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

— (2005). How many words do you need? A note on corpus and lexicon size. Electronic document. http://emeld.org/school/classroom/text/lexicon-size.html

Nichols, Johanna & Balthasar Bickel (2005a). Locus of marking in the clause. In Haspelmath et al. (eds.), 98–101.

— (2005b). Locus of marking: Whole-language typology. In Haspelmath et al. (eds.), 106–109.

Nichols, Johanna, David A. Peterson, & Jonathan Barnes (2004). Transitivizing and detransitivizing languages. *Linguistic Typology* 8: 149–211.

Correlating phonological complexity: Data and validation

by IAN MADDIESON

1. Introduction

In a forthcoming paper I examine the patterns of correlation found in a large sample of languages between several fairly simple indices of phonological complexity. The indices examined all relate to phonological contrast and include the number of contrastive consonant segments, the number of contrastive vowel qualities, the total number of vowel nuclei, the presence of tonal contrast and its degree of elaboration, and the complexity of permitted syllable structure. Of course, other properties of phonological systems not considered here, such as word structure templates or the degree of transparency in alternations, also contribute to complexity. Across the language sample as a whole, the analysis shows that of the ten possible pairwise comparisons between the selected measures five show positive correlations, four show no correlation, and one shows a negative correlation. The results are summarized in Table 1, and discussed in more detail in Maddieson (forthcoming).

The data in Table 1 provide *prima facie* evidence against an apparently widely-held belief among linguists that complexity in one sub-part of the grammar of a language is likely to be compensated for by simplification in another. Such a view seems to be based on the humanistic principle that all languages

Table 1. *Direction of correlation between complexity measures*

	syllable complexity	# consonants	# V qualities	# V nuclei
# consonants	positive			
# V qualities	uncorrelated	uncorrelated		
# V nuclei	uncorrelated	uncorrelated	positive	
tone complexity	negative	positive	positive	positive

are to be held in equal esteem and are equally capable of serving the communicative demands placed on them. In rejecting the notion of "primitive" languages linguists seem to infer that a principle of equal complexity must apply. As Akmajian et al. (1979: 4) put it "all known languages are at a similar level of complexity and detail – there is no such thing as a primitive human language". From this point of view it might seem a logical further step to hypothesize that languages will undergo adjustments to equalize their overall complexity across different subsystems.

As far as phonology is concerned it seems to be widely believed (though such beliefs are rarely expressed in print) that, for example, a language with a large number of distinct consonants is likely to have a small number of vowels, or that a language with an elaborated tone inventory will have a tendency to simple syllable structure. These are two of the issues that the analyses summarized in Table 1 were designed to examine. But the appropriate way to conduct such an analysis of a sample of languages and what confidence to place in the results remain very open questions for linguists. The focus of this contribution will be on one of the correlations noted in the table, the positive correlation between syllable complexity and size of consonant inventory, and on some of the possible strategies to test its reliability. Before proceeding further it may be worth emphasizing that this correlation is not an automatic consequence of any necessity to have a large consonant inventory in order to make complex syllables. A small inventory – and the smallest contains six consonants – could be employed to create many varied strings of consonants in both onset and coda positions of syllables. A complex syllable can be created from an inventory of two consonants, as in English *stoats* (/stoʊts/). With just these consonants even more complex syllabic structures could easily be constructed, such as /stsotst/.

The manner of compilation of the language sample used will first be described, then the global result for the correlation in question presented. The remainder of the article will discuss various strategies for testing the validity of the result. In this discussion the importance of identifying outliers when means of numerical indices are being examined will emerge, as well as issues concerning sampling procedures. Some initial steps towards the use of re-sampling strategies will be proposed.

2. The language sample

The sample of languages analyzed in the present study is essentially an expanded version of the UCLA Phonological Segment Inventory Database (UPSID) sample described in Maddieson (1984). This was earlier extended from 317 to 451 languages for a version distributed as an interactive database (Maddieson & Precoda 1990). At both these stages, selection of languages for inclusion was governed by a quota principle seeking maximum genetic diversity among extant languages (and those spoken recently enough to have been described by linguistically sophisticated observers from direct experience). More recently, the author's participation in the *World Atlas of Language Structures (WALS)* project (Haspelmath et al. (eds.) 2005) led to the merging of the UPSID sample with the 200 languages chosen as the core sample for *WALS*. Although genetic diversity was a major goal of the *WALS* sample, a number of other factors also played a role in the selection, including the political importance of a language, the availability of a full-length grammar, the presence of the language in other typological samples and a desire for wide geographical coverage. Due to overlapping membership of the UPSID and *WALS* samples this resulted in a merged sample of about 520 languages. As part of a project to more fully examine geographical distribution of phonological characteristics of languages, this database is continuing to be enlarged, with the intention of reaching a total of 1000 or so languages. At the time of writing the database includes 621 languages, with the data for different languages at varying levels of completeness.

The addition of more languages, and in particular the relaxation of the requirement that no pair of closely related languages be included, changes the representativity of the sample. On the one hand, including more languages means of course that the range of diversity among languages has a greater chance of being better represented. Naturally, properties that are rare in the world's languages are more likely to be found in at least one of the sample languages as more and more are included. On the other hand, the relative frequency with which particular characteristics are represented may be increasingly distorted by factors such as the aim of including languages from geographical areas that would otherwise be sparsely sampled, and the inequality of availability of information on languages from different families and parts of the world. These factors are likely to result in certain language families being overrepresented in comparison with the numbers that would be included in the kind of genetically stratified sample that is most often proposed as the foundation of quantitative typological research (Bell 1978, see also discussions by Perkins 2001, Cysouw forthcoming). Such concerns make it even more critical to consider ways of testing how meaningful any particular result extracted from the data is.

3. Overall correlation of syllable structure type and consonant inventory size

The result that will be the focus of this article is the finding of a positive overall correlation between the two parameters of syllable complexity and consonant inventory size. Syllable complexity is a ranked categorial value based on the maximally complex syllable type that is permitted in the language. It has three values, Simple, Moderate, and Complex. The structure of onsets and of codas are both taken into account in this evaluation, but nuclei are not considered as contributing to syllable complexity. Languages with Simple syllable structure are those which do not permit any elaboration of the onset beyond a single consonant and do not permit any codas. In other words they permit only (C)V syllables. Among languages in this category are Yoruba, Ju|’hoan, Maori, and Guaraní. For example, verbs in Yoruba typically consist of a single CV syllable, such as /bɔ/ ‘come’ or /k͡pa/ ‘hit, kill’. Languages with a Moderate level of syllabic complexity are those in which the coda may contain not more than one consonant, and the onsets are limited to a single consonant or a restricted set of two-consonant clusters having the most common structural patterns, typically an obstruent followed by a glide or liquid (the 30 language sample used in Maddieson (1992) had shown that simple codas and minimally elaborated onsets tend to co-occur in languages). Among languages in this class are Ewe, Dadibi, Nahuatl, Somali, Yanyuwa, Sundanese, Thai, and Mandarin. Ewe and the New Guinea language Dadibi allow onsets with a glide or liquid in second position, but no codas. Nahuatl, Somali, and Yanyuwa have no tautosyllabic clusters, but allow a single coda consonant. Sundanese, Thai, and Mandarin allow single codas as well as onset clusters with liquid or glide as C_2. In Mandarin the “medials” of the traditional analysis of Chinese syllables are treated as glides in an onset cluster. The most complex Mandarin syllable thus has a structure like that in words like /ljâŋ/ ‘bright’. Languages which permit a wide range of onset clusters, or which permit clusters in the coda position are classified as belonging to the Complex syllable category. Languages in this class include Georgian, Lushootseed, Soqotri, and French. Some Georgian examples of complex syllables are /prxiv/ ‘on foot’ and /vxetkt/ ‘we are splitting’; Lushootseed examples include /sqibk^{w}’/ ‘octopus, squid’, /q’tsab/ ‘lack’, /sx^{w}q^{w}’up/ ‘Green River’, and /sup’qs/ ‘hair seal’; French examples include /stʁikt/ ‘strict’, /klwatʁ/ ‘cloister’, and /tʁɥizm/ ‘truism’. It should be noted that these classifications are only in terms of the maximal syllable elaboration permitted, and languages may differ considerably in how frequently the more complex patterns occur in the lexicon or in text. Recent international loan vocabulary is excluded from evaluation of a language’s syllable complexity category.

Languages in the Moderate class are considerably more frequent than either of the other types. Of the 515 languages in the current sample for which values of both syllable complexity and consonant inventory size are entered, 294 are Moderate (57.1 %), while just 62 are Simple (12.0 %), and 159 are Complex (30.9 %).

The relationship of syllable type to the number of consonants in the inventory is examined in some detail here. The size of the inventory is a straightforward numerical index – in the current sample ranging from a low of 6 to a high of 128. For many languages the consonant inventory size is a relatively uncontroversial datum, although there are plenty of cases which are open to varying interpretations and for some languages the data may be unreliable. An advantage of using a large sample of languages is that errors in either direction are more likely to cancel each other out in an analysis of the aggregate. The comparison of the two variables is performed in two ways, using either all the 515 available languages or excluding a small number with what are considered to be outlier values for the number of consonants. An outlier is a value that is far from the range of the majority of the data; as a rule of thumb an outlier is often defined as a value more than two standard deviations from the mean. In the entire sample the mean number of consonants per language is 22.72 with a standard deviation of 10.65.

When the set of languages is divided by syllable category, there are on average more consonants in the inventories of languages with complex syllable structures than in languages with moderately complex syllable structures and the least in languages with simple syllable structure. This result is shown graphically in the left-hand panel of Figure 1. In an analysis of variance, syllable category has a significant effect on consonant inventory size ($F(2,512) = 11.097$, $p < .0001$), but in a post-hoc comparison of means using Fisher's PLSD (Protected Least Significant Difference) test adjusted for unequal cell sizes the difference between Moderate and Simple categories does not reach significance. Since outlier values can have distorting effects on analyses based on means, the analysis was repeated excluding those languages with consonant inventories more than two standard deviations from the mean following the common rule of thumb used to define outliers. This results in excluding the 18 languages with 45 or more consonants, or about 3.5 % of the sample. Not surprisingly, this exclusion results in lower means for the number of consonants, as shown in the right-hand panel of Figure 1, with a particularly marked reduction in the mean for Simple languages. Analysis of variance of this reduced set of languages shows a highly significant effect of syllable type on size of consonant inventory ($F(2,494) = 28.87$, $p < .0001$). In this case all post-hoc comparison of means are significant at $p < .002$ or better. Finding a significant difference between Moderate and Simple categories is not due to exclusion of a disproportionate number of outlier languages with Simple syllable structure. In fact, only two languages are excluded from this category (3.2 %), compared to ten removed from the Moderate category (3.4 %) and six from the Complex category (3.8 %). However, since the number of Simple languages is the smallest the exclusion of any language has a greater effect on the mean, and among the languages excluded is Ju|'hoan, which has the second highest number of consonants (92) of any language in the sample.

Since means can be misleading indicators of group differences if distributions of values within the groups are non-normal, the data were also examined

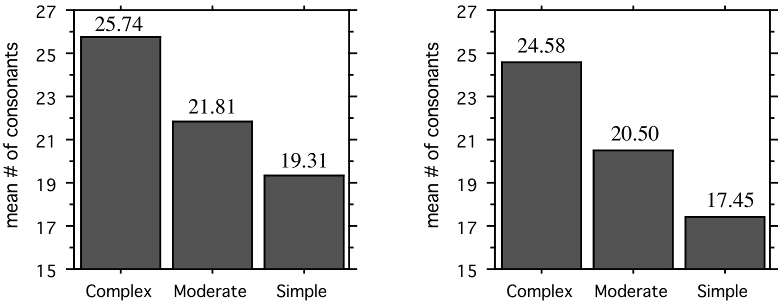

Figure 1. *Mean number of consonants by syllable complexity class. On the left, analysis with all languages included; on the right with exclusion of languages with more than 44 consonants.*

to see how well the three groups of languages meet the assumption of normality. Histograms of the number of languages with given sizes of consonant inventories for each syllable category are presented in Figure 2 together with the values for skewness and kurtosis. Outliers, as defined above, are excluded from these figures and calculations. A best fit to a normal distribution curve is superimposed on each of the histograms. These plots show that the normal fit is good for Complex and Moderate languages, and there are low values for skewness and kurtosis in these groups. Skewness is a measure of how equally the values are distributed above and below the mean, kurtosis a measure of of how peaked the distribution is. In a perfectly normal distribution both are zero; a value above 3 is usually considered excessive. A normal fit is somewhat less appropriate for Simple languages. Nonetheless, the skewness and kurtosis of this distribution are not excessive. The platykurtic (flat) peak is in part attributable to the fact that the number of languages in this category is the lowest.

Figure 2 also confirms that the "center of gravity" of the distribution is lower as syllable structure is simpler. The midpoints of the distributions shown in Figure 2 can be expressed by the medians, which increase with an increase in syllable complexity. For the Complex languages the median number of consonants in the inventory is 24, for Moderate languages the median is 20, and for Simple language the median is 17. If outliers are included, the median for Simple languages increases to 18, but the others do not change. For the entire sample the median is 20 excluding outliers, which increases to 21 when the outliers are included.

The analyses presented above demonstrate that in this large sample of languages rather than complexity in syllable structure being compensated by simplicity in consonant inventory, there is a tendency for consonant inventories to be more elaborated in languages with increasing levels of syllabic complexity.

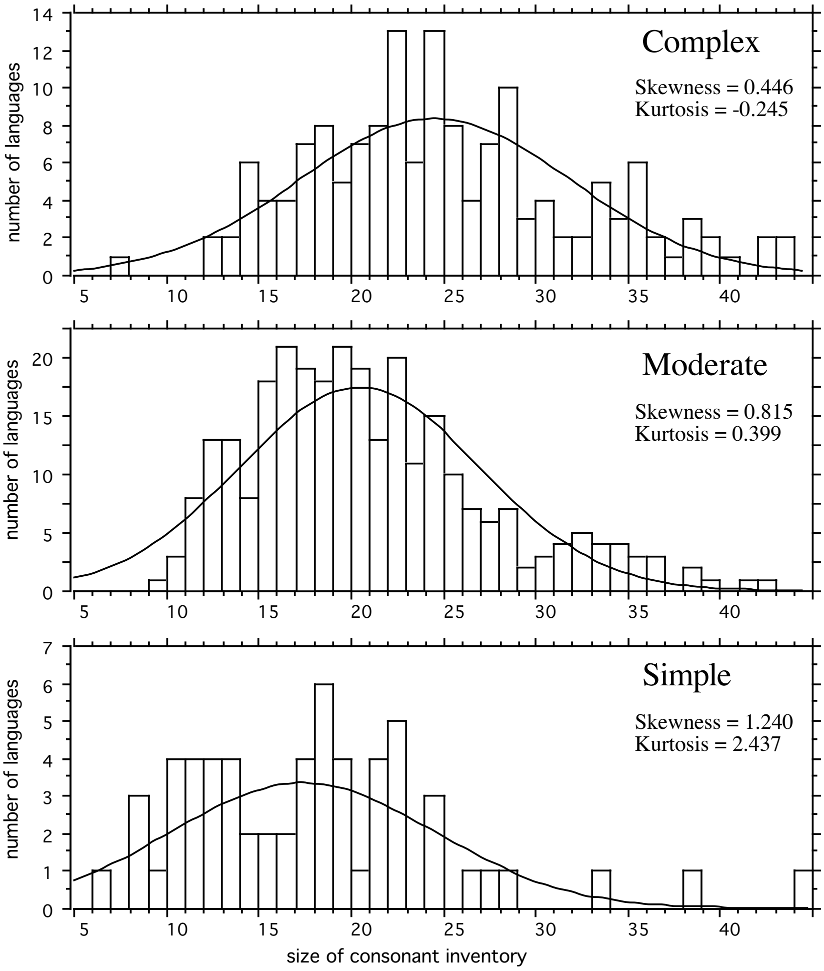

Figure 2. *Histograms of the number of languages with given consonant inventory size, by syllable complexity category*

It has been shown that the differences between the means of consonant inventory size across the syllable complexity categories are significant, and that it is appropriate to take the means as representing the within-group distributions.

This result is over the surveyed languages as a whole, but what does this mean? Can the trend discovered be taken as general or does it reflect strong but local tendencies – patterns that are particularly strong in certain areas or language families? Are there imbalances in the sample which risk creating distortions in the results?

4. Validation by areal sub-samples

Various methods have been suggested to test the generality and validity of results in large language samples. One is to divide the sample on areal/genetic lines and look for the repetition of a pattern across the divisions, either in the aggregate of the languages in each division or within genetic units judged to represent comparable historical depths to each other. This approach has been particularly advocated by Dryer (e.g., 1989, 1992, 2003) and also widely employed by Nichols (e.g., 1992). Dryer (1989) originally suggested using five divisions (cf. Maddieson 1991) but in later work has used six. Nichols (1992) employed a variety of groupings ranging from just three macro-areas to as many as twelve smaller areal groups, as well as using alternate cross-cutting purely genetic and contact-based groupings. In the present analysis six areal/genetic groupings are used which are similar to Dryer's but differ in some details.

The groups are based primarily on the continental landmasses and the major language families which are indigenous to and "anchored" on these landmasses. All members belonging to a given language family are assigned to a single area even when the dispersion of the family crosses the areal divisions employed. In other words, in cases where strictly geographical and genetic affiliations are in conflict, the genetic relationship between languages are given priority. The six groups and the number of languages in each group in the 515 language sample analyzed here are shown in Table 2.

The "Europe, West & South Asia" area covers Europe, Asia Minor, and the Indian subcontinent as well as Siberia and the rest of Asian Russia. This is home to several large language families such as Indo-European, Dravidian, Altaic, and Uralic, as well as smaller groups such as the Caucasian families and Chukchi-Kamchatkan and isolates like Basque and Burushaski. Japanese, Korean, and Ainu are taken to be Altaic languages and so fall in this group despite being outside the geographical limits. Creoles lexically based on Indo-

Table 2. *Areal/genetic groupings*

Area	Number of languages in sample
Europe, West & South Asia	74
East & South-East Asia	102
Africa	125
North America	66
South & Central America	70
Australia & New Guinea	78

European languages are grouped with Indo-European for want of a better classification.

The "East & South-East Asia" area covers China, the Indo-Chinese peninsula, and island Asia. The main language families are Sino-Tibetan, Austro-Asiatic, Tai-Kadai, and Austronesian. Since the Austronesian family is anchored in this area, the many Austronesian languages spoken outside the area (e.g., Malagasy, Iaai, Maori, Rapanui) are included in this group.

"Africa" is simply the African continent and its adjacent islands. All languages in the large Niger-Congo, Nilo-Saharan and Afro-Asiatic families are included in this group as well as the languages in the Southern African "Khoisan" groups and the probable isolates Hadza and Sandawe. The main mismatch with the geographical limits concerns those Afro-Asiatic languages spoken in Asia Minor or adjoining areas of Europe, which are included in this group.

"North America" is geographically that part of the Americas north(-west) of the Isthmus of Tehuantepec in southern Mexico. It includes Greenland and the Aleutian islands and the major islands of the Caribbean. There are numerous language families which are entirely indigenous to this geographical area, including Na-Dene, Eskimo-Aleut, Uto-Aztecan, Algonquian, Tanoan, Wakashan, Hokan, and Totonacan among others. Oto-Manguean, Mixe-Zoque, and Huavean languages are also classed with the North American languages although some are found at or just south(-east) of the Isthmus.

"South & Central America" is that part of the Americas to the south(-east) of the Isthmus of Tehuantepec. This area includes the Yucatán peninsula of Mexico as well as the countries traditionally considered to form Central and South America. A number of well-recognized indigenous language families such as Cariban, Arawakan, Chibchan, Je, Tupian, and Mayan are included here as well as other languages and groups whose genetic affiliations are less well understood.

The final area, labeled "Australia & New Guinea", geographically covers the area often referred to as Oceania in British usage, namely Australia, New Zealand, and New Guinea and the other islands of Melanesia, as well as Micronesia and Polynesia. Since many of the languages spoken in this area are Austronesian languages and these are included in the East and South-East Asia group, the only languages in this group are the indigenous languages of Australia and the "Papuan" languages. Papuan is a cover term for all the non-Austronesian languages of island Asia and Melanesia including the large Trans-New Guinea family as well as other smaller groupings spoken in parts of Indonesia, Papua New Guinea, and the Solomon Islands whose historical relationships are not yet well understood.

Table 2 shows that the numbers of languages in the current sample from the different areas are quite unequal. In particular, there are almost twice as many languages from Africa as from North America. This specific imbalance reflects

Table 3. *Numbers of languages in areal/genetic groups by syllable complexity category*

	Europe, West & South Asia	East & South-East Asia	Africa	North America	South & Central America	Australia & New Guinea
Complex	55	20	24	35	10	15
Moderate	19	72	83	29	39	52
Simple	0	10	18	2	21	11

partly the fact that the African area has had more effort invested in compilation of the sample at its present stage, but also the fact that a very large number of languages have been recognized in Africa and most of them continue to be spoken daily whereas many indigenous languages of North America were already lost from daily use before being adequately described.

The breakdown of the languages in the sample by areal/genetic groupings and syllable complexity category is given in Table 3. The table reveals salient differences among the groupings in terms of relative frequency of the syllable types. In four groups, Moderate languages are predominant, as in the entire sample, but in the Europe, West & South Asia group and in North America Complex languages are most frequent. No area has a majority of Simple languages, but their proportion is greatest in South and Central America.

In order to examine the relationship between the two variables of interest in each area a sufficent number of cases must occur in each cell of this table. Since there are no cases of Simple syllable structure in the languages in the first area (Europe, West & South Asia), and only two in North America, these areas must be omitted from any area-by-area comparison involving all three syllable complexity classes. The mean consonant inventory size by syllable category in the remaining four groupings is shown in Figure 3. In the upper panel all languages in the four groups are included, a total of 375, whereas in the lower panel outliers, as defined earlier, are excluded. The exclusion removes two languages from the East & South-East Asia group, one from Australia & New Guinea, but eight from Africa. A two-way analysis of variance with area and syllable category as main effects shows that there is, as expected, a highly significant effect of area on consonant inventory size ($F(3, 363) = 13.209$, $p < .0001$), with all pairwise comparisons being significant at better than $p < .004$ except for that between South America and Australia & New Guinea, which are not significantly different. Syllable category does not have a significant effect. However, when outliers are excluded there is a significant main effect of syl-

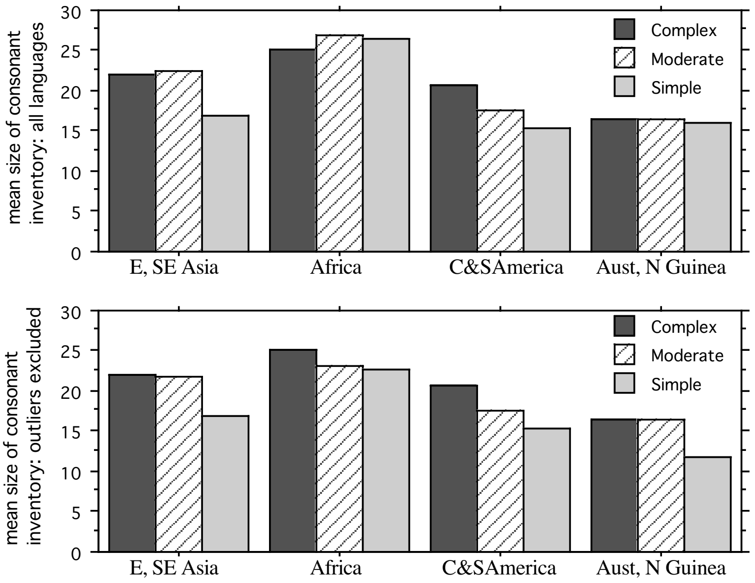

Figure 3. *Histograms of mean consonant inventory size by syllable complexity category across four areal/genetic language groupings. Upper panel: all languages included; lower panel: outliers excluded.*

lable complexity categories ($F(2, 351) = 8.867$, $p < .0002$). With the outliers included both the East & South-East Asia and Africa groups show exceptions to the monotonic decrease of mean consonant inventory with simpler syllable type, and the Australia & New Guinea group shows virtually no difference across the syllable complexity classes. When outliers are excluded, there are no direct counterexamples to the pattern found in the overall results as shown in Figure 1, and replicated most clearly in the Central & South American group in Figure 3.

Moderate and Complex languages can be compared in the two remaining areal groupings, as shown in Figure 4, on the left with outliers included, and on the right with outliers excluded. Exclusion of outliers removes only one language from the North American group, and five (all from Caucasian families) from the Europe, West & South Asia grouping. In a two-way analysis of variance, there is no significant difference in mean consonant inventory size between these areas, but a highly significant difference between the Complex and Moderate syllable categories, whether outliers are included or excluded (in the latter case, $F(1, 128) = 16.519$, $p < .0001$).

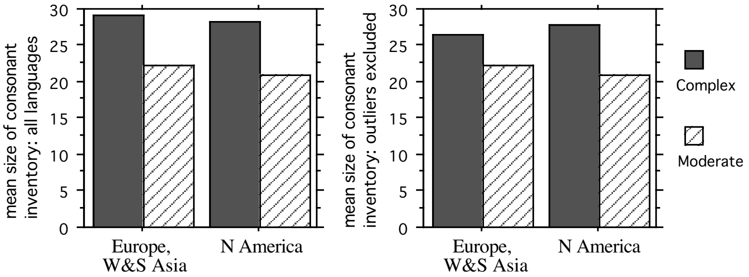

Figure 4. *Histograms of mean consonant inventory size by syllable complexity category for remaining areal/genetic language groupings. Left panel: all languages included; right panel: outliers excluded.*

In summary, although the syllable complexity patterns are unequally distributed among the areal/genetic groupings, there is good evidence that the correlation between consonant inventory size and syllable complexity is not a localized pattern. When the distorting effect of outliers is removed, Complex languages have a saliently higher mean consonant inventory size than Moderate ones in four of the six areas (in Europe, West & South Asia, Africa, North America, and Central & South America), while in the remaining two areas the means are roughly equal. In the three of the four areas in which Moderate and Simple languages can be compared, the Moderate languages have higher mean consonant inventory size than Simple ones (in East & South-East Asia, Central & South America, and Australia & New Guinea), the exception being Africa where Moderate and Simple language means are similar. It follows from the foregoing that in each of these four areas Complex languages have higher mean consonant inventory size than Simple ones. Of the ten possible within-area comparisons between a given syllable complexity level and the next lower-ranked level, seven show that mean consonant inventory size is greater in the more complex syllable class, and in five of these the difference is statistically significant (using Fisher's PLSD, with significance levels ranging from .0341 to .0016) despite the small number of cases in several of the cells being compared. These statistically significant comparisons occur in five different areas, Africa being the only area in which a significant difference is not found. The remaining three within-area comparisons show roughly equal mean numbers of consonants. No comparison shows a larger mean consonant inventory size in the less complex syllable class.

In employing a within-area analysis procedure the analysis above is similar in certain respects to the approach taken by Dryer, but it differs significantly in not employing any subsetting by genera. Dryer looks for a "majority vote" among genera rather than among languages in each area. A genus is loosely defined as a genetic group “roughly comparable in time depth to the subfamilies of Indo-European” (Dryer 1992: 83–84). Since Dryer is dealing with relations between binary categories his method is to count the number of genera in which a hypothetically related pair of values occurs and compare that with the number where other pairings occur. If the languages in a genus are not uniform, the genus will be counted more than once. Dealing with a numerical variable, as in the present study, it is not possible to follow a directly analogous procedure. However, given a classification into genera, it would be possible to calculate the mean for each genus and then average these means, so that each genus contributes equally to the areal mean. Alternatively, the consonant inventory size variable could be divided into binary categories (languages above the median, vs. those below) and a procedure similar to Dryer’s applied. Neither of these steps have been pursued, in part because I am less confident than Dryer (and less confident than I used to feel) that comparable genera can in fact be established around the world on the basis of current knowledge.

The results in this section can be taken as sufficient to reject any suggestion that the overall correlation found in the sample is the result of a particularly strong trend in the languages of just one or two given areas or families which swamps weaker contrary trends outside those groups. This strengthens the case for rejecting the hypothesis of compensation between the two parameters being compared. But can this result be taken to indicate that there is actually a general trend for increase in syllabic complexity and elaboration of consonant inventory to go together in natural spoken human languages? Although individual languages vary a great deal in how the two parameters match up (and languages with simple syllable structure but large consonant inventories or with complex syllable structure and small consonant inventories certainly both occur), is it reasonable to conclude that the paths of natural historical linguistic change are in general more likely to create positively correlated patterns of complexity? To support such a conclusion some of the well-known problems concerning construction and analysis of large language samples (which led Dryer to adopt the strategy he employs) must be confronted. These will be briefly discussed in the following section. Various sub-sampling strategies designed to address (some of) these problems will then be presented and their outcomes examined in an attempt to evaluate if the general trend found in the aggregate data should be considered a meaningful and reliable result.

5. Sampling and re-sampling strategies

The principal concern raised with large language samples has been over the issue of independence of the entries selected. Independence has a number of meanings in this context, and which of these aspects are relevant – and the extent to which they raise concern – depends on how the sample is to be used. One meaning concerns whether the chance of any given language being included in the sample is independent of the chance of another language being included. Random sampling can ensure independence of this kind. Given a list of the currently extant (or recently spoken) languages, it should be possible to construct a random sample of this population. However, there are two serious problems in constructing such a sample. The first is that there is in fact no reliable list of languages available, and no immediate possibility of this being constructed, despite admirable attempts to approach this goal, notably in the compilation of the *Ethnologue*, now in its fifteenth edition (Gordon (ed.) 2005). The difficulties are twofold. First, knowledge of linguistic diversity "on the ground" remains quite unequal across different areas of the world. Secondly, and perhaps more critical, any enumeration of the number of languages depends on the establishment and consistent application of criteria as to what constitutes a separate language (if that is even possible). It seems apparent that in available lists, quite uneven criteria have been applied. For example, the *Ethnologue* lists a total of 35 Arabic languages, but just 14 Sinitic languages in China. Many observers feel that Chinese subsumes a greater divergence of linguistic varieties than Arabic.

Not only is there a dearth of knowledge about the number of languages in existence, but biases are introduced into the information that is available by the fact that linguists frequently tend to keep working in an area or language family they know, so that related or adjacent languages are more likely to gain documentation. Because not merely knowledge of the existence of some language but the appropriate data on the language for the intended analysis is required, this factor compromises the independence of the available sample points to some difficult to quantify degree.

A second meaning of independence of samples concerns whether the data value or values for a chosen sample language is or are independent of those of other languages sampled. This factor is the one that has received the most discussion, and has most influenced the efforts to design sampling strategies. A pair of languages that descend from a common parent, or a pair whose speakers have been in contact with each other over a period of time, are generally assumed to be more likely to share features than a pair of languages not known to have a common parent, nor to have been in contact. The importance given to this sense of independence derives from the purpose for which most language samples are assembled. Relatively few researchers in the domain of language typology and universals are interested in the frequency or contingency relations of properties as they are distributed across the universe of extant languages – which is the universe that language samples can most realistically hope to represent. Instead it is often hoped to provide a foundation to make, as Perkins (2001: 423) puts it, "inferences about LANGUAGE, the general faculty of which individual languages are instances". It may be useful to conceptualize this goal as a search for those patterns that would still hold true if the history of humankind since the birth of language could be re-run multiple times with quite different probablities for the survival and extinction of given languages over the millennia. Since we only have one history to deal with, we cannot actually compare multiple outcomes. Most typologists seem to feel that the best that can be done is to structure the sample so that relatively high nodes in the genetic trees of extant language families determine sampling strata (Nichols 2003 demurs in proposing a primarily areal sampling strategy – as used by Shosted 2006). Beyond this general agreement, a variety of strategies has been proposed, ranging from those whose selection criteria is a largely intuitive determination of historical depth (Maddieson 1984, Bybee et al. 1994) to those which use a computation of branching complexity of family trees (Rijkhoff & Bakker 1998), and those which include prior tests for the association of variables of interest before sample size is determined (Perkins 1989, Stephens & Justeson 1984). Here we wish to suggest that sub-sampling and re-sampling techniques should be more seriously considered by typologists, and we take a few tentative steps in that direction.

Using a small randomly-drawn sub-sample from the larger sample may be the best that can be done to approximate a random sample of languages. If the sub-sample's size is small it is likely to be constructed in a way that also satisfies the desire for only largely unrelated and non-adjacent languages to be included. The process can be repeated and an average result over a number of trials obtained in a similar fashion to the way that jackknifing techniques are used in cases where few (independent) data points are available. The purpose is to estimate how great the chance variation contributed by sample structure is, and to use this to determine the reliability of any result. The mean result obtained by selecting five random sub-samples of 30 languages each is shown in Figure 5. Random number variables were generated using a function in the Statview program (SAS Institute 1992–1998) for the purpose and the list of

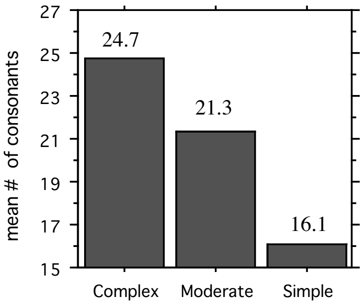

Figure 5. *Mean number of consonants by syllable category in five random samples of 30 languages*

languages then sorted by ordering the random numbers and picking the 30 languages aligned with the lowest numbers. Outliers which would have been picked by this process were excluded and the next language taken (exclusions ranged from 0 to 2 in the 5 trials). Other than this, any language had an equal chance of being selected, that is, no constraint against the same language occurring in more than one sub-sample was enforced, nor was the selection "edited" to meet any areal or genetic diversity criteria. Except for the fact that two of the five sub-samples included just one North American language (against an expected four or so), the sub-samples generally had a good balance across areas and languages families. A monotonically decreasingly mean number of consonants was found with decreasing syllable complexity. Across this relatively small number of trials, the grand means of the Complex and Moderate sets are both significantly different (by Fisher's PLSD) from that of the Simple sets at at least the .007 level. The individual sub-sample means for the Complex sets range between 20.8 and 28.0 (s.d. 2.76), for the Moderate sets between 19.3 and 25.8 (s.d. 2.64), and for the Simple sets between 14.5 and 19.0 (s.d. 2.04). This variation provides the estimate of the random component due to choice of sampled languages.

However, there is a potentially serious flaw in the foregoing procedure. Because of the small number of languages in the Simple syllable class in the entire sample, only 2 to 4 Simple languages are included in each of the random sub-samples, against 6 to 16 Complex and 10 to 22 Moderate languages. Such a small number of data points is a poor basis for constructing a meaningful average. A second method was therefore tried which constrained the sub-samples to include ten languages from each syllable category. The selection within-category is random, giving a hybrid random-stratified sample. Outliers (between 0 and 1 per sub-sample) were excluded as before, but no other modifications to the random selections made. Mean results from five trials with this constraint are shown in Figure 6. At least two languages from each area occur in each sub-sample. Again, a monotonic decrease in mean consonant inventory size with decreasing syllable complexity is found. The difference between the Complex mean and both those of the Moderate and Simple sets is significant (by Fisher's PLSD) at at least the .02 level. The range for Complex sets is between 22.2 and 25.0 (s.d. 1.10), for Moderate sets between 14.4 and 21.7 (s.d. 2.86), and for Simple sets between 14.4 and 20.2 (s.d. 2.69). In this exercise, a truer estimate of the variation contributed by choice of sampled languages in the Simple category is probably obtained, and the more limited number of Moderate languages in each sub-sample makes it less likely that this category will simply converge to the overall mean of the whole sample.

Although re-sampling techniques would normally be carried out with a much larger number of iterations, the limited analyses of sub-samples included here suggest that the trend found in the overall sample for larger consonant inventory

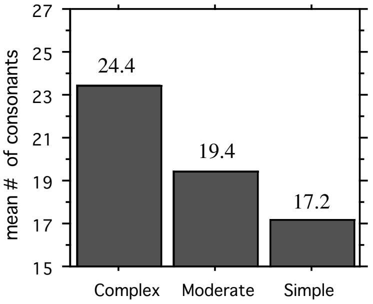

Figure 6. *Mean number of consonants by syllable category in five random samples of 30 languages constrained to include 10 languages of each syllable category*

to co-occur with more complex syllabic possibilities is quite robust, and is not dependent on a biased choice of sample languages. Of course, this finding only relates directly to the universe of extant and recent languages. The extent to which it might be projected to the universe of possible languages is another debate altogether.

6. Summary

This paper makes two principal points. First, it presents a statistical finding: that increasing syllable complexity shows an overall positive correlation with increasing size of consonant inventory in modern spoken languages. Secondly, it examines the reliability of this finding using several techniques, including dividing the total sample into areal/genetic sets and using re-sampling. Though re-sampling methods have been used increasingly often as an analytical tool in other disciplines, their application in typological linguistic studies to date appears negligible. It has also highlighted the need to consider the nature of the variables being examined and the distribution of the values in any sample used. In this case, the fact that one of the variables is a real number (an interval scale) with a near-normal distribution opens up certain possibilities, such as reliance on means, but closes off others. Major inequalities in the frequency of languages across the syllable complexity categories needed to be taken into account in designing appropriate tests.

Received: 19 August 2005
Revised: 16 March 2006

University of California at Berkeley

Correspondence address: Department of Linguistics, University of California at Berkeley, Dwinelle Hall 1203, #2650, Berkeley CA 94720-2650, U.S.A.; e-mail: ianm@berkeley.edu

Acknowledgment: This work was supported by the National Science Foundation through grant BCS-034578 to the author.

References

Akmajian, Andrew, Richard A. Demers, & Robert M. Harnish (1979). *Linguistics: An Introduction to Language and Communication*. Cambridge, Mass.: MIT Press.

Bell, Alan (1978). Language samples. In Joseph H. Greenberg, Charles A. Ferguson, & Edith A. Moravcsik (eds.), *Universals of Human Language*, Volume 1: *Method and Theory,* 123–156. Stanford, Cal.: Stanford University Press.

Bybee, Joan L., Revere D. Perkins, & William Pagliuca (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Cysouw, Michael (forthcoming). Quantitative methods in typology. In Reinhard Köhler, Gabriel Altmann, & Rajmund G. Piotrowski (eds.)*, Quantitative Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

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

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

— (2003). Significant and non-significant implicational universals. *Linguistic Typology* 7: 108–128.

Gordon, Raymond G. Jr. (ed.) (2005). *Ethnologue: Languages of the World*. Fifteenth Edition. Dallas: SIL International. http:/www.ethnologue.com/

Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Maddieson, Ian (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

— (1991). Testing the universality of phonological generalizations with a phonetically-specified segment database: Results and limitations. *Phonetica* 48: 193–206.

— (1992). The structure of segment sequences. In John J. Ohala, Terrance M. Nearey, Bruce L. Derwing, M. M. Hodge, & Grace E. Wiebe (eds.), *Proceedings of the Second International Conference on Spoken Language Processing, ICSLP-2, Banff, Alberta*, Addendum, 1–4.

— (forthcoming) Interrelationships between measures of phonological complexity. In Maria-Josep Solé, Pam Beddor, & Manjari Ohala (eds.), *Experimental Approaches to Phonology*. Oxford: Oxford University Press.

Maddieson, Ian & Kristin Precoda (1990). Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104–114.

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

— (2003). Samples for comparative grammar: A practical guide. Unpublished manuscript, University of California at Berkeley.

Perkins, Revere (1989). Statistical techniques for determining language sample size. *Studies in Language* 13: 293–315.

— (2001). Sampling procedures and statistical methods. In Martin Haspelmath et al. (eds.), *Language Typology and Language Universals: An International Handbook*, 419–434. Berlin: De Gruyter.

Rijkhoff, Jan & Dik Bakker (1998). Language sampling. *Linguistic Typology* 2: 263–314.

SAS Institute (1992–1998). Statview (Data analysis and presentation system). Cary, N.C.

Shosted, Ryan K. (2006). Correlating complexity: A typological approach. *Linguistic Typology* 10: 1–40.

Stephens, Laurence D. & John S. Justeson (1984). On the relationship between numbers of vowels and consonants in phonological systems. *Linguistics* 22: 531–545.