J. Linguistics **41** (2005), 117–131.
doi:10.1017/S002222670400307X

On the power-law distribution of language family sizes[1]

SØREN WICHMANN

Max Planck Institute for Evolutionary Anthropology, Leipzig

(Received 27 April 2004; revised 6 September 2004)

When the sizes of language families of the world, measured by the number of languages contained in each family, are plotted in descending order on a diagram where the x-axis represents the place of each family in the rank-order (the largest family having rank 1, the next-largest, rank 2, and so on) and the y-axis represents the number of languages in the family determining the rank-ordering, it is seen that the distribution closely approximates a curve defined by the formula $y = ax^{-b}$. Such 'power-law' distributions are known to characterize a wide range of social, biological, and physical phenomena and are essentially of a stochastic nature. It is suggested that the apparent power-law distribution of language family sizes is of relevance when evaluating overall classifications of the world's languages, for the analysis of taxonomic structures, for developing hypotheses concerning the prehistory of the world's languages, and for modelling the future extinction of language families.

1. Introducing the observation

When one takes a glance at the sizes of the language families in the world recognized as genetic units by most linguists, an interesting pattern emerges. There are only a few very large families, some middle-sized ones, increasingly more small ones, and some isolates. Based on the data given in table 1[2], one can list the language families in the order of descending numbers of languages contained within each family. A curve then emerges (see figure 1) which is a good approximation to a curve produced by an equation of the type $y = ax^{-b}$ (the high R^2 value, representing the goodness-of-fit, shows the approximation to be statistically significant).

[1] Some of the ideas in this paper were helped along by brief conversations with Michael Lachmann, Roy King, and Sidney Frankel. Balthasar Bickel provided some useful comments on an oral presentation at the Max Planck Institute for Evolutionary Anthropology, 11 May 2004, and so did several of the participants in the workshop on phylogeny at the Max Planck Institute for Mathematics in the Sciences, 24 May 2004, where I also presented the paper. Finally, two anonymous *JL* referees as well as Lada Adamic and Merritt Ruhlen commented on the manuscript. All errors and inadequacies are mine.

[2] Unclassified languages (as well as creoles, pidgins, mixed languages, sign languages, and artificial languages) are excluded from consideration here.

(1489) Niger-Congo	(33) Geelvink Bay	(7) Left May	(2) Lower Mamberamo
(1262) Austronesian	(33) Penutian	(6) Maku	(2) Harakmbet
(552) Trans-New Guinea	(32) Macro-Ge	(6) Muskogean	(2) Peba-Yaguan
(443) Indo-European	(32) Hmong-Mien	(6) Kwomtari-Baibai	(2) Yenisei Ostyak
(372) Afro-Asiatic	(30) Panoan	(6) Kiowa-Tanoan	(2) Arutani-Sape
(365) Sino-Tibetan	(29) Carib	(6) Tacanan	(2) Amto-Musan
(258) Australian	(29) Khoisan	(6) Witotoan	(2) Zamucoan
(199) Nilo-Saharan	(28) Hokan	(5) Caddoan	(2) Alacalufan
(172) Oto-Manguean	(27) Salishan	(5) Chukotko-Kamchatkan	(2) Araucanian
(168) Austro-Asiatic	(26) West Papuan	(5) Mascoian	(2) Yukaghir
(104) Sepik-Ramu	(25) Tucanoan	(5) Guahiban	(2) Yuki
(75) Dravidian	(22) Chibchan	(5) South Caucasian	(2) Uru-Chipaya
(70) Tupi	(17) Siouan	(5) Wakashan	(2) Keres
(70) Tai-Kadai	(16) Mixe-Zoque	(5) Nambiquaran	(2) Cahuapanan
(69) Mayan	(13) Andamanese	(5) Chapacura-Wanham	(2) Salivan
(65) Altaic	(12) Japanese	(4) Huavean	(2) Chon
(62) Uto-Aztecan	(11) Totonacan	(4) Gulf	(2) Bayono-Awbono
(60) Arawakan	(11) Mataco-Guaicuru	(4) Misumalpan	(1) Coahuiltecan
(48) Torricelli	(11) Eskimo-Aleut	(4) Subtiaba-Tlapanec	(1) Paezan
(47) Na-Dene	(10) Choco	(4) Jivaroan	(1) Lule-Vilela
(46) Quechuan	(10) Iroquoian	(4) Yanomam	(1) Chimakuan
(40) Algic	(8) Arauan	(3) Katukinan	(1) Mura
(38) Uralic	(7) Chumash	(3) Basque	(1) Mosetenan
(36) East Papuan	(7) Sko	(3) Aymaran	(1) Cant
(34) North Caucasian	(7) Zaparoan	(3) East Bird's Head	(30) Other language Isolates
	(7) Barbacoan		

Table 1

A ranking of the world's language families in terms of number of languages according to data from *Ethnologue* (Grimes 2000)

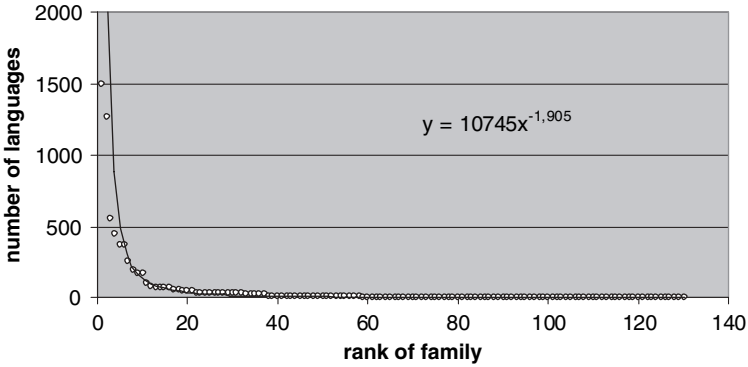

Figure 1
Language family sizes in *Ethnologue* (Grimes 2000) ($R^2 = 0.957$)

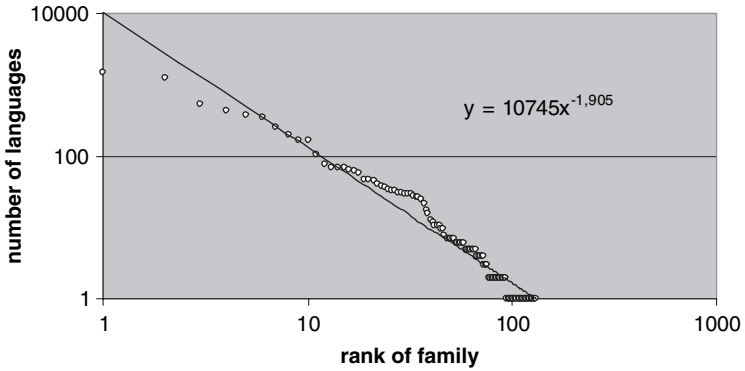

Figure 2
Language family sizes in *Ethnologue* (Grimes 2000) plotted on a log–log scale ($R^2 = 0.957$)

When plotted on a log–log scale, the distribution will approximate a straight line, as seen in figure 2.

A distribution of this kind is called a 'power-law' distribution, known to characterize a wide range of phenomena in both the natural world and the social world. Thus – citing a major, early publication for each individual field – power laws have been found in urban conglomerations (Auerbach 1913), the abundance of biological taxa (Yule 1924), word frequencies (Zipf 1949), the distribution of personal income (Champernowne 1953), the size of earthquakes (Kanamori & Anderson 1975), the popularity of Internet sites (Glassman 1994), the activity of genes (Ueda et al. 2004), and many other areas.

A comparison of figure 2 with figure 3 gives a graphic illustration of the non-coincidental nature of the distribution of language family sizes. Figure 3

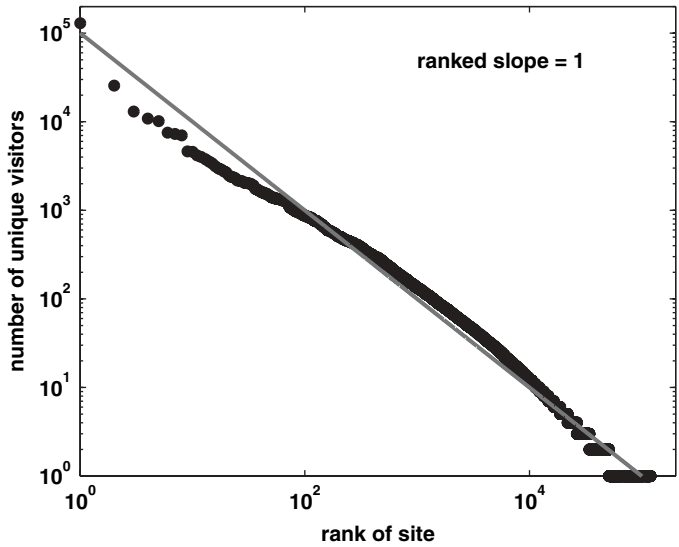

Figure 3
Sites ranked by the number of unique AOL visitors they received on 1 December 1997 (after Adamic & Huberman 2002: figure 2)

represents an example of a power-law distribution in the world of social interaction. It shows the distribution of Internet sites ranked by popularity, i.e. the relationship between the rank-order of internet sites in terms of popularity and the number of visitors which determine the rank-order. The x-axis represents the rank-order. The most popular site has rank No. 1 and is thus placed at the far left end of the x-axis. The farther to the right a site is plotted on the axis, the less popular it is. Popularity is measured by the number of visitors that a given site received during a single day (1 December 1997); these numbers are plotted on the y-axis. The same principle of size–rank organization is used in both figures. In figure 2 the x-axis ranks language families according to size and the y-axis shows the number of languages in the families ranked; in figure 3 the x-axis ranks Internet sites according to number of hits and the y-axis plots the corresponding numbers. The figures are remarkably similar even with regard to the types of small deviations from the straight line.

2. Some background

Power laws are not something alien to linguistics. An instance of the phenomenon is also known as 'Zipf's law', named after the linguist George Kingsley Zipf, who first observed that absolute word frequencies (P) are inversely proportional to their rank (r): $P \sim r^{-b}$, where b is close to 1. He found

the law to apply to a variety of social phenomena and ascribed it (without really presenting substantiating evidence) to a principle of least effort (Zipf 1949). Zipf's law is a special instance of a power law, where $y = ax^{-b}$. In Zipf's law $b = 1$. A kindred type of quantitative observation in linguistics is due to Menzerath, who stated that a sound is the shorter the longer the whole in which it occurs and that the more sounds in a syllable the smaller its relative length (Menzerath 1928: 104). Several scholars have since developed Menzerath's ideas, including Altmann (1980, 1986).

Since $y = ax^{-b}$ is equivalent to $\log(y) = -b \log(x) + \log(a)$, a power-law distribution can be depicted as a straight line with a slope –b on a log–log plot, accounting for the kinds of patterns seen in figures 2 and 3. There are mathematical developments which are claimed to allow us to approximate even more closely the sorts of distributions seen in figures 2 and 3, which are characterized by 'fat tails' deviating from the straight line. Such slightly 'deviant' distributions not only characterize the popularity of Internet sites and (apparently) language family sizes but also many other phenomena, such as the distributions of radio and light emission from galaxies, country population sizes, temperature variations, citations of physicists, etc. (Laherrere & Sornette 1998). However, for the purposes of the present paper, which is concerned with the identification of a single power-law-related phenomenon and less with the ultimate explanation for all such phenomena, it would be excessive to discuss these developments further.

Among the various explanations for power-law distributions that have been proposed, the most recent ones seem to be informed by data from studies of Internet usage and derive such distributions from networks that expand continuously and where new connections (vertices) attach preferentially to nodes that are already well connected (Barabási & Albert 1999).

In the work of the physicist Per Bak and his associates, conveniently explained to the lay audience in Bak (1996), power-law distributions are viewed as characteristic of self-organizing systems, and models for deriving such distributions are described, notably the sand-pile model. In simulations and actual experiments, it can be shown that the continuous addition of grains of sand to a pile will produce avalanches which, when the system has entered into a state of so-called 'self-organized criticality', are of varying sizes. Rank-ordering of avalanche sizes will produce a power-law distribution. Even though the process can be simulated on a computer it has, surprisingly, not been possible to devise a mathematical formula which would produce the same results as the simulation (Bak 1996: 63).[3]

A third way of arriving at power-law distributions is based on stochastic branching models and thus, intuitively, seems similar to the kind of process

[3] Bak (1996: 63) also mentions the difficulties in understanding mathematically the exponent that determines the slope of the line on the log–log plot. With regard to this problem there has been some progress in recent years, cf. Bollobás et al. (2001).

we are dealing with when studying the developments of language families. The distributions are derived from the simple branching process known as the Galton-Watson process. This process is named after Rev. Henry William Watson, who posed the problem of why British surnames tended to disappear, and Sir Francis Galton (a half-cousin of Charles Darwin), who provided a solution to the problem (Watson & Galton 1875).

The scenario is as follows. At any given taxonomic level an entity has the probability P_0 of producing no offspring, the probability P_1 of producing one offspring, the probability P_2 of producing two offspring, etc. The mean number of offspring, m, is the sum of the set of probabilities P_i times offspring i. This may be expressed by the formula in (1) (cf. Chu & Adami 1999).

(1) $$m = \sum_{i=0}^{\infty} i \cdot P_i$$

As an aid to understanding for readers unfamiliar with mathematical symbols we can give an example. If the probability set is as in (2) the mean number of offspring m will be 1.65.

(2) *Example of a probability set*

i	0	1	2	3	4	5
P_i	0.1	0.5	0.2	0.1	0.05	0.05

If $m > 1$ (as in example (2)), the family will likely grow; if $m = 1$, it will converge towards extinction over infinite time; and if $m < 1$, the family is certain to eventually become extinguished. Essentially, a power-law distribution is arrived at by an iteration of the initial branching process. According to Chu & Adami (1999), the closer m is to 1, the closer the curve will approximate a power law, while the further away it is from 1, the closer it will approximate an exponential curve. On the other hand, different kinds of probability sets will not affect the shape of the rank-frequency curve, as long as m remains the same. That is, numbers other than those in (2) could apply and we would still obtain a power-law distribution.[4]

Processes of preferential attachment and processes of branching seem to tell the same story, only from opposite perspectives: branching theory describes the branching off of nodes and network theory describes the connections of branches to nodes. The two types of explanations can probably be shown mathematically to be equivalent, although a formal proof of this exceeds my competence. Both describe continuous processes that do not allow for catastrophic behavior of the systems. In this regard they

[4] Reed & Hughes (2002) criticize the model of Chu & Adami for being unrealistic and present their own, somewhat more involved model of macroevolution which, nevertheless, also builds on branching theory.

may be less adequate descriptions of reality than the sand-pile model. The point of this little discussion, however, is not to show how we might arrive at a perfect mathematical model for deriving the kind of distribution that characterizes language family sizes. Rather, the bottom line is simply that different models exist and that, despite differences in the approaches, they all are based on the assumption that the process is in essence stochastic and not due to external causation or to qualitative factors inherent in the various systems that exhibit power-law distributions.

3. Implications

The following list of implications of the above findings is not exhaustive, nor is it conclusive. The list is rather intended as a preliminary set of research questions which might fruitfully be addressed in the future.

3.1 *Size-rank distributions as a tool for assessing the consistency of across-the-board classifications*

Most readers will have taken note of the fact that the empirical data on which the initial observations of this study were based were taken from Grimes (2000) – henceforth '*Ethnologue*'. Why *Ethnologue* and not some other classification? This was accidental in the sense that I did not look for a power-law distribution of language family sizes, but simply happened to see it as I was plotting the data from *Ethnologue* for other purposes. On the other hand, the reason why I found that dataset interesting to look at in the first place is not accidental. *Ethnologue* is one of the few complete listings of the world's languages, it may be the only one which is continuously updated, and it represents as close as one may hope to come to a consensus view regarding the classification of these languages. Although many linguists (including myself) will have some differences with *Ethnologue* concerning either classifications or the numbers of languages given for each family, it is probably true to say that this source roughly represents the state-of-the-art of the application of standard methods of comparative linguistics as well as the current knowledge concerning numbers of languages for each family – or at least counts that represent a consistent set of considerations. There are no universally accepted criteria for distinguishing languages from dialects. Often the *Ethnologue* will list as languages entities that some other classifications consider to be dialects. This means that the number of languages in each family will sometimes be lower in other sources. As long as the tendency is consistent such differences will not, however, affect the usefulness of the dataset for most statistical purposes.

While the apparent power-law distribution of language family sizes was initially surprising to me, I have now come to realize that this is exactly what we would expect. That the *Ethnologue* data fit the expected distribution

(1175) Austric	(241) Afro-Asiatic	(34) Na-Dene	(5) Chukchi-Kamchatkan
(1064) Niger-Kordofanian	(170) Australian	(31) Khoisan	(1) Basque
(731) Indo-Pacific	(144) Indo-Hittite	(28) Elamo-Dravidian	(1) Burushaski
(583) Amerind	(138) Nilo-Saharan	(24) Uralic-Yukaghir	(1) Ket
(258) Sino-Tibetan	(63) Altaic	(9) Eskimo-Aleut	(1) Nahali
	(38) Caucasian		

Table 2
A ranking of the world's language families in terms of number of languages according to data from *A Guide to the World's Languages* (Ruhlen 1987)

thus inspires some confidence in the overall statistical picture even if this distribution in itself does not vouch for the correctness of all aspects of the classification. The fact that an overall statistical distribution may remain constant even if a classification is improved has been noted in a paper by Sepkoski (1993). On two occasions, separated by a span of ten years, this author studied the origination and extinction rates of biological taxa found in the fossil record. Even though several reclassifications and a lot of new knowledge in general had accumulated during the lapse of the decade, the author found almost identical distributions on both occasions. The question now arises how far one's classifications may deviate from accuracy and yet conform to a neat, straight line on a log–log scale. While it seems impossible to give a precise answer to this question, it may at least be tested empirically whether other classifications do or do not produce power-law distributions. In this regard the classification of Ruhlen (1987) is very useful and interesting since it conveniently provides a classification of all the world's languages which in many respects differs from that of *Ethnologue*, and at the same time provides data for the number of languages within each family. Moreover, the classification is of intrinsic interest since it has had a significant impact in the scientific world, for instance among population geneticists. The figures given by the author for numbers of languages in each family and his labels for the genetic units proposed are reproduced in table 2.[5]

Just as we did for *Ethnologue*, we now plot the data from *A guide to the world's languages* in a normal x/y diagram, where x is the descending rank in terms of numbers of languages for a given family and y is the number that determines that rank. Next, we convert these axes to logarithmic ones. The results are shown in figures 4 and 5, respectively.

A visual comparison of figures 1–2 and figures 4–5 shows that Ruhlen's data diverge more from the expected distribution than the *Ethnologue* data. It is readily seen in figure 5 that the best approximation to the points in

[5] As in table 1, unclassified languages as well as creoles, etc. are excluded from consideration.

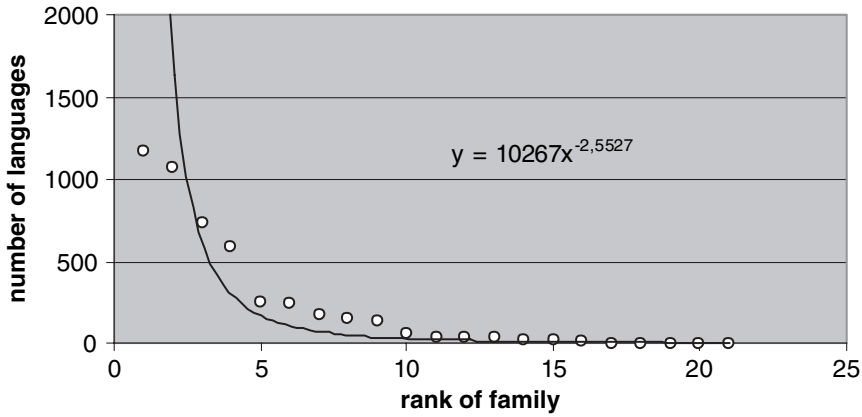

Figure 4
Language family sizes from Ruhlen (1987) ($R^2 = 0.782$)

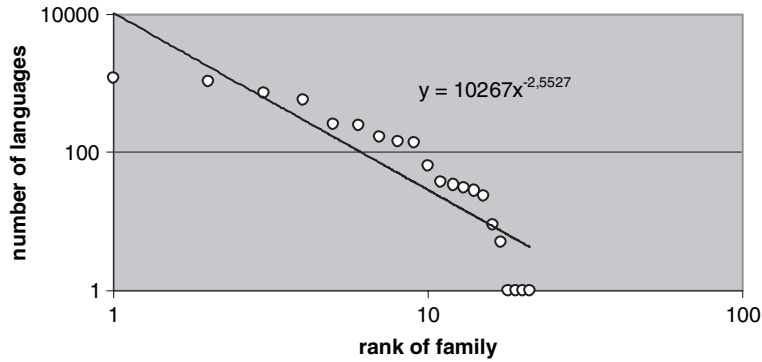

Figure 5
Language families from Ruhlen (1987) plotted on a log–log scale ($R^2 = 0.782$)

the diagram is clearly not a straight line but, rather, a curve. The visual impression is sustained by the low R^2-value, representing the goodness-of-fit. This value is 0.782, which compares unfavorably with the value 0.957 of the *Ethnologue* data. There are two possible explanations for this. Either some of Ruhlen's language families are in fact not valid genetic units or the units do not all belong to the same taxonomic stratum (both explanations may, of course, apply simultaneously).

Independently of one's opinion of the various specific genetic classifications that have been proposed, I think one is forced to draw the conclusion that mapping the rank-by-size distribution is in fact useful for gauging the overall plausibility and consistency of a genetic classification. I would like to add, however, that the results of my exercise with the particular data in

Ethnologue and *A guide to the world's languages* are not results that I had predicted or aimed at when I embarked upon this study, nor should they be considered more than results open to interpretations.

3.2 *Taxonomic structure*

An important property of power-law distributions is that they are self-similar or 'scale-free'. That is, we can zoom in on any stretch of the line in the log–log diagram and we still just see a line.[6] In the case of the development of taxa, whether biological or linguistic, this means that we should expect to find similar relationships as we move successively up and down different taxonomic levels. Thus, across the range of language families we should expect the number of languages contained within the lowest subgrouping level to also represent a power-law distribution when they are plotted on an x/y scale in ranked order, and similarly for successively higher taxonomic levels. In particular, it is interesting that we may expect for the highest taxonomic level generally recognized, i.e. reconstructed nodes such as proto-Indo-European, proto-Niger-Congo, proto-Sino-Tibetan, proto-Austronesian, etc., that a few of these will have many branches, some an intermediate number of branches, and many just a few. I have made a preliminary test of the results of the genera-to-family and language-to-genera distributions using the list of genera proposed by Matthew Dryer.[7] Both distributions approach power-law distributions. More work on the issue of distributions at different taxonomic levels is required, however. One issue in particular needs to be addressed, namely, whether we should expect same or different values of $-b$ (the exponent) as we move up and down taxonomic levels and, in general, what the exponent tells us regarding the structure of individual trees or the ecologies in which the world's languages participate during different historical periods.

The fact that statistical distributions replicate themselves at different taxonomic levels suggests that historical linguists should pay more attention to mathematical models of branching structures and that simulations need to be developed in order to improve our understanding of branching.

---

[6] Obviously, for extrinsic reasons, most systems have an upper bound. For instance, there can be no earthquakes larger than one which would result in the fragmentation of the entire earth, nor can there be more language families than there are individuals on earth. The deviations from power-law distributions often seen at the top extreme of the curve are usually explained as symptoms of the approximation to a situation where scale-freedom breaks down and absolute quantities become an issue.

[7] Dryer has published the list on his home page (see http://wings.buffalo.edu/soc-sci/linguistics/people/faculty/dryer/dryer/genera). As a source for languages Dryer uses *Ethnologue*, while the classification into genera is his own. Eventually a similar list of genera will be included in the publication of the World Atlas of Language Structures, carried out at the Max Planck Institute for Evolutionary Anthropology. Thanks go to Hans-Jörg Bibiko for initial assistance in plotting these data.

3.3 *Does the existence of large language families need explanation?*

The answer to the question posed in the heading is simple: no. It is an inherent feature of the expected distribution of language family sizes that there will be just a few large families, given the nature of power laws. As shown in the work of Bak and his associates (cf. Bak 1996), self-organized systems will eventually arrive at power-law distributions without any outside influence. Intuitively, one might think that the existence of certain very large families is something that cries out for an explanation, while it is actually the case that the ABSENCE of such 'freak' families would be unexpected. Now, an explanation for the existence of certain large language families has actually been proposed, namely the theory that such families tend to correlate with early farming (Bellwood 2001, 2002; Renfrew 2002). Independently of the strength of the evidence of such a theory, one could ask whether external causation is at all compatible with the existence of a system conforming to Bak's model of 'self-organized criticality'. That is, does the absence of a need for an explanation in terms of external causation imply that such an explanation would necessarily be wrong? The answer is of course that we cannot preclude that there were external causes that influenced the sizes of different language families during the history of the evolution of languages. In fact, we know that some families have been greatly reduced during historical times because of conquests. But just as we do not NECESSARILY need external factors like meteors to explain biological mass extinctions like that of the dinosaurs (Bak 1996: 151–153), we also do not NECESSARILY need agriculture to explain why, say, Niger-Congo is so big.

3.4 *The world's languages participate in a global ecology*

A common characteristic of phenomena obeying power-law distributions is that they involve sets of entities that are qualitatively similar (i.e. as similar as one grain of sand to another or one biological species to another) and, crucially, that they interact so as to make up an ecology. This means that we should view the languages in the world as an ecological system where the extinction of certain languages or the development of others will have repercussions elsewhere in the system. Just as earthquakes cannot be predicted by studying the history of earthquakes, we cannot predict the development of language families from observations of current changes in distributions; but we can at least infer that the developments are interlaced.

3.5 *Are power laws ubiquitous in quantitative distributions relating to languages?*

Given that power laws are so widespread in both the physical world and the social world, they could be widespread in many quantitative distributions

relating to languages as well. Accordingly, pointing out the power-law distribution of language family sizes might be argued not to constitute a significant scientific step forward. Undeniably, the observation was indeed within close reach and would have been made sooner or later. But stating what, in retrospect, appears obvious is important, since it points the direction of inquiry to less obvious facts and raises questions about phenomena that earlier did not seem to require special explanations. In the present context, quantitative distributions relating to languages which DO NOT follow power-law distributions would constitute a novel issue.

As it turns out, power laws are not the only type of distribution that characterizes quantitative phenomena relating to the world's languages. Prompted by my findings regarding language family sizes, I investigated whether it might be the case that the sizes of the world's languages in terms of number of speakers had the same kind of distribution. This is not the case. Figure 6a shows the curve obtained when the numbers of speakers of each of the world's languages (on the y-axis) are plotted against their rank (on the x-axis). Figure 6b represents the same data plotted on a log–log diagram. The data were, again, obtained from *Ethnologue*, which provides estimates of numbers of speakers for most languages in the catalogue.[8] The curve obtained does not even approximate a power-law distribution, nor is it exponential, logarithmic or otherwise mathematically simple. As is perhaps most clearly seen in figure 6a, the shape is due to a divide between a small number of very large languages and a large number of small languages. Thus, 5.3% of the languages have from a million (10^6) to a billion (10^9) or slightly more speakers, and the rest have less than a million speakers. Depending on the viewpoint, the deviation from the power-law distribution may be seen as due either to an overly small number of languages in the middle range or to an overly high number of languages in the lower range. If the power-law distribution is the normal, expected kind of distribution, then the picture we see is one of imbalance. Indeed, as is well known, we are currently in a phase – mostly due to the continuing effect of European colonization – where very many of the world's languages are being extinguished or verge on extinction. When this phase is over, i.e. when the currently endangered languages are gone, it may be the case that the number-rank distribution of speakers per language will assume a power law. Such a

[8] *Ethnologue* gives estimates of numbers of speakers for 6,142 languages. For another 768 languages estimates are lacking. When plotting the data I have made some small modifications. When estimates are given as a range (e.g. 500–1,000) I have used the mid-point of the range (e.g. 750) for the plot. In rare cases estimates are formulated non-numerically. The estimate 'few' has been converted to the number 50, and the estimate 'very few' to 5. Whenever the estimates distinguish between mother-tongue and second-language speakers, the figure plotted is that of mother-tongue speakers. I heartily thank Michael Lachmann for creating a script for exporting the on-line *Ethnologue* data to a file and also Luise Vörkel for her heroic assistance with the data.

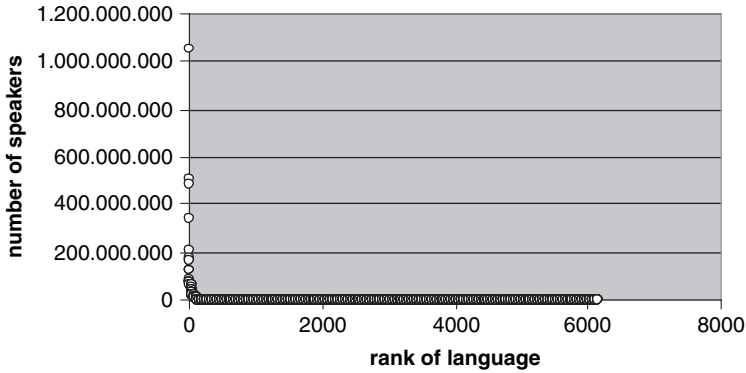

Figure 6a
Rank-ordering of language sizes measured as numbers of speakers according to *Ethnologue* (Grimes 2000)

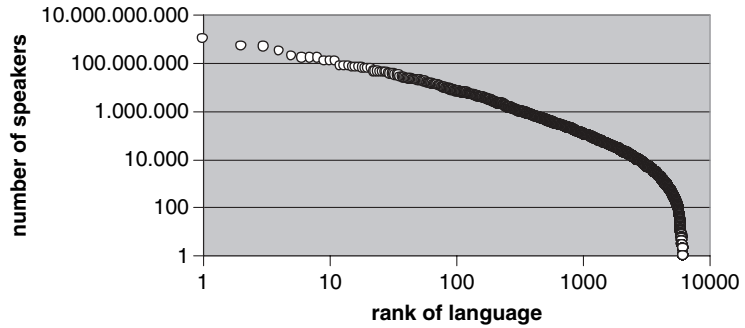

Figure 6b
Rank-ordering of language sizes on a log–log scale

distribution may have existed in earlier times as well. As a matter for future research, computer simulations might be used to test whether this hypothesis is viable.

It seems to be the case that the rank-size curve for language families (measured in terms of numbers of languages) has a greater degree of inertia than the corresponding curve for language sizes (measured in terms of numbers of speakers). Possibly, the former will in the future assume the current shape of the latter.

In this section we have seen that while power-law distributions are widespread they are not ubiquitous. Thus, we need to learn more about the conditions that govern their presence not only to explain the power laws themselves, but also to explain cases where they are absent.

3.6 *Modelling the prehistory and future extinction of language families*

A simple prediction that follows from our results is that throughout most of the history of the evolution of human languages there was a distribution similar in nature to that of modern language families, with just a few large ones, some intermediate ones, and several small ones. Obviously, the families must overall have been smaller than today, but proportionately the same sort of distribution is to be expected. It is likely that neolithic revolutions across the globe reshuffled the rankings of language families with respect to their sizes. Perhaps some of the world's large hunter-gatherer families were the largest families in the times immediately preceding global neolithics. Minimally, it now seems possible to at least begin to try to imagine what the global language situation would have looked like, say, in the period 20,000–10,000 BP, after language had evolved sufficiently so as to result in several language families and before agriculture became widespread and began to cause radical changes in the global population distribution.

In connection with the observations of the previous two subsections, we might also raise the question of whether it is possible to model FUTURE distributions of language families. Again it should be emphasized that the mechanics inherent in their development do not allow for predictions with regard to which language families will grow and which will become extinct. Nevertheless, it is possible to predict that despite possible catastrophic events, in some distant future the overall distribution will again come to resemble a power law. Not only the generation of taxonomic entities but also their extinction tend to follow power laws (Sepkoski 1993). Indeed, as already mentioned, the first study to provide a mathematical foundation for the kinds of distributions that we have been observing addressed the question of extinction (Watson & Galton 1875). Thus, based on the likely development of certain healthy families, like Indo-European, and the likely extinction of less viable entities – small families scattered throughout the world – we might create simulations of the overall future development of linguistic diversity. It is possible, though this has yet to be tested, that one might in the future expect the curve representing the rank-order of language family sizes (measured in terms of number of languages) to enter into an unstable phase where it perhaps assumes the shape of the curve representing the current shape of language sizes (measured in terms of numbers of speakers).

REFERENCES

Adamic, L. A. & Huberman, B. A. (2002). Zipf's law and the Internet. *Glottometrics* **3**. 143–150.

Altmann, G. (1980). Prolegomena to Menzerath's law. In Grotjahn, R. (ed.), *Glottometrika* **2**. Bochum: Studienverlag Dr. N. Brockmeyer. 1–10.

Altmann, G. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Studienverlag Dr. N. Brockmeyer.

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geografische Mitteilungen* **59**. 73–76.

Bak, P. (1996). *How nature works: the science of self-organized criticality*. New York: Copernicus.
Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**. 509–512.
Bellwood, P. (2001). Early agriculturalist population diasporas? Farming, languages and genes. *Annual Review of Anthropology* **30**. 181–207.
Bellwood, P. (2002). Farmers, foragers, languages, genes: the genesis of agricultural societies. In Bellwood & Renfrew (eds.), 17–28.
Bellwood, P. & Renfrew, C. (eds.) (2002). *Examining the farming/language dispersal hypothesis*. Cambridge: McDonald Institute for Archaeological Research.
Bollobás, B., Riordan, O., Spencer, J. & Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures and Algorithms* **18**. 279–290.
Champernowne, D. G. (1953). A model of income distribution. *The Economic Journal* **63**. 318–351.
Chu, J. & Adami, C. (1999). A simple explanation for taxon abundance patterns. *Proceedings of the National Academy of Sciences of the United States of America* **96**. 15017–15019.
Glassman, S. (1994). A caching relay for the World Wide Web. *Computer Networks and ISDN Systems* **27**. 165–173.
Grimes, B. F. (2000). *Ethnologue: languages of the world* (14th edn.). Dallas, TX: Summer Institute of Linguistics.
Kanamori, H. & Anderson, D. L. (1975). Theoretical basis of some empirical relations in seismology. *Bulletin of the Seismological Society of America* **65**. 1073–1095.
Laherrere, J. & Sornette, D. (1998). Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales. *European Physics Journal* **B2**. 525–539.
Menzerath, P. (1928). Über einige phonetische Probleme. *Actes du premier congrès international de linguistes*. Leiden: Sijthoff. 104–105.
Reed, W. J. & Hughes, B. D. (2002). On the size distribution of live genera. *Journal of Theoretical Biology* **217**. 125–135.
Renfrew, C. (2002). 'The emerging synthesis': the archaeogenetics of farming/language dispersals and other spread zones. In Bellwood & Renfrew (eds.), 3–16.
Ruhlen, M. (1987). *A guide to the world's languages*, vol. 1: *Classification*. Stanford, CA: Stanford University Press.
Sepkoski, J. J., Jr. (1993). Ten years in the library: new data confirm paleontological patterns. *Paleobiology* **19**. 43–51.
Ueda, H. R., Hayashi, S., Matsuyama, S., Yomo, T., Hashimoto, S., Kay, S. A., Hogenesh, J. B. & Iino, M. (2004). Universality and flexibility in gene expression from bacteria to human. *Proceedings of the National Academy of Sciences of the United States of America* **101**, **11**. 3765–3769.
Watson, H. W. & Galton, F. (1875). On the probability of extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland* **4**. 138–144.
Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London* (B) **213**. 21–87.
Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Author's address: *Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-014103 Leipzig, Germany.*
E-mail: soerenw@hum.ku.dk