MICHAEL CYSOUW (Leipzig)

Using the *World Atlas of Language Structures*

The *World Atlas of Language Structures* is a recently published resource providing data on the geographical distribution of grammatical structures for a large sample of the world's languages. The articles in this special issue are the result of some first attempts to use the wealth of data available in the atlas. This introduction discusses some general aspects of the data available in the Atlas and summarises the articles.

1. Introduction

After about five years of preparations, the *World Atlas of Language Structures* (henceforth WALS, HASPELMATH *et al.* 2005) appeared in the summer of 2005. The atlas contains a wealth of information on the world's languages, including about data from 2,600 languages and about 140 different structural characteristics. The characteristics covered in WALS range throughout various subdisciplines of linguistics, including aspects of phonology, morphology, syntax, lexicography and linguistic categories. There is even some information included on sign languages and writing systems. Looking at it from a slightly different perspective, WALS also presents a survey of the kind of interests that have been pursued in the field of linguistic typology over the last decades. Both perspectives offer many new angles into the investigation of the world's linguistic diversity. This special issue of STUF offers some first attempts to obtain new insights from this tremendous resource.

The articles collected in this issue originated with a pre-WALS-launch workshop held in December 2004 at the *Max Planck Institute for Evolutionary Anthropology* in Leipzig, although only a few of the papers in this issue (viz. the ones from BAKKER and from MASLOVA) are a direct spin-off of the presentations given at that workshop. Another workshop related to the launching of WALS was held in July 2005 at the sixth meeting of the *Association for Linguistic Typology*, held in Padang on Sumatra, Indonesia. At this occasion, the paper by DAHL was first presented. The other papers in this issue are all written by people that were present at either (or both) of those occasions and went home inspired by the possibilities offered by WALS.

As more and more linguists are starting to use WALS for typological explorations it is important to remember to properly cite any usage of this resource. As the editors suggest in the introduction, please consider WALS to be an edited volume with individual chapters that each have their own authors. So, when referring to data from WALS, each individual chapter consulted should be cited explicitly, and not the atlas as a whole. Also, when the *Interactive Reference Guide* (BIBIKO 2005) is used extensively, please consider to acknowledge the effort that went into preparing this practical tool by citing it.

There is some variation among current linguists regarding the usage of a definite article in English when referring to WALS in running text. The editors of WALS expressed their preference (in personal communication) for not using a definite article before the acronym WALS (a preference followed through the articles in this issue). Of course, when the acronym is used in an adjectival sense, the complete noun phrase should have an article (e.g. "the WALS data").

2. Available data

The amount of data available in WALS is tremendous. In total, there are almost 60,000 data points included. Each data point consists of one particular grammatical information on one particular language—and this does not yet include the wealth of metadata available, like geographical location and literature references. However, a quick calculation based on the previously mentioned 2,600 languages and 140 features gives an expected number of $2{,}600 \cdot 140 = 364{,}000$ data points in a completely filled data table. This means that the actual data table in WALS is only filled by about $60/364 = 16.5\%$. The reason is that most languages are only mentioned incidentally (cf. Figure 1).

Such a relatively empty data table leads to various problems for quantitative analysis. Fortunately, the available data is not distributed completely random through out the data table. The editors of the WALS provided a base list of 100 languages and an extended list of 200 languages for which the contributors were asked to minimally include information. At least for these languages the data should be complete. Actually, when restricting the data to the top 200 of most coded languages, the data table turns out to be filled (only) for about 74%, which is of course much better than 16%, though still far from complete.

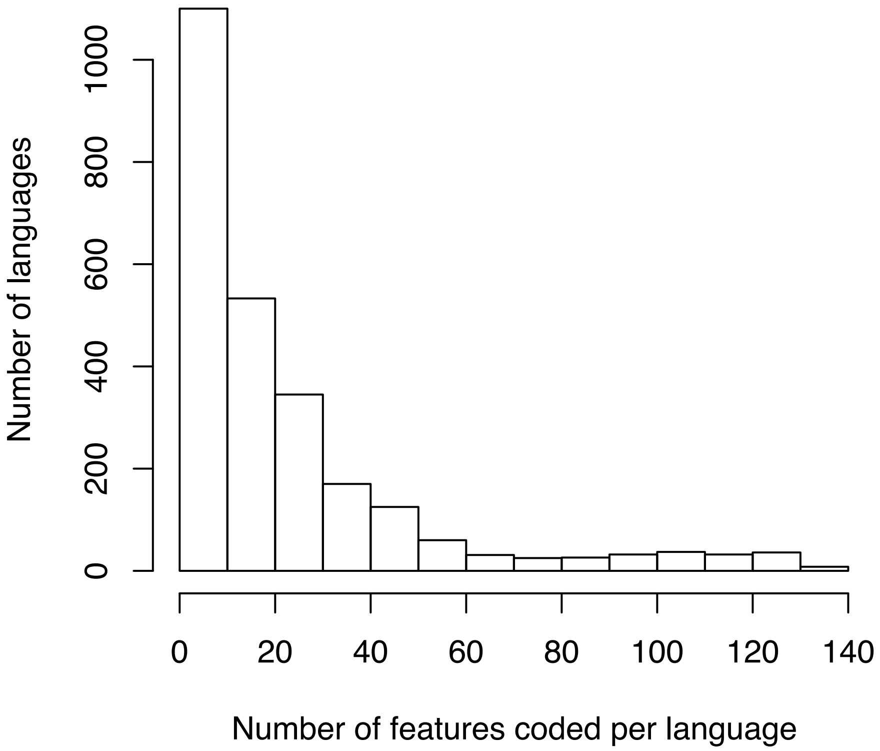

Figure 1. Histogram of the number of features available per language in WALS. Most languages are only mentioned incidentally.

The presence of missing data is to a large extent inevitable in a typological survey. The available sources for a particular language are never complete—the more so when dealing largely with strongly underdescribed languages—and it is not always possible (or feasible) to collect information directly through specialists or native speakers. The problem of missing data becomes more urgent when different features are combined into a single analysis. To be able to combine features, information on the same languages is needed for all features under investigation (though some corrections are possible to deal with missing data). However, the size of the sample drops down dramatically when more than two features are combined, as shown in Figure 2. This figure should be read as follows. When taking only one feature, there will always (100%) be a sample of more than 100 languages available (though not always the same sample). In other words, all features in WALS have information on more than 100 languages (disregarding the maps on sign language and on writing systems). However, when cross-secting the samples from two maps, then not all combinations will have information on 100 or more languages. There are almost 20,000 two-feature combinations possible, but only 76.5% of those combinations result in a sample with more than 100 languages. For three feature combinations this percentage drops below 50% and already by combining only eight features, the chances that any such combination would yield a language sample with more than 100 languages falls below 1%.

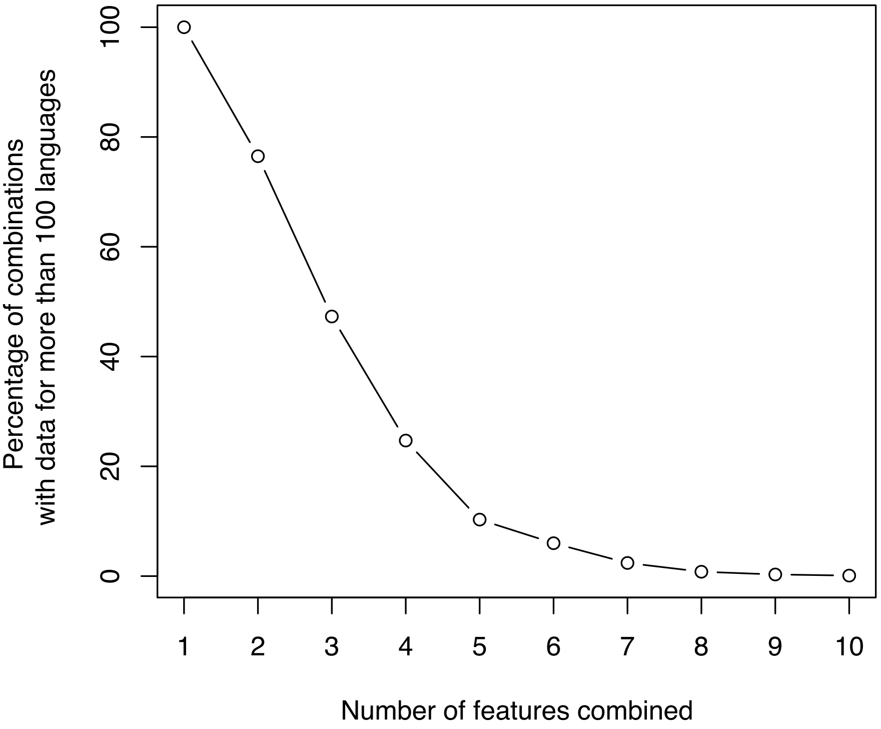

Figure 2. Rapidly falling language availability when multiple features are combined.

3. Survey of this issue

In the first paper, DIK BAKKER exemplifies his custom-made software package LINFER by applying it to the WALS data. With LINFER, BAKKER searches for (implicational) universals in the WALS data. Under suitable conditions, about one in 5,000 correlations between two WALS values turns out to be statistically significant. However, it is difficult to assess the meaning of such a number. Some significant correlations are bound to turn up even in random data. By using a randomization procedure, BAKKER shows that the number of significant correlations in the WALS data are about four times as frequent as in comparable random data. So there is indeed something to be gained from typological comparisons.

ELENA MASLOVA pursues another *meta*-typological question: if a typology distinguished a certain number of language types, what kind of expectation should we have about the number of languages in each type? *A priori*, one might assume that each type should be equally frequent. The stronger a particular typology would deviate from this assumption, the more interesting the actual state of affairs among the world's language would be. However, by looking at the distribution of languages over types in the dozens of typologies in WALS, MASLOVA argues that linguistic typologies do not seem to be evenly distributed, but are more alike to a *pareto* (or *power law*) distribution. In such distributions, there are a few large types and many smaller types with a continuous cline between large and small types. This finding indicates that typologists might have to reconsider the statistical basis of the interpretation of typological frequencies.

ÖSTEN DAHL uses the WALS data to evaluate—*a posteriori*—a longstanding assumption about language sampling. For the investigation of the world-wide diversity of a certain typological parameter, typologists assume that all linguistic families and all geographical areas have to be sampled on an equal basis, because we have no reason to assume *a priori* that one family or area is more interesting than another for the parameter under investigation. However, DAHL shows that, when summarizing over all maps in WALS, there are actually some macro-areas that show much less overall diversity (specifically this holds for South-East Asia and New Guinea), and some that show much more diversity (specifically, DAHL points to Australia and the Americas). Interpreting this as a lesson learned from past experiences, future samples might maybe better have a bias towards including languages from those macro-areas expected to have more diversity.

BALTHASAR BICKEL has another take on *a posteriori* sampling. Instead of preestablishing a sample for a typological study, he suggests that (when possible) larger samples should be collected which can then *post hoc* be investigated for any genealogical or areal bias. He specifically proposes an algorithm to minimize effects of known genealogical and areal biases in the sample. By investigating some WALS data, he shows that it actually makes a difference how the sample is delimited. This paper also makes clear that most samples in WALS are not ideally stratified samples of the world's linguistic diversity—a fact that has to be taken into account when using the frequencies of occurrence of linguistic phenomena as documented in WALS.

The final three papers in this issue all deal with the same question: is it possible to establish a notion of stability for typological parameters? both the paper by MIKAEL PARKVALL and the paper by SØREN WICHMANN and DAVID KAMHOLZ start from the assumption that *diachronic* stability of typological parameters can be measured by investigating the internal consistency of known genealogical units. So, if—for a particular feature—genealogical units are generally homogeneous (meaning that most languages in the group are of the same type), then this feature is considered to be relatively stable. To measure homogeneity, PARKVALL uses the *Herfindahl-Hirschman index*, better known to linguists as the *Greenberg index* (see the Appendix to the paper by CYSOUW, ALBU & DRESS for a note on the mathematical background of this index). WICHMANN and KAMHOLZ base their estimates of homogeneity on the assumption that feature-values are binomially distributed (which actually assumes that every feature-value is equally likely, which is criticised in the paper of MASLOVA, this issue).

In contrast to these two papers, the notion of stability as proposed in the paper by MICHAEL CYSOUW, MIHAI ALBU and ANDREAS DRESS is based on a different principle. They assume that all the data from WALS combined give an indication of overall linguistic similarity between languages. Individual features from WALS are then correlated with the overall similarity to find those individual features that best predict this overall similarity. Because the phylogenetic information in the WALS data is probably rather limited, this approach actually identifies features that are important to the typological profile of languages (as far as the available characteristics in WALS are concerned) and not features that are diachronically important for the structure of the WALS languages.

References

.

BIBIKO, HANS-JÖRG 2005. Interactive reference tool to the *World Atlas of Language Structure*. MPI-EVA, Leipzig.

Correspondence address

Max Plank Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
cysouw@eva.mpg.de