Commentary

Counting genera vs. counting languages

by *Matthew S. Dryer*

1. Introduction

This paper is a response to the preceding paper in this journal, by Elena Maslova, and assumes familiarity with that paper. I will focus on a number of issues that are associated with the relative value of counting genetic groups of a time depth of 3,500 to 4,000 years (henceforth *genera*) and of counting actual numbers of languages. In many respects, Maslova's paper considerably raises the level of discussion on the issue of testing typological generalizations, and is suggestive of more sophisticated approaches to such problems. However, I will argue that (i) Maslova makes rather unlikely assumptions about the number of languages 3,700 years ago; (ii) her model is seriously inaccurate in failing to capture the high frequency in the actual world both of genera containing a single language, or less than five languages, and of genera containing more than 100 languages; (iii) her model fails to capture the effect that such large genera can have on altering the frequency of a linguistic type after 3,700 years; (iv) her discussion confuses frequency of a type among genera with frequency of a type 3,700 years ago; (v) counting genera provides a better basis for testing typological generalizations than counting actual numbers of languages; and (vi) counting genera is not enough.

2. How many languages were there 3,700 years ago?

While the issue is ultimately not essential to Maslova's arguments, her discussion does at times assume that the number of languages spoken 3,700 years ago was probably less than 1,000. Her apparent view is that over much of the past 3,700 years the number of languages has been increasing rapidly, along with overall population increases, and that only in recent centuries has this increase stopped (and reversed). While I can only discuss the issue briefly here, I think

Linguistic Typology 4 (2000), 334–356

1430–0532/2000/004-0334

such a view is seriously mistaken, and that there is no reason to believe that the number of languages spoken 3,700 years ago was not as high as it is now, and may well have been higher. When we examine languages spoken today and in recent centuries, we find a clear correlation between political structure and numbers of speakers. Languages spoken by hunter-gatherer societies typically have very small numbers of speakers, fewer than 10,000 and often considerably less. Conversely, where there are languages spoken by larger numbers of speakers, especially over half a million, there are typically political units containing large numbers of people. Over the past 3,700 years, there has been a huge increase in the number of people living in such large political units, and a decrease in the number of hunter-gatherer societies. This leads to the conclusion that while the population of the world was only a small fraction of what it is today, the number of languages could easily have been as many as there are today.

Maslova argues, based on the figures in her Table 3, that if there were as many as 5,500 languages 3,700 years ago, then, assuming the groups in *Ethnologue*, the time depth of major groups would be over one million years, which is clearly implausible. What she is really saying is that her model predicts that it would take this long for major genetic groups to become as large as they are. But an alternative inference to make from this is that there is a problem with her model. I will argue below that her model independently suffers from failing to account for the number of genetic groups with over 100 languages after only 3,700 years.

3. A summary of Maslova's position

In Dryer (1989), I explained and defended an approach to testing typological generalizations that involves (in part – see Section 9 below) counting what I call genera, genetic groups of a time depth of 3,500 to 4,000 years, rather than counting actual numbers of languages in the world today. The primary rationale behind this was that for many typological parameters, languages within a genus are typically the same, and numbers of languages are distorted by large genera. Much of Maslova's paper can be construed as directly challenging this methodological claim. Her argument can be summarized as follows, though what I say here probably oversimplifies her position in some respects. First, counting genera is similar to counting frequencies of languages 3,700 years ago. Hence the difference between counting genera and counting languages is, she claims, roughly the difference between counting languages 3,700 years ago and counting languages today. Given this view, she raises the legitimate question: why should counting languages spoken 3,700 years ago provide a better basis for testing typological generalizations than counting languages spoken today? She argues in addition that her model shows that it is unlikely that

the frequency of language types will change significantly during this period of time. If they are not significantly different, then what reason is there to count genera rather than to count languages? She further argues that in some respects, counting languages is BETTER than counting genera. If, as she assumes, the numbers of languages 3,700 years ago was considerably less than it is today, then the frequency is likely to be LESS representative of the actual probability of types, since elementary principles of statistics tell us that smaller populations are more likely to deviate from the norm than larger populations. However, this last argument assumes that the number of languages 3,700 years ago was less than it is today, an assumption I challenged in the preceding section.

4. A computer simulation of Maslova's model

In order to evaluate Maslova's model, I have written a computer program that simulates it. By running this simulation a number of times, we can determine a probability distribution corresponding to the model. The data in Figure 1 gives the distribution found over 1,000 trials of the frequency of a type that occurs with an initial frequency of 50 %. Since what we are interested in is the likelihood of a particular type changing in frequency solely due to the effects of "births" and "deaths" of languages (rather than to type shifts), the simulation assumes no type shifts. These trials assume that the initial number of languages is 600 (one of the possibilities Maslova considers most likely, but contrary to what I argued in Section 2 above), and assume the probabilities of birth and death that Maslova assumes when the initial number of languages is 600 (namely 0.097 and 0.035 respectively for each 100-year period). It is worth mentioning that the average number of languages after 3,700 years and the average number of surviving genera (or languages from the initial set of 600 with surviving descendants) over these trials are roughly what Maslova claims: the average number of languages in the present over these 1,000 trials was 5,545 (just a bit less than the 6,000 Maslova claims) and the average number of genera was 395, very close to the 400 assumed by Maslova.

Since a set of trials this large should give us a very good approximation of the probability, we can use it to provide a good estimation of the frequency of a type changing. Table 1 summarizes Figure 1 by giving a number of frequency intervals and the likelihood of a type that starts with 50 % frequency ending up with a frequency within that frequency interval. Table 1 shows that the probability of a type changing from 50 % to something within the frequency interval of 45 % to 55 % inclusive is 0.96; in other words, there is only a 0.04 chance of the type changing in frequency so that it was less than 45 % or more than 55 %. This accords well with the spirit of Maslova's claim: that under the assumed initial number of languages and birth and death frequencies, the

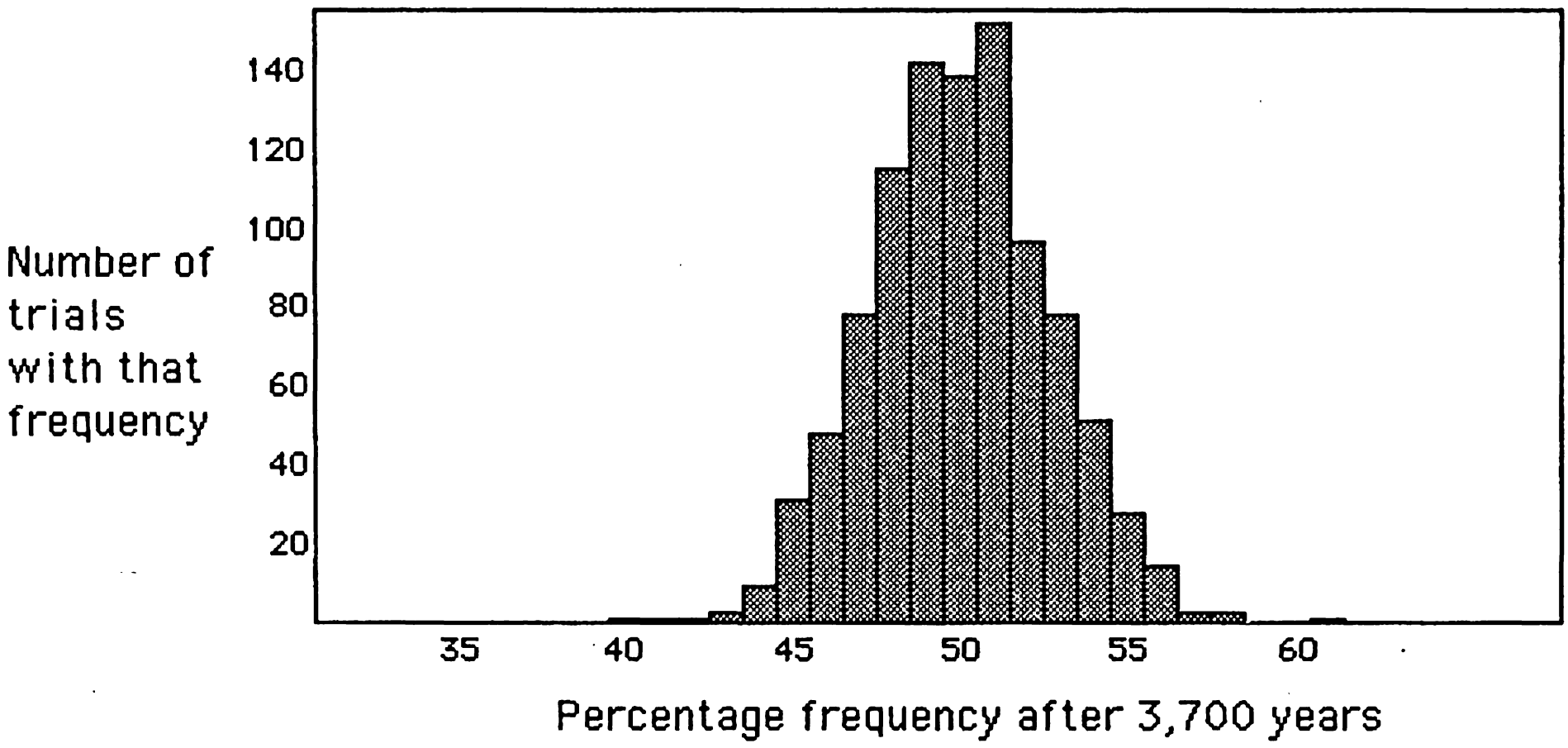

Figure 1. *Number of trials in which a type with initial frequency of 50 % occurred with the frequency given after 3,700 years, where initial number of languages is 600*

Table 1. *Likelihood of a type with initial frequency of 50 % having a frequency after 3,700 years within the frequency interval given, where initial number of languages is 600*

Frequency interval	Percentage of trials within interval in %	Percentage of trials outside interval in %
[43...57]	99	1
[44...56]	98	2
[45...55]	96	4
[46...54]	90	10
[47...53]	80	20

frequency of a type after 3,700 years (as a proportion of all languages) will not be significantly different from its initial frequency.

Figure 1 and Table 1 are based on the assumption that the number of languages spoken 3,700 years ago was 600. I ran a similar simulation for the alternative assumption that the number of languages was 6,000, approximately the same as in the present. The assumed birth and death rate were, following Maslova, both 0.38. The results, based again on 1,000 trials, are not significantly different from those shown in Figure 1 and Table 1. The average number of genera after 3,700 years was 368, a little less than the 400 claimed by Maslova. Table 2 and Figure 2 give data comparable to Table 1 and Figure 1.

Although similar to the results for an initial state of 600 languages, the probabilities shown in Table 2 show that with an initial state of 6,000 languages,

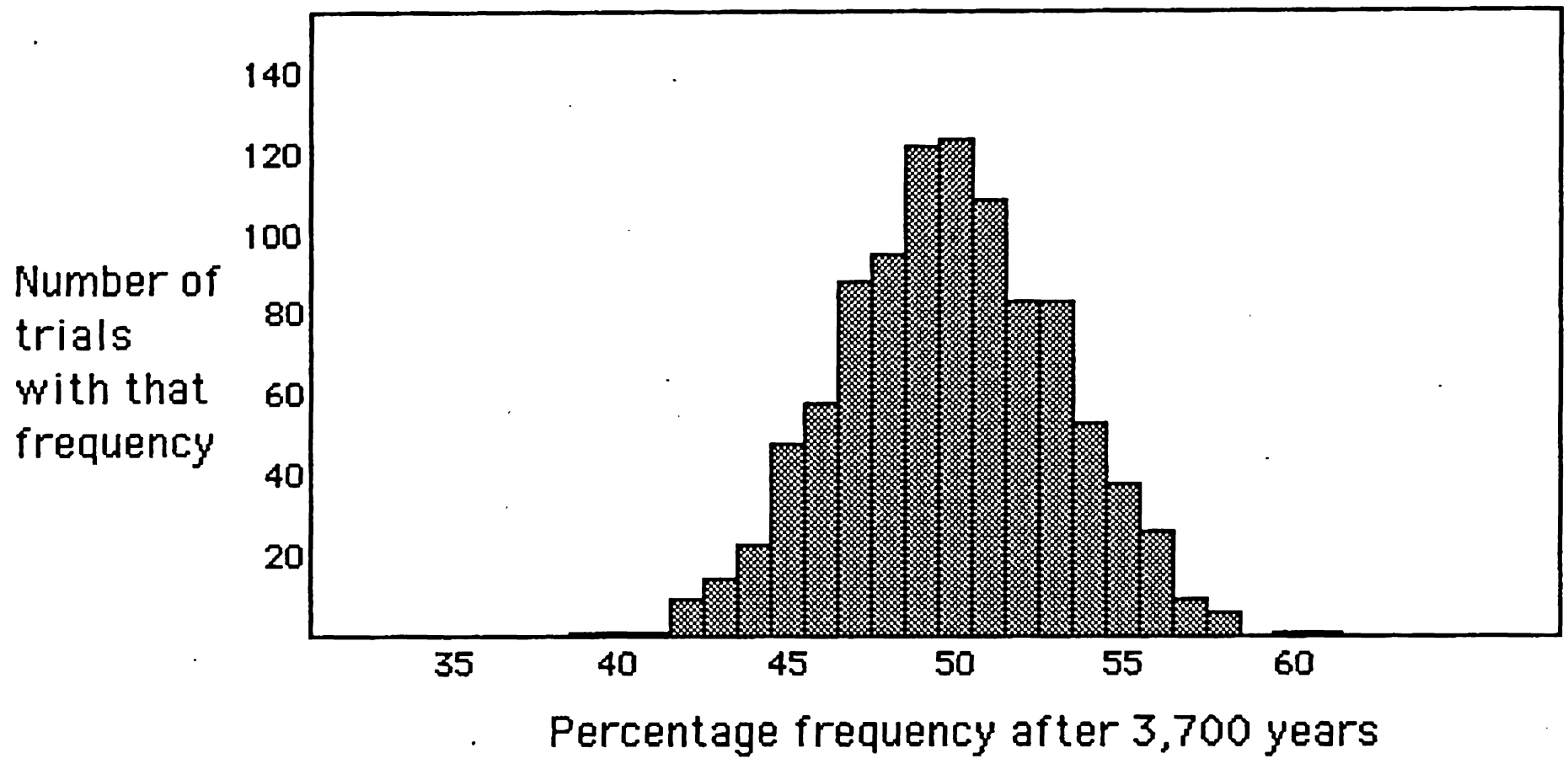

Figure 2. *Number of trials in which a type with initial frequency of 50 % occurred with the frequency given after 3,700 years, where initial number of languages is 6,000*

Table 2. *Likelihood of a type with initial frequency of 50 % having a frequency after 3,700 years within the frequency interval given, where initial number of languages is 6,000*

Frequency interval	Percentage of trials within interval in %	Percentage of trials outside interval in %
[42...58]	99	1
[43...57]	97	3
[44...56]	95	5
[45...55]	90	10
[46...54]	81	19
[47...53]	70	30

there is a slightly greater chance of the frequency changing. For example, while the probability of remaining within the frequency interval [45...55] is 0.96 for an initial state of 600 languages, Table 2 shows that the probability of remaining within this frequency interval is only 0.90 for an initial state of 6,000 languages.

5. The distribution of different sizes of genera

While the results shown above accord with the spirit of Maslova's claims, there are a number of problems both with her model and with her argumentation. The first problem is that her model seriously underestimates the degree of variation in sizes of genera after 3,700 years. During a period of 3,700 years, some

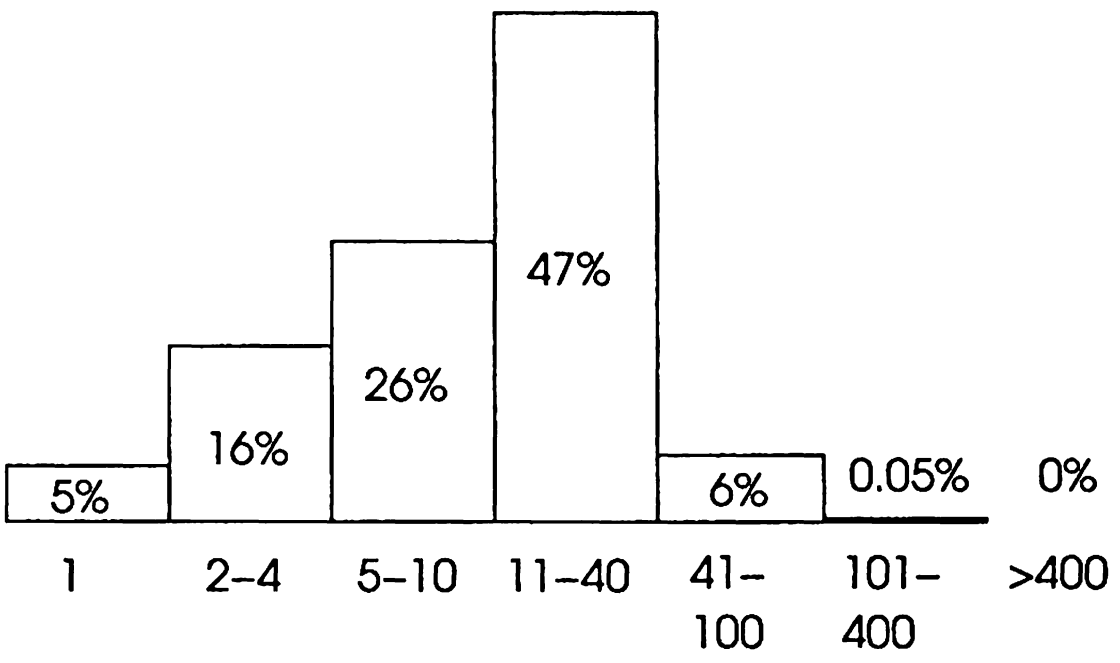

Figure 3. *Percentage of genera with this many languages according to Maslova's model*

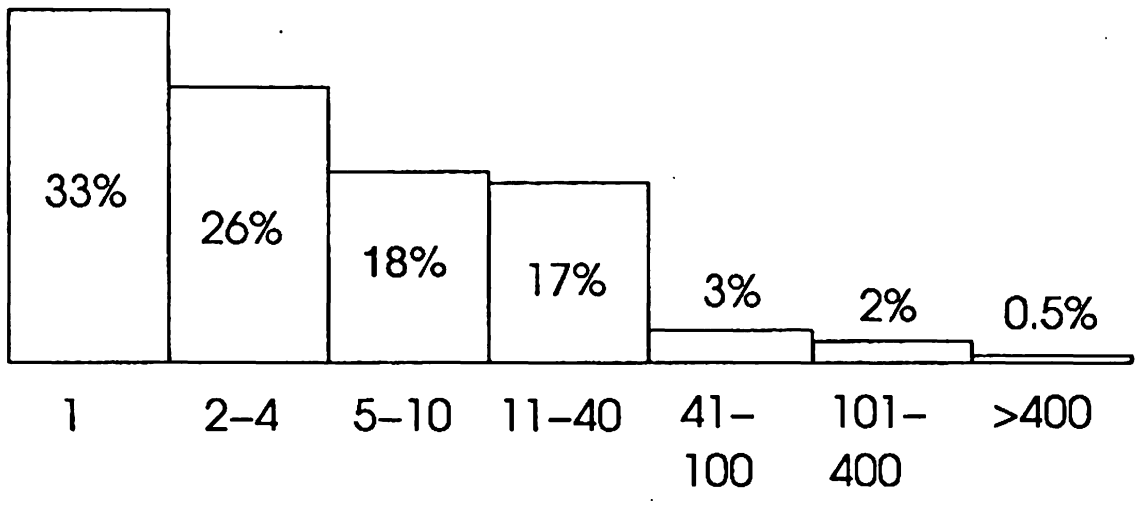

Figure 4. *Percentage of genera with this many languages in the real world*

genetic groups will die out, some will contain a single language, some will contain a small number of languages and some will contain many languages. It turns out that if we compare the distribution of numbers of genera of different sizes under her model, we find a distribution that is RADICALLY different from what we find in the real world. Figure 3 shows, as a bar graph, the average percentage, over 1,000 trials of the simulation of her model, of genera of different sizes.

Figure 3 shows that under Maslova's model, on average only 5 % of genera will contain exactly one language after 3,700 years, 16 % will contain two to four languages, and so on. At the upper end, we find that only 0.05 % (i.e., 5 in 10,000) of groups will have more than one hundred languages: in fact in only about one fourth of the trials did any group contain more than 100 languages after 3,700 years, and the largest genus over all 1,000 trials contained only 172 languages.

Figure 4 shows the comparable frequencies of different sizes of genera for the actual world, based on the numbers of languages for groups listed in *Ethnologue*. The decisions as to which groups constitute genera is based on my own educated guesses as to which levels are most likely to be of a time depth comparable to the subfamilies of Indo-European. For most genera, this is based

Table 3. *Genera containing 100 or more languages, with number of languages*

Bantoid	646	Bodic	134
Oceanic	493	Philippine Austronesian	130
Indic	219	Sulawesi Austronesian	112
Pama-Nyungan	176	Sundic	109
Adamawa-Ubangian	157	Baric	102
Central Malayo-Polynesian	149	Gur	100
Borneo Austronesian	137		

partly on examination of the languages in question and on discussions in the literature bearing on the question, as discussed in Dryer (1989).

The most obvious difference between Figures 3 and 4 is that there are far more small genera in the actual world, as shown in Figure 4, than is predicted under Maslova's model: while her model predicts that about 5 % of genera will contain one language, we find that about 33 % in the actual world do. While in her model only about 21 % (5 % + 16 %) have fewer than 5 languages, Figure 4 shows that in the actual world these constitute the majority, about 59 % (33 % + 26 %).

But the more significant difference between Maslova's model and the actual world is at the opposite end, and is less obvious just glancing at the two graphs: her model predicts that only 0.05 % of groups will contain over 100 languages, while in the actual world we find about 2.5 % (2 % + 0.5 %), or 50 times as many. Table 3 lists the genera with 100 or more languages, with the number of languages according to *Ethnologue*.

Not only are there these thirteen genera with 100 or more languages, but two of them contain many more than 400 languages. Bantoid is the largest, with 646 languages, while Oceanic contains 493 languages. This is in sharp contrast to the size of the largest genus in the 1,000 trials based on Maslova's model, which, as noted above, contained only 172 languages.

Two questions that arise are: why is this difference between Maslova's model and the actual world important and what is it about Maslova's model that leads to this difference? I will return to the first of these questions below, but a brief answer is that the primary reasons for counting genera rather than counting languages is that large genera can distort the number of languages of a particular type considerably from the probability of that type. But since on her model, genera are rarely larger than 100 languages, her model underestimates the extent to which large genera can have this effect.

The second question was why Maslova's model yielded such a different distribution of genus sizes from what we find in the real world. What is wrong about Maslova's assumptions? What different assumptions would we have to

make to change her model to get language groups as large as those found in the actual world? It turns out that the crucial assumption that Maslova makes which is the source of the problem is that the birth and death probabilities are constant. I claim that NO model based on constant birth and death probabilities can reflect the range of genus sizes that we find, while still accurately representing the number of languages and the number of genera. A more complex model would require CONDITIONAL PROBABILITIES. A simple example of such a model would be one in which the probability of a language splitting into two would not be a constant, but would be a function of the history of that language. For example, we could construct a model in which the probability of a language splitting into two languages would be higher if that language had resulted from a language split within the past 1,000 years. For example, we could revise the model so that the probability of splitting into two languages is 0.1 if the language has split into two languages within the past 1,000 years but only 0.07 if the language has not. When we change the model in this way, we find that this does increase somewhat the number of larger genera.

Using computer simulations of the sort described above, I have played with various models with conditional probabilities, but have thus far not been able to find a model which yields a range of genus sizes that approximates the distribution found in the real world. While models of the sort described in the preceding paragraph did increase the frequency of genera with over 100 languages, what I found was that they did so at the expense of decreasing the number of genera and of decreasing the number of genera containing a single language, but what we need is a model that yields at the same time a higher number of genera containing more than 100 languages and a higher number of genera containing a single language. And while I was able to construct models that yielded ranges of genus sizes somewhat more like what we find in the actual world, the models still fell short and even these models required a large number of ad hoc features that ultimately raised questions about the value of the enterprise. Ideally, the probability function ought to be motivated by features of the world that have yielded the distribution of genus sizes we find. Intuitively, if a language has gone for 2,000 years without dying or without splitting into two languages, its probability of continuing in this way is considerably higher than for a language which has recently split, especially if it has recently split into many languages. The range of genus sizes shown in Figure 4 presumably reflects the fact that particular areas of the world remain relatively stable and unchanged over long periods of time, while other areas undergo massive changes when a people move into the area, which will typically result in an increase in the deaths of languages already spoken in that area, and a greater than average increase in the rate of "births" in the group moving into that area. What is needed is a model that captures the nature of the historical situations associated with the huge increases in size

of the two largest genera in the world, Bantoid and Oceanic. The general moral is that mathematical models are of interest to the extent that they resemble the real world, and it seems likely that mathematical models will only achieve this if they are considerably more complex than the one proposed by Maslova.

6. An alternative approach

Rather than continue searching for a probability function which would yield a distribution of genus sizes similar to that of the actual world, I pursued the following alternative approach. I took a list of genera, with the number of languages in each genus according to *Ethnologue*, and wrote a computer program which randomly assigned one of two types to each genus, such that two types had equal likelihood, and again assuming that all languages within a genus are of the same type. For each assignment of types to genus, we can then compute the percentage of languages of each type.[1] By repeating this procedure many times, we can obtain a frequency distribution of the percentages of one of the two types over the set of trials. The data in Figure 5 gives the frequency distribution over 1,000 trials for one of the two types.

Even at a glance, Figure 5 looks very different from Figures 1 and 2 based on Maslova's model: the shape of the distribution is much flatter in Figure 5, reflecting the fact that the frequency distribution for a world with a greater number of large genera is much broader than under Maslova's assumptions. Table 4 summarizes the data from Figure 5 by showing the percentage of trials in which the percentage of one type fell within the frequency intervals indicated.

Table 4 shows a much broader frequency distribution from those given above in Tables 1 and 2 based on Maslova's model, and shows that a distribution of genus sizes like that found in the real world makes it possible for a type to show a significant change in frequency. In fact, in 15 % of the trials, the type either increased in frequency to over 60 % or decreased in frequency to less than 40 %; in other words, there is a 15 % chance of two types starting with equal frequency 3,700 years ago and ending up with one type more than 50 % more frequent than the other type. This shows that Maslova's conclusion that a type cannot change significantly in frequency over 3,700 years does not apply to the real world and is an artifact of features of her model that make it different from the real world.

The model just described is in fact rather conservative relative to the actual world, since it does not take into consideration the fact that genera within the same family are more likely to be of the same type. In the actual world, genera within the same family often share typological characteristics, and the historical factors leading to one genus within a family being large often lead to other

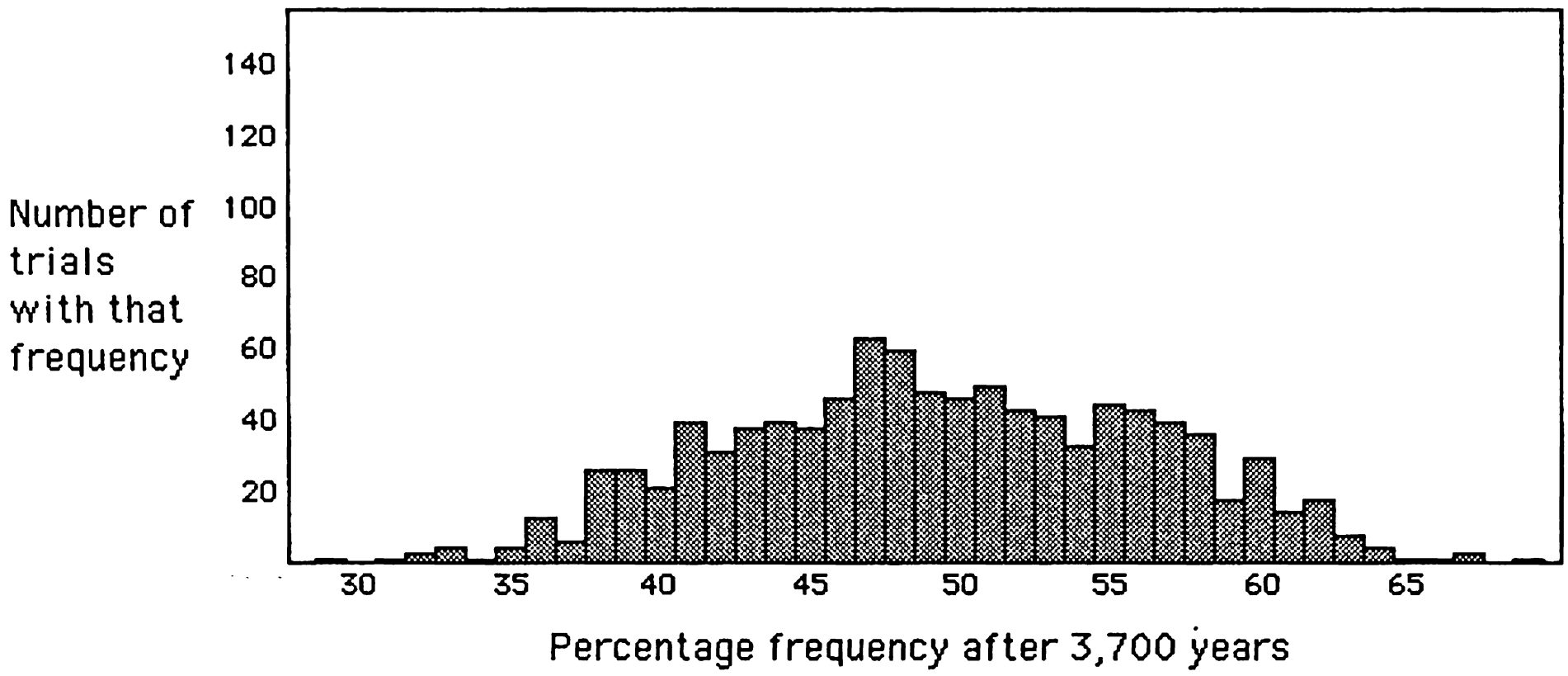

Figure 5. *Number of trials in which a type with initial frequency of 50 % occurred with the frequency given after 3,700 years, for a world where the distribution of genus sizes is the same as in the real world*

Table 4. *Likelihood of a type with initial frequency of 50 % having a frequency after 3,700 years within the frequency interval given, where the distribution of genus sizes is the same as in the real world*

Frequency interval	Percentage of trials within interval in %	Percentage of trials outside interval in %
[33...67]	99.4	0.6
[34...66]	98.5	1.5
[35...65]	98.1	1.9
[36...64]	97.5	2.5
[37...63]	95.6	4.4
[38...62]	94	6
[40...60]	85	15

genera in the same family being large. This is reflected by the fact that of the thirteen genera in the real world containing 100 or more languages listed above in Table 2, eleven are from just three families: six are Austronesian, three are Niger-Congo, and two are Sino-Tibetan. This consideration is not reflected in Figure 5 and Table 4.

We can add the significance of families to the model by weighting the probabilities so that although the first genus in each family has an even chance of being either of the two types, all other genera in the family have a greater than even chance of being the same type as the first genus. The data in Figure 5 shows the frequency distribution over 1,000 trials for this revised model, where

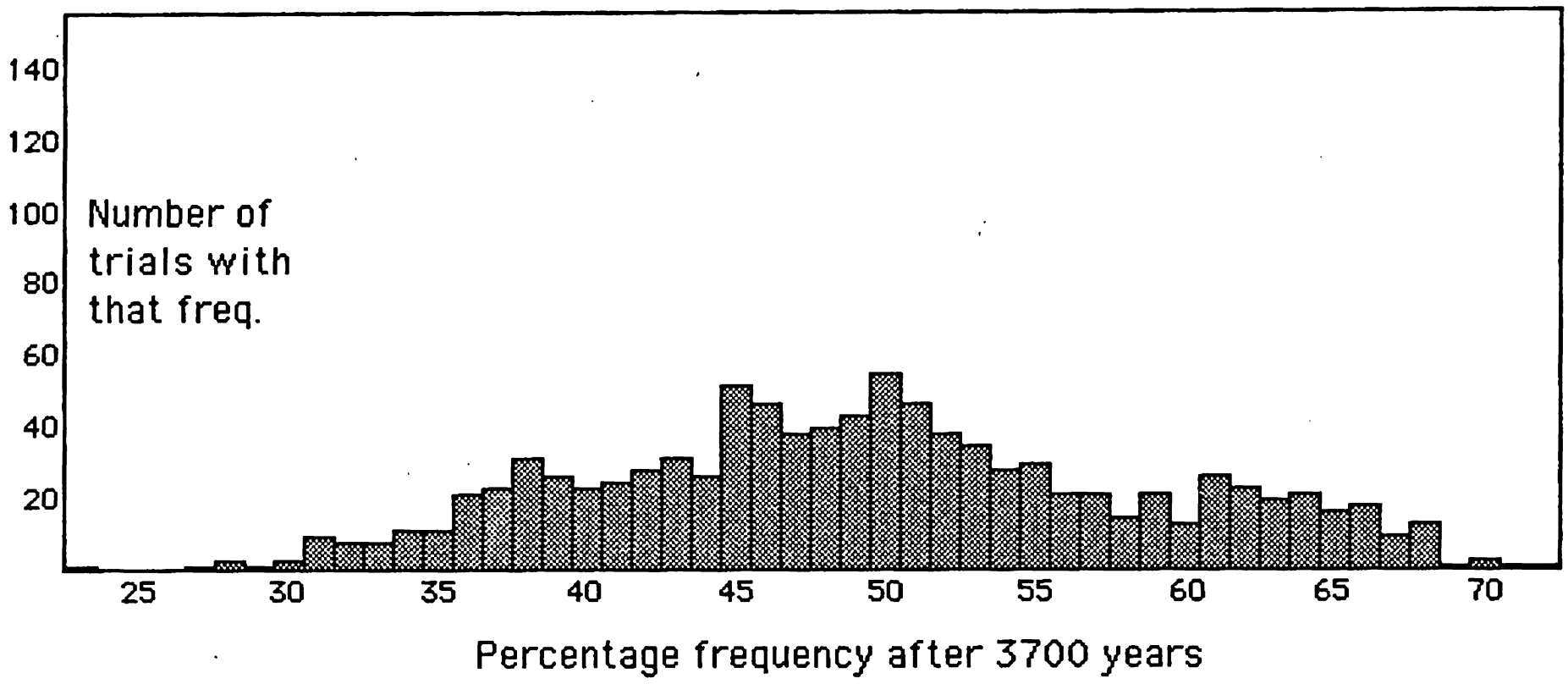

Figure 6. *Number of trials in which a type with initial frequency of 50 % occurred with the frequency given after 3,700 years, for a world where the distribution of genus sizes is the same as in the real world, with probabilities weighted so that genera within the same family tend to be the same*

Table 5. *Likelihood of a type with initial frequency of 50 % having a frequency after 3,700 years within the frequency interval given, where the distribution of genus sizes is the same as in the real world, with probabilities weighted so that genera within the same family tend to be the same*

Frequency interval	Percentage of trials within interval in %	Percentage of trials outside interval in %
[30...70]	99	1
[32...68]	97	3
[33...67]	94.9	5.1
[34...66]	93	7
[40...60]	69	31

each genus in a family other than the first one is given an 80 % chance of being of the same type as the first genus in the family. As before, this frequency distribution is summarized with different frequency intervals in Table 5.

Figure 6 and Table 5 show an even broader distribution from the preceding ones and show clearly how broad a frequency distribution we get as we construct models that more closely approximate the distribution of genus and family sizes found in the actual world, with many large genera, often within the same family. Table 5 shows that in only 93 % of the trials was one type less than twice as frequent as the other; in the other 7 % of cases, one type was more than twice as frequent, despite the fact that the number of genera of

the two types is the same. This shows that it is relatively easy for two types to be of equal frequency at one point in time, but for one type to be twice as frequent after 3,700 years. Table 5 also shows that in almost one third of the trials, one type was more than 50 % more frequent than the other after 3,700 years (outside the interval [40...60]).

The model that provides the basis of the figures in Figure 6 and Table 5 takes into consideration the fact that a few huge language families can sharply skew different types of languages in the actual world. It still, however, completely ignores the effects of areal phenomena; two adjacent families in that model are no more likely to be of the same type than two families in different parts of the world. What such a model would need to do would be to capture the fact that genera in different families that are geographically adjacent to each other have a greater than chance probability of sharing typological characteristics. Because of the complexities associated with constructing a model that captures this, I have not done this. But the addition of linguistic areas would presumably have an effect similar to that of moving from a model based entirely on genera (as in Figure 5 and Table 4) to one that considers families (as in Figure 6 and Table 5): namely it would increase even more the probability of a type changing significantly in frequency during a period of 3,700 years. In short, we have ample reason to reject Maslova's claim that a type is unlikely to change significantly in frequency during a period of 3,700 years.

7. Counting genera is not the same as counting languages 3,700 years ago

I have up to this point been following Maslova in assuming, for the sake of argument, that counting genera is equivalent to counting frequency 3,700 years ago. Her argument that there is no need to count genera rather than languages is based on her claim that changes in frequency over the past 3,700 years due to births and deaths are unlikely to be significant and her assumption that counting genera is equivalent to counting languages 3,700 years ago. I have argued above against the first half of this; I will argue here that there are significant differences between counting genera and counting languages 3,700 years ago. Maslova also argues that if there are significant differences between numbers of genera and numbers of languages, then these are unlikely to be due to births and deaths, but rather to type shifts. Again, this assumes that numbers of genera represent frequency 3,700 years ago.

The primary reason that numbers of genera cannot be equated with numbers of languages 3,700 years ago is that BOTH numbers of languages AND numbers of genera reflect type shifts during the past 3,700 years. If the protolanguage of a genus was of a particular type, and some of the languages in that genus have undergone a type shift so that they are of a different type from the protolanguage, then that genus will be counted among the genera containing languages

of the type towards which there has been a shift. For example, if the protolanguage of a genus was SOV and some of the languages in the genus are now SVO, then that genus will be included both among the number of genera containing SOV languages and among the number of genera containing SVO languages. And in some cases, none of the contemporary languages will be of the type of the protolanguage. Historical evidence suggests that proto-Semitic was VSO, but the contemporary spoken Semitic languages are either SVO or SOV. In counting genera for contemporary languages, Semitic is included in the counts for SVO and SOV but not for VSO. Hence, when there is a significant difference between numbers of languages and numbers of genera, these cannot be due primarily to type shifts, contrary to Maslova's claim, since type shifts will be reflected in both numbers. Rather such differences must be due primarily to births and deaths.

8. An example: the frequency of SVO word order

It is worth making the discussion more concrete by illustrating with a specific example. Maslova cites an example I discuss in Dryer (1989), that of the frequency of SVO word order. Tomlin (1986) estimates the proportion of languages with SVO order as being around 42 %. This figure is based on the frequency among actual languages, and his sampling technique deliberately includes more languages from genetic groups which contain many languages. However, I observe in Dryer (1989) that the frequency in terms of number of genera that are SVO is only around 26 %.[2] Thus in the example discussed above of Semitic, Semitic is both among the genera containing SOV languages and among the genera containing SVO languages. Hence the sum of the proportions of GENERA containing SVO, SOV, etc. will be more than 100 %. If, however, we count the SVO languages within Semitic as one subgenus and the SOV languages within Semitic as a second subgenus, then the sum of the proportions of SUBGENERA of the different types will total 100 %. The discussion here also systematically ignores the fact that many languages have sufficiently flexible word order that they cannot be assigned to one of the six traditional types, as discussed by Dryer (1997). Maslova claims, based on her own model, that this difference must be due to type shifts. However, I have argued against this above, both because her model underestimates the extent to which types can change in frequency and because both the number of languages and the number of genera reflect the effect of type shifts.

It is furthermore possible to demonstrate that the specific case of the different frequencies for SVO is due to births and deaths, more specifically to the historical accident of a huge number of births of SVO languages, leading to SVO being the dominant word order in the two largest genera in the world, Bantoid and Oceanic. According to *Ethnologue*, the number of Bantoid languages

is 646. Among the 15 Bantoid languages in my database, 14 (or 93 %) are SVO. If we take this frequency as representative of the frequency of SVO order among Bantoid languages, this leads to an estimate of 603 SVO languages in Bantoid. Similarly, *Ethnologue* lists 493 Oceanic languages. Among the 37 Oceanic languages in my database, 23 (or 62 %) are SVO. Again, we can use this figure to provide an estimate of the number of SVO languages in Oceanic as 306. We can thus estimate the number of SVO languages in these two genera as 909. Now, if we assume a figure of 6,700 as an estimate of the total number of languages in the world, this means that these 909 SVO languages in Bantoid and Oceanic constitute approximately 14 % of the languages of the world. That means that among the 42 % of languages that are SVO, about 14 % of the total are in Bantoid and Oceanic and the remaining 28 % are in the rest of the world. But this figure of 28 % is close to the 26 % proportion of genera that are SVO. That means that most of the difference between the 26 % proportion of genera and the 42 % proportion of languages is directly attributable to the large number of SVO languages in Bantoid and Oceanic, and hence to the historical factors that led to the huge expansion of these two genera. Hence we cannot only conclude that the difference between these two figures of 26 % and 42 % is due to births and deaths rather than type shifts, but we can specifically trace the source of the difference to the large number of births of SVO languages in these two genera.

9. Counting genera is not enough

The discussion so far formulates the question in terms of whether it is better to count genera or count languages. However, while I have argued here that it is better to count genera than to count languages, I argue in Dryer (1989) that counting genera is not enough. If we are testing a typological generalization involving a preference for one type over another, it is not sufficient just to show that there are more genera of that type, since one type may be represented by more genera due to historical accidents leading to that type being common in a particular linguistic area. Consider the data in Table 6 showing the relative frequency of Genitive-Noun (GN) and Noun-Genitive (NG) order among SVO languages.

Table 6. *The order of genitive and noun in SVO languages*

	Africa	Eurasia	Southeast Asia and Oceania	Australia and New Guinea	North America	South America	Total
SVO&GN	5	3	7	6	1	4	26
SVO&NG	26	5	11	1	2	0	45

In terms of the total number of genera, SVO&NG outnumbers SVO&GN by 45 genera to 26, a difference approaching 2 to 1. However, I proposed in Dryer (1989) that in order to conclude that there is a linguistic preference for one type over another, it must outnumber the other type in all continental linguistic areas. In my 1989 paper, I assumed five continental areas, but in more recent work (e.g., Dryer 1992), I have assumed six areas: rather than a Eurasian area covering all of mainland Eurasia and a Pacific area including all of Austronesian, New Guinea, and Australia, I now assume an area Southeast Asia and Oceania that includes the languages of southeast Asia, including all of Sino-Tibetan, plus all Austronesian languages, with a Eurasian area which contains the remaining languages of Europe and Asia and with a new Australia-New Guinea area. The data in Table 6 show that SVO&NG outnumbers SVO&GN in only four of the six areas, and in fact the other two areas are overwhelmingly SVO&GN. Hence the overall higher frequency of SVO&NG is not sufficient basis for concluding that there is a linguistic preference for this order. In fact, closer examination of the data in Table 6 reveals that the higher overall number of SVO&NG languages is due entirely to the large number of genera of this type in Africa: outside of Africa, SVO&GN actually outnumbers SVO&NG slightly, by 21 genera to 19. In this case, we have reason to believe that the overall higher number of genera is not indicative of a linguistic preference.

The use of genera rather than languages is therefore not based on an assumption that the overall frequency of genera is a valid basis for estimating the probability of a given linguistic type. Rather, the claim is that counting genera WITHIN EACH AREA is better than counting languages within each area. The reason for this is that within an area a single family can swamp the other families in that area. Note that in some families, a single family may contain a large proportion of languages in that area: the majority of languages in Africa are Niger-Congo, and the majority of languages in Southeast Asia and Oceania are Austronesian. Hence, it is only if we count genera WITHIN EACH AREA that we can test generalizations about whether a particular linguistic type is preferred over another.

10. Conclusion

I have argued in this paper that Maslova's model does not represent accurately features of the real world. These arguments are based, however, on assumptions as to which genetic groups should be counted as genera. As noted above, the decisions as to which groups are genera are based on my own educated guesses. An obvious objection to any claims based on these genera is that the conclusions might be artifacts of my own decisions as to which groups are genera, and someone else making their own educated guesses might come to

different conclusions as to which groups should be counted as genera. There is no denying that the lack of solid criteria for determining genera is a weakness in the methodology.

Let me address briefly the question of whether the conclusions of this response to Maslova might depend on my decisions as to what are genera, most specifically the claim that the real world contains far more large genera than her model predicts. Is it possible that the real world does not contain such large genera and that the large groups I assume to be genera are actually groups with a time depth greater than 4,000 years and that each group really consists of a number of genetic groups with this time depth, so that there are in fact few if any instances of genera containing more than 100 languages? Let me focus attention on the possibility of this being the case with the two groups that I claim to be especially large genera, namely Bantoid and Oceanic. My brief response is that in the case of these two genera, we actually have more archaeological evidence bearing on their time depth than we have for most genera. Since I am not an expert on this literature, I will not cite the relevant literature here, but Oceanic represents the easternmost spread of Austronesian languages into areas that were in many cases not previously inhabited, and archaeologists associate fairly specific dates less then 3,500 years for specific points in this spread. In the case of Bantoid, there is also extensive archaeological evidence of a spread of iron age technology through the majority of the area in which these languages are spoken, again at a time considerably less than 3,500 years ago. Thus while my guesses may be inaccurate in many instances, there does seem to be clear archaeological evidence suggesting that it is not plausible that these two groups really both consist of many subgroups all with a time depth of more than 3,500 years and all with fewer than 100 languages. In short, while there are legitimate overall concerns about the reliability of my guesses as to which groups are genera, there does appear to be ample reason to conclude that the real world does contain two huge genetic groups with a time depth of less than 4,000 years.

In the final paragraph to her paper, Maslova says "... an approach to statistical analysis of typological data cannot be verified or falsified by specific applications; it must be shown to be theoretically justified before it can be applied ..." While this may be true in principle, it is often the case that in practice the flaws in a particular approach to statistical analysis only become clear when one examines their specific applications. When I first read Maslova's paper, many of her arguments seemed quite sound, and it was only when I implemented a computer simulation of her model and compared the properties of her model with those of the actual world that the problems described above became clear to me.

I should emphasize that I have argued here only against certain claims of Maslova's; there is much else in her paper of merit and interest. The view

of linguistic preferences in terms of transitional probabilities rather than static probabilities seems fundamentally right. The endeavour of constructing mathematical models of the sort she proposes is worth pursuing, though I would immediately add that such work only becomes useful when it is shown that the models resemble the real world and when the method is applied to actual problems. But this will at the very least require a model more complex than Maslova's, one that represents the number of very small genera and the number of very large genera that we find in the real world.

Received: 2 October 2000

SUNY Buffalo

Notes

Correspondence address: Department of Linguistics, State University of New York at Buffalo, Buffalo, NY 14260-0001; e-mail: dryer@acsu.buffalo.edu

1. Since each type is assigned randomly to each genus, the relative frequency of the two types in the initial state (i.e., over genera) might deviate from 50 %–50 %. To avoid this problem, all trials in which the frequencies of the two types were not the same over genera were discarded.
2. Technically, this is actually the frequency among subgenera, where a subgenus is a set of languages within a genus of a particular type.

References

Dryer, Matthew S. (1989). Large linguistic areas and language sampling. *Studies in Language* 13: 257–292.

— (1992). The Greenbergian word order correlations. *Language* 68: 81–138.

— (1997). On the 6-way word order typology. *Studies in Language* 21: 69–103.

Grimes, Barbara F. (ed.) (1997). *Ethnologue: Languages of the World.* 13th edition. Dallas: Summer Institute of Linguistics.

Tomlin, Russell S. (1986). *Basic Word Order: Functional Principles.* London: Croom Helm.

The view from hologeistic linguistics

by *Revere D. Perkins*

In her paper Elena Maslova takes as her motivation the presumed inadequacy of the languages-as-trials approach in universals research. Whether she correctly reflects the views of authors from Greenberg to Rijkhoff and Bakker and including Croft, Dryer, and Hawkins, among others, I am unable to judge conclusively. Many of her motivating assumptions appear to me to be simply incorrect or, at least, misguided. Take, for instance, her statement, "The problem is that there seem to be no criteria that would allow for an empirical distinction between genuine distributional universals and accidental statistical tendencies" (p. 308). The idea of a need for an empirical distinction between GENUINE

universals and STATISTICAL tendencies does not seem at all reasonable to me (though there does seem some support for the idea that this distinction is held to obtain by some typologists). I take statistical tendencies to be the results produced by sampling particular languages to test one's hypotheses about universals. Those tendencies might not reflect universals if other factors adversely affect the test of the hypothesis, thereby biasing the results due to there being too many cases of a particular type. The hologeistic method as developed by Raoul Naroll (see Naroll et al. (1974) for a summary of the method) provides systematic statistical techniques for dealing with the problems that the author concludes (and quotes others as concluding) are insurmountable.

Like other authors in this domain, Maslova appears not to appreciate two related points: First, the requirement of statistical independence of cases is not independence in some absolute or non-statistical sense. Second, complete statistical independence is not required to make inferences from a sample and, in fact, the concept of "complete statistical independence" is a misguided concept; statistical independence is a matter of degree. There are several multivariate techniques that allow one to control for complicating factors like areal and genetic effects and determine the size of the effect that is hypothesized to exist. See Perkins (forthcoming) for more details. These points make the motivation for the alternative "solution" questionable at best.

Naroll worked out the general approach for multivariate analysis of complicating variables in cultural anthropology but the statistics he proposed have been superseded to a substantial extent by the introduction of more accurate statistics made possible by the speed of current computers. As an example of a multivariate approach for controlling for macro-areas, I use some data from Perkins (1992: 222–223) and further classify it by macro-area as defined in Nichols (1992: 26) and given in Table 1.

Using the software package MIM 3.0 as explained in Edwards (1995) and the data in Table 1 and starting with a model that includes relationships between all three variables, the model, developed by elimination, that is most corroborated by the data is one that includes relationships between Cultural Complexity and Deictic Elaboration as well as one between Deictic Elaboration and Macro-Area.

This model is based on exact tests involving exhaustive enumeration of all the possible distributions of cell counts and adjusted degrees of freedom are used for calculating probabilities because of scarcity of data in some cells.

The variables Deictic Elaboration and Cultural Complexity are associated with a probability of 0.0020 of having occurred by chance and Deictic Elaboration and Macro-Area are associated with a probability of 0.0266 of having occurred by chance and Cultural Complexity and Macro-Area are associated with a probability of 1.0 of having occurred by chance. These results confirm that although there is a statistically significant association between Deic-

Table 1. *Distribution of Cultural Complexity and Deictic Elaboration values across Macro-Areas*

Cultural Complexity	Deictic Elaboration	Macro-Area	Number of Languages
Not Complex	Minimal	Old World	2
Not Complex	Minimal	New World	0
Not Complex	Minimal	Oceania	1
Not Complex	Moderate	Old World	5
Not Complex	Moderate	New World	0
Not Complex	Moderate	Oceania	5
Not Complex	Extensive	Old World	2
Not Complex	Extensive	New World	5
Not Complex	Extensive	Oceania	9
Complex	Minimal	Old World	10
Complex	Minimal	New World	1
Complex	Minimal	Oceania	1
Complex	Moderate	Old World	4
Complex	Moderate	New World	0
Complex	Moderate	Oceania	4
Complex	Extensive	Old World	0
Complex	Extensive	New World	0
Complex	Extensive	Oceania	0

tic Elaboration and Macro-Area, this relationship does not explain or account for the relationship between Deictic Elaboration and Cultural Complexity. The latter association is statistically independent of the area variable. By means of a multivariate approach the influence of Macro-Area on the focal relationship may be determined. By extension of the technique other control variables may be included as well. With exact calculations scarcity of data is no longer a problem.

Maslova's proposed alternative solution introduces the technique of Markov modeling to typological research. It is a welcome addition to the methods available to linguists but does not require the author's critical stance toward a languages-as-trials approach. More explanation of the technique and its applicability to the problem dealt with would have been welcome. The author admits that the proposed solution requires many questionable (and probably incorrect) assumptions, such as the independence of linguistic and non-linguistic events, and the notion that at some time depth linguistic families were truly independent. Their explicit formulation, however, does make it possible to test

some of those assumptions and hence the viability of the proposal, which is laudable. On the other hand, I think, the questions that typologists hope to deal with extend beyond the proportions of types, whether now or 4,000 or 10,000 years ago.

Received: 7 October 2000

CYMFONY, INC.

Notes

Correspondence address: 39 Pinewood Drive, Springville, NY 14141, USA; e-mail: reveredp@adelphia.net

Many thanks to Bill Pagliuca for substantial assistance in producing this response.

References

Edwards, David (1995). *Introduction to Graphical Modelling*. New York: Springer.

Naroll, Raoul, Gary L. Michik, & Frada Naroll (1974). Hologeistic theory testing. In Joseph G. Jorgensen (ed.), *Comparative Studies by Harold E. Driver and Essays in his Honor*, 121–148. New Haven: HRAF Press.

Nichols, Johanna (1992). *Linguistic Diversity in Space and Time*. Chicago: University of Chicago.

Perkins, Revere D. (1992). *Deixis, Grammar, and Culture*. Amsterdam: Benjamins.

— (forthcoming). Sampling procedures and statistical methods. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, & Wolfgang Raible (eds.), *Language Typology and Linguistic Universals: An International Handbook*. Berlin: de Gruyter.

Few languages in a short time interval?

by *Fritz Schweiger*

These comments are restricted to the attempt to define distributional universals as the stable distribution of a Markov process. Maslova introduces the concept of LANGUAGE POPULATION. Formally, $A(t)$ is a finite set whose elements are called LANGUAGES, and $N(t)$ denotes the number of languages at time t. A TYPOLOGY $\mathbf{T}=\{\mathrm{T}_1,\ldots,\mathrm{T}_M\}$ is a partition of $A(t)$ together with a set of probabilities $p_1(t),\ldots,p_M(t)$ which should be interpreted as $p_i(t)$ being the probability that a language $L \in A(t)$ belongs to T_i. This set of probabilities is called the A-DISTRIBUTION (where "A" can be read as either "actual" or "accidental"). They are given by Laplace's definition (1).

$$(1) \qquad p_i(t) = \frac{\#\{L \in A(t) : L \in \mathrm{T}_i\}}{N(t)}, \; 1 \leq i \leq M$$

Clearly, a different problem is to give estimates for the values $p_i(t)$ by considering only a sample of languages, i.e., a subset of $A(t)$. This leads to the well known problem how to find suitable subsets of a language population for typological research. Since Maslova wants to define probabilities which do not

depend on the selected time t she postulates that there are stationary transition probabilities p_{ki} which are seen as the probability that a language of type T_k changes to type T_i "after a small time interval". Mathematically this means:

$$(2) \qquad p_i(t+\Delta t) = \sum_{k=1}^{M} p_k(t)\, p_{ki},\ 1 \le i \le M$$

Since languages do not change very fast, $\Delta t \approx 500$ years may be a reasonable guess. Maslova calls a time interval "small" if a type-shift is possible, but not highly probable. Note, that the size $N(t)$ of the language population is not relevant any more (and the discussion of birth-and-death processes in the paper should provide some arguments that for "large language populations" this sounds reasonable).

Therefore the mathematical model reduces to a finite Markov chain with transition probabilities p_{ki}, $1 \le k,\ i \le M$. The observation that some changes follow preferred directions is reflected by higher probabilities p_{ki}. If this Markov chain is ergodic, then a stable distribution $p(\mathrm{T}_1), \ldots, p(\mathrm{T}_M)$ exists which mathematically is uniquely defined by the equations:

$$(3) \qquad p(\mathrm{T}_i) = \sum_{k=1}^{M} p(\mathrm{T}_k)\, p_{ki}, 1 \le i \le M$$

The essential condition for ergodicity is that after a finite number of steps any change is possible, i.e., for any pair (i,k) there is an $S = 1$ such that $p_{ki}^{(S)} > 0$. Here $p_{ki}^{(S)}$ are the entries of the S-th power of the matrix $((p_{ki}))$.

Maslova claims that the stable distribution should be seen as an appropriate definition of the DISTRIBUTIONAL UNIVERSAL related to the typology **T**. Unfortunately Maslova does not provide a single example of a typology for which this model is tested.

However, some methodological questions arise. First of all the application of time-dependent probabilistic models is questionable due to the fact that the numbers $N(t)$ are actually "small numbers" and the observable time interval $[t_0, t_0 + \tau]$ also is "short".

At the first moment an estimated average size $N(t) \approx 6{,}000$ and a time period $\tau \approx 6{,}000$ years, say, seem to be "large numbers". The number $N(t) \approx 6{,}000$ is large enough to make data collection and their interpretations by statistical methods reasonable but a probabilistic model is quite a different thing. This can be illustrated by the paradox of large blocks of events.

Let us consider the well known toss of a coin. Clearly, in the standard model we suppose the probability of tail and head are equal, $p(T) = p(H) = 0.5$. By the same reasoning we find $p(TT) = p(TH) = p(HT) = p(HH) = 0.25$. Therefore (one may see this as doing justice to any block) any block of length

n has the same positive probability 2^{-9}. Now let us start playing with a coin. If the sequence of the first 9 outcomes is *TTHTHHTHH* no one will suspect anything. If the sequence starts with *TTTTTTTT* no one will believe that the coin is "fair". But the model itself predicts that a block of 9 consecutive "tails" has a small but positive probability 2^{-n}. Even more, if the coin is "fair" this block must appear infinitely often when the game goes on infinitely long. After 2^{12} throws say, approximately $2^3 = 8$ blocks *TTTTTTTT* should appear and no one could exclude the possibility that the sequence of outcomes starts with this block. Therefore, a sequence of 100 throws, cannot be seen as a sample for the "ideal coin" but only as a sample for some properties of the "ideal coin", e.g., the appearance of blocks of length 1, 2, or 3. However, the belief that an experiment with a coin confirms the idea of an "ideal coin" as a model is supported by the fact that this experiment can be repeated and one can easily enlarge the number of tosses. But the interpretation of short sequences of outcomes as samples for an "ideal coin" is based not only on empirical reasons but on theoretical considerations about equal likelihood and independence of outcomes (Laplacian model). This means that any probabilistic model which sees a language population as a sample must be rooted in theoretical considerations about the possible validity of the model. Probabilistic models in biology as in population genetics, say, also deal with great numbers (of bacteria or rabbits). The observed time is considerably longer (compared with the lifetime of the observed species) and the experiment can be repeated too. Nothing comparable is true for a probabilistic model of language populations. The uniqueness of the historical development of languages therefore requires extreme care in applying probabilistic models to typological change. Clearly this does not imply that such a viewpoint cannot be interesting or that it would not be worthwhile to consider the possibility of such a model.

It is claimed that the transition probabilities p_{ki} do not depend on time and that the underlying Markov chain is ergodic. But if the transition probabilities are stationary this implies that there is a "language inherent" probability that a language changes from type T_k to type T_i. This is not welcome for any attempt to explain language change. At least there is evidence that typological changes strongly interact with areal distributions, which fact is not covered by the model. Furthermore, almost no language is known which really substantiates ergodicity in its known short life cycle (of, say, 4,000 years). Some accusative languages became ergative languages and some ergative languages became accusative languages but I do not know of any language which is known as an ergative language at time t_0, as accusative at time $t_1 > t_0$, and as again ergative at time $t_2 > t_1 > t_0$. Here one sees that Maslova's interpretation of ergodicity is a vulnerable concept for language history. Maslova writes: "The probability $p(T_i)$ can be thought of either as the probability that a language will be found in state T_i at a randomly selected moment of its HISTORY or as the

probability for a randomly selected member of a large language POPULATION to be in this state." This leads to the problem of the IDENTITY OF A LANGUAGE through history. Latin clearly is of a different morphological type than Italian, but to which extent is Italian the same language as Latin? However, for the mathematical model this interpretation of ergodicity is not needed.

This leads to the question of the usefulness of probabilistic (and other mathematical) models. I mention EXPLANATION and PREDICTION. Probabilistic models may be used to predict observable events in the future. There is no doubt that gambling by and large follows the proposed rules. Since the days of Gregor Mendel we know that the procreation of species is governed by probabilistic laws. Last but not least the calculations of insurance companies are built on probabilistic models. Since typological changes seem to be happen in longer time intervals in the moment no prediction can be controlled. Explanation is a different thing. Here one assumes that the past is governed by essentially the same rules as the present. In a very strict sense this is possible when laws of physics and chemistry are applied. In biology we are less certain and dealing with historical facts often different and even competing explanations are offered. In any case explanation still must have a kind of controllability. Here I mean that observation over a time interval (of the past!) offers a prediction for the following (already past!) time interval. A simple example could be a linguistic statement like "If remote languages come into contact they will become similar due to linguistic diffusion". No prediction for the future is possible but this statement seems to offer some explanation for language changes in the past.

But let us be optimistic. Suppose we get some crude estimates for

$$p_{ki} \approx \frac{p_i(t+\Delta t)}{p_k(t)}$$

and calculate the stationary probabilities $p(\mathrm{T}_k)$, $1 \leq k \leq M$. Let $t_p = t_0 + \tau$ be the present time. If $p_k(t_p) = p(\mathrm{T}_k)$ then the language population of the present world has already reached its equilibrium. If $p_k(t_p) \neq p(\mathrm{T}_k)$ the stable distribution will eventually reached in some remote future but since the number of languages will be drastically reduced shortly this distribution will never be observed. Unfortunately this does not bode well for the richness and variety of languages.

Received: 28 September 2000

Universität Salzburg

Correspondence address: Institut für Mathematik, Universität Salzburg, 5020 Salzburg, Austria; e-mail: fritz.schweiger@sbg.ac.at