Studies in Language 17-1. 169-203 (1993).

A METHOD OF LANGUAGE SAMPLING*

JAN RIJKHOFF, DIK BAKKER,
KEES HENGEVELD & PETER KAHREL
University of Amsterdam

ABSTRACT

In recent years more attention is being paid to the quality of language samples in typological work. Without an adequate sampling strategy, samples may suffer from various kinds of bias. In this article we propose a sampling method in which the genetic criterion is taken as the most important: samples created with this method will reflect optimally the diversity of the languages of the world. On the basis of the internal structure of each genetic language tree a measure is computed that reflects the linguistic diversity in the language families represented by these trees. This measure is used to determine how many languages from each phylum should be selected, given any required sample size.

0. Introduction

Recently the relevance of cross-linguistic research has cropped up in several publications (e.g. Hawkins 1988; Hawkins ed. 1988), demonstrating that there are several arguments to motivate a cross-linguistic approach in grammatical theory. The most obvious reason is that a general theory of grammar must provide a framework for all languages and not just for, say, Dutch or English. These are just two manifestations of possible languages, and there is no reason to assume a priori that by studying one or two languages we can account for linguistic phenomena in every other language as well (cf. also Comrie 1978; Comrie 1981b; Mallinson and Blake 1981; Hawkins 1988). Another reason is that certain facts, notably implicational universals, can only be found through cross-linguistic research (Greenberg 1966; Keenan and Comrie 1977; Comrie 1981b: 5-9; Foley 1980; Hawkins 1983; Croft 1990).

To illustrate the last point: if grammatical properties are related by implication (if A, then B), one of four possibilities is logically excluded:[1] Let A be inflection and B derivation; Universal 29 ("If a language has inflection, it always has derivation" — Greenberg 1966: 93) excludes languages with inflection but without derivation (i.e. *A & not B). It does permit languages in (1):

(1) i. with inflection and derivation — A & B
ii. without inflection but with derivation — not A & B
iii. without inflection and derivation — not A & not B

Linguistics must be able to find and account for the distribution of features and the correlations between them, if any. That is, it must explain the occurrence and non-randomness of linguistic facts, and this is usually not possible if the data base consists of one or two languages. Cross-linguistic research is one of the most important ways to find out more about linguistic facts and, indirectly, about the universal system that underlies all natural languages.

In view of the growing interest for cross-linguistic research it may be expected that sampling techniques will play an important role in future studies on language universals. This article puts forward a sampling procedure which contains some original ideas while at the same time retaining valuable insights of earlier proposals. In fact, our procedure leans heavily on ideas put forward in Bell (1978), but wants to provide a more objective way to avoid genetic bias.

Furthermore our procedure is not so much designed to suit statistical or probabilistic purposes (more on this below), but is rather meant to reveal as much as possible the different ways in which languages can give form to a certain meaning (e.g. **negation**) or underlying structure (such as that of the **noun phrase**). Therefore our method recognizes most of all that the sample must display the greatest possible variety, which implies that the universe consists of all natural extant and extinct languages that are presently known; and that at least every phylum (independent language family) must be represented. A weighted procedure is then used to add more languages from the phyla with relatively great internal diversity.

1. Preliminaries

1.1 *Two approaches to language sampling*

Basically there are two approaches to language sampling; which of these is most suitable largely depends on the kind of question one tries to answer. If one is mainly interested in finding tendencies or possible correlations, then the languages in the sample must be independent. In other words, these languages should be unrelated in terms of genetic affiliation, geographic distribution, etc. This is because only in the case of independent units (here: languages) can one make statistically valid generalizations. Perkins (1980) designed a method to construct probability samples, which makes it possible to apply statistical tests and to determine if there are correlations (see also Perkins 1989 and Dryer 1989 on testing statistical claims). However, in view of recent proposals which suggest still larger genetic groupings, resulting in fewer independent language families, it is clear that it will become increasingly difficult to design representative probability samples in which languages are not genetically related.

If, on the other hand, one tries to account for all possible realizations of a certain meaning or structure across languages, like **definiteness** or **relative clause**, then the sample should display the greatest possible diversity. This approach is particularly relevant in the greater context of a theory of grammar. In a variety sample (as opposed to a probability sample) it is very important to have cases of the rarest type, since "exceptional types test the theory" (Perkins 1988: 367). If a general theory of grammar is to be universally valid it has to provide for the grammars of **all** languages, whatever their genetic origin, linguistic type, or geographical location. Thus, for such a theory to be typologically and descriptively adequate it is necessary to explore as much as possible the full range of forms or constructions as they occur in natural languages. In this way one can be reasonably certain that the theory covers all variants and that there are no counterexamples to the rules and principles of that theory. The sampling procedure to be outlined below is appropriate for this second approach: it is designed to maximize the amount of variation in the data in samples of any given size.

1.2 *Genetic bias*

In establishing a language sample there are at least five different kinds of bias that should be avoided: genetic, geographic, typological, cultural, and bibliographic bias. It is probably safe to assume that the quality of a language sample is affected worst if languages are too closely related genetically, as when their common ancestor language was spoken in the not too distant past. The reason why this kind of bias should be avoided most is that it may generate other sources of bias: if languages are closely related in time, chances are that these languages are also related in space (geographically), in type (having inherited, for instance, the basic word order pattern of their common ancestor), and are spoken by people sharing the same kind of culture. Therefore our proposal is mainly an attempt to avoid genetic bias in language samples, although we also recognize that in the actual sampling procedure one should also take into account other kinds of bias (see e.g. Bell 1978; Perkins 1989; Dryer 1989).

Obviously the best way to avoid genetic bias is to make sure that all languages in the sample are from different phyla, but this results in a rather poor sample of less than thirty languages (if we accept Ruhlen's (1987) classification; see below). This would also include the Language Isolates (languages that cannot be related to any other language or language family), each of which can be regarded as constituting a phylum by itself. More often than not, however, cross-linguistic studies aim for more than global results and require larger samples, so that it becomes impossible to avoid including languages that are genetically related. Our proposal specifically relates to this problem.

1.3 *Genetic language classification*

The sampling method to be outlined below is mostly (though not exclusively) concerned with genetic relations among languages. Genetic affiliations are commonly represented in the form of a so-called tree diagram. Figure 1 gives the relations between the small group of Baltic languages, itself part of the independent Indo-Hittite language family (or: phylum).

Such diagrams graphically reflect how entities (here the Baltic languages) are related in time. Historical relations between languages are usually established by comparing lists of vocabulary items (lexical and grammatical) from different languages, and if semantic and phonological similarities are found these are taken as indicative of a common genetic origin.

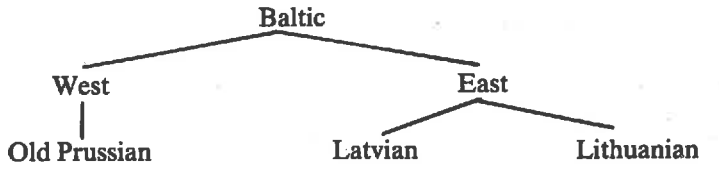

Figure 1. Classification of Baltic languages

1.3.1 *Ruhlen's classification*

To illustrate our sampling procedure we have used the latest classification of languages, which is provided by Ruhlen (1987). As indicated in Table 1, Ruhlen groups all languages in seventeen phyla and recognizes nine languages as Language Isolates. Additionally he mentions thirty-eight Pidgin and Creole languages (which we treat as one phylum), while sixteen languages remain unclassified.[2]

1.3.2 *Genetic classification and stratification*

Perhaps the most important reason why genetic classification is used in language sampling is that it provides us with a good (and — in the absence of serious alternatives — probably the best) category to stratify languages. This means that languages are divided into non-overlapping categories or strata, which in this case are genetic groupings. Stratification not only makes sampling more efficient (it generally allows one to reduce the size of the sample; Perkins 1989: 289f.), it also helps to produce samples with a broad scope in terms of linguistic variation, which is what we aim for here. Furthermore stratification is only possible when all units (languages) are categorized in an exhaustive manner, which is precisely what genetic classification offers (bar the sixteen unclassified languages in Ruhlen's classification which we shall ignore in the rest of this paper).

Some of Ruhlen's genetic (sub)groupings, however, are not uncontroversial and in some cases there is even serious disagreement among the experts.[3] But the fact that at present there is no general consensus on this matter should not keep us from using genetic classifications in a sampling procedure. In this context it may be relevant to refer to Bell (1978: 138), who stated that

Table 1. Phyla according to Ruhlen (1987)

Phylum	extant	extinct	all
Afro-Asiatic	241	17	258
Altaic	63	3	66
Amerind	583	271	854
Australian	170	92	262
Austric	1175	11	1186
Caucasian	38	0	38
Chukchi-Kamchatkan	5	0	5
Elamo-Dravidian	28	1	29
Eskimo-Aleut	9	0	9
Indo-Hittite	144	36	180
Indo-Pacific	731	17	748
Khoisan	31	2	33
Na-Dene	34	7	41
Niger-Kordofanian	1064	4	1068
Nilo-Saharan	138	0	138
Sino-Tibetan	258	10	268
Uralic-Yukaghir	24	3	27
Language Isolates	5	4	9
Pidgins and Creoles	37	1	38
Unclassified lgs.	16	0	16
Totals	4794	479	5273

> [..] stratification [..] is [not] greatly sensitive to the fine details of the partition. Thus, if genetic affiliation is used as a category, the researcher need not agonize about the justification of a particular classification.

In theory other strata categories might be geographic location or linguistic type (or, if necessary, cultural classification; see Perkins 1989). As to the former, it is true that for most languages we also know where they are spoken, but areal stratification is much more problematic (cf. Dryer 1989; Perkins 1989: 303) and consequently rather difficult to incorporate in a (mechanical) sampling procedure. Whereas in genetic classification relations and distances (in time) between languages are more or less fixed, in areal classification one has to deal with migration and variable distances (in space). For instance, distances between language communities in sparsely populated deserts are greater than between language communities in densely populated areas such as Western Europe. This does not mean that one should not take into account geographic information altogether. Even if languages are genetically unrelated they may share certain properties simply because they are spoken in the same area. It is a well known fact that in the course of time neighbouring languages may influence each other and it seems that there are no restrictions as to the kind of linguistic feature that can be borrowed (cf. Mallinson and Blake 1981: 425; Comrie 1988: 86ff.; Matisoff 1990: 109fn.8).[4]

Given the circumstances in which geographic bias may occur, the sample should exclude pairs of languages which (i) are spoken in adjacent regions, (ii) are used in different situations in the same speech community, or (iii) are spoken in the same linguistic area (see note 4). In some cases the only way to avoid possible geographic bias would be to exclude an entire language family. For instance, the languages of the Miao-Yao family (itself a primary branch of the Austric phylum) are spoken in small pockets that are scattered over a large area on the map of South-East Asia. This may make it a difficult task to select other (even genetically unrelated) languages from that region without risking some degree of bias that is due to language contact. Nevertheless, we assume that the quality of a language sample is affected worst if languages are too closely related genetically. For this reason we let the genetic distribution of languages have precedence over the geographic distribution of languages whenever the two are in conflict.

Stratification on the basis of typological properties such as basic word order is simply impossible in the present situation, because this kind of information is only available for the small subsection of the world's languages that have been described in any detail.

1.3.3 *Genetic language trees and linguistic diversity*

So far we have stated several times that our approach is aimed at producing language samples in which differences between individual languages are maximal. As a first step one could select languages that all belong to different phyla, but we saw earlier that such samples can never contain more than twenty-seven languages, including all nine Language Isolates and one Pidgin or Creole language. Thus, if we use a larger sample ($n > 27$) the problem is twofold:

1. How many languages of the same phylum should be included in the sample?
2. How should genetically related sample languages be selected?

Before these questions can be answered (see Section 2), we must devote some attention to the notion **linguistic diversity**. This is because in our view not only must a variety sample contain one representative from each phylum, but also the number of languages in a sample that belong to the same phylum should be proportional to the linguistic diversity in that particular phylum.

Let it be mentioned first of all that linguistic diversity (or linguistic variety) is a relative notion, which only makes sense in relation to the problem under investigation; languages may be very similar in one respect (e.g. basic word order), but totally different in another (e.g. the way participants are coded in the main predicate, if at all). The point is that one can only create the conditions which are likely to create diversity in the sample, but one cannot predict in any detail exactly how the languages are going to be different; at least not if they are chosen at random within the appropriate strata categories (more on this below).

Let us now briefly turn to the question: how can we determine the degree of linguistic diversity in some phylum? Taking into account the considerations on linguistic diversity above, it seems best to have an objective criterion that works for all language families alike.

At first sight one might expect there to be a direct relationship between the degree of linguistic diversity in a phylum on the one hand and the absolute number of languages in that phylum on the other: the more languages a phylum contains, the greater the chance of finding variant forms and structures. However, matters are not as straightforward as that. It is generally assumed that certain small families (e.g. Afro-Asiatic — 258 lgs.) are relatively more diverse than numerically larger groupings (such as the Broad Bantu family — 500 lgs., or the Oceanic family — 426 lgs.; Bell 1978: 146). This suggests that one should not rely too heavily on the absolute number of genetically related languages in an attempt to create maximum linguistic diversity in the sample.

Instead we will exploit the internal structure of the genetic language tree, i.e. the hierarchical structure of the different levels (**generations**), the number of nodes at each of these levels (**parents**), and finally the number of branches under each node (**children**). Thus we do not just take into account the absolute number of languages, but rather the way these languages are

Table 2. Node ratios (nt=non-terminal (=(sub)phylum), pt=preterminal, t=terminal (=language))

Phylum	nt	pt	t	pt/nt	t/nt	t/pt
Afro-Asiatic	153	80	258	0.52	1.69	3.22
Altaic	38	21	66	0.55	1.74	3.14
Amerind	400	258	854	0.64	2.13	3.31
Australian	94	75	262	0.80	2.79	3.49
Austric	484	268	1186	0.55	2.45	4.43
Caucasian	19	10	38	0.53	2.00	3.80
Chukchi-Kamchatkan	4	3	5	0.75	1.25	1.67
Elamo-Dravidian	26	14	29	0.54	1.12	2.07
Eskimo-Aleut	7	4	9	0.57	1,29	2.25
Indo-Hittite	108	68	180	0.63	1.67	2.65
Indo-Pacific	255	161	748	0.63	2.93	4.65
Khoisan	17	13	33	0.76	1.94	2.54
Na-Dene	36	23	41	0.64	1.14	1.78
Niger-Kordofanian	371	236	1068	0.64	2.88	4.53
Nilo-Saharan	89	51	138	0.57	1.55	2.71
Pidgins and Creoles	15	14	38	0.93	2.53	2.71
Sino-Tibetan	100	63	268	0.63	2.68	4.25
Uralic-Yukaghir	14	8	27	0.57	1.93	3.38

historically related. It should be noted in this context that different criteria have been used in establishing phyla and that the internal structure assigned to a phylum is rather dependent on our present knowledge of genetic relations between (groups of) languages, which may vary considerably for each purported genetic grouping (see also section 3.1). Compare in this respect the data in Table 2.[5] Nodes symbolize subgroupings in a genetic language tree (see Figure 1; also Figure 2 below). Preterminal nodes are also nonterminal nodes, of course, the difference being that the former represent the lowest genetic groupings in a phylum.

What strikes us is that the number of languages (t) per non-terminal (nt) and preterminal (pt) node is low for relatively well-explored phyla like Indo-Hittite (ratios 1.67 and 2.65), and rather high for the phyla for which

our knowledge still leaves much to be desired, such as Indo-Pacific (ratios 2.93 and 4.65). On the other hand, the ratio between preterminals and nonterminals seems to be rather stable: for the phyla containing over a hundred languages[6], and disregarding the extremely flat Australian phylum, the figures are between 0.52 and 0.64. For Indo-Pacific and Indo-Hittite the figures are exactly the same. This shows that the differences between node ratios are mainly caused by divisions under the preterminal node (at the level of the lowest genetic subgroupings), and provides support for our view that the level of the individual languages should be left out of consideration in determining the degree of linguistic variation. It also suggests that there is no strict relationship between the mere depth of a genetic language tree and the degree of linguistic diversity within the phylum that is represented by that tree. This latter point is also accounted for in our sampling procedure.

Thus, it will appear that the internal structure of any genetic grouping (as represented in the form of a tree diagram) can be exploited to measure linguistic diversity among genetically related languages. The following section is a detailed description of the sampling procedure.

2. A new method of language sampling

There are two ways to make sure that a language sample is genetically diverse. One is to take into account the variation across language families; the other is to consider the variation within individual phyla (some are more diverse than others). It is these two factors that make up the genetic criterion. Both are captured by the (computerized) sampling method that we propose and which is elaborated below. It can be summarized as follows:

(i) the universe from which the sample is taken contains all known extant and extinct languages
(ii) all phyla are represented in the sample by at least one member;
(iii) additional languages are selected on the basis of the so-called Diversity Value of a phylum;
(iv) the Diversity Value of a phylum is determined on the basis of an objective measure (which replaces Bell's age-criterion; see below).

The inclusion of extinct languages is motivated by the fact that we aim for maximal diversity in the language sample. There is no reason to assume that an extinct language is a less representative instance of a language system than an extant language.

2.1 *Variation across phyla: minimal representation*

The first component of the sampling procedure is relatively simple: every phylum is represented by at least one member. A consequence of this approach is that every Language Isolate, constituting a phylum by itself, will be represented in the sample.[7] This may seem strange, but is fully in line with our basic goal: creating diversity in the sample. Language Isolates are classified as such precisely because they are so different from all other known languages. This approach becomes even more understandable if we consider Language Isolates to be the last surviving members of previously existing larger families. In this connection Comrie (1981a: 238, 261ff.) points at the case of Ket, now listed as a Language Isolate by Ruhlen (1987), but in fact a member of a larger family whose other members (such as Arin, Assan, Kott) became extinct before the end of the last century; see also Nichols (1990: 479) on Burushaski as the sole survivor of a phylum.

There is one drawback in connection with this aspect of our approach. Bybee (1985: 25), when discussing Perkins' (1980) sample, notes that a sample in which all Language Isolates are included tends to suffer from geographic bias. This is true of the samples created with our method too. Nearly all of Ruhlen's Language Isolates (i.e. Basque, Burushaski, Gilyak, Ket, Nahali, Etruscan, Hurrian, Meroitic, and Sumerian) are or were spoken in Eurasia.[8] However, as was stated before (Section 1.3.2), in such cases we let the genetic criterion have precedence over geographic considerations.

2.2 *Variation within phyla: proportional representation*

In our view the number of languages by which a phylum is represented in a sample must correlate proportionally with the linguistic diversity within that phylum (cf. Bell 1978); consequently we need a technique that measures the diversity in a group of genetically related languages.

Our method is based on the assumption that the graph theoretic structure of a genetic language tree reflects the linguistic diversity within the phylum it represents (Section 1.3.3) and consists of the computation of a factor, called Diversity Value (henceforth DV). This method takes into

consideration both the depth and the width of a genetic language tree (see Figure 1; cf. also Ruhlen 1987: ch. 1).

2.2.1 *Preparation of the genetic language trees*

The genetic language trees in Ruhlen (1987) have to undergo some preparatory transformations in order to permit DV computations. For these computations we ignore the top node of the tree, where we find the name of the phylum, as well as the terminal nodes at all bottom levels, where we find the names of the individual languages belonging to that phylum. Top nodes do not add any extra information; and by disregarding all terminal nodes we restrict the influence of the actual number of languages which make up a phylum, shifting the weight entirely to its internal structure. Thus DV is computed over the number of nodes at the intermediate levels between the top node of the tree and the terminal nodes at the bottom end.

To determine the width of a phylum at some intermediate level, we have extended all higher level **preterminal nodes** to the deepest level of the representation of the tree, thus accounting for the fact that every separate branch adds to the width of the phylum as a whole. Figure 2 gives the resulting constellation of intermediate levels of the Uralic-Yukaghir phylum. Hyphens indicate extensions of higher level preterminal nodes. The figures at the bottom line indicate the number of (projected) nodes per level.

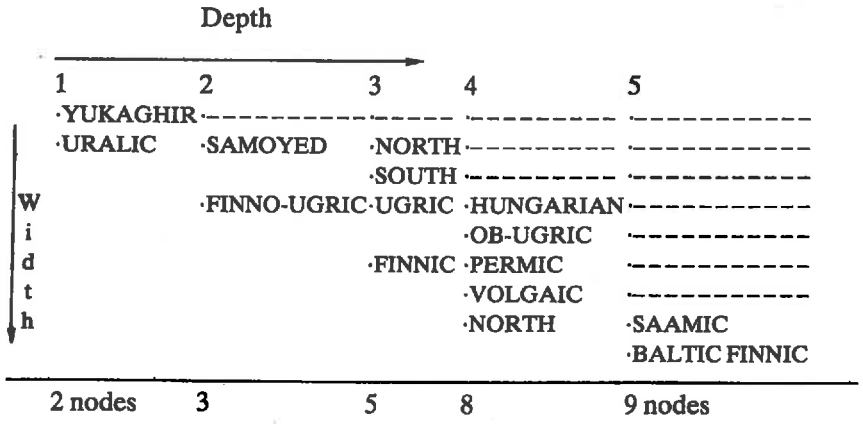

Figure 2. Width of the five intermediate levels of the Uralic-Yukaghir phylum

The width at some level of a genetic language tree is equal to the number of nodes plus the number of preterminal nodes above that level. Uralic-Yukaghir has five levels in between the top node (Uralic-Yukaghir) and the terminal nodes at the deepest level, where we find the individual languages. The width of the Uralic-Yukaghir language tree at level 3 is five; at level 5 it is nine.

In case the set of daughters of some node N consists of both non-terminal and terminal nodes (i.e. groups and individual languages), a phenomenon quite common in Ruhlen (1987), we have inserted an extra preterminal node in between node N and terminal node t. This is done for terminal node t7 in Figure 3 (in which non-terminal node F and terminal node t7 are daughters of C), where Tree I is transformed into Tree I'.

2.2.2 *The computation of DVs*

There are many ways in which the structure of a genetic language tree may be converted into a figure indicating a diversity value. Perhaps the simplest method would be to let DV be the average number of nodes per intermediate level. However, one may assume that high-level splits in the tree are more significant in terms of variation than low-level splits. This is

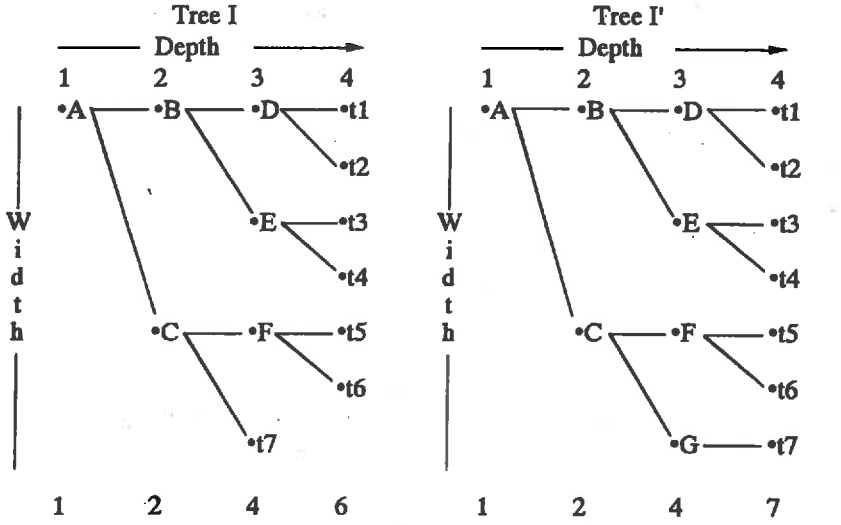

Figure 3. Introduction of a preterminal node (A-G symbolize non-terminal notes; t stands for terminal node).

because high-level splits have occurred earlier in time than low-level splits, which means that two languages separated by a high-level split have had more time to develop into distinct languages than languages separated by a low-level split. So on the assumption that the distinguishing power of levels diminishes when going down the genetic language tree, deeper levels should be assigned less significance in the computation of DV. We decided to let the contribution of the **extra** nodes at each deeper intermediate level decrease by steps of 1/n, where n is the highest number of intermediate levels found in any phylum; in Ruhlen's classification Niger-Kordofanian is the phylum which has most intermediate levels: 16.[9] If n is the maximum number of intermediate levels found in any one phylum, the contribution Cy of intermediate level y with number of nodes Ny is given by the formula

(2) $Cy = Cx + ((n-x)/n * (Ny - Nx))$

where x = y-1. In other words, Cy is obtained by adding the contribution of the level immediately above level y (or: Cx) to the extra nodes of level y (as compared to the number of nodes at level x, i.e. Ny - Nx), the latter being multiplied by a fraction decreasing over the levels according to the row n/n, (n-1)/n, ..., 1/n (i.e. 16/16, 15/16, 14/16, ..., 1/16). By definition, the contribution of the first intermediate level (immediately below the top node) is equal to the number of nodes at that level (C1 = N1). The DV of phylum Z can now be calculated as the mean value of the contributions of all intermediate levels of phylum Z.

Table 3. Computation of DV of the Uralic-Yukaghir phylum (Contr. = Contribution)

Level	Nodes			Contr.
		Cx	+ ((n-x)/n * (Ny - Nx))	= Cy
1	2			2
2	3	2	+ ((16-1)/16 * (3 - 2))	= 2.9375
3	5	2.9375	+ ((16-2)/16 * (5 - 3))	= 4.6875
4	8	4.6875	+ ((16-3)/16 * (8 - 5))	= 7.1250
5	9	7.1250	+ ((16-4)/16 * (9 - 8))	= 7.8750
				24.6250

DV = 24.625 / 5 = 4.925

Table 4. Number of languages per phylum in a weighted and in an unweighted 100-language sample (n=number of languages per phylum, d=difference between weighted and unweighted 100-language sample)

Phylum	n	Unweighted sample	Weighted sample	d
Afro-Asiatic	258	5	6	+1
Altaic	66	1	2	+1
Amerind	854	16	18	+2
Australian	262	5	7	+2
Austric	1186	23	14	−9
Caucasian	38	1	1	0
Chukchi-Kamchatkan	5	0	1	+1
Elamo-Dravidian	29	1	1	0
Eskimo-Aleut	9	0	1	+1
Indo-Hittite	180	3	4	+1
Indo-Pacific	748	14	13	−1
Khoisan	33	1	1	0
Sumerian	1	0	1	+1
Ket	1	0	1	+1
Nahali	1	0	1	+1
Hurrian	1	0	1	+1
Burushaski	1	0	1	+1
Meroitic	1	0	1	+1
Basque	1	0	1	+1
Etruscan	1	0	1	+1
Gilyak	1	0	1	+1
Na-Dene	41	1	1	0
Niger-Kordofanian	1068	20	9	−11
Nilo-Saharan	138	3	5	+2
Pidgins and Creoles	38	1	2	+1
Sino-Tibetan	268	5	4	−1
Uralic-Yukaghir	27	0	1	+1
Totals	5257	100	100	0

To illustrate our sampling procedure, the computation of the DV of the Uralic-Yukaghir phylum is given in Table 3. The DVs arrived at through this method provide us with a straightforward measure to compute the proportional number of languages per phylum that should be in a sample of some predetermined size. It is completely based on the graph-theoretic properties of the genetic tree employed, and not on other, external criteria.

To illustrate the effect we give in Table 4 the number of languages per phylum in a 100-language sample according to this method as compared with the number of languages per phylum if one only takes into account the absolute number of languages. The exact way in which the number of languages per phylum in a sample is computed will be treated below.

Table 4 shows that there are important differences between weighted samples and those based on the absolute number of languages. The most striking difference is that in the case of the weighted sample the large phyla Austric and Niger-Kordofanian have to pay for a stronger representation of the small phyla. This is entirely in line with our major objective, which is to create diversity in the sample.[10]

2.3 *The distribution of sample languages over the phyla*

Once the DVs have been calculated, the question is how these values are to be turned into numbers that indicate how many languages from each phylum a sample should contain.

It will be remembered that at least one language from each phylum should be included in the sample. This implies that there is a lower limit to the sample size, equal to the number of phyla. Following Ruhlen's division and considering each of his Language Isolates as representing a singleton phylum, this lower limit is set at twenty-seven languages. Remaining languages, the number of which depends on the required sample size, are distributed proportionally according to the DVs of the phyla. Due to rounding this procedure may not always lead to the required sample size, but this is made up for by assigning languages to or taking languages from the phyla that divert most from the ideal division, again under the restriction that every phylum is represented.

Applying this procedure to the 100-language sample presented above, we arrive at the final number of languages per phylum in the way indicated in Table 5. In the first phase one language is assigned to each phylum which

Table 5. Computation of languages per phylum in a 100-language sample

Phylum	DV	Phase 1	Phase 2	Phase 3
Afro-Asiatic	55.53	0	6	6
Altaic	14.79	0	2	2
Amerind	178.44	0	18	18
Australian	67.58	0	7	7
Austric	137.41	0	14	14
Caucasian	8.54	0	1	1
Chukchi-Kamchatkan	2.47	1	1	1
Elamo-Dravidian	7.43	1	1	1
Eskimo-Aleut	3.34	1	1	1
Indo-Hittite	39.71	0	4	4
Indo-Pacific	124.79	0	13	13
Khoisan	6.97	1	1	1
Sumerian	0.00	1	1	1
Ket	0.00	1	1	1
Nahali	0.00	1	1	1
Hurrian	0.00	1	1	1
Burushaski	0.00	1	1	1
Meroitic	0.00	1	1	1
Basque	0.00	1	1	1
Etruscan	0.00	1	1	1
Gilyak	0.00	1	1	1
Na-Dene	9.44	0	1	1
Niger-Kordofanian	90.38	0	9	9
Nilo-Saharan	42.18	0	4	→ 5
Pidgins and Creoles	13.47	0	1	→ 2
Sino-Tibetan	38.52	0	4	4
Uralic-Yukaghir	4.93	1	1	1
Totals	845.92	14	98	100

Table 6. Number of languages in samples of different sizes using DV

Phylum	Sample size 30	40	50	60	70	80	90	100	125	150	175	200	225	250
Afro-Asiatic	1	2	2	3	4	5	5	6	8	9	11	12	14	16
Altaic	1	1	1	1	1	1	2	2	2	3	3	3	4	4
Amerind	2	5	7	9	12	14	16	18	24	29	35	40	45	51
Australian	1	2	3	4	4	5	6	7	9	11	13	15	17	19
Austric	2	4	5	7	9	11	12	14	19	23	27	31	35	39
Caucasian	1	1	1	1	1	1	1	1	1	2	2	2	2	3
Chukchi-Kamchatkan	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Elamo-Dravidian	1	1	1	1	1	1	1	1	1	1	1	2	2	2
Eskimo-Aleut	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Indo-Hittite	1	1	2	2	3	3	4	4	5	7	8	9	10	11
Indo-Pacific	2	3	5	7	8	10	11	13	17	20	24	28	32	35
Khoisan	1	1	1	1	1	1	1	1	1	1	1	2	2	2
Sumerian	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ket	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Nahali	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hurrian	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Burushaski	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Meroitic	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Basque	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Etruscan	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Gilyak	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Na-Dene	1	1	1	1	1	1	1	1	1	2	2	2	2	3
Niger-Kordofanian	1	3	4	5	6	7	8	9	12	15	18	20	23	26
Nilo-Saharan	1	1	2	3	3	4	4	5	6	7	8	10	11	12
Pidgins and Creoles	1	1	1	1	1	1	2	2	2	2	3	3	4	4
Sino-Tibetan	1	1	2	2	3	3	4	4	5	6	7	9	10	11
Uralic-Yukaghir	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Totals	30	40	50	60	70	80	90	100	125	150	175	200	225	250

on the basis of its DV alone would not be represented in a 100-language sample (i.e. phyla with DVs lower than 845.92 / 100 = 8.46; cf. Table 5). This is to ensure that all phyla will be represented in the sample. In the second phase remaining languages are assigned to the phyla on the basis of DVs, and the result of this step is added to the outcome of the first phase. In the third phase the effects of rounding are corrected and the sample acquires the intended size.

The technique illustrated in Table 5 requires the size of the sample to be determined first; each sample size requires a separate calculation. In Table 6 we give figures for samples of varying sizes.

2.4 *The selection of genetically related languages*

Samples arrived at by the method described above will usually contain more languages from the same phylum. In order to have a representative contribution from the respective subphyla of these phyla, the number of languages to be assigned to each of the subphyla has to be determined. This is achieved by repeating the procedure outlined above, but with a different value for n, whose value is now determined by the maximum number of intermediate levels found under a **sister** node of the subphylum in question.

Thus, first DVs are computed for all **primary** branches (or: major subphyla; e.g. Uralic and Yukaghir in the Uralic-Yukaghir phylum; see Figure 2) according the formula given in (2). Then fractions of the total number of languages assigned to the phylum are assigned to these subphyla according to DV figures of the major subphyla. This procedure is applied recursively until the situation arises in which there are more subphyla than there are languages to be sampled from these subphyla. At this point a decision has to be taken from which subphyla languages are to be selected. This will be discussed in Section 3.

This recursive application of the procedure is illustrated in Table 7, which gives the representation of the different subphyla of the Amerind phylum in a 250-language sample. In a sample of this size Amerind is represented with 51 languages, to be selected from 854 languages that are distributed over six subphyla. First one language must be assigned to each subphylum which on the basis of its DV alone would not have been represented in a sample of this particular size; in this case none of the subphyla qualify. Then the remaining languages are distributed over the six major subphyla on the basis of their DVs. They are then distributed over the **subphyla** of these subphyla according to the same method. In the case of Andean the procedure is repeated only once: the three languages assigned to this major subphylum can be selected from three of its six subphyla (see also Section 3). In the case of the other five major subphyla the procedure has to be repeated once again. Finally, in the case of Ge-Pano-Carib and Northern Amerind it has to be applied a third time.

Table 7. Representation of the Amerind phylum and its subphyla in a 250-language sample

Amerind (DV=178.44) 51 of 854 languages over 6 subphyla	
Central Amerind (DV=19.05)	6 of 60 over 3 subph.
Oto-Manguean (DV= 9.17)	2 of 19 over 7 subph.
Uto-Aztecan (DV=11.67)	3 of 33 over 8 subph.
Tanoan (DV= 2.88)	1 of 8 over 2 subph.
Ge-Pano-Carib (DV=29.25)	9 of 193 over 2 subph.
Macro-Carib (DV=13.29)	3 of 77 over 5 subph.
Ge-Pano (DV=22.35)	6 of 116 over 2 subph.
Macro-Panoan (DV=12.72)	3 of 72 over 6 subph.
Macro-Ge (DV=16.38)	3 of 44 over 14 subph.
Northern Amerind (DV=45.48)	14 of 232 over 3 subph.
Hokan (DV=17.20)	4 of 43 over 8 subph.
Penutian (DV=21.46)	6 of 92 over 8 subph.
Almosan-Keresiouan (DV=15.98)	4 of 97 over 2 subph.
Almosan (DV=10.08)	2 of 62 over 3 subph.
Keresiouan (DV=10.13)	2 of 35 over 4 subph.
Equatorial-Tucanoan (DV=44.96)	14 of 268 over 2 subph.
Macro-Tucanoan (DV=24.21)	6 of 59 over 19 subph.
Equatorial (DV=30.89)	8 of 209 over 12 subph.
Chibchan-Paezan (DV=16.91)	5 of 71 over 2 subph.
Chibchan (DV=10.89)	2 of 39 over 7 subph.
Paezan (DV=11.67)	3 of 32 over 10 subph.
Andean (DV= 9.50)	3 of 30 over 6 subph.

As has been mentioned earlier, each sample size requires a separate calculation. By way of an example, we give a full specification of a 100-language sample in Table 8.[11] Figures in parentheses refer to: (i) the DV of the (sub)phylum, (ii) the number of primary branches (i.e. subphyla), and (iii) the number of languages in the (sub)phylum. For example, the DV of Afro-Asiatic is 55.53; the phylum has 6 primary branches and contains 258 languages. Branching of the genetic language trees is shown by indentation.

Table 8. A 100-language sample

Afro-Asiatic (55.53/6/258)	6
Chadic (19.18/4/123)	1
Omotic (5.42/2/34)	1
Semitic (7.16/3/31)	1
Cushitic (9.04/2/36)	1
Ancient Egyptian (0.00/0/2)	1
Berber (7.72/3/32)	1
Altaic (14.79/2/66)	2
Altaic Proper (11.03/2/62)	1
Korean-Japanese (3.00/3/4)	1
Amerind (178.44/6/854)	18
Central Amerind (19.05/3/60)	2
Ge-Pano-Carib (29.25/2/193)	3
Macro-Carib (13.29/5/77)	1
Ge-Pano (22.35/2/116)	2
Macro-Panoan (12.72/6/72)	1
Macro-Ge (16.38/14/44)	1
Northern Amerind (45.48/3/232)	5
Hokan (17.20/8/43)	2
Penutian (21.46/8/92)	2
Almosan-Keresiouan (15.98/2/97)	1
Equatorial-Tucanoan (44.96/2/268)	5
Macro-Tucanoan (24.21/19/59)	2
Equatorial (30.89/12/209)	3
Chibchan-Paezan (16.91/2/71)	2
Chibchan (10.89/7/39)	1
Paezan (11.67/10/32)	1
Andean (9.50/6/30)	1
Australian (67.58/30/262)	7
Austric (137.41/3/1186)	14
Austro-Tai (106.03/2/1027)	10
Austronesian (118.17/4/970)	9
Malayo-Polynesian (131.05/2/950)	6
CE Malayo-Polynesian (69.26/2/576)	3
E Malayo-Polynesian (63.67/2/486)	2
Oceanic (64.61/18/430)	1
S Halmahera-NW New G (8.93/2/56)	1
C Malayo-Polynesian (17.36/5/90)	1
W Malayo-Polynesian (78.17/11/374)	3
Paiwanic (5.00/5/14)	1
Tsouic (2.45/2/4)	1
Atayalic (0.00/0/2)	1
Daic (4.67/2/57)	1
Austroasiatic (28.08/2/155)	3
Mon-Khmer (23.33/3/138)	2
Munda (4.29/2/17)	1
Miao-Yao (2.00/2/4)	1
Caucasian (8.54/2/38)	1
Chukchi-Kamchatkan (2.47/2/5)	1

Table 8 continued

Elamo-Dravidian (7.43/2/29)	1
Eskimo-Aleut (3.34/2/9)	1
Indo-Hittite (39.71/2/180)	4
Indo-European (36.94/9/175)	3
Anatolian (4.00/4/5)	1
Indo-Pacific (124.79/13/748)	13
Torricelli (9.50/7/48)	1
Trans-New Guinea (68.97/21/508)	1
West Papuan (5.33/4/24)	1
Sepik-Ramu (15.25/5/98)	1
East Papuan (6.89/3/27)	1
Arai (0.00/0/6)	1
Sko (2.00/2/8)	1
Andaman Islands (2.42/2/13)	1
Geelvink Bay (2.00/2/5)	1
Kwomtari-Baibai (3.00/3/5)	1
Amto-Musian (0.00/0/2)	1
East Bird's Head (2.00/2/3)	1
Tasmanian (0.00/0/1)	1
Khoisan (6.97/3/33)	1
Sumerian (0.00/0/0)	1
Ket (0.00/0/0)	1
Nahali (0.00/0/0)	1
Hurrian (0.00/0/0)	1
Burushaski (0.00/0/0)	1
Meroitic (0.00/0/0)	1
Basque (0.00/0/0)	1
Etruscan (0.00/0/0)	1
Gilyak (0.00/0/0)	1
Na-Dene (9.44/2/41)	1
Niger-Kordofanian (90.38/2/1068)	9
Niger-Congo (90.07/2/1036)	8
Niger-Congo Proper (89.68/2/1007)	7
Central Niger-Congo (91.16/2/961)	6
South C Niger-Congo (51.76/3/755)	3
Eastern (47.73/9/703)	1
Western (7.07/2/47)	1
Ijo-Defaka (2.00/2/5)	1
North C Niger-Congo (49.59/4/206)	3
West Atlantic (10.05/3/46)	1
Mande (9.30/3/29)	1
Kordofanian (9.51/2/32)	1
Nilo-Saharan (42.18/9/138)	5
Pidgins and Creoles (13.47/13/38)	2
Sino-Tibetan (38.52/2/268)	4
Tibeto-Karen (28.71/2/255)	3
Tibeto-Burman (33.58/3/241)	2
Karen (3.67/2/14)	1
Sinitic (4.10/2/13)	1
Uralic-Yukaghir (4.93/2/27)	1

2.5 *Ideal and actual samples*

The figures for the distribution of sample languages over phyla and subphyla presented so far give a representation of ideal samples. There are several reasons why the actual sample may divert from this ideal situation.

The first reason is a practical one: there may be no adequate descriptions of languages from some of the (sub)phyla which should be represented in a sample of a certain size. Since there are, to our knowledge, no adequate descriptions of the extinct Language Isolates Meroitic and Etruscan, the first time we are confronted with this situation is at the first application of our procedure. Although every ideal sample should contain these two Language Isolates, no actual sample will contain them. We prefer not to assign the two vacancies in the sample to other phyla, since this would distort the proportions within the sample, and obscure the fact that the sample is no longer the one that would have been used under ideal circumstances. The same goes for any other gap that might occur in a sample due to bibliographic restrictions.

The second reason why an actual sample may divert from the ideal sample is a side effect of the way DV is computed. The situation may arise that in a large ideal sample a (sub)phylum should be represented by more languages than it actually contains. Since our procedure favours small subphyla and is not based on the actual number of languages in a phylum, but rather on genetic relations, smaller (sub)phyla may get exhausted. This occurs in samples containing more than 1213 languages, which shows that for realistic sample sizes, this phenomenon is no factor of great significance. Furthermore, since the missing languages at stake here are those which have been assigned to (sub)phyla represented exhaustively, these cannot be said to be underrepresented. Again we prefer not to assign the vacancies created by the missing languages to other (sub)phyla.

2.6 *Relation to earlier approaches*

The two steps of the procedure outlined above can be related in an interesting way to two earlier approaches of language sampling, Perkins (1980) and Bell (1978).

Our first step, assigning one language to each phylum, corresponds with the genetic criterion proposed and applied in Perkins (1980), who investigated the relationship between culture and grammar and, in doing

Table 9. Representation of the Amerind phylum in 50-language sample based on different language classifications

	Perkins' method	Our method
Voegelin & Voegelin (1966)	17	≥ 17
Ruhlen (1987)	1	7

so, not only applied a genetic criterion, but also a cultural and a geographic one. He selected one language per phylum, which, since he used Voegelin and Voegelin's (1966) classification, resulted in a sample of 50 languages. Approximately the same sample was used by Bybee (1985).

Against Perkins' (1980) approach it may be objected that it does not take into account the internal diversity of phyla, and thus is more dependent on the state of the art in language classification than our method. This is demonstrated by the figures in Table 9. It must be added that this is not really a fair comparison, in that Perkins primarily aims at a probability sample rather than a variety sample (see Section 1).

Ruhlen (1987) has an Amerind phylum which comprises languages from what Voegelin and Voegelin (1966) treated as seventeen different phyla. If Ruhlen's (1987) classification had been available to Perkins and if he had used this classification rather than Voegelin and Voegelin (1966), his sample would have contained one Amerindian language rather than the seventeen that it actually contains. On the other hand, if we had applied our method to Voegelin and Voegelin (1966) rather than to Ruhlen (1987), a 50-language sample would have contained at least seventeen languages, rather than the seven languages that it actually contains. Both methods are dependent on the particular classification used, but our method mitigates the effects of new insights in this area.

It appears that our DV figures correspond to a large extent with Bell's (1978) age-criterion. Bell's important paper addresses the problem of homogeneity within phyla; we already mentioned that on the whole languages from large and deep phyla (e.g. Niger-Kordofanian, Austric) tend to be rather homogeneous. To tackle this problem Bell arbitrarily chose 3500 years divergence as a breakpoint to determine language groups within phyla. The size of the groups ranges from one language (Language Isolates)

Table 10. Bell's language groups

Phylum	Languages	Groups	Average
Ket	1	1	1.0
Burushaski	1	1	1.0
Khoisan	20	5	4.0
Eurasiatic	70	13	5.4
Nilo-Saharan	100	18	5.5
Amerind	900	est. 150	6.0
Indo-Pacific	700	est. 100	7.0
Australian	200	ca. 27	7.4
Indo-European	90	12	7.5
Na-Dene	30	4	7.5
Ibero-Caucasian	35	4	8.7
Afroasiatic	200	23	8.7
Sino-Tibetan	250	ca. 20	12.5
Austric	800	ca. 55	14.5
Dravidian	20	1	20.0
Niger-Kordofanian	900	44	20.5
Totals	ca. 4300	478	9

to 300 (Bantu subphylum; Niger-Kordofanian); in total Bell distinguished 478 groups. Table 10 (*ibid.* p.148) gives in the first column the major families according to Bell (1978); the second column states the estimated number of languages in each family; the third column states for every family the estimated number of groups separated by 3500 years or more. For the sake of clarity, we have added a fourth column, which gives the average number of languages per group, and we have ordered the phyla on the basis of the latter in ascending order.

Bell estimated that there are about 4300 languages distributed over 478 groups separated by 3500 years or more; the average number of languages in these groups is nine. Compare Niger-Kordofanian with Amerind, both of which have 900 languages; Niger-Kordofanian has 44 groups separated by Bell's criterion (average: 20.5), Amerind has 150 (average 6.0). From this it can be concluded that Amerind is genetically more diverse than Niger-Kor-

dofanian. On the whole Bell concludes on the basis of Table 10 that Dravidian, Niger-Kordofanian and Austric are less diverse than average and that Eurasiatic, Nilo-Saharan and Khoisan are more diverse than average.

It is interesting to note that our DV calculations show more or less the same tendencies. In Table 11 we give Ruhlen's (1987) phyla in the first column, the number of languages in these phyla in the second column, and their DV in the third column. The fourth column gives the number of languages per DV-unit. Again the phyla are ordered in ascending order according to the figures in the fourth column. Notice that we have omitted the Language Isolates in Table 11, since their DVs are all 0.00.

If one compares the outcome of Bell's estimates with our DV calculations, as represented in Tables 10 and 11, respectively, the overall tendencies seem to coincide, in so far as the genetic groupings can be compared. Particularly striking is that in both tables Austric, Niger-Kordofanian and Sino-Tibetan show up as not very diverse, and Nilo-Saharan as highly diverse (cf. also Table 4). Bell's not very diverse Dravidian in Table 10 has been replaced in Ruhlen's classification by the more complex Elamo-Dravidian phylum, which is reflected in a higher position in Table 11. The high position of Pidgin and Creole languages in Table 11 is due to the fact that it is not a genetically based grouping.

Table 11. Genetic diversity of phyla according to DV

Phylum	Languages	DV	Average
Chukchi-Kamchatkan	5	2.47	2.025
Eskimo-Aleut	9	3.34	2.692
Pidgins & Creole	38	13.47	2.821
Nilo-Saharan	138	42.18	3.272
Australian	262	67.58	3.877
Elamo-Dravidian	29	7.43	3.903
Na-Dene	41	9.44	4.344
Caucasian	38	8.54	4.449
Indo-Hittite	180	39.71	4.533
Afro-Asiatic	258	55.53	4.646
Altaic	66	14.79	4.646
Khoisan	33	6.97	4.735
Amerind	854	178.44	4.786
Uralic-Yukaghir	27	4.93	5.482
Indo-Pacific	748	124.79	5.994
Sino-Tibetan	268	38.52	6.957
Austric	1186	137.41	8.631
Niger-Kordofanian	1068	90.38	11.817
Totals	5248	845.92	4.978

Bell used the number of groups within a phylum to determine whether families are overrepresented or underrepresented in a sample, and he illustrated this by two example samples, which are given in Table 12 (*ibid.* p.149). Sample A is a random sample drawn from the list of languages in Voegelin and Voegelin (1966) and sample B is stratified by the sixteen families; the number of languages selected in each family is proportional to the number of groups in each family.

Table 12. Comparison of samples (Bell 1978)

Family	Groups	A	B
Dravidian	1	0	0
Eurasiatic	13	0	1
Indo-European	12	1	1
Nilo-Saharan	18	3	1
Niger-Kordofanian	44	8	3
Afroasiatic	23	0	1
Khoisan	5	0	0
Amerind	150	8	9
Na-Dene	4	0	1
Austric	55	8	4
Indo-Pacific	100	1	6
Australian	27	0	2
Sino-Tibetan	20	1	1
Ibero-Caucasian	4	0	0
Ket	1	0	0
Burushaski	1	0	0
Totals	478	30	30

It appears that there are five errors in sample A: Niger-Kordofanian, Nilo-Saharan and Austric are overrepresented, while Indo-Pacific and Australian are underrepresented.

The merit of Bell's approach is that it recognizes genetic diversity among the languages of the world, but a drawback is that groups should be separated by 3500 years. This criterion is completely arbitrary (readily admitted by Bell himself), but perhaps even more problematic is that it seems very difficult to assess the number and size of language groups on the basis of this criterion for all phyla. The history of a language family is well recorded or reconstructible in the case of the European languages, for instance, whose history is relatively well-known. The history of many phyla, however, is much more obscure, and any estimate as to the internal diversity on the basis of time-depth alone is mere conjecture. Indeed, Bell (1978: 147) mentions that, due to the absence of sufficient material, his estimates for New Guinea and South America are only guesses. Our DV calculations can be seen as an objectivization of Bell's language groups, and as such represent an objective measure which replaces his age criterion.

Another undesirable feature of Bell's approach is that not all language families are represented in the sample, since the number of language groups in each family is the main criterion. This approach leads to a situation in which small phyla such as Caucasian, Chukchi-Kamchatkan and Khoisan and all Language Isolates will never be represented in smaller samples. This second objection is countered in our approach by means of our first step, which ensures the presence of all phyla in a sample.

In summary, one may say that the problems posed by Perkins' approach are solved by Bell's approach and vice versa. Our sampling method combines the positive elements of both approaches. That is, the first aspect of the genetic criterion (minimal representation) takes into account variation across phyla, as does Perkins' sampling technique; the second aspect (proportional representation) accounts for variation within phyla, as does Bell's method.

3. Sample design

The method of determing the composition of a sample of a predetermined size presented so far can be used whatever the size of the sample and the way in which sample languages are selected. In this section we address some issues of sample design in so far as they are related to the application of our method.

The sampling method described in Section 2 can be applied randomly or non-randomly. In a non-random procedure languages are selected by the investigator, obeying the distribution of languages over the (sub)phyla as established on the basis of our method (Section 2). Using this procedure one probably chooses the languages on the basis of the availability and quality of language descriptions.

In a random procedure two phases need to be taken into consideration. Obviously, a random procedure should be used to select sample languages from (sub)phyla once it has been established which (sub)phyla are to be represented in a sample. But apart from that, there may be situations in which the (sub)phyla from which the sample languages are to be chosen have to be selected randomly first.

This situation arises if the number of daughters of a (sub)phylum M is larger than the number of languages assigned to M. To give an example, as indicated in Table 7, the Andean subphylum of the Amerind phylum has to be represented with three members in a 250-language sample. These three members are to be selected from the six subphyla of Andean. In order to decide from which three of these six subphyla the sample languages should be chosen, and provided that every language to be selected from M must belong to a different daughter, there are, in principle, three options:

(i) arrange the daughters according to their DVs and first choose the daughters with the highest DVs;
(ii) randomly select the required number of daughters of M, irrespective of their DVs;
(iii) randomly select the appropriate number of daughters of M, while taking into account their DVs.

Method (i) would systematically exclude languages from the daughters with low DVs in all samples not large enough to include all daughters of M. Method (ii) would give languages in very small subphyla a far greater chance of being selected than those belonging to relatively large subphyla. This means that in the long run languages of the very small subphyla would be overrepresented in samples below a certain size. We therefore adopt method (iii). Thus we randomly select the appropriate number of daughters of M, while using their respective DVs as a weighting factor. Consequently the chance of a language being selected from a particular subfamily of M is proportional to the DV of that subfamily. See for example Table 13, which gives the three Andean languages randomly selected according to this pro-

Table 13. Distribution of languages over the subphyla of Andean (Amerind) in a random 250-language sample

Andean (9.50/6/30)	3
Northern (5.00/5/7)	1: Catacao
Southern (4.00/4/9)	1: Tehuelche
Aymaran (0.00/0/2)	1: Jaqaru

cedure: Catacao, Tehuelche, and Jaqaru. This particular selection does not include a representative of the Cahuapanan-Zaparoan subphylum (2.00/2/7), although this grouping had a better chance of being represented than the Aymaran subphylum (0.00/0/2), which does have a language in the sample. The Urarina-Waorani (0.00/0/3) and Quechuan (0.00/0/2) subphyla are not represented in this sample either.

In order to counter bibliographic problems created by this random procedure one might stipulate that only those languages for which an adequate description exists may be selected (cf. also Perkins 1980). This would, of course, require an extension of the database with bibliographic information. In order to counter geographic problems created by this procedure, every randomly selected sample should be checked on its consistency with the geographic criterion (see 1.3.2).

4. Conclusion

In this paper we presented a method to create language samples in which the genetic distance between individual languages is maximal.

The method takes into account genetic diversity both across and within phyla. So as to account for the variation across phyla, every sample contains at least one representative from each phylum. In order to account for variation within phyla additional sample languages are assigned to phyla on the basis of their Diversity Value, an objective measure based on the internal structure of the genetic language tree.

The recursive application of the method ensures that the genetic criterion is met at all relevant levels of the genetic language tree.

Authors' adresses:

Jan Rijkhoff, Peter Kahrel
Dept. of General Linguistics
Spuistraat 210
1012 VT Amsterdam
The Netherlands

Dik Bakker
Dept. of Computational Linguistics
Spuistraat 134
1012 VB Amsterdam
The Netherlands

Kees Hengeveld
Dept. of Spanish
Spuistraat 134
1012 VB Amsterdam
The Netherlands

(Received December 12th, 1991)

NOTES

* Thanks are due to Keith Allan, Peter Bakker, Matthew S. Dryer, Martin Haspelmath, Revere D. Perkins and Rob Schoonen for useful comments on an earlier draft. The usual disclaimers apply.

1. Some authors have argued, however, that it would be more realistic to avoid committing oneself to exceptionless universals and speak only of cross-linguistic generalizations instead. Apart from the fact that the widespread occurrence of some phenomenon must be accounted for too, strong tendencies (or: statistical universals) tend to be more interesting than absolute universals (see Mallinson and Blake 1981: 8).

2. The list of languages is still growing. The 1988 edition of Grimes' *Ethnologue: languages of the world* contains entries for 6170 languages, 725 more than the 1984 edition.

3. This probably holds especially for Ruhlen's classification of the indigenous languages of the Americas, which is adopted from Greenberg 1987 (cf. Campbell 1988 and Greenberg's reply, Greenberg 1989; also Adelaar 1989; Kaufman 1990; Matisoff 1990; Nichols 1990; see also e.g. Foley (1986: 3), who denies that there is an Indo-Pacific phylum). But then, as Blake (1988: 261) noted, "[..] naturally any worldwide genetic classification will be controversial, and anyone who publishes one puts themselves in the unenviable position of being the target of expert criticism from every corner of the globe". In any case, since our proposal does not crucially hinge on any genetic classification in particular, it can easily be adapted to other proposals.

4. Language contact may not only be due to social and commercial relations between members of neighbouring language communities, but may also occur because another language was or continues to be used for prestigious, political, religious or scientific purposes (see for instance Okell 1965, Kahane 1986).

   The danger of geographic bias is particularly great in a linguistic area, a region where languages share non-trivial features that do not occur immediately outside that area

(Campbell et al. 1986: 536; Mallinson and Blake 1981: 17). A good example is the absence of an infinitive in the Balkan-languages (see Joseph 1983; Schaller 1975). Apart from the well known Balkan Sprachbund at least the following regions have been mentioned as constituting a linguistic area: India/South Asia (Emeneau 1956; Masica 1976), Ethiopia (Ferguson 1976), Meso-America (Campbell et al. 1986), Arnhem-Land and other parts of Australia (Heath 1978; Dixon 1980: 238-251), New Guinea (Foley 1986: 25-9), the Western Amazon (Payne 1987: 21). The uncertainty about the number and location of linguistic areas makes geographic bias difficult to control for.

5. For our sampling procedure we had to slightly modify the format of the genetic language trees as given in Ruhlen (1987); this is explained in Section 2.2.1. Figures in Table 2 are based on the modified trees.

6. Smaller phyla have too little internal structure to allow for reliable calculations.

7. Alternatively, one could, as in the case of Pidgins and Creole languages, create an artificial phylum containing all language isolates and select one isolate from this phylum. Our sampling method could be applied using such a classification, but we prefer the more principled approach defended here.

8. We are grateful to Matthew Dryer for drawing our attention to this fact.

9. It appears that if we were to consider **all** nodes (rather than just the extra nodes) at the intermediate levels, this would result in samples containing a disproportionally high number of languages from relatively small (sub)phyla. For instance, in a 250-language sample, Indo-Hittite would be represented by 9 Indo-European and 2 Anatolian languages if all nodes would be taken into account. The Indo-European branch contains 175 languages, but the Anatolian branch includes only 5 languages.

   The fact that we take into consideration the maximum number of levels found in any phylum has the additional advantage that corresponding levels in each phylum are treated alike.

10. But note that this effect is due not to the size of Austric and Niger-Kordofanian, but to their relative homogeneity.

    A large-scale evaluation of our method by computer shows that the effect of stratification on the basis of DV figures is remarkably constant. Small samples based on the absolute number of languages divert some 50 percent of the ideal distribution following our method. Samples of over a 100 languages divert around 40 percent. This percentage is rather constant for samples up to 1500 languages, and only slowly declines above that number.

11. Full specifications of samples of other sizes are provided by the authors on request.

REFERENCES

Adelaar, Willem F.H. 1989. Review of Joseph H. Greenberg, *Language in the Americas*. *Lingua* 78: 249-255.

Bell, Alan. 1978. "Language samples". In Greenberg, Joseph H. (ed.) 123-156.

Bender, M.L.; Bowen, J.D.; Cooper, R.L.; and Ferguson, C.A. (eds). 1976. *Language in Ethiopia*. London: Oxford University Press.

Blake, Barry. 1988. Review of Merritt Ruhlen, *A guide to the world's languages. Vol. I: Classification*. *Journal of Linguistics* 24-1: 261-262.

Bybee, Joan L. 1985. *Morphology: a study of the relation between meaning and form*. Amsterdam: Benjamins.

Campbell, Lyle. 1988. Review of Joseph H. Greenberg, *Language in the Americas*. *Language* 64-3: 591-615.

Campbell, Lyle; Kaufman, Terrence; and Stark-Smith, Thomas C. 1986. "Meso-America as a linguistic area". *Language* 62-3: 530-570.

Comrie, Bernard. 1978. "Linguistics is about languages". *Studies in the Linguistic Sciences* 8-2: 221-236.

Comrie, Bernard. 1981a. *The languages of the Soviet Union*. Cambridge: Cambridge University Press.

Comrie, Bernard. 1981b. *Language universals and linguistic typology: syntax and morphology*. Oxford: Blackwell (2nd printing 1983).

Comrie, Bernard. 1988. "Genetic classification, contact, and variation". In Walsh, Thomas J. (ed.), 81-93.

Croft, William. 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Dixon, Robert M.W. 1980. *The languages of Australia*. Cambridge: Cambridge University Press.

Dryer, Matthew S. 1989. "Large linguistic areas and language sampling". *Studies in Language* 13-2: 257-292.

Emeneau, Murray B. 1956. "India as a linguistic area". *Language* 32-1: 3-16.

Ferguson, Charles. 1976. "The Ethiopian language area". In Bender et al. (eds), 63-76.

Foley, William A. 1980. "Toward a universal typology of the noun phrase". *Studies in Language* 4-2: 171-199.

Foley, William A. 1986. *The Papuan languages of New Guinea*. Cambridge: Cambridge University Press.

Greenberg, Joseph H. 1966. "Some universals of grammar with particular reference to the order of meaningful elements". In Greenberg, Joseph H. (ed.), 73-113.

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

Greenberg, Joseph H. 1989. "Classification of American Indian languages: A reply to Campbell". *Language* 65-1: 107-114.

Greenberg, Joseph H. (ed.). 1966. *Universals of language*. Cambridge, Mass.: The MIT Press.

Greenberg, Joseph H. (ed.). 1978. *Universals of human language. Volume I: Method & Theory*. Stanford, Ca.: Stanford University Press.

Grimes, Barbara F. (ed.). 1988. *Ethnologue: languages of the world (plus supplement: Ethnologue index)*. Dallas, Texas: Summer Institute of Linguistics. (11th edition.)

Hammond, M.T.; Moravcsik, E.A.; and Wirth, J.R. (eds). 1988. *Studies in syntactic typology*. Amsterdam: Benjamins.

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

Hawkins, John A. 1988. "On generative and typological approaches to Universal Grammar". *Lingua* 74: 85-100.

Hawkins, John A. (ed.). 1988. *Explaining language universals*. Oxford: Basil Blackwell.

Heath, Jeffrey. 1978. *Linguistic diffusion in Arnhem Land*. Canberra: Australian Institute of Aboriginal Studies.

Joseph, B.D. 1983. *The synchrony and diachrony of the Balkan infinitive*. Cambridge: Cambridge University Press.

Kahane, Henry. 1986. "A typology of the prestige language". *Language* 62-3: 495-508.

Kaufman, Terrence. 1990. "Language history in South America: what we know and how to know more". In Payne, Doris L. (ed.), 13-73.

Keenan, Edward; and Comrie, Bernard. 1977. "Noun phrase accessibility and universal grammar". *Linguistic Inquiry* 8: 63-99.

Mallinson, Graham; and Blake, Barry J. 1981. *Language typology: cross-linguistic studies in syntax*. Amsterdam: North-Holland.

Masica, Colin P. 1976. *Defining a linguistic are: South Asia*. Chicago: University of Chicago Press.

Matisoff, James A. 1990. "On megalocomparison". *Language* 66-1: 106-120.

Nichols, Johanna. 1990. "Linguistic diversity and the first settlement of the New World". *Language* 66-3: 475-521.

Okell, John. 1965. "Nissaya Burmese: a case of systematic adaptation to a foreign grammar and syntax". *Lingua* 15: 186-227.

Payne, Doris L. 1987. "Noun classification in the Western Amazon". *Language Sciences* 9-1: 21-44.

Payne, Doris L. (ed.). 1990. *Amazonian linguistics: studies in lowland South American languages*. Austin: University of Texas Press.

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

Perkins, Revere D. 1988. "The covariation of culture and grammar". In Hammond et al. (eds), 359-378.

Perkins, Revere D. 1989. "Statistical techniques for determining language sample size". *Studies in Language* 13-2: 293-315.

Ruhlen, Merritt. 1987. *A guide to the world's languages. Vol. 1: Classification*. London: Edward Arnold.

Schaller, Helmut Wilhelm. 1975. *Die Balkansprachen: eine Einführung in die Balkanphilologie*. Heidelberg: Carl Winter.

Voegelin, Charles F.; and Voegelin, Florence M. 1966. "Index to languages of the world". *Anthropological Linguistics* Vol. 8, nos. 6-7.

Walsh, Thomas J. (ed.). 1988. *Synchronic and diachronic approaches to linguistic variation and change. Georgetown University Round Table on Language and Linguistics 1988*. Washington D.C.: Georgetown University Press.