{m!}e^{-\lambda}, m = 0,1,2,3,... \quad (2)$$

The Poisson distribution has only one parameter, the rate λ, which represents both the mean and the variance. In this model, a word's frequency is Poisson(Lognormal(μ, σ^2))-distributed. The right panels of Figure 1 show a scatterplot and a density plot for two random variables X and Y from a bivariate Poisson-Lognormal(1, 4) distribution. This plot was created by generating 1000 random reals from a Lognormal(1, 4) general distribution, resulting in a vector of usage rates λ_i, i = 1, 2, ..., 1000. Each usage rate λ_i defined a sampling distribution, for which two random Poisson(λ_i)-distributed integers were generated, representing the word frequencies in two different corpora of the same size. In this way we

obtained two vectors of 1000 word frequencies, pairwise sharing the same usage rate. Just as in real word frequency counts, there are many words in this simulated data set with low frequencies, and a small number of high-frequency outliers. For visualization purposes, we therefore plotted $\log(Y + 1)$ against $\log(X + 1)$ in the right panels of Figure 1, adding 1 to both X and Y in order to include the zero-frequencies in the plot.

First note that the data points closely cluster around the line $\log(Y + 1) = \log(X + 1)$, represented by a dashed line in the upper right panel. Also note that we do not have a spherical scatter as in the left panels of Figure 1, but a variance structure that decreases with increasing frequency. For the lowest frequencies we have a striated pattern in the upper right panel that is due to the discrete nature of the Poisson distribution. Unlike in the bivariate normal case, where each point (X, Y) is a combination of a unique value of X with a unique value of Y, a coordinate pair (X, Y) of a bivariate Poisson-Lognormal distribution may be instantiated by a great many words, as illustrated by the density plot in the lower right panel of Figure 1. Note that there are many words in this distribution that occur zero times in both the C_X and C_Y corpora. When comparing actual corpora, word frequency distributions are trunctated as the counts of words with zero frequency are not available.

It makes no sense to calculate the Pearson correlation for X and Y given the gross violation of the sphericity condition. Non-parametric regression lines (Cleveland, 1981) fit to bivariate Poisson-Lognormal distributions begin slightly above the main diagonal, then dip slightly under the main diagonal, and end at or slightly above the main diagonal (see the solid line in the upper right panel). What we have here, in other words, are two highly but not completely correlated variables for which the slope of the regression line is nevertheless very near to 1. This is impossible in the case of bivariate normal distributions: If the slope is 1, then ρ is also 1, and all scatter has disappeared. While in the case of bivariate normal distributions ρ is available as a parameter that regulates the extent to which Y can be predicted from X and that at the same time describes the amount of regression towards the mean, bivariate Poisson-Lognormal distributions have no such parameter. The values of X and Y for a given word ω_i with usage rate λ_i predict each other within the bounds set by λ_i, independently of whether λ_i is extremely small (near 0) or extremely large.

A scatterplot such as shown in the upper right panel of Figure 1 shows that the values of Y are very similar to those of X. What is the precise relation between Y and X? In other words, what is the expected value of Y

given a particular value of X, $E[Y \mid X]$? In the case of a bivariate standard normal distribution with slope not greater than 1,

$$E[Y \mid X] = \rho\, E[X]. \tag{3}$$

The corresponding expression for word frequency distributions can be found in Good (1953). Let N denote the size of the sample (corpus) in tokens, i.e., the summed token frequencies of all types, and let $V(m, N)$ denote the number of words that occur with frequency m in a corpus of N tokens. Good, acknowledging Turing, showed that the expected value of Y given that X has the (discrete) frequency m, $E[Y \mid X = m]$, is

$$E[Y \mid X = m] = \frac{m+1}{1+1/N} \frac{E[V(m+1, N+1)]}{E[V(m, N)]}. \tag{4}$$

Crucially, the value of $m^* = E[Y \mid X = m]$ is slightly smaller than the value of m itself for all $m \geq 1$. For corpora with 1 million word tokens, the adjustment for the highest frequency words is roughly 1 token. For words occurring once, the adjustment is from 1 to 0.68 (see Baayen, 2001: 63). Note that even for the lowest frequency words the adjustment is down, from the mean, instead of up, towards the mean. In other words, the lowest frequency words in one corpus are expected to occur with even lower frequencies in another corpus, instead of with higher frequencies.

Summing up, what this second simulation shows is that regression towards the mean is a phenomenon that does not generalize from bivariate normal distributions to other bivariate distributions. The bivariate Poisson-Lognormal distribution (and the same holds for the other distributions proposed in the literature for word frequency distributions) have a different property, namely, the Good-Turing relation. Just as regression towards the mean is the hallmark of bivariate normal distributions, the Good-Turing adjustment is the hallmark of word frequency distributions.

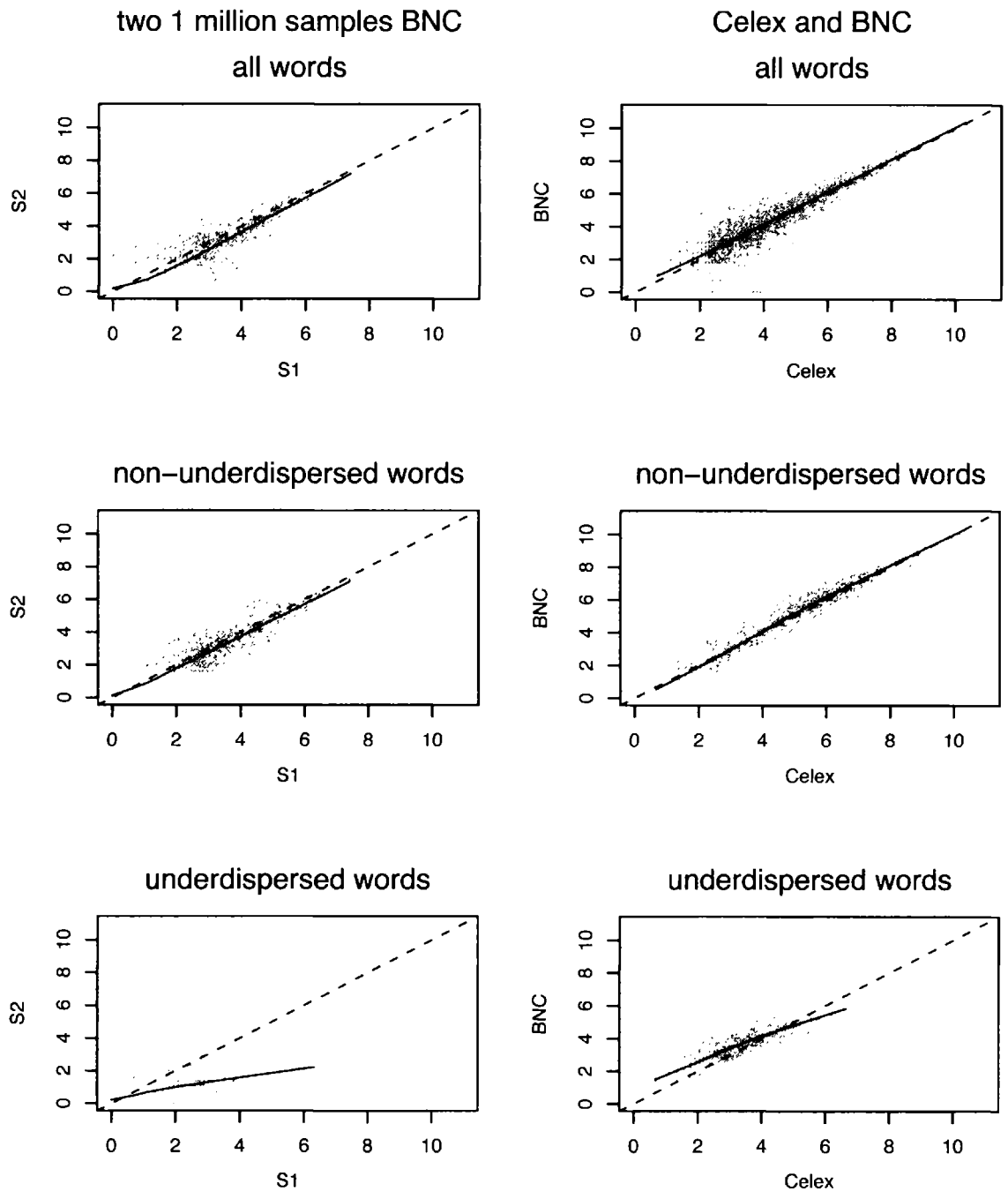

Figure 2. Scatterplots for log(frequency+1) in two subcorpora of one million words of the British National Corpus (left panels) as well as for the CELEX frequencies (standardized to 18 million) and the summed frequencies of 18 subcorpora of the BNC (18 million words, right panels). The top panels concern all monomorphemic nouns, the center panels the nouns that are not underdispersed, and the bottom panels the nouns that are underdispersed.

4. Word frequencies in corpora

Now that we have demonstrated with simulations that regression towards the mean is a phenomenon that is not to be expected for word frequencies, we turn to actual corpora. Is regression towards the mean evident when we examine corpora? To answer this question, we created 20 corpora of 1 million words from the 100 million word British National Corpus (BNC, http: //www.hcu.ox.ac.uk/BNC/). These corpora were created by going sequentially through the files of the BNC and assigning words to a given corpus until 1 million words had been read and assigned to this corpus. These 20 corpora make it possible to carry out 190 comparisons of two different corpora of 1 million words, the corpus size of the Brown corpus (Kučera and Francis, 1967) on which Gernsbacher (1984) based her conclusions.

We then selected as a test set all English words listed as monomorphemic singular nouns with a length greater than one letter and a frequency greater than 0 in the CELEX frequency counts (Baayen, Piepenbrock, and Gulikers, 1995), which are based on the Cobuild corpus at the time that corpus comprised 18.6 million words (see Renouf, 1987, for a description of this corpus at that time; currently, the Cobuild corpus comprises some 450 million words, see http://titania.cobuild.collins.co.uk/boe_info.html). The frequencies of homographs were collapsed. For each of the resulting 4410 nouns, we computed its frequency in each of the twenty subcorpora of the BNC. We also recorded the frequencies of these nouns as available in CELEX.

The upper left panel of Figure 2 is a pairs plot of the frequencies in the first two of our BNC subcorpora. As before, we plot the logarithmic transforms of the frequencies, again adding 1 in order to include words with zero frequency in the bilogarithmic plot. Word frequency values from one corpus ($S1$) were used to predict the frequency values of those words in a second corpus ($S2$). The first thing to note is that the plot has a shape that is quite similar to the shape generated by our Poisson-Lognormal simulation, and that is quite different from the shape generated by the bivariate normal simulation. The striated pattern for the lowest frequencies is, just as in the upper right panel of Figure 1, due to the discrete nature of word frequency counts. The solid line shows a nonparametric regression line, and the dashed line is the line $Y=X$. The second thing to note that there is no evidence whatsoever of regression towards the mean. This same pattern is seen across the different combinations of our 20 subcorpora

of the BNC.

We also created one corpus of 18 million words by combining the first 18 of the BNC subcorpora, in order to obtain a corpus size that is comparable in magnitude to that of the corpus underlying the frequency counts in the CELEX lexical database for English. The upper right panel of Figure 2 shows how well the CELEX frequencies (scaled from 18.6 to 18 million) can predict the frequencies from the BNC. We observe a very similar pattern, with again no trace of any visible indication of regression towards the mean. (There are no words with zero frequency in our data set of simplex English nouns, however, hence there is an asymmetry in the plot in the sense that there are nouns that occur in CELEX but not in the BNC, while there are no nouns that occur in the BNC but not in CELEX.) The upper panels of Figure 2 clearly demonstrate that the phenomenon of regression towards the mean, which for theoretical reasons is expected not to occur, does not occur in practice when comparing actual corpora, large or small.

The absence of regression towards the mean for word frequencies does not imply that there is no problem of sampling error. The bivariate Poisson-lognormal model is based on the assumption that words are used randomly in texts. This is a useful simplification, but words are generally not used at random. Many words are not distributed smoothly throughout different texts and corpora. Rather, they have a tendency to occur many times in one corpus, and may occur very few (or even zero) times in other corpora of the same size. This is a function of topicality, and will lead to certain words being used a lot in some of the texts that are sampled for frequency corpora, but very few times (or not at all) in many of the other texts sampled. This 'bunching' can lead to substantial 'sampling error' with zero frequency underestimating what happens in many other corpora and non-zero frequency overestimating what happens elsewhere. In other words, for the subset of words that occur bunched up in texts, a pattern similar to that of regression towards the mean in bivariate normal distributions might arise.

One way of making this notion of words being bunched up in texts more precise is to make use of occupancy theory (Johnson and Kotz, 1977) and the concept of underdispersion. The dispersion of a word is the number of different texts or subcorpora in which a word appears at least once. In our case, if a word appears in all 20 subcorpora of the BNC, its dispersion equals 20. If it occurs in only 1 of these subcorpora, its dispersion equals 1. A dispersion of 1 is not surprising for a word with a

frequency of 1, but it is very surprising for a word with a frequency of, say, 5000. To quantify the extent to which we should be surprised, we first calculate the expected dispersion E[*d*] for a given word with frequency *f* in a corpus with *k* equally-sized subcorpora:

$$E[d] = k(1-(1-\frac{1}{k})^{f}). \tag{5}$$

Next, we calculate the corresponding variance,

$$VAR[d] = k(1-\frac{1}{k})^{f} + (k(k-1)(1-\frac{2}{k})^{f} - k^{2}(1-\frac{1}{k})^{2f} \tag{6}$$

which allows us to calculate a *Z*-score:

$$Z = \frac{d-E[d]}{\sqrt{VAR[d]}}. \tag{7}$$

We found that roughly 42% of the words used in this study were significantly underdispersed ($Z < -1.96$). That is, they had significantly more "zero" entries than are expected given their overall frequencies. The remaining 58% of the nouns in our data set are not underdispersed at the low level of granularity (of only 20 subcorpora) that we have used here. The center panels of Figure 2 show that for the non-underdispersed words the non-parametric regression line and the line $Y = X$ are virtually indistinguishable to the eye. However, as shown in the bottom panels of Figure 2, for the underdispersed words, the frequencies in one corpus do not provide an accurate estimate of the corresponding frequencies in a second corpus.

First consider the bottom left panel, which compares two 1 million corpora from the BNC. For these small corpora, the nonparametric regression line is, except for the lowest frequency, below the line $Y = X$. In other words, a word with a frequency greater than one in subcorpus *S*1 tends to be paired with a lower frequency in *S*2. Given the granularity with which we have calculated the *Z*-scores for underdispersion, using subcorpora of one million words, there is no underestimation for the very lowest frequency words. This asymmetric pattern of sampling error differs from the symmetric pattern of regression towards the mean illustrated in Figure 1 for bivariate normal distributions.

Turning to the right panel, we see a symmetric pattern that is more similar to regression towards the mean, with the lowest frequencies in *S*1 tending to have somewhat higher frequencies in *S*2 and the higher frequencies in *S*1 tending to be somewhat lower in *S*2. With corpus sizes of 18 million, the level of granularity of our dispersion measure is now sufficient to render visible the underestimation for the lower frequency words in addition to the overestimation of the higher frequency words. It is crucial to keep in mind that this pattern emerges only by conditioning explicitly on words being significantly underdispersed. Given that we know a word is underdispersed, and given the knowledge that its frequency is extreme, we know that in another corpus its frequency will be less extreme. However, without prior knowledge about the topicality of a word, the fact that it has an extreme frequency in one corpus is not predictive about its frequency in another corpus, as demonstrated by the upper panels of Figure 2.

Finally, a comparison of the bottom panels shows that for small corpora the problem with underdispersed words is more substantial than for large corpora. This illustrates that studies using frequency norms based on small corpora are especially prone to sampling error.

5. Corpora of different size

Thus far, we have considered corpora of the same size. We now consider the correlation of word frequencies in corpora of different size. The upper panel of Figure 3 addresses this issue by means of a simulation. The horizontal axis plots the simulated frequencies of 1000 words following a Poisson-Lognormal(1, 4) distribution. The summed token frequency of these words is 1354830. The vertical axis plots the corresponding frequencies in a corpus of 338591 tokens, a reduction in size of 1/4. The simulated frequencies in the smaller corpus were obtained by reducing by a factor 4 the lognormal(1, 4)-distributed usage rates underlying the frequencies in the large corpus. The dashed line represents $Y = X$, the line around which the data points would have clustered if both corpora would have had the same size as in Figure 2. The straight solid line, $Y = -\log(4) + X$, shows the shift due to the reduction in corpus size. A word with frequency rate λ has a reduced rate of $\lambda/4$ in the small corpus, which in a bi-logarithmic plot shows up as a shift of log(4) to the right (entailing an intercept of –log(4)). The curved line represents a non-parametric

regression. Note that is is indistinguishable from $Y = -\log(4) + X$ for the higher frequencies. For the lowest frequencies, however, we see a deviation towards the origin. This is a simple consequence of the discrete

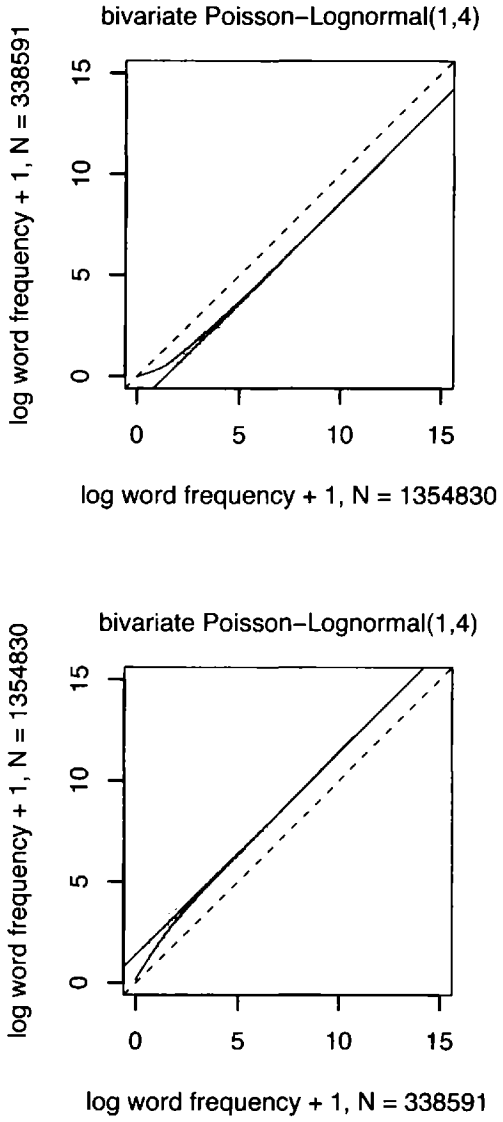

Figure 3. Scatterplots for log (frequency + 1) for corpora of different size. The upper panel plots the frequencies for 1000 words following a Poisson-Lognormal (1, 4) distribution on the horizontal axis, with on the vertical axis the corresponding frequencies in a corpus roughly a quarter in size. The lower panel reverses the axes. The dashed line represents the line $Y = X$, the straight solid line the expected frequency. The curved solid line is a non-parametric estimate.

nature of word frequencies. A word cannot have an observed frequency of, say, 0.2, which on a logarithmic scale would map onto a negative value on the vertical axis of −1.61, but only a frequency of 0, 1, 2, ... As a result, the lowest frequencies in the small corpus are slightly higher than would be expected given the big corpus.

As shown in the lower panel of Figure 3, the pattern for the lowest-frequency words reverses when we plot the frequencies in the large corpus on the vertical axis and those in the small corpus on the horizontal axis.

Due to discretization, the frequencies in the big corpus are not as large as one would expect given the corresponding frequencies in the small corpus. Thus, the discrete nature of word frequencies has as its consequence a slightly upward curvature when the frequencies from the small corpus are plotted against those of the large corpus, and a slightly downwards curvature when the axes are reversed. For the higher frequencies, no such divergence emerges. This pattern is incompatible with that expected for regression towards the mean, which predicts a straight regression line with a slope less than 1.

6. Singulars and plurals

In Galton's study, the heights of sons (Y) was plotted against the heights of their fathers (X). In his study, the heights of sons regressed towards the mean. It might be argued that a similar situation should obtain when we study the relation between the frequencies of singular nouns (the fathers) and the frequencies of their plurals (the sons). If a singular noun has a very high frequency, relative to other singular nouns, then the frequency of its plural should not be as extreme, relative to the frequencies of other plural nouns.

The assumption underlying this line of reasoning is that a given singular and its plural share the same probability of use p, but that their usage rates differ due to a difference in corpus size, N_{sg} for the singulars, and N_{pl} for the plurals, with $N_{sg} > N_{pl}$. Singulars would then have usage rate $\lambda_{sg} = N_{sg}\,p$ and plurals would have usage rate $\lambda_{pl} = N_{pl}\,p$. We will refer to this view of the relation between singular and plural frequency as the shared probability model.

Is regression towards the mean expected given the shared probability model? And is the shared probability model a sensible model for singular and plural frequencies? Given that there is no regression towards the mean

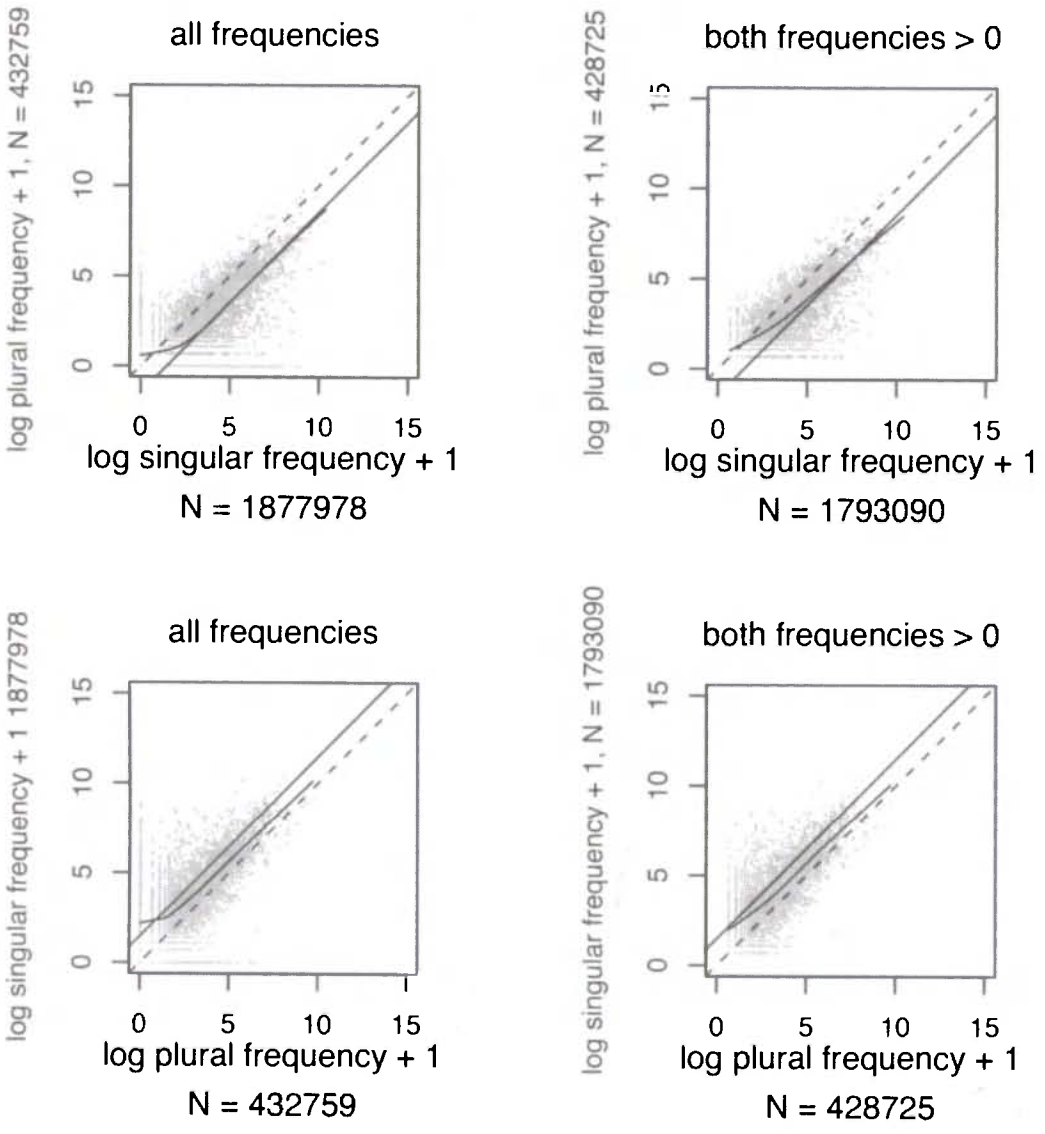

Figure 4. The frequencies of singulars and plurals of the 6431 monomorphemic English nouns in the CELEX lexical database. The upper panels plot the frequencies of plurals against those of their singulars. The lower panels reverse the axes. The left panels plot the full distribution, the right panels exclude nouns with zero singular or plural frequency.

for the frequencies of words in small and large corpora, as shown in the preceding section, no regression towards the mean is expected. This expectation is born out by the upper left panel of Figure 4, which plots the frequencies of singulars (on the horizontal axis) and plurals (on the vertical axis) for the 4631 monomorphemic English nouns (including pluralia tantum) in the CELEX lexical database.

The dashed line represents $Y = X$, the curved solid line represents a non-parametric regression. The straight solid line is $Y = \log(1877978/432759) + X$, and represents the expected regression line corrected for the difference in the size of the singular 'corpus' (1877978) and the size of the plural 'corpus' (432759). This graph is very similar to the upper panel of Figure 3, which plotted 1000 words in the plane spanned by their simulated frequencies in a large and a small corpus. This suggests that the sample of plurals might indeed be regarded as a small corpus sampled from the same population of lemmata from which the large corpus of singulars was sampled.

However, upon closer scrutiny, it turns out that the shared probability model is incorrect. Recall that plotting the frequencies in the large sample against the corresponding frequencies in the small sample, instead of the other way round, resulted in plots that are each other's mirror image, as shown in Figure 3. A plot of singular frequencies against plural frequencies, however, does not result in a mirror image, as shown by the lower left panel of Figure 4.

Note that the nonparametric regression line for singulars predicted from plurals shows an upward curvature where the corresponding lower panel in Figure 3 shows a downward curvature. Also note that the non-parametric regression line nowhere coincides with the dashed line representing the expected regression line under the shared probability hypothesis, except for the point where the two lines cross. The shared probability model incorrectly predicts higher singular frequencies given the plural frequencies than are actually observed.

In fact, there are various linguistic reasons for rejecting the shared probability model, and to regard singular and plural nouns as constituting relatively independent populations of their own. For instance, some nouns only occur in the plural (e.g., *trousers*). For these nouns, it makes no sense to model them as having a usage rate λ for singulars and $\lambda/4.34$ for plurals. The appropriate way to model such nouns is to assign them a zero usage rate as singulars and some non-zero usage rate as plurals. The presence of these pluralia tantum, represented by the leftmost column of points in the

upper left panel of Figure 4, causes the non-parametric regression line in this panel to end well above zero, while in the corresponding panel of Figure 3, the non-parametric regression line reaches the origin.

Similarly, there are many nouns that are used exclusively in the singular (e.g., *wool*), and, as can be seen in the bottom left panel of Figure 4, their frequencies (represented by the leftmost column of points) are also clearly higher than one would expect from their plural frequencies. Again, it makes no sense to model the singular frequency of words such as *wool* as having a usage rate that is four times that of the plural.

In addition, while unmarked singulars tend to be more frequent than their marked plurals, there are nouns that show a markedness reveral (Tiersma, 1982, Baayen, Dijkstra, and Schreuder, 1997). For these nouns, the plural form is semantically unmarked, and the singular semantically marked. For instance, the noun *eye* is more than three times as frequent in the plural than in the singular, which is no surpise as eyes tend to come in pairs. Again, it is incorrect to model the frequency of *eyes* as four times the frequency of *eye* observed in a smaller sample.

A final question is why, given the shared probability model, the prediction from the singular frequency to the plural frequency is right on target, while the reverse prediction results in an overestimation. The key to the answer is an asymmetry in the numbers of words that occur only in the singular (the singularia tantum) and the number of words that occur only in the plural (the pluralia tantum). There are 1254 singularia tantum in our data set, and only 162 pluralia tantum, in a total of 4631 nouns. When we remove the singularia and pluralia tantum from the data set and redraw the left-hand plots of Figure 4, the plots shown in the right-hand panels of Figure 4 are obtained. Note that the two graphs are now more like each other's mirror image, in that the non-parametric regression line is intermediate between the solid and dashed lines for a wide range of predictor values.

Comparing the upper two panels, we find that it is thanks to the large number of singularia tantum that the non-parametric regression line and the dashed line representing the predictions of the shared-probability model largely coincide in the upper left panel. (The removal of pluralia tantum has only a small effect on the non-parametric regression line, which starts off at nearly the same position in the lower left corner of both graphs).

Comparing the lower two panels, we observe that the number of pluralia tantum is far too small to allow the non-parametric predictions to coincide with those of the shared probability model. The two non-

parametric regression lines look very similar, except at the extreme left hand side of the graphs. In the lower left panel, the effect of the presence of singularia tantum is clearly seen compared to the lower right panel.

In the preceding section, we showed that when the shared probability model is correct, regression towards the mean does not take place for word frequency distributions. In the present section, we have seen for singular and plural frequencies, firstly, that the shared probability model is a non-optimal approximation at best, and secondly, that the regression towards the mean again does not take place.

7. Regression towards the mean in similarity neighborhoods

Landauer and Streeter (1973) observed that higher-frequency words tend to have high-frequency neighbors (lexical competitors at Hamming distance 1). However, as the frequency of a word increases, the likelihood that it will have neighbors that are even more frequent decreases, which has been described as regression towards the mean (Frauenfelder, Baayen, Hellwig, and Schreuder, 1993). Why is it that there is regression towards the mean in this case, and not in the examples studied in the preceding sections?

In the previous examples, we always were concerned with paired observations: the frequency of a given word in one sample paired with the frequency of the same word (or lemma, in the case of singulars and plurals) in a second sample. In other words, for one and the same word we have always had two measurements, each yielding a frequency with some measurement error. For neighbors, however, we are dealing with a given word the frequency of which is compared to the frequencies of other words that happen to be similar to it. So we need a different statistical model.

An appropriate statistical model is a Markov model generating word forms from transitional probabilities (Mandelbrot 1953, Nusbaum 1985, Frauenfelder et al. 1993, Baayen, 2001). This model defines words as strings of letters or phonemes from a set of pre-defined transitional probabilities. The higher the transitional probabilities of a word, the higher the probability of that word, and hence the higher its frequency. Neighbors that differ from a given target word in only one letter or phoneme share most transitional probabilities with the target word, except two: the transition into the differing segment and the transition from the differing segment back into the shared segments. Consider the neighbors *must* and *mast*, which differ at the second segment position. The set of all

neighbors sharing *m_st* is defined by the set of transitions *m_* followed by the transitions *_s*. Let *S* denote the set of all such paired transitional propabilities. The higher a given probability $p \in S$, the higher the frequency of a target word will be. At the same, the higher *p* is, the less likely it becomes that there will be another $p' > p$ for a neighbor with a frequency greater than that of the target word. Thus, the comparison of the frequencies of a target word and one of its neighbors boils down to comparing two probabilities drawn from the same set of probabilities *S*. This leads directly to regression towards the mean (see Frauenfelder et al. 1993 for detailed discussion). What this example shows, then, is that regression towards the mean is observed for the frequencies of different pairs of bigrams underlying the frequencies of different words. It can similarly be observed for the frequencies of different words in the same sample or corpus. If we sample a very high frequency word from a corpus, the likelihood of sampling a word with an even higher frequency becomes smaller with increasing frequency. This sampling situation should be carefully distinguished from the situation in which the same word is sampled in different corpora.

8. Conclusions

This study addressed the reliability of word frequency counts. Gernsbacher (1984) argued that sampling error renders printed word frequency unreliable, and linked sampling error with regression towards the mean. We have shown that this link between sampling error and regression towards the mean is unjustified, both in theory and in practice. There is no a-priori reason to assume that a high-frequency word will on average have a lower frequency in another corpus, or that a low-frequency word will have on average a higher frequency in another corpus. Both probability theory and detailed comparisons of corpora of 1 million words as well as corpora of 18 million words show that there is no regression towards the mean, and that on average a word with some frequency *f* in a corpus of a given size will occur in another corpus of the same size with a frequency that is very similar to *f*.

These results show that factorial designs contrasting high and low frequency words cannot be discredited on the basis of the argument that extreme frequencies are likely to regress towards the mean. Interestingly, even the exclusion of the more extreme frequencies in a regression design,

as in the study of Ford et al. (present volume), is unwarranted. Just as removal of the central observations from a scattercloud leads to an overestimation of the true correlation, the unwarrent exclusion of the higher and lower frequencies results in an underestimation of the correlation.

Our demonstration that there is no regression towards the mean for word frequencies does not imply that there is no problem of sampling error. We have shown that some 42% of monomorphemic nouns in English are significantly underdispersed, i.e., they occur in fewer subcorpora than one would expect given their frequency of use. Within the subset of the underdispersed, topical words, a phenomenon similar but not identical to regression towards the mean as defined for bivariate normal distributions can be observed. However, whether a word is a topical word or not cannot be predicted from its frequency alone.

Topical underdispersed words constitute a special problem for experimental studies of lexical processing. Carroll (1970) proposed to adjust frequency counts for their dispersion. However, when detailed information about the distribution of the tokens over texts and subcorpora is missing, as in for instance the CELEX lexical database, Carroll's adjusted frequency measure cannot be calculated. An additional problem with topical words is that their mean frequency is not a valid estimate of how such words are really used — the median frequency is a more appropriate point estimator. A more principled approach to this problem has recently been developed by McDonald and Shillcock (2001), who argue that an entropy-based measure of lexical variation is superior to the simple word frequency count. Pursuing this approach would take us beyond the goal of the present paper, however, which is to clarify the confusion in the current literature about the supposed regression towards the mean in lexical statistics.

References

Baayen, R. Harald

2001 *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. Harald, Ton Dijkstra, and Robert Schreuder

1997 Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, 37:94-117.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers
1995 *The CELEX lexical database* (CD-ROM). Linguistic Data Consortium, Philadelphia, PA.: University of Pennsylvania.

Carroll, John B.
1967 On sampling from a lognormal model of word frequency distribution. In *Computational analysis of present-day American English*, H. Kučera, W.N. Francis (eds.) Providence: Brown University Press.

Carroll, John B.
1970 *An alternative to Juilland'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.

Fisher, Ronald
1918 On the correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399-433.

Frauenfelder, Uli .H., R. Harald Baayen, Frauke M. Hellwig, and Robert Schreuder
1993 Neighborhood density and frequency across languages and modalities.
Journal of Memory and Language, 32:781-804.

Gernsbacher, Morton A.
1984 Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113:256-281.

Good, I.J.
1953 The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237-264.

Johnson, N.L. and S. Kotz
1977 *Urn models and their application: An approach to modern discrete probability theory*. New York: John Wiley & Sons.

Kučera, Henry and W.Nelson Francis
1967 *Computational analysis of present-day American English.* Providence: Brown University Press.

Mandelbrot, Benoit
1953 An information theory of the statistical structure of language. In *Communication theory*, W.E. Jackson (ed.). New York: Academic Press

McDonald, Scott and Richard Shillcock
2001 Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44:295-323.

Nusbaum, Howard C.
1985 A stochastic account of the relationship between lexical density and word frequency. Technical report, Indiana University. *Research on Speech Perception*, Progress Report #11.

Renouf, Antoinette
1987 Corpus development. In *Looking up: An account of the Cobuild Project in lexical computing*, J.M. Sinclair, (ed.). London: Collins.

Sereno, Joan, and Allard Jongman
1997 Processing of English inflectional morphology. *Memory and Cognition*, 25:425-437.

Taft, Marcus
1979 Recognition of affixed words and the word frequency effect. *Memory and Cognition*, 7:263-272.

Tiersma, P.M.
1982 Local and general markedness. *Language*, 58:832-849.