

Counting Genera vs. Counting Languages:
A Response to Maslova

Matthew S. Dryer
SUNY Buffalo

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 3500 to 4000 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 (1) Maslova makes rather unlikely assumptions about the number of languages 3700 years ago; (2) 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; (3) her model fails to capture the effect that such large genera can have on altering the frequency of a linguistic type after 3700 years; (4) her discussion confuses frequency of a type among genera with frequency of a type 3700 years ago; (5) counting genera provides a better basis for testing typological generalizations than counting actual numbers of languages; and (6) counting genera is not enough.

1. How many languages were there 3700 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 3700 years ago was probably less than 1000. Her apparent view is that over much of the past 3700 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 such a view is seriously mistaken, and that there is no reason to believe that the number of languages spoken 3700 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 3700 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 5500 languages 3700 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 3700 years.

2. 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 8 below) counting what I call genera, genetic groups of a time depth of 3500 to 4000 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 3700 years ago. Hence the difference between counting genera and counting languages is, she claims, roughly the difference between counting languages 3700 years ago and counting languages today. Given this view, she raises the legitimate question: why should counting languages spoken 3700 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 3700 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 3700 years ago was less than it is today, an assumption I challenge in the preceding section.

3. 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 Fig. 1 gives the distribution found over 1000 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 argue in section 1 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 3700 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 1000 trials was 5545 (just a bit less than the 6000 Maslova claims) and the average number of genera was 395, very close to the 400 assumed by Maslova.

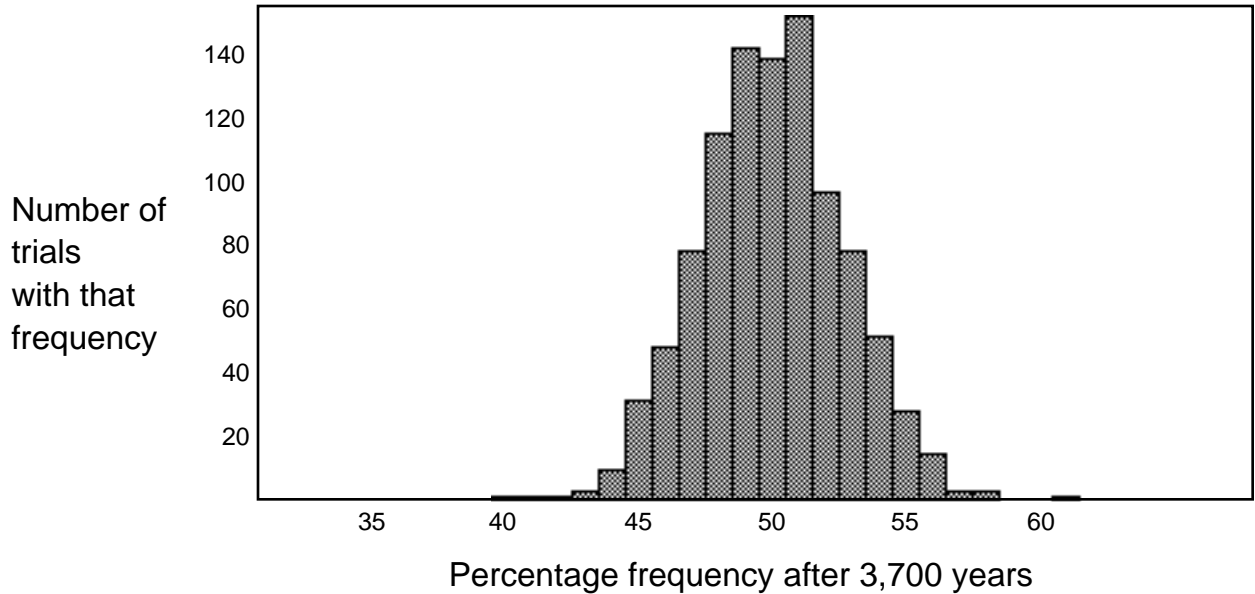


Figure 1

Number of trials in which a type with initial frequency of 50% occurred with the frequency given after 3700 years, where initial number of languages is 600

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 Fig. 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.

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

Table 1

Likelihood of a type with initial frequency of 50% having a frequency after 3700 years within the frequency interval given, where initial number of languages is 600

Table 1 shows that the probability of a type changing from 50% to something within the frequency interval of 45% to 55% inclusive is .96; in other words, there is only a .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 frequency of a type after 3700 years (as a proportion of all languages) will not be significantly different from its initial frequency.

Fig. 1 and Table 1 are based on the assumption that the number of languages spoken 3700 years ago was 600. I ran a similar simulation for the alternative assumption that the number of languages was 6000, approximately the same as in the present. The assumed birth and death rate were, following Maslova, both 0.38. The results, based again on 1000 trials, are not significantly different from those shown in Fig. 1 and Table 1. The average number of genera after 3700 years was 368, a little less than the 400 claimed by Maslova. Table 2 and Fig. 2 give data comparable to Table 1 and Fig. 1.

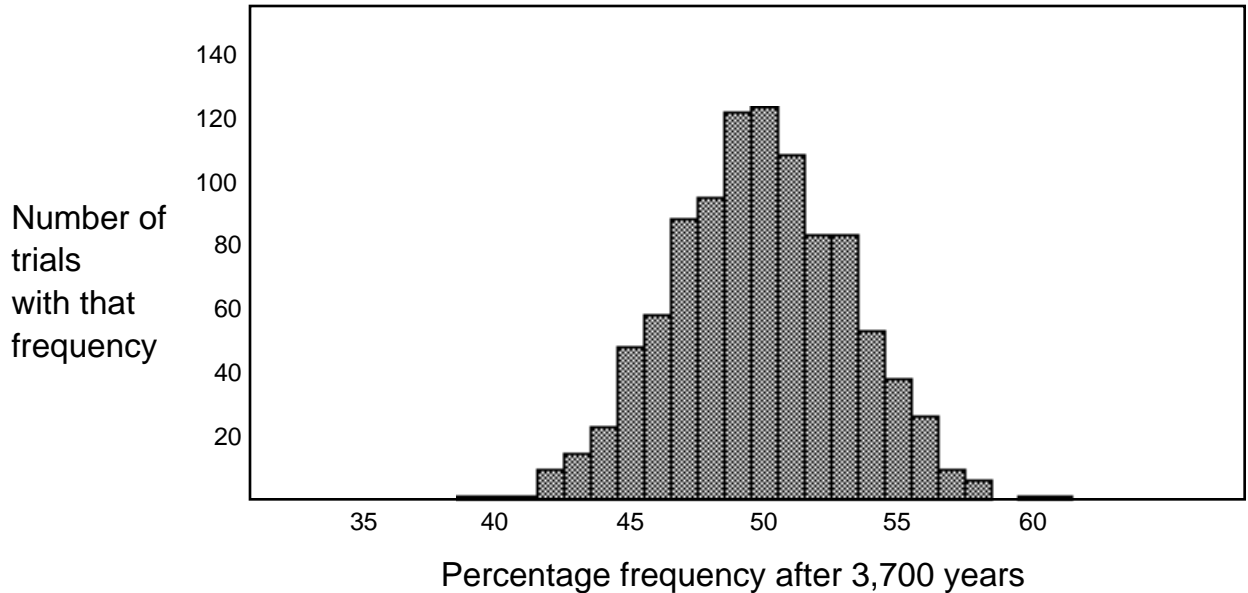


Figure 2

Number of trials in which a type with initial frequency of 50% occurred with the frequency given after 3700 years, where initial number of languages is 6000

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

Table 2

Likelihood of a type with initial frequency of 50% having a frequency after 3700 years within the frequency interval given, where initial number of languages is 6000

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 6000 languages, there is a slightly greater chance of the frequency changing. For example, while the probability of remaining within the

frequency interval [45 ... 55] is .96 for an initial state of 600 languages, Table 2 shows that the probability of remaining within this frequency interval is only .90 for an initial state of 6000 languages.

4. 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 3700 years. During a period of 3700 years, some 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. Fig. 3 shows, as a bar graph, the average percentage, over 1000 trials of the simulation of her model, of genera of different sizes.

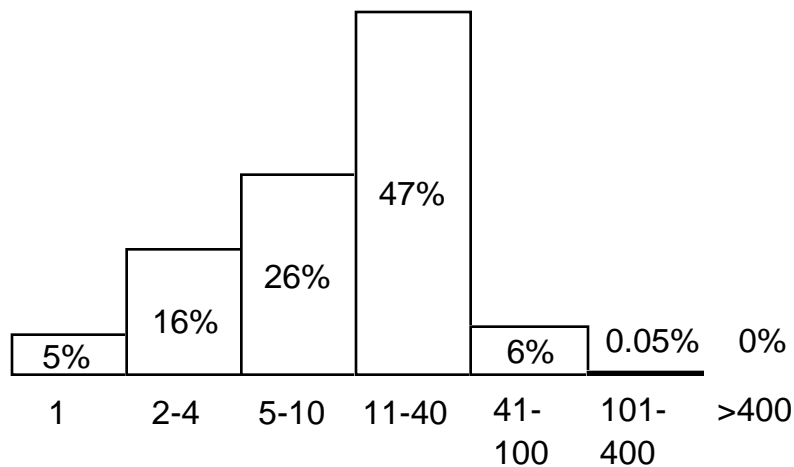


Fig. 3
Percentage of genera with this many languages
according to Maslova's model

Fig. 3 shows that under Maslova's model, on average only 5% of genera will contain exactly one language after 3700 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 3700 years, and the largest genus over all 1000 trials contained only 172 languages.

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

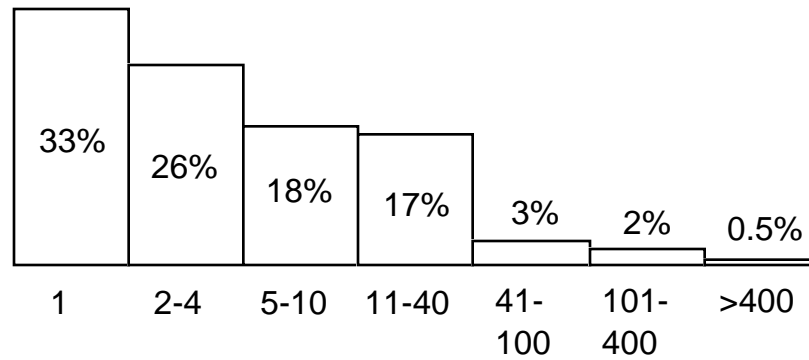


Fig. 3
Percentage of genera with this many languages
in the real world

The most obvious difference between Figures 3 and 4 is that there are far more small genera in the actual world, as shown in Fig. 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, Table 12 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.

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

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

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 1000 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 1000 years. For example, we could revise the model so that the probability of splitting into two languages is .1 if the language has split into two languages within the past 1000 years but only .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 2000 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 Fig. 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.

5. 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. [Footnote: 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.] 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 Fig. 5 gives the frequency distribution over 1000 trials for one of the two types.

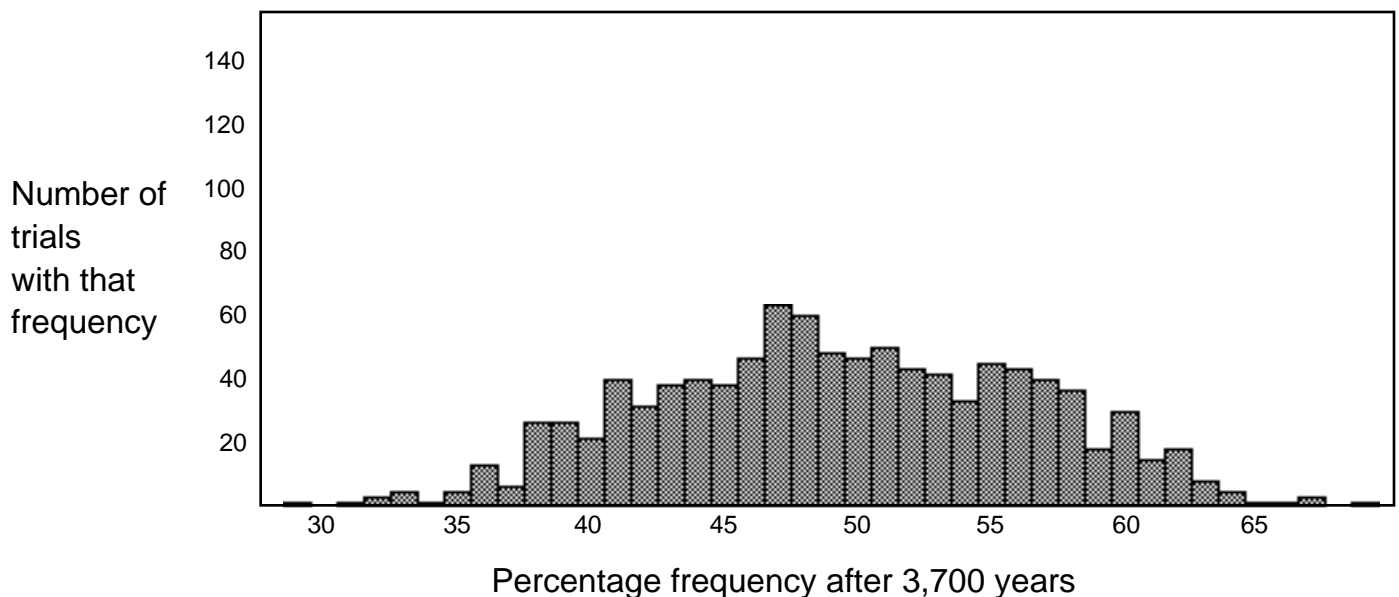


Figure 5
Number of trials in which a type with initial frequency of 50% occurred with the frequency given after 3700 years, for a world with a distribution of genus sizes is the same as the real world

Even at a glance, Fig. 5 looks very different from Figs. 1 and 2 based on Maslova's model: the shape of the distribution is much flatter in Fig. 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 Fig. 5 by showing the percentage of trials in which the percentage of one type fell within the frequency intervals indicated.

frequency interval	% of trials within interval	% of trials outside interval
[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%

Table 4

Likelihood of a type with initial frequency of 50% having a frequency after 3700 years within the frequency interval given, where the distribution of genus sizes is the same as the real world

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 3700 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 3700 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 genera in the same family being large. This is reflected by the fact that of the 13 genera in the real world containing 100 or more languages listed above in Table 2, 11 are from just 3 families: 6 are Austronesian, 3 are Niger-Congo, and 2 are Sino-Tibetan. This consideration is not reflected in Fig. 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 Fig. 5 shows the frequency distribution over 1000 trials for this revised model, where each genus in a family other than the first one is given an 80% of being of the same type as the first genus in the family.

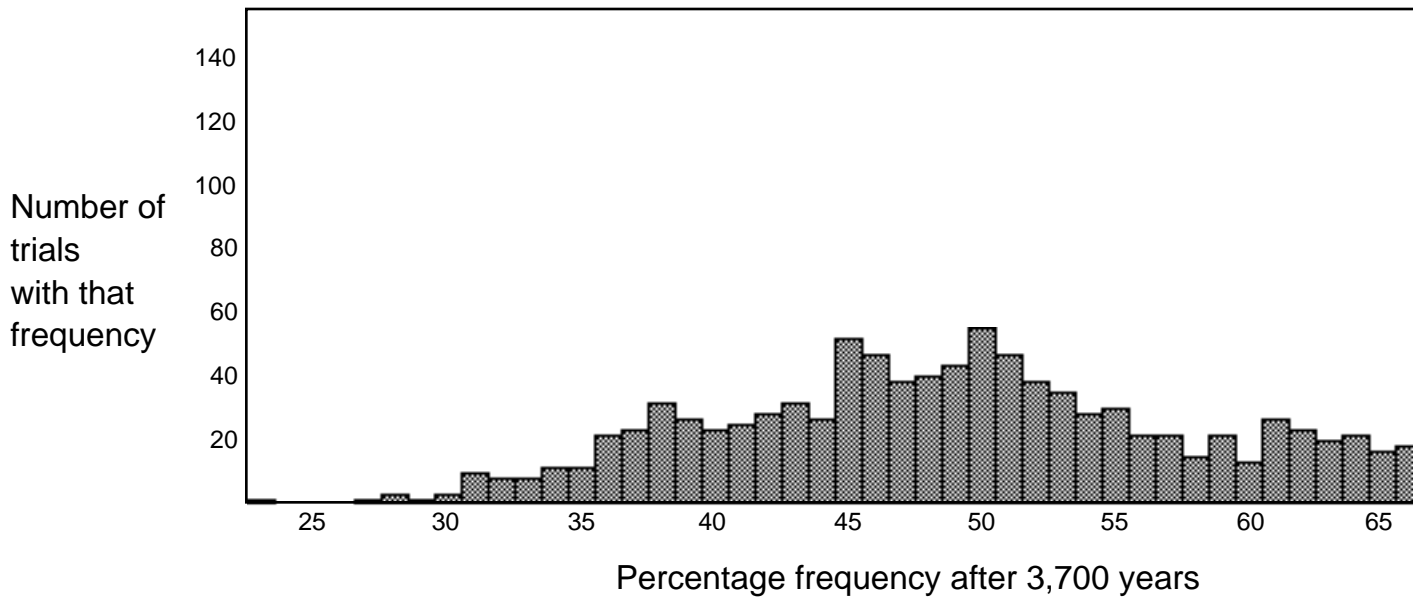


Figure 6

Number of trials in which a type with initial frequency of 50% occurred with the frequency given after 3700 years, for a world with a distribution of genus sizes is the same as the real world, with probabilities weighted so that genera within the same family tend to be the same

As before, this frequency distribution is summarized with different frequency intervals in Table 5.

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

Table 5

Likelihood of a type with initial frequency of 50% having a frequency after 3700 years within the frequency interval given, where the distribution of genus sizes is the same as the real world, with probabilities weighted so that genera within the same family tend to be the same

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 3700 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 3700 years (outside the interval [40 .. 60]).

The model that provides the basis of the figures in Fig. 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. Incorporating areal phenomena into a model is a far more complex matter than genetic factors, and I have not attempted to construct a model that does this. 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 Fig. 5 and Table 4) to one that considers families (as in Fig. 6 and Table 5): namely it would increase even more the probability of a type changing significantly in frequency during a period of 3700 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 3700 years.

6. Counting genera is not the same as counting languages 3700 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 3700 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 3700 years due to births and deaths are unlikely to be significant and her assumption that counting genera is equivalent to counting languages 3700 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 3700 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 3700 years ago.

The primary reason that numbers of genera cannot be equated with numbers of languages 3700 years ago is that *both* numbers of languages *and* numbers of genera reflect type shifts during the past 3700 years. If a 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.

7. 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%. [footnote: Technically, this is actually the frequency among subgenera, where a subgenus is a set of languages within a genus of a particular type. 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 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 6700 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.

8. 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.

	Africa	Eurasia	SEAsia&Oc	Aus-NewGui	NAmer	SAmer	Total
SVO&GN	5	3	7	6	1	4	26
SVO&NG	26	5	11	1	2	0	45

Table 6
The order of genitive and noun in SVO languages

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.

9. 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 4000 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 than 3500 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 3500 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 3500 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 4000 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.

References

- Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13: 257-292.
- Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68: 81-138.

- Dryer, Matthew S. 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.
- Maslova, Elena. 2001. A dynamic approach to the verification of distributional universals. *Linguistic Typology* (this issue).
- Tomlin, Russell S. 1986. *Basic Word Order: Functional Principles*. London: Croom Helm.