# Language Sampling (Bakker 2007)

## 1. Introduction

Linguists interested in exploring the variation and distribution of linguistic phenomena in the languages of the world face a problem of selection. The number of languages currently spoken is around 7,000. This number is too large for any research project. As in other cases where a population is too large to be studied in full, one has to restrict oneself to a small but representative subset by taking a sample. However, there are a number of complications which prevent a linguist from applying a random selection and stratifying the sample for parameters known to interact with the research variables, as is common in many domains of empirical research.

The most outstanding problem is that for about two thirds of the existing languages, no grammar is available. Most of these languages are spoken in isolated areas and belonging to under-investigated language groups, which potentially show unique features. Because of this bibliographical gap, the representativeness of any sample drawn from the total collection is threatened by the lack of data for a large subset of the population.

A second point is whether the extant languages are indeed the population that we want to study. Although 7,000 languages is a lot, it is only a fraction of the languages that have ever been spoken. Most languages have either become extinct or become another language due to internal diachronic processes and language contact. A rough estimate of their number yields 240,000 extinct and still existing languages as the overall population studied by linguistic typology. Thus, the database of linguistics is fundamentally restricted to a sample of under 3%. Since this sample is diachronically and culturally biased by the fact that over 90% of the languages are spoken in today's world, this immediately poses the question of how representative the existing languages are of human language in general.

An answer to the question of representativeness can only be very provisional. For some linguistic areas — where writing systems have been available for a considerable amount of time — we can go back a maximum of four or five language generations in order to see whether 5,000 years ago languages were the way they are now. There is no reason to conclude that for these cases anything fundamental has changed. In the face of this, historical linguists often adopt the principle of *uniformitarianism*, which extends this conclusion to the whole era of human language. Under this assumption, our 7,500 sample of known languages may serve as a reasonable representative of the 240,000 languages that were ever spoken, and as a basis for inferences about human language in general.

However, for most purposes, we consider the 7,500 existing languages as overall population for typol. studies, without the illusion that we will ever have data on the other 97%. In a more practical sense, only those languages qualify for which we have a description of the relevant research variable. The maximum we get out of a sample is an idea about what is possible in the languages of the world, though not a very reliable idea about what is impossible. Even more care should be taken when extending conclusions to language in the more abstract sense.

## 11. Types of samples

Two different classes of typological questions may be distinguished, each requiring its specific type of sample. The first class of questions concerns the probability that a language is of a specific type. E.g. what is the chance for a language to have postpositions, prepositions, or both. To find out about the real preferences, we want only independent cases in our sample. Thus, the fact that both Spanish and French have prepositions rather than postpositions is due to the fact that both inherited the majority of their adpositions from Latin and therefore represent only one instance of prepositionality. In general, in a sample, one wants only one language from a group of languages that shares both the relevant feature and a common ancestor, provided that this ancestor also had the feature under consideration and did not itself inherit it from one of its ancestors. So, we have to control our sample for genetic relationships at the right historical level.

Another cause of sharing may be areal, as in a Sprachbund, where languages as a result of contact and bilingualism all acquired some feature, possibly absent in the original languages. Obviously, the amount of restriction one puts on a sample depends on the relative stability of the linguistic variable in question. For instance, alignment of case marking is a rather conservative parameter: it is highly resistant to internal change as well as change under contact. Constituent order, in comparison, is less stable, as witnessed by the variation in basic and alternative orders among the Indo-European languages. In general terms, a sample for this type of research – *a probability sample* in Rijkhoff et al. (1993) – will be relatively small in size, typically between 50 and 200 languages, and will vary, depending on what is known beforehand about the range of values for the relevant linguistic variables and their stability. This is the preferred type of sample if one wants to apply conclusions drawn from the sample directly to the population in terms of the distribution of the phenomena observed.

A fundamentally different situation arises when linguistic variables are explored about which not much is known in advance. ...