A note on language sampling

(by Kees Hengeveld, Department of Spanish, U Amsterdam)

When one is not in a position to study all the units (in our case: languages) that make up the universe of one's research (in our case: the extant and extinct languages of geographic Europe), then one cannot do without an adequate sampling strategy. Within EUROTYP little or no attention has been given so far to the relevance of such a strategy. There may at least be two reasons for this situation:

(i) From a linguistic point of view, the boundaries of Europe are relatively arbitrary. However, if the results of a project that studies the typology of European languages are to lead to valuable insights and interesting hypotheses concerning the typology of the world's languages, then one should have an adequate sampling technique right from the start.

(ii) Many projects aim at full coverage of the languages of Europe. If these projects succeed in their aim then indeed no sampling strategy is necessary. However, as long as some languages remain uncovered due to the unavailability of data, we will indeed be working with samples. It is crucial to note that it is quite possible to have a sample of 100 out of the approximately 150 European languages which is heavily biased genetically and geographically. Thus even if one strives for full coverage it is advisable to build up the database along the lines prescribed by some sampling procedure.

Since typological research isprimarily concerned with the range of variation found across languages (cf. Comrie 1981:31), the samples one uses within this type of research should display the highest possible degree of diversity. The primary aim of the (computerized) sampling method we propose in Rijkhoff et al. (fc.) is therefore to create samples in which the differences between individual sample languages are maximal. In order to achieve this aim, 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). The method consists of two components, which together account for variation both across and within phyla, i.e. major genetic groupings. 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. 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) has drawn attention to various degrees of internal complexity of phyla and has pointed out that, for example, we are likely '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 phylum.

The technique Bell uses is to determine for each phylum how

many groups of languages it contains that are separated from a common ancestor by a time-depth of 3 500 years. The number of groups within a phylum then serves as the basis for determining the number of languages by which that phylum is to be represented in a sample. A problem with this approach is that for many phyla too little is known about their history to determine the number of groups they contain.

The alternative technique proposed in Rijkhoff et al. (fc.) circumvents this problem. It is based on the idea that the algebraic structure of a genetic language tree reflects the linguistic diversity within the phylum it represents and consists of the computation of a factor, called Relative Weight (RW), which takes into consideration both depth and width of genetic language trees. The actual procedure in which the RW values of phyla is calculated is described and motivated ex-tensively in Rijkhoff et al. (fc.). 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.

The method as a whole is applied recursively, that is, once it has been determined how the sample languages should be distributed over the major phyla, the method is repeated in order to determine how the languages assigned to each phylum should be distributed over its subphyla, sub-subphyla, etc.

This sampling method can be applied quite easily to the (genetic classification of) European languages. I hope to report more extensively on the outcome of this application in the near future. Here I will restrict myself to illustrating some aspects of the resulting samples.

European languages represent six of the major phyla recognized in Ruhlen (1987): AFRO-ASIATIC, ALTAIC, BASQUE, CAUCASIAN, INDO-HITTITE, and URALIC-YUKAGIR. In several cases only particular branches of these phyla are relevant for the classification of European languages. The major European phyla to be recognized are ARABIC, TURKIC, BASQUE, CAUCASIAN, INDO-EUROPEAN, and FINNO-UGRIC. The first component of the sampling method makes sure that each of these phyla will be represented in the sample by at least one member. Note that, consequently, Basque and Maltese, being the sole members of their respectivephyla (BASQUE and ARABIC), will be represented in every European sample. The second component of the sampling method distributes the remaining languages over the phyla proportionally, i.e. according to the relative weight of these phyla. By way of illustration I give the number of languages by which each phylum is represented in a 25-language sample:

ARABIC (RW=0.000)	1
TURKIC (RW=3.000)	2
BASQUE (RW=0.000)	1
CAUCASIAN (RW=6.667)	4
INDO-EUROPEAN (RW=23.901)	15
FINNO-UGRIC (RW=3.000)	2
TOTAL	25

Repeated application of the method is called for in order to determine how the sample languages should be distributed over the subphyla, sub-subphyla etc. The final result for a 25-language sample is shown below. In this overview a '>' should be read as 'to be subdivided into'.

ARABIC (RW=0.000)	1
TURKIC (RW=3.000)	2 >
BOLGAR	1
COMMON TURKIC	1
BASQUE (RW=0.000)	1
CAUCASIAN (RW=6.667)	4 >
SOUTH (RW=3.000)	1
NORTH (RW=6.120)	3 >
NORTHWEST (RW=3.000)	1
NORTHEAST (RW=4.813)	2 >
NAX	1
DAGHESTAN	1
INDO-EUROPEAN (RW=23.901)	15 >
ARMENIAN (RW=0.000)	1
INDO-IRANIAN (RW=2.875)	2 >
INDIC	1
IRANIAN	1
ALBANIAN (RW=0.000)	1
GREEK (RW=0.000)	1
ITALIC (RW=5.656)	4 >
OSCO-UMBRIAN (RW=0.000)	1
LATINO-FALISCAN (RW=6.653)	3 >
FALISCAN	1
LATIN	1
ROMANCE	1
CELTIC (RW=2.250)	1
GERMANIC (RW=4.444)	3 >
EAST	1
NORTH	1
WEST	1
BALTO-SLAVIC (RW=3.889)	2 >
BALTIC	1
SLAVIC	1
FINNO-UGRIC (RW=3.000)	2 >
UGRIC	1
FINNIC	1
TOTAL	25

Note that a sample should not only be genetically representative, but also geographically. Roughly speaking, it should be avoided, wherever this is possible, to include languages that are spoken in adjacent regions in a sample. For every sample size a separate calculation is required. A specification of European samples of different sizes will be available in the near future.

Bell, Alan. 1978. "Language samples." In: Greenberg, Joseph H. (ed.) *Universals of human language. Vol. I: Method & theory*, 123-156. Stanford: Stanford University Press

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

Rijkhoff, Jan & Bakker, Dik & Hengeveld, Kees & Kahrel, Peter. Forthcoming. "A method of language sampling."

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