Some Principles on the use of Macro-Areas in Typological Comparison

Harald Hammarström and Mark Donohue
Max Planck Institute for Psycholinguistics &
Centre for Language Studies
Radboud Universiteit Nijmegen
harald.hammarstroem@mpi.nl
and
Department of Linguistics
Australian National University
mark.donohue@anu.edu.au

Abstract
While the notion of the ‘area’ or ‘Sprachbund’ has a long history in linguistics, with geographically-defined regions frequently cited as a useful means to explain typological distributions, the problem of delimiting areas has not been well addressed. Lists of general purpose largely independent ‘macro-areas’ (typically continent size) have been proposed as a step to rule out contact as an explanation for various large-scale linguistic phenomena. This squib points out some problems in some of the currently widely-used pre-determined, those found in the World Atlas of Language Structures (Haseplmath et al. 2005). Instead, we propose a principled division of the world’s landmasses into six macro-areas that arguably have better geographical independence properties.

Keywords
Areality, Macro-areas, typology, linguistic area.

1. Areas and Areality
The investigation of language according to area has a long tradition (dating to at least as early as Hervás y Panduro 1971 [1799] who draws few historical conclusions, or Kopitar 1829 who is more interested in historical inference), and is increasingly seen as just as relevant for understanding a language’s history as the investigation of its line of descent, as revealed through the application of the comparative method.

The underlying idea is that human interaction (as well as descent) is a major factor in shaping the languages involved (Weinreich 1953, Thomason and Kaufman 1988). Since, however, data on human interaction in prehistory is scarce except for some regions and some periods, we cannot always trace interaction patterns to their linguistic outcome. Rather, the study of areas and linguistic features is bidirectional: sometimes the linguistic data is known, and the interaction is inferred; sometimes the interaction patterns are known or hypothesised, and their role in shaping the languages is inferred. Either
inference is justified by (probabilistic) laws extractable from cases where both interaction and linguistic patterns are known, and the principle of uniformitarianism.

Most studies are devoted to the linguistic outcomes of a specific present contact situation. Essentially, the two options are to extract an area from the linguistic data empirically (Daume & Campbell 2007, Donohue & Whiting 2011, Muysken et al in press, and others), or to pre-determine the bounds of a region by some non-linguistic criterion and then test for the significance of these regions against linguistic data (Bickel and Nichols 2012, Muysken et al in press, Lucas et al. 2011, Michael et al. 2014). Such cases do not usually require enumerating all effects of interaction, typically not the effects involving languages peripheral to the language in scope. Alternatively, to phrase it differently, to describe one fragment of history is a contribution that does not necessarily require describing every fragment of history. However, equally important are studies that are devoted to ruling out language contact as an explanation for some linguistic data in favour of a genealogical and/or universal account (cf. Bakker 2010). Doing so is typically a much heavier task, since it requires encompassing all effects of contact. The task would be made simpler if it were possible to divide the world into independent discrete micro-regions to be used as stratification.

Human interaction is enormously dynamic, involving a number of factors that seem intractable (but see section 3). However, one strong limiting factor on human interaction is that, in pre-industrial times, adjacency was a prerequisite for any significant interaction. Further refinements could possibly be made with reference to mountains, rivers, deserts or other geographically bounded factors. Be that as it may, there is no hope of ever achieving an all-purpose list of independent discrete micro-regions this way, since most micro-regions would be adjacent to other micro-regions, and we have no means of rejecting the assumption of interaction between them. There are, however, some principled possibilities for macro-regions, given the continent shapes and the water bodies dividing them.

In section 2 we discuss the principles and outcome areas of the widely-used World Atlas of Language Structures (Haspelmath et al. 2005, henceforth WALS). In section 3, we offer a macro-areal division which is based purely on the principles of geographical independence, without reference to linguistic data. We argue that this division, though not dramatically different, has advantages over that found in the WALS.

2. WALS Areas and the principles behind them

Given the existence of continents with large bodies of water separating them, macro-regions offer some possibilities to achieve a general-purpose list of discrete and
independent areas. Dryer (1989:264) takes up this line and sets up five large continental areas: Africa, Eurasia, Australia-New Guinea, North America and South America. In Dryer (1992: 84-85), this list is revised to six areas: Africa, Eurasia (excluding southeast Asia), Southeast & Oceania, Australia-New Guinea, North America and South America; these areas are now widespread in typological work, including their adoption in the World Atlas of Language Structures (though see footnote 1). Dryer (1989: 83-85, 1992: 267-270) contains ample discussion of various difficulties involved in testing for areal (and genealogical) biases in linguistic feature distributions, and the macro-areal division clearly has the ambition to “control for large-scale areal phenomena” (Dryer 1992: 84-85) which requires the “assumption ... that these areas are independent of each other” (Dryer 1992:84). At the same time, however, his macro-areal division also aims to have “areas ... roughly comparable in genetic and typological diversity” (Dryer 1989:268-269, 1992:84). The WALS classification into macro-areas directly continues Dryer’s (1989, 1992) work. We shall decompose the two requirements for areal inclusion set up by Dryer:

- **Areal-genealogical control** ("control criterion"): that the areas can serve as (part of a scheme for) a control for areal and genealogical effects, and
- **Comparable in genealogical and typological diversity** ("balance criterion"): that the areas should be roughly comparable in genetic and typological diversity.

Examples of what concrete choices of either kind in the context of WALS will be given in the reminder of this section. We then discuss the theoretical motivations thereof, and propose a different solution.

In WALS the world is divided into six areas, as listed in (1).¹

(1) Africa:
  - Australia-New Guinea:
  - Eurasia:
  - North America;
  - Southeast Asia and Oceania;
  - South America.

The basis for assigning a language into one area and not another is generally straightforward, especially in the centre of each area. There are, however, numerous cases, predictably towards at the borders between the areas, in which assignment to areas

---
¹ We note that this macro-areal classification is found only in the CD-ROM accompanying the 2005 WALS edition (available online: [http://www.eva.mpg.de/lingua/research/tool.php](http://www.eva.mpg.de/lingua/research/tool.php), accessed 1 June 2012). It is not discussed in the 2005 book, and the WALS online (Dryer et al. 2011) offers no alternative areal classification(s).
does not follow from the geography. Map 1 shows the distribution of these six areas, as found in the coding of languages included in *WALS*, mapped according to the coordinates declared there.

Map 1. Regions pre-defined as ‘macro-areas’ in *WALS*

Key: from west to east, the regions are: Africa (black); Eurasia (white); Southeast Asia and Oceanic (black); Australia and New Guinea (white); North America (black); South America (white)

A first point of interest is the Bering Strait which divides the macro-areas Eurasia and North America. The relevant portion of the *WALS* classification is shown in Map 2, with three languages of mainland Asia being classed as part of North America, not mimicking the water-division.
The assignment of the three mainland Asian languages to the North American macro-area is deliberate (Dryer 1989:268, 1992:83). Whenever there are genealogically related (on the genus level) languages that span across macro-areas, all the member languages are assigned to exactly one of the areas. Here, the three languages concerned are Naukan, Siberian and Sirenik Yupik, part of the Eskimo-Aleut family and the Eskimo genus, that is for the most part found in North America.

Another example is the Middle East where the genus Semitic stretches from Africa into Eurasia. The locations of the languages concerned are shown in Map 3.
In this case, the intrusion of languages from the Africa-assigned genus reaches further into the alien continent, so that there are Eurasia-languages which are closer to the African continent than Africa-assigned ones, e.g., on Cyprus where ‘Arabic (Kormakiti)’ is classed as African, while Greek (Cypriot) is Eurasian, and, e.g. Neo-Mandaic classed as African versus various Iranian languages classed as Eurasian.

The very extreme of the genus assignment principle concern languages very far from their congeners due to colonial expansion. They can be found at extraordinary distances from their macro-area, such as ‘Afrikaans’ (located in South Africa) being part of the Eurasia macro-area.

Another case of non-trivial border-area assignment illustrates a different principle of the Dryer (1989:269) division. Map 4 shows the extent to which ‘Southeast Asia and Oceania’ is found far from the usual geographic use of the term. Here, all Tibeto-Burman languages have been assigned away from Eurasia, including Mandarin Chinese in the north of China and Bodish languages as far west as Balti in Pakistan (as well as numerous languages of India, and all languages of Nepal except (Indo-Aryan) Nepali and the isolate Kusunda). This appears to have been done in order to increase the genealogical and/or typological diversity of the Southeast Asia area, and not because the languages in question form one genus. The Sino-Tibetan languages in question belong to many different genera and have some important typological differences to the other Southeast Asian families, wherefore their assignment to Southeast Asia increases the genealogical and typological diversity of that region (and more so than their assignment so Eurasian would have achieved).

Map 4. ‘Southeast Asia and Oceania’ affiliation on mainland Asia

The Darien gap is the widely accepted geographic boundary between South America and the rest of the Americas. The WALS languages in that region are shown in Map 5.
As in the previous examples, genera found on both sides of the gap, are assigned to one area each. But there is also an inconsistency with previous examples here: Miskito Coast English Creole is part of ‘South America’, yet the languages genealogically related to Miskito Coast Creole English are found in the Eurasian and African macro-areas.

The macro-area ‘Southeast Asia and Oceania’ extends from the mainland of (south-east) Asia into the eastern Pacific, but excludes those languages in the ‘Australia and New Guinea’ macro-area, in the centre of the ‘Southeast Asia and Oceania’ range. The point at which these two macro-areas interact is shown in Map 6.
Here we find another likely inconsistency in the macro-area assignments. The genus Oceanic, making up the eastern half of the white dots on Map 6, is clearly adjacent to the languages of the ‘Australia and New Guinea’ area rather than the Southeast Asia area. The assignment of the Oceanic genus to ‘Australia and New Guinea’ would also have increased the typological diversity of that region.

We have seen that while many of the ‘macro-areas’ listed in WALS follow approximately continental norms, there are many cases in which the ‘macro-area’ groupings are simply another way to represent families. This would seem to be counterproductive to achieving statistical independence between macro-areas, yet at the same time they are explicit rules in the schema underlying WALS. We will therefore take a closer look, recalling the two desiderata from the beginning of this section.

Regarding areal-genealogical control (the “control criterion”), the classic scheme for testing whether some discrete property a language can have (henceforth “single-language property”) is overrepresented after controlling for genealogical and areal factors, is to divide languages of the world into genealogical cells (genera, families, etc.) and geographical cells (areas or macro-areas of some sort), cf. Bakker (2010, and references cited therein). To the extent that the cells are independent, classical techniques from probability theory can be validly applied (Cochran 1963). This is indeed also Dryer’s (1989, 1992) tactic, but, as we have seen, it is applied with various imperfections. This means that even if the genealogical and geographical cells are assumed to be independent, the cross-area assignments allow for the possibility that, even if one selects languages from different genera and macro-areas they are not guaranteed to be independent. (If the unit of sampling is genera rather than languages from genera, the same holds, i.e., selecting genera from different macro-areas does not guarantee that those genera are areally independent.) With a strict geographical division and a strict genealogical division, selecting languages from different areas and genealogies (be it families, genera or some other division) does preserve independence.

Regarding the desideratum of comparability in genealogical and typological diversity (the “balance criterion”), a stipulation that the geographical cells should be of roughly equal size can not be motivated from sampling theory, if the matter of interest is a single-language property. It may be motivated if the matter of interest is some property (henceforth “language-set property”) a set of languages can have (rather than a single language) or a (possibly continuous) value which requires more than one language to estimate. To make this difference clearer, consider the following example. Suppose we have a series of identical coins and we want to find out whether this type of coin is biased. If we can flip the coins independently, it is straightforward to estimate the bias of the type coin by combining the evidence from each coin flip. If, however, someone sprayed glue between various subsets of the coins, flipping one coin may flip other coins
connected by glue. We may then wish to divide the series of coins into cells that are connected by glue, such that no two different cells are connected by glue. If we then combine the evidence from flips of coins from different cells (rather than every coin) we will be able to use classical techniques from probability theory to estimate the bias of the type of coin itself, eliminating any effects of the glue distribution. This is the situation parallel to testing for a biased distribution of some single-language property. Coins correspond to languages, glue to areal or genealogical influence, and the single-language property is heads/tails in the case of coins, but may be, e.g., Noun-Numeral order in the case of languages. We reiterate that for this case, it makes no difference if the cells unconnected by glue are different or equal in size. Two independent coins (two cells, one member each) is equivalent evidence to one independent coin and one thousand coins connected by glue (two cells, one member and thousand-member), when applying classical techniques in probability theory (Cochran 1963). Suppose now that we are not interested in a discrete property of a coin, but we are interested in, e.g., the interaction of glue and coins – perhaps the glue sticks less to the head side of the coin than the tail side, or, perhaps there are different kinds of glue – then there may be a reason two divide the coins into cells so that every cell contains coins with glue on both the head side of some coin and the tail side of some coin. It may even be so that it is motivated to have cells that are not completely glue-separated, if this allows for more cells of the required size and distribution.

Since the two desiderata ‘control criterion’ and ‘balance criterion’ can be opposing, it is not possible to devise a general strategy that optimises both aims simultaneously. It would be possible if there was a weighting of the two aims, but the appropriate weight clearly depends on the exact feature in question. For this reason it may be an ill-founded idea to devise an all-purpose list of areas to be used for language-set properties. For single-language properties, however, the ‘balance criterion’ is irrelevant and thus a geographically optimal all-purpose area-division is feasible. We pursue this in the next section. (It contains only a minimal requirement on the number of languages in a cell and thus the amount of potentially underused independent data points discarded is very small.) We then compare the resulting areal division, and it turns out that empirically, this area division is at least as good in terms of ‘balance criterion’ as that in WALS, whose division has some of the ‘control criterion’ compromised. We thus argue that the macro-areal division in this paper is preferable to the one in WALS, even using Dryer’s (1989, 1992)’s desiderata and regardless of which of the ‘control criterion’ and ‘balance criterion’ is more important.

3. Principles for Macro-Areal Division

If geographic regions are intended to model socially significant ‘interaction zones’, then water poses a particular problem, one that emphasises social factors over Euclidean ones
when determining areas, and macro-areas. For instance, Malagasy, spoken on the island of Madagascar, south-east of the African mainland, is classified in WALS as part of the Southeast Asia and Oceania macro-area, despite being clearly closer to Africa. Can this be justified? One argument in favour of this grouping is that, historically, the language of Madagascar is descended from (Austronesian) languages of Southeast Asia, and that trade and interaction is recorded involving long-term contact with Southeast Asia. On the other hand, the island is part of the African region, and Malagasy contains loanwords that have Bantu, not Austronesian, sources, and the Malagasy population has as strong an African biological source as a Southeast Asian one (e.g., Hurles et al. 2005), implying very strong social contact with the African mainland, and a very strong case for Madagascar to be considered at least as African as it is Southeast Asian. In the present paper, we only assume that, ceteris paribus, the larger the body of water the more difficult the social interaction across it. We do not assume that interaction is impossible across any body of water.

Many other possible classifications could be suggested; one other objectively verifiable one is that obtained by classing the world into areas based on social features (Burton et al. 1996); this does not address actual interaction, but at least takes account of social similarity (Map 7).

Map 7. Social ‘macro-areas’

A truly socio-geographic classification of the world into (macro-)areas seems, for the moment, to be an intractable problem. A more modest, and importantly attainable goal is to devise macro-areas on geographical grounds. On purely geographical grounds, we suggest the following list of macro-areas, with locations shown in Map 8 (Donohue et al. 2013):
(2) Africa; Australia; Eurasia; Multinesia;\(^2\) North America; South America.

Map 8. Suggested revised ‘macro-areas’ (listed in (2))

These areas have the virtue of being maximally defined by physical geography, and being completely independent of linguistic affiliations. It also avoids any subjectivity required to decide what is a genus of not, as required for consistency with the cross-areal genera found in WALS. In the remainder of this section we offer a mathematical proof for the fitness of the classification we have advanced with Map 8.

We propose the following theoretical underpinning for diving the world into linguistic areas on geophysical grounds. From a purely landmass versus water perspective, a division into areas might take the following form.

First, we define the terms *partition* and *distance metric*:

**Partition:** A partition \( C=[C_1,\ldots,C_n] \) is a division of the world’s landmass into cells. A partition must exhaust the world’s landmass (place any landmass in some cell), all cells must be non-empty and no two cells may overlap.

\(^2\) We suggest the novel name ‘Multinesia’ for the region which encompasses Indo-Malaysia, Melanesia, Micronesia and Polynesia. We do not use the existing label ‘Pacific’ as this label sometimes includes Australia.
**Distance Metric:** A distance metric $\delta$ is a function that provides a distance (in kms) for any two pieces of landmass.

For example, one partition may be $C_1$ as the landmass defined by the official borders of Sweden, $C_2$ comprising of Madagascar and Greenland, $C_3$ of mainland South America (cut off at the Panama canal and leaving out any adjacent islands), and $C_4$ the rest of the world’s landmass. A possible $\delta$ is simply the shortest distance (in kms) between the coasts of two landmasses (henceforth *coast-distance*). Another possible $\delta$ is the distance (in kms) between the centrepoints of two landmasses (henceforth *centre-distance*).

We may characterise any partition by its *internal coherence* and its *disjointness*:

**Internal Coherence:** The internal coherence of a cell $ICC(C_i)$ is defined as the shortest distance required to connect all members of the cell. The internal coherence of an entire partition is defined as the sum internal coherence of all its cells:

$$ICP(C) = \sum_i ICC(C_i)$$

**Disjointness:** The disjointness of a partition is the the length (in kilometers) of the border found between adjacent landmasses in different cells. If $ADJ(C_i, C_j)$ is the number of kilometres of adjacent landmass between landmasses in $C_i$ and $C_j$, the disjointness of the entire partition is:

$$DP(C) = \sum_{i \neq j} ADJ(C_i, C_j)$$

For example, the Internal Coherence of a cell consisting of mainland South America is 0 since it is one coherent landmass. The Internal Coherence of a cell consisting of Madagascar and Greenland, on the other hand, is much worse, since Madagascar and Greenland are many kilometres apart. The disjointness of a cell of Madagascar and Greenland is, however, 0 since both are islands and thus have no land adjacent to any other landmass. The disjointness of mainland South America is the length of the Panama Canal\(^3\) since the corresponding strip of South American mainland is adjacent to the North American landmass along that line.

Naturally, an optimal partition of the world’s landmass is one that minimises $ICP(C) + DP(C)$.

Next, we add languages to the problem. By definition, all languages are spoken by humans, and all humans live on land (even if some groups are known to spend a significant amount of time on the sea (Sopher 1965), they still have a home base on land). The task of diving the world into (linguistic) macro-areas then reduces to dividing the human-inhabited landmass into areas. A further constraint is that we are interested only in partitions where all cells have at least a certain size. For linguistic areas, this size constraints may be naturally expressed in terms of number of languages, i.e., the areas in each cells have to be the home to at least $k$ languages.

---

\(^3\) This example assumes that the Panama canal has not been dug yet or that it is too thin to count.
The problem of finding the optimal geophysical division into linguistic macro-areas is thus: Given a number $n$ of areas, and a minimum number of languages $k$, what is the partition $C=\left[C_1, \ldots, C_n\right]$ of the inhabited landmass that minimises $ICP(C)+DP(C)$?

Since the problem is of combinatorial nature, there is no guarantee that the solution can be computed quickly. However, in WALS $n=6$ and if we set only a small size restriction, e.g., $k=250$, it is possible to find the optimal partition. It is derived through a series of steps as outlined below.\(^4\) Figure 1 shows a simplified schema of the same inference.

1. There are a total of 204 coherent landmasses large enough to appear with the resolution used in our experiments (see footnote 4).
2. 95 of those are inhabited with at least one language. A selection of these 95 landmasses are shown in Table 1.

### Table 1: A selection of coherent landmasses and their respective area and number of languages. The numbers deviate by minor amounts from official figures due to the resolution used in the computations.

<table>
<thead>
<tr>
<th>ID</th>
<th># languages</th>
<th>Area (km(^2))</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>3 685</td>
<td>79 866 683</td>
<td>Eurasia-Africa</td>
</tr>
<tr>
<td>9</td>
<td>1 261</td>
<td>39 230 623</td>
<td>Mainland Americas</td>
</tr>
<tr>
<td>156</td>
<td>877</td>
<td>804 588</td>
<td>New Guinea</td>
</tr>
<tr>
<td>124</td>
<td>307</td>
<td>7 636 894</td>
<td>Australia</td>
</tr>
<tr>
<td>123</td>
<td>147</td>
<td>745 879</td>
<td>Borneo</td>
</tr>
<tr>
<td>134</td>
<td>142</td>
<td>287 335</td>
<td>Luzon</td>
</tr>
<tr>
<td>132</td>
<td>83</td>
<td>160 851</td>
<td>Sulawesi</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>91</td>
<td>11</td>
<td>592 481</td>
<td>Madagascar</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td>2 390 173</td>
<td>Greenland</td>
</tr>
</tbody>
</table>

3. There are 4 landmasses that are in themselves large enough ($k=250$ languages or more) to be a cell of their own. The requirement is that, in the end, 6 cells need to have at least $k=250$ each. In the optimal solution, those two further cells, should they consist of splits of the big-4 in combination with other landmasses or solely of combinations of the other landmasses? It is always suboptimal to break coherent

---

\(^4\) All computations were done using the elevation/water data points from the Shuttle Radar Topography Mission (SRTM), http://dds.cr.usgs.gov/srtm/ accessed 16 March 2011, at a resolution of 0.5 degrees. Maps 9-10 reflect this resolution. The location of languages is from a database of centre-coordinates derived from the Ethnologue (Lewis 2009) with extinct languages and other amendments added by the first author.
landmasses, unless this is needed to meet the requirement that other cells should be large enough.\(^5\)

4. We may check all the combinations of the non-big-4 landmasses versus all the possible ways to split the big-4 anywhere that yields two large enough areas. As is also intuitively obvious, the two least expensive of the big-4 cuts that yield large enough areas on both sides are the Panama canal (82 km) and Suez canal (164 km). As for combinations of the non-big-4 landmasses, brute-force computations show that the least expensive way to get at least \(k=250\) languages is to connect Borneo to the Philippine islands via the small inhabited islands intervening. That is, to connect\(^6\) Borneo to Bonggi island (167 kms), Bonggi to Palawan (156 kms), and Palawan to Luzon (273 kms) totalling 597 kms. In any other constellation where those islands do not form a cell of their own, the cheapest way to connect them anywhere totals 421 kms. So to cut Eurasia-Africa at the Suez canal and attaching Borneo-Philippines to their closest neighbors will incur a cost of \(164+421 = 585\) kms in disjointness and internal coherence. The solution with an independent cell with Borneo-Philippines at its core will incur a slightly higher cost \(0 + 597 = 597\) kms. (How to allocate the remaining landmasses is completely parallel for these two possibilities, so the calculation presented is the crucial one.) The Borneo-Philippines cell was the least expensive way to achieve a large enough cell, so the dividing at the Suez and Panama canals is better than any partition that makes six large enough cells without dividing at Suez and Panama.

5. After the division at Suez and Panama, there are six landmasses which are large enough. Any further split is suboptimal, because that would only incur an adjacency cost, without a gain anywhere else in the constraint or optimality formula. All that remains is to optimally allocate the remaining 91 landmasses. In other words, to partition each one into the six landmasses such that the total ICP is minimised. This is a computationally tractable problem known as the minimal spanning tree (MST) problem for undirected graphs. In our case, the nodes are the landmasses and edges the distances between them. We used Kruskal’s algorithm (Kruskal 1956) with a small modification for solving the MST problem at hand. The small modification is that any solution which connects the six cells is disallowed and that the algorithm stops when there are six connected components (rather than one).

---

\(^5\) This is also true if there are fewer coherent landmasses than cells required; this is clearly not the case here.

\(^6\) The example uses coast-distances. The results for centre-distances are parallel.
Figure 1: A simplified schema of the inference of the optimal division into macro-areas where \( n=6 \) and \( k=250 \).

The resulting optimal divisions into macro-areas where \( n=6 \) and \( k=250 \) are shown in Maps 9 and 10.

Maps 9 and 10. The optimal division into macro-areas where \( n=6 \) and \( k=250 \) using coast-distances (top) and centre-distances (bottom). The maps are shown in Mercator projection but all distances used in the computations are the actual distances, i.e., with earth as a sphere.
There is a difference if coast-distances or centre-distances are used. In the case of coast-distances, Greenland-Iceland belongs to North America and Eurasia stretches all the way to the Weber line, close to the western shores of New Guinea. In the case of centre-distances, the New Guinea area takes over the entire archipelago off Southeast Asia, including Taiwan, and Greenland-Iceland is drawn into Eurasia. Which one of centre- or coast distances is relevant? This depends on the exact nature of the areal phenomenon in question. The pivotal point of an area is the centre of gravity. If the areal phenomenon is very diffusible, the physical obstacles, i.e., large bodies of water matter more than distance from the centre. On the other hand, if the phenomenon is not very diffusable, the distance to the centre is more relevant than the physical obstacles.

The differences in the outcome areas (be it centre or coast distance induced) compared to WALS are not dramatic regarding Africa, North America and South America. There is, however, a significant difference regarding Eurasia where WALS has separated Southeast Asia from this landmass (yielding a very long adjacent border) as well as the position of Australia. This amounts to something like 2 differences out of 6. If one uses the approach to statistical testing of typological universals found in Dryer (1992), where a trend is deemed significant if present in at least 5 of 6 (macro-)areas, then even the flip of one area can push a trend out of conventional levels of significance (Cysouw 2005).

Finally, recall that one of the desiderata in Dryer (1989, 1992) was that the macro-areas should contain roughly similar amount of genealogical and typological diversity. A comparison of the number of language families according to Glottolog 2.2 (Nordhoff et al. 2013) for the WALS areas (Map 1) and the areas of Map 9 in the present paper are given in Table 2. As can be seen, the genealogical diversity is more balanced in the Map 9 classification, even though this was not a design criterion for it. The most diverse area in Map 9 is less diverse than that in the WALS classification, and the least diverse area in Map 9 is much more diverse than the least diverse WALS area.

---

7 The Weber line is a significant biogeographical boundary line with respect to mammalian fauna (Simpson 1977), lying east of the Wallace line and west of the Lydekker line.

8 We use this classification since it is the only one to both (i) (briefly but) systematically disclose all justifications for the family divisions given, and (ii) classify all known languages in the world.
Table 2. A comparison of genealogical diversity in the \textit{WALS} areas vs. the areas suggested here.

<table>
<thead>
<tr>
<th>Area</th>
<th>\textit{WALS}</th>
<th>Map 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>50</td>
<td>50.67</td>
</tr>
<tr>
<td>Australia-New Guinea</td>
<td>158</td>
<td>33</td>
</tr>
<tr>
<td>Eurasia</td>
<td>28</td>
<td>33.67</td>
</tr>
<tr>
<td>North America</td>
<td>72</td>
<td>71.33</td>
</tr>
<tr>
<td>SE Asia &amp; Oceania</td>
<td>8</td>
<td>128.33</td>
</tr>
<tr>
<td>South America</td>
<td>110</td>
<td>109</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>426</td>
<td>426</td>
</tr>
</tbody>
</table>

As for typological diversity we cannot perform a quantitative comparison for lack of systematic typological data;\footnote{In particular, the data in the \textit{WALS} itself is not sufficient (Hammarström 2009).} an informal comparison, however, suggests that the differences are not dramatic. For example, neither the typological diversity of our ‘Australia’ nor that of \textit{WALS}’s ‘SE Asia & Oceania’ is large, and the remainder seems to balance out.

4. Conclusions

We have shown that it is possible to come up with a list of objectively defined pre-defined areas that can be used as normative controls in cross-linguistic work. These areas are blind to language affiliation, and can be justified on a continental basis. The areas so defined are not of even size, but are all sizeable enough to allow for meaningful comparison. The number of macro-areas used in this paper was set to six, following assumptions in previous work. Although beyond the scope of the present paper, the scheme used for dividing the world into areas described in this paper may, in principle, be applied with any other number as well.

A list of the languages found in \textit{WALS}, with their areal affiliations as classified in this paper, can be found in the appendix to this paper.

Acknowledgements

The authors wish to thank three anonymous reviewers for their highly useful comments on a previous draft. The first author acknowledges support from ERC Advanced Grant 230310 ‘Traces of Contact’. The second author acknowledges the ARC Discovery Grant DP1093191 ‘New Guinea’s place in Southeast Asia’, and the ARC grant FT100100241 ‘Understanding Human History in Asia through linguistic analysis’.
References


