Paper for the Symposium
Language Contact and the Dynamics of Language: Theory and Implications
10-13 May 2007
Leipzig

WHAT DOES WALS TELL US ABOUT THE DIFFUSION OF STRUCTURAL FEATURES?

Bernard Comrie
Max Planck Institute for Evolutionary Anthropology
and University of California Santa Barbara
comrie@eva.mpg.de

1. Introduction

In this paper, I will try to show that the *World Atlas of Language Structures* (Haspelmath et al. 2005), hereafter WALS, provides an important new tool in the investigation of the areal diffusion of structural features of language.

But first, let me say a few words about the nature of WALS. It is, first, a printed atlas, which shows the worldwide distribution of the feature values of around 142 structural features of languages. It consists of 142 maps, each accompanied by explanatory text, and each map is devoted to one structural feature (with some minor provisos, some of which figure in what follows). On each map, each language in the sample is identified by a small circle, and the color of that circle identifies the feature value that is evinced by that language with respect to the feature in question. Any of the maps accompanying this paper can be used as illustration. Second, WALS is an interactive research tool that enables the user to manipulate and combine maps, for instance to test whether particular feature values correlate with one another (e.g. whether languages with the order Object–Verb tend also to have the order Genitive–Noun – it turns out that they do); the interactive research tool comes on a CD-ROM accompanying the printed atlas. Finally, WALS is a database of languages and feature values, with several additional kinds of information (e.g. geographical location and genealogical affiliation of languages); while the database is not yet, for copyright reasons, in the public domain, it is hoped that from 2008 it will be so available.

A WALS map thus enables one to see at a glance the geographical distribution of a particular feature value, e.g. where on the world languages with the order Adjective–Noun are located (see the accompanying Map 1). This can then be used as input into studies of the diffusion of structural features on the basis of language contact. I would emphasize that WALS does not directly give any information on areal diffusion or language contact. All it provides is a static, roughly contemporary picture of the geographical distribution of feature variables. But this is important information if we want to carry out studies of areal diffusion, especially if we are interested in the establishment of large linguistic areas. (In what follows, I will make reference to Southeast Asia as one such large linguistic area – perhaps the best exemplar for which we currently have evidence; for a more detailed discussion, see Comrie (in press).) However, I emphasize that the identification of an area of typological similarity on the basis of WALS can only be the first stage in the investigation of language contact and areal diffusion; it provides, as it were, prima facie evidence for the existence of a linguistic area (sprachbund) that has arisen through language contact. Further research is needed to validate that areal diffusion is indeed responsible for the geographical pattern involved, and to work out the detailed processes of language contact involved. In some cases, it may not be possible to verify the initial assumption because of lack of relevant data. In

some cases, detailed investigation may suggest an alternative explanation, e.g. chance or universal preferences.

An initial illustration may be provided by a more detailed examination of the accompanying Map 1, reproducing Map 87 by Matthew Dryer from WALS. This map shows the relative order of Adjective and Noun within the Noun phrase, with four possibilities. Noun–Adjective and Adjective–Noun are transparent. "No dominant order" refers to languages for which the chapter author was unable to come to a clear decision that either of these two orders is clearly dominant over the other. Finally, and irrelevant to the following discussion, "Only internally headed relative clauses" covers languages that lack a category of attributive adjectives distinct from internally headed relative clauses. It will be noted that the order Noun–Adjective is more than twice as frequent as that of Adjective–Noun, but more interestingly that the order Adjective–Noun is frequent in really only one part of the world, namely Eurasia. (Here, as usually turns out to be the case in typological work, "Eurasia" excludes Southeast Asia, which, as will be noted below, often differs typologically from the rest of Eurasia in the geographical sense. In the case of the order of Adjective and Noun, there are a couple of other geographically peripheral exceptional areas, namely southwestern Europe, southwestern Asia, and the Himalayas – the last arguably an extension of Southeast Asia in this case – with the order Noun–Adjective.) This particular map is interesting in that it was used, in a pre-WALS version, by Dryer (1988) to suggest a particular hypothesis concerning the areal distribution of the feature value Adjective–Noun in Eurasia, namely that Eurasia is potentially available as a large diffusion area and that the feature value in question might indeed have diffused within this area, thus accounting for the striking imbalance between its rarity in the world as a whole versus its near-ubiquity in Eurasia.

2. Using WALS

In work since the appearance of WALS there have been basically two directions in the exploitation of the materials to establish potential linguistic areas. First, some linguists have worked directly with the database, employing primarily statistical analyses, based for instance on calculating typological distances between languages and then correlating these with such parameters as geographical distance or genealogical distance (e.g. languages within the same genus, languages within the same family, languages from different families). A good illustration of the application of this methodology is Dahl (MS). Second, others have worked directly with the maps, effectively "eye-balling" them to come up with clear patterns of areal distribution. In some cases this probably simply reflects lack of competence in the relevant statistical methods; this is certainly my own case, the reason why I have adopted this methodology in my own work, for instance Comrie (in press) and the present paper. Comparison of the discussions of Southeast Asia in Comrie (in press) and Dahl (MS) do, however, suggest that the two methodologies at least sometimes lead to strikingly similar conclusions.

Whichever methodology is employed, some basic issues need to be taken into consideration in the actual exploitation of WALS materials, and it is to these that I will now turn, with illustrations.

2.1. Discreteness of feature values

In WALS, a given feature is always listed as having a number of discrete feature values, with hierarchical classification or cross-classification of feature values. In some cases, this is clearly an adequate input to further investigations. For instance, in considering the order of Adjective and Noun, we clearly want Noun–Adjective and Adjective–Noun to be holistically distinct feature values. But other cases are either

clearly not of this type, or are unclear, and some modification needs, at least ideally, to be made in recognition of this.

The clearest cases that are not of the discrete type are those features whose values are arranged on a scale. For instance, Map 1 in WALS, by Ian Maddieson, deals with the size of the consonant inventory of a language, dividing the languages of the world into those with small (6–14), moderately small (15–18), average (19–25), moderately large (26–33), and large (34 or more) consonant inventories. The map thus identifies five discrete types. However, it is clear that "small" is closer to "moderately small" than it is to "large", so one would prefer a classification that captures this fact. (Going beyond this, one might note that a "small" language with 14 consonants is actually closer to a "moderately small" language with 15 consonants than it is to another "small" language with only 6 consonants; though whether such a language is also closer to an "average" language with 19 consonants is less obvious, since the scale is presumably not linear.) In this case, one would probably want to reflect directly the scalar nature of the feature, perhaps even quantify the distance between different numerical values in terms of a non-linear scale.

Other cases that linguists would probably agree on include the various instances of "No dominant order"; in the case of Map 1, for instance, this value presumably in some sense lies between the clear cases of Noun–Adjective and Adjective–Noun.

But there are other cases that become more controversial, perhaps involving particular theoretical positions that are not accepted by all linguists. For instance, my own Maps 98–99 in WALS dealing with alignment typology with regard to case marking present two discrete types: Nominative-accusative (standard), where either nominative and accusative are both case marked (e.g. Latin *filius* 'son (NOM)', *filium* 'son (ACC)'), or where the accusative is marked and the nominative unmarked (e.g. Hungarian *ház* "house (NOM)', *házat* 'house (ACC)'), and Nominative-accusative (marked nominative), where the nominative is marked and the accusative unmarked (e.g. Harar Ororo *sárée* 'dog (ACC)', *sáréen* 'dog (NOM)'. As my terminology may betray, I would actually have liked to include both of these as sub-types of a single "Nominative-accusative" type. Given the way the WALS database was constructed, this was excluded for purely technical reasons. But Dixon (1994: 64) argues that the Marked nominative type is "radically different" from the general run of Nominative-Accusative. One advantage of WALS here is the possibility, using the interactive reference tool, to combine feature values. Thus, if one wants to modify Maps 98–99 to show a single "Nominative-accusative" type, one can do so, i.e. WALS overall allows the best of both worlds (or the worst, for the more pessimistically inclined).

2.2. Logical dependencies among features and feature values

In the work that underlies Comrie (in press), the first stage was to look through the maps in WALS and identify those that show a clear boundary separating Southeast Asia from the rest of Eurasia. In principle, then, the number of maps that show such a boundary can be used as one piece of evidence in favor of identifying Southeast Asia, prima facie, as a linguistic area. However, a number of problems arise, as will be discussed in sections 2.2 and 2.3.

In the present sub-section I note that there are some purely logical relations between some maps, i.e. between some features, and between some feature values that precludes their being used as independent evidence. The clearest cases are where one map simply combines information from two other maps, as with Matthew Dryer's Maps 95–97 in WALS, which illustrate the possible combinations of feature values for, on the one hand, order of Object of Verb and, on the other, in order: order of Adposition and

Noun phrase, order of Relative clause and Noun, order of Adjective and Noun. In principle, these maps could have been left to be compiled by the interested user of WALS using the interactive research tool, but the editors felt that it would be worthwhile illustrating how maps can be combined in this way by including a very restricted number of such maps in the printed atlas. But needless to say, these maps cannot be used to provide evidence independent of the one-feature maps on which they are based (but see further section 2.3).

Some such logical dependencies involve only some feature values across maps. A good example of this is the series of maps by Greville Corbett, 30–32 in WALS, dealing with questions of grammatical gender. Map 30 deals with the number of genders in a language, with one possible value being, of course, zero; Map 31 asks whether the gender system of a language is based on sex or not; Map 32 asks whether the gender system is purely semantic or involves a combination of semantic and formal criteria. The new feature values introduced by Maps 31 and 32 are independent of one another, but the feature value "no gender" runs through all three, meaning that these maps are partially logically dependent on one another. But the logical dependence can be clearly identified and factored into the investigation, including one based on statistical methods applied to the database.

More insidious are cases where one feature value restricts the range of choices for some other feature, but without predicting precisely which value of this other feature holds. This question arose in a significant way in the work that underlies Comrie (in press). One of the typological characteristics of languages of Southeast Asia is that they have little or no inflectional morphology, and this distinguishes them from most languages of the rest of Eurasia, as is illustrated in Matthew Dryer's Map 110 in WALS, where languages of Southeast Asia show up clearly as having little or no inflectional affixation. Now, if a particular semantic opposition is shown primarily by affixation across the world's languages, then one might expect such a category to be rare or non-existent in languages of Southeast Asia. A case in point is the marking of plurality, where well over half the languages in the sample for Matthew Dryer's Map 33 in WALS have a plural affix (usually a suffix, with Africa, more especially Niger-Congo, providing most of the examples of prefixes). How is plurality marked in Southeast Asia? Logically predictable is the observation that plural affixes are virtually absent. But this does not mean that languages of Southeast Asia typically lack plural marking; in fact, they divide between languages that do indeed lack plural marking and those that use plural words, i.e. separate words (not affixes) that indicate plurality. In a case like this, it is harder to incorporate the "fractional" dependence of one feature value on another – the absence of affixation excludes the possibility of plural affixes, but does not determine which of the other possibilities will be chosen.

2.3. Correlations among features and feature values

Since the pioneering work of Greenberg (1966), a major interest among typologists has been the search for correlations among logically independent features. An example can be provided by examination of the accompanying Maps 2 and 3, dealing respectively with the order of Object and Verb within the Clause and with the order of Genitive and Noun within the Noun phrase. The hypothesis to be tested would be whether the order Object–Verb correlates with the order Genitive–Noun, the order Verb–Object with the order Noun–Genitive. Using the interactive research tool, one comes up with the following numbers of languages for each combination, showing basically that the correlation is indeed valid:

Genitive–Noun and Object–Verb	434
Noun–Genitive and Verb–Object	352
Genitive–Noun and Verb–Object	113
Noun–Genitive and Object–Verb	30

To the extent that such a correlation is valid across the languages of the world as a whole, the value of the individual features is diminished as independent evidence for the establishment of a linguistic area. This is a problem that continued to beset me in the work that underlies Comrie (in press), since a good number of the features differentiating Southeast Asia from Eurasia that stand out by eye-balling WALS maps relate to constituent order features that show high degrees of correlation (of the head-final versus head-initial patterns illustrated above). They are not completely independent, but since the correlations are never 100% they are not completely dependent either.

In fact further examination of these partially correlating features throws important light on Southeast Asia as a linguistic area, because although a series of constituent order maps highlight a difference between Southeast Asia and the rest of Eurasia, the boundary between the two areas is never exactly the same from one map to another, and indeed can sometimes vary quite considerably between a pair of maps, with "Southeast Asia" in a linguistic sense now appearing larger, now smaller. This can be seen by comparing the accompanying Maps 2 and 3 – the red (Southeast Asian) area is considerably smaller in Map 3 than it is in Map 2. Thus, the geographical distributions in Maps 2 and 3 cannot be accounted for solely in terms of correlation, indeed peripheral Southeast Asia harbors a fair number of the languages of the rare combination "Genitive–Noun and Verb–Object". Rather, the difference is consistent with a pattern of diffusion in which different features spread to differing extents, thus giving rise to a penumbra around core Southeast Asia that shares some but not all of the feature values that differentiate between Southeast Asia and the rest of Eurasia. Indeed, Comrie (in press), taking mainland Southeast Asia as the geographical core, shows that the southern periphery (e.g. Indonesian) shares a high number of feature values with the core, with the western periphery (e.g. Burmese) occupying an intermediate position, and the northern periphery (e.g. Chinese) showing fewest similarities to Southeast Asia and in many respects being intermediate between Southeast Asia and the rest of Eurasia (especially northern Asia).

Examination of exceptions to correlations can also be rewarding in that they may point to areas on the boundary of two diffusion areas that have borrowed values of different features from different directions. I suspect that this accounts for much of the intermediate status of Chinese, including the fact that southern varieties of Chinese (like Cantonese) are somewhat closer to Southeast Asia than are northern varieties like Mandarin (especially northern varieties of Mandarin, like that of Beijing). One striking example of this is illustrated by Map 4. Chinese is like Southeast Asian languages in having the constituent order Verb–Object, whereas most of the rest of Asia is Object–Verb; but Chinese is like the rest of Asia and unlike Southeast Asia in having Relative clauses before their head noun. What is interesting is that this combination – though readily understandable as the result of diffusion from different directions – is extremely rare among the languages of the world. Indeed, in Matthew Dryer's sample of 756 languages for his Map 96 in WALS, identical in relevant respects to the accompanying Map 4, exactly 5 languages combine the orders Verb–Object and Relative clause–Noun. Three of these are varieties of Chinese. One of the other two is Bai, a Sino-Tibetan language whose precise genealogical status – as a Tibeto-Burman language heavily influenced by Chinese or the other way round – is controversial. The last is Amis, an Austronesian language spoken in Taiwan, and which therefore cries out for more detailed

investigation to see whether this unusual combination may be attributed to contact with Chinese.

2.4. Statistical preponderance

Finally, in using WALS as prima facie evidence for language contact it is important to bear in mind the overall frequency of particular feature values across the languages of the world. Click consonants are very rare across the languages of the world, so the fact that they are virtually restricted to southern Africa is highly significant areally. In the patterns established in Comrie (in press), more emphasis should therefore be placed on rarer features that characterize Southeast Asia, such as complex tone systems or obligatory numeral classifiers, than on those that are frequent across the world as a whole, such as verb-like predicative adjectives, or even an overwhelmingly majority, such as the order Noun–Adjective.

3. Conclusion and prospects

In one sense, the answer to the question posed in the title. "What does WALS tell us about the diffusion of structural features?" is: Nothing, since WALS indeed tells us nothing directly about the process of diffusion, indeed it cannot even tell us unequivocally whether a particular geographical distribution is the result of diffusion. However, WALS does provide a rich set of data relating to the areal distribution of structural features of language than can, especially when used with appropriate care, form an important empirical base for studies of diffusion of structural features of language.

For instance, the identification of Southeast Asia as a linguistic area does not tell us how it came to be a linguistic area, indeed its identification as a linguistic area arguably opens up as many questions as it solves. For instance, both Dahl (MS) and Comrie (in press), using different methodologies applied to WALS, identify Thai as the most typically Southeast Asian language, which is surprising, given that Thai (and Tai languages more generally) are relatively recent arrivals in Southeast Asia from southern China and that many typically Southeast Asian features predate contact with Thai/Tai in many of the other languages of the area. Has Thai been a more voracious borrower than any of its lenders? Or is there a non-Tai substrate underlying Thai that passed on these features to Thai? WALS provides us with no evidence concerning these or other possibilities. But it does point us towards a problem crying out to be investigated.

Acknowledgment

For discussion of the use and interpretation of WALS I am grateful in particular to Balthasar Bickel, Michael Cysouw, and Östen Dahl, though none of them should be held responsible for the claims made in the present paper, for which I bear full responsibility.

References

Comrie, Bernard. In press. Areal typology of mainland Southeast Asia: what we learn from the *WALS* maps. To appear in *Manusya*.

Dahl, Östen. MS. An exercise in a posteriori language sampling.

Dixon, R.M.W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Dryer, Matthew S. 1988. Object-Verb order and Adjective-Noun order: dispelling a myth. *Lingua* 74: 77-109.

Greenberg, J.H. 1966. Some universals of grammar with particular reference to the order of meaningful elements, in J.H. Greenberg (ed), *Universals of Language*, 2nd ed. Cambridge MA: MIT Press, 73–113.

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

[The unpublished papers by Comrie and Dahl should be available from the respective authors.]