Language sampling

JAN RIJKHOFF and DIK BAKKER

Abstract

This article has two aims: first, to present a revised version of the sampling method that was originally proposed in 1993 by Rijkhoff, Bakker, Hengeveld, and Kahrel, and second, to discuss a number of other approaches to language sampling in the light of our own method. We will also demonstrate how our sampling method is used with different genetic classifications (Voegelin & Voegelin 1977; Ruhlen 1987; Grimes (ed.) 1997) and argue that—on the whole—our sampling technique compares favourably with other methods, especially in the case of exploratory research.

Keywords: Diversity Value, genetic classification, sampling

1. Introduction

The last few decades have seen a significant rise in typological studies that are based on data from a more or less representative language sample, especially after the publication of Greenberg's (1963) seminal article on language universals—including Greenberg et al. (eds.) (1978), Hawkins (1983), Dryer (1992), Stassen (1985), Nichols (1992) (and before Greenberg 1963 see, e.g., Forchheimer 1953). However, despite the growing tendency to consider data from a wide range of languages, most typologists have paid little attention to sampling techniques; some notable exceptions are Bell (1978), Perkins (1980, 1989), Bybee (1985), Dryer (1989, 1992), Nichols (1992), Bybee, Perkins, & Pagliuca (1994), and Whaley (1997: 36–43).

This article has two aims. In the first place it seeks to contribute to the issue of language sampling by showing that the sampling method proposed earlier by Rijkhoff, Bakker, Hengeveld, & Kahrel (1993) can be used with different genetic language classifications, thus eliminating the misconception that it only works in tandem with Ruhlen's classification of the world's languages (Ruhlen 1987). We will first apply our method to

1430–0532/1998/002–263
© Walter de Gruyter

Voegelin & Voegelin's classification (Voegelin & Voegelin 1977) and then to the classification used in the 13th and most recent edition of the *Ethnologue* (Grimes (ed.) 1997).[1]

In the second place we will discuss some alternative approaches to language sampling in the light of the method explained here. We will compare the samples in Bybee et al. (1994), Dryer (1989, 1992), Stassen (1997), and Nichols (1992) to samples of the same size but constructed according to our method. We will make these comparisons on the basis of a set of criteria that we believe any sampling method should meet. To the extent that the various methods can be compared, we will argue that our method is preferable in a number of ways, for example, because it systematically covers all language families and subfamilies, whatever the size of the sample. Another important feature of our method is that it produces samples in which the genetic (historical) distance between languages is always maximal and in which languages are distributed proportionally across families and subfamilies on the basis of linguistic variety rather than of quantitative considerations such as the number of languages that belong to a certain (sub)family. Not only does our method force the typologist to take into account data from a wide range of languages, thus ensuring that the results are based on a representative language sample, it also provides a reliable starting point for those who first want to establish a sound cross-linguistic basis before expanding their sample to investigate certain questions in greater detail. Finally we will argue that, in contrast to other techniques, our method is completely transparent and guarantees that all samples are constructed in exactly the same way, which makes it possible to test and compare the validity of the results.

Before applying our method to some of the better-known language classifications (Section 3) and comparing it to other sampling methods used in typological research (Section 4), we will first sketch its main features.

2. The DV sampling method

2.1. *Types of samples and kinds of questions*

Samples that are constructed according to our sampling method are best characterized as VARIETY SAMPLES. Two other types of samples that are commonly distinguished are PROBABILITY SAMPLES and RANDOM SAMPLES. Which of these types of samples is most suitable depends on the kind of questions that one seeks to answer, or rather, on the type of linguistic statements that one wishes to make. For example, if one is interested in correlation pairs or in the probability of occurrence of some linguistic phenomenon, then one wants to have a probability sample, i.e.,

a sample that is free of any kind of bias. In establishing such a sample it is especially important to avoid the inclusion of languages that are genetically related, since they may share certain properties of the ancestor language. Other known forms of bias to be avoided are, for example, cultural bias (due to the fact that languages are spoken by people sharing the same cultural traits which may be reflected in the language; cf. Perkins 1992) and areal bias (as when the results are influenced by language contact).

The requirement that the languages be independent units makes it rather difficult to construct a good probability sample. For example, even in a relatively small sample it is practically impossible to avoid the inclusion of languages that are not somehow genetically related or spoken in the same region.[2] Several attempts have been made to deal with this problem (Perkins 1980; Dryer 1989; Nichols 1992), but basically there are only two ways out. Either a small sample is used which, however, is not quite representative with respect to genetic, areal, and/or cultural diversity (cf. Perkins 1980), or a large sample is used and genetic, areal, and/or cultural relationships are manipulated so as to meet the requirements of statistical tests (e.g., Dryer 1992: 83). Essentially, however, there does not seem to be a real solution.

Sometimes the term "probability sample" is also used in a more general way to refer to a sample that does not take into account any form of (genetic, typological, geographic, cultural, etc.) stratification. Such samples can nevertheless be used to formulate statistical generalizations regarding the frequency of occurrence of some linguistic phenomenon.[3] To avoid confusion, we will call such samples "random samples" and reserve the term "probability sample" for the kind of sample described above. Depending on the degree of (known, expected, or anticipated) variation displayed by the linguistic phenomenon in question, a random sample must be relatively large in size to be able to produce reliable results.

The samples produced by the method we developed are best characterized as "variety samples" in that the method is designed to generate samples that display the greatest possible linguistic variety. Variety samples are mainly used for explorative research: when little is known about the form or construction under investigation it is important that the sample offers a maximum degree of variation for the linguistic parameters involved.

Apart from the main criterion, i.e., the nature of the questions to be answered, several other factors may play a role in determining the best sampling strategy. To give an example, in the case of a well-explored linguistic domain, existing knowledge can be very helpful in determining the nature of the sample. Thus, if certain linguistic variables are known

to interact with the research variable—e.g., certain syntactic and morphological characteristics on the one hand and the existence of serial verb constructions on the other—then samples may be stratified for the former.

Other factors that can play a role in constructing a sample are the stability (or conservativeness) of the relevant variables and the ease with which they can be borrowed (cf. Bakker 1997). When a variable is relatively time-resistant, such as morphological type, genetic considerations are more crucial than when variables are involved that tend to change relatively easily, such as certain word order patterns. Note also that any kind of sample may suffer from bibliographic bias, i.e., the tendency to select languages for which good descriptions are available. Only a minority of the world's languages have been studied thoroughly, and that subset, taken as a sample, violates any of the above criteria. This fact alone is likely to put constraints on any sampling method.

Another topic that we want to address briefly is the size of a sample. One of the few typologists who seriously addressed the issue of sample size is Perkins (1989). On the basis of a large set of languages with known scores for a certain linguistic variable, Perkins applied the $tau_{R|C}$ test in order to establish to what extent these scores were independent vis à vis some areal or genetic grid. Subsequently the maximum size was determined for a subset of the languages in a sample, which were more or less independent as regards the specific variable under consideration. For this purpose he examined two samples: (i) the 317-language sample in Maddieson (1984), which is concerned with segmental phonemes, and (ii) the 1,063-language sample that is used in Tomlin's study about the relative order of Subject, Object, and Verb (Tomlin 1986). Depending on the kind of research question and taking into account different areal and genetic classifications, Perkins (1989: 312) then recommends "[u]sing around a hundred languages for most linguistic samples to balance the requirements for representativeness and independence in samples. The results from using samples of this size should always be checked, however, to determine if the variables under consideration significantly vary across languages". (See also Section 4.2 below, which deals with the sample in Bybee, Perkins, & Pagliuca 1994.)

Since Perkins' test (which can of course only be applied post hoc) requires languages to be independent, it is clear that his method is less relevant for variety samples than for probability samples. Notice also that he only tested samples which were used to investigate one variable (such as the relative order of Subject, Object, and Verb). When multiple variables are involved one usually needs a (much) larger sample, especially if one is also interested in the interaction between different variables.

Since the above has no immediate relevance for the kind of sample that is generated by our sampling method, we have developed a procedure

which calculates the theoretical minimum sample size, given some variable or set of variables, their possible or expected values, and the distribution of these values (if an estimate is available). For example, the variable Basic Word Order has six possible values: SOV, SVO, VSO, VOS, OVS, and OSV. On the assumption that all values have the same chance of occurrence in a random sample, we get the figures of Table 1.

This is just an example, of course, but it is important to note that there are probably few linguistic phenomena that have an equal chance of occurrence. For instance, if we assume that Tomlin's study gives us the actual distribution of basic word order patterns, our procedure yields the sample sizes given in Table 2. Thus, only 50% of the (random) samples containing 350 languages will include all six basic word order patterns. And in order to have a 95% chance of having all six patterns, including the rare OSV order, in a random sample it should contain over 1000 languages. In the case of relatively unexplored variables, whose possible values and their distribution have not been studied in great detail, one can just hope that the sample size that is chosen is large enough to display the total range of variant forms and patterns that exist in the languages of the world. Only a method that steers a sample in the direction of maximum variety may set a manageable limit to its size.

The sampling method discussed in this paper is designed first and foremost to produce variety samples for exploratory linguistic research. Since languages are assigned to different genetic groups precisely because

Table 1. *Minimum sample size when all six values have the same chance of occurrence*

Chance of occurrence of all six values in sample	Minimum sample size
.5	13
.75	17
.90	23
.95	26
.99	37
.999	43

Table 2. *Minimum sample size when all (six) values do not have the same chance of occurrence*

Chance of occurrence of all six values in sample	Minimum sample size
.5	350
.75	610
.90	830
.95	>1000

they differ on a number of linguistic parameters, we have developed a method in which genetic diversity is given precedence over any of the other parameters along which languages can be stratified and which can be a source of bias (such as geographic location). The method provides for maximal genetic distance between languages, irrespective of sample size and classification chosen. This is achieved in two steps. First we take into account linguistic variation ACROSS independent language families (i.e., across all maximal genetic groupings or phyla); then we consider variation WITHIN each individual language family. This is explained in Sections 2.2 and 2.3 respectively.

2.2. *Linguistic variation across language families*

In order to account for linguistic variation across independent genetic groupings our sampling method requires that a sample must (ideally) contain at least one representative from each independent language family. This requirement does not pose a problem when the number of languages in the sample equals or exceeds the number of language families. For instance, Ruhlen (1987) recognizes only 27 such families, so that a sample of 30 languages would already be fully representative with respect to the genetic diversity at the level of language families. This is shown in Table 3. Notice that Pidgin and Creole languages are treated as one family here, but that each of the nine Language Isolates is regarded as an independent language family (i.e., as a so-called singleton phylum); unclassified languages will be ignored in all three classifications that we are concerned with here.

However, when the number of language families exceeds the number of languages in a sample, it is necessary to have a procedure to determine which families will and which will not be represented in the sample. This problem is likely to occur, for instance, when one uses a more conservative genetic classification of the world's languages, such as those of Voegelin & Voegelin (1977) or of the *Ethnologue* (Grimes (ed.) 1997), counting some 50 families or over 120 families respectively.[4] We will return to the classifications by Voegelin & Voegelin (1977) and Grimes (ed.) (1977) in Section 3. Next we will explain on the basis of Ruhlen's classification how our sampling method takes into account linguistic variation within language families.

2.3. *Measuring linguistic variation in language families*

If one is interested in the broadest possible range of linguistic forms and structures one would like to see that the number of sample languages from the same family is proportional to the linguistic diversity in that

Table 3. *Language families according to Ruhlen (1987)*

Family	Number of languages
Afro-Asiatic	258
Altaic	66
Amerind	854
Australian	262
Austric	1186
Caucasian	38
Chukchi-Kamchatkan	5
Elamo-Dravidian	29
Eskimo-Aleut	9
Indo-Hittite	180
Indo-Pacific	748
Khoisan	33
Na-Dene	41
Niger-Kordofanian	1068
Nilo-Saharan	138
Sino-Tibetan	268
Uralic-Yukaghir	27
Language Isolates	9
Pidgins and Creoles	38
Unclassified languages	16
Total	5273

particular family. For this reason measuring the (potential) linguistic diversity in a group of genetically related languages is an important feature of our sampling method. It is based on the assumption that the graph theoretic structure of a genetic language tree adequately reflects the linguistic diversity of the family it represents, and consists of the computation of a factor called DIVERSITY VALUE or DV.

The DV of a family is computed over the number of nodes at the intermediate levels between the top node (i.e., the name of the language family) and the terminal nodes at the bottom end of the genetic language tree (i.e., the names of the individual languages of that family). Top nodes are ignored because they do not add any extra information: each family has exactly one such node and they all look the same. By disregarding all terminal nodes we restrict the influence of the actual number of languages in a family, shifting the weight entirely to the internal structure of the tree.[5]

When we measure the DV we take into consideration both the depth and the width of a genetic language tree. To calculate the width of such a tree at some intermediate level x, we extend all higher-level preterminal

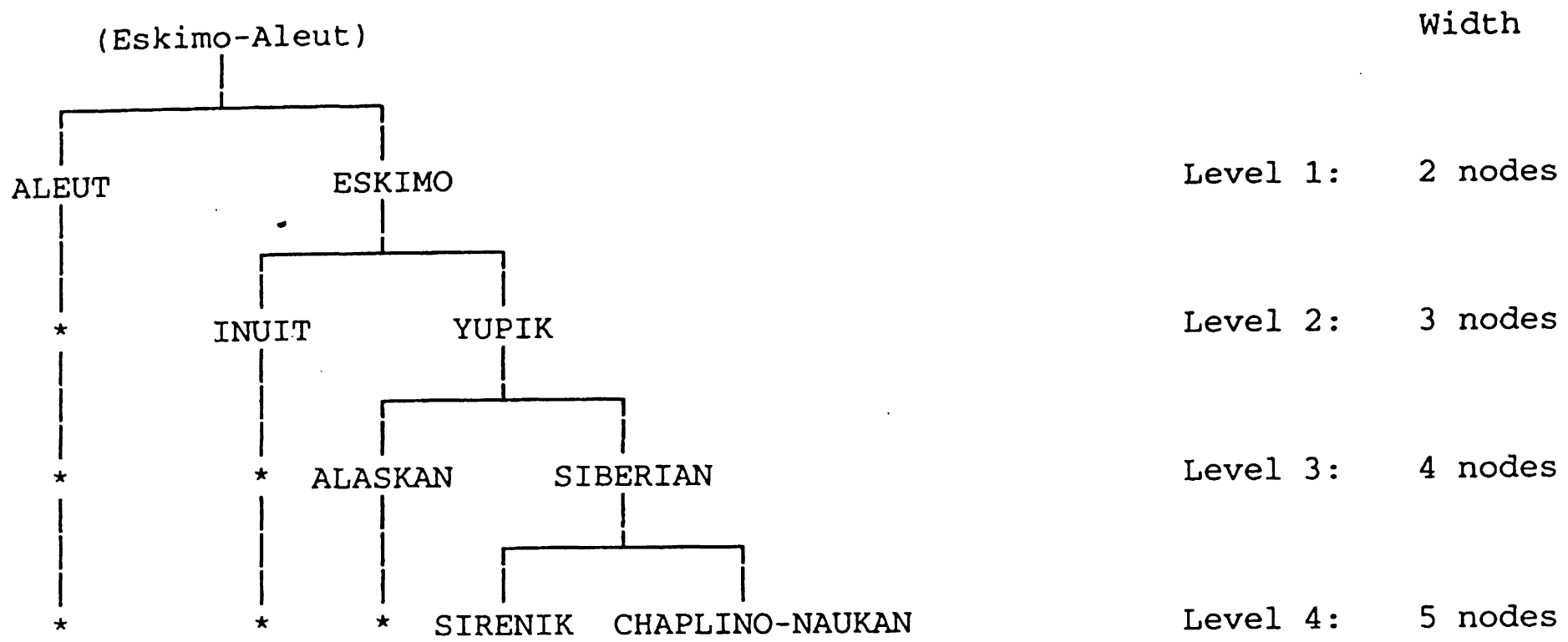

Figure 1. *Width (in number of nodes) at the four intermediate levels of the Eskimo-Aleut family*

nodes to level x. In the case of the Eskimo-Aleut family this leads to the constellation of intermediate levels and the corresponding number of (projected) nodes per level as represented in Figure 1. The width at some intermediate level of a genetic language tree is equal to the number of nodes on that level plus the number of extended preterminal nodes at higher intermediate levels (indicated by an asterisk). Thus, Eskimo-Aleut has four levels in between the top node (level 0) and the terminal nodes at the deepest level (i.e., the individual languages, not shown in Figure 1). Its width at level 1 is two; at level 4 the width is five.

There are various ways to convert the structure of a genetic language tree into a figure that indicates a diversity value (DV). Perhaps the simplest method would be to let DV be the average number of nodes per intermediate level, but we contend that high-level nodes should be assigned greater significance than lower-level nodes in the computation of DV. This is because high-level (early) split-ups may be assumed to have a greater impact on the linguistic diversity than low-level (recent) splits. For example, one expects greater differences between Eskimo and Aleut languages than between Alaskan and Siberian languages (Figure 1). So on the assumption that the distinguishing power of levels diminishes when going down the genetic language tree, we decided to let the contribution of the EXTRA nodes at each deeper intermediate level decrease by steps of $1/N_{max}$, where N_{max} is the highest number of intermediate levels found in any independent family. In Ruhlen's classification Niger-Kordofanian is the family which has most intermediate levels, namely 16.

These considerations are accounted for in the following formula, which calculates the contribution C_x of level x to the overall DV of a genetic grouping (N_x is the total number of nodes at level x, including the

Table 4. *Computation of the Diversity Value (DV) of the Eskimo-Aleut family*

Level	Nodes	C_{x-1}		C_x
1	2			2.0000
2	3	$2+((3-2)\times 15/16)$	=	2.9375
3	4	$2.9375+((4-3)\times 14/16)$	=	3.8125
4	5	$3.8125+((5-4)\times 13/16)$	=	4.6250+
				13.3750
				DV = 13.3750/4 = 3.34375

virtual ones):

(1) $C_x = C_{x-1} + ((N_x - N_{x-1}) \times (N_{max} - (x-1))/N_{max})$

By definition, the contribution of the first intermediate level (immediately below the top node) is equal to the number of nodes at that level ($C_1 = N_1$). C_x is recursively obtained by adding the contribution of the level immediately above level x (i.e., C_{x-1}) to the EXTRA nodes of level x (as compared to the total number of nodes one level up, i.e., $N_x - N_{x-1}$), the latter being multiplied by a fraction decreasing over the N_{max} levels according to the row N_{max}/N_{max}, $(N_{max}-1)/N_{max}$, ..., $1/N_{max}$). Thus, the deeper the level, the less the significance of the actual number of nodes. This is illustrated in Table 4 for the Eskimo-Aleut family ($N_{max} = 16$).

Table 4 shows that the contribution of level 1 is equal to the number of nodes at that level (viz., 2), but the contribution of the level 4 is 4.625 (rather than 5). DV is the mean value of the contributions of all intermediate levels. In the case of Eskimo-Aleut this is 13.375/4 = 3.34. In sum, since every language family has a fixed DV we can now have a proportional representation of families in the sample on the basis of their linguistic diversity. Table 5 gives the DVs for all of Ruhlen's families (LI = Language Isolate). In order to let Language Isolates participate in the sampling procedure they all receive the minimum value of 1.00 (rather than 0.00). We will explain in Section 2.4 that in our sampling procedure DVs are also calculated for genetic groups at all levels WITHIN each language family. Consequently sample languages can be proportially distributed over the various subfamilies as well.

2.4. *The distribution of languages over (sub)families*

In our sampling method the actual distribution of languages over families and subfamilies is achieved in three consecutive phases. Let n be the sample size and S_{DV} the sum of the DVs of all independent language

Table 5. *Diversity Values (DVs) of Ruhlen's language families*

Family	DV	Family	DV
Afro-Asiatic	55.53	Nilo-Saharan	42.18
Altaic	14.79	Pidgins and Creoles	13.47
Amerind	178.44	Sino-Tibetan	38.52
Australian	67.58	Uralic-Yukaghir	4.93
Austric	137.41	Basque (LI)	1.00
Caucasian	8.54	Burushaski (LI)	1.00
Chukchi-Kamchatkan	2.47	Etruscan (LI)	1.00
Elamo-Dravidian	7.43	Gilyak (LI)	1.00
Eskimo-Aleut	3.34	Hurrian (LI)	1.00
Indo-Hittite	39.71	Ket (LI)	1.00
Indo-Pacific	123.39	Meroitic (LI)	1.00
Khoisan	6.97	Nahali (LI)	1.00
Na-Dene	9.44	Sumerian (LI)	1.00
Niger-Kordofanian	90.38	Total	853.20

families. First one language is assigned to each family that would not have been represented on the basis of DVs alone, i.e., families with DVs lower than S_{DV}/n. This is to make sure that all families will be represented in the sample. In the second phase all remaining languages are proportionally distributed over the families according to their DVs. Due to rounding it may happen that in the end fewer or more languages are selected than the number of languages that the sample is supposed to contain. Consequently in the third and final phase the sample is padded until it contains the right number of languages. These phases are exemplified in Table 6 (using Ruhlen's classification), where n is fixed at 100 (i.e., the sample will contain 100 languages); hence the value of S_{DV}/n is 8.53 (853.2/100).

Evidently our method requires that the size of the sample must be determined before sample languages are selected. We would also like to point out once again that the distribution of languages over families is not so much determined by the size of the family as by the genetic relations between individual languages. For example, the Amerind family, which contains 854 languages, is represented by twice as many languages as the Niger-Kordofanian family, which contains as many as 1068 languages. Amerind divides into six primary branches, whereas Niger-Kordofanian has only two. Table 7 gives the distribution of languages over families in samples of varying sizes.

Since Ruhlen recognizes only 27 independent language families (including nine Language Isolates and a group of Pidgins & Creoles), samples based on this classification will often contain more languages from the

Table 6. *The distribution of languages over language families in a 100-language sample (Ruhlen's classification)*

Family	DV	Phase 1	Phase 2		Phase 3
Afro-Asiatic	55.53	0	6		6
Altaic	14.79	0	1	→	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	123.39	0	12	→	13
Khoisan	6.97	1	1		1
Language Isolates:					
Basque	1.00	1	1		1
Burushaski	1.00	1	1		1
Etruscan	1.00	1	1		1
Gilyak	1.00	1	1		1
Hurrian	1.00	1	1		1
Ket	1.00	1	1		1
Meroitic	1.00	1	1		1
Nahali	1.00	1	1		1
Sumerian	1.00	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 & Creoles	13.47	0	1	→	2
Sino-Tibetan	38.52	0	4		4
Uralic-Yukaghir	4.93	1	1		1
Totals	853.20	14	96		100

same family. Hence our sampling method also contains a procedure that deals with the selection and distribution of genetically related languages, i.e., WITHIN each language family. Since we want the genetic distance between sample languages to be maximal, genetically related languages are always selected from different same-level subfamilies. And in order to have a REPRESENTATIVE contribution from the subfamilies, the number of sample languages to be selected from a subfamily is determined by the DV of that subfamily. So, analogous to the calculations of Section 2.3, we repeat the procedure outlined above, with an adapted value for N_{max}, which is now determined by the maximum number of intermediate levels

Table 7. *The distribution of languages in samples of different sizes using DVs*

Families (Ruhlen 1987)	Sample size 30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
1. Afro-Asiatic	1	2	2	3	4	5	5	6	8	9	11	12	16	19	26	32	65
2. Altaic	1	1	1	1	1	1	2	2	2	3	3	3	4	5	7	9	17
3. Amerind	2	5	7	9	12	14	16	18	24	29	35	40	51	61	83	104	209
4. Australian	1	2	3	4	4	5	6	7	9	11	13	15	19	23	31	39	79
5. Austric	2	4	5	7	9	11	13	14	19	23	27	31	39	47	64	80	161
6. Basque (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7. Burushaski (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8. Caucasian	1	1	1	1	1	1	1	1	1	2	2	2	3	3	4	5	10
9. Chukchi-Kamchatkan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3
10. Elamo-Dravidian	1	1	1	1	1	1	1	1	1	1	1	2	2	3	3	4	9
11. Eskimo-Aleut	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	4
12. Etruscan (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13. Gilyak (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14. Hurrian (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15. Indo-Hittite	1	1	2	2	3	3	4	4	5	7	8	9	11	14	18	23	47
16. Indo-Pacific	2	3	5	7	8	10	11	13	17	20	24	28	35	42	57	72	145
17. Ket (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18. Khoisan	1	1	1	1	1	1	1	1	1	1	1	2	2	3	3	4	8
19. Meroitic (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20. Na-Dene	1	1	1	1	1	1	1	1	1	2	2	2	3	3	4	5	11
21. Nahali (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
22. Niger-Kordofanian	1	3	4	5	6	7	8	9	12	15	18	20	26	31	42	53	106
23. Nilo-Saharan	1	1	2	3	3	4	4	5	6	7	8	10	12	15	20	25	50
24. Pidgins & Creoles	1	1	1	1	1	1	1	2	2	2	3	3	4	5	6	8	16
25. Sino-Tibetan	1	1	2	2	3	3	4	4	5	6	7	9	11	13	18	22	45
26. Sumerian (LI)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27. Uralic-Yukaghir	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	3	6

found under any daughter node of the mother node of the subfamily in question.

Thus, DVs are first computed according the formula given in (1) for the subfamilies immediately under the top node, the so-called primary branches (e.g., Aleut and Eskimo in the Eskimo-Aleut family; see Figure 1). Then fractions of the total number of languages assigned to a family are assigned to the primary branches according to DVs. This procedure is applied recursively in a top-down fashion until there are more subfamilies than there are languages to be sampled from these subfamilies. At this point a decision has to be taken which of these subfamilies are to be represented in the sample and which subfamilies are going to be left out; this will be discussed in Section 2.5.

The recursive application of this part of the sampling procedure is illustrated in Table 8, which gives the representation of the different subfamilies of the Amerind family in a 250-language sample. Figures in parentheses indicate the DV, the number of daughter nodes (branches), and the total number of languages under the corresponding node in the tree. In a sample of this size Amerind is represented with 51 languages,

Table 8. *Representation of the Amerind family and its subfamilies in a 250-language sample*

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

to be selected from the 854 languages that are distributed over six subfamilies (the primary branches). First one language must be assigned to each subfamily which on the basis of its DV alone would not have been represented in a sample of this particular size. This is not the case for any of the subfamilies. Then the remaining languages are distributed over the six primary branches on the basis of their DVs. Next the languages are distributed over the secondary branches (i.e., subfamilies immediately under the primary branches) according to the same method. In the case of Andean the procedure is repeated only once: we end up with three languages, to be distributed over its six subfamilies (see Table 8, bottom line). In the case of the five other primary branches of the Amerind group 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 to distribute six and four languages over the Ge-Pano and Almosan-Keresiouan subfamilies, respectively.

When there are not enough languages to have a representative of each subfamily in the sample (as in the case of Andean in the example above), one can either make one's own selection, or let a pre-established procedure decide which subfamilies will be represented in the sample. If no such procedure is available, it is to be expected that typologists will tend to select only those languages whose descriptions are readily available. This could mean, of course, that due to bibliographic restrictions certain subfamilies will be represented time and again, whereas other subfamilies, those containing languages that are less well described, will be systematically ignored. The next section desribes how the sampling program deals with this problem.

2.5. *The selection of subfamilies*

When we use the DV method to distribute sample languages over subfamilies, there is always a moment (whatever the size of the sample) when not every subfamily can be represented by a language in the sample, simply because the number of subfamilies exceeds the number of available languages. For example, a 250-language sample contains three Andean languages but since Andean has six subfamilies at least three of them will not be represented in the sample. There are basically three strategies to consider in such a case:

(a) arrange the subfamilies according to their DVs and select the subfamilies with the highest DVs;
(b) randomly select the subfamilies, irrespective of their DVs;
(c) randomly select the appropriate number of subfamilies, while taking into account their DVs.

Table 9. *Distribution of languages over the subfamilies of Andean in a 250-language sample according to strategy (c)*

Andean (9.50/6/30)	3	
Northern (5.00/5/7)		1
Southern (4.00/4/9)		1
Cahuapanan-Zaparoan (2.00/2/7)		—
Urarina-Waorani (1.50/0/3)		—
Aymaran (1.50/0/2)		1
Quechuan (1.50/0/2)		—

Strategy (a) has a serious drawback in that subfamilies with low DVs are systematically ignored and will never show up in samples below a certain size. In procedures (b) and (c), on the other hand, every subfamily has a chance of being represented with a language in the sample. The only difference is that in the long run subfamilies with low DVs would be overrepresented if we would choose strategy (b). This is not the case with strategy (c), where the chance of a particular subfamily to be selected is proportional to the DV of that subfamily. This is exemplified in Table 9, which gives the three Andean subfamilies that our computer program selected by applying strategy (c). This particular selection does not include a representative of the Cahuapanan-Zaparoan subfamily (2.00/2/7), although it had a better chance of being represented than Aymaran (1.50/0/2), which does have a language in the sample.

3. Sampling with other classifications

3.1. *Family counts*

So far we have used only Ruhlen's classification of the world's languages to illustrate how sample languages are distributed over the various language families and subfamilies. Therefore certain problems remained undiscussed. For instance, we already saw that Ruhlen recognizes no more than 27 families, whereas Voegelin & Voegelin (1977) and Grimes (ed.) (1997) distinguish around 50 and 120 families, respectively, again counting each Language Isolate as a singleton phylum and the Pidgins & Creoles as one family. In other words, using Ruhlen's classification, a 30-language sample is already large enough to contain a representative from each family. If one uses one of the other classifications, however, the sample must be about two or even four times as large in order to contain at least one language from each family. Since it is not always possible—or necessary—to have samples of that size, our sampling method contains a procedure which determines which families are going to be represented in the sample and which families will be left out when there are more

families than languages in the sample. This is discussed in Section 3.4. First we will deal with some details of the classifications by Voegelin & Voegelin (1977) and Grimes (ed.) (1997).[6]

3.2. *Voegelin & Voegelin (1977)*

Voegelin & Voegelin (1977) distinguish 48 families, including a non-genetic group of Pidgin & Creole languages and 15 Language Isolates (but see below).[7] These families are listed in Table 10, where the figures in parentheses after each family name refer to Diversity Value, the number of primary branches, and the number of languages.

In order to let families without an internal structure (e.g., Chimakuan: no branches, only individual languages) participate in the sampling procedure we give each of them a DV of 1.50. This puts them in between the Language Isolates (DV=1; no branches, only one language) and families with only one intermediate level and two primary branches, such as Wakashan (2.00/2/6).[8]

Table 10. *Language families in Voegelin & Voegelin (1977)*

1. Afroasiatic (25.04/6/209)
2. Andean-Equatorial (39.90/4/250)
3. Australian (56.42/28/259)
4. Austro-Asiatic (10.33/4/109)
5. Austronesian (65.67/3/778)
6. Aztec-Tanoan (4.19/2/30)
7. Caucasian (5.25/3/33)
8. Chimakuan (1.50/0/2)
9. Chukchee-Kamchatkan (1.50/0/5)
10. Dravidian (3.00/3/22)
11. Eskimo-Aleut (1.50/0/6)
12. Ge-Pano-Carib (20.06/6/269)
13. Hokan (21.31/20/36)
14. Indo-European (28.53/13/153)
15. Indo-Pacific (112.45/49/754)
16. Khoisan (3.88/3/48)
17. Macro-Algonquian (9.00/9/30)
18. Macro-Chibchan (18.19/16/60)
19. Macro-Siouan (9.92/6/26)
20. Na-Dene (3.00/3/32)
21. Niger-Kordofanian (55.93/2/1036)
22. Nilo-Saharan (22.61/6/114)
23. Oto-Manguean (7.88/7/25)
24. Penutian (25.38/21/78)
25. Pidgins & Creoles (4.58/2/34)
26. Salish (5.00/5/23)
27. Sino-Tibetan (21.31/5/289)
28. Ural-Altaic (11.29/2/89)
29. Waicurian (1.50/0/2)
30. Wakashan (2.00/2/6)
31. Yeniseian (1.50/0/4)
32. Yukaghir (1.50/0/2)
33. Yuki (1.50/0/2)

Language Isolates:

34. Ainu (1.00/0/1)
35. Aricapu (1.00/0/1)
36. Baenna (1.00/0/1)
37. Basque (1.00/0/1)
38. Beothuk (1.00/0/1)
39. Burushaski (1.00/0/1)
40. Callahuaya (1.00/0/1)
41. Calusa (1.00/0/1)
42. Gilyak (1.00/0/1)
43. Juma (1.00/0/1)
44. Keres (1.00/0/1)
45. Kutenai (1.00/0/1)
46. Nahali (1.00/0/1)
47. Natu (1.00/0/1)
48. Tarascan (1.00/0/1)

Because Voegelin & Voegelin's classification is older than those by Ruhlen (1987) and Grimes (ed.) (1997), it is often not so much a straightforward presentation of genetic relations between the world's languages (the ideal input for our sampling procedure), but rather an overview of alternative and often conflicting classification proposals regarding certain languages or language groups. Rather than somehow trying to find a way to incorporate all these alternatives in our computer database, we simply accepted the genetic groupings that Voegelin & Voegelin provide following the sections in which they discuss the various proposals. Consequently Voegelin & Voegelin's classification is on the whole less detailed than those put forward by Ruhlen and Grimes (Table 11). Notice also that Voegelin & Voegelin have more Language Isolates than Ruhlen, but the list would have been even longer if we had also counted all extinct Language Isolates and the languages with "unknown affiliation" (the difference between the two categories is not always clear). For most of these languages only a few inscriptions remain and sometimes linguistic data are lacking altogether, as when the only evidence we have is the name of a language in an ancient text. This holds especially for the numerous extinct languages that were spoken in Asia Minor.[9] Furthermore, as at least some of the other languages (such as Karankawa, Hishkaryana, and Pontic) have been classified since the publication of Voegelin & Voegelin's book, we have decided to ignore these languages in the sampling procedure.

3.3. *The* Ethnologue *(Grimes (ed.) 1997)*

In the older editions of the *Ethnologue* the editor closely followed Ruhlen's classification, but in later editions the more conservative classification of the *International Encyclopedia of Linguistics* is used (cf. Grimes & Grimes 1993: iii–v). The 1997 edition of the *Ethnologue* contains many more families than Ruhlen's classification, mostly because families that are part of a larger genetic group in Ruhlen are listed here as separate families (see also Table 14). Furthermore, in the more recent *Ethnologue* editions priority is given to "classifications based on shared structural and lexical innovations, as opposed to lexicostatistical, purely typological, or impressionistic subgroupings" (Grimes & Grimes 1993: iii).

Table 11 gives the families as listed in the classification of the 13th edition of the *Ethnologue* (Grimes (ed.) 1997). We made some minor adjustments in the version that was made available to us (see Note 1) in that we left out the Sign Languages and Pitcairn-English (which was listed by itself under the name of "English-Tahitian cant"). We also decided to treat all singleton phyla as Language Isolates; in the *Ethnologue*

Table 11. *Language families in* Ethnologue *(Grimes (ed.) 1997)*

Afro-Asiatic (52.76/7/371)
Alacalufan (1.50/0/2)
Algic (7.38/3/33)
Altaic (14.51/3/65)
Amto-Musan (1.50/0/2)
Andamanese (2.46/2/13)
Araucanian (1.50/0/2)
Arawakan (9.15/5/74)
Arutani-Sape (1.50/0/2)
Australian (58.92/20/257)
Austro-Asiatic (35.54/2/180)
Austronesian (141.76/2/1236)
Aymaran (1.50/0/3)
Barbacoan-Paezan (3.92/3/7)
Basque (1.50/0/3)
Caddoan (2.90/2/5)
Cahuapanan (1.50/0/2)
Carib (8.23/2/29)
Chapacura-Wanham (2.00/2/5)
Chibchan (10.46/10/22)
Choco (5.46/5/10)
Chon (1.50/0/2)
Chukotko-Kamchatkan (2.46/2/5)
Pidgins & Creoles (23.38/20/77)
Daic (9.42/3/68)
Dravidian (11.18/5/78)
East Bird's Head (2.00/2/3)
East Papuan (7.05/3/36)
Eskimo-Aleut (2.90/2/11)
Geelvink Bay (8.35/5/34)
Gulf (1.50/0/4)
Hmong-Mien (5.31/3/32)
Hokan (10.92/5/27)
Huavean (1.50/0/4)
Indo-European (50.34/9/425)
Iroquoian (4.69/2/9)
Japanese (2.90/2/12)
Jivaroan (2.00/2/5)
Katukinan (1.50/0/3)
Keres (1.50/0/2)
Khoisan (6.85/3/35)
Kiowa Tanoan (2.92/2/5)
Kwomtari-Baibai (3.00/3/6)
Left May (1.50/0/7)
Macro-Ge (16.77/14/32)
Maku (1.50/0/5)
Mascoian (1.50/0/6)
Mataco-Guaicuru (2.00/2/11)
Mayan (11.54/5/68)
Misumalpan (1.50/0/4)
Mixe-Zoque (3.85/2/16)
Muskogean (2.00/2/6)
Na-Dene (8.28/2/42)
Nambiquaran (1.50/0/5)
Niger-Congo (130.36/3/1436)
Nilo-Saharan (46.65/10/194)
North Caucasian (6.64/3/34)
Oto-Manguean (12.15/7/173)
Panoan (9.38/8/29)
Peba-Yaguan (1.50/0/2)
Penutian (9.35/5/27)
Quechuan (2.92/2/47)
Salishan (8.23/5/27)
Salivan (1.50/0/2)
Sepik-Ramu (17.27/5/105)
Sino-Tibetan (39.13/3/360)
Siouan (5.50/2/17)
Sko (2.00/2/8)
South Caucasian (3.00/3/5)
Subtiaba-Tlapanec (1.50/0/4)
Tacanan (2.90/2/6)
Torricelli (9.77/7/48)
Totonacan (2.00/2/11)
Trans-New Guinea (89.69/21/539)
Tucanoan (8.15/4/26)
Tupi (15.15/11/70)
Uralic (5.93/2/34)
Uru-Chipaya (1.50/0/2)
Uto-Aztecan (9.31/2/60)
Wakashan (2.00/2/5)
West Papuan (6.92/4/27)
Witotoan (2.90/2/6)
Yanomam (1.50/0/4)
Yenisei Ostyak (1.50/0/2)
Yukaghir (1.50/0/2)
Yuki (1.50/0/2)
Zamucoan (1.50/0/2)
Zaparoan (1.50/0/7)

Language Isolates:
Quileute (1.00/0/0)
Tonkawa (1.00/0/0)
Ainu (1.00/0/0)
Andoque (1.00/0/0)

Table 11. *(Continued)*

Burmeso (1.00/0/0)	Puinave (1.00/0/0)
Burushaski (1.00/0/0)	Purépecha (1.00/0/0)
Busa (1.00/0/0)	Ticuna (1.00/0/0)
Camsá (1.00/0/0)	Tol (1.00/0/0)
Cayubaba (1.00/0/0)	Trumaí (1.00/0/0)
Gilyak (1.00/0/0)	Tuxá (1.00/0/0)
Itonama (1.00/0/0)	Warao (1.00/0/0)
Karkar-Yuri (1.00/0/0)	Warembori (1.00/0/0)
Kibiri (1.00/0/0)	Yámana (1.00/0/0)
Korean (1.00/0/0)	Yale (1.00/0/0)
Kutenai (1.00/0/0)	Yuchi (1.00/0/0)
Mbugu (1.00/0/0)	Yuracare (1.00/0/0)
Nihali (1.00/0/0)	Zuni (1.00/0/0)
Pankararú (1.00/0/0)	Vilela (1.00/0/0)
Pauwi (1.00/0/0)	Tsimané (1.00/0/0)
Puelche (1.00/0/0)	Múra-Pirahã (1.00/0/0)

a distinction is made between Language Isolates proper and phyla with one language, but which until recently were still a small family: Múra-Pirahã (=Mura), Chimakuan (=Quileute), Tonkawa (=Coahuiltecan), Vilela, and Mosetenan (=Tsimané).

3.4. *The distribution of languages over families in samples based on Voegelin & Voegelin 1977 and the* Ethnologue

In Section 2.5 we proposed a method that determines which subfamilies are going to be represented when the sample is not large enough to include a representative from each subfamily at some level in the family tree. According to this method we randomly select the appropriate number of subfamilies while taking into account their DVs, so that subfamilies with a relatively high DV have a greater chance of being represented in the sample than subfamilies with a relatively low DV. We use the same strategy when the sample is too small to contain a language from each independent family, a problem that is likely to occur if one uses the classifications by Voegelin & Voegelin or Grimes (recall that Ruhlen has less than 30 families, including the Language Isolates). Tables 12 and 13 give the distribution of languages over families in samples of various sizes based on the classifications of Voegelin & Voegelin (1977) and the 13th edition of the *Ethnologue* (Grimes (ed.) 1997). For samples that are not large enough to contain a language from each family we calculated the chance of a family being represented in the sample. To give an example based on Voegelin & Voegelin's classification (Table 12), 969 out of 1000

Table 12. *Samples on the basis of Voegelin & Voegelin (1977)*

Families	Sample size 30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
Afroasiatic	0.853	0.969	1	1	2	2	3	3	4	5	6	7	10	12	16	20	41
Ainu (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Andean-Equatorial	0.953	0.996	1	2	3	3	4	5	7	8	10	12	15	19	25	32	65
Aricapu (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Australian	0.987	1.000	1	3	4	5	6	7	9	12	14	17	21	26	36	46	92
Austro-Asiatic	0.545	0.763	1	1	1	1	1	1	2	2	3	3	4	5	7	8	17
Austronesian	0.994	1.000	1	3	4	6	7	8	11	14	17	19	25	30	41	53	107
Aztec-Tanoan	0.273	0.444	1	1	1	1	1	1	1	1	1	1	2	2	3	3	7
Baenna (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Basque (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Beothuk (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Burushaski (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Callahuaya (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Calusa (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Caucasian	0.329	0.520	1	1	1	1	1	1	1	1	2	2	2	3	4	4	9
Chimakuan	0.108	0.190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Chukchee-Kamchatkan	0.108	0.190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Dravidian	0.204	0.343	1	1	1	1	1	1	1	1	1	1	1	2	2	3	5
Eskimo-Aleut	0.108	0.190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3
Ge-Pano-Carib	0.784	0.938	1	1	1	2	2	3	3	4	5	6	8	9	13	16	33
Gilyak (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hokan	0.804	0.948	1	1	1	2	2	3	4	5	6	6	8	10	14	17	35

Indo-European	0.887	0.981	1	2	2	3	3	4	5	6	7	9	11	13	18	23	47
Indo-Pacific	1.000	1.000	3	5	7	9	12	13	18	23	28	33	43	52	71	91	184
Juma (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Keres (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Khoisan	0.255	0.419	1	1	1	1	1	1	1	1	1	1	2	2	3	3	6
Kutenai (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Macro-Chibchan	0.751	0.920	1	1	1	2	2	2	3	4	5	6	7	9	12	15	30
Macro-Siouan	0.530	0.749	1	1	1	1	1	1	2	2	3	3	4	5	6	8	16
Macro-Algonquian	0.496	0.715	1	1	1	1	1	1	2	2	2	3	4	4	6	7	15
Na-Dene	0.204	0.343	1	1	1	1	1	1	1	1	1	1	1	2	2	3	5
Nahali (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Natu (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Niger-Kordofanian	0.987	1.000	1	3	4	5	6	7	9	12	14	16	21	26	35	45	91
Nilo-Saharan	0.822	0.956	1	1	2	2	3	3	4	5	6	7	9	11	14	18	37
Oto-Manguean	0.451	0.667	1	1	1	1	1	1	2	2	2	3	3	4	5	6	13
Penutian	0.856	0.970	1	1	2	2	3	3	4	6	7	8	10	12	16	21	42
Pidgins & Creoles	0.294	0.474	1	1	1	1	1	1	1	1	1	2	2	2	3	4	8
Salish	0.316	0.503	1	1	1	1	1	1	1	1	1	2	2	2	3	4	8
Sino-Tibetan	0.804	0.948	1	1	2	2	2	3	4	5	6	6	8	10	14	17	35
Tarascan (LI)	0.073	0.131	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ural-Altaic	0.577	0.793	1	1	1	1	1	2	2	3	3	3	4	5	7	9	19
Waicurian	0.108	0.190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Wakashan	0.141	0.245	1	1	1	1	1	1	1	1	1	1	1	1	2	2	3
Yeniseian	0.108	0.190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Yukaghir	0.108	0.190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Yuki	0.108	0.190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2

Table 13. *Samples on the basis of* Ethnologue *(Grimes (ed.) 1997)*

Families	Sample size 30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
Afro-Asiatic	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	3	5	6	9	11	17	22	47
Ainu	0.07	0.11	0.15	0.24	0.24	0.38	0.52	0.73	1	1	1	1	1	1	1	1	1
Alacalufan	0.09	0.14	0.22	0.29	0.35	0.52	0.58	0.73	1	1	1	1	1	1	1	1	2
Algic	0.36	0.52	0.68	0.78	0.90	0.91	1.00	1.00	1	1	1	1	2	2	3	4	7
Altaic	0.61	0.79	0.92	0.95	1.00	1.00	1.00	1.00	1	1	2	2	3	4	5	6	13
Amto-Musan	0.10	0.15	0.23	0.34	0.45	0.52	0.68	0.64	1	1	1	1	1	1	1	1	2
Andamanese	0.16	0.21	0.33	0.41	0.55	0.66	0.78	0.91	1	1	1	1	1	1	1	2	2
Andoque	0.06	0.10	0.15	0.22	0.28	0.33	0.42	0.46	1	1	1	1	1	1	1	1	1
Araucanian	0.09	0.13	0.20	0.29	0.41	0.52	0.65	0.73	1	1	1	1	1	1	1	1	2
Arawakan	0.41	0.52	0.68	0.83	0.90	0.95	1.00	1.00	1	1	1	1	2	2	3	4	8
Arutani-Sape	0.09	0.13	0.18	0.20	0.28	0.28	0.45	0.64	1	1	1	1	1	1	1	1	2
Australian	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	3	5	7	10	12	18	24	53
Austro-Asiatic	0.85	0.97	0.97	1.00	1.00	1.00	1.00	1.00	1	2	3	4	6	8	11	15	32
Austronesian	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2	8	12	15	22	29	44	57	127
Aymaran	0.09	0.14	0.22	0.32	0.41	0.52	0.71	0.91	1	1	1	1	1	1	1	1	2
Barbacoan-Paezan	0.21	0.33	0.48	0.61	0.69	0.76	0.87	1.00	1	1	1	1	1	1	2	2	4
Basque	0.09	0.15	0.22	0.29	0.38	0.43	0.61	0.73	1	1	1	1	1	1	1	1	2
Burmeso	0.06	0.10	0.15	0.22	0.31	0.38	0.48	0.64	1	1	1	1	1	1	1	1	1
Burushaski	0.06	0.10	0.15	0.22	0.31	0.43	0.52	0.73	1	1	1	1	1	1	1	1	1
Busa	0.07	0.11	0.17	0.25	0.34	0.48	0.51	0.64	1	1	1	1	1	1	1	1	1
Caddoan	0.17	0.26	0.34	0.41	0.52	0.62	0.65	0.73	1	1	1	1	1	1	1	2	3
Cahuapanan	0.09	0.14	0.20	0.32	0.41	0.57	0.68	0.91	1	1	1	1	1	1	1	1	2
Camsá	0.06	0.10	0.15	0.22	0.31	0.38	0.52	0.64	1	1	1	1	1	1	1	1	1
Carib	0.41	0.62	0.75	0.88	0.93	1.00	0.97	1.00	1	1	1	1	2	2	3	4	8
Cayubaba	0.06	0.10	0.15	0.22	0.28	0.38	0.52	0.73	1	1	1	1	1	1	1	1	1

Chapacura-Wanham	0.12	0.19	0.25	0.34	0.48	0.52	0.58	0.73	1	1	1	1	1	1	1	1	2
Chibchan	0.45	0.59	0.73	0.81	0.93	1.00	1.00	1.00	1	1	1	2	2	3	4	5	10
Choco	0.29	0.41	0.58	0.76	0.86	0.95	0.97	1.00	1	1	1	1	1	2	2	3	5
Chon	0.08	0.13	0.19	0.24	0.38	0.48	0.52	0.64	1	1	1	1	1	1	1	1	2
Chukotko-Kamchatkan	0.15	0.22	0.34	0.46	0.59	0.71	0.87	0.91	1	1	1	1	1	1	1	2	2
Daic	0.48	0.61	0.75	0.85	0.96	1.00	1.00	1.00	1	1	1	2	2	3	3	4	9
Dravidian	0.49	0.67	0.78	0.93	0.97	0.95	0.97	1.00	1	1	1	2	2	3	4	5	10
East Bird's Head	0.11	0.20	0.27	0.39	0.52	0.62	0.77	1.00	1	1	1	1	1	1	1	1	2
East Papuan	0.35	0.51	0.65	0.81	0.93	0.90	1.00	1.00	1	1	1	1	2	2	3	3	6
Eskimo-Aleut	0.15	0.25	0.33	0.47	0.55	0.67	0.71	0.82	1	1	1	1	1	1	1	2	3
Geelvink Bay	0.42	0.57	0.67	0.83	0.90	0.95	1.00	1.00	1	1	1	1	2	2	3	4	8
Gilyak	0.06	0.10	0.15	0.22	0.31	0.43	0.55	0.73	1	1	1	1	1	1	1	1	1
Gulf	0.09	0.14	0.18	0.29	0.34	0.48	0.55	0.64	1	1	1	1	1	1	1	1	2
Hmong-Mien	0.29	0.44	0.58	0.70	0.83	0.95	0.97	1.00	1	1	1	1	1	2	2	3	5
Hokan	0.50	0.63	0.77	0.90	0.93	1.00	1.00	1.00	1	1	1	2	2	3	4	5	10
Huavean	0.10	0.15	0.22	0.29	0.48	0.52	0.61	0.73	1	1	1	1	1	1	1	1	2
Indo-European	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	3	4	6	8	11	16	21	45
Iroquoian	0.23	0.35	0.48	0.64	0.69	0.86	0.90	1.00	1	1	1	1	1	2	2	3	4
Itonama	0.07	0.11	0.17	0.22	0.28	0.33	0.52	0.55	1	1	1	1	1	1	1	1	1
Japanese	0.15	0.24	0.36	0.49	0.55	0.71	0.90	0.91	1	1	1	1	1	1	1	2	3
Jivaroan	0.12	0.20	0.30	0.44	0.58	0.62	0.74	0.91	1	1	1	1	1	1	1	1	2
Karkar-Yuri	0.06	0.10	0.14	0.19	0.28	0.38	0.52	0.46	1	1	1	1	1	1	1	1	1
Katukinan	0.09	0.14	0.20	0.29	0.38	0.53	0.64	0.82	1	1	1	1	1	1	1	1	2
Keres	0.09	0.14	0.20	0.29	0.41	0.52	0.71	0.73	1	1	1	1	1	1	1	1	2
Khoisan	0.35	0.53	0.60	0.76	0.86	0.95	1.00	1.00	1	1	1	1	2	2	3	3	6
Kibiri	0.06	0.10	0.15	0.22	0.31	0.43	0.52	0.55	1	1	1	1	1	1	1	1	1
Kiowa Tanoan	0.16	0.26	0.35	0.46	0.62	0.81	0.94	1.00	1	1	1	1	1	1	2	2	3
Korean	0.06	0.09	0.15	0.17	0.24	0.43	0.42	0.45	1	1	1	1	1	1	1	1	1
Kutenai	0.06	0.11	0.15	0.22	0.34	0.38	0.48	0.55	1	1	1	1	1	1	1	1	1
Kwomtari-Baibai	0.17	0.26	0.38	0.51	0.65	0.71	0.84	0.91	1	1	1	1	1	1	2	2	3

Table 13. *(Continued)*

Families	Sample size 30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
Left May	0.09	0.14	0.22	0.32	0.31	0.48	0.58	0.82	1	1	1	1	1	1	1	1	2
Macro-Ge	0.67	0.83	0.93	1.00	1.00	1.00	1.00	1.00	1	1	2	2	3	4	6	7	15
Maku	0.09	0.14	0.19	0.27	0.38	0.52	0.58	0.73	1	1	1	1	1	1	1	1	2
Mascoian	0.09	0.14	0.22	0.24	0.35	0.47	0.55	0.82	1	1	1	1	1	1	1	1	2
Mataco-Guaicuru	0.13	0.20	0.32	0.44	0.58	0.67	0.78	0.91	1	1	1	1	1	1	1	1	2
Mayan	0.53	0.71	0.77	0.85	0.96	0.95	1.00	1.00	1	1	1	2	2	3	4	5	10
Mbugu	0.06	0.10	0.15	0.22	0.31	0.38	0.49	0.55	1	1	1	1	1	1	1	1	1
Misumalpan	0.08	0.13	0.20	0.24	0.34	0.48	0.61	0.73	1	1	1	1	1	1	1	1	2
Mixe-Zoque	0.21	0.32	0.45	0.54	0.66	0.81	0.87	1.00	1	1	1	1	1	1	2	2	4
Múra-Pirahã	0.07	0.11	0.15	0.22	0.28	0.43	0.48	0.55	1	1	1	1	1	1	1	1	1
Muskogean	0.12	0.20	0.30	0.39	0.48	0.62	0.74	0.91	1	1	1	1	1	1	1	1	2
Na-Dene	0.41	0.56	0.75	0.80	0.96	0.95	1.00	1.00	1	1	1	1	2	2	3	4	8
Nambiquaran	0.08	0.13	0.18	0.27	0.34	0.48	0.61	0.64	1	1	1	1	1	1	1	1	2
Niger-Congo	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	7	11	14	21	27	40	53	117
Nihali	0.06	0.09	0.14	0.19	0.28	0.33	0.45	0.55	1	1	1	1	1	1	1	1	1
Nilo-Saharan	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	2	4	5	8	10	15	19	42
North Caucasian	0.34	0.51	0.72	0.80	0.93	0.95	0.97	1.00	1	1	1	1	1	2	3	3	6
Oto-Manguean	0.49	0.67	0.86	0.98	1.00	1.00	1.00	1.00	1	1	1	2	2	3	4	6	11
Pankararú	0.06	0.09	0.13	0.20	0.28	0.38	0.45	0.36	1	1	1	1	1	1	1	1	1
Panoan	0.46	0.65	0.82	0.90	0.93	1.00	1.00	1.00	1	1	1	1	2	2	3	4	9
Pauwi	0.07	0.11	0.15	0.22	0.24	0.38	0.48	0.55	1	1	1	1	1	1	1	1	1
Peba-Yaguan	0.09	0.14	0.20	0.29	0.38	0.52	0.68	0.73	1	1	1	1	1	1	1	1	2
Penutian	0.44	0.62	0.80	0.90	0.93	0.95	1.00	1.00	1	1	1	1	2	2	3	4	9
Pidgins & Creoles	0.77	0.92	0.97	1.00	1.00	1.00	1.00	1.00	1	1	2	3	4	5	8	10	21
Puelche	0.06	0.10	0.14	0.20	0.28	0.33	0.42	0.55	1	1	1	1	1	1	1	1	1
Puinave	0.05	0.09	0.13	0.19	0.28	0.33	0.45	0.64	1	1	1	1	1	1	1	1	1

Purépecha	0.06	0.10	0.15	0.22	0.31	0.43	0.55	0.64	1	1	1	1	1	1	1	1	1
Quechuan	0.17	0.28	0.38	0.54	0.69	0.76	0.84	0.91	1	1	1	1	1	1	2	2	3
Quileute	0.06	0.10	0.15	0.19	0.31	0.33	0.42	0.55	1	1	1	1	1	1	1	1	1
Salishan	0.39	0.58	0.71	0.81	0.86	1.00	0.97	1.00	1	1	1	1	2	2	3	4	8
Salivan	0.09	0.14	0.20	0.29	0.35	0.52	0.68	0.82	1	1	1	1	1	1	1	1	2
Sepik-Ramu	0.67	0.82	0.97	1.00	1.00	1.00	1.00	1.00	1	1	2	2	3	4	6	8	16
Sino-Tibetan	0.87	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1	2	3	5	6	9	12	16	35
Siouan	0.26	0.40	0.57	0.68	0.79	0.81	1.00	1.00	1	1	1	1	1	2	2	3	5
Sko	0.12	0.18	0.25	0.36	0.48	0.62	0.71	0.82	1	1	1	1	1	1	1	1	2
South Caucasian	0.18	0.26	0.38	0.52	0.55	0.77	0.74	1.00	1	1	1	1	1	1	2	2	3
Subtiaba-Tlapanec	0.09	0.14	0.22	0.32	0.41	0.57	0.71	0.82	1	1	1	1	1	1	1	1	2
Tacanan	0.17	0.27	0.39	0.49	0.62	0.71	0.84	1.00	1	1	1	1	1	1	1	2	3
Ticuna	0.06	0.10	0.15	0.17	0.24	0.38	0.42	0.64	1	1	1	1	1	1	1	1	1
Tol	0.07	0.10	0.15	0.22	0.28	0.38	0.52	0.73	1	1	1	1	1	1	1	1	1
Tonkawa	0.06	0.10	0.15	0.22	0.31	0.43	0.45	0.55	1	1	1	1	1	1	1	1	1
Torricelli	0.49	0.66	0.80	0.90	0.97	1.00	1.00	1.00	1	1	1	2	2	3	4	5	9
Totonacan	0.13	0.20	0.28	0.41	0.52	0.62	0.68	0.82	1	1	1	1	1	1	1	1	2
Trans-New Guinea	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	5	8	10	14	19	28	36	80
Trumaí	0.06	0.09	0.15	0.20	0.28	0.33	0.45	0.46	1	1	1	1	1	1	1	1	1
Tsimané	0.06	0.09	0.13	0.22	0.21	0.29	0.39	0.36	1	1	1	1	1	1	1	1	1
Tucanoan	0.39	0.53	0.68	0.81	0.90	0.91	1.00	1.00	1	1	1	1	2	2	3	4	7
Tupi	0.60	0.84	0.95	0.98	0.97	1.00	1.00	1.00	1	1	2	2	3	4	5	7	14
Tuxá	0.06	0.10	0.15	0.22	0.31	0.38	0.48	0.64	1	1	1	1	1	1	1	1	1
Uralic	0.32	0.45	0.63	0.71	0.86	0.95	0.97	1.00	1	1	1	1	1	2	2	3	5
Uru-Chipaya	0.09	0.15	0.23	0.34	0.41	0.53	0.64	0.73	1	1	1	1	1	1	1	1	2
Uto-Aztecan	0.49	0.66	0.85	0.88	0.97	1.00	1.00	1.00	1	1	1	1	2	2	3	4	9
Vilela	0.06	0.09	0.13	0.17	0.21	0.28	0.32	0.45	1	1	1	1	1	1	1	1	1
Wakashan	0.11	0.16	0.23	0.37	0.45	0.52	0.71	0.91	1	1	1	1	1	1	1	1	2
Warao	0.06	0.10	0.15	0.22	0.31	0.38	0.52	0.64	1	1	1	1	1	1	1	1	1
Warembori	0.07	0.11	0.15	0.20	0.28	0.43	0.52	0.55	1	1	1	1	1	1	1	1	1

Table 13. *(Continued)*

Families	Sample size 30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
West Papuan	0.37	0.51	0.67	0.78	0.86	1.00	0.97	1.00	1	1	1	1	2	2	3	3	6
Witotoan	0.16	0.25	0.32	0.44	0.55	0.62	0.81	0.91	1	1	1	1	1	1	1	2	3
Yámana	0.06	0.10	0.15	0.22	0.28	0.38	0.48	0.73	1	1	1	1	1	1	1	1	1
Yale	0.06	0.10	0.15	0.22	0.28	0.38	0.45	0.55	1	1	1	1	1	1	1	1	1
Yanomam	0.09	0.14	0.21	0.22	0.38	0.38	0.58	0.82	1	1	1	1	1	1	1	1	2
Yenisei Ostyak	0.09	0.15	0.22	0.29	0.38	0.48	0.64	0.73	1	1	1	1	1	1	1	1	2
Yuchi	0.06	0.09	0.13	0.20	0.28	0.38	0.45	0.55	1	1	1	1	1	1	1	1	1
Yukaghir	0.10	0.15	0.21	0.29	0.41	0.52	0.61	0.64	1	1	1	1	1	1	1	1	2
Yuki	0.09	0.15	0.23	0.34	0.45	0.53	0.77	0.82	1	1	1	1	1	1	1	1	2
Yuracare	0.07	0.11	0.15	0.20	0.27	0.34	0.42	0.64	1	1	1	1	1	1	1	1	1
Zamucoan	0.09	0.15	0.23	0.30	0.41	0.57	0.68	0.82	1	1	1	1	1	1	1	1	2
Zaparoan	0.08	0.14	0.20	0.27	0.41	0.48	0.68	0.82	1	1	1	1	1	1	1	1	2
Zuni	0.06	0.10	0.15	0.22	0.31	0.34	0.45	0.55	1	1	1	1	1	1	1	1	1

Notice that families with the same DV (notably the Language Isolates) do not all have the same chance of being included in a sample that is not large enough to contain a representative from every independent language family. This is a side-effect of the randomizer, a mechanism to simulate pseudo-chance in a computer.

40-language samples generated on the basis of the DV method will contain a representative from the Afro-Asiatic family.[10]

3.5. *The influence of the classification on a sample*

Given the significant differences between the various classifications, one may wonder what influence a particular classification has on the distribution of languages over independent families in a sample. To investigate this, we compared the latest *Ethnologue* edition with Ruhlen's classification. For this purpose we often had to reclassify the families in the *Ethnologue* according to Ruhlen's more comprehensive proposal. This is illustrated in Table 14.

In this comparison we ignored Basque and Yenisei-Ostyak, which are treated as small families in the *Ethnologue* but are listed as the Language Isolates Basque and Ket in Ruhlen (1987). In order to be able to make a fair comparison between the two classifications, we also reduced the influence of the Language Isolates—9 in Ruhlen and 31 in the *Ethnologue*—by treating them as a single family. Finally, recall that the comparison is somewhat distorted by the fact that Ruhlen's classification includes both

Table 14. *Comparing Ruhlen and* Ethnologue *classifications*

Ruhlen	*Ethnologue*
Afro-Asiatic	Afro-Asiatic
Altaic	Altaic, Japanese
Australian	Australian
Austric	Austro-Asiatic, Austronesian, Daic, Hmong-Mien
Caucasian	North Caucasian, South Caucasian
Chukchi-Kamchatkan	Chukotko-Kamchatkan
Elamo-Dravidian	Dravidian
Eskimo-Aleut	Eskimo-Aleut
Indo-Hittite	Indo-European
Indo-Pacific	Amto-Musan, Andamanese, East Bird's Head, East Papuan, Geelvink Bay, Kwomtari-Baibai, Left May, Sepik-Ramu, Sko, Trans-NewGuinea, West Papuan
Khoisan	Khoisan
Language Isolates	Language Isolates
Na-Dene	Na-Dene
Niger-Kordofanian	Niger-Congo
Nilo-Saharan	Nilo-Saharan
Pidgins and Creoles	Pidgins & Creoles
Sino-Tibetan	Sino-Tibetan
Uralic-Yukaghir	Uralic, Yukaghir
Amerind	Alacalufan, Algic, ∴, Zamucoan, Zaparoan (i.e., all remaining groups in Table 13)

extant and extinct languages, whereas the *Ethnologue* only lists extant languages.[11]

Table 15 shows the distribution of languages in samples based on the classification in the *Ethnologue* (in samples of 100, 200, 300, 400, and 500 languages) after they have been reclassified according to Ruhlen's proposal. The first column under the figure indicating the sample size (Ru) gives the number of languages in a sample based on Ruhlen's classification; the second column (Et) specifies the number of languages in a sample of the same size that is based on the classification of the *Ethnologue*, but which is reclassified in terms of Ruhlen's classification (taking into account the modifications discussed above). The figure in the third column (under the symbol #) gives the proportional difference between the *Ethnologue* sample and the Ruhlen sample. For example, in a 100-language sample the DV method assigns six languages to Afro-Asiatic when we use Ruhlen's classification, and two languages when we take the *Ethnologue* classification; the difference is 2 : 6 = 0.3 (the closer to 1, the smaller the difference). We have omitted figures between 0.9 and 1.1, i.e., a difference of 10% or less, as well as those cases in which the difference is only one language in absolute terms.

Table 15 shows that the greatest differences occur in the relatively small samples. This is mostly due to the large number of families in the *Ethnologue* and the requirement that at least one language must be assigned to each of them. As the sample size increases the differences become smaller. This is particularly clear for Afro-Asiatic, Amerind, Nilo-Saharan, and Sino-Tibetan. Some of the figures are strikingly similar, as in the case of Indo-Hittite, Indo-Pacific, and Niger-Kordofanian. For samples containing over 300 languages, the mean difference for all families is under 3%. Note, finally, that Amerindian languages are always better represented in samples based on the *Ethnologue* classification than in samples based on Ruhlen's classification, apparently at the expense of the Afro-Asiatic and Australian families.

3.6. *Reducing the influence of Language Isolates in the sample*

In Section 2.3 we mentioned our initial policy to treat each Language Isolate as an independent family and to include each Language Isolate in a sample that is large enough to contain a representative from each independent family. However, since Language Isolates are predominantly found in Eurasia and the Americas this might lead to areal bias. Furthermore, one could argue that Language Isolates form a disproportionally big part of the families, especially in smaller samples. The *Ethnologue* classification lists over thirty Language Isolates (i.e., approximately 25% of the

Table 15. *Comparing samples based on Ruhlen (Ru) and* Ethnologue *(Et)*

Sample size:	100			200			300			400			500		
	Ru	Et	#	Ru	Et	#	Ru	Et	#	Ru	Et	#	Ru	Et	#
Afro-Asiatic	6	2	0.3	13	8	0.6	20	13	0.7	26	19	0.7	33	24	0.7
Altaic	2	2	·	4	3		5	5		7	7		9	9	
Amerind	20	52	2.6	42	67	1.6	63	85	1.3	84	100	1.2	105	122	1.2
Australian	8	2	0.3	16	9	0.6	24	15	0.6	32	21	0.7	40	27	0.7
Austric	15	8	0.5	32	29		48	49		65	69		81	88	
Caucasian	1	2		2	2		3	3		4	5		5	5	
Chukchi-Kamchatkan	1	1		1	1		1	1		1	1		2	1	
Elamo-Dravidian	1	1		2	2		3	3		4	4		4	5	
Eskimo-Aleut	1	1		1	1		1	1		2	1		2	2	
Indo-Hittite	5	2	0.4	9	8		14	13		19	18		23	23	
Indo-Pacific	14	14		29	28		44	43		58	58		73	73	
Khoisan	1	1		2	1		2	2		3	3		4	3	
Language Isolates	1	1		1	1		1	1		1	1		1	1	
Na-Dene	1	1		2	2		3	3		5	3	0.6	6	4	0.7
Niger-Kordofanian	10	4	0.4	21	19		32	32		43	46		53	59	
Nilo-Saharan	5	2	0.4	10	7	0.7	15	12	0.8	20	17	0.8	25	21	0.8
Pidgins & Creoles	2	1		3	4		5	6		6	9	1.5	8	11	1.4
Sino-Tibetan	5	1	0.2	9	6	0.7	14	10	0.7	18	14	0.8	23	18	0.8
Uralic-Yukaghir	1	2		1	2		2	3		2	4	2.0	3	4	

Table 16. *Number of Language Isolates in samples of various sizes when the Language Isolates are treated as one group (rather than singleton phyla)*

Size:	30	40	50	60	70	80	90	100
V&V:	0.83	1	1	1	2	2	2	2
Ethnologue:	0.10	0.17	0.25	0.40	0.67	0.91	1	1
Ruhlen:	1	1	1	1	1	1	1	1

Size:	125	150	175	200	250	300	400	500	1000
V&V:	3	4	4	5	6	7	10	12	20
Ethnologue:	1	1	1	1	1	1	1	1	2
Ruhlen:	1	1	1	1	1	1	1	1	2

families are Language Isolates), whereas Ruhlen has nine of them (i.e., approximately 30%). Therefore we have provided the sampling program with an option that constructs samples (following the same procedure that we explained above) in which all Language Isolates are treated as members of a family without an internal structure (hence this "family" has a DV of 1.50; see Section 3.2). Table 16 gives the number of Language Isolates in samples of different sizes. Complete matrices are provided in Appendices 1 and 2 (but only for the classifications in Ruhlen 1987 and the *Ethnologue* 1997).

4. Other approaches to language sampling

4.1. *Criteria for sampling*

Our focus has so far been on our own sampling method, which is especially designed to produce variety samples. Now we will turn to other approaches to language sampling in recent typological work and, where possible, compare them with our method. For this purpose we have generated DV samples of the same size as those used in the typological studies mentioned below. First, however, we will introduce some criteria that we think are relevant for any sampling strategy. Apart from the fact that we use these criteria so as to be able to compare different approaches to language sampling, we also hope to provide some material for a more general discussion about language sampling.

(a) EXPLICITNESS. Whatever the nature of one's sampling method, it must be clear what kind of classification is used and how sample languages are selected. Only then is it possible for others to reasonably assess the method and to decide whether this method suits their own typological investigations. Note furthermore that one can only make a meaningful comparison

between samples if they are constructed according to the same procedure and that only fully explicit sampling procedures can be formalized and computerized. A computer program for language sampling has several advantages. Apart from the fact that it speeds up the sampling procedure, it produces samples that are more consistent in their internal make up and consequently more reliable. Furthermore, it allows the typologist to experiment with different samples before deciding on the actual sample that (s)he is going to use. For example, one can experiment with samples of different sizes to see how languages are going to be distributed and to what extent certain families and areas will be covered; or one could try out different samples that vary with respect to the inclusion or exclusion of extinct languages.

(b) SAMPLE TYPE. We already mentioned in Section 2 that different research questions require different kinds of samples. For this reason it is important to know what kind of samples are generated by the method. Hence it must be clear in what respect a sample that is constructed by a particular method is representative for the languages of the world. In other words: Does the sampling method provide the researcher with a probability sample, a random sample, or a variety sample?

(c) CLASSIFICATION. The method must be able to generate samples on the basis of different classifications rather than depend on one particular proposal. The researcher must have the freedom to select his or her classification independently of the sampling method, which must only refer to formal properties of the classification.

(d) DOMAIN. The method must generate samples that have general applicability, i.e., it should not be directed at solving a highly specific research question or at problems pertaining to one particular linguistic domain (e.g., constituent order).

(e) SIZE. The method should not put any constraints on the size of the sample. Although large samples tend to suffer from various kinds of bias, it is ultimately the typologist who is responsible for the quality of the data.

(f) TREATMENT OF SPECIAL LANGUAGE TYPES. Another requirement that we want to mention here is that a sampling method must be entirely clear about its universe, i.e., the set of languages that will be sampled from. Although this is partially dependent on the kind of classification that is used (for example, the *Ethnologue* basically ignores extinct languages) it must be made explicit how the method deals with, for example, Language Isolates, Pidgin and Creole languages, mixed languages, unclassified languages, extinct languages, sign languages, secret languages, and artificial languages.

(g) FIT FOR THE JOB. Finally the method must be able to produce a sample that suits the problem under investigation. This may seem obvious, but

the method should also cater for the researcher who does not want to use a sample which is representative of ALL the world's languages. For example, one may only be interested in languages of a certain type (like ergative or VSO languages) and/or from a particular area (e.g., Europe). In such cases it should be specified what modifications were made and how they (may) have affected the results.

We believe that the DV method does well on most of these criteria. For example, every step in the sampling procedure is fully explicit and we indicated that it is designed to produce variety samples. We have shown that our sampling technique does not depend on any particular classification and that it can be used for a wide variety of research questions in all linguistic domains. As far as sample size is concerned, we mentioned that the lower limit is basically set by the number of independent families in the classification, but we also showed how to work around this problem if this is not possible or feasible (Section 3.4). Currently the only restriction concerns the format of the classification, which should have the graph theoretic properties of a tree structure (see Note 6).

Of course, whether sampling method x is better than method y is ultimately an empirical matter. More specifically, for variety samples it should be demonstrated for any concrete research question that a sample of size S generated by method x is consistently more varied with respect to certain linguistic parameter(s) than a sample constructed by method y.

In the following sections we will use criteria (a)–(g) to compare our method with the samples that were recently used in four typological studies.

4.2. *Bybee, Perkins, & Pagliuca's sample*

The main objective of *The Evolution of Grammar* (Bybee, Perkins, & Pagliuca 1994) was formulated as follows:

> The present work focuses on a formally defined set of linguistic elements—the grammatical morphemes associated with verbs. [...] Formally, grammatical morphemes may be affixes, stem changes, reduplication, auxiliaries, particles, or complex constructions such as English *be going to*. We refer to all of these types equally as grammatical morphemes and for convenience shorten this term to "gram" [note omitted]. Our study examines only those grams which have a fixed position with respect to the verb. Given a set of linguistic elements identified formally in this way, we ask whether there are any regularities or commonalities across languages in the meanings expressed by these elements.
>
> In particular our analysis focuses on some of the semantic domains most often marked morphologically on verbs—tense, aspect, and modality. (pp. 2–3)[12]

Ideally the sample used in Bybee et al. should have contained 94 languages, but due to the fact that 18 of the selected languages were not sufficiently documented the actual sample was reduced to 76 languages. These languages were stratified according to Voegelin & Voegelin (1977), because in the view of Bybee et al. this classification "provides the most complete and uniformly applied listing of the world's languages to date" (1994: 303). The classification has 80 families (including the 32 unclassified languages and extinct Language Isolates that we ignored in our listing in Table 10). In the sample languages were distributed over (sub)families "roughly proportional to their numbers in the universe of all languages" (1994: 303).

Before languages were selected, Bybee et al. divided the families into two groups: very small phyla ("minimal groups") and other phyla (families ranging in size between 22 and 1046 languages). The minimal groups contain one to six languages. Out of a total of 55 minimal groups (47 isolate/unclassified languages and eight groups with four to six members) only two languages were selected: Basque (Language Isolate) and Inuit. Apart from the one representative from the group of Pidgin and Creole languages (Tok Pisin) all remaining languages were selected from the 24 so-called "major groups" in the following way (1994: 29):[13]

> Phyla with fewer than forty members each contributed one language to the sample. [...] Phyla with more than forty languages contributed more than one language each, the number and affiliation of the selected languages depending on the subgrouping within the phylum.
>
> If the phylum has first-level or primary subgroups with more than twenty member languages, then one language was selected from each of these primary subgroups. If one of these primary subgroups is further divided into secondary subgroups of twenty languages or more, one language is selected from each of these subgroups.

Table 17 illustrates this part of the procedure (1994: 30). Thus, one language was selected from both Chadic subfamilies and one from each of the remaining primary subgroups. Primary subgroups that contained fewer than 20 languages were put together and languages were selected at the rate of one language for every 30 languages in the residual group. Only one language was selected from large secondary subgroups (Niger-Kordofanian, Austronesian, Indo-Pacific).

The method employed by Bybee et al. is in many ways remarkably similar to the method proposed by Rijkhoff et al. (1993), especially with respect to the way languages are distributed over (sub)families (i.e., proportionally across language families in a top-down fashion). But since the DV method is consistently applied to ALL families and subfamilies

Table 17. *The distribution of languages in the Afro-Asiatic family according to Bybee et al. (1994)*

Maximal Group	Primary Subgroup	Secondary Subgroup
Afro-Asiatic (209)	Chadic (108)	East Chadic (32)
		West Chadic (76)
	Berber (24)	
	Cushitic (29)	
	Omotic (23)	
	Semitic (24)	
	Egyptian* (1)	

* Primary residue.

there is no exact match between the Bybee et al. sample and a sample of the same size construed by our method (using the same classification, of course). This is shown in Table 18.

Table 18 shows that 16 languages (21.1%) in the actual sample of Bybee et al. would be assigned to a different family in a 76-language DV sample when languages are distributed over families according to Voegelin & Voegelin's classification. The major differences between the two samples are that, according to our method, Austronesian, Indo-European, and Sino-Tibetan are overrepresented, whereas Indo-Pacific and (cumulated) Amerindian are underrepresented.

A more detailed comparison between the two methods yields the following list of observations.

First, in both methods genetic affiliation is the major parameter for stratification.

Second, Bybee et al. want a representative from each family, but they only select two languages from the minor groups (small phyla and Language Isolates). Our method always includes at least one representative from each small family and allows the researcher to choose between including all Language Isolates, or putting them together, treating them as if they constituted an independent family.

Third, in our method languages are distributed over families and subfamilies according to the degree of linguistic diversity. Bybee et al. distribute languages over (sub)families on a quantitative basis. It is generally known, however, that certain small families such as Afro-Asiatic are relatively more diverse than numerically larger groupings such as the Bantu languages or languages of the Oceanic group (Bell 1978: 146). This suggests that one must not rely too heavily on the absolute number of languages in a family in an attempt to create linguistic diversity in the sample.

Table 18. *Bybee et al.'s actual (n=76, left column) and ideal (n=94, right column) samples compared with DV samples of the same size*

Family	n=76		DV	n=94		DV
Afro-Asiatic	5	→	3	6	→	3
Andean-Equatorial	5	→	4		5	
Australian		6		6	→	7
Austro-Asiatic	3	→	1	4	→	2
Austronesian	10	→	7	13	→	9
Aztec-Tanoan		1			1	
Caucasian		1			1	
Chimakuan	0	→	1	0	→	1
Chukchee-Kamchatkan	0	→	1	0	→	1
Dravidian		1			1	
Eskimo-Aleut		1			1	
Ge-Pano-Carib	1	→	2		3	
Hokan	1	→	2	1	→	3
Indo-European	5	→	3	5	→	4
Indo-Pacific	6	→	11		15	
Khoisan		1			1	
Language Isolates	1	→	2	1	→	2
Macro-Algonquian		1			1	
Macro-Chibchan	1	→	2	2	→	3
Macro-Siouan		1		1	→	2
Na-Dene		1			1	
Niger-Kordofanian		6			7	
Nilo-Saharan		3			3	
Oto-Manguean		1			1	
Penutian		3		3	→	4
Pidgins & Creoles		1			1	
Salish		1			1	
Sino-Tibetan	6	→	2	6	→	3
Ural-Altaic	3	→	1	3	→	2
Waicurian	0	→	1	0	→	1
Wakashan		0		0	→	1
Yeniseian	0	→	1	0	→	1
Yukaghir	0	→	1	0	→	1
Yuki	0	→	1	0	→	1
	n=76			n=94		
Change Family:	16 (21.1%)			14 (14.9%)		

Fourth, Bybee et al. only consider primary and secondary branches of language families, ignoring any lower-level divisions in the genetic trees. The DV method takes into consideration both the width and the depth of genetic trees.

Fifth, it is not clear how the method used by Bybee et al. can be applied to other classifications, nor how they would construct larger or smaller samples. Our method can be applied to different classifications and generates samples of varying sizes.

Sixth, Pidgin and Creole languages are regarded as one language family by both methods.

Seventh, both methods use mathematical criteria to determine the distribution of languages over (sub)families, thus making it theoretically possible to calculate for each language the probability of being included in the sample. In this regard either sampling method generates samples that can also be regarded as probability samples in the sense of Blalock (see Note 3). So although Bybee et al. speak of a stratified probability sample (1994: 27), we believe that their sample is better characterized as a variety sample from which probabilities could be extrapolated.[14]

These observations alone do not prove that the DV method would have produced a more varied sample than the one used in Bybee et al. This can only be demonstrated if we were to investigate the same phenomena in a DV sample. Nevertheless, some of the information in Bybee et al. suggests that a DV approach may have resulted is a richer sample as far as linguistic diversity is concerned. For example, the ideal 94-language sample contains thirteen languages from the Austronesian group, which are distributed over the three primary subgroups in the following way: one language belongs to the South Halmahera-West New Guinea subgroup, nine languages (but only seven in the actual sample) belong to the Oceanic branch, and three languages (two in the actual sample) belong to the Malayo-Polynesian family. ADV sample of the same size would have contained only two Oceanic languages and four representatives from the Malayo-Polynesian family. Judging from the information contained in the thirty or so tables in Bybee et al. (1994) it seems that relatively speaking the Malayo-Polynesian languages contribute more to the variety in the sample than the Oceanic languages (notably Tangga and Molu). This leads us to suspect that the sample would have displayed more diversity if it had contained more languages from Malayo-Polynesian, the Austronesian subgroup with the highest DV, and fewer Oceanic languages.

In sum, the sampling procedure in Bybee, Perkins, & Pagliuca, (1994) superficially shares many features with our method. However, there are also several important differences, and in the light of the criteria mentioned above we believe that the DV sampling method compares favourably with the method used by Bybee, Perkins, & Pagliuca.

4.3. *Dryer's sample*

Dryer has constructed a large probability sample that controls for both areal and genetic bias by grouping languages into equivalent genetic clusters or "genera", which are then grouped into six independent linguistic areas: Africa, Eurasia, South-East Asia & Oceania, Australia–New Guinea, North America, and South America (Dryer 1992: 83). Genera have a time depth of approximately 4000 years, a distance which is more or less equivalent to the Indo-European subfamilies such as Germanic and Celtic and which is considered to be sufficiently large to warrant areal independency between the genera. Dryer's sample is still expanding; the version that we used for the current investigation contains 736 languages which are distributed over 257 genera.[15] His sample includes Language Isolates, but Pidgin and Creole languages are not taken into account.

Dryer's method differs in a number of important ways from the methods designed by Bybee, Perkins, & Pagliuca and ourselves. Dryer's research is mostly concerned with Greenbergian correlations, and so as to avoid genetic and areal bias he counts genera (clustered in six non-overlapping geographic areas) rather than individual languages. Each genus is typically represented by a number of languages (between one and eight), and if a genus contains languages with different values for a certain parameter it is counted twice (e.g., Semitic has languages with SVO and SOV basic word order). A pattern is universal if there are no counterexamples in the genera of each of the six linguistic areas and it is a statistical preference if it prevails in the genera of all areas; otherwise a pattern is ascribed to genetic preferences or areal factors. To give an example, basic word order SOV is statistically preferred since it prevails over SVO order in the genera in all areas.

By using a time depth of approximately 4000 years Dryer employs a bottom-up approach to stratify his languages over different genetic groups. (Recall that individual languages are at the bottom end of the genetic language tree and that the top node stands for the proto-language from which all languages in the tree are supposed to have developed.) In contrast the methods employed by Bybee, Perkins, & Pagliuca and ourselves distribute languages over the subfamilies in a top-down fashion: first independent families, then primary branches, then secondary branches, etc. Since Dryer considers all groups with a time depth of approximately 4000 years, his sample also has a lower limit. That is to say, his smallest sample would be one in which each genus ($n=257$) is represented by a single language. Apart from empirical considerations, there appears to be no upper limit to the number of languages per genus. Presumably this is a consequence of the fact that Dryer wanted a

probability sample rather than a sample that is representative for the linguistic diversity among the world's languages.

Dryer has designed an interesting procedure to construct a very large sample that does not suffer from genetic and areal bias. Neverthless there also are certain drawbacks. For example, apart from the problem that (due to bibliographic restrictions) every large sample is likely to contain the same set of languages, Dryer's method is not fully explicit on the number of languages that a genus may or must contain. Furthermore, the size of Dryer's sample as well as the fact that his method is more or less tailored to his research questions (testing universals of the kind formulated in Greenberg (1963) and Hawkins (1983)) make it practically impossible for others to use the same method or the same kind of probability sample for their own typological investigations.

When, in spite of the obvious differences in methodology and sampling objective, we compare Dryer's sample with a DV sample of the same size to see to what extent his sample is representative of the linguistic variety, we get the result in Table 19 (using Ruhlen's classification; cf. Dryer 1989: 286). Table 19 shows that as many as 103 languages (14%) would be

Table 19. *Dryer's sample compared with a DV sample of the same size*

Family	Dryer		DV
Afro-Asiatic	39	→	48
Altaic	15	→	13
Amerind	220	→	152
Australian	51	→	58
Austric	96	→	118
Caucasian	12	→	8
Chukchi-Kamchatkan		3	
Elamo-Dravidian	12	→	7
Eskimo-Aleut	2	→	3
Indo-Hittite	49	→	34
Indo-Pacific	68	→	106
Khoisan	3	→	6
Na-Dene	10	→	8
Niger-Kordofanian	70	→	77
Nilo-Saharan	35	→	36
Pidgins & Creoles	0	→	12
Sino-Tibetan	38	→	33
Uralic-Yukaghir	7	→	5
Language Isolates	6	→	9
	n = 736		
Change Family:	103 (14.0%)		

distributed over different families in a DV sample of the same size. Considerably overrepresented from the DV perspective are Amerindian (+45%) and Indo-Hittite (+44%), whereas Indo-Pacific (−36%) and Austric (−19%) would be seriously underrepresented if Dryer's sample were to be used for variety purposes. One of the more conspicuous aspects of Dryer's sample is the absence of Pidgin and Creole languages whereas a 736-language DV sample would have contained twelve such languages. Since Dryer decided to leave these languages out of his sample, we instructed our sampling program to disregard all Pidgin and Creole languages and made a new DV sample of the same size, but this did not change the outcome in any interesting way.[16]

This exercise demonstrates that it is very important to use a sampling procedure that fits the purpose of the typological investigation. Probability samples such as those constructed by Dryer cannot simply be copied for exploratory research, since they are not likely to be representative of the linguistic diversity in the languages of the world to the extent that variety samples are. There is, however, a way to derive a subset of languages out of a non-variety sample which is representative for the linguistic diversity in the languages of the world. We will return to this option in the next section, which deals with the kind of sample that Leon Stassen uses in his typological investigations.

4.4. *Stassen's sample*

The next sample to be discussed here is the one used by Stassen in his study on intransitive predication (Stassen 1997). Although he does not follow a strict methodology, it is clear that he aims at a sample that is more or less representative for the linguistic diversity regarding various aspects of intransitive predication.

Again we compared the sample (410 languages) with a DV sample of the same size on the basis of Ruhlen's classification, which Stassen used along with Voegelin & Voegelin's (1977) and the *Ethnologue*'s. We decided to select Ruhlen's classification rather than one of the two others for the comparison, because our sampling program was most successful in matching the language names in Stassen's sample with the names in Ruhlen's classification: 85% of the languages in Stassen's sample were correctly identified (the success rate was only 75% when we tried to match the same sample with the *Ethnologue* classification).[17]

Before we were able to make the comparison we had to slightly modify Stassen's sample. For example, there are a number of languages in his sample (like Middle Egyptian and Moroccan Arabic) that are simply absent in Ruhlen's classification. In all such cases we replaced the missing

language with a language in Ruhlen's classification that has the same genetic history, a GENETIC SUBSTITUTE, so as not to distort genetic relations in the sample (for example, Middle Egyptian was replaced with Ancient Egyptian, and Moroccan Arabic with Western Arabic). However, this procedure did not work when certain subfamilies were already exhaustively represented in terms of Ruhlen's classification, i.e., when there were no genetic substitutes left to take the place of languages not listed in Ruhlen's classification. This was the case with Cairene Arabic and Seychellois: both languages are not mentioned in Ruhlen's classification and could not be replaced by a genetic substitute, since all languages with the same genetic specifications in Ruhlen's classification had already been selected in Stassen's sample. Therefore we decided to ignore these two languages in the comparison and to reduce the sample to 408 languages. Table 20 gives the results of this comparison for the language families.

Table 20 shows that 61 languages (almost 15%) belong to another family in a DV sample of the same size. Compared to a DV sample of 408 languages, families such as Altaic (+100%), Caucasian (+225%), Indo-Hittite (+100%), and Uralic-Yukaghir (+333%) are seriously

Table 20. *Stassen's sample compared with a DV sample of the same size*

Family	Stassen		DV
Afro-Asiatic	29	→	26
Altaic	14	→	7
Amerind	91	→	84
Australian	20	→	32
Austric	65	→	64
Caucasian	13	→	4
Chukchi-Kamchatkan		1	
Elamo-Dravidian	6	→	4
Eskimo-Aleut	1	→	2
Indo-Hittite	38	→	19
Indo-Pacific	25	→	58
Khoisan	1	→	3
Language Isolates	5	→	9
Na-Dene	4	→	5
Niger-Kordofanian	40	→	42
Nilo-Saharan	21	→	20
Pidgins & Creoles	4	→	7
Sino-Tibetan	16	→	18
Uralic-Yukaghir	13	→	3
	n=408		
Change Family:	61 (14.9%)		

overrepresented in Stassen's sample. Families like Australian (−38%) and Indo-Pacific (−57%), on the other hand, need a much stronger representation according to the DV method.

We already mentioned that Stassen aims at a more or less representative variety sample, but he has stated (personal communication) that he tends to select relatively more languages from (sub)families that he considers to be particularly interesting in the light of the phenomenon under investigation. Consequently his sample is somewhat domain-dependent (in that it is geared towards specific research questions) and biased with respect to certain (sub)families. Our sampling program can, however, also be used to find in an existing sample the largest subset of languages which together constitute a DV sample. When we applied this option to Stassen's original sample we found that the largest DV sample that can be derived from his sample contains 184 languages.

The possibility to create variety samples out of existing samples offers some obvious advantages for linguistic typology. To give an example, it allows us to derive variety samples from both variety and non-variety samples that were used in previous typological research, thus making them comparable to other (derived) variety samples.[18]

4.5. *Nichols' sample*

Nichols (1992) uses a sample of 174 languages to examine the areal distribution of selected morpho-syntactic patterns. Her sampling procedure is in several respects similar to Dryer's sample discussed above, hence the same general observations apply as made in Section 4.3. For example, she employs a bottom-up procedure to stratify languages among genetic groups and languages are selected from well-defined geographic "sample areas". Since Nichols locates her families at a somewhat greater time depth than Dryer, the languages are distributed over fewer genetic clusters than the languages in Dryer's sample, and because she makes finer areal distinctions the world is divided into ten (rather than six) so-called sample areas: Africa, Ancient Near East, Northern Eurasia, South and Southeast Asia, Oceania, New Guinea, Australia, North America, Mesoamerica, and South America. Nichols deals with the issue of language sampling in considerable detail and she argues that bottom-up samples combined with areal stratification are particularly suited for her type of investigation (i.e., areal patterning).

Although she aims for genetic (and areal) completeness, her sample has a number of gaps, which are mostly due to the lack of "usable descriptions" (notably involving language families of South America and New Guinea). Another reason why not every family is represented is that there

is an upper limit of six languages per independent family or "stock", which restricts the contribution of Indo-European, Niger-Congo, and Austronesian.[19] Each Language Isolate is taken to be a family in its own right, but the sample contains no Pidgin and Creole languages.

In selecting the languages, Nichols generally chose "the best-described" language of the family (1992: 37) and tried to exclude colonial languages such as Spanish (thus avoiding bias through language contact) as well as "languages judged by specialists in their families to be atypical of the family" (to this she adds: "[e]xcept that when selecting families from multi-branch stocks from which I could take several languages I made an affort to include the known diversity").

Since Nichols (like Dryer) uses families of a fixed time depth rather than some existing classification to stratify her sample languages, we had to stratify her sample languages according to one of the genetic classifications discussed earlier before we could make a comparison. For this purpose we decided to use Ruhlen's classification rather than the one published in the most recent edition of the *Ethnologue*, mainly because the latter does not list all languages in Nichols' sample.

Table 21. *Nichols' sample compared with a DV sample of the same size*

Family	Nichols		DV
Afro-Asiatic	5	→	11
Altaic	5	→	3
Amerind	66	→	34
Australian	19	→	13
Austric	10	→	26
Caucasian	3	→	2
Chukchi-Kamchatkan		1	
Elamo-Dravidian		2	
Eskimo-Aleut	0	→	1
Indo-Hittite	6	→	8
Indo-Pacific	26	→	24
Khoisan	3	→	1
Language Isolates	7	→	9
Na-Dene	3	→	2
Niger-Kordofanian	6	→	17
Nilo-Saharan	6	→	8
Pidgins & Creoles	0	→	3
Sino-Tibetan	2	→	8
Uralic-Yukaghir	4	→	1
	n = 174		
Change Family:	47 (27.0%)		

The most striking differences (see Table 21) concern Eskimo-Aleut (no languages), Afro-Asiatic, Austric, Niger-Kordofanian, and Sino-Tibetan, which are all underrepresented in terms of the DV method; Amerind, Australian, and Uralic, on the other hand, are all overrepresented. Finally, it is interesting to note that Nichols' sample has almost the same deviation from the DV figures (27.0%) as the reduced version of the only other bottom-up sample we have discussed, i.e., the one constructed by Dryer (28.4%; see Note 16).

5. Conclusions

In this article we have presented a revised version of the sampling method that was first proposed in Rijkhoff, Bakker, Hengeveld, & Kahrel (1993). The samples that are generated by this method are variety samples, which are most suitable for exploratory typological research. We have shown that this method can be used in combination with any classification that is presented in the form of a family tree.

We have discussed four other approaches to language sampling in the light of some general methodological criteria and tried to show that, to the extent that a comparison could be made, the DV method compares favourably with other sampling techniques, especially in the context of explorative research.

The final point we want to make is that it is more important for linguistic typology to strive for agreement on what constitutes a good sampling strategy (for a particular problem) than to defend or promote one particular method. Trying to understand the rules and principles that are responsible for the differences and similarities among the world's languages remains the main goal of linguistic typology. Sampling methods are only instruments that can bring us closer to that goal.

Received: 8 June 1998
Revised: 20 August 1998

University of Texas at Austin
Universiteit van Amsterdam

Appendix 1. *DV samples on the basis of Ruhlen (1987)*

	Sample size (all Language Isolates treated as one single family)																
	30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
Afro-Asiatic	2	2	3	4	4	5	6	6	8	10	11	13	16	20	26	33	66
Altaic	1	1	1	1	1	2	2	2	2	3	3	4	4	5	7	9	17
Amerind	4	6	9	11	14	15	18	20	26	31	36	42	52	63	84	105	211
Australian	2	3	4	4	5	6	7	8	10	12	14	16	20	24	32	40	80
Austric	4	5	7	8	11	12	14	15	20	24	28	32	40	48	65	81	162
Caucasian	1	1	1	1	1	1	1	1	1	2	2	2	3	3	4	5	10
Chukchi-Kamchatkan	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	3
Elamo-Dravidian	1	1	1	1	1	1	1	1	1	1	2	2	2	3	4	4	9
Eskimo-Aleut	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	4
Indo-Hittite	1	2	2	3	3	4	4	5	6	7	8	9	12	14	19	23	47
Indo-Pacific	3	5	6	8	9	11	12	14	18	21	25	29	36	44	58	73	146
Khoisan	1	1	1	1	1	1	1	1	1	1	2	2	2	2	3	4	8
Language Isolates	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
Na-Dene	1	1	1	1	1	1	1	1	1	2	2	2	3	3	5	6	11
Niger-Kordofanian	2	3	5	6	7	8	9	10	13	16	18	21	27	32	43	53	107
Nilo-Saharan	1	2	2	3	4	4	4	5	6	7	9	10	13	15	20	25	50
Pidgins and Creoles	1	1	1	1	1	1	2	2	2	2	3	3	4	5	6	8	16
Sino-Tibetan	1	2	2	3	3	4	4	5	6	7	8	9	11	14	18	23	45
Uralic-Yukaghir	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	6

Appendix 2. *Samples on the basis of* Ethnologue *(Grimes (ed.) 1997)*

	Sample size (all Language Isolates treated as one family)																
	30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
Afro-Asiatic	0.96	0.99	1.00	1.00	1.00	1.00	1	2	4	5	7	8	11	13	19	24	51
Alacalufan	0.10	0.17	0.29	0.40	0.50	0.82	1	1	1	1	1	1	1	1	1	1	1
Algic	0.42	0.61	0.83	0.83	1.00	1.00	1	1	1	1	1	1	2	2	3	4	7
Altaic	0.69	0.82	0.92	1.00	1.00	1.00	1	1	1	2	2	2	3	4	5	7	14
Amto-Musan	0.11	0.19	0.29	0.43	0.67	0.82	1	1	1	1	1	1	1	1	1	1	1
Andamanese	0.15	0.23	0.37	0.53	0.61	0.91	1	1	1	1	1	1	1	1	1	2	2
Araucanian	0.11	0.18	0.29	0.40	0.61	0.82	1	1	1	1	1	1	1	1	1	1	1
Arawakan	0.46	0.60	0.71	0.77	0.83	1.00	1	1	1	1	1	2	2	3	4	4	9
Arutani-Sape	0.10	0.18	0.29	0.43	0.56	0.91	1	1	1	1	1	1	1	1	1	1	1
Australian	0.99	1.00	1.00	1.00	1.00	1.00	1	2	4	6	7	9	12	15	21	27	56
Austro-Asiatic	0.92	0.99	1.00	1.00	1.00	1.00	1	1	3	4	5	5	7	9	13	16	34
Austronesian	1.00	1.00	1.00	1.00	1.00	1.00	2	4	9	13	17	21	28	35	50	63	136
Aymaran	0.11	0.18	0.29	0.43	0.61	0.82	1	1	1	1	1	1	1	1	1	1	1
Barbacoan-Paezan	0.24	0.37	0.56	0.70	0.89	0.91	1	1	1	1	1	1	1	1	2	2	4
Basque	0.10	0.18	0.29	0.43	0.67	1.00	1	1	1	1	1	1	1	1	1	1	1
Caddoan	0.19	0.32	0.46	0.60	0.83	0.82	1	1	1	1	1	1	1	1	1	2	3
Cahuapanan	0.10	0.17	0.23	0.37	0.50	0.82	1	1	1	1	1	1	1	1	1	1	1
Carib	0.41	0.56	0.71	0.97	0.94	1.00	1	1	1	1	1	2	2	3	3	4	8
Chapacura-Wanham	0.12	0.22	0.29	0.43	0.67	0.82	1	1	1	1	1	1	1	1	1	1	2
Chibchan	0.56	0.78	0.88	1.00	1.00	1.00	1	1	1	1	2	2	2	3	4	5	10
Choco	0.33	0.50	0.65	0.80	0.94	1.00	1	1	1	1	1	1	1	2	2	3	5
Chon	0.11	0.18	0.29	0.40	0.50	0.73	1	1	1	1	1	1	1	1	1	1	1
Chukotko-Kamchatkan	0.15	0.22	0.33	0.50	0.61	0.82	1	1	1	1	1	1	1	1	1	2	2
Daic	0.47	0.67	0.87	0.97	1.00	1.00	1	1	1	1	2	2	2	3	4	5	9
Dravidian	0.57	0.74	0.91	1.00	1.00	1.00	1	1	1	1	2	2	3	3	4	5	11

Appendix 2. *(Continued)*

	Sample size (all Language Isolates treated as one family)																
	30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
East Bird's Head	0.14	0.22	0.38	0.50	0.72	0.91	1	1	1	1	1	1	1	1	1	1	2
East Papuan	0.39	0.57	0.77	0.93	1.00	1.00	1	1	1	1	1	1	2	2	3	4	7
Eskimo-Aleut	0.19	0.31	0.50	0.60	0.72	0.82	1	1	1	1	1	1	1	1	1	2	3
Geelvink Bay	0.42	0.65	0.81	0.90	1.00	1.00	1	1	1	1	1	2	2	3	3	4	8
Gulf	0.10	0.18	0.27	0.43	0.61	0.73	1	1	1	1	1	1	1	1	1	1	1
Hmong-Mien	0.30	0.48	0.64	0.83	0.89	1.00	1	1	1	1	1	1	1	2	2	3	5
Hokan	0.53	0.72	0.83	0.97	1.00	1.00	1	1	1	1	2	2	3	3	4	5	11
Huavean	0.11	0.18	0.27	0.37	0.56	0.82	1	1	1	1	1	1	1	1	1	1	1
Indo-European	0.97	1.00	1.00	1.00	1.00	1.00	1	2	3	5	6	8	10	13	18	23	48
Iroquoian	0.27	0.40	0.58	0.70	0.89	0.91	1	1	1	1	1	1	1	2	2	2	5
Japanese	0.18	0.27	0.42	0.60	0.78	1.00	1	1	1	1	1	1	1	1	1	2	3
Jivaroan	0.12	0.21	0.33	0.46	0.67	0.73	1	1	1	1	1	1	1	1	1	1	2
Katukinan	0.10	0.17	0.23	0.40	0.50	0.82	1	1	1	1	1	1	1	1	1	1	1
Keres	0.09	0.15	0.25	0.33	0.50	0.64	1	1	1	1	1	1	1	1	1	1	1
Khoisan	0.37	0.54	0.75	0.87	0.94	1.00	1	1	1	1	1	1	2	2	3	3	7
Kiowa Tanoan	0.20	0.28	0.44	0.60	0.67	0.91	1	1	1	1	1	1	1	1	1	2	3
Kwomtari-Baibai	0.17	0.27	0.42	0.50	0.67	0.91	1	1	1	1	1	1	1	1	2	2	3
Language Isolates	0.10	0.17	0.25	0.40	0.67	0.91	1	1	1	1	1	1	1	1	1	1	2
Left May	0.10	0.18	0.27	0.37	0.56	0.64	1	1	1	1	1	1	1	1	1	1	2
Macro-Ge	0.74	0.94	0.98	1.00	1.00	1.00	1	1	1	2	2	3	4	5	6	8	16
Maku	0.10	0.18	0.27	0.43	0.67	0.82	1	1	1	1	1	1	1	1	1	1	2
Mascoian	0.11	0.18	0.29	0.47	0.67	0.91	1	1	1	1	1	1	1	1	1	1	2
Mataco-Guaicuru	0.14	0.22	0.38	0.53	0.61	1.00	1	1	1	1	1	1	1	1	1	1	2
Mayan	0.54	0.80	0.90	1.00	1.00	1.00	1	1	1	2	2	2	3	3	4	6	11
Misumalpan	0.09	0.15	0.23	0.30	0.50	0.82	1	1	1	1	1	1	1	1	1	1	1
Mixe-Zoque	0.26	0.39	0.58	0.77	0.89	1.00	1	1	1	1	1	1	1	1	2	2	4

Muskogean	0.14	0.22	0.33	0.40	0.67	0.64	1	1	1	1	1	1	1	1	1	1	2
Na-Dene	0.44	0.61	0.75	0.93	1.00	1.00	1	1	1	1	1	2	2	3	3	4	8
Nambiquaran	0.10	0.17	0.25	0.43	0.61	0.73	1	1	1	1	1	1	1	1	1	1	1
Niger-Congo	1.00	1.00	1.00	1.00	1.00	1.00	1	4	8	12	16	19	26	32	46	58	125
Nilo-Saharan	0.95	1.00	1.00	1.00	1.00	1.00	1	1	3	5	6	7	9	12	17	21	45
North Caucasian	0.38	0.55	0.77	0.90	0.89	1.00	1	1	1	1	1	1	2	2	3	3	6
Oto-Manguean	0.59	0.85	0.92	1.00	1.00	1.00	1	1	1	2	2	2	3	3	5	6	12
Panoan	0.46	0.68	0.81	0.97	0.94	1.00	1	1	1	1	2	2	2	3	4	5	9
Peba-Yaguan	0.10	0.18	0.25	0.40	0.56	0.73	1	1	1	1	1	1	1	1	1	1	1
Penutian	0.46	0.67	0.88	0.97	1.00	1.00	1	1	1	1	2	2	2	3	4	5	9
Pidgins & Creoles	0.84	1.00	1.00	1.00	1.00	1.00	1	1	2	3	3	4	5	6	9	11	22
Quechuan	0.19	0.31	0.46	0.60	0.78	0.91	1	1	1	1	1	1	1	1	1	2	3
Salishan	0.43	0.60	0.75	0.93	1.00	1.00	1	1	1	1	1	2	2	2	3	4	8
Salivan	0.09	0.15	0.23	0.37	0.50	0.73	1	1	1	1	1	1	1	1	1	1	1
Sepik-Ramu	0.73	0.88	0.96	1.00	1.00	1.00	1	1	2	2	2	3	4	5	6	8	17
Sino-Tibetan	0.93	1.00	1.00	1.00	1.00	1.00	1	1	3	4	5	6	8	10	14	18	38
Siouan	0.32	0.52	0.65	0.83	1.00	1.00	1	1	1	1	1	1	2	2	2	3	5
Sko	0.13	0.20	0.29	0.36	0.72	0.91	1	1	1	1	1	1	1	1	1	1	2
South Caucasian	0.18	0.31	0.38	0.57	0.67	0.82	1	1	1	1	1	1	1	1	1	2	3
Subtiaba-Tlapanec	0.11	0.17	0.25	0.40	0.55	0.64	1	1	1	1	1	1	1	1	1	1	1
Tacanan	0.19	0.24	0.41	0.57	0.67	0.82	1	1	1	1	1	1	1	1	1	2	3
Torricelli	0.51	0.72	0.85	0.97	1.00	1.00	1	1	1	1	2	2	2	3	4	5	9
Totonacan	0.13	0.22	0.34	0.43	0.61	0.82	1	1	1	1	1	1	1	1	1	1	2
Trans-New Guinea	0.99	1.00	1.00	1.00	1.00	1.00	1	3	6	8	11	13	18	22	32	40	86
Tucanoan	0.41	0.60	0.81	0.93	0.94	1.00	1	1	1	1	1	1	2	2	3	4	8
Tupi	0.67	0.85	0.94	1.00	1.00	1.00	1	1	1	2	2	2	3	4	6	7	15
Uralic	0.33	0.54	0.71	0.83	0.94	1.00	1	1	1	1	1	1	2	2	3	3	6
Uru-Chipaya	0.10	0.17	0.27	0.37	0.56	0.64	1	1	1	1	1	1	1	1	1	1	1
Uto-Aztecan	0.49	0.70	0.81	0.97	1.00	1.00	1	1	1	1	1	2	2	3	4	5	9
Wakashan	0.13	0.19	0.31	0.43	0.56	0.82	1	1	1	1	1	1	1	1	1	1	2

Appendix 2. *(Continued)*

	Sample size (all Language Isolates treated as one family)																
	30	40	50	60	70	80	90	100	125	150	175	200	250	300	400	500	1000
West Papuan	0.38	0.57	0.81	0.93	1.00	1.00	1	1	1	1	1	1	2	2	3	3	7
Witotoan	0.18	0.28	0.44	0.57	0.67	0.91	1	1	1	1	1	1	1	1	1	2	3
Yanomam	0.10	0.15	0.25	0.30	0.50	0.73	1	1	1	1	1	1	1	1	1	1	1
Yenisei Ostyak	0.09	0.17	0.23	0.36	0.56	0.64	1	1	1	1	1	1	1	1	1	1	1
Yukaghir	0.11	0.17	0.29	0.40	0.61	0.73	1	1	1	1	1	1	1	1	1	1	1
Yuki	0.10	0.18	0.29	0.43	0.67	0.91	1	1	1	1	1	1	1	1	1	1	1
Zamucoan	0.10	0.18	0.27	0.40	0.61	0.82	1	1	1	1	1	1	1	1	1	1	1
Zaparoan	0.09	0.15	0.23	0.37	0.61	0.82	1	1	1	1	1	1	1	1	1	1	2

Notes

Correspondence addresses: (Rijkhoff) 11803-B Argonne Forest Trail, Austin, Texas 78759-2224, USA; e-mail: Jan.Rijkhoff@hum.uva.nl, seibt@mail.utexas.edu; (Bakker) Instituut voor Algemene Taalwetenschap, Universiteit van Amsterdam, Spuistraat 210, NL-1012 VT Amsterdam, The Netherlands; e-mail: Dik.Bakker@hum.uva.nl

1. We are extremely grateful to Joseph and Barbara Grimes, who gave us permisson to use a digital version of the classification that is used in the 13th edition of the *Ethnologue*, and to Matthew Dryer and Leon Stassen, who provided us with their language samples. We would also like to thank Anna Siewierska for helpful comments on an earlier draft.
2. For example, Ruhlen's classification has only 17 independent language families (plus nine Language Isolates and a group of Pidgin and Creole languages); Dryer (1989) contends that certain linguistic areas are continental in size.
3. Cf. Blalock (1972: 509), who defines a probability sample as follows: "The distinguishing characteristic of a probability sample is that every individual must have a known probability of being included in the sample. In making statistical inferences it is not absolutely necessary that all probabilities be equal, since, if the probability of a selection is known, it will be possible to adjust for unequal probabilities by a weighting procedure of some kind" (cf. also Bybee et al. 1994: 304).
4. In both cases the number of families includes the Language Isolates.
5. Cf. also Perkins (1989: 311), who suggests that the distribution of languages across families should be based "on the degree of internal variation within families".
6. Apart from the fact that we are not aware of a complete, global classification that is based on the wave model (Ruhlen 1987: 256–257), we must point out that our sampling method is only suitable for classifications in the form of the family tree model. Note, however, that our method is not inherently restricted to genetic classifications. It will also work, for example, with classifications that stratify languages on the basis of geographical, typological, or cultural parameters—provided the classification has the hierarchical properties of a tree structure. (We are grateful to Edith Moravcsik for pointing this out.) For a recent criticism of the family tree model see Dixon (1997: 29), who argues that "the family tree model, while appropriate and useful in many circumstances, is not applicable everywhere and cannot explain every type of relationship between languages. We need a more inclusive model, which integrates together the ideas of family tree and diffusion area".
7. The main reason why we discuss Voegelin and Voegelin's twenty year old classification is that it is used in Bybee, Perkins, & Pagliuca (1994).
8. See also the DVs of Urarina-Waorani, Aymaran, and Quechuan in Table 9.
9. For instance, Bithynian is characterized as follows (Voegelin & Voegelin 1977: 81): "*Bithynian. A non-INDO-EUROPEAN language of ancient Asia Minor, known chiefly from Greek traditions (Meyer 1897; Holl 1908). *AFFILIATION*: none known."
10. Due to rounding the chance is in some instances as high as 1 (=100% chance), which is of course theoretically impossible; the real figure would be, e.g., 0.99999.
11. Nevertheless the number of languages listed in the *Ethnologue* (nearly 7000) is much higher than in Ruhlen (1987) (almost 5300 languages, about 10% extinct). Since this difference is mainly due to the fact that the *Ethnologue* lists more variants (read: dialects), this will mostly contribute to the number of terminal nodes (recall that our method ignores the top node and the terminal nodes in determining the DV value).
12. A little further we find (Bybee et al. 1994: 25): "Even though it is not possible in all cases discussed in this book, our ultimate goal would be to examine in detail the actual mechanisms of change from one meaning to another".

13. 55 (minimal groups)+24 (major groups)+1 (group of Pidgins & Creoles)=80 languages.
14. The sample in Bybee, Perkins, & Pagliuca (1994) is based on a method that was designed earlier by Perkins for a very specific purpose (see Perkins 1980). Since Perkins wanted to investigate a possible correlation between culture and grammar, he constructed a 50-language sample that was free of cultural bias. In other words, genetic representation was overruled by cultural factors (cf. also Bybee 1985: 25). However, in the expanded version of Perkins' probability sample that is used in Bybee et al. genetic considerations are more important (1994: 28): "Since it is known that one source of similarity among languages is genetic relation, and since we wanted our sample to represent all languages of the world, we wanted a sample which has languages from each of the known language families. In fact, independent research indicates that the optimal sample would contain [...] languages that are all as distantly related to one another as possible. Consequently, the sample had to be selected from a universe [...] that is stratified for genetic relations." In sum, it seems that Perkins' probability sample was turned into a sample that was also representative for genetic (linguistic) diversity.
15. The sample in Dryer (1989) contains 542 languages which are distributed over 218 genera. In Dryer (1992) the data base has been extended to 625 languages which are distributed over 252 genera.
16. If we take the bottom-up method as a viable alternative to the top-down approach, then the minimum sample would contain precisely one language for each genus. We constructed such a language sample by randomly selecting one language from each of Dryer's 257 genera. Interestingly, this sample turned out to deviate much more (28.4%) from the regular DV sample of the same size than Dryer's original sample. This difference is mostly due to the fact that the internal structure of the genera (subfamilies) is not taken into account by Dryer whereas it does play a role in the DV sampling method.
17. Since there is a certain degree of overlap in the list of language names (for instance, there are at least half a dozen languages called Pahari; see the list of language names in the *Ethnologue* Index), we verified that each language in the sample was indeed linked to its proper counterpart in Ruhlen's classification.
18. To determine the maximum size of a potential DV sample in an existing non-DV sample we have provided the sampling program with a procedure that controls the influence of families that—due to their poor representation in the original sample—keep the size of the maximum DV sample extremely low. For instance, Stassen's sample contains only one language from the Khoisan family, which means that a straightforward application of this extra procedure yields a proper DV sample that contains only 184 languages, i.e., this is the largest possible DV sample with just one Khoisan language. It is possible, however, to specify a certain degree of deviation from a proper DV sample. For example, one can instruct the sampling program to construct a DV sample with a maximum deviation of one language (per family). In that case the program will produce a sample in which no language family deviates by more than one language from a proper DV sample. Relaxing the lower limit makes it possible to increase the size of the subset and to preserve better the variety in the original sample.
19. There are, however, "two groups lower than the stock from which two languages were taken: Semitic [...] and Finno-Ugric [...]" (Nichols 1992: 35). She also states that the incomplete coverage of South American families does not seriously affect the results: "[...], the genetic diversity in South America is such that covering even a minority of the families gives a sample comparable in size to that for Africa, Northern Eurasia, or Australia; hence it can be regarded as adequately covered for purposes of sample

design" (Nichols 1992: 26). Note also that Nichols' "stocks" are no older than 8,000 years (1992: 25).

References

Bakker, Dik (1997). Flexibility and consistency in word order patterns in the languages of Europe. In Siewierska (ed.) 1997, 383–419.

Bell, Alan (1978). Language samples. In Greenberg et al. (eds.) 1978, Volume 1, 123–156.

Blalock, Hubert M., Jr. (1972). *Social Statistics*. 2nd edition. New York: McGraw Hill.

Bybee, Joan L. (1985). *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: Benjamins.

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.

Dixon, Robert M. W. (1997). *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

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

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

Forchheimer, Paul (1953). *The Category of Person in Language*. Berlin: de Gruyter.

Greenberg, Joseph H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of Language*, 73–113. Cambridge, Mass.: MIT Press.

Greenberg, Joseph H., Charles A. Ferguson, & Edith A. Moravcsik (eds.) (1978). *Universals of Human Languages. Vol. 1: Method—Theory. Vol. 2: Phonology. Vol. 3: Word Structure. Vol.4: Syntax*. Stanford: Stanford University Press.

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

Grimes, Joseph & Barbara F. Grimes (1993). *Ethnologue: Language Family Index*. Dallas: Summer Institute of Linguistics.

Hammond, Michael, Edith A. Moravcsik, & Jessica R. Wirth (eds.) (1988). *Studies in Syntactic Typology*. Amsterdam: Benjamins.

Hawkins, John A. (1983). *Word Order Universals*. New York: Academic Press.

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

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

Perkins, Revere D. (1980). The evolution of culture and grammar. Ph.D. dissertation, State University of New York at Buffalo.

—(1988). The covariation of culture and grammar. In Hammond et al. (eds.) 1988, 359–378.

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

—(1992). *Deixis, Grammar, and Culture*. Amsterdam: Benjamins.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld, & Peter Kahrel (1993). A method of language sampling. *Studies in Language* 17: 169–203.

Ruhlen, Merritt (1987). *A Guide to the World's Languages. Volume 1: Classification*. London: Edward Arnold.

Siewierska, Anna (ed.) (1997). *Constituent Order in the Languages of Europe*. Berlin: Mouton de Gruyter.

Stassen, Leon (1985). *Comparison and Universal Grammar*. Oxford: Blackwell.

—(1997). *Intransitive Predication*. Oxford: Oxford University Press.

Tomlin, Russell S. (1986). *Basic Word Order: Functional Principles*. London: Croom Helm.
Voegelin, Charles F. & Florence M. Voegelin (1977). *Classification and Index of the World's Languages*. New York: Elsevier.
Whaley, Lindsay J. (1997). *Introduction to Typology: The Unity and Diversity of Language*. London: Sage.