STUF, Berlin **61** (2008) 3, 208–220

Östen Dahl (Stockholm)

An exercise in *a posteriori* language sampling

Abstract

A central methodological issue in language typology is sampling – how to choose a representative set of languages for a typological investigation. Most proposed typological sampling methods are *a priori* in the sense that they are based on assumed, rather than observed, effects of biasing factors – such as genealogical and areal proximity. The advent of the *World Atlas of Language Structures* (WALS) creates for the first time a chance to attempt *a posteriori* sampling. The basic idea is to create a sample by removing from the set of available languages one member of each pair of languages whose typological distance – as defined in terms of the features in WALS – does not reach a predefined threshold. In this way, a sample of 101 languages was chosen from an initial set of the 222 languages that are best represented in WALS. The number of languages from different macroareas in this sample can be taken as an indication of the internal diversity of the area in question. Two issues are discussed in some detail: (i) the high diversity of the indigenous languages of the Americas and the tendency for these to be underrepresented by previous sampling methods; (ii) the extreme areal convergence of Mainland South East Asian languages. It is concluded that areal factors cannot be neglected in typological sampling, and that it must be questioned whether the creation of elaborate sampling algorithms makes sense.

1. Introduction

As language typology has matured as a field, there has been a growing awareness of methodology. A central methodological issue is sampling – how to choose a representative set of languages for a typological investigation. In simpler terms, how do you find a hundred languages that reflect the diversity of human languages in an optimal way? For a number of reasons, the problems connected with the sampling of languages are rather different from the typical sampling problems encountered, for example, in the behavioural sciences. This is probably one of the reasons why typologists are still quite insecure about their sampling methods. Another reason is that whereas typologists are in general keenly aware of the necessity to avoid various kinds of biases in their data, the patchiness of the available data is such that it is not only practically very difficult to obtain unbiased samples, it is also very hard to evaluate samples as to their representativeness. If you have got hold of one ear and one foot of the proverbial elephant, how can you tell if what you got is at all representative of the rest? In this situation, it is natural that typological sampling methods are almost totally *a priori*. By this I mean that they are based on assumed, rather than observed, effects of biasing factors. Thus, essentially everyone agrees that it is genealogical and areal biases that primarily have to be avoided, but there are hardly any attempts to verify how similarities and differences between languages actually relate to genealogical and areal connections, let alone other possibly relevant factors such as community size, population density, or literacy. However, as Rijkhoff & Bakker (1998) note, whether

sampling method *x* is better than method *y* "is ultimately an empirical question" and if the aim is to maximize linguistic variety within a sample, "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*".[1]

The advent of the *World Atlas of Language Structures* (WALS, Haspelmath et al. 2005) gives us for the first time a chance to start doing something of this kind – to attempt to do what could be called *a posteriori* sampling. The database behind the 142 maps in the printed version of WALS provides information about a large number of features pertaining to the major subdivisions of linguistic description, for a large set of languages, and we should thus be able to draw some empirically well-founded conclusions about linguistic diversity and its dependence on various factors. In this paper, I shall try to formulate a measure of the typological distance between two languages computed on the basis of the WALS data. This measure will then serve as a basis for the attempt at *a posteriori* sampling. The idea is simple: if we have a way of determining the typological distance between two languages, we could define a minimal distance that any two languages in a sample should have to each other, thus ensuring that the sample meets the minimal requirements of diversity.

It should be added immediately, however, that although the wealth of data in WALS is most impressive, for the purposes of *a posteriori* sampling it is just barely sufficient for a pilot investigation. The total number of languages included in the WALS is 2,561. This would mean that about one third of all languages in the world are represented in WALS.[2] However, only 115 languages appear in at least 100 maps.[3] Since a global measure of typological distance should be based on a comparison of a large number of features, and preferably the same features for every language, the available set of languages that can be assigned such a measure is not really very large. (Below, I shall discuss how one can get some extra mileage out of the WALS data.) However, one has to begin somewhere, and I think it is still possible, within the limits of the WALS database, to draw some fundamental conclusions about typological diversity and its dependence on factors of different kinds.

2. Some problems with *a priori* sampling methods

Since language sampling was discussed seriously for the first time by Bell (1978), various scholars have constructed language samples with claims of representativeness, in most cases in connection with their own typological investigations. The problem of language sampling is treated from a more general point of view in two papers by a group of Dutch scholars, Rijkhoff et al. (1993) and Rijkhoff & Bakker (1998). In these papers, six previously proposed samples – Bell (1978), Perkins (1980), Bybee et al. (1994), Dryer

[1] Actually, this requirement is probably a bit too strong, in that we cannot assume that variation is consistently distributed over languages, which is also acknowledged in other places in the same paper.

[2] We may here disregard the fact that some of the WALS languages would rather be varieties of one language in the *Ethnologue* database.

[3] The authors of the WALS were given a set of 100 languages which should be included, if possible.

(1992), Nichols (1992), and Stassen (1997) – are discussed and evaluated against the authors' own proposal to select languages, which I shall refer to as the Amsterdam method. In all language sampling methods proposed in the literature, genealogical relationships play a major role. Sampling methods differ, it appears, mainly in two respects: whether other information (primarily about geographical areas) is also taken into account, and whether assumptions about the time-depth of genealogical relationships are given any weight. The "purest" proposal is the Amsterdam method, which disregards anything but "the graph theoretic structure" of the genealogical trees associated with the family affiliation, the main idea being that this structure "adequately reflects the linguistic diversity of the family it represents." Rijkhoff & Bakker (1998) say that the method "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." However, no concrete example of how this could be done is provided.

Any sampling method that relies on genealogical relationships has to make assumptions about those relationships. This is not a trivial problem. Linguists do not agree on the genealogical classification of the world's languages, and for some parts of the world the proposals differ to an extreme degree, most notably in the Americas. At one end, Greenberg's (1987) classification postulates only three families among the indigenous American languages: Amerind, Na-Dene, and Eskimo-Aleut. At the other extreme, the *Ethnologue* (Gordon 2005) lists no less than 57 independent language families and in addition a considerable number of isolates in the Americas. In between these extremes, Voegelin & Voegelin (1977) postulate nineteen phyla in the Americas. It is unavoidable that a choice between these classifications will influence a language sample. Rijkhoff et al. (1993) compare the outcomes of applying the sampling method proposed in Perkins (1980) to the classifications proposed by Voegelin & Voegelin (1977), on the one hand, and by Ruhlen (1987), on the other. The languages subsumed under "Amerind" will have seventeen representatives in a 50-language sample following Voegelin & Voegelin but only one single language in a similar sample following Ruhlen's classification. Rijkhoff et al. (1993) contrast this to the Amsterdam method, which yields the same number of languages (seventeen, which is 34 % of the sample) for the Voegelin & Voegelin classification but as much as seven languages (which is 14 % of the sample) if it is applied to Ruhlen's classification. They comment: "Both methods are dependent on the particular classification used, but our method mitigates the effects of new insights in the area." However, a difference between 34 and 14 per cent of a sample is still quite serious. In fact, with the *Ethnologue* classification, which is even more "splitting" than that of Voegelin & Voegelin, the discrepancies are even greater. In Rijkhoff & Bakker (1998), the results of applying the Amsterdam method to the classifications of Ruhlen and the *Ethnologue* are compared. In a 100-language sample, the Ruhlen classification yields 20 Amerind languages and the *Ethnologue* classification results in 52 languages, that is, the difference amounts to a third of the sample. In favour of the Amsterdam method, it has to be added that the influence of the classification used decreases as the sample size increases. The 20 and 52 languages in the 100-language sample correspond to 105 for Ruhlen and 122 for the *Ethnologue* in a 500-language sample. But this also means that the Amsterdam method is quite sensitive to sample size: depending on whether a sample of 100 or 500 languages is constructed, the same classification may assign between 52/100 = 52 % and 122/500 = 24 % of the sample to the same grouping.

Some sampling methods, for example those proposed by Bell (1978) and Dryer (1992), put an upper time limit (e. g. 3,500 years) on the genealogical relationships employed in the classification. Since genealogical units are much less controversial within this time range, the results tend to be in between the extreme values quoted above. Thus, Bell's method would yield nine Amerind languages in a 30-language sample, which is 30 per cent. According to the calculations in Rijkhoff & Bakker (1998), Dryer's 1992 sample contains 220 Amerind sampling points (genera rather than languages in his case) out of a total of 736. This in fact yields exactly the same percentage – 30 %. We may further note the following numbers for the representation of Amerind in some other proposed samples (all according to the figures given in Rijkhoff & Bakker 1998): Bybee et al. (1994): 16 out of 76 (= 21 %), Stassen (1997): 91 out of 408 (= 22 %), and Nichols (1992): 66 out of 174 (= 38 %). What we see here is thus quite significant variations between different typological samples, depending to a large extent on assumptions about the genealogical classification of languages and about the effects of genealogical relationships. A look at the actual similarities and dissimilarities between the languages involved seems highly motivated.

3. Defining and computing the distance measure

As already stated, the basis for my exercise is the WALS database. However, not all the maps were suitable for inclusion. All maps, except two, show spoken languages only. In contrast, the maps 139 and 140 (Zeshan 2005 a, b) show variation in sign languages. These two sign-language maps have a totally disjoint sample from the rest of the atlas, and it would not make sense to compare them with the others. An investigation of the typological distance between sign languages is certainly a desideratum, which, however, cannot be realized with our present knowledge. A somewhat more controversial question is what kind of features should be considered for spoken languages. Intuitively, properties of languages that involve "Saussurean arbitrariness" with respect to the relationship between meaning and form do not seem to qualify as typological. This principle would exclude, for example, features such as "the word for 'horse' begins with *h*". But it is not entirely obvious where to draw the line. Maps 136 and 137 (Nichols & Peterson 2005) show the distribution of consonantal patterns in personal pronoun systems (viz. "M-T pronouns" and "N-M pronouns"). I have excluded these, although they might be seen as borderline cases. My own map on the word for 'tea' (Dahl 2005 = WALS 138) was a clearer case for exclusion, like the map on writing systems (Comrie 2005 = WALS 141), which does not show individual languages but rather areas. After this excision, the database contained 138 maps.

The intuitive idea behind the typological distance measure is simple: How large a proportion of the features that are defined for both members of a language pair have different values? This measure could be applied directly to the WALS database as it stands, but there are a number of reasons why one would like to revise it. It is only when a feature is binary that it is wholly unproblematic to simply ask if the value is the same or different. However, some WALS features are graded – one example is map 49 on the number of cases (Iggesen 2005). Suppose that we have a language L with two cases. Then, a binary choice between "same" and "different" equates the distance between, on the one hand, L and a language with three cases, and, on the other, L and a language with ten cases, although a language with three cases could very well be considered more similar to L than

a language with ten cases. Other features are really composite, for example map 19 on the presence of uncommon consonants (Maddieson 2005). In this map, one value is labelled "[The language has] pharyngeals", another value is "th-sounds", and a third value is "pharyngeals and th". Pharyngeals and th-sounds do not seem to have anything intrinsic in common except that it is practical to show their distribution in one map. If we want to compute typological distances, however, we should rather treat them as entirely separate features.

In the revised version of the WALS database that I used, the values of graded features were redefined as fractions of 1, and other features were, if possible, redefined as sets of binary features. The distance between two languages L1 and L2 then is the average difference between the values for L1 and L2 for all the features defined for both of them. In some cases, I added a new feature to the old one. This occurred when some of the values of a feature seemed to form a natural class. For example, map 32 on systems of gender assignment (Corbett 2005) had three values: "no gender", "semantic", and "semantic and formal". It stands to reason that languages with gender should be seen as more similar to each other than languages without gender, motivating a binary feature in which the two last-mentioned values of the original one were merged. This could not be done without introducing some redundancy in the system, but I made the decision that it was more important to include as many features as possible, at the expense of some redundancy. The final result of the revisions was a database with 219 features on which all distance measures were calculated. In principle, the measure should be a fraction of 1. I found it more convenient, however, to use percentages, which means that two maximally different languages would have a distance of 100.

As was noted above, computing typological distances makes sense only if there is a sufficient number of features to compare. It is not immediately clear how to determine what would be a sufficient number for the current purpose. One possible way of checking that a distance measure makes sense is to see how it relates to areal and genealogical relationships. If comparisons are made with too few features then languages that have no geographical or genealogical connection come out as "typological twins", that is, they obtain very low values of the distance measure. The higher the number of features compared, the smaller the proportion of such accidental similarities becomes. After some experimenting, I decided on 101 as a suitable minimal number of features that should be available for each language – this decision yielded 222 languages that could be used as a base for the further calculations. Some caution is still warranted in view of the fact that there are still some unexpected "typological twins", the most extreme case being Tigak in New Guinea (101 features) and Wolof in West Africa (100 features) with a distance of 13.6. If one wants to draw conclusions about the real typological distance between individual languages, it may be wise to choose pairs where the members are coded for a higher number of features.

4. The base sample

Let us take a closer look at the base sample of 222 languages with a minimal number of 101 coded features, as introduced in the preceding section. The typological distance between the languages in this set of languages varies between 9.8 (between Dutch and German) and 74.6 (between Ju|'hoan and Central Yup'ik). The arithmetic mean of all

distances is 41.73 (and the median is 42). This mean is probably higher than one would get by looking at arbitrary language pairs, since the base 222-language sample is not a randomly selected subset of the world's languages, but represents the editors' and authors' strivings to display the diversity of human language as well as possible.[4]

Although genealogically related languages (not unexpectedly) have smaller distances to each other than genealogically unrelated languages, it is possible for two languages to belong to the same genealogical grouping and still be quite different from each other typologically. In about 3 per cent of the language pairs in the 222-sample, the members belong to the same family. Of these, roughly seven per cent have a distance that exceeds the average value of 41.73. The most extreme case is the distance between the Australian languages Maung and Yidiny, which is 53.3. Not far behind are Bagirmi and Dongolese Nubian (Nilo-Saharan), Murle and Koyraboro Senni (also Nilo-Saharan), Irish and Kashmiri (Indo-European) and Zulu and Yoruba (Niger-Congo), all with distances above 50. Not even belonging to the same genus is a guarantee of typological similarity. For example, the language pairs Drehu and Kilivila (both Oceanic languages within Austronesian) and Amharic and Modern Hebrew (both Semitic languages within Afro-Asiatic) have distances of about 40, meaning that they are only slightly below the average from the overall sample.

In most cases, genealogically related languages with large typological distances are also geographically distant from each other. Apparently, related languages that are no longer in contact with each other can in a few thousand years develop typological profiles that are no longer indicative of a common origin. For example, the distance between English and Persian (43) is close to the average distance in the overall sample. Just looking at the features listed in WALS for these two languages there is thus no way of telling that they are both Indo-European.

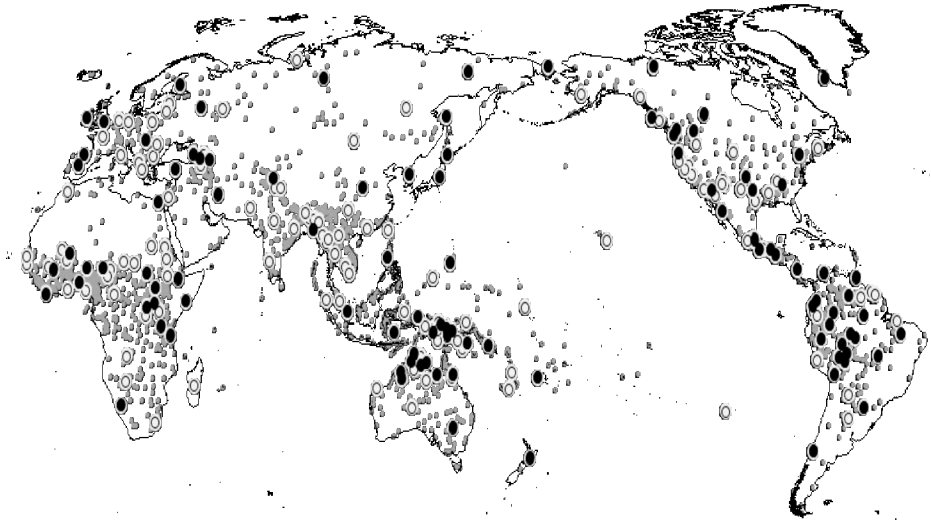
5. Creating smaller samples

I proceeded as follows to obtain smaller samples out of the base sample. Having chosen a threshold value for the typological distance measure, I excluded one member of any language pair whose distance was lower than the threshold. The member excluded was the one with the smallest number of features available in the database. The advantage is that the languages that remained at the end would have a maximal number of features at the possible cost that languages prioritized in WALS would be unduly favoured. We shall soon see to what extent this is true. The essential effect of performing this cleansing operation on a sample is to eliminate overrepresentation of any kind, that is, it weeds out "typological twins" whether the resemblance depends on factors such as genealogical or areal association or whether it is purely accidental. What the method cannot do, however, is to make amends for under-representation. Thus, if the base sample contains too many Australian languages, their number will be reduced, but if there are no or only a handful Australian languages to start with, the method does not help. This is important to have in mind, since the languages in some parts of the world are still not well described and may not have made their way into the typological literature to the extent they would deserve.

[4] Certain other considerations also played a role in the choice of languages. For example, the language sample recommended by the editors included a larger number of major languages of Eurasia than would be motivated by representativeness concerns, cf. the introduction to Haspelmath et al. 2005).

This method does not take into account where the languages compared come from. In other words, if a language from South America by accident happens to be similar enough to a language from Siberia, one of them will be removed. This means that the number of languages retained in the final sample from a certain area or a certain family does not necessarily reflect the internal diversity of that area or family. But it is possible to apply the method also to subsets of the base sample, as long as they are reasonably large. That is, we can use the languages of one macro-area as the base for the selection procedure, which means that only similarities within the area will be considered.

My aim was to define a sample of 100 languages. As it turned out, a threshold for the distance measure of 25.4 yielded a sample of 101 languages, but raising it ever so slightly resulted in 99 languages, so I had to be content with the 101-language sample (see Map 1). The fear that the procedure described above would just return the 100-language core WALS sample (cf. the introduction to Haspelmath et al. 2005) turned out to be unfounded. My 101-language sample and the WALS core 100-language sample share only as few as 60 languages. The distribution of these samples over macroareas is shown in Table 1, which also shows what happens if one applies the method to the macroareas separately (with the same threshold). This kind of reduction increases the number of languages in a few cases, though without changing the general picture. We can also see that on the whole, the distribution is quite similar to that found in the WALS core 100-sample, with two clear exceptions: Asia has somewhat fewer languages in my sample than in the core WALS sample, and South America has quite a few more. Both these cases deserve some discussion.



Map 1: Filled circles: 101-language sample; unfilled circles: other members of base 222-language sample, grey dots: other languages in WALS database

Macroarea	WALS core 100-sample	Base 222-sample	Reduced 101-sample	Reduction by macroarea
Africa	16	38 (17%)	16	17
Europe	7	23 (10%)	8	8
Asia	24	46 (21%)	17	19
Oceania	4	8 (4%)	3	3
New Guinea	10	21 (9%)	9	11
Australia	7	15 (7%)	9	9
North America	19	42 (19%)	19	21
South America	13	29 (13%)	20	21

Table 1: Distribution of the various samples over macroareas

6. The Americas

As discussed previously in Section 2, the Americas are the part of the world where opinions about the genealogical classification of languages vary most dramatically, and consequently also the representation of these continents in typological samples shows large differences. The number of languages in my 101-sample that are subsumable under Greenberg's proposed "Amerind" family is 38. Although this is not as high as what one gets applying the Amsterdam method to a 100-sample based on the *Ethnologue* classification, it is higher than in all other proposed samples quoted in Section 2, with the exception of Nichols' sample.

To claim on the basis of my method alone that the indigenous languages of America have been underrepresented in most typological samples is indeed to stick out one's neck. It would be good to have some way of corroborating my results. One simple way of estimating the internal diversity of macroareas is to see to what extent the possible values of the WALS features are represented in the languages of that area, which can be taken as a rough measure of how much of the total diversity of the world's languages is found in the area in question. The idea is that an area which has, say, languages with VSO, SVO, and SOV languages is *ceteris paribus* more diverse than an area with only VSO and SVO languages. The measure is rough because it does not take account of combinations of features, just the features themselves. In my revised version of the WALS database, the 219 features have together 753 possible values. Table 2 shows how many of these features are represented in each macroarea. What we can see here is that the macroareas that have between 15 and 20 representatives in the 101-sample are also the ones that have the highest number of different features in WALS. This supports the claim that the four areas Africa, Asia, North America, and South America, are indeed comparable as to their internal diversity. There is in fact a highly significant correlation between the number of languages per area in the 101-sample and the number of different features per area (Pearson's $r = 0.92$, $p < .005$). However, there are differences in the details. For example, Asia and New Guinea have slightly more different features than would be predicted from its number of languages in the 101-sample.

Macroarea	No. of languages in the 101-sample	No. of different feature values
Africa	16	640
Europe	8	545
Asia	17	675
Oceania	3	466
New Guinea	9	613
Australia	9	540
North America	19	660
South America	20	646

Table 2: Distribution of languages in the 101-sample and WALS feature values over macroareas

Further, I have investigated two other possible measures of diversity in the 222-sample. The second column in Table 3 shows the distribution of languages whose average distance to the other languages within the sample exceeds 43 (i.e. the general average distance in the sample). These are the languages with an unusual typological profile. The third column in Table 3 shows the number of languages per macroarea that have a minimum distance of 26 to the nearest language in the rest of the sample. This picks out the "typological isolates", that is languages whose profile is extremely uncommon. These two measures are clearly related to the internal diversity of areas, but have the advantage of being more independent of how the WALS languages were sampled. The following observations, due to Michael Cysouw (p.c.), are relevant here. There are relatively good correlations between the number of languages in the 222-sample and the number of oddities. Summing them up (i.e. taking the typological isolates and the typological unusuals together), there is a significant correlation between the number of languages and the number of oddities (Pearson's $r = 0.85, p < .01$), but the most interesting effect is that the Americas have clearly more oddities than expected from the number of languages in the 222-sample (residuals after regression of + 27%). However, the correlation with the 101-sample is almost perfect ($r = 0.97, p < .00001$), indicating that this sample gives a better reflection of the typological diversity.

In all of this, what we see is that the American continents come out as essentially on a par with Africa and Asia together, if not above them. There is thus nothing in all these measures that contradicts the conclusion that the Americas together should be represented in a sample with at least the same number of languages as Asia and Africa together, which makes something like 40 per cent a realistic level. My claim, then, is that any method that gives the Americas less than 35 per cent of the languages in a sample underrepresents them.

Cysouw (forthcoming) has looked at the WALS data from a slightly different angle. He investigated the distribution of unusual characteristics among the WALS languages. A comparison of his figures with mine is not without interest. To start with, on Cysouw's list of the 15 languages with the highest "mean rarity level", eight are from the Americas. Looking at families with at least three languages included in the WALS data, out of the top ten having the highest "weighted rarity", six are from North America and one from South

Macroarea	No. of languages in the 222-sample	No. of languages with an average distance > 43	No. of languages with a minimal distance > 26
Africa	38	9	8
Europe	23	9	0
Asia	46	10	7
Oceania	8	1	0
New Guinea	21	2	3
Australia	15	5	1
North America	42	17	7
South America	29	13	9

Table 3: Distribution over macroareas of languages in the 222-sample and languages in that sample with a high average or minimal distance

America. Finally, of the 15 "areas of high rarity", four are in North America and one in South America. Although rarity and diversity are not necessarily the same thing, the high representation of the Americas on Cysouw's lists shows that the languages spoken in these continents are not only internally diverse but also tend to have a high number of unusual features.

According to the statistics on the *Ethnologue* website, there are 1,002 living languages in the Americas, that is 14.5 per cent of the total of 6,912 languages in the world. Africa and Asia have together 4,361 languages or 63.9 per cent of the total. If we assume that the contribution of the two regions to typological diversity is about the same, then there is an enormous discrepancy between internal diversity and number of languages. The discrepancy is even larger if we consider the total number of speakers of the languages in question. Again according to the *Ethnologue*, the languages of Asia and Africa have almost a hundred times as many speakers as the indigenous languages of America. In contrast, when we look at genealogical diversity, the typological diversity found in the Americas fits the number of families in Africa and Asia (in the view of a moderate "splitter") together quite well. Likewise, in terms of physical area, the relationship is much more even: 42 million square kilometres for the Americas vs. 75 million for Africa and Asia combined. Relating the figures for numbers of speakers to the areas, we can see that a major difference between the Americas and the two Old World continents is that there are many more speakers of indigenous languages per areal unit in the latter. Indeed, the indigenous languages of the Americas are largely spoken in areas that are (or recently were) extremely sparsely populated, to a great extent by "peoples who traditionally have not embraced or have only partially embraced an agricultural way of life" (Brown 2005: 527). That is, these languages are or were spoken in small communities in sparsely populated areas.

A similar pattern is also found when comparing Australia to New Guinea. The 230 Australian languages have nine representatives in the 101-sample. This stands in contrast to New Guinea, which has the same representation in the 101-sample (i.e. nine languages), but where it is taken from a much larger population of languages – 1200, that is, more

than in the two American continents together. But then these languages are all spoken in a relatively small and densely populated area (less than one million square kilometres). It stands to reason that areal pressure will tend to be higher in such an environment than, for example, in Amazonia, where a population of comparable size is spread out over an area ten times as large. I hope to be able to develop these thoughts in more detail in future research.

7. Mainland South East Asia: the ultimate Sprachbund

It was noted above that Asia had fewer languages in my 101-sample than in the WALS core sample, viz. 17 rather than 24. Now, Asia is a very large continent, and treating it as a unit is actually misleading. If we look at Map 1, we can see that most of the Asian languages in the sample are at the periphery (almost a third of the languages are from the sparsely populated northeastern part), and that the densely populated areas in the southern half of the continent have very few representatives. Thus, from five countries in Mainland South East Asia – Myanmar, Thailand, Laos, Cambodia, Vietnam and mainland Malaysia – there is not a single language in the 101-sample, as compared to eight languages in the 222-sample and four languages in the WALS core 100-sample. The values for the typological distance measure reveal that the languages in this area are really very close to each other typologically, although they belong to several unrelated families. If there ever was a Sprachbund, this is it. In particular, this is true of a core area, where Thai seems to be the pivot (for a recent discussion of Mainland South East Asia as a language area, see Enfield 2005).

To appreciate how close the languages in this area are, we can compare them to the Germanic languages. There are three West-Germanic languages in the 222-sample, viz. Dutch, English, and German. The distances between them vary between 9.8 (German – Dutch) and 21.1 (German – English). We find five South East Asian languages, representing three different language families, whose internal distances do not exceed those found in the West Germanic group: Hmong Njua (Hmong Mien),[5] Khmer, Khmu', Vietnamese (all Austroasiatic), and Thai (Tai-Kadai)[6]. The distance values all fall between a minimum of 11.4 (Thai – Vietnamese) and a maximum of 22.5 (Hmong Njua – Khmu'). In addition, there is a representative of a fourth family, Eastern Kayah Li (Sino-Tibetan), whose distance to the other languages mentioned ranges from 20.6 to 25. The two remaining languages in Myanmar (Burmese and Lahu) have significantly higher distances to the others and clearly do not belong to the core area. On the other hand, there is also reason to assume that the Austronesian languages in the area (e.g. Cham and Chru) partake in this convergence (cf. Thurgood 2000 and Enfield 2005), although there is insufficient information about them in WALS to further develop this thought here.

We thus have a total of five language families, the representatives of which have converged in a very strong way. The typological properties shared by the languages in the area – which include the classical examples of the "isolating" language type – are also rather uncommon in the rest of the world. If we combine the WALS map on the position of tense-

[5] Hmong Njua is actually spoken in southern China, outside the countries listed above.
[6] Each of these languages are coded for at least 139 features in the extended database.

aspect affixes (Dryer 2005a = WALS 69) and the map on the expression of pronominal subjects (Dryer 2005b = WALS 101), we obtain a sample of 552 languages. Sixteen of these languages, or 2.9 per cent, share the values "Optional pronouns in subject position" and "No tense-aspect affixes". The first one implies that there are no affixes, clitics or free pronouns used obligatorily to mark the subject of a clause. The second characteristic means that there is no tense-aspect morphology. These properties appear to be quite general in the core Mainland South East Asian area, and of the sixteen languages in the sample, eight, or 50 per cent, come from this part of the world. It thus seems safe to conclude that this is the largest concentration of such languages in the world.

Irrespective of whether it is advisable to exclude Mainland South East Asia completely from a typological sample, the figures discussed here show the dangers of neglecting the results of areal pressure or of assuming that they are always weaker than the effects of common descent. A sampling method that looks only for genealogical relationships tells us to prefer a sample that includes the pair Khmer – Thai rather than the pair Khmer – Mundari, since Khmer and Mundari are both Austroasiatic languages whereas Thai is from the Tai-Kadai family. However, according to the measure presented in this paper the distance between Khmer and Thai is 12.3 (which is about the same as the distance between Polish and Russian, which is 12.8). In contrast, the distance between Khmer and Mundari is 48. In other words, if we want to maximize the diversity of the sample, including both Khmer and Mundari makes more sense than including both Khmer and Thai.

8. Final discussion

I think it is worth emphasizing once more that although WALS has made the life of areal typologists much easier, working with the data from WALS quickly makes one aware of its limitations. It is my hope that this exercise in *a posteriori* sampling performed here can be repeated with more languages, more features and more refined methods. Still, a number of conclusions can be drawn from the current results.

One general conclusion concerns the role of areal factors. The proponents of the Amsterdam method motivate their decision not to take areal factors into account with at least two considerations, if I understand them correctly. One is that "areal stratification is much more problematic […] and consequently rather difficult to incorporate in a (mechanical) sampling procedure" (Rijkhoff et al. 1993). The second consideration is that "the quality of a language sample is affected worst if languages are too closely related genetically" (Rijkhoff et al. 1993: 175). What my analysis of the WALS data suggests is that areal stratification is indeed problematic, since areal pressure is highly variable, but, on the other hand, it is sometimes strong enough to overshadow genealogical relationships, as we saw in South East Asia. In other words, sample constructors who ignore areal pressure do so at their peril.

Taking this reasoning one step further, I think it must be seriously questioned if it is at all possible to construct a general sampling algorithm for the world's languages that does not take into account observations of actual typological distances between languages. At least, elaborating the details of such algorithms risks creating a false impression of exactitude. Until we get a firmer empirical basis for sampling, I would propose to do "common sense sampling" along the following lines. First, try to see to it that the macroareas are represented in roughly the proportions found in my 101-sample (cf. Table 1). Second, when choosing

languages within each area, try to maximize genealogical and geographical diversity, but also see to it that both large and small genealogical groupings are represented. A sample that follows these principles cannot be totally wrong.

References

Bell, Alan (1978): Language samples, in: Greenberg, Joseph. H.; Ferguson, Charles A. & Moravcsik, Edith A. (eds.) *Universals of human language*. Stanford: Stanford University Press, 123–156.

Bybee, Joan L.; Perkins, Revere & Pagliuca, William (1994): *The evolution of grammar. Tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.

Brown, Cecil H. (2005): Finger and hand, in: Haspelmath et al., 526–529.

Comrie, Bernard (2005): Writing systems, in: Haspelmath et al., 568–571.

Corbett, Greville G. (2005): Systems of gender assignment, in: Haspelmath et al., 134–137.

Cysouw, Michael (forthcoming): Quantitative explorations of the world-wide distribution of rare characteristics, or: the exceptionality of north-western European languages, in: Simon, Horst & Wiese, Heike (eds.), *Expecting the unexpected – Exceptions in grammar.* Berlin: Mouton de Gruyter.

Dahl, Östen (2005): Tea, in: Haspelmath et al., 554–557.

Dryer, Matthew (1992): The Greenbergian word order correlations, in: *Language* 68, 81–138.

Dryer, Matthew (2005a): Position of tense-aspect affixes, in: Haspelmath et al., 282–285.

Dryer, Matthew (2005b): Expression of pronominal subjects, in: Haspelmath et al., 410–413.

Enfield, Nick J. (2005): Areal linguistics and Mainland Southeast Asia, in: *Annual Review of Anthropology* 34, 181–206.

Gordon, Raymond G. (ed.) (2005). *Ethnologue: languages of the world*. SIL International, Dallas, TX.

Greenberg, Joseph (1987): *Language in the Americas.* Stanford: Stanford University Press.

Haspelmath, Martin; Dryer, Matthew; Gil, David & Comrie, Bernard (eds.) (2005): *The World Atlas of Language Structures.* Oxford: Oxford University Press.

Iggesen, Oliver (2005): Number of cases, in: Haspelmath et al., 202–205.

Maddieson, Ian (2005): Presence of uncommon consonants, in: Haspelmath et al., 82–85.

Nichols, Johanna (1992): *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nichols, Johanna & Peterson, David (2005): Personal pronouns, in: Haspelmath et al., 546–553.

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

Rijkhoff, Jan & Bakker, Dik (1998): Language sampling, in: *Language Typology* 2(3), 263–314.

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

Ruhlen, Merritt (1987): *A guide to the world's languages*. Stanford: Stanford University Press.

Stassen, Leon (1997): *Intransitive predication*. Oxford: Oxford University Press.

Thurgood, Graham (2000): Learnability and direction of convergence in Cham: the effects of long-term contact on linguistic structures, in: *Proceedings of the Western Conference On Linguistics 2000*. Fresno: California State University, 507–527.

Voegelin, Charles F. & Voegelin, Florence M. (1977): *Classification and index of the world's languages*. New York: Elsevier.

Zeshan, Ulrike (2005 a): Irregular negatives in sign languages, in: Haspelmath et al., 560–563.

Zeshan, Ulrike (2005 b): Question particles in sign languages, in: Haspelmath et al., 564–567.

Östen Dahl
Department of Linguistics
Stockholm University
106 91 Stockholm
SWEDEN
oesten@ling.su.se