ELENA MASLOVA (Stanford University & University of Bielefeld)

Meta-typological distributions

1. Introduction

The goal of this paper is to demonstrate that the typological database of the *World Atlas of Language Structures* (WALS, HASPELMATH *et al*. 2005) can be used to address some questions which lie at the very core of the methodological foundations of linguistic typology. These questions are essentially various modifications of the following one:

> What is the probability $p(n;N)$ of a language type having exactly n representatives in a set of N languages?

In other words, this paper is an attempt to analyze the quantitative representation of language types as a random variable and examine the distribution of this variable. The WALS gives us, for the first time ever, an empirical foundation to pursue this line of research, which can be referred to as *meta-typology*.

The reason why such meta-typological questions are of interest to typologists is, roughly speaking, that the quantitative analysis of cross-linguistic data would strongly benefit from a clear picture of what is expected *a priori* and what is not. This is needed to assess whether the typologist's findings indicate something unusual and possibly linguistically interesting with respect to the specific parameter under investigation. Indeed, the contrast between the expected and the observed plays a major role in typological argumentation, yet the commonly received assumptions about what is to be expected remain unsubstantiated (Section 2). A study focused on a single typological parameter tends to view its cross-linguistic distribution as a unique property of this specific parameter. In contrast to this, a meta-typological study can help us discern some shared properties of such distributions. If such shared properties exist, they are quite likely to have a common explanation.

The reason why meta-typological issues can be approached by examining WALS is that this database accumulates statistical information on a wide range of different language types. Given the broad international participation in the project, WALS comes rather close to a collection of all language types currently of interest to the typological community. Obviously, neither the database as a whole, nor its parameter-specific components has been designed with a meta-typological usage in mind, which means that it will not give conclusive answers to this class of questions. Rather, any meta-typological results should be interpreted as more or less useful heuristics (Section 3).

My own belief is that the approach suggested here opens a useful line of inquiry. However, this belief is largely based on promising preliminary results. Accordingly, a description of these results, however tentative, constitutes an integral part of this presentation (Section 4). My greatest hope in undertaking the minor study

reported here was to discover some distribution patterns which would, on the one hand, make sense on independent grounds and, on the other, provide some new insights into the general properties of typological distributions in the language population. It seems to me that such patterns can indeed be discovered. Most significantly, it appears that the quantitative representation of language types conforms to the so-called Pareto (or power-law) distribution, a distribution observed in a wide variety of real-world situations and better known to linguists under the name of Zipf's law. In this class of distributions, the probability of a type having n representatives decreases with n as a negative power of n (the shape of this distribution is illustrated by the solid line in Figure 1).

2. Some common assumptions in typology

A prototypical typological study investigates a small set of parameters of cross-linguistic variation, and attempts to interpret the distribution of linguistic properties as attested for these particular parameters. It will be convenient to represent the quantitative result of a typological study as a vector of k numbers, $<n_1,\ldots, n_k>$, where k is the number of values of the parameter and n_i is the number of representatives of value i. Any linguistic interpretation of such results is based on how the findings differ from what we expect to find if a typological distribution is determined solely by non-linguistic random factors. The common typological wisdom is to assume that all n_i are expected to be roughly equal:

> 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. (Comrie 1989: 20).

The concept of typological distribution, as invoked by Comrie in this quotation, falls under the general probability-theoretic concept of probability distribution if each linguistic parameter is viewed as a discrete random variable, which can have one of k values with certain probabilities $<p_1,\ldots,p_k>$ ($\Sigma p_i=1$). Comrie's general principle implies that if no universal is involved, i.e. if there are no parameter-specific linguistic pressures, then this distribution must be uniform ($p_i = 1/k$). The hypothesis of uniform distribution serves, in effect, as the basic 'null hypothesis' of linguistic typology that can, or cannot, be rejected on the basis of statistical data.

The use of statistical criteria for this purpose necessarily involves additional assumptions about the potential effects of non-linguistic random processes on the actual numbers of languages of each type, $<n_1,\ldots,n_k>$ as observed in a representative sample of languages (cf. Maslova 2000). At this level of analysis, another probability distribution, $p(n_1,\ldots, n_k; N)$, must be assigned to the set of all possible vectors $<n_1,\ldots,n_k>$, since we have to decide which deviations from uniformity are unlikely enough to reject the null hypothesis. In actual typological practice, $p(n_1,\ldots, n_k; N)$ is assumed to be the multinomial distribution, that is, the implicit assumption is that the non-linguistic random effects on typological distributions can be equated with those involved in random drawing of N languages from an in-

finite pool of languages. This entails the formula as shown in (1) for the probability $p(n;N)$ of a language type having exactly n representatives in a set of N languages.

$$(1) \qquad p(n;N) = \frac{N!}{n!(N-n)!}(\alpha)^n(1-\alpha)^{N-n}$$

In this formula, α is the linguistically determined likelihood of the language type, which is expected to be equal to $1/k$ for all types of a k-ary classifications that is not affected by language universals. For large N and reasonably small fixed k, this distribution can be approximated by the normal (bell-curve) distribution with $\mu = N/k, \sigma = \sqrt{N(k-1)/k}$. The dash-dot line in Figure 1 visually represents this distribution for $k = 5$ and $N = 100$. If an observed value differs from μ by, say, more than 3σ, this should be taken, according to Comrie's suggestion above, as an indication of a language universal to be stated and accounted for.[1]

There are several problems with this approach. For example, we tend to take the very existence of typological distributions that significantly depart from uniformity as the core piece of evidence for a significant role of universal linguistic factors. However, this is not necessarily the case. Let us assume, for the sake of argument, that language universals play no role in shaping typological distributions, i.e. these distributions arise solely by virtue of non-linguistic random processes in the language population. Obviously, this does not mean that strong deviations from uniformity are impossible; it just means that they are rare. Moreover, these deviations are not just accidental properties of a particular sample. Strong deviations are present in the language population from which the sample is drawn. In other words, these deviations are not just due to sampling errors, but also to non-linguistic random processes in the language population. Therefore, strong deviations will be observed (oversimplifying the matter to some extent) in any representative sample. This entails that some strongly non-uniform distributions are sure to be found as more and more parameters are investigated. To put it the other way round, given that the typological community has already studied quite a number of typological parameters, it would have indeed been a miracle if all of them had happened to conform to the hypothesis of uniformity, even if language universals would have no major impact. These considerations suggest that the very existence of some strongly skewed typological distributions need not indicate anything linguistically significant at all. In order to verify the hypothesis that language universals can manifest themselves statistically in cross-linguistic distributions, we need to investigate the quantitative representation of multiple language types. If we happen to detect something similar to the expected bell-curve distribution, then both cross-linguistically rare and cross-linguistically widespread types will find their place at the margins of this distribution.

---

[1] The same basic assumptions are implicitly at work not only when statistical methods are used to infer universal linguistic preferences, but also whenever statistically established dependencies between different linguistic variables are interpreted in terms of language universals.

Another assumption of the received approach is, informally, that the non-linguistic random processes in the language population work as sampling of sorts, in that they randomly push the frequency of a type in any direction, and are therefore unlikely to bring about strongly skewed typological distributions (Bell 1978: 171). To be more accurate, it is assumed that those processes that possibly do not work like this, e.g. those that determine the rise and fall of language families, can be compensated for by appropriate sampling procedures (Bell 1978, Perkins 1989).

These assumptions, however intuitively plausible they may seem, are far from being self-evident. Consider, for example, the size of language families. It is often assumed in the literature, implicitly or explicitly, that if the relevant events (i.e. splits and shifts of language communities) had been statistically independent of historical and geographical circumstances, then we would expect all genetic groupings to have roughly an equal number of members, which is obviously not what is observed. However, this assumption is just wrong: a simple model of the language population with a single constant probability of split and a single constant probability of shift for each language predicts something very similar to the observed distribution of family sizes (Maslova 2000). This example demonstrates that our ideas of what should be expected of a random typological distribution strongly depend on the model of the underlying random processes. Most importantly, 'random' need not mean 'roughly equal'. For example, both distributions shown in Figure 1 are random. The difference between them demonstrates just how drastically our expectations can change if another model of non-linguistic random processes is found more plausible.

While our knowledge of the random processes at work in the language population is sufficient to question the empirical validity of the multinomial model of random drawing from an infinite pool of languages (e.g. Dryer 1989), it does not seem enough to choose another one without additional empirical data. Ideally, the best empirical foundation would be provided by data on multiple linguistic classifications *a priori* known not to involve any language universals (which might have skewed the corresponding cross-linguistic distributions). Obviously, this possibility is out of the question. However, the statistical data from multiple linguistic parameters as available in WALS gives us an opportunity to analyze the frequency of language types as a random variable and determine whether this variable conforms to any known distributional pattern. This meta-typological analysis might provide important insights for developing an appropriate model of randomness in the language population and the implicit effects on typological distributions.

3. Methodological issues

The general goal of the approach suggested here is to find out whether it is possible to detect some known distribution functions in the quantitative representation of linguistic features, so that the value of $p(n;N)$ could be expressed as a function of n and N, $p(n;N) = f(n,N)$. For example, the formula in (1) above is such a function. The method of approaching the search for such a formula empirically is fairly straightforward: one has to count the number of language types represented by exactly n languages (for $n = 1,\ldots,N$) and compare the results with those expected un-

der the assumption that $p(n;N) = f(n,N)$ for different distribution functions that might seem to provide a plausible model (see Woods *et al.* 1986: 132-150; Bain and Engelhardt 2000: 442-462 for details).[2] Obviously, this procedure requires two (types of) samples: a representative sample of language types, and one or more representative samples of languages.

The set of language types represented in WALS is obviously not a random sample drawn from a hypothetical general population of objectively existing linguistic features. Rather, it can be thought of as a representative sample of types that are currently of interest in the typological and, possibly, broader linguistic community. This, in itself, does not undermine the potential relevance of the results for future typological studies (after all, any such study, by definition, would focus on other linguistic features from the same 'interesting' set). Yet, it has to be taken into account in any interpretation of these results.[3] Moreover, we cannot assume that universal linguistic pressures (if any) have had no effect on the frequencies of these types, which would be ideal for estimating the effects of non-linguistic random processes. However, if language universals are involved, then they 'push' the frequencies of language types represented in WALS in both directions. In this sense, the sample will not be biased in favor of 'preferred' or 'dispreferred' types because of the workings of language universals. This means that a meta-typological study based on WALS can be thought of as a reversal of the usual typological approach. Normally, typology focuses on how a specific typological distribution is affected by linguistic pressures and construes the non-linguistic random processes as a source of random errors. What I suggest is to look for the effects of these random processes (which, by their very nature, affect all cross-linguistic distributions in the same way), whereas the potential effects of parameter-specific language universals would be taken into account as a potential source of deviations from distributions determined by non-linguistic processes.

As far as the samples of languages are concerned, our data is obviously limited to the samples used in WALS, which are, with some exceptions, different for different parameters. To begin with the most evident problem, these samples differ considerably in size (N). It does not seem to be wise to throw away the additional information provided by larger samples in order to keep N constant for all language types. For this reason, I decided to work with proportions ($s = n/N$) of language types rather than with absolute frequencies.[4] The switch from a discrete random variable to a continuous one means that the basic concept of probability is not di-

---

[2] Here and below I give references to these two introductory descriptions of the relevant statistical methods, one written specifically for linguists, the other somewhat more mathematically sophisticated. However, the methods are general enough to be described in practically any other textbook as well.

[3] On the other hand, some parameters present in the database should obviously be excluded from a study like this. This concerns, primarily, those parameters whose values are deliberately defined in accordance with their quantitative representation, as is the case, for example, for the classification of consonant inventories into 'small', 'average', and 'large', where the values are defined in such a way as to achieve a relatively even distribution (Maddieson 2005: 10).

[4] This solution involves some methodological problems of its own, yet they do not seem particularly relevant in the present context and will not be discussed in this paper.

rectly applicable to specific values of the variable. Instead, we have to deal with the *cumulative* distribution function $F(s)$, i.e. the probability that the frequency of a language type does not exceed s, and its derivative, the *probability density* function $f(s) = F'(s)$, which serves as the continuous counterpart to the probability distribution function for discrete variables. Accordingly, instead of counting the number of types represented by exactly n languages, we have to count the number of types whose frequencies lie within a certain interval ($s_1 \leq s < s_2$) and compare the results with the corresponding numbers predicted by $F(s)$, that is, with $(F(s_2)\text{-}F(s_1)) \cdot M$, where M is the total number of types in the set of language types under consideration (Woods *et al.* 1986: 132-150; Bain and Engelhardt 2000: 442-462).

The shift from absolute frequencies to proportions does not, of course, resolve all the problems associated with differences between parameter-specific samples, and, more importantly, between sampling procedures. The greatest hazard for reliability of the results lies in the fact that some samples might have been chosen with the specific parameter and its cross-linguistic distribution in mind, so that the presence of a language in the sample is not independent of the type it represents.[5] Unfortunately, at the time of this study I had no access to descriptions of sampling procedures for all studies represented in the atlas, which might have allowed for a more accurate approach to this issue. For this paper, I used the following procedures intended to compensate for possible overrepresentation of some genetic groupings in some samples. I used the two levels of genetic classification provided within the database itself: stock and genus. If we denote the total WALS sample for a given parameter **T** as $\mathbf{L}_W(\mathbf{T})$, then a *stock-level randomized sample*, $\mathbf{L}_S(\mathbf{T})$, contains a single language from each genetic stock represented in $\mathbf{L}_W(\mathbf{T})$, which has been randomly drawn from $\mathbf{L}_W(\mathbf{T})$. Similarly, a *genus-level randomized sample*, $\mathbf{L}_G(\mathbf{T})$, contains a single language from each genus represented in $\mathbf{L}_W(\mathbf{T})$, which has been randomly drawn from $\mathbf{L}_W(\mathbf{T})$. For each parameter **T**, I compared the distributions observed in the samples thus obtained.

The differences between stock-level and genus-level samples are insignificant for all parameters in WALS, i.e. these samples can be taken to represent the same underlying distribution in all cases. Note that this point can have methodological consequences beyond the scope of this paper, since stock-level samples are commonly considered superior to genus-level samples in statistically oriented typological studies because they increase the genetic distance between languages in the sample and thus enlarge the presumed mutual independence of their linguistic properties. In the case of WALS, this choice would not significantly affect the result. Yet, genus-based samples have the obvious advantage of being larger and therefore more informative.[6]

---

[5] It goes without saying that such sampling procedures can be absolutely justified in the context of WALS itself. The problems can arise only when the resulting samples are used for other purposes.

[6] Based on the estimates presented in Maslova (2000), I believe that such samples can be taken as representative of the language population as a whole. The same estimates suggest that the lack of significant differences between genus-level and stock-level samples is not an accidental property of WALS, but rather a general property of the typological distributions in the language population.

In contrast to this, the distributions exhibited by the total WALS samples proved to be significantly different from those exhibited by randomized samples for many parameters. Since such differences are most likely determined by overrepresentation of some genera in some WALS samples for parameter-specific reasons, the genus-level randomized samples emerge as the most appropriate choice for the purposes of a meta-typological study.

4. Some preliminary results

Finally, I would like to present some preliminary results, which, in my view, demonstrate the heuristic potential of the meta-typological use of WALS. The first result is based on samples containing a single randomly chosen language type for each parameter represented in WALS. The considerations behind this sampling procedure are two-fold. First, it keeps the measurements mutually independent (which cannot be assumed when two values from the same parameter are included). Second, it gives equal representation to parameters with different number of possible values. The general idea was to discover a distribution function which would adequately describe the frequency of a language type viewed as a random variable. Put differently, if we study a language type, the question to be answered is what is the probability that its proportion s will be less than x ($0 < x \leq 1$), assuming that we don't know the number of possible typological alternatives. It turns out that the observed distribution can be approximated by Pareto distribution. The corresponding probability density function is shown in Figure 2, along with a column chart of the observed distribution (each column represents the number of parameter values whose frequencies lie within the corresponding interval, normalized for visualization purposes). If we analyze parameters with the different number of values separately, it turns out that for binary parameters the observed distribution of frequencies is most closely approximated by the uniform distribution, i.e. all frequencies are equally probable. As the arity of the parameter increases, the distribution shifts towards a Pareto-like distribution, so that some variant of a Pareto distribution can be said to approximate the frequency distribution for parameters with four or more values.[7]

---

[7] Since we are interested only in whether the general model has an appropriate form, the values of parameters were estimated for each randomized sample by means of maximum likelihood estimation, and the chi-squared goodness-of-fit test was used to test the hypothesis of Pareto distribution with unspecified parameters. The intervals were chosen in such a way to maximize the number of degrees of freedom, yet to keep the expected values around 5 or so to ensure that the chi-squared approximation is accurate. Depending on the sample, the resulting chi-squre values correspond to significance levels ranging from 0.10 to 0.05. Arguably, this is not a very good fit. However, since the shape of the observed distribution becomes closer and closer to Pareto as the arity of the typological parameters increases, it seems reasonable to hypothesize that the fit would have been much better if the maximum arity had not been limited by the obvious cartographic considerations, and the typological parameters with actual values grouped in such a way as to equalize the quantitative representation of types (see Section 3) had been consistently excluded from the data.

These results are suggestive primarily because the same distribution pattern is known to occur in a wide range of real-world situations that are in important respects quite similar to the language population (e.g. it also describes the distribution of wealth in a community, settlement sizes, and the sizes of language families). For this reason, it seems highly likely that the non-linguistic random processes in the language population work in such a way as to bring about a Pareto-like distribution. In other words, the assumption that random processes in the language population work in such a way as to bring about some variant of Pareto distribution would have been *a priori* more plausible than the assumption of a normal-like distribution (as commonly assumed in typology); for example, it would be expected if language contacts play the major role in language change. As far as I know, however, this hypothesis has never even been entertained by typologists. In this sense, even the current preliminary results of a meta-typological analysis of WALS appear to provide useful heuristics. They can by no means be taken as proof, yet seem to point in a promising direction as far as development of a feasible model of the random processes in the language population is concerned.

References

BAIN, LEE J. & MAX ENGELHARDT (2000) *Introduction to probability and mathematical statistics*, 2nd edition. Duxbury.

BELL, ALAN (1978). Language sampling. In: Joseph H. Greenberg et al. (eds.) *Universals of Human Languages*. *Vol. 1. Method & Theory*. 125-156. Stanford: Stanford University Press.

COMRIE, BERNARD (1989). *Language Universals and Linguistic Typology*. (2 ed.) Oxford: Blackwell Publishers.

DRYER, MATTHEW S. (1989). Large linguistic areas and language sampling. *Studies in Language* 13: 257-292.

HASPELMATH, MARTIN, MATTHEW S. DRYER, DAVID GIL & BERNARD COMRIE (eds.) (2005) *The World Atlas of Language Structures*. Oxford: Oxford University Press.

MADDIESON, IAN. (2005). Consonant Inventories. In (Haspelmath et al. 2005), p.10-13.

MASLOVA, ELENA (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4: 307–333.

PERKINS, REVERE D. (1989). Statistical techniques for determining language sample size. *Studies in Language* 13: 293-315.

WOODS, ANTHONY, PAUL FLETCHER & ARTHUR HUGHES (1986). *Statistics in language studies.* Cambridge: Cambridge University Press.

Correspondence address

Elena Maslova
2000 Walnut Ave, #J307
Fremont, CA 94538, U.S..A.
maslova@jps.net

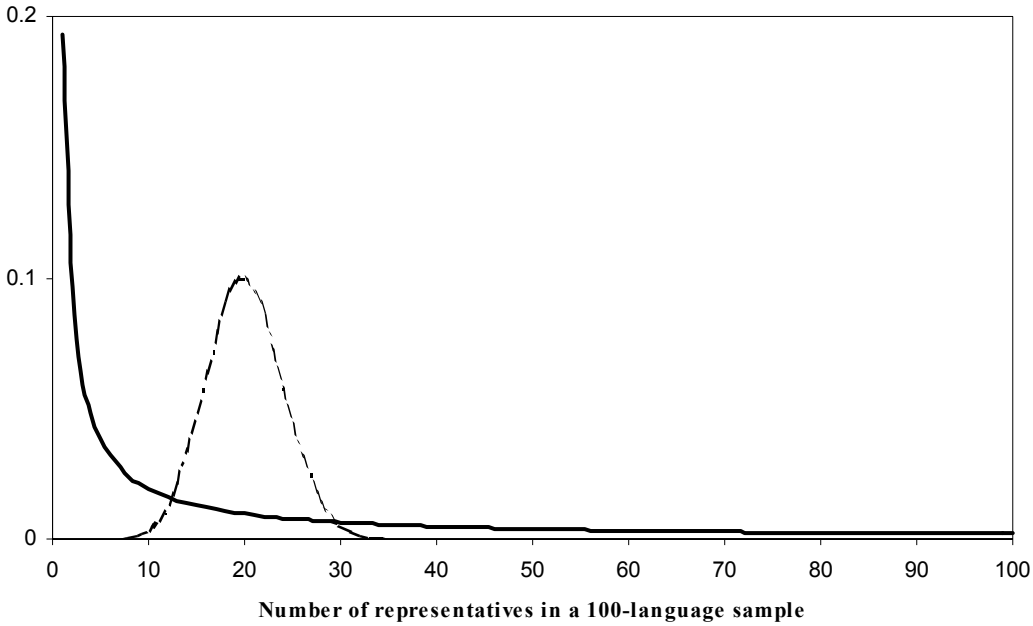

Figure 1. Comparison of two possible distributions for quantitative representation of a language type: a Pareto distribution (solid line) and a binomial distribution (dash-dot line).

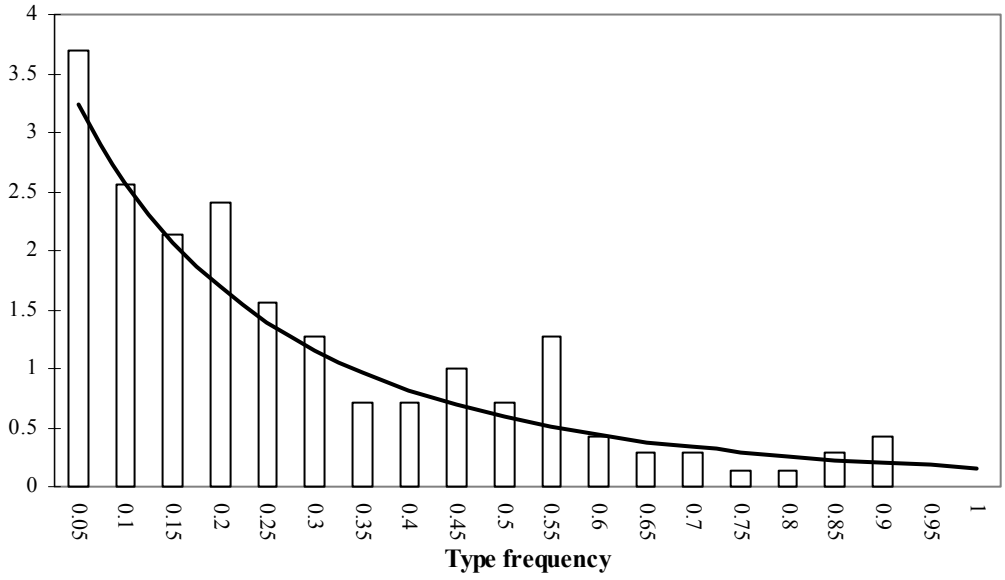

Figure 2. Comparison of the empirical distribution of type frequencies with a Pareto probability density function.