On the probability distribution of typological frequencies

Michael Cysouw

Max Planck Institute for Evolutionary Anthropology, Leipzig
cysouw@eva.mpg.de

Abstract. Some language types are more frequent among the world's languages than others, and the field of linguistic typology attempts to elucidate the reasons for such differences in type frequency. However, there is no consensus in that field about the stochastic processes that shape these frequencies, and there is thus likewise no agreement about the expected probability distribution of typological frequencies. This paper explains the problem and presents a first attempt to build a theory of typological probability purely based on processes of language change.

1 Probability distributions in typological research

A central objective of the typological study of linguistic diversity is to explain why certain kinds of linguistic structures are much more frequently attested among the world's languages than others. Unfortunately, such interpretations of empirically attested frequencies often rely on purely non-mathematic intuitions to judge whether observed frequencies are in any sense noteworthy or not. In the typological literature, this frequently leads to a tacit assumption that typological frequencies are evenly distributed, i.e. that *a priori* all language-types should be equally frequent, and any observed skewing of frequencies is thus in need of an explanation. Such an argumentation can be found, for example, in the widely read typological textbook by Comrie [2]:

"In a representative sample of languages, if no universal were involved, i.e. if the distribution of types along some parameter were purely random, then we would expect each type to have roughly an equal number of representatives. To the extent that the actual distribution departs from this random distribution, the linguist is obliged to state and, if possible, account for this discrepancy" (p. 20)

Various more sophisticated approaches to the interpretation of empirical frequencies often assume an underlyingly normal (or, more precisely, multinomial) distribution, as for example indicated by the regular use of χ^2 statistics or Fisher's exact test (e.g. in Cysouw [3]). Janssen et al. [6] and Maslova [12] explicitly discuss the problem of tacit assumptions of the underlying probability distributions as made in linguistic typology. As a practical solution to circumvent this problem for the assessment of statistical significance, Janssen et al. [6]

propose to use randomization-based significance tests. Such tests do not make any assumptions about the underlying probability distribution. This of course still leaves open the question about the nature of these distributions.

There is a small literature that explicitly deals with the question of the underlying probability distribution of typological variables, but there is no agreement whatsoever. The first paper to make any such proposal was Lehfeldt [9], who proposes a gamma distribution for the size of phoneme inventories. In a reaction to this claim Justeson and Stephens [7] proposed a log-normal distribution for the same data. The size of the phoneme inventory is a clear case of linguistic complexity. More general, Nichols [14] (and more recently Nichols et al. [15]) proposed a normal distribution for linguistic complexity. Similarly, Maddieson [10] hints at a normal distribution for the size of consonant inventories. Finally, Maslova [12] argues that the frequencies of types in the *World Atlas of Language Structures* (henceforth WALS, [5]) follows a pareto distribution. There are thus at least proposals for gamma, log-normal, normal and pareto distributions of typological variables.

Most of these proposals arrive at their distribution on the basis of the inspection of empirical values. For example, the only argument Lehfeldt (1975) offered for the gamma distribution is a (rather speculative) interpretations of some moment-like characteristics of the empirical frequencies. Nichols et al. [15] only observe bell-shaped distribution, and propose a normal distribution on that meagre basis. Though empirical distributions might *suggest* a particular underlying probability distribution, but they can never be used to *argue* for a particular distribution. Both the gamma distribution and the log-normal distribution can be fitted easily to the empirically observed phoneme-size distribution. The proper argument for a particular probability distribution is the explication of the stochastic process that causes the distribution to arise. This approach was used by Justeson and Stephens [7] in their plea for a log-normal distribution of phoneme inventory size. Phoneme inventories, they argue, are based on phonological feature inventories. Given n binary features, it is possible to compose 2^n phonemes. Now, assuming that feature inventories are normally distributed (a claim they do not further elucidate), phoneme inventories will thus be log-normally distributed (i.e. the logarithm of the phoneme inventory size will be normally distributed). Irrespective of the correctness of their claim, this argument is a good example of an attempt to find a stochastic reason for a particular distribution. The actual proposal of Justeson and Stephens is not convincing, because they do not substantiate the normal distribution of feature inventories. Still, their approach is an important step in the right direction.

2 The stochastic process of language change

To investigate the nature of any underlying probability distribution it is necessary to consider the stochastic process that causes the phenomenon at hand. For typological frequencies there are at least two (non-exclusive) kind of processes that can be considered. The frequencies of linguistic types in the world's

languages are partly shaped by cognitive processes, and partly by diachronic processes. In this paper I will restrict myself to further investigate the latter idea, namely that the *process of language change* determines the probability distribution of typological frequencies.

The synchronic frequencies of a typological variable (for example the word order of verb and object cf [4]) can be seen as the result of the diachronic processes of language change (cf. Plank and Schellinger [16]). More precisely, the current number of languages of a particular linguistic type can be analyzed as the result of a Markov process in which language change from one type to another sequentially through time (cf. Maslova [11]). For example, a verb-object language can change into a object-verb language, and vice-verse, and this process of change from one type to the other determines the probability distribution of the linguistic type.

As a first (strongly simplified) approach to the stochastic nature of this process of type-change, I will in this paper consider type-change as a simple birth-death process: a verb-object language is "born" when an object-verb language changes to a verb-object language, and a verb-object language "dies" when this language changes to an object-verb language.[1] Also as a first approximation, I will assume that such type-changes take place according to a Poisson process. A Poisson process is the stochastic process in which events occur continuously and independently of one another, which seems to be a suitable assumption for language change.

Such a basic birth-death model with events happening according to a Poisson distribution is known in queueing theory as an M/M/1 process (using the notation from Kendall [8] in which M stands for a "Markovian" process). Normally, queueing models are used to describe the behavior of a queue in a shop. Given a (large) population of potential buyers, some will once in a while come to a cash register to pay for some goods (i.e. a "birth" in the queue) and then, possibly after some waiting time, pay and leave the queue again (i.e. a "death" in the queue). The queueing model also presents a suitable metaphor to illuminate the dynamics of typological variables. Consider all the worlds languages throughout the history of *homo loquens* as the (large) population under investigation. Through time languages change from one type to another type, and vice-versa. Metaphorically, one can then interpret the number of languages of a particular type at a particular point in time as the length of a queue.

A central parameter of a queueing model is the traffic rate t, which is defined as the fraction of the arrival rate λ and the and the departure rate μ: $t = \lambda/\mu$. The arrival and the departure rate designate the average number of arrivals and

[1] Altmann [1] also uses a birth-death model to investigate distributions in a cross-linguistic contexts. However, he investigates another kind of phenomenon, namely the probability distribution of the number of different types in dialectological maps. This aspect is closely related to the number of types per map in a typological atlas like WALS, though there is more arbitrariness in the number of types in a typological map compared to the number of types in a dialectological map. Altmann convincingly argues that the number of types on dialectological maps should be negatively binomially distributed.

departures in the queue per time unit. However, time is factored out in the traffic rate t, so this rate is just a general indication of the dynamics of the queue. In a stable queue, the traffic rate must be between 0 and 1 (a traffic rate larger than one would result in an indefinite growing of the queue). Now, in an M/M/1 model with traffic rate t, the probability distribution of the queue length q is distributed according to (1), which is a slight variation on a regular (negative) exponential distribution (cf. Mitsenmacher and Upfal [13], p. 212). Following this model, typological frequencies should accordingly be roughly (negatively) exponentially distributed.

$$P(q = n) = (1 - t) \cdot t^n \tag{1}$$

3 Meta-typological fitting

To get an impression how these assumptions fare empirically, I will present a small meta-typological experiment (for an introduction to meta-typology, cf. Maslova [12]. As described earlier, such an experiment is no argument for a particular distribution; it will only show that it is possible to model empirical frequencies by using a negative exponential distribution. Whether this is indeed the right distribution can never be proved by a well-fitted curve. The theoretical derivation of the distribution has to be convincing, not the empirical adequacy.

For the meta-typological experiment, I randomly selected one cross-linguistic type from each chapter of WALS [5]. A histogram of the number of languages per type is shown in Figure 1. Based on a similar distribution, Maslova [12] proposed that the size of cross-linguistic types follows a pareto distribution. However, as described in the previous section, it seems to make more sense to consider this an exponential distribution as defined in (1).

To fit the empirical data to the proposed distribution in (1) I divided the types from WALS into bins of size 10, i.e. all types with 1 to 10 languages were combined into one group, likewise all types with 11-20 languages, etc. For each of these bins, I counted the number of types in it. For example, there are 18 types that have between 1 to 10 languages, which is 12.9% of all 140 types. So, the probability to have a"queue" with a length between 1 and 10 languages is 0.129. Fitting these empirical probabilities for type-sizes to the proposed distribution in (1) results in a traffic rate $t = .85 \pm .01$.[2]

It is important to note that I took the bare number of languages per type as documented in each chapter in WALS. This decision has some complications, because the set of languages considered (i.e. the "sample") is rather different between the different chapters in WALS. The number of languages in a particular type will normally differ when the researcher considered 100 languages or 500 languages as a sample. Still, I decided against normalizing the samples,

[2] I used the function *nls* ("non-linear least squares") from the statistics environment R [17] to estimate the traffic rate from the data, given the predicted distribution in (1). Also note that these fitted values represent a random sample of types from WALS, and the results will thus differ slightly depending on the choice of types.

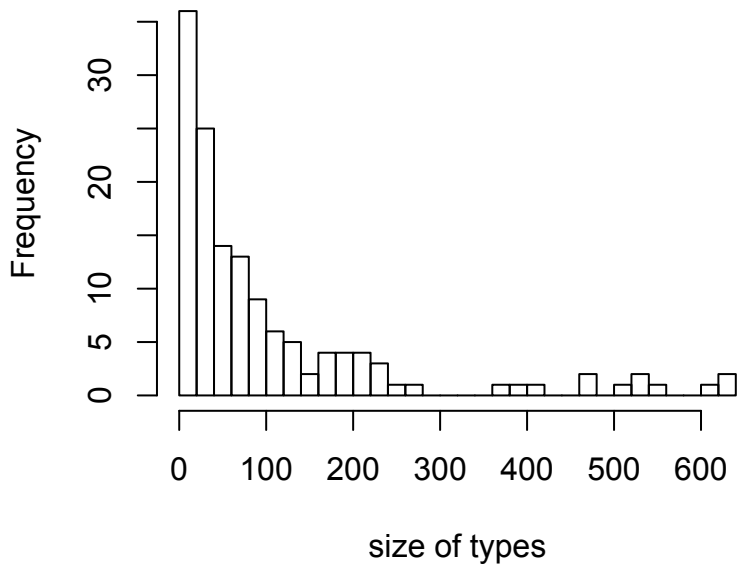

Fig. 1. Histogram of type sizes from WALS

because that would introduce an artificial upper boundary. This decision implies, however, that the distribution of type-size in the current selection of data from WALS is also influenced by yet another random variable, namely the size of the sample from each chapter. From the perspective of typology, this is a rather strange approach, because a typological samples attempt to sample to current world's languages. For the current purpose, however, the population to be sampled is not the current world's languages, but all languages that were ever spoken, or will ever be spoken through time and space. From that perspective, any restriction on sample size will only restrict the number of languages, but it will not influence the traffic rate, nor the type-distribution.

The relation between the empirically observed probabilities and the fitted probabilities is shown in Figure 2. It is thus easily possible to nicely fit the empirical data to the proposed distribution in (1). However, as argued in the previous section, this is not a proof of the proposal, but only an illustration. The nature of a probability distribution can never be empirically proven, but only made more plausible by a solid analysis of the underlying processes.

4 Outlook

The model presented in this paper is restricted to a very simplistic birth-death model of typological change. More complex models will have to be considered to also cover the more interesting cases like the size of phoneme inventories.

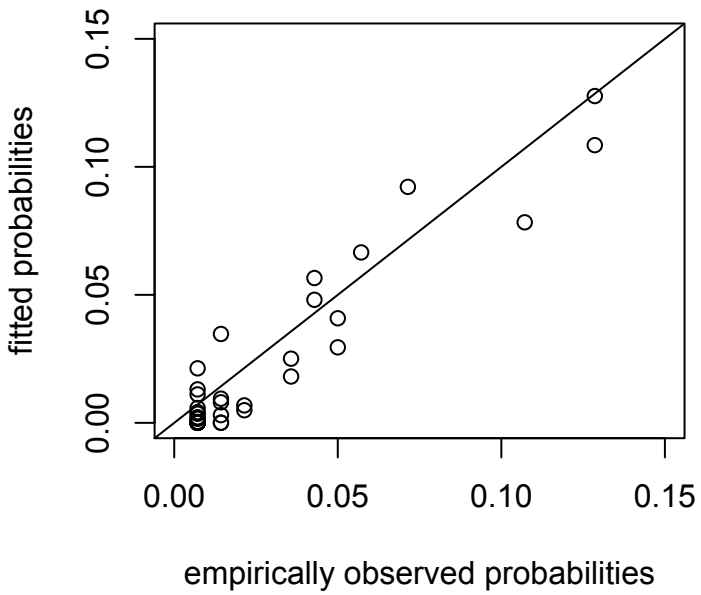

Fig. 2. Fit of empirical distribution to predicted distribution

Basically, to extend the current approach, Markov models involving multiple states and specified transition probabilities between these states are needed. For example, phoneme inventories can be considered a linearly ordered set of states, and the process of adding or losing one phoneme can also be considered a poisson process. At this point, I do not know what the resulting probability distribution would be in such a model, but it would not surprise me if Lehfeldt's [9] proposal of a gamma distribution would turn out to be in the right direction after all.

References

1. Altmann, G. (1985). Die entstehung diatopischer varianten. *Zeitschrift für Sprachwissenschaft*, 4(2):139–155.
2. Comrie, B. (1989). *Language Universals and Linguistic Typology.* Blackwell, Oxford.
3. Cysouw, M. (2003). Against implicational universals. *Linguistic Typology*, 7(1):89–101.
4. Dryer, M. S. (2005). Order of object and verb. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *World Atlas of Language Structures*, pages 338–341. Oxford University Press, Oxford.
5. Haspelmath, M., Dryer, M. S., Comrie, B., and Gil, D., editors (2005). *The World Atlas of Language Structures.* Oxford University Press, Oxford.
6. Janssen, D. P., Bickel, B., and Zúñiga, F. (2006). Randomization tests in language typology. *Linguistic Typology*, 10(3):419–440.

7. Justeson, J. S. and Stephens, L. D. (1984). On the relationship between the numbers of vowels and consonants in phonological systems. *Linguistics*, 22:531–545.
8. Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354.
9. Lehfeldt, W. (1975). Die verteilung der phonemanzahl in den natürlichen sprachen. *Phonetica*, 31:274–287.
10. Maddieson, I. (2005). Consonant inventories. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *World Atlas of Language Structures*, pages 10–13. Oxford University Press, Oxford.
11. Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology*, 4(3):307–333.
12. Maslova, E. (2008). Meta-typological distributions. *Sprachtypologie und Universalienforschung*, 61(3):199–207.
13. Mitzenmacher, M. and Upfal, E. (2005). *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge Uniersity Press, Cambridge.
14. Nichols, J. (1992). *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago.
15. Nichols, J., Barnes, J., and Peterson, D. A. (2006). The robust bell curve of morphological complexity. *Linguistic Typology*, 10(1):96–106.
16. Plank, F. and Schellinger, W. (2000). Dual laws in (no) time. *Sprachtypologie und Universalienforschung*, 53(1):46–52.
17. R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.