2. The sample

2.0. Introduction

The sample on which the typological part of this study is based was created using the method proposed in Rijkhoff et al. (in press). A brief description of this method is given in 2.1. The resulting sample is presented in 2.2. Some presentational matters are dealt with in 2.3.

2.1. Sampling method

Rijkhoff et al. (in press) proposes a (computerized) sampling method, the primary aim of which is to create samples in which the differences between individual sample languages are maximal. In order to achieve this, pride of place is given to a genetic criterion, rather than to a geographic or typological one. It is assumed that the quality of a language sample is affected worst if languages are too closely related genetically. The genetic classification used is Ruhlen (1987). Note that although aspects of this classification might itself be questioned, it is here taken for granted.

In order to create maximal genetic diversity within samples the method consists of two components, which together account for variation both across and within phyla. The first component makes sure that every major phylum is represented by at least one member. This step is fully in line with the major objective of the sampling method: creating diversity within the sample, since a major phylum is posited only in those cases in which it is supposed not to have any genetic affiliations with another major phylum. Notice that as a consequence of this approach every language isolate, constituting a phylum by itself, will be represented in the sample. The second component makes sure that the number of languages by which a phylum is represented correlates proportionally with the linguistic diversity within that phylum. Bell (1978) was probably the first to draw attention to the problem of the different degrees of internal complexity of phyla. He notes that, for instance, one expects "to learn more from the 200 or so AFROASIATIC languages than from the much more homogeneous BANTU languages, though they number over 300" (Bell 1978:146). In order to tackle this problem one needs a technique that measures the diversity within a group of genetically related languages.

The technique proposed in Rijkhoff et al. (in press) is based on the assumption that the algebraic structure of a genetic language tree reflects the linguistic diversity within the phylum it represents. It consists of the computation of a factor, called Diversity Value (DV), which takes into consideration both depth and width of a genetic language tree. Consider the representation of the Uralic-Yukaghir sample in Figure 6.

	Depth → 1	2	3	4	5
W	Yukaghir	•	•	•	•
i	Uralic	Samoyed	North	•	•
d			South	•	•
t		Finno-Ugric	Ugric	Hungarian	•
h				Ob-Ugric	•
			Finnic	Permic	•
↓				Volgaic	•
				North	Saamic
					Baltic-Finnic
	2	3	5	8	9

Figure 6. The Uralic-Yukaghir phylum.

Figure 6 gives the internal structure of the genetic language tree of the Uralic-Yukaghir phylum, ignoring the top node as well as the terminal nodes (the individual languages). It shows that every separate branch adds to the width of the phylum as a whole. The width of every level can be calculated, as has been done in Figure 6. The resulting figures may then be used to calculate the DV of the phylum as a whole, yielding an objective measure of the linguistic diversity within the Uralic-Yukaghir phylum, which may be compared with the DV values of other phyla.

The actual procedure by means of which the DV values of phyla are calculated is described extensively in Rijkhoff et al. (in press). Here it may suffice to say that, since the distinguishing power of levels diminishes when going down the genetic language tree, a decreasing weight is assigned to the contribution (in terms of nodes) of deeper levels. A 40-language sample construed on the basis of this two-step procedure is given in Table 1. In this table the following information between brackets follows each (sub)phylum name: (i) the DV of the (sub)phylum, (ii) the number of daughter nodes of the (sub)phylum, and (iii) the number of languages within the (sub)phylum. Branching of the genetic tree is shown by means of indentation.

As Table 1 shows, in some cases the two-step procedure has to be applied more than once. This situation occurs when the sample languages assigned to a phylum outnumber the primary branches of that phylum, as, for instance, in the case of Niger-Kordofanian. Three sample languages should be selected from this phylum, which has only two primary branches. In order to determine the number of languages to be selected from each primary branch the procedure described above is repeated: each subphylum is represented by at least one member, and the remaining languages are distributed over the subphyla according to their DV.

Table 1. A 40-language sample

Afro-Asiatic (55.53/6/258)	2		
Altaic (14.79/2/66)	1		
Amerind (178.44/6/854)	5		
Australian (67.58/30/262)	2		
Austric (137.41/3/1186)	4		
Austro-Tai (106.03/2/1027)		2	
Daic (4.67/2/57)			1
Austronesian (118.17/4/970)			1
Austroasiatic (28.08/2/155)		1	
Miao-Yao (2.00/2/4)		1	
Basque (0.00/0/1)	1		
Burushaski (0.00/0/1)	1		
Caucasian (8.54/2/38)	1		
Chukchi-Kamchatkan (2.47/2/5)	1		
Elamo-Dravidian (7.43/2/29)	1		
Eskimo-Aleut (3.34/2/9)	1		
Etruscan (0.00/0/1)	1		
Gilyak (0.00/0/1)	1		
Hurrian (0.00/0/1)	1		
Indo-Hittite (39.71/2/180)	1		
Indo-Pacific (124.79/13/748)	3		
Ket (0.00/0/1)	1		
Khoisan (6.97/3/33)	1		
Meroitic (0.00/0/1)	1		
Na-Dene (9.44/2/41)	1		
Nahali (0.00/0/1)	1		
Niger-Kordofanian (90.38/2/1068)	3		
Niger-Congo (90.07/2/1036)		2	
Niger-Congo Proper (89.68/2/1007)			1
Mande (9.30/3/29)			1
Kordofanian (9.51/2/32)		1	
Nilo-Saharan (42.18/9/138)	1		
Pidgins and Creoles (13.47/13/38)	1		
Sino-Tibetan (38.52/2/268)	1		
Sumerian (0.00/0/1)	1		
Uralic-Yukaghir (4.93/2/27)	1		

In other cases the primary branches of a phylum outnumber the sample languages to be selected from that phylum. For instance, five languages have to be selected from the Amerindian phylum, which has six primary branches. Here the five sample languages can simply be chosen from five different primary branches.

Although this whole procedure is based on a genetic criterion, additional care should be taken to avoid geographic bias. The geographic restriction proposed in Rijkhoff et al. is that no two languages that are spoken in contiguous regions be included in a sample. In cases in which the genetic and geographic criterion are in conflict, precedence is given to the genetic one.

It is important to note that Table 1 represents an ideal sample. The actual sample may be different from this ideal one due to bibliographic restrictions. For instance, a 40-language sample should contain the extinct isolates Etruscan and Meroitic, but too little is known about these languages to allow for their inclusion in any sample, and too little is known about the system of non-verbal predication of Hurrian, another extinct isolate, to allow for its inclusion in the present sample. The gaps resulting from the absence of these three languages from the sample are not filled with other languages, since this would distort the proportions within the sample. Thus, an ideal 40-language sample corresponds to an actual 37-language sample in this study.

2.2. Description of the sample

The languages in the sample on which this study is based are distributed over the phyla as indicated in Table 1. The languages selected are listed in Table 2. These languages were selected non-randomly, the major criteria for their inclusion being the availability of reliable descriptions on the one hand, and an acceptable geographic distribution on the other. The sources of information on the sample languages are listed in Table 3, and their geographic distribution is shown in the map on pages 20-21.

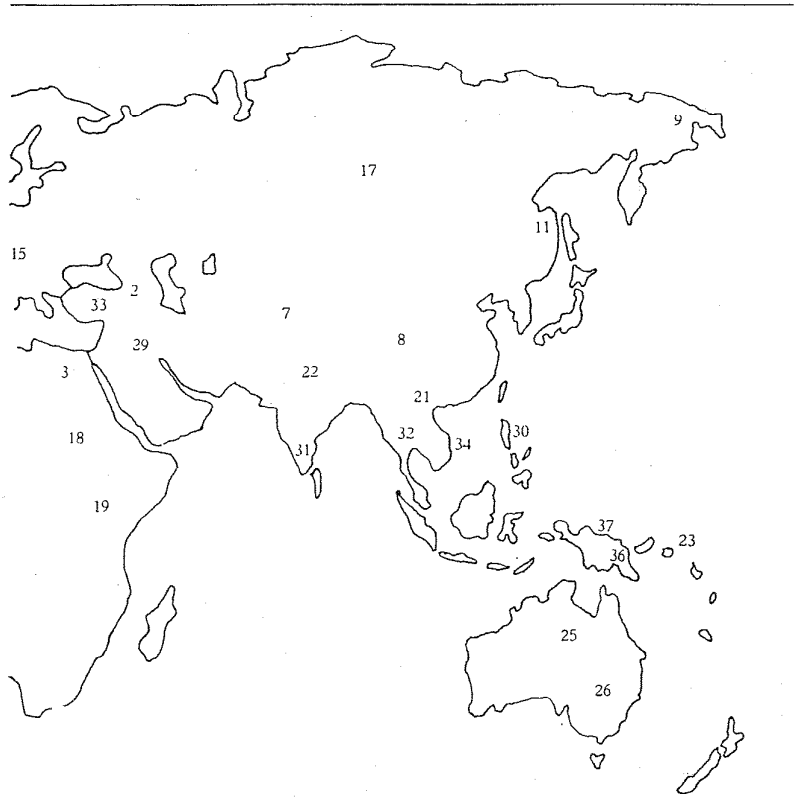
A serious problem with respect to the geographic distribution occurs in South-East Asia. The Miao-Yao languages are spoken in small regions scattered all over Northern Vietnam, Laos, Thailand, and Southern China. It is virtually impossible to select a language from this family without its being in contact with another language that has to be included in the sample on the basis of the genetic criterion. Two other languages spoken in contiguous regions are Abkhaz and Turkish. I included these two languages in the sample for bibliographic reasons.

Table 2. Genetic affiliations of the languages in the sample

Afro-Asiatic (2)	Chadic (1)		Hausa
	Semitic (1)		Arabic, Egyptian
Altaic (1)			Turkish
Amerind (5)	Northern (1)		Mam
	Andean (1)		Quechua
	Equatorial-Tucanoan (1)		Guaraní
	Ge-Pano-Carib (1)		Hixkaryana
	Central Amerind (1)		Pipil
Australian (2)	Gunwinyguan (1)		Ngalakan
	Pama-nyungan (1)		Ngiyambaa
Austric (4)	Austro-Tai (2)	Daic (1)	Thai
		Austronesian (1)	Tagálog
	Austroasiatic (1)		Vietnamese
	Miao-Yao (1)		Miao
Basque (1)			Basque
Burushaski (1)			Burushaski
Caucasian (1)			Abkhaz
Chukchi-Kamchatkan (1)			Chukchee
Elamo-Dravidian (1)			Tamil
Eskimo-Aleut (1)			West Greenlandic
Etruscan (1)			—
Gilyak (1)			Gilyak
Hurrian (1)			—
Indo-Hittite (1)			Dutch
Indo-Pacific (3)	Trans New Guinea (1)		Yagaria
	Sepik-Ramu (1)		Yessan-Mayo
	East Papuan (1)		Nasioi
Ket (1)			Ket
Khoisan (1)			!Xũ
Meroitic (1)			—
Na-Dene (1)			Navaho
Nahali (1)			Nahali
Niger-Kordofanian (3)	Niger-Congo (2)	N.-C. Proper (1)	Babungo
		Mande (1)	Bambara
	Kordofanian (1)		Krongo
Nilo-Saharan (1)			Lango
Pidgins and Creoles (1)			Jamaican Creole
Sino-Tibetan (1)			Mandarin Chinese
Sumerian (1)			Sumerian
Uralic-Yukaghir (1)			Hungarian

Map. Approximate location of the languages in the sample

1. !Xũ (Namibia)
2. Abkhaz (Caucasus)
3. Arabic, Eg. (Egypt)
4. Babungo (Cameroon)
5. Basque (S France, N Spain)
6. Bambara (Mali, Gambia, Senegal)
7. Burushaski (N Pakistan)
8. Chinese, Mandarin (China)
9. Chukchee (NE Siberia)
10. Dutch (Netherlands, Belgium)
11. Gilyak (E Siberia, N Japan)
12. Guaraní (Paraguay)
13. Hausa (Nigeria, Niger)
14. Hixkaryana (N Brasil)
15. Hungarian (Hungary)
16. Jamaican Creole (Jamaica)
17. Ket (C Siberia)
18. Krongo (N Sudan)
19. Lango (Uganda)



20. Mam (Guatemala, Mexico)
21. Miao (China, Laos, Vietnam)
22. Nahali (NE India)
23. Nasioi (Bougainville Isl.)
24. Navaho (SW United States)
25. Ngalakan (N Australia)
26. Ngiyambaa (SE Australia)
27. Pipil (El Salvador)
28. Quechua, Imb. (Ecuador)
29. Sumerian (Mesopotamia)
30. Tagálog (Philippines)
31. Tamil (S India, Sri Lanka)
32. Thai (Thailand)
33. Turkish (Turkey)
34. Vietnamese (Vietnam)
35. W.-Greenlandic (Greenland)
36. Yagaria (NE New Guinea)
37. Yessan-Mayo (N New Guinea)

Table 3. Sources of information on the languages in the sample

!Xũ	Snyman (1970, personal communication), Köhler (1981).
Abkhaz	Hewitt (1979), Spruit (1986, personal communication).
Arabic, Egyptian	Anwar (1979), Olmstedt Gary—Gamal-Eldin (1982), Eid (1983), Cuvalay-Haak (personal communication).
Babungo	Schaub (1985).
Bambara	Brauner (1974).
Basque	Lafitte (1944), Saltarelli (1988).
Burushaski	Lorimer (1935-1938), Berger (1974).
Chinese, Mandarin	Li—Thompson (1977, 1981), van den Berg (1989).
Chukchee	Bogoras (1922), Nedjalkov, V. (personal communication).
Dutch	Author.
Gilyak	Nakanome (1927), Nedjalkov, I. (personal communication).
Guaraní	Gregores—Suárez (1967).
Hausa	Kraft—Kraft (1973), Cowan—Schuh (1976).
Hixkaryana	Derbyshire (1979).
Hungarian	Kiefer (1968), de Groot (1989, personal communication).
Jamaican Creole	Bailey (1966), Beck (personal communication), Veenstra (personal communication).
Ket	Castrén (1858).
Krongo	Reh (1985).
Lango	Noonan (1981).
Mam	England (1983).
Miao	Miao Language Team (1972), Lyman (1979).
Nahali	Kuiper (1962).
Nasioi	Rausch (1912), Hurd—Hurd (1966).
Navaho	Schauber (1979), Young—Morgan (1987).
Ngalakan	Merlan (1983).
Ngiyambaa	Donaldson (1980).
Pipil	Campbell (1985).
Quechua, Imbabura	Cole (1982).
Sumerian	Thomsen (1984).
Tagálog	Schachter—Otanes (1972).
Tamil	Asher (1982).
Thai	Noss (1964), Kuno—Wongkhomthong (1981).
Turkish	Lewis (1967), Lees (1972), Ersen-Rasch (1980), van Schaaik (1983, personal communication), Tura (1986).
Vietnamese	Le-van-Ly (1948), Thompson (1965), Nguyên Dăng Liêm (1969, 1975).
West Greenlandic	Fortescue (1984, personal communication), Kristoffersen (personal communication).
Yagaria	Renck (1975), Haiman (1980).
Yessan-Mayo	Foreman (1974)

2.3. Matters of presentation

In following chapters sample languages will be referred to as belonging to the (sub)phylum that triggered their inclusion in the sample. Thus, Miao will be referred to as a Miao-Yao language rather than as an Austric language, Dutch will be referred to as an Indo-Hittite language rather than as a Germanic language. Languages from singleton phyla will be referred to as Isolates.

In the course of this study reference will be made to languages other than those included in the sample. These will be referred to as if they were sample languages. Thus, Spanish will be referred to as an Indo-Hittite rather than as a Romance language, and Tongan as an Austronesian rather than Austric one.

Example sentences are preceded by a line giving the following information: *Language name* ((Sub-)phylum-name; Source references). For the interlinear morphemic translations I largely follow the directions given in Lehmann (1982b). An exception concerns the use of a dot to separate parts of the interlinear morphemic translations, which covers Lehmann's semi-colon, which separates distinct parts of an interlinear morphemic translation in those cases in which the exact morpheme boundaries are not or cannot be established in the original text, and his dot, which separates distinct parts of multi-word translations of a single morpheme in the original text.

Literal translations between double qotation marks precede free translations between single quotation marks in those cases in which this may facilitate the processing of the sentence under consideration.