Power laws and other heavy-tailed distributions in linguistic typology

GERHARD JÄGER
Institute of Linguistics, University of Tübingen, Wilhelmstr. 19
72074 Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

September 23, 2011

Abstract

The paper investigates the quantitative distribution of language types across languages of the world. The studies are based on three large-scale typological data bases: The World Color Survey, the Automated Similarity Judgment Project data base, and the World Atlas of Language Structures. The main finding is that a surprisingly large and varied collection of linguistic typologies show power law behavior. The bulk of the paper deals with the statistical validation of these findings.

1 Introduction

Power laws are a common feature of many complex systems. They are typical of physical systems near a phase transition, but they also frequently emerge in biological and social systems. Examples are Yule's observation about the number of biological species within a genus, Pareto's Law about income distributions, Zipf's Law about word frequencies, or the distribution of city sizes. Within recent years, it has been observed that several numerical parameters of self-organized networks—like the internet—show a power law distribution.

Linguistics is no exception here. Zipf's Law, which states that word frequency distributions follow a certain power law, has already been mentioned. Wichmann ([19]) argues that the size of language families—measured in the number of languages per families—show power law behavior as well. Maslova recently made a case that *meta-typological distributions* have the same property (cf. [14]). A meta-typological distribution concerns the size of language types (that are defined in terms of inherent properties of a language, rather than in terms of their historical affiliation).

In reaction to Maslova, Cysouw argues that the size of typological classes follows an exponential distribution rather than a power law (cf. [8]). This debate points at an important issue: Identifying power laws in empirical data is a non-trivial issue. There is a variety of probability distributions that may have a similar shape like power law distributions for appropriate parameter values. To identify a power law behavior in a certain empirical domain with a certain confidence, it is important to ensure (a) that the proposed power law fits the observed data reasonably well, and (b) it provides a better fit than other candidate distributions.

In this paper I will look at data from two sizeable typological data bases: the World Color Survey ([13]) and the data form the *Automated Similarity Judgment Program* ([20]). It will be shown that the classification of languages according to quite diverse criteria—partially semantically and partially phonetically motivated—lead to power law distributions. A large part of the paper is devoted to the methodological issue how the power law hypothesis can be justified statistically.

Finally, Maslova's and Cysouw's proposals—which are based on data from the World Atlas of Language Structures ([10])—will be evaluated according to the same standards.

2 What is a power law distribution?

A probability distribution over a set of positive real numbers X is a power law distribution if its density function p has the form

$$p(x) \propto x^{-\alpha}, \tag{1}$$

where $\alpha > 1$ is a real number. If p is a continuous function that is defined for all reals $\geq x_{min}$ (for $x_{min} > 0$), a power law distribution has the density function

$$p(x) = (\alpha - 1)x_{min}^{\alpha-1} \cdot x^{-\alpha}. \tag{2}$$

The cumulative distribution function is then given by

$$\begin{aligned} Pr(X \geq x) &= \int_x^\infty p(y)dy & (3)\\ &= x_{min}^{\alpha-1}x^{1-\alpha} & (4)\end{aligned}$$

Now suppose A is a sufficiently large random sample of numbers drawn from the power law distribution $p(\cdot|\alpha, x_{min})$, with $|A| = N$. Let $x \in A$. According the law of large numbers,

$$Pr(X \geq x) \approx \frac{|\{y \in A|y \geq x\}|}{N}. \tag{5}$$

The quantity $r_x = |\{y \in A|y \geq x\}|$ is the rank x within A; i.e. if the elements of A are ordered according to size in descending order, x will occupy the r_xth position. Putting this together, we get

$$r_x \approx N x_{min}^{\alpha-1}x^{1-\alpha}, \tag{6}$$

and thus

$$x \approx x_{min}N^{\frac{1}{\alpha-1}}r_x^{\frac{1}{1-\alpha}} \tag{7}$$

So if a variable is distributed according to a power law, the size of an observation is approximately proportional to a power of its rank. This fact is often employed for discovering and visualizing power laws. Obviously it holds that

$$\log x \approx \frac{1}{1-\alpha}\log r_x + log(x_{min}N^{\frac{1}{\alpha-1}}) \tag{8}$$

In a doubly logarithmic plot, the size is thus approximately a linear function of the rank. The dependency thus shows up as a straight line with slope $1/(1-\alpha)$.

If the distribution is only defined for positive integers (which makes sense in all empirical domains where we are dealing with counts, like the number of city inhabitants or of word occurrences), it takes the form

$$p(x) = \frac{1}{\sum_{n=0}^{\infty}(n + x_{min})^{-\alpha}}x^{\alpha} \tag{9}$$

The normalizing constant ensures that all probabilities sum up to 1.

To ensure comparability of power law models with other "heavy-tailed" distributions, I will always assume an underlying continuous distribution according to (1) though. If the observed values are integers, I will assume they are the result of rounding. (Values $< .5$ are then rounded to 0 and, accordingly, unobservable.) Generally, for an underlying continuous probability density function p, I will assume the following discrete approximation:

$$p_{disc}(k) = \frac{1}{\int_{0.5}^{\infty}p(y)dy}\int_{k-0.5}^{k+0.5}p(x)dx \tag{10}$$

In [7] a detailed exposition is given how to test whether a sample of empirical data really confirms to a power law distribution. I will largely follow this proposal.

In a first step, we will check whether a power law provides a reasonably good fit to the data. This is done in the following way:

1. Estimate the parameters α and x_{min} for the data in question that assign them the highest likelihood.

2. Measure the divergence between the empirical data and the power law distribution that is governed by these parameters. A suitable way to do so is to calculate the Kolmogorov-Smirnov statistic.

3. Generate a collection C of random numbers which are distributed according to the power law that is fitted to the empirical data.

4. Estimate the parameters for C and determine the Kolmogorov-Smirnov statistic between the random sample and its maximum likelihood model and compare it to the corresponding value for the empirical data.

The last two steps are repated 1,000 times. The relative frequency p of random sample that have a higher divergence from their fitted model than the empirical data gives an estimate how likely it is to observe such a distribution as we have it in our data if the underlying distribution is in fact a power law. In the logic of empirical tests, this procedure gives us a p-value, where the null hypothesis is that the data are generated by a power law. Note that unlike in most cases of statistical testing, here the null hypothesis is the "interesting" case, and thus high p-values are "good". The authors of [7] propose that $p > 0.1$ indicates that a power law distribution is a plausible model for the data at hand.

The mentioned authors also consider the case that not all data points are governed by a power law, and they therefore also devote attention how to estimate x_{min}. I will make the more restrictive assumption here that the minimal observed value is always 1, i.e. that the power law holds for all data. According to the approximation of discrete data by a continuous model (cf. 10), I will thus assume $x_{min} = .5$.

In a second step we will perform the same test for other heavy-tailed distributions to check whether there is a better model for the data than the power law.[1]

Candidate distributions that are considered in this paper are

- the exponential distribution:
$$p(x|\lambda) = \lambda e^{\lambda x_{min}} e^{-\lambda x} \tag{11}$$

- the log-normal distribution:
$$p(x|\mu,\sigma) = \sqrt{\frac{2}{\pi\sigma^2}}[x \cdot \mathrm{erfc}(\frac{\log x_{min} - \mu}{\sqrt{2}\sigma})]^{-1} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \tag{12}$$

- the power law with exponential cutoff:
$$p(x|\alpha,\lambda) = \frac{\lambda^{(1-\alpha)}}{\Gamma(1-\alpha, \lambda x_{min})} x^{-\alpha} e^{-\lambda x} \tag{13}$$

- the Sichel distribution (also called *Generalized Inverse Gauss-Poisson* distribution; see [1, 16, 17]):[2]
$$p(x|a,b,\gamma) = \frac{(a/b)^{\frac{\gamma}{2}}}{2K_\gamma(\sqrt{(ab)})} x^{\gamma-1} e^{-(ax+b/x)/2}, \tag{14}$$
where $K_\gamma(\cdot)$ ist the modified Bessel function of the second kind.

[1] As the computational effort for parameter estimation is quite high for some of these models, I sometimes only used 100 random samples rather than 1,000, which is usually sufficient.

[2] Thanks to Harald Baayen (p.c.) for drawing my attention to this model.

The exponential distribution is mainly included to facilitate comparison with Cysouw's proposal. The log-normal distribution has a similar shape than a power law for certain parameter settings, and it is sometimes difficult to tell samples from these distributions apart. The logarithms of the probabilities assigned by the log-normal distribution are distributed according to a normal distribution. According to the central limit theorem, macroscopic variables that are the sum of many independent microscopic variables follow approximately a normal distribution. Accordingly, macroscopic variables that are the product of many independent microscopic variables follow a log-normal distribution. Therefore this distribution is *prima facie* a plausible model for the kind of emergent phenomena that we observe in linguistic typology.

The power law with exponential cutoff is a mixture of a power law and an exponential distribution. For small x, this distribution is close to a power law, while larger values follow an exponential distribution. Finally, the Sichel distribution is a further refinement of the power law with exponential cutoff that includes the term $-b/2x$ into the exponent, thus shifting probability mass from high to low values. It has been successfully applied for instance to word frequency distributions.

3 Color term systems

The World Color Survey is a large scale typological questionnaire study about the diversity of color naming systems across languages of the world. It was conducted by researchers around Paul Kay and Brent Berlin within the past decades (see [13]). The study worked with more than 2,000 participants from 110 typologically diverse languages around the globe. On average, 24 participants per language where consulted.

The investigation made use of the Munsell Chart, a collection of 330 colored chips which cover 320 shades of maximally saturated colors (forty different hues and eight different levels of brightness) plus ten achromatic chips (black, white, and eight shades of gray in between). In one of the tasks, these 330 chips were presented to the participants in a random order. They had to name the color of each chip in their native language. The thus obtained data are freely available from the homepage of the World Color Survey project at http://www.icsi.berkeley.edu/wcs/.

Using Principal Component Analysis as a dimensionality reduction techniques, [11] shows that 91.6% of the cross-linguistic variance among the extensions of color terms can be accounted for by just 15 features. Each of these features is a 330-dimensional vector, assigning each of the 330 Munsell chips a real number. Eleven of these 15 features can be identified with the fuzzy extensions of the basic color terms that Berlin and Kay identified in their seminal study [4]. The other four features can be described as *olive green*, *light blue*, *pastel green* and a shade of color between blue and purple.

The extension of a certain color term for a certain speaker is likewise a 330-dimensional vector, consisting only of 1s (for Munsell chips falling under the extension of this term) and 0s. The similarity (intuitively: the degree of overlap) between a feature f and the extension of a term t can now be computed as the vector product between f and the vector corresponding to t. Therefore, the classification data for a given participant uniquely define a partition over the 15 features.[3]

The recent literature on the typology of color terms mainly focuses on the classification of the six primary colors *red*, *green*, *blue*, *yellow*, *black* and *white* across languages (see for instance [12]). Therefore I will concentrate on these six features here.

Among the 1,771 participants for which the data base contains complete data, the most frequent partitions types are:

- {white}, {red}, {yellow}, {green, blue}, {black} (742 occurrences)
- {white}, {red}, {yellow}, {green}, {blue}, {black} (447 occurrences)
- {white}, {red, yellow}, {green, blue, black} (111 occurrences)

[3] A more detailed description of this feature extraction procedure is given in [11].

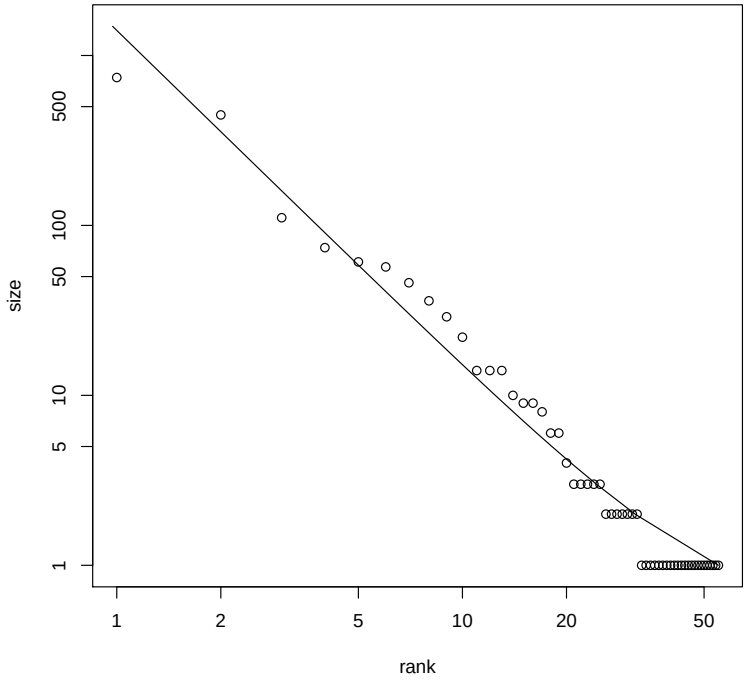

Figure 1: Doubly logarithmic rank-size plot of partition types over the six primary colors

- {white}, {red}, {yellow}, {green}, {black, blue} (74 occurrences)
- {white}, {red}, {yellow, green, blue}, {black} (61 occurrences)

In a doubly logarithmic rank-size plot, the distribution of partitions types looks approximately like a straight line, so it is *prima facie* a candidate for a power law distribution (Figure 1). Using numeric optimization, the maximum likelihood estimation for the exponent α (according to approximation by a continuous model according to equation (10)) is approximately 1.505. The line in Figure 1 displays the predicted values for the power law with this exponent. (The curve is slightly concave for high values; this is due to the discretization according to (10)).

The p-value for the color categorization data—for the maximum likelihood power law model with $\alpha = 1.504$—is estimated as 0.66.

The p-values for the exponential distribution and the log-normal distribution are 0.00, so both models can be rejected for our data.

The power law is a special case of the power law with exponential cutoff (for $\lambda = 0$), and the latter in turn is a special case of the Sichel distribution (for $\gamma = 1 - \alpha$, $a = 2\lambda$, and $b = 0$).[4] Therefore the fit of the power law with exponential cutoff and the Sichel distribution are necessarily as good as the fit of the power law, i.e. the two more complex models are initially plausible candidates to fit the empirical data as well.

The three plausible candidate models are thus *nested*. To compare the goodness of fit of two nested models, I will use the likelihood ratio test. This is a general way to compare the fit of two statistical models M_0 and M_1 to a data set. Suppose M_0 is a special case of M_1. The likelihood ration test follows the logic of statistical hypothesis testing, the null hypothesis being that M_0

[4]Strictly speaking the normalization terms are undefined for these parameter settings; a more precise formulation is to say that the power law is the limit of a power law with exponential cutoff if λ converges to 0, and likewise for the relation between the power law with exponential cutoff and the Sichel distribution.

is correct and the alternative hypothesis being that M_1 is correct. Let LL_0 and LL_1 be the log-likelihoods of the data according to M_0 and M_1 respectively. The test statistic

$$D = 2LL_1 - 2LL_0$$

approximately follows a chi-square distribution, with the degrees of freedom being the difference in the number of free parameters between M_1 and M_0. This in turn allows to compute a p-value, i.e. the likelihood of the observed data given the null hypothesis.

For the data and models at hand, the p-value for comparing the power law to the power law with exponential cutoff is 0.16 and for power law vs. the Sichel distribution 0.38. Assuming the usual significance level of 0.05, we can conclude that neither of the two refinements of the power law significantly improves the fit of the data.

The *Akaike Information Criterion* (AIC, cf. [5]) is another suitable measure of the relative goodness of fit of statistical models. Compared with the likelihood ratio test, it has the two advantages that is is also applicable to non-nested models, and it does not assume that any of the two models compared is the correct one. It simply tells us which of the two models provides a better fit.

For a given model with k parameters, the AIC measure is defined as

$$AIC = 2k - 2\log(LL), \tag{15}$$

where LL is the log-likelihood of the data given the model. The AIC estimates the divergence of the model to the underlying distribution of the empirical data. The best model is thus the one with the lowest AIC. This measure rewards a good fit of the model to the observations, and it penalizes the number of model parameters. The rationale for the latter is the fact that the risk of overfitting increases with the number of parameters. As the power law model has only one parameter while the other two candidate models have two and three respectively, the power law has *ceteris paribus* an advantage.

The AIC value for the power law is 334.63, the value for the power law with exponential cutoff is 334.70, and the one for the Sichel distribution is 336.70. So this criterion also favors the power law over its two refinements.

It has been suggested (by Amir Zeldes, p.c.) that the power law distribution of color naming systems might a side effect of the power law distribution of linguistic family sizes (because speakers of languages that belong to the same famility are likely to use similar systems).

The 102 languages that were used to obtain the color categorization data belong to 70 different families. To exclude the potentially confounding effect of family size, I repeated the steps described above for a collection of 70 languages, each being drawn at random from a different family. The rank-size plot is shown in the left panel of Figure 2. It suggests that these data also follow a power law.

This impression is supported by the statistical analysis. The p-values for power law ($\alpha = 1.51$) is 0.44, and the ones the exponential distribution and the log-normal distribution are both 0.00. So the latter two distributions can be rejected as possible models.

However, the p-value for comparing the power law to the power law with exponential cutoff according to the likelihood ratio test is 0.13, and the one comparing the power law to the Sichel distribution is 0.34. This indicates that these two models do not significantly improve the fit as compared to the power law.

The AIC values for the three candidate models are 294.39 (power law), 294.07 (power law with exponential cutoff) and 295.15 (Sichel) respectively. So the AIC predicts that the power law with exponential cutoff will perform slightly better when applied to unseen data, while the Sichel distribution will perform worse.

As the number of language families per genus also follows a power law, I repeated the procedure for a sample of languages each of which is a randomly picked representative of a different genus. The languages in the WCS belong to 45 different genera (see right panel of Figure 2). Here the p-values for the power law is 0.95 (for $\alpha = 1.56$), while the p-values both for the exponential and the log-normal distribution are 0.00.

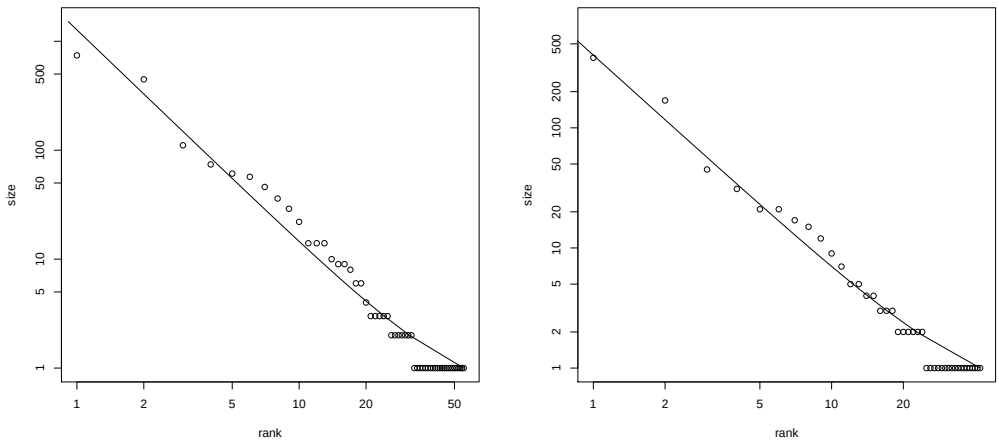

Figure 2: Doubly logarithmic rank-size plot of partition types; one language per family/genus

The likelihood ratio test for comparing the power law to the power law with exponential cutoff yields a p-value of 0.25, and for comparint the power law to the Sichel distribution 0.51. So neither significantly improves the fit of the power law.

This finding is confirmed by the AIC. The AIC values for power law, power law with exponential cutoff, and Sichel distribution are 231.8, 232.5 and 234.5 respectively. Again the power law distribution provides the best fit for the data. We can thus safely conclude that the power law distribution of color categorization types is not a side effect of the distribution of family size or genus size.

These qualitative results do not depend on the somewhat arbitrary decision to bin color naming systems according to their partition over the six most informative features. Including a seventh feature (roughly corresponding to purple) into the analysis leads to a rather similar pattern. The rank-size plot is shown in Figure 3. The p-values for the power law distribution ($\alpha = 1.61$) is as high as 0.94 here, while the p-values for the exponential and the log-normal distributions are both 0.00.

Again the two refinements of the power law do not significantly improve the fit of the model. The p-value for the likelihood ratio test comparing the power law with the power law with exponential cutoff is 0.17, and the p-value for comparing the power law with the Sichel distribution is 0.43

The AIC leads to the same conclusion. The AIC values for the power law, the power law with exponential cutoff, and the Sichel distribution are 548.79, 548.90, and 551.09 respectively, so the simple power law comes out best here as well.

4 Phonological templates

The *Automated Similarity Judgment Program* (ASJP; see [20] and the project home page `http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm`) has set up a large data base of core vocabulary items across many (in fact, most) languages of the world. It works with a set of 40 core concepts (being derived from the Swadesh list). Examples of these concepts are *I, you, we, one, two, water, stone, fire, skin, leaf, tree*. The expressions of these concepts in 4,802 languages are collected in a uniform coarse-grained phonetic transcription. Excluding artificial languages and creole languages leaves data from 4,748 languages. The following investigation is based on the expressions for the 40 core concepts in these languages (in cases where more than one expression was given, only the first one was used), which are 168,015 word forms in total. (Some expressions

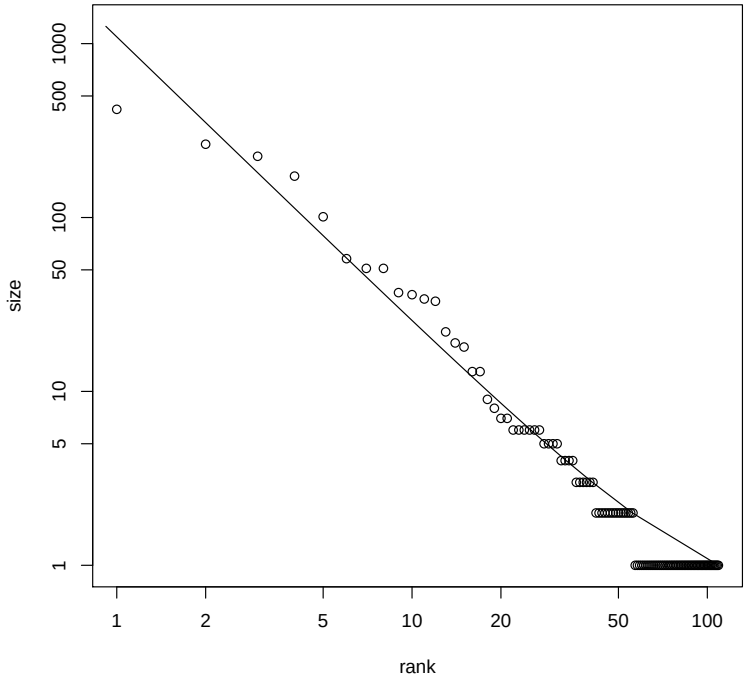

Figure 3: Doubly logarithmic rank-size plot of partition types over seven colors

are missing, which explains why this number is slightly smaller than $4,748 \cdot 40 = 189,920$.)

The main purpose of setting up the ASJP data base is the aim to automatize the estimation of similarity between languages by comparing the phonetic form of core vocabulary items, and thus to facilitate the computer aided identification of genetic relations between languages. However, it also provides ample material to investigate the phonological structure of core vocabularies across languages. The following barely scratches the surface of the potential of this resource.

The word forms were categorized according to the sequence of consonants (C) and vowels (V) they instantiate. For instance, the English words "I", "you", "we", "one" and "two" are categorized as VV, CV, CV, CVC, and CV respectively. The most frequent CV templates across languages are given in Table 1.

The size-frequency plot (with doubly logarithmig axes) for all CV templates is given in Figure 4.

CVCV	30,958
CVC	20,278
CV	16,395
CVCVC	12,275
VCV	9,971
CVCVCV	6,915
CVCCV	5,428
CVV	5,293
VCVC	4,338
CCV	4,140

Table 1: Most frequent CV templates

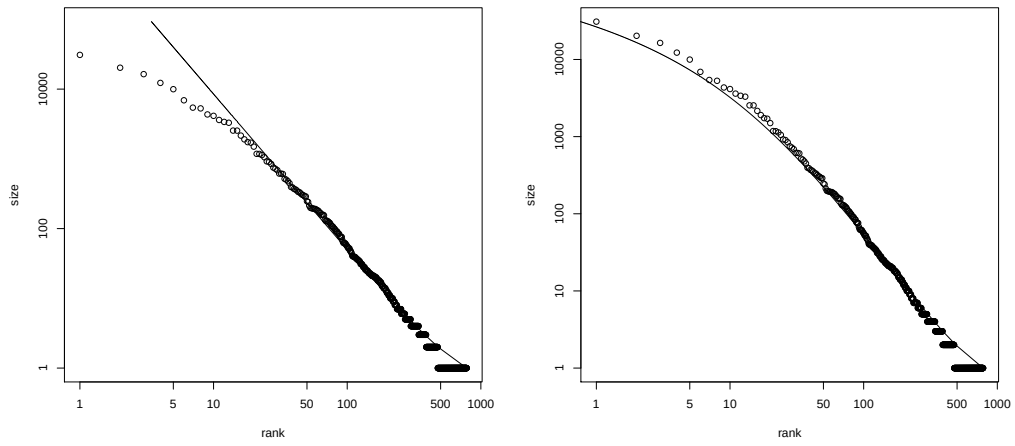

Figure 4: Size-frequency plot of CV templates in ASJP data base; fitted power law (left) and power law with exponential cutoff (right)

It seems that most data points (those with a rank larger than ca. 30) can be approximated quite well with a stright line (see left panel), so *prima facie* we are dealing with a candidate for a power law distribution. The p-value for a power law ($\alpha = 1.45$) is 0.54, so this hypothesis is in fact plausible. The p-value for the exponential and the log-normal distributions are again 0.00.

The likelihood ratio test reveals, however, that here the power law with exponential cutoff (with $\alpha = 1.42$ and $\lambda = 2.69 \cdot 10^{-5}$) provides a significant improvement. The p-value for comparing the two models is 0.0018, so it is highly significant. Comparing the power law with the Sichel distribution yields a p-value of 0.0079, which is also significant. However, comparing the power law with exponential cutoff with the Sichel distribution leads to a p-value of 0.98, so there is no noticable improvement. In fact, the fitted Sichel distribution has a parameter value of $b = 0$, so it is in fact a power law with exponential cutoff.

These findings are confirmed by the AIC. The AIC values for the power law, the power law with exponential cutoff, and the Sichel distribution are $5,309.5$, $5,301.9$, and $5,303.9$ respectively. Clearly the power law with exponential cutoff provides the best fit here.

The very small value for λ indicates that this model is still very close to a pure power law. Nonetheless the difference is not negligible, as can be seen from the right panel of Figure 4, where this model is shown (solid line).

To exclude family size effects, I carried out the same tests for the word forms from a sample of 529 languages, each taken from a different language family. The rank-size plot for the frequencies of CV-templates from this sample are shown in Figure 5.

As the previous distribution, this distribution seems to follow a power law except for the most frequent items. The statistical test for a power law distribution gives a p-value of 0.44, while the p-values for the exponential and the log-normal distributions at 0.00.

The p-values for the likelihood ratio test for power law vs. power law with exponential cutoff are 0.005, for power law vs. Sichel distribution 0.018, and for power law with exponential cutoff vs. Sichel distribution 1.0. We can thus again conclude that the power law with exponential cutoff (with $\alpha = 1.46$ and $\lambda = 2.7 \cdot 10^{-4}$) provides the best fit of the data.

This is also confirmed by the AIC. The AIC values for power law, power law with exponential cutoff and Sichel distribution are $2,428.3$, $2,422.4$, and $2,424.4$ respectively.

Sampling one language from each of the 219 genera that are represented in the ASJP data base leads to a qualitatively similar result (see Figure 6).

The p-values for the power law is 0.17, the p-values for the log-normal and exponential distributions are 0.00 respectively. The likelihood ratio test for comparing the power law with the

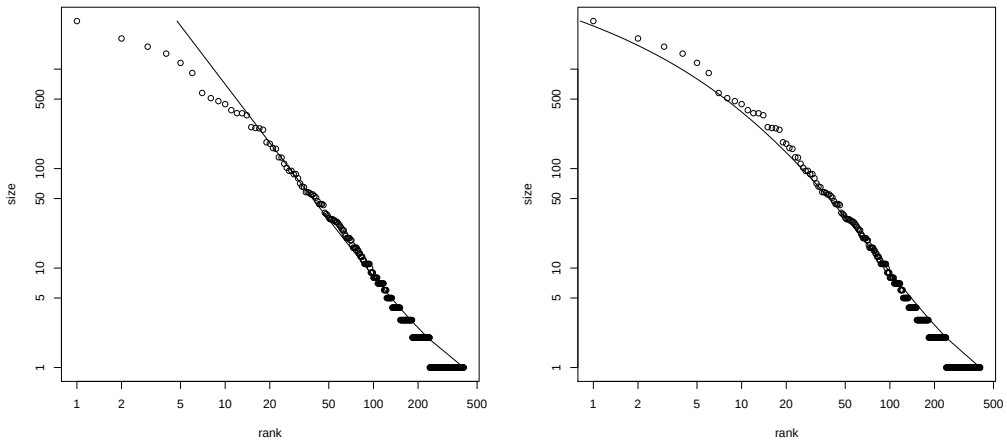

Figure 5: Size-frequency plot of CV templates, one language per family; fitted power law (left) and power law with exponential cutoff (right)

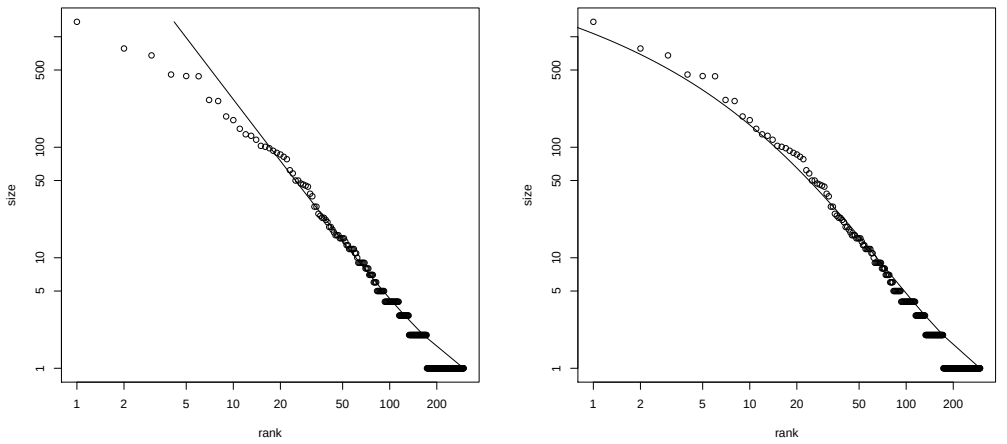

Figure 6: Size-frequency plot of CV templates, one language per genus; fitted power law (left) and power law with exponential cutoff (right)

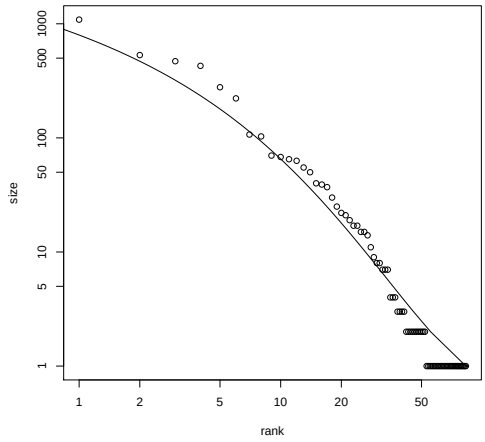

Figure 7: Size-frequency plot of CV templates for first person plural pronoun, one language per genus; fitted power law with exponential cutoff

power law with exponential distribution yields a p-value of 0.005, so the difference is significant. The comparison of the power law with exponential cutoff to the Sichel distribution yields a p-value of 1.0, so the power law with exponential cutoff comes out as the model providing the best fit. This is confirmed by its AIC value of $1,680.8$, as compared to the value of $1,686.6$ for the pure power law and $1,682.8$ for the Sichel distribution.

We can thus conclude that the distribution of CV-templates in the core vocabularies of the languages of the world follow a distribution that is very close to a power law and can best be modeled by a power law with exponential cutoff with an exponent close to 1.5 and a decay rate close to 0.

A similar result is obtained if the frequencies of templates is confined to a particular meaning. For instance, the frequencies of CV-templates in the words for the first person singular pronoun are distributed as shown in Figure 7. The p-value for the power law is 0.18, and it is 0.00 for both the exponential and the log-normal. The likelihood ratio test yields a p-value of 0.014 for comparing the power law to the power law with exponential cutoff ($\alpha = 1.32$ and $\lambda = 8.0 \cdot 10^{-4}$), and 0.99 for comparing the power law with exponential cutoff to the Sichel distribution. The AIC values are 585.5 for the power law, 581.4 for the power law with exponential cutoff and 583.4 for the Sichel distribution. Again, the power law with exponential cutoff clearly provides the best fit of the data.

Analogous results obtain for the other 39 concepts covered in the ASJP data base as well. Likewise, categorizing languages according to the first two segments in their word for a particular core concept, or according to the last two segments, gives rise to power law-like distributions.

Since several quite diverse and randomly selected categorizations of typological variables give rise to power law distributions (or something very close to it), one might wonder how surprising this outcome really is. Therefore I conducted another study, using a completely arbitrary categorization criterion to see whether anything close to a power law will emerge. I classified the word forms for the numeral "two" in the ASJP data base according to the first and the third segment in the ASJP phonetic transcription.[5] The resulting distribution is also "heavy-tailed", with a small number of frequent and a large number of rare types, as can be seen from the linearly scaled rank-size plot that is shown in the left panel of Figure 8.

The right panel of Figure 8 displays the same information in a doubly logarithmic plot. Here the visual impression is deceptive. The data points seem to lie close to a straight line. However, the p-value for the power law is 0.032, so this model can be rejected with high confidence. The

[5]Words with up to two segments were all classified according to their initial segment.

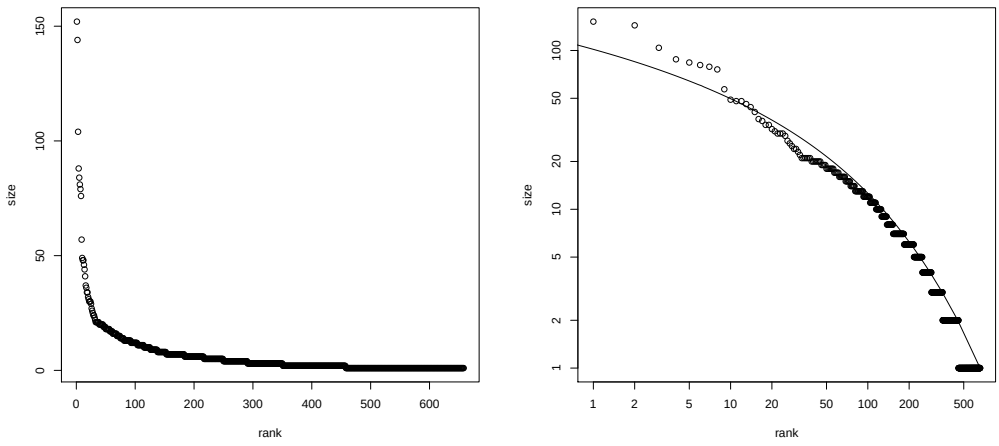

Figure 8: Size-frequency plot of frequencies of first/third segment combinations of word for "two" in ASJP data base; linear (left) and log-log scaled with fitted Sichel distribution (right)

exponential and the log-normal distributions have p-values of 0.00. The power law with exponential cutoff has a p-value of 0.57, so it cannot be rejected. The Sichel distribution (with $a = 0.013$, $b = 4.71$, and $\gamma = -1.07$ fits the data even better. The p-value for the likelihood ratio test between the power law with exponential cutoff and the Sichel distribution is 0.007. The AIC value for the power law with exponential cutoff is $3,502.6$, while the Sichel distribution reaches a value of $3,497.44$.

5 Meta-typological distributions

Let us finally briefly consider the meta-typological distributions that where discussed by Maslova and Cysouw. Both authors consider the typologies that are collected in the World Atlas of Language Structures (WALS; see [10] and the WALS homepage `http://wals.info`). In WALS, large numbers of languages are categorized according to grammatical features that are considered interesting by linguists. Examples of the features covered are "Syllable Structure", "Number of Genders", "Plurality in Independent Personal Pronouns", "Nonperiphrastic Causative Constructions" and the like. In total, 142 such features are covered.[6] The number of languages that are categorized varies considerably between features: the minimum is 6 ("Writing Systems") and the maximum 1,370 ("Order of Object and Verb"). The mean is ca. 409 and the median 299. In total 2,561 languages are categorized by at least one feature. The number of different values that the features can have varies between two (for instance "Vowel Nasalization", "Obligatory Possessive Inflection" or "Zero Copula for Predicate Nominals") and nine (for "Coding of Nominal Plurality", "Number of Cases", and "Comitatives and Instrumentals").

Cysouw picked one random value for each feature and counted the number of languages that have this feature. He claims that the sizes of these types follows an exponential distribution. I conducted the same experiment. The rank-size plot for the resulting distribution is shown in Figure 9. If the data follow an exponential distribution, they should lie on a straight line in a plot where the x-axis is logarithmically and the y-axis is linearly scaled. Judging from the visual impression, this seems a plausible hypothesis. However, the p-value for this model is 0.00, so this hypothesis can be rejected with very high confidence. Likewise, the p-value for the power law is 0.00. The log-normal distribution (with $\mu = 3.54$ and $\sigma = 1.52$) and the power law with exponential cutoff (with $\alpha = 0.40$ and $\lambda = 7.9 \cdot 10^{-3}$) are *prima facie* plausible models. The former

[6] This applies to the version of WALS from 21 February, 2010.

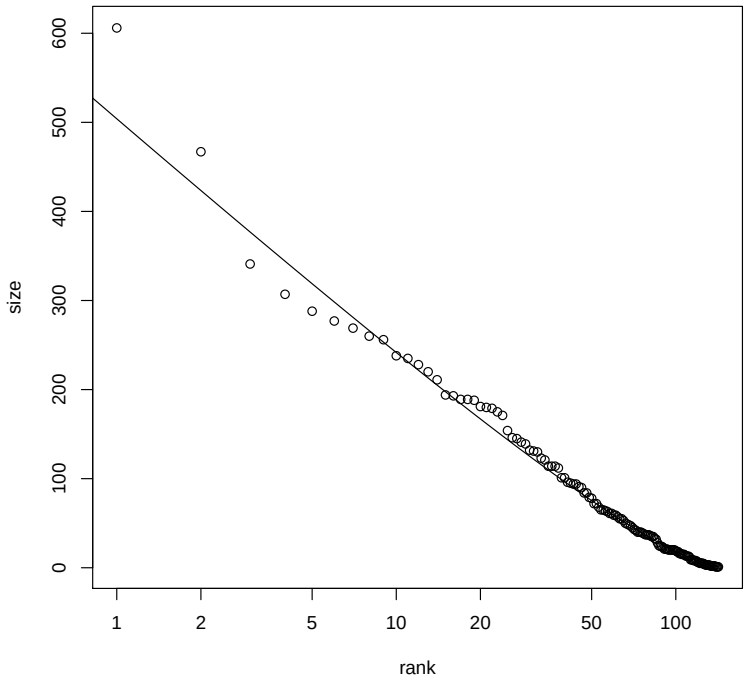

Figure 9: Rank-size plot for a Cysouw style meta-typological distribution. The x-axis is logarithmically and the y-axis is linearly scaled; fitted power law with exponential cutoff

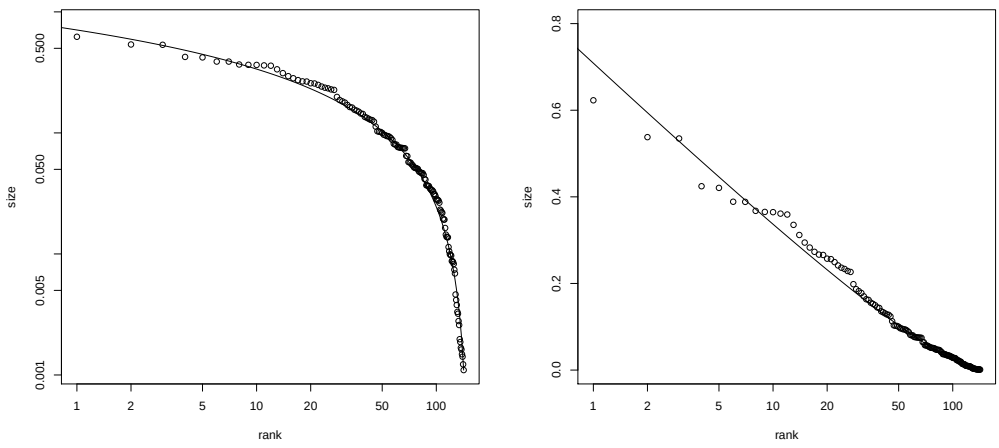

Figure 10: Rank-size plot for a Maslova style meta-typological distribution; doubly logarithmic (left) and simply logarithmic (right) rank-size plot; fitted power law with exponential cutoff

has a p-value of 0.406 and the latter of 0.532. The Sichel distribution does not increase the fit of the data significantly over the power law with exponential cutoff; the p-value for the likelihood ratio test comparing the two models is 0.98. The AIC values for log-normal distribution, power law with exponential cutoff, and Sichel distribution are $1,530.6$, $1,512.2$, and $1,514.2$ respectively. The power law with exponential cutoff thus clearly provides the best fit of the data. Note that the value for α is relatively low and the value for λ relatively high in comparison with the models that were fitted to the data from the previous section. While those distributions were rather close to a pure power law, the Cysouw style meta-typological distribution is closer to a pure exponential distribution.

Maslova uses a slightly different approach in her investigation of meta-typological distributions (see [14]). She uses the same type of sample as Cysouw, but normalizes the observed numbers by dividing by the total number of languages that are categorized for the feature at hand in WALS. In other words, she investigates the distribution of relative frequencies, rather than absolute sizes, of language types. Using the same sample as above, I replicated her experiment by normalizing my data in this way. The resulting distribution is shon in Figure 10.

Maslova hypothesizes that the resulting distribution follows a power law. If this is correct, the data points in the rank-size plot should lie close to a straight line in a doubly logarithmic plot. As can be seen from the left panel, this is evidently not correct. The right panel shows a plot with a logarithmic x-axis and a linear y-axis. As the data points do lie close to a straight line here, it is a plausible initial hypothesis that the data follow follow an underlying exponential distribution.

This impression is not confirmed by the statistical tests though. The p-value for the power law distribution is 0.00, while the p-value for the exponential distribution is 0.057, so both models can be rejected with high confidence.

The two two-parameter distributions achieve higher values: the p-value for the log-normal distribution is 0.85 (at $\mu = -3.01$ and $\sigma = 1.52$), and the p-value for the power law with exponential cutoff is 0.80 (at $\alpha = 0.42$ and $\lambda = 5.53$). The p-value for the likelihood ratio test comparing the power law with exponential cutoff to the Sichel distribution is 0.98, so the Sichel distribution does not significantly improve the fit of the model to the data.

The AIC value of the power law with exponential cutoff is -357.8, as compared to a value of -331.2 for the log-normal distribution.[7] So the AIC clearly identifies the power law with exponential cutoff as the best model for the data.

[7]Since the distribution at hand is continuous, the log-likelihood and thus the AIC values are negative now, while they are usually positive for discrete distribution such as the examples we considered so far.

We can thus conclude that these two experiments disconfirm both Maslova's and Cysouw's hypotheses about the nature of meta-typological distributions. Cysouw's hypothesis that these distributions are exponential is closer to the truth than Maslova's assumption that they follow power laws.

6 Conclusion

To briefly summarize the main finding of this article: power law distributions—or power law distributions with exponential cutoff that are very close to pure power laws—appear to be surprisingly common in cross-linguistic comparisons. Given that this result is based on a small number of typologies and just two typological data bases, it is arguably premature to draw far-reaching conclusions from this. Nonetheless, these result are perhaps interesting enough to make further studies of this issue worthwhile.

The obvious question to ask is of course which stochastic processes are responsible for typological power laws. As power laws are very common in the physical, biological and social domain, there is no shortage of models that generate power law distributions. Discussing them would go far beyond the scope of this article, and an informed decision which model is best-suited for the cross-linguistic domain will require extensive further investigation. Therefore I will not discuss this issue in depth and just entertain some general considerations.

Among the power law generating mechanisms that have been discussed in the literature are:[8]

- **Preferential attachment:** Suppose you have a collection of indivudals that are grouped into n clusters. At each time step, a new individual is added to the population. With a small probability p, the new item will start a new cluster. Otherwise, it will be added to one of the existing clusters, with a probability that is monotonically rising with the number of individuals already within this cluster. So in short, the mechanisms follows the *Matthew Effect*.[9] The sizes of the clusters will converge towards a power law distribution. This mechanism has been invoked to account for the power law distributions of city sizes, the number of biological species per genus, the number of citations of scientific papers, and many other natural and social variables. In the linguistic context, this model is arguably relevant to account for the number of speakers per language. which follows a power law with $\alpha = 1.98$ for the 73 largest languages (and a log-normal distribution for those remaining 98 languages for which these numbers are available at http://www.ethnologue.com). However, it can be shown analytically that preferential attachment always leads to power laws with $\alpha \geq 2$. (See for instance [18], page 371 for the derivation of this result.) The estimate for the number of speakers per language may still be within the margin of error. The estimates for the value of α in the typological distributions considered here all fall between 1.3 and 1.7. Consequently, we can exclude preferential attachment as an explanatory model.

- **Self-organized criticality:** Physical systems near a phase transition—like water near boiling point—exhibit power law behavior in various traits. While this holds mostly holds only for very specific parameters, there is a class of dynamical systems that are attracted towards the critical point due to the inherent dynamics of the system. This phenomenon was first discussed in [3] and dubbed *self-organized criticality*. The behavior of wild fires (see [2]) may serve as illustration. Suppose a certain area is covered with trees. The seeds of trees fall at random locations. If such a location is not not covered by trees yet, a new tree will grow there. Occasionally lightning strikes at a random location. If there is a tree at this location, it will be set ablaze, and the fire will spread to neighboring trees. If the area is densely covered with trees, a single lightning will initiate a large wild fire. If trees are scarce, only small fires will occur. At the critical point, the area contains tree clusters of all orders of magnitude, such that wild fires of all size occur. After a large fire, the tree density will be

[8] See [15, 18] for comprehensive presentations.

[9] Named after the biblical line "For to all those who have, more will be given, and they will have an abundance; but from those who have nothing, even what they have will be taken away." (Matthew 25:29).

below the critical point. Hence many seeds will grow into trees, thus increasing the density towards the critical point until fires reduce it again etc. The combination of growth and random lightning thus leads to a dynamics that is constantly attracted towards the critical point. The size of fires at the critical point is distributed according to a power law.

Self-organized criticality has been evoked as an explanation for power law behavior in systems as diverse as the magnitude of earthquakes, the intensity of wars, avalanche sizes, traffic jams, and many others.

It is intriguing to hypothesize that self-organized critical behavior lays at the root of power laws in typology as well. Justifying such an approach will require more detailed investigations though. For instance, it is a hallmark of systems in a critical state that they show *long-range correlations*. For instance, the correlation of the states of sites in the forest fire model sketched above decreases with the distance between the sites, and this dependency is again power law distributed. So a suitable next step would be to investigate whether linguistic types display a similar spatial correlation pattern.

- **Multiplicative processes:** If a random variable X is the product of many identically and independently distributed random variables Y_i with $Y_i > 0$, X will be distributed according to a log-normal distribution. (This is a direct consequence of the central limit theorem, which predicts that $\log(X) = \sum_i \log(Y_i)$ will be normally distributed.) Suppose a dynamical system is defined as follows:

$$\begin{aligned} X_0 &= Y, \\ X_{t+1} &= Y \cdot X_t, \end{aligned}$$

where Y is a random variable with $Y > 0$. For large t, X_t will be log-normally distributed.

Now consider a minor modification of this system (due to [6]), where a lower bound ε on the values of X_t is imposed:

$$\begin{aligned} X_0 &= \max(Y, \varepsilon), \\ X_{t+1} &= \max(Y \cdot X_t, \varepsilon). \end{aligned}$$

In the limit of large t, X_t will be distributed according to a power law, rather than a log-normal distribution. A similar shift from a log-normal to a power law regime also obtains if a multiplicative process is combined with a small additive noise variable.

The number of languages per type undergoes change over time due to language splitting, language death, language contact, and internal language change dynamics. It seems *prima facie* plausible to assume that these processes can be modelled by multiplicative stochastic variables. Therefore this approach is also a good candidate to account for the observed pattern in typology.

As the example in Figure 8 shows, not every heavy-tailed distribution that is generated by a partition of a large number of languages according to internal features can be approximated by a power law. It took some trial and error though to find such a feature; simpler and more natural features invariably led to distributions that could plausibly be modeled by power laws (with exponential cutoff). A reasonable starting point to gain a deeper understanding of the issues involved could be to investigate systematically which features lead to power law like distributions and which do not. The rigorous methodology from [7] to test power law hypotheses is invaluable here.

Furthermore, it is an interesting issue what consequences these findings—if they can be generalized—have for the interpretation of quantitative linguistic results. The tacit or explicit assumption of this kind of research since Greenberg's work in the 1960s (like [9]) is usually that feature values that are frequent among the languages of the world are intrinsically more likely than rare values—be it for cognitive of functional reasons. If, however, typological distributions converge towards power laws regardless of the nature of the feature involved, this conclusion does not hold

any longer. Even if all values of a given feature are cognitively and functionally equally good, we would expect to find a power law distribution with a small number of frequent feature values and a large number of rare ones. It will require some mathematical research in applied statistics as well as empirical research to figure out whether and how it is possible to draw conclusions about intrinsic asymmetries between feature values from quantitive typologies.

7 Acknowledgments

I thank Harald Baayen and Michael Cysouw for valuable comments on a previous version of the article.

References

[1] Baayen, R. H., *Word Frequency Distributions* (Kluwer Academic Publishers, Dordrecht, 2001).

[2] Bak, P., Chen, K., and Tank, C., A forest-fire model and some thoughts on turbulence, *Physics Letters A* **147** (1990) 297–300.

[3] Bak, P., Tang, C., and Wiesenfeld, K., Self-organized criticality: an explanation of $1/f$ noise, *Physical Review Letters* **59** (1987) 381–384.

[4] Berlin, B. and Kay, P., *Basic color terms: their universality and evolution* (University of California Press, Chicago, 1969).

[5] Burnham, K. P. and Anderson, D. R., *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach* (Springer, New York, 1998).

[6] Champernowne, D. G., A model of income distribution, *The Economic Journal* **63** (1953) 318–351.

[7] Clauset, A., Shalizi, C. R., and Newman, M. E. J., Power-law distributions in empirical data, *SIAM Review* **51** (2009) 661–703.

[8] Cysouw, M., On the probability distribution of typological frequencies, in *The Mathematics of Language. 10th and 11th Biennial Conference. Revised Selected Papers*, eds. Ebert, C., Jäger, G., and Michaelis, J. (Springer, 2010), pp. 28–35.

[9] Greenberg, J., Some universals of grammar with special reference to the order of meaningful elements, in *Universals of Language* (MIT Press, Cambridge, Mass., 1963), pp. 73–113.

[10] Haspelmath, M., *The World Atlas of Language Structures* (Oxford University Press, 2005).

[11] Jäger, G., Using statistics for cross-linguistic semantics: a quantitative investigation of the typology of color naming systems, ms., to appear in *Journal of Semantics* (2011).

[12] Kay, P., Berlin, B., Maffi, L., and Merrifield, W., Color naming across languages, in *Color categories in thought and language*, eds. Hardin, C. L. and Maffi, L. (Cambridge University Press, Cambridge, 1997), pp. 21–58.

[13] Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., and Cook, R., *The World Color Survey* (CSLI Publications, Stanford, 2009).

[14] Maslova, E., Meta-typological distributions, *STUF — Language Typology and Universals* **61** (2008) 199–297.

[15] Newman, M. E. J., Power laws, Pareto distributions and Zipf's law, *Contemporary Physics* **46** (2005) 232–251.

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

[17] Sichel, H. S., Word frequency distributions and tye-token characteristics, *Mathematical Scientist* **11** (1986) 45–72.

[18] Sornette, D., *Critical Phenomena in Natural Sciences* (Springer, Heidelberg, 2006).

[19] Wichmann, S., On the power-law distribution of language family sizes, *Journal of Linguistics* **41** (2005) 117–131.

[20] Wichmann, S., Müller, A., Velupillai, V., Brown, C. H., Holman, E. W., Brown, P., Sauppe, S., Belyaev, O., Urban, M., Molochieva, Z., Wett, A., Bakker, D., List, J.-M., Egorov, D., Mailhammer, R., Beck, D., and Geyer, H., The ASJP Database (version 13), http://email.eva.mpg.de/˜wichmann/ASJPHomePage.htm (2010).