Studies in Language 13-2. 257-292 (1989).

LARGE LINGUISTIC AREAS AND LANGUAGE SAMPLING

MATTHEW S. DRYER
University of Alberta & State University of New York at Buffalo

1. Introduction

Claims are often made by linguists that certain properties are **typical** or **normal** properties of language, or that certain properties are found more often than others, or that there is a linguistic preference for some property over another. Some of these claims involve associations or correlations between one linguistic property and another, such as the claim that OV languages tend to be postpositional, while VO languages tend to be prepositional (*cf.* Greenberg 1963a). Other claims refer to a single property, such as the claim that SOV order is preferred over other orders (*cf.* Keenan 1979). The question addressed in this paper is how we can test such claims. I will argue that previous attempts to address such questions have underestimated the effect of areal phenomena, and that large-scale areal phenomena may be more widespread than is generally thought. Finally, I will propose a method for testing claims regarding linguistic preferences that controls for such areal phenomena.

2. Previous methods of sampling

Consider how we might test the hypothesis that there is a linguistic preference for SOV order over other orders. As a first approximation, suppose we were to collect data on the basic clause order for a convenience sample of 40 languages, *i.e.* 40 languages for which data is readily available, while attempting to include languages from as many areas of the world as possible. And suppose further that in this sample of 40 languages, there were 20 SOV languages, 15 SVO languages, and 5 VSO languages. Given that the sample contains more SOV languages than any other type, could we there-

fore conclude that there is a linguistic preference for SOV order? For a variety of reasons, the answer is no. An initial question would be whether the sample was representative: perhaps it is just a coincidental property of this particular sample, and SOV order is not most common in other samples of 40 languages that we could have chosen. A related question is whether the higher number of SOV languages in the sample is statistically significant. Intuitively, the difference between 20 SOV languages and 15 SVO languages is a small difference, and could very easily be an accidental property of this particular sample. After all, if one flips an unbiased coin 100 times, one doesn't expect to necessarily get exactly 50 heads and 50 tails; the fact that the number of heads and the number of tails are not identical does not mean one should conclude that the coin is biased.

A recent study by Tomlin (1986) provides data that would seem to answer the question of whether there is a linguistic preference for SOV order. He collected data on basic clause order for a sample of 402 languages that can be described as a **proportionally representative sample** of the world's languages. Following a methodology discussed by Bell (1978), Tomlin constructed his sample so that each language family and each linguistic area is represented to an extent that is proportional to the number of languages in the family. Assuming that the number of languages in the world is approximately 4800, his sample included about one twelfth of the languages of the world, and about one twelfth of the languages in each family and area. Thus for example, since there are approximately 439 Bantu languages, he constructed his sample so that it contained 33 Bantu languages, about one twelfth of the total. Using this methodology, he arrived at the following relative frequencies of the six possible basic orders:

SOV	44.78%
SVO	41.79%
VSO	9.20%
VOS	2.99%
OVS	1.24%
OSV	0.00%

Although it is not clear what the margin of error for these figures is, Tomlin's methodology assures us that we can take these figures as good estimates of the relative frequency among the entire set of existing languages. And although SOV order is the most frequent order here, the difference between the number of SOV languages and the number of SVO languages

is sufficiently small that it seems, intuitively, to be the kind of difference that could be due to chance, and hence that there is no reason to believe that there is a linguistic preference for SOV order, and that even if there is some preference for SOV order, the preference must be sufficiently small as to be largely uninteresting. I will argue, however, that Tomlin's data is misleading, and that there is evidence of a linguistic preference for SOV order over other orders.

There are dangers in attempting to draw inferences about linguistic preferences from the actual frequency of different language types, even when, as in the case of Tomlin's data, these frequencies can be considered reliable estimates of the relative frequency among all of the languages of the world. The actual frequency of different language types among the languages of the world is due to three factors. First of all, the frequency is partly due to what we might call linguistic factors or principles. The rarity of object-initial languages is most likely due to some such principles (*cf.* Tomlin 1986 for some suggestions), and seems unlikely to be a coincidence. In other words, there is probably a linguistic **dispreference** for object-initial order. On the other hand, the actual frequency of different types is also due, in part, to nonlinguistic historical factors. The intuition that the slightly higher frequency of SOV order as compared to SVO order in Tomlin's data could well be due to chance reflects an assumption that it is as likely due to nonlinguistic historical factors as to linguistic ones, and that had the history of the world been different, we might just as likely have found a slightly higher frequency of SVO order. In other words, the actual frequency of different language types is partly due to random variation.

A third factor influencing the actual frequency of different language types among the languages of the world is that of the historical factors that have resulted in certain large language families. The point is best illustrated by a hypothetical example. Imagine a world with 1000 languages. Suppose these 1000 languages are distributed over eleven language families such that there is one large family containing 900 languages and ten small families each containing 10 languages. Suppose further that all 900 languages in the large family are SVO but that the languages in the ten small families are all SOV. *I.e.* in this world, 90% of the languages are SVO, while only 10% are SOV. What could we conclude about the possibility of a linguistic preference for SVO over SOV? Would the fact that 90% of the languages are SVO provide a reason for concluding that there is a linguistic preference for SVO order? Clearly not. It should be clear in this hypothet-

ical case that the higher frequency of SVO order is not due to linguistic factors or principles, but rather to the fact that the one language family with SVO languages happens to be very large. And since the size of a family is due, presumably, to nonlinguistic historical factors, the greater frequency of SVO order in such a world would not provide a basis for concluding that there is a linguistic preference for SVO order. In fact, if we assume that each of the families is areally distinct in this hypothetical world (*i.e.* if we assume that the geography of this world is such that the effects of contact can be discounted), then we could conclude that there is a linguistic preference for SOV order, despite the fact that it is found in only 10% of the languages in that world. The reason we could conclude such is that the number of SOV **families** would outnumber SVO families by 10 to 1, and such a difference is statistically significant (by a simple binominal test).

Now although the actual frequencies of the different word order types in the real world has not been affected by large language families to the extent that it is in the above hypothetical case, there is evidence that those frequencies are affected in major ways by certain large families. For example, extrapolating from Tomlin's data we can estimate that about 71% of the VSO languages in the world are in the Austronesian family. Hence the actual frequency of VSO order in the world is considerably higher than it would be had it not been for the nonlinguistic factors that resulted in the large size of the Austronesian family. It is simply a matter of historical accident that certain people in southeast Asia happened, presumably, to speak a verb-initial language at the time certain nonlinguistic factors caused them to move out into an area that, again by historical accident, happened to consist of a very large number of islands, this being a primary reason for the large size of the Austronesian family. Had it not been for this fortuitious combination of circumstances, the frequency of VSO order among the languages of the world might well have been less than 3%, rather than over 9%, as Tomlin's evidence suggests it actually is.

Secondly, and more crucially to the matter at hand, it can also be extrapolated from Tomlin's data that about 40% of the SVO languages in the world are Niger-Congo languages. If it were not for whatever historical factors led to the large size of this family, particularly those leading to the relatively recent expansion of speakers of Bantu languages, the number of SVO languages in the world would have been considerably lower, in fact not much more than half the number of SOV languages. Hence is is dangerous to try to conclude from Tomlin's data that there is no linguistic preference for SOV order. What we would like to do is determine whether the difference between the frequencies of SOV and SVO order is statistically significant. But we cannot apply the relevant statistical tests, at least straight-forwardly, to Tomlin's data, because such tests require that the items in the sample be independent. But many of the languages in Tomlin's sample are not independent. As noted, for example, his sample contains 33 Bantu languages. In other words, although Tomlin's methodology allows one to obtain reliable estimates of the relative frequency of different language types among the languages of the world, it does not allow one to determine the extent to which those frequencies are due to linguistic factors, as opposed to nonlinguistic ones, and hence no way to determine whether there are statistically significant linguistic preferences for one language type over another.

The problem can be restated in terms of the sampling notions discussed by Bell (1978). Bell distinguishes the *universe* (the class of objects which is the object of investigation), the *frame* (the subset of the universe which one has access to), and the *sample* (the collection of objects that is actually observed). If one is investigating the relative frequency of different types, then one's universe is the set of existing languages and Tomlin's methodology is called for. However, if one is investigating linguistic preference or correlations between linguistic variables, then one's universe is the set of possible human languages and Tomlin's methodology is inappropriate because the set of existing languages does not constitute a reliable frame from which inferences about the universe can be drawn, primarily because of the distorting effects of large families.

One solution to this problem is to construct a sample of **independent** languages. This is the methodology employed by Perkins (1980, 1985) in testing hypotheses about the relationship between language characteristics and culture characteristics. He constructed a sample (a variant of which is used by Bybee 1985) consisting of 50 languages, no two of which are from the same family or from the same culture area, assuming a taxonomy of culture areas based on Kenny (1974). With such a set of independent languages, it is possible to apply statistical tests to test for linguistic preferences or associations between linguistic variables.

Questions arise, however, whether the 50 languages in Perkins' sample are really independent. Three of them, for example, are generally classified as Nilo-Saharan: Ingassana, Maasai, and Songhai. Nilo-Saharan is a very remote grouping, however, and the inclusion of Songhai within it is con-

troversial. It is probably not unacceptable to include within a sample pairs of languages which are very remotely related. Independence is a relative notion. Where two closely related languages share a given characteristic, they will generally share the characteristic because of their common origin. Hence for the purposes of a statistical test on data that includes this characteristic, the two languages will be non-independent and thus it would be inappropriate to include both in a sample. However, if two languages are only remotely related, then even if they share a given characteristic, the phenomena of these two languages possessing that characteristic can be viewed as independent phenomena. Even if the shared characteristic is a common retention, the two languages can still be considered independent with respect to the characteristic, since after sufficient time, it will be largely a matter of chance that they still share the characteristic. In such cases, the causal factors contributing to such languages' possessing the characteristic include not only whatever caused them to have those characteristics in the first place, but also whatever caused each language to retain the characteristic. After sufficient time, the main reason that a language has a given characteristic is not whatever caused it to be that way in the first place, but whatever has caused it to remain that way. If one shuffles a deck of cards repeatedly, the configuration of the cards will eventually be independent of the original configuration in the sense that the difference between the two configurations will not be significantly different from the difference between a random pair of configurations. If a given pair of cards in the deck happen to remain adjacent after many shuffles, the fact that those two cards were adjacent in the original configuration and the fact that they are adjacent in the final configuration are independent phenomena, and the fact that the two configurations are identical in that respect is a matter of chance. What is not clear is just how remotely related two languages have to be in order to count as independent. Furthermore, the answer depends in part on the type of linguistic characteristic being investigated. Since word order characteristics change fairly easily, it is possible that languages like those from Nilo-Saharan in Perkins' sample can be considered independent.[1] However, morphological characteristics are more conservative, and a pair of languages that one might consider independent for the purposes of word order might not be properly treated as independent for the purposes of morphology. Exactly how remote a relationship is necessary for other types of linguistic characteristics is a matter for investigation.

Perkins' sample includes other languages that are more closely related. Perhaps most severe is the inclusion of six languages which are considered Mon-Khmer (in the widest sense of that term): Car (Nicobarese), Semai, Khasi, Khmer, Palaung, and Vietnamese. Unlike the three Nilo-Saharan languages, these six languages are very similar typologically (though to varying extents), and hence their inclusion within the same sample, especially six of them, is a problem.

A more severe problem with Perkins' sample, however, is the inclusion of multiple languages from well-defined linguistic areas. Four of the six Mon-Khmer languages noted in the preceding paragraph are spoken on mainland southeast Asia: Khmer, Palaung, Semai, and Vietnamese. In fact their typological similarity may be due at least as much to diffusion as to their common genetic origin, since non-Mon Khmer languages of this general area, like Thai, Karen, Miao-Yao and Chinese languages, are also similar in many respects. Other inclusions of multiple languages from well-defined linguistic areas in Perkins' sample include Haisla and Quileute from the Pacific Northwest of North America (as well as Kutenai just outside this area) (*cf.* Boas 1929, Jacobs 1954, Thompson and Kinkade to appear); Nahua, Zapotec, and Tarascan from Meso-America (*cf.* Campbell, Kaufman, and Smith-Stark 1986); and Ainu and Korean, both of which exhibit similarities to Japanese that are probably due either to diffusion or to common origin. Perkins did attempt to control for areal phenomena by not including two languages from the same culture area (of Kenny 1974), but it is clear that for the purposes of linguistic areas, the areal grid he uses is far too fine.

Now it might be thought that the problem with Perkins' sample is simply a problem in practice rather than a problem of principle, that the problem could be corrected by simply selecting a new set of 50 languages that are at best remotely related and that are all from different linguistic areas. It is not clear, however, that one **can** construct a sample with as many as 50 languages that meets this criterion. The major problem is the possibility of linguistic areas even larger than the ones mentioned here. If one is conservative in one's assumptions in this regard, it may not be possible to construct a sample with many more than ten languages. But it is difficult to achieve significant or convincing results with a sample of such a small size. Two of the primary purposes of this paper are to provide evidence that suggests the possibility of the existence of very large linguistic areas and to propose a method for testing hypotheses despite this problem.

Before addressing these tasks, however, I would first like to discuss an example from the literature that illustrates the need for care in drawing inferences from language samples. Nichols (1986) proposes a distinction between head-marking languages and dependent-marking languages that seems to be a fundamental typological parameter; head-marking languages are ones in which there is more often grammatical marking on heads that indicates properties of their dependents (such as pronominal affixes on verbs indicating subject and/or objects), while dependent-marking languages are ones in which there is more often grammatical marking on dependents, generally indicating the nature of the relationship between the dependent and the head (such as case marking). But Nichols also argues for a relationship between this typology and word order type. Namely, she claims (among other things) that the head-marking type is more common among verb-initial languages than among other word order types, and she uses the chi-square test to compute a level of statistical significance of less than .005 for this association. An examination of her sample, however, suggests that this apparently high level of statistical significance is simply an artifact of the fact that the languages in her sample are not independent, a condition for the appropriateness of the chi-square test. The thirteen verb-initial languages in her sample include four instances of pairs of languages from the same family, and an even more dramatic areal skewing. The following table illustrates the areal and genetic distribution of these thirteen languages:

North America:
- Pacific Northwest:
  - Salish: Shuswap, Squamish
  - Chinookan: Wishram
  - Wakashan: Nootka
- Oregon (both Sahaptian): Sahaptin, Nez Perce
- "Central" Canada: Cree
- California: Barbareño Chumash
- Guatemala (both Mayan): Sacapultec, Tzutujil

Pacific (both Polynesian within Austronesian): Hawaiian, Samoan
Southwest Asia: Arabic

Apart from the four instances of language families which are represented by two languages (Salish, Sahaptian, Mayan, and Austronesian), ten of the thirteen languages are spoken in North America (in the broad sense that includes Central America), and four of these are spoken in the Pacific Northwest; furthermore the Sahaptian languages of eastern Oregon are just outside this area, and speakers of these languages are known to have had extensive contact with speakers of Interior Salish languages (*cf.* Aoki 1975). However, since these two Sahaptian languages are not head-marking and thus go against the association Nichols argues for, their inclusion does not present a problem for the particular issue under discussion. But since the languages in this sample are clearly not genetically and areally independent, Nichols' claim of a relationship between head-marking type and verb-initial order remains undemonstrated. In fact there is reason to suspect that the supposed association between head-marking type and verb-initial order is an artifact of the areal skewing of the languages in Nichols' sample. First note that none of the three verb-initial languages in her sample from outside North America (Arabic, Hawaiian, and Samoan) are head-marking while eight of the ten verb-initial languages from North America are. Second, note the following distribution of languages in her sample:

	North America	Not North America
Head-Marking	15	2
Not Head-Marking	8	34

This table shows that among the languages in Nichols' sample, about two thirds of the languages from North America are head-marking while only two out of 36 languages from outside North America are. While it is not clear how representative this sample is, it strongly suggests that the head-marking type is considerably more common in North America than it is elsewhere in the world (as Nichols herself notes). Thus it would appear that the supposed association between head-marking type and verb-initial order is simply an artifact of the fact that most of the verb-initial languages in Nichols' sample are from North America and the fact that the head-marking type is considerably more common in North America than it is elsewhere in the world. Of course, it is possible that the association claimed by Nichols does exist; but her evidence fails to demonstrate it. This example illustrates two things. First, it shows how including languages that are not independent genetically and areally can lead to unsupported conclusions. Nor is this case unique; various examples could be cited from the literature where conclusions are reached, often with levels of statistical significance cited, which can be shown to be artifacts of the nonindependence of the languages in the sample.

A second point illustrated by this example is that it demonstrates the possibility of a linguistic area encompassing most of North America. If it is really the case that the head-marking type is common throughout North America but uncommon elsewhere, this may suggest that its multiple occurrence throughout this area is not an accident, but rather a reflection of diffusion or remote genetic relationship, or both. If so, then a sample of independent languages that would be needed to test the hypothesis that there is a relationship between head-marking type and verb-initial order should only contain one language from this entire area. And since it is not known what other characteristics may be shared throughout this area due to areal or genetic factors, this suggests that in general, language samples should contain only one language from this area. But if there exist language areas as large as North America, the possibility exists that there may exist language areas of comparable size elsewhere in the world. And in fact I will provide evidence below that suggests that much of Eurasia and much of Africa may form linguistic areas. But if so, then it is plausible that it may not be possible to construct samples containing more than ten independent languages. Nor is it clear whether the frequency of the head-marking type is limited to North America. It may extend into South America as well; Nichols' sample includes only two languages from South America, both Quechuan, and neither of these is head-marking. Evidence for a possible linguistic area subsuming much of the New World is discussed in sect. 4.3 below.

Two points should be emphasized at this point. First, by *linguistic area* I mean something rather different from what is often intended by the term. Generally, the term is used to refer to an area in which a large number of typological characteristics have diffused among languages which are genetically unrelated or at best remotely related. However, by *linguistic area*, I intend an area in which at least one linguistic property is shared more often than elsewhere in the world to an extent which is unlikely to be due to chance, but which is probably due either to contact or remote genetic relationships. In other words, the number of typological characteristics shared may not be enough to satisfy the normal notion of linguistic area; all I demand is that at least one property be shared to an extent that is likely to be due to areal or genetic factors. Furthermore, I remain uncommitted to what extent the existence of properties in a large area is likely to be due to diffusion as opposed to genetic relationship; hence by *linguistic area* I do not preclude the possibility that the underlying cause is partly or largely genetic.

A second point to be emphasized is that I do not claim to be presenting evidence that **proves** that the areas in question are linguistic areas in the sense used here, but rather evidence that suggests that they **may** be linguistic areas. It should be emphasized that the conservative assumption for language sampling is that languages are related, either genetically or areally. The reason that this is the conservative assumption is that if we overestimate the degree to which languages are non-independent, *i.e.* if we assume languages are not independent when in fact they are, the result is that we may fail to achieve statistically significant results demonstrating some conclusion, and we may fail to conclude something which in fact we could conclude. On the other hand, if we underestimate the extent to which languages are related (genetically or areally), then we may reach conclusions that are in fact unsupported if not false. I assume the latter danger is the more serious one, and examples like the hypothesis of Nichols' discussed above suggest that it may not be uncommon.

3. The method

The method to be proposed here for testing hypotheses assumes the possibility of very large linguistic areas. Since the evidence that suggests the existence of these large areas is best presented in terms of the method to be proposed, I will discuss the method first and then turn to the evidence for such areas.

I will illustrate the method with data from a cross-linguistic study of word order universals, from an overall sample of 542 languages (*cf.* Dryer 1986, 1987, 1988a, 1988b).[2] Clearly, for the reasons discussed already in this paper, one cannot simply count languages in this overall sample. Rather, what is done is as follows. First, the languages are grouped into genetic groups roughly comparable to the subfamilies of Indo-European, like Germanic and Romance. I refer to each of these groups as a *genus* (following a suggestion by Bill Croft), since they are rather analogous to the taxonomic level of *genus* in biology. In some areas of the world, these genera are the maximal level of grouping whose genetic relationship is uncontroversial. By counting genera rather than languages, we control for the most severe genetic bias, since languages within genera are generally fairly similar typologically. Although languages in separate but related genera are often similar as well, this similarity may often be as much due to diffusion as to genetic relationship. It must be borne in mind that diffusion

not only results in unrelated languages acquiring new characteristics, but also reinforces typological similarity between related languages. Although the latter effect of diffusion is less often discussed (since its effect is less obvious), I suspect it is the major role that language contact plays.

The second step in the method is to divide the genera into five large continental areas: Africa, Eurasia, Australia-New Guinea, North America, and South America. The divisions between these five areas are rather well-defined physically, though sometimes I allow genetic groupings to define the actual boundary. Thus I treat the Semitic languages as part of Africa, since their genetic relationships go in that direction. For similar reasons, I treat the division between North and South America as roughly in Honduras, treating the Chibchan languages of Central America with South America. I treat the Austronesian languages as part of Eurasia, since they exhibit some typological similarity to the languages of southeast Asia, possibly due to remote genetic relationship (*cf.* Benedict 1975). This includes the Austronesian languages of New Guinea, although a number of them have clearly been influenced by non-Austronesian New Guinea languages.

The only assumption I make about independence is that these five areas are independent of each other. Although there are possible genetic and areal relationships across these continental boundaries (*e.g.* the case of Austronesian New Guinea languages just noted, and Chukchi-Kamchatkan with Eskimo-Aleut), I assume that their effect is sufficiently small that it can be ignored. The most serious question regarding the independence of the five areas surrounds North and South America: Greenberg (1987) argues that most of these languages form a single Amerind language family. While his claim is controversial, one of the objections that has been raised is that some of the lexical resemblances he notes might be due to borrowing. However, if there are lexical resemblances between North and South America that are due to borrowing, then there could just as easily be typological similarities due to diffusion, in which case the two areas would not be independent. Thus, even if one does not accept his evidence as sufficient for demonstrating a genetic relationship, it does suggest that North and South America may not be sufficiently independent for the purposes of the sampling methodology of this paper. In fact, I will present evidence below that suggests that the New World may constitute a single linguistic area. Despite this, I will treat North and South America as separate areas. One reason for this is that on the whole, each of North and South America seem to exhibit as much typological diversity and genetic diversity (in terms

of the number of genera) as the other areas. A couple of the other areas could be split up into two smaller areas, for example Australia separate from New Guinea, or southeast Asia (including Austronesia) from the rest of Eurasia, but the areas that would result would exhibit considerably less typological diversity. But the particular choice of areas remains tentative.

The total number of genera represented in the sample from each of the five areas is as follows:

Afr	Eura	A-NG	NAm	SAm	Total
45	52	30	60	31	218

(Afr=Africa, Eura=Eurasia, A-NG=Australia-New Guinea, NAm=North America, SAm= South America.) A weakness of the methodology is that it depends crucially on assumptions as to what genetic groupings constitute genera. My conclusions in this regard have been very impressionistic and tentative. Unfortunately, it is very difficult to determine from the literature just which groups constitute genera. A list of the genera assumed for the languages in my sample is given in an appendix. My estimate of the total number of genera in each of these five areas, including ones not represented in my sample, is as follows:

Afr	Eura	A-NG	NAm	SAm	Total
59	56	80	70	57	322

Bell (1978) gives estimates for the number of genetic groups separated by 3500 years, analogous to my genera, and his estimates are notably higher then mine; he estimates, for example, that there are 478 such groups in the entire world, considerably higher than my estimate of 322 genera.

The third step in the method is to count the number of genera of each of the relevant linguistic types being investigated within each continental area. The final step is to determine how many of the five areas conform to the hypothesis being tested. If all five conform, then the hypothesis is considered to be confirmed. Let me illustrate the method by examining the question of whether there is a linguistic preference for SOV order over other orders. Since the only other common order is SVO, I will compare SOV with SVO. The relevant data is as follows:

	Afr	Eura	A-NG	NAm	SAm	Total
SOV	22	26	19	26	18	111
SVO	21	19	6	6	5	57

The numbers indicate, within each area given across the top, the number of genera containing languages of the type given on the left-hand side. For example, the leftmost column indicates that there are 22 genera in Africa containing SOV languages, and 21 genera containing SVO languages.[3] In each case, I have enclosed the larger of the two figures within each area in a box. It can be seen that within each of the five areas, the number of genera containing SOV languages is greater than the number of genera containing SVO languages. The greater number of SOV genera in Africa is marginal, but in the other four areas, there are clearly more SOV genera than SVO genera. Since SOV exceeds SVO in all five areas, the hypothesis that SOV is preferred over SVO is confirmed. The logic of the statistical test is very simple: it is a binominal sign test, and is analogous to flipping a coin five times. The chance of flipping an unbiased coin five times and getting five heads is one in thirty-two. If there were not a linguistic preference for SOV order over SVO order, then the chance of all five areas containing more SOV genera would also be one in thirty-two, if we assume that the five areas are genetically and areally independent. The hypothesis is thus confirmed at a level of statistical significance less than .05. Note that no assumption need be made about the independence of the languages within each area. The margin in Africa is sufficiently small that we can say that the conclusion barely achieves statistical significance. Clearly, a few more languages in the sample might tip the margin the other way, but such is always true with statistical tests. A potentially more serious problem is that, as noted above, conclusions depend on assumptions as to what genetic groupings constitute genera.

It is worth emphasizing that the conclusion that there is a linguistic preference for SOV order over SVO is at odds with what Tomlin's data discussed in sect. 2 might suggest. His evidence shows that SVO order is almost as common as SOV among the languages of the world. But as noted there, about 40% of the SVO languages of the world are apparently Niger-Congo, and there are also a large number of SVO Austronesian languages. Hence the number of SVO languages in the world is inflated by a couple of large families, thereby obscuring the linguistic preference for SOV order.

Tomlin's data is sect. 2 would suggest that SVO is preferred over VSO, especially in light of the fact that about 71% of the VSO languages in his data are in one family, Austronesian. It is not the case, however, that there are more SVO genera than VSO genera in each of the five areas:

	Afr	Eura	A-NG	NAm	SAm	Total
SVO	[21]	[19]	[6]	6	[5]	57
VSO	5	3	0	[12]	2	22

As this table shows, there are more SVO genera than VSO genera in four areas, but in North America, the number of VSO genera is higher. Thus, by the test being assumed here, the preference for SVO order over VSO order falls short of statistical significance. Nevertheless, since four of the five areas conform, we can say that there is a *trend* in favour of SVO over VSO. Since the test we are using is rather conservative, many such trends probably do indicate real linguistic preferences. The evidence for such a trend here is further bolstered by the fact that in the four areas in which there are more SVO genera, the number of SVO genera is at least twice the number of VSO genera.

Many of the claims of word order typology (*cf.* Greenberg 1963a) involve relationships between word order parameters, such as the claim that OV languages tend to be postpositional. We can test such a claim by comparing the number of genera containing languages that are OV and postpositional with the number of genera containing languages that are OV and prepositional:

	Afr	Eura	A-NG	NAm	SAm	Total
OV&Po	[13]	[27]	[15]	[20]	[12]	87
OV&Pr	2	2	1	0	0	5

The data clearly support the claim. Not only is the number of genera containing OV&Po languages higher in each of the five areas, but it is considerably higher, and there are relatively few exceptions. This constitutes, therefore, a particularly strong statistical universal. The following data show that this association goes both ways: VO languages also exhibit a strong tendency to be prepositional:

	Afr	Eura	A-NG	NAm	SAm	Total
VO&Pr	[14]	[23]	[5]	[15]	[5]	62
VO&Po	4	1	0	2	2	9

The method proposed in this paper is useful for testing implicational universals in general. Consider the universal "If a language places the demonstrative after the noun, then it will place the adjective after the noun as well." This is part of Greenberg's Universal 18 and is Universal V' of

Hawkins' (1983). This universal is exceptionless in Hawkins' sample, but there are six exceptions in my sample, falling into four genera, two of them Tibeto-Burman:

Ubangian: Gbaya, Sango, Nzakara
Burmic: Lahu
Tibetic: Dafla
Tsimshian: Coast Tsimshian

However, although there exist exceptions to it, it does constitute a strong statistical universal:

	Afr	Eura	A-NG	NAm	SAm	Total
NDem&NAdj	[28]	[14]	[8]	[8]	[5]	63
NDem&AdjN	1	2	0	1	0	4

We see that in all five areas, there are clearly more NDem&NAdj genra than there are NDem&AdjN genera, so that we can conclude that there is a linguistic preference for NDem languages to be NAdj. The correlation goes only in one direction, however. Among languages in which the demonstrative precedes the noun, both orders of adjective and noun are common:

	Afr	Eura	A-NG	NAm	SAm	Total
DemN&AdjN	5	[27]	[7]	[19]	[8]	66
DemN&NAdj	[7]	10	6	14	6	43

There is at most a weak trend favouring AdjN order among languages in which the demonstrative precedes the noun. Nevertheless we can still say that there is clear evidence of a correlation or association between the order of demonstrative and noun and the order of adjective and noun. This can be shown more clearly by comparing the **proportion** of DemN genera that are AdjN (as opposed to NAdj) with the proportion of NDem genera that are AdjN:[4]

	Afr	Eura	A-NG	NAm	SAm
DemN	[.42]	[.73]	[.54]	[.58]	[.57]
NDem	.03	.09	.00	.11	.00

Proportion of AdjN among DemN/NDem languages

The same method provides possible evidence for an association or correlation between the order of object and verb and the order of noun and article, although the following count data do not, at first sight, seem to provide much basis for saying that there is an association:

	Afr	Eura	A-NG	NAm	SAm	Total
OV&ArtN	0	2	2	3	[3]	10
OV&NArt	[3]	[4]	2	3	0	12
VO&ArtN	2	[11]	[1]	[10]	[2]	26
VO&NArt	[7]	2	0	1	0	10

In OV languages, the two orders of article and noun are about equally common. Among VO languages, ArtN order is more common in four areas, but in two of these areas (Australia-New Guinea and South America), the difference is marginal since there are few genera with VO languages in my sample from these areas that employ articles. Furthermore, in one area (Africa), there are clearly more VO&NArt genera. However, if we compare the **proportion** of OV genera that are ArtN (as opposed to NArt) with the proportion of VO genera that are ArtN, a somewhat clearer pattern emerges:

	Afr	Eura	A-NG	NAm	SAm
OV	.00	.33	.50	.50	1.00
VO	[.22]	[.85]	[1.00]	[.91]	1.00

Proportion of ArtN among OV/VO languages

In four areas, the proportion of ArtN is higher among VO languages than it is among OV languages. In the remaining area (South America), the proportions are the same. Because the proportion is not higher in all five areas, we fall short of statistical significance. Nevertheless, since the proportions are the same in South America, we come close to statistical significance, and have clear evidence of a trend. Thus the evidence is likely indicative of an association between the order of object and verb and the order of noun and article, such that articles tend to precede the noun more often in VO languages than they do in in OV languages.

4. Evidence for large areas

4.1. *OV and Adjective-Noun order in Eurasia*

It is widely thought that there is a linguistic preference for OV languages to place modifying adjectives before the noun. But the following data do not support such an hypothesis:

	Afr	Eura	A-NG	NAm	SAm	Total
OV&AdjN	6	[22]	5	9	6	48
OV&NAdj	[17]	9	[15]	[17]	[10]	68

Only in Eurasia are there more SOV&AdjN genera than SOV&NAdj genera. In the other four areas it is more common for SOV languages to place the adjective after the noun; in fact outside Eurasia, there are 59 OV&NAdj genera but only 26 OV&AdjN genera. Thus if anything there is a trend in the opposite direction from that generally thought. The widely held belief that there is a preference for SOV languages to place the adjective before the noun seems to be due to the fact that this is the dominant pattern in Eurasia. Not only is SOV&AdjN more common in Eurasia, but the exceptions tend to be geographically peripheral (*cf.* Dryer 1988b). The nine SOV&NAdj genera in Eurasia are the following: Basque, spoken in western Europe; Sumerian, an ancient language of the Middle East; Iranian, in which Persian (Farsi) is NAdj while the other Iranian languages in my sample are AdjN (the NAdj order of Persian may be due to contact with Arabic); Northwest Caucasian; three subgroups of Tibeto-Burman, which extends into southeast Asia, an area in which the dominant order is SVO&NAdj; Andamanese, off the coast of Burma; and a subgroup of Austronesian including OV languages spoken on New Guinea. But in the core of Eurasia, in an area that extends from Turkey to Japan, and from south India to northern Russia to Siberia, almost all of the SOV languages place the adjective before the noun (*cf.* also Masica 1976). Since there is a clear trend outside this area for SOV languages to place the adjective after the noun, this core of Eurasia forms a linguistic area in the sense in which that term is being used here: it is an area in which a particular linguistic characteristic (SOV&AdjN) is widespread (in fact overwhelmingly dominant), while the same linguistic characteristic is significantly less common outside that area. It is unlikely that it is a coincidence that most of the SOV languages of this area are AdjN. Rather it seems likely to reflect either

remote genetic relationships or diffusion, or a combination. It is possible, of course, that it is partly coincidental, and that there are in fact two areas, such as the Indian subcontinent and north-central Eurasia (corresponding roughly to the Soviet Union, but extending south into Japan, Iran, and Turkey), and that it is a coincidence that these two areas share this characteristic. And although it would be valuable to investigate whether there are other characteristics shared within this area, the evidence from the SOV&AdjN languages of this area is sufficient to raise the possibility of a single linguistic area here and thus to cast doubt on the results of any study that uses a sample containing more than one language from this area. Since the method proposed here does not assume independence of any groups within Eurasia, the possibility of this large area does not pose a problem for the results presented here, or other results that employ the method proposed here.

It is worth noting that there does seem to be further evidence of characteristics shared throughout much of this area. First, not only are most of the OV languages of this core area of Eurasia AdjN, but the languages of this area are entirely SOV. Second, Crothers (1976) claims that front round vowels are uncommon outside a large part of Eurasia, all but southeast and southwest Asia and the Indian subcontinent. Finally, although the order of relative clause and noun correlates with that of adjective and noun, it is worth noting that OV&RelN order is also considerably more common in Eurasia than it is elsewhere in the world:

	Afr	Eura	A-NG	NAm	SAm	Total
OV&RelN	4	[12]	2	1	1	20
OV&NRel	[7]	5	[5]	[12]	[2]	31

It is often thought that there is preference for SOV languages to place the relative clause before the noun. But again, this preference is only found among the languages of Eurasia. Rather strikingly, however, the area in Eurasia in which SOV&RelN languages are found is **not** identical with the area in which SOV&AdjN languages are found. A number of the SOV&RelN languages are ones which are SOV&NAdj: Basque, Abkhaz (in Northwest Caucasian), and all of the Tibeto-Burman languages of my sample for which I have the relevant data. Although this might be partly due to coincidence, the conservative assumption is again that there may be an area within Eurasia that includes these languages as well, in short an area that includes all but southeast Asia and the Middle East.

4.2. *Noun-Numeral order among VO languages in Africa*

The order of numeral and noun in VO languages provides evidence that suggests the existence of a linguistic area covering much of Africa. The relevant data is given in the following table:

	Afr	Eura	A-NG	NAm	SAm	Total
VO&NumN	2	[19]	[4]	[14]	[4]	43
VO&NNum	[19]	5	0	0	0	24

In Africa, there is an overwhelming tendency, by 19 genera to 2, for VO languages to place the numeral **after** the noun. Outside of Africa, there is an almost equally strong tendency, by 41 genera to 5, for VO languages to place the numeral **before** the noun. In fact, all VO languages in my sample in Australia-New Guinea and the New World place the numeral before the noun. The extent to which the VO languages of Africa behave differently from VO languages elsewhere in the world is striking, and provides evidence that is clearly suggestive again of a large linguistic area. Significantly, the two genera in Africa which fail to conform to this pattern are Berber and Semitic, the two most northerly genera in Africa. All of the VO languages in my sample south of these two genera place the numeral after the noun. And they are widespread both genetically and areally. All four of Greenberg's families are represented: 1 genus is in the Khoisan family, 11 are Niger-Kordofanian, 5 are Nilo-Saharan, and 1 (Chadic) is Afro-Asiatic. The Khoisan group and Chadic are both likely candidates for influence by Niger-Congo languages, but the fact that the Nilo-Saharan groups exhibit the same pattern is more surprising. (See Dryer (in preparation) for details.)

4.3. *Possessive prefixes in the New World*

The third case suggestive of a large linguistic area is less convincing. In fact, if right, it would present a problem for the way in which hypotheses have been tested in this paper, since it suggests that North and South America are not independent. Nevertheless, it is incompatible with the view that there are a large number of linguistic areas in the New World, and renders suspect any conclusion reached on the basis of a sample containing too many languages from the New World.

In many languages, nouns can be inflected with possessive affixes indicating the person and/or number of the possessor. While both prefixes and suffixes are common for this category (in contrast to many other affix categories for which suffixes are more common), the New World exhibits a different pattern from the Old World:

	Afr	Eura	A-NG	NAm	SAm	Total
PossSuff	[16]	[12]	10	12	3	53
PossPref	4	8	10	[33]	[12]	67

While possessive suffixes are as common as prefixes in Australia-New Guinea and more common in Africa and Eurasia, possessive prefixes are clearly more common in both North and South America. In fact approximately half of the genera with possessive prefixes are in North America. 38 out of 60 genera with possessive affixes in the Old World employ suffixes while only 15 out of 60 genera in the New World do. It is also worth noting that 8 of the 12 genera in North America which employ possessive suffixes fall into two well-defined linguistic areas, the Pacific Northwest and Meso-America. Outside these two areas but within North America, 27 out of 31 genera use prefixes rather than suffixes. This is in striking contrast to Africa, where 16 out of 20 genera use suffixes. The extent to which patterns like this might be due to chance should not be underestimated. Nevertheless it is not unreasonable to suppose that they may reflect remote areal or genetic factors.

5. A detailed examination of one example

It should be stressed that I am not claiming that samples containing more than one language from a given continental area necessarily contain non-independent languages. For one thing, there are undoubtedly many instances in which there are pairs of languages within a single area that are essentially indepenent for sampling purposes. Furthermore, if a given linguistic characteristic is found sufficiently sporadically within an area, in a number of small areas far apart, then these areas can probably be considered independent with respect to that particular phenomenon. Of course even in such cases it is possible that two areas exhibiting a phenomenon at great distances are instances of a common retention, but we can be more confident of the independence of two such areas if they are separated by languages lacking the characteristic in question than if they are at opposite ends of a large area over which most languages exhibit the characteristic.

Thus if a given characteristic were found in Eurasia in Basque, Chukchee, Dravidian and Polynesian, but not elsewhere, then we would probably be justified in assuming that these four instances are independent phenomena. In general, the suitability of a particular sample for testing a particular hypothesis will depend on the distribution of the particular linguistic characteristics relevant to the hypothesis.

The following example illustrates how one might go beyond the five areas assumed in the methodology described in this paper in attempting to assess how many independent instances of a given phenomenon we have. The hypothesis to be tested is Greenberg's (1963a) Universal 5: "If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun." This universal is exceptionless, not only in Greenberg's sample, but also in the expanded sample of Hawkins (1983) and in my sample. This universal is logically equivalent to Hawkins' Universal I. The hypothesis I wish to discuss is not whether this universal is exceptionless but simply whether it is a statistically significant **statistical** universal. In other words, is there evidence of a linguistic preference among languages which are SOV and NGen to be NAdj? If there is not, then the question of whether this universal is exceptionless is moot. The following table gives the data from my sample for the four possible orders of noun and genitive and noun and adjective among SOV languages:

	Afr	Eura	A-NG	NAm	SAm	Total
SOV&GenN&AdjN	6	[17]	4	9	5	41
SOV&GenN&NAdj	[9]	7	[8]	[15]	[8]	47
SOV&NGen&AdjN	0	0	0	0	0	0
SOV&NGen&NAdj	4	3	2	0	0	9

Apart from the lack of SOV&NGen&AdjN languages, perhaps the most obvious observation here is that the first two types, those in which the genitive precedes the noun, are considerably more common than either of the latter two types, in which the genitive follows the noun. In all five areas, the number of genera containing languages of each of the first two types is greater than the number of genera containing languages of each of the last two types. This reflects, of course, the fact that SOV languages tend to be GenN. As a result, the type of language which Greenberg's Universal 5 refers to, SOV&NGen, is itself an infrequent type. But this makes it difficult to test hypotheses about SOV&NGen languages. In fact, by the methodology proposed in this paper, the evidence for a preference among SOV&NGen languages to be NAdj rather than AdjN falls short of statistical significance, since in only three areas is SOV&NGen&NAdj more common than SOV&NGen&AdjN, there being no languages of either sort in North or South America in my sample.

By assuming independence between these continental areas, but non-independence within these areas, we essentially assume three independent instances of SOV&NGen order. But it is worth examining the actual instances of SOV&NGen order to see whether this assumption is overly strict, whether we might not safely conclude that there are more than three independent instances of this order. The languages of this sort in my sample, divided by genus and area, are as follows:

AFRICA
- KORDOFANIAN PROPER: Rashad
- SAHARAN: Tubu
- EASTERN CUSHITIC: Geleba, Oromo
- SOUTHERN CUSHITIC: Iraqw

EURASIA
- IRANIAN: Persian
- Sumerian
- CENTRAL-EASTERN MALAYO-POLYNESIAN: Manam

AUSTRALIA-NEW GUINEA
- GUNWINYGUAN: Dalabon
- PAMA-NYUNGAN: Guugu Yimidhirr, Alyawarra

Could one argue, for any of these three areas, that they contain more than one independent occurrence of SOV&NGen order? In Eurasia, the answer is clearly yes. The occurrence of this order in Manam, a New Guinea Austronesian language, is clearly independent of its occurrence in Persian and Sumerian. Furthermore, the apparent explanation for this order in Manam is that it has lost its original Austronesian VO order and has borrowed the OV order from Papuan languages, while retaining the NGen order from Austronesian. On the other hand, the occurrences of this order in both Persian and Sumerian, an ancient language of the Middle East, are more plausibly non-independent. Greenberg (1963a) furthermore lists Elamite, another ancient language of the area, as well as the North Semitic language Akkadian, as SOV&NGen&NAdj. Hence this seems to

be common in this area. On the other hand, Hawkins (1983) lists Old Persian as primarily GenN and AdjN. This suggests that the NGen&NAdj order in Modern Persian is a recent development, possibly due to influence by Arabic. If so, then, despite the geographical proximity, Sumerian and Persian might constitute independent instances of the phenomenon. Nevertheless, since both are spoken in an area which is solidly NGen&NAdj, the conservative assumption is that they are not independent.

Turning to Australia-New Guinea, there are two genera containing SOV&NGen&NAdj languages, both in Australia. Word order is very flexible in many Australian languages, even between nouns and their modifiers, and there is considerable variation among related languages as to which order is unmarked. For many languages both orders of noun and modifier are common for one or both modifiers; I leave these languages unclassified with respect to the characteristics in question. Among those languages for which a dominant order can be assigned for these characteristics, there is a general tendency for the Pama-Nyungan languages (which cover most of the area of Australia) to be GenN&NAdj and for the Northern (non-Pama Nyungan) languages to be either GenN&NAdj or GenN&AdjN. The Gunwinyguan language Dalabon is spoken well into the area of Northern languages, and is separated from Pama-Nyungan languages by a number of other Northern languages. However, all of the languages in my sample between Dalabon and Alyawarra, one of the Pama-Nyungan SOV&NAdj&NGen languages in my sample, are NAdj. This, plus the fact that Dalabon is remote ly related to the Pama-Nyungan languages, renders it at least plausible that its NGen&NAdj order is not independent of the occurrence of this order in Pama-Nyungan. The two Pama-Nyungan SOV&NGen&NAdj languages in my sample are geographically separated: Alyawarra is an Arandic language spoken in the southern part of the Northern Territory; Guungu Yimidhirr is a Maric language spoken in east Queensland. Since both fall within the same genus, they are plausibly non-independent. It is conceivable that we have as many as three independent occurrences of SOV&NAdj&NGen in Australia, but again the conservative assumption is that we have only one.

The four SOV&NGen genera in Africa fall into three different genetic stocks: Kordofanian belongs to Niger-Kordofanian, Saharan to Nilo-Saharan, and the two Cushitic genera to Afro-Asiatic. Although other orders of noun and genitive and noun and adjective are found within Cushitic, I

assume that these two branches are sufficiently close as to constitute instances of the same phenomenon. But one might plausibly argue that we nevertheless have three independent phenonema, corresponding to the three genetic stocks. I have argued in sect. 4.2, however, that the high incidence of noun-numeral order among the VO languages of Africa provides evidence that much of Africa, excluding the Semitic and Berber languages in the north, constitutes a single linguistic area. That in itself does not show, of course, that the three instances of SOV&NGen order are non-independent. At most it suggest they might be non-independent. The fact that the hypothesis under discussion refers to SOV languages, while the evidence for a large linguistic area came from VO languages, would suggest that the evidence from VO languages may not be relevant. At the very least such evidence provides no reason to believe that the areas of Africa in which OV languages are spoken are included in this large linguistic area. Furthermore the SOV&NGen languages tend to be somewhat peripheral to the area in which VO&NNum is common. Although the Southern Cushitic languages are spoken in northern Tanzania in an area adjacent to Nilotic and Bantu languages which do exhibit VO&NNum order, the Central Cushitic languages are spoken in Ethiopia outside the VO&NNum area. Rashad, an SOV&NGen Kordofanian language, is spoken in a linguistically complex area in the central Sudan in which other Kordofanian languages and a variety of East Sudanic Nilo-Saharan languages are spoken. Since there are both Kordofanian and East Sudanic VO&NNum languages in this area, it is fair to say that Rashad falls within the VO&NNum area. The Saharan languages, spoken in Chad, are clearly peripheral to the area of VO&NNum languages, though they border the Chadic languages, which are in this area. It is thus plausible to argue that the evidence for a VO&NNum area in Africa is consistent with the claim that we have three independent instances of SOV&NGen in Africa.

On the other hand, the Kordofanian language Rashad is spoken only 300 miles from the Central Cushitic language Oromo. Furthermore, the languages spoken between are VO&NGen&NAdj. Hence Oromo and Rashad both fall within a solid NGen&NAdj area. Similarly, the Saharan languages, though peripheral to the VO&NNum area, are surrounded primarily by Berber, Semitic, and Chadic languages, all of which are NGen&NAdj as well. They thus belong to a large NGen&NAdj area that includes Semitic and Berber languages to the north and west and extends down through Chadic and the Niger-Congo languages of Nigeria down into

Bantu. Of course, since NGen&NAdj is a frequent type in other areas of the world, we do not have a well-defined area, in the way we do for VO&NNum; the argument in the latter case depends on the infrequency of VO&NNum order elsewhere in the world. Nevertheless it is worth noting that about half of the NAdj&NGen genera in my sample — 21 out of 43 — are in Africa.

But particularly crucial is the fact that the high frequency of NNum order among VO languages is simply the clearest manifestation of a general pan-African tendency to place modifiers after nouns more often than elsewhere in the world (*cf.* Dryer in preparation). This tendency varies from modifier category to modifier category and is weaker among OV languages, but the fact that almost half the SOV&NGen genera — 4 out of 9 — are in Africa is just another manifestation of this tendency. It is significant that the linguistic type in question, SOV&NGen&NAdj, involves both modifiers following the noun. For these reasons, I remain unconvinced that the different occurrences of SOV&NGen&NAdj order in Africa are really independent. It is possible that further evidence from these languages, especially diachronic evidence, might provide a basis for believing that these three instances of SOV&NGen&NAdj order are independent phenomena, or even that two of them are independent. But in the meantime the conservative assumption is that they are not.

It is also possible that the occurrences of SOV&NGen&NAdj order in Persian and Sumerian are not independent of its occurrence in Africa. Two of the four genera in Africa, the Cushitic groups, are Afro-Asiatic, while a third one, Saharan, is almost surrounded by Afro-Asiatic languages. But both Persian and Sumerian are languages with contact with Semitic Afro-Asiatic languages, and as noted above, Greenberg (1963) lists the Semitic language Akkadian as an instance of SOV&NGen&NAdj as well. Hence the conservative assumption is that the occurrence of SOV&NGen&NAdj order in Persian and Sumerian is not independent of its occurrence in Africa. The basic reason for this is that all are spoken in an area in which NGen&NAdj order is more common than elsewhere in the world.

We are left then with only three clearly independent instances of SOV&NGen&NAdj: Africa-Southwest Asia, Manam (in New Guinea), and Australia. If there is no preference for SOV&NGen languages to be NAdj, then there would be one chance in eight of three such groups being NAdj. Since this falls short of conventional levels of statistical significance, we must conclude that there is not convincing evidence for such a preference. Nevertheless, this is true only under the most conservative assumptions. It is quite possible that we do in fact have more than one independent occurrence of SOV&NGen&NAdj either in Africa-Southwest Asia or in Australia. Thus we can certainly say that there is a trend that suggests that there may be a preference for SOV&NGen languages to be NAdj.

There are a number of general conclusions to be made on the basis of this example. First, the fact that the universal is exceptionless even in a large sample of languages does not mean that it constitutes a significant generalization about language, since it is not clear whether the universal is a significant **statistical** universal. The fact that all of the SOV&NGen languages in this sample are NAdj may simply be due to a combination of areal factors and chance.[5] Conversely, the fact that the generalization is supported in only three continental areas does not mean that there are only three independent instances of it. The minimal three instances of it do not correspond to the three continental areas assumed in this paper. First, there are at least two instances of it in Eurasia. And second, the instances of it in southwest Asia may not be independent of Africa. Furthermore, it is quite possible that there are multiple independent instances in Australia, among languages that are remotely genetically related, and even possibly within a single genus, Pama-Nyungan. Close attention to the distribution of linguistic characteristics relevant to a particular hypothesis may provide insights that the general methodology of this paper fails to.

6. Conclusion

I have argued in this paper that there may exist large linguistic areas that are continental or almost continental in size and that apparently statistically significant results based on samples that include many languages from each continent may simply reflect areal phenomena rather than linguistic preferences. This problem is well-known in cross-cultural studies in anthropology, where it is known as Galton's Problem. Naroll (1970) discusses a variety of solutions and the method proposed in this paper might be viewed as a possible general solution to Galton's Problem. It is not clear to what extent the nature of the problem may be different for language sampling in contrast to culture sampling. It seems that in general linguistic characteristics are not borrowed as easily as cultural ones. Driver and Chaney (1979) note the example of the Yurok, Karok, and Hupa, three tribes in California whose languages belong to different stocks — Algic,

Hokan, and Nadene respectively — but whose cultures are almost identical. Although there may be some linguistic borrowing among these three languages, the extent is clearly less than that of cultural borrowing. At first sight, one might think that this implies that Galton's Problem is less severe for linguistic sampling: if linguistic borrowing is less common than cultural borrowing, then (it might be thought) diffusion should present less of a problem for language sampling. Two points must be stressed, however. First, the similarity among languages over large areas may be due partly to remote genetic relationships rather than diffusion. And second, the fact that language characteristics are less easily borrowed partly reflects the conservativeness of language characteristics. Thus languages which are separated by great geographic distances are more likely to share characteristics due to historical factors — either genetic or diffusional — than cultures separated by the same distance. As a result, it may be more difficult to construct reasonably-sized samples of independent languages than of independent cultures.

An obvious question to ask, if there can exist linguistic areas as large as continents, is whether the entire world might not constitute a single large linguistic area. How do we know that certain languages types are more common throughout the world, not because of any truly linguistic preference but simply because the entire world forms a single linguistic area and certain language types are more common in all five areas simply because of remote genetic and areal factors? In fact, in general we have no way of knowing that such is not the case. The kind of evidence presented here that suggests the existence of large linguistic areas depends on the possibility of showing that one continental area patterns differently from the rest of the world. But clearly there is no other area to which we can compare the entire world. And it is quite conceivable, for example, that the evidence cited in sect. 3 for a linguistic preference for SOV order over other orders is simply due to the residue of an original SOV word order. In fact it is conceivable that the lower frequency of VSO order is simply due to the fact that it takes a greater change to get from SOV order to VSO order than it takes to get from SOV order to SVO. One might argue that we know that basic clause order changes fairly easily, so that it seems more plausible that word order has changed often enough that any original order could not have survived enough to still be most frequent. However, there are large areas in which the languages are overwhelmingly SOV and in which any genetic factors accounting for the distribution of word order would have to

assume that they have retained SOV order for a period of at least 10000 years. Two such areas are the core area of Eurasia extending from Turkey to Japan and from South India to northern Russia to western Siberia, and New Guinea. It seems more likely that diffusion has played a major role in these areas, especially given the evidence for diffusion in India, the nomadic cultures of northern Asia, and the apparent extent of intertribal interaction in New Guinea prior to contact with European culture. But it is still possible that the evidence for a linguistic preference for SOV order reflects no more than the combination of an original SOV order combined with extensive diffusion which has served to reinforce the high frequency of SOV order. It must be borne in mind, again, that contact probably reinforces existing typological characteristics more often than it causes languages to acquire new characteristics.

On the other hand, **associations** or correlations between typological parameters are more difficult to explain by appealing to the idea that the entire world might be a single linguistic area. The fact that OV order correlates with postpositions while VO order correlates with prepositions seems immune to explanation in terms of a single world-wide linguistic area. For this reason, one might argue that hypotheses about linguistic preferences for one type over another, such as SOV over SVO, are untestable, and that we should restrict attention to hypotheses about the relationship between two or more parameters.

I have argued in this paper that the sampling methods employed in many recent cross-linguistic studies underestimate the possibility of large-scale areal phenomena, and that the reliability of their conclusions is suspect, since they may be the result of non-independent languages in the sample. It might be felt that the evidence I have given for large linguistic areas is less than entirely convincing. The patterns observed may be partly due to coincidence. After all, if one examines enough typological parameters one is bound to eventually find some property shared by any arbitrary set of languages more than other languages in the world. The fact remains, however, that unless the possibility of large areas is considered, we may reach conclusions through language sampling which are in fact false. And even if I have overestimated the extent of the problem, the methodology proposed here represents an excellent way to test hypotheses regardless of whether there are large areas or not. In other words, more conventional methods of constructing samples with thirty or more allegedly independent languages depend on the nonexistence of large areas. The method proposed here does not depend on the existence or nonexistence of such areas.

7. Appendix

This appendix contains a listing of the 218 genera into which the 542 languages in the sample have been grouped. As noted in the main text, this grouping is highly tentative and based on meagre initial impressions. I would appreciate information leading to improvements in this list. The groups are intended to be comparable to the subfamilies of Indo-European, with a time depth of separation in the area of about 3500 to 4000 years. There are two possible errors in deciding which genetic groups are genera. On the one hand, if two groups of languages are classified here as belonging to the same genus when they are more remotely related than subfamilies of Indo-European, then the genus in question ought to be split up. A special case of this is instances of languages or groups of languages whose genetic relatedness is controversial. It is generally the case that languages whose genetic relatedness is controversial are separated at a distance that is at least as great as that separating subfamilies of Indo-European. The opposite type of error would be groups of languages which are treated here as separate genera but which are as closely related as languages within subfamilies of Indo-European, particularly deeper subfamilies like Celtic. My hunch is that most errors in the taxonomy below are of the former sort. In a number of cases I have resolved doubtful cases by treating a more familiar or well-established genetic grouping as the genus. For example, I treat Salish as a genus rather than its subgroups. This may prove in need of revision. In general, the classification follows Ruhlen (1987), although I deviate in a few instances. My decision to use groups separated at a distance comparable to the separation of the subfamilies of Indo-European is partly arbitrary. However, languages **within** such groups are generally fairly similar typologically, so shallower groupings would probably not be appropriate. Deeper groupings become increasingly controversial, both in terms of genetic relatedness and in terms of actual grouping. Hence the particular level chosen as defining a genus.

The classification into genera is subject to dispute not only in terms of whether the groups I have treated as genera are of the appropriate level but also whether the groups in question are in fact valid genetic groups at all. This problem is due not only to the fact that classification in some parts of the world remains very tentative because of the amount of work that has been done (as in Australia), but also due to the fact that certain families present particularly difficult problems for subgrouping (as in the case of Austronesian).

The genera below are listed by area. The order within each area generally reflects probable genetic relationships between genera, but these are not specified here. There are, as noted in the text, many additional genera not represented in my sample.

AFRICA

North Southern African Khoisan, Central Southern African Khoisan, Kadugli, Kordofanian Proper, Mande, Northern West Atlantic, Kru, Gur, Ubangian, Ijo-Defaka, West-Central Niger-Congo, Yoruba, Edo, Igbo, Lower Cross, Cora, Bantoid, Songhai, Saharan, Maban, Fur, Kuliak, Surma, Nubian, Nera, Nyimang, Temein, Tama, Daju, Nilotic, Kresh, Sara-Bagirmi, Mangbutu-Efe, Balendru, Berta, Kunama, Komuz, Berber, Chadic, Omotic, Beja, Central Cushitic, Eastern Cushitic, Southern Cushitic, Semitic.

EURASIA

Basque, Armenian, Indic, Iranian, Albanian, Greek, Italic, Celtic, Germanic, Baltic, Slavic, Samoyedic, Finno-Ugric, Yukaghir, Mongolian, Tungus, Turkic, Japanese, Korean, Chukchee-Kamchatkan, Nivkh (Gilyak), Ket, Sumerian, Hurrian, Kartvellian, Northwest Caucasian, Nax, Dagestan, Burushaski, Northwest Dravidian, Dravidian Proper, Chinese, Karen, Tibetic, Baric, Burmic, Miao-Yao, Munda, Khasi, Palaung-Khmuic, Viet-Muong, Bahnaric, Khmer, Aslian, Nicobarese, Kam-Tai, Atayalic, Paiwanic, Philippine Austronesian, Sundic, Central-Eastern Malayo-Polynesian, Andamanese.

AUSTRALIA-NEW GUINEA

Finisterre-Huon, East New Guinea Highlands, Central and South New Guinea, Angan, Marind, Sentani, Dani-Kwerba, Wissel Lakes-Kemandoga, Binanderean, Central and Southeast New Guinea, Madang, Adelbert Range, Trans-Fly-Yelmek-Maklew, Kolopom, Torricelli, Sepik-Ramu, Bougainville, Yele-Solomons, Mangarayi, Nunggubuyu, Tiwi, Yiwaidjan, Gunwinyguan, Maran, West Barkly, Garawan, Daly, Wororan, Tangkic, Pama-Nyungan.

NORTH AMERICA

Eskimo-Aleut, Haida, Tlingit, Athapaskan-Eyak, Kutenai, Algic,

Chimakuan, Wakashan, Salish, Keresan, Yuchi, Siouan, Caddoan, Iroquoian, Tsimshian, Chinookan, Takelman, Yakonan, Klamath, Sahaptin-Nez Perce, Maidu, Yokuts, Costanoan, Miwok, Zuni, Chitimacha, Tunica, Muskogean, Yukian, Totonacan, Mixe-Zoquean, Mayan, Karok, Chimariko, Palaihnihan, Pomo, Washo, Chumash, Salinan, Esselen, Seri, Yuman, Tonkawa, Karankawa, Coahuiltecan, Tequistlatecan, Tarascan, Tanoan, Numic, Takic, Hopi, Pimic, Taracahitic, Aztecan, Coric, Otomian, Mixtecan, Popolocan, Zapotecan, Tlapaneca.

SOUTH AMERICA

Yanomam, Misumalpan, Guaymi, Itonama, Warao, Mura, Barbacoan, Cahuapanan, Zaparoan, Quechua, Aymara, Iranxe, Movima, Ticuna, Tucanoan, Cayuvava, Saliva, Jivaroan, Cariri, Tupi, Guahiban, Arawakan, Andoke, Carib, Guaicuruan, Mataco, Pano-Tacana, Rikbaktsa, Bororo, Chiquito, Ge-Kaingang.

Author's address:

Matthew S. Dryer
Dept. of Linguistics
SUNY
Buffalo, NY 14260
USA

ACKNOWLEDGEMENTS

I wish to thank the participants in the Symposium on Language Sampling at the 1987 Annual Meeting of the Linguistic Society of America in San Francisco, particulary Revere Perkins, William Croft, John Justeson, Ian Maddieson, and Johanna Nichols. I have also benefitted from comments from M. Lionel Bender and Edith Moravcsik. Talks incorporating the basic ideas of this paper were presented at Stanford University, UCLA, the University of Michigan, and the University of Manitoba; I am indebted to the audiences of those talks for relevant comments. The research for this paper was made possible by three research grants from the Social Sciences and Humanities Research Council of Canada, #410-81-0940, #410-83-0354, and #410-85-0540.

NOTES

1. The inclusion of both Ingassana (also called Gaam) and Maasai is more questionable since both are now classified as falling within the East Sudanic branch of Nilo-Saharan (Ruhlen 1987, M.L. Bender, p.c.).

2. It should be noted that I do not have data for all characteristics discussed in this paper for all of the languages in my sample. Relevant information for certain characteristics is lacking for many languages in my sample, and in other cases, languages may not fit clearly into one of the relevant typological categories; not all languages, for example, can be assigned a basic order of the sort SOV, SVO, etc.

3. Note that it is possible for a particular genus to be included in both figures, if it contains a language of both sorts. Usually this does not happen, since languages within genera are generally similar, but Semitic is an example of a genus that contains both SVO and SOV languages (as well as VSO languages).

4. Note that the term "proportion of genera" is somewhat misleading since a given genus may contain languages of both types. For example, suppose that within an area, there are 6 genera containing languages of one sort and 4 genera containing languages of the other sort. Then I would describe the proportion of genera of the first sort as .6 (6 out of 10). There may, however, not be a total of 10 genera here, since a genus may contain languages of both sorts. For example if two genera are both represented in both types, there will only be a total of 8 genera, not 10, and the proportions of genera will really be .75 and .5 rather than .6 and .4. Thus what I refer to as **proportions of genera** are really **proportions of subgenera**, where a genus is divided into subgenera with respect to a given typological parameter if the languages of the genus differ with respect to that parameter.

5. The conclusions of this section are based on my own sample and it is possible that there are other instances of SOV&NGen&NAdj order that are independent of the ones discussed here. Hawkins (1983) lists a number of additional possible instances of this order most of which I have not investigated. Apart from a number of further examples of languages in the same general areas as those in my sample, such as Africa and Australia, the most crucial instances are Ladakhi, a Tibeto-Burman language, and Khamti, a Tai language. According to my investigations, Ladakhi is GenN, not NGen. I have not investigated Khamti, though it should be noted that Khamti belongs to a large area that is predominantly NGen&NAdj that extends into Austronesian, including the SOV&NGen&NAdj language Manam.

REFERENCES

Aoki, Haruo. 1975. "The East Plateau linguistic diffusion area." *IJAL* 41: 183-199.

Bascom, William R.; and Herskovits, Melville J. (eds). 1959. *Continuity and change in African cultures*. Chicago: University of Chicago Presss.

Bell, Alan. 1978. "Language samples." In: Joseph H. Greenberg (ed.) 1978.

Benedict, Paul K. 1975. *Austro-Thai language and culture*. Human Relations Areal Files Press.

Boas, Franz. 1929. "Classification of American Indian languages." *Language* 5. (Reprinted in *Race, language, and culture*, New York: Macmillan, 1940.)

Bybee, Joan. 1985. *Morphology: a study of the relation between meaning and form*. Amsterdam: John Benjamins [*Typological Studies in Language* Vol. 9].

Campbell, Lyle; Kaufman, Terrence; and Smith-Stark, Thomas C. 1986. "Meso-America as a linguistic area." *Language* 62: 530-570.

Crothers, John. 1976. "Areal features and natural phonology: the case of front rounded vowels." *BLS* 2.

Delancey, Scott; and Tomlin, Russell (eds). 1986. *Proceedings of the Second Annual Pacific Linguistics Conference*. Eugene, Oregon: Dept. of Linguistics, Univ. of Oregon.

Driver, Harold E.; and Chaney, Richard P. 1970. "Cross-cultural sampling and Galton's problem." In: Naroll and Cohen (eds) 1970.

Dryer, Matthew S. 1986. "Word order consistency and English." In: Delancey and Tomlin (eds.) 1986.

Dryer, Matthew S. 1987. Final Report to Social Sciences and Humanities Council of Canada on Research Grant No. 410-85-0540: A statistical study of word order universals.

Dryer, Matthew S. 1988a. "Universals of negative position." In: Hammond, Moravcsik and Wirth (eds) 1988: 93-124. Amsterdam: John Benjamins.

Dryer, Matthew S. 1988b. "Object-verb order and adjective-noun order: dispelling a myth." *Lingua* 74: 185-217.

Dryer Matthew S. In preparation. "Noun-numeral order in Africa."

Greenberg, Joseph H. 1959. "Africa as a linguistic area." In: Bascom and Herskovits (eds) 1959.

Greenberg, Joseph H. 1963a. "Some universals of grammar with particular reference to the order of meaningful elements." In: Greenberg (ed.) 1963b.

Greenberg, Joseph H. (eds). 1963b. *Universals of language*. Cambridge, Mass.: MIT Press.

Greenberg, Joseph H. (ed.). 1978. *Universals of human language, Volume 1: Method and theory*. Stanford: Stanford University Press.

Greenberg, Joseph H. 1987. *Language in the Americas*. Stanford: Stanford University Press.

Hammond, Michael; Moravcsik, Edith; and Wirth, Jessica (eds). 1988. *Studies in syntactic typology*. Amsterdam: John Benjamins [*Typological Studies in Language* Vol. 17].

Hawkins, John A. 1983. *Word order universals*. New York: Academic Press.

Jacobs, Melville. 1954. "The areal spread of sound features in the languages north of California." In: *Papers from the Symposium on American Indian Linguistics. [University of California Publications in Linguistics* 10].

Keenan, Edward L. 1979. "On surface form and logical form." *Linguistics in the Seventies: Directions and Prospects, Studies in the Linguistic Sciences* 8:2.

Kenny, James A. 1974. *A numerical taxonomy of ethnic units using Murdock's 1967 word sample*. Unpublished Indiana University Ph.D. Dissertation.

Kim, Alan Hyun-Oak. 1985. Preverbal focusing and Type XXIII languages. Presented at the 14th Annual Linguistic Symposium of Language Typology and Universals, University of Wisconsin-Milaukee.

Masica, Colin P. 1976. *On defining a linguistic area*. Chicago: University of Chicago Press.

Naroll, Raoul. 1970. "Galton's problem." In: Naroll and Choen (eds) 1970.

Naroll, Raoul; and Cohen, Ronald (eds). 1970. *A handbook of method in cultural anthropology*. Garden City, NY: Natural History Press.

Nichols, Johanna. 1986. "Head marking and dependent marking grammar." *Language* 62: 56-119.

Perkins, Revere. 1980. *The evolution of culture and grammar*. Unpublished SUNY Buffalo dissertation.

Perkins, Revere, 1985. "The covariation of culture and grammar." Paper presented at the Symposium on Language Typology and Universals, University of Wisconsin at Milwaukee. In: Hammond, Moravcsik and Wirth (eds) 1988: 359-378.

Ruhlen, Merritt. 1987. *A guide to the world's languages. Volume 1: Classification*. Stanford: University Press.

Thompson, Laurence C.; and Kinkade, M. Dale. Forthcoming. "Linguistic relations and distributions." In: *Handbook of American Indians, Volume VII: The Northwest Coast.*

Tomlin, Russell. 1986. *Basic constituent orders: functional principles*. London: Croom-Helm.

Studies in Language 13-2. 293-315 (1989).

STATISTICAL TECHNIQUES FOR DETERMINING LANGUAGE SAMPLE SIZE

REVERE D. PERKINS
SUNY at Buffalo

1. Introduction

About a century ago Edward Tylor read a paper at the Royal Anthropological Institute that introduced the cross-cultural survey method. Tylor correlated several anthropological traits for a number of cultures and drew some conclusions about the theoretical relationships between the traits. Francis Galton objected that since anthropological traits often spread by migration and borrowing it was uncertain how many independent cases supported Tylor's conclusions (Naroll 1973a: 893f, 1973b: 974). The objection he raised is still relevant. Anthropologists, led by Raoul Naroll, have dealt with Galton's objection, or Galton's problem, as it has come to be known, extensively over the last thirty years and provided a variety of 'solutions.' Linguists and other social scientists seem not to have understood Galton's problem or have chosen to ignore it, even though the results of several recent cross-linguistic studies are open to very similar objections.

Anthropologists, following Naroll's lead, have dealt with Galton's problem indirectly by measuring the overall extent to which diffusion of traits is evident in the cultures in a sample. If diffusion is not sufficient to account for the association between the traits being investigated, Galton's problem is considered solved. By looking at the results of testing for Galton's problem in a number of studies, Naroll recommends 50 to 75 cultures as being the most that one can sample without Galton's problem being severe enough to invalidate one's results, at least when traits that diffuse widely are being studied. For traits that do **not** diffuse widely Naroll suggests that three or four hundred cultures are optimal (Naroll 1973a: 889). Previously, it has been unclear which of these sets of numbers should