

# Pairwise comparisons of typological profiles

*Søren Wichmann & Eric W. Holman*

## 1 Introduction

Rare linguistic features, languages possessing many rare characteristics or areas possessing many languages with rare characteristics are all relative phenomena. Studying *rara* takes the researcher into a complex area, an area which gets the more complex the more languages are involved. To develop a sense of the determinants of this complexity it is worthwhile to conduct a thought experiment where a maximally simple situation is considered first, and where various additional layers are only subsequently added. In a maximally simple situation there would only be one language in the world, and in this situation rarity would be a non-existing phenomenon. In a situation with just two languages, rarity still could not be defined, but we could talk about similarities and differences between the two languages. Two languages can be more or less similar. It is probably a truism that two major factors contributing to similarities among languages are common descent and diffusion. Other factors, whose effects are less easy to identify, are universal tendencies and chance. Leaving these last two factors aside, we can predict that, all else being equal, geographic propinquity and relatedness are expected to enhance similarities among the members of any pair of languages. If we enlarge the picture to include three languages, then, we would predict that among a set of three unrelated languages the two most proximate geographically would be the most similar; and within a set of three languages spaced equally apart where language A and B are related and language C unrelated to the two others we would expect A and B to be more similar to one another than either is to C. Once we have entered a situation where three languages are considered the concept of *rara* becomes relevant. A trait found in only one language to the exclusion of the two others is, by any definition, rare. Such a trait could exist in any language within a set of three, but among the three languages the least closely related or most geographically remote language would have a higher chance of possessing such a rare trait. Given a world-wide sample of languages we would still expect geography and relatedness to be at the root of the phenomenon of *rara*, but with a large set of languages the confounding

factors accumulate. Genealogical and areal biases in the sampling becomes an issue, peculiarities of geography enter the picture, differential degrees of feature attestation must be considered, as well as biases in the selection of traits. One may try to take all or most of these factors into account when developing a statistical approach to *rara*, cf. Cysouw (in press) for an attempt, but, in the end, if a given feature is promoted to the status of *rara*, or if a given language is identified as being quirky or if an area is found to be typologically unusual it will be hard to interpret the result because so many factors are in play: diffusion, relatedness, chance, universal tendencies, peculiarities of geography, sampling of languages and features, and degrees of attestation. In addition, what is rare today may have been common yesterday and may not tell us anything interesting about languages. In other words, the global linguistic typological profile is in constant flux, and there is no doubt that various historical contingencies which are beyond the observational reach of the investigator contribute to a large extent to these dynamics.

In this paper we are interested in the phenomenon of “language rarity”, i. e. the degree to which languages differ from one another as a whole. In the belief that a complex phenomenon is best approached by isolating its components, we will consider only pairs of languages. As mentioned, it is really only in the situation where three or more languages are compared that rarity is a relevant concept, but since the multilateral comparisons necessary for making observations on rarity decompose into a set of pairwise comparisons, a fundamental approach to the problem of why some languages stand apart from all or most others is to ask why pairs of languages exhibit differential degrees of similarity.

Our hypothesis is that there are two major factors which contribute to similarities among languages: relatedness and propinquity. The latter factor, i. e. the influence of geography on typological similarity, was investigated in Holman et al. (2007). In this paper, then, we focus more on the role that genealogical relatedness plays with respect to similarities among language pairs. In particular, we are interested in knowing whether there is a cut-off point  $S_{high}$  in the amount of similarities such that we can be sure that language pairs that have more than  $S_{high}$  similarities are all generally thought to be related, and we also want to know whether there is a cut-off point  $S_{low}$  at the other end of the scale such that all languages having less similarities than  $S_{low}$  are thought to be *unrelated*. In other words, if a language is relatively similar to some other language, as Burushaski is to Telugu, just to name an example, does this imply that the two languages are related according to commonly

accepted classifications? Or, if two languages are mutually very exotic, as Burushaski and Samoan, for instance, does this imply that they are thought not to be related in commonly accepted classifications?

The data we use, as well as the genealogical classification, are from *The World Atlas of Language Structures* (Haspelmath et al. 2005, eds., henceforth *WALS*). The conclusions must of course be seen in relation to this particular dataset. Thus, when we observe a certain amount of typological similarity between two languages, this is strictly and only similarity in terms of the kinds of features investigated in *WALS*. The dataset includes 134 non-redundant features, each of which distinguishes two to nine discrete values. All of these are quite generic typological features. Our conclusions are also limited to the amount of data available. We have required that for any language pair in our sample there should be 45 or more features attested for both members of the pair (a motivation for this precise number follows shortly). This has limited our sample to 320 languages and 29,810 pairs of languages compared. Among these pairs, there are 1,099 which are considered to be related, according to the classification in *WALS*. This classification, described by Dryer (2005), has two taxonomic levels. Families are defined as the most inclusive groups believed by a majority of specialists to have descended from a common ancestral language. Genera are defined as the most inclusive groups whose common ancestor is believed to have existed no more than about 3,500 to 4,000 years ago. Languages in the same *WALS* family are henceforth called “related”.

## 2 Results

Figure 1 on the next page presents the overall results of the investigation. As can be seen, the more similar languages get, the greater the probability is that they are related. The figures on which the curve is based are presented in Table 1 on the following page. Percent similarity was defined as the percentage of available features for which both languages have the same value. We have binned language pairs in 5% intervals from 10% to 90% similarity. Figure 1 plots the percentage of the pairs in each bin that are related, as a function of their mean percent similarity. Table 1 gives some additional information: it also shows how many language pairs belong in each interval. This is important for the interpretation of the results, as we shall see shortly.

Before giving our interpretation let us explain why we have chosen the criterion that language pairs should have 45 or more features attested for both

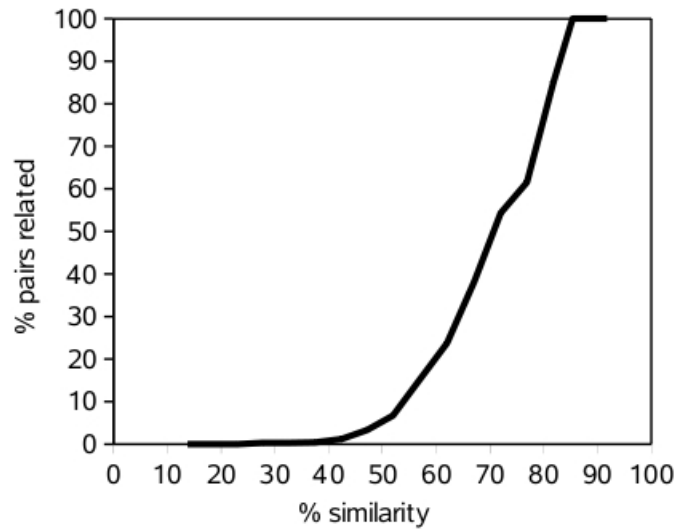


Figure 1. The probability of finding related languages as a function of their similarity

Table 1. Mean percent similarity between members of pairs, percent of language pairs that are related, and number of language pairs, in each similarity interval

% similarity interval	Mean % similarity	% related	Pairs
10.0–14.9	13.8	0.00	11
15.0–19.9	18.1	0.00	91
20.0–24.9	23.0	0.00	443
25.0–29.9	27.8	0.26	1,566
30.0–34.9	32.7	0.33	3,904
35.0–39.9	37.5	0.40	6,019
40.0–44.9	42.4	1.20	6,772
45.0–49.9	47.2	3.26	4,873
50.0–54.9	52.0	6.68	3,520
55.0–59.9	57.1	15.41	1,551
60.0–64.9	62.0	23.72	666
65.0–69.9	67.1	38.24	238
70.0–74.9	72.0	54.26	94
75.0–79.9	76.9	61.54	39
80.0–84.9	81.8	85.00	20
85.0–89.9	85.4	100.00	2
90.0–94.9	91.8	100.00	1

languages. We tested results for different numbers of features (30 or more, 45 or more, 60 or more, 75 or more). It turns out that for a criterion of 30 or more features the curve is rather similar but not quite as steep, showing less dependence between the amount of similarity and the probability of finding related pairs. This indicates that the fewer features one operates with, the more prominent is random sampling variability in percent similarity. When operating with a criterion of 60 or more attested features the curve becomes uneven, indicating that the higher criterion passes too few pairs for stable results. This becomes even more pronounced when the criterion is 75 or more features. Obviously, with a more extended database the number of features taken to be criterial could be raised, but around 45 is a number that suits the data available in WALS because a number in the vicinity of 45 maximizes the combined information in terms of both the number of languages and features available.

It may be of interest to mention the language pairs that fall in the lower and upper ranges of the percentage of shared values. Collectors of linguistic trivia may find it interesting that the members of the most divergent language pair in the world (in our dataset), i. e. Tümpisa Shoshone and Wari', are both native American languages, that someone who wants a radical alternative to Romance linguistics should turn to Nivkh, and that someone wanting to study a language as different as possible from Swedish should visit the Koasatis. Lists of the 20 most divergent language pairs and the 20 most similar ones are provided in Tables 2 on the next page and 3 on page 247.

While Table 2 does not point in any specific direction and remains a curiosity, Table 3 provides fragments of information which fits into the larger picture that emerges from our study. We note that two pairs of unrelated languages, Vietnamese–Thai and Khmer–Thai, turn up in this list, which otherwise consists of genealogically related language pairs. Furthermore, the rest of the pairs represent a mixture of languages related to different degrees.

Returning to Figure 1 and the associated data in Table 1 let us proceed to overall interpretations. We set out asking whether there is some degree of similarity in typological profiles beyond which it is certain that languages are related. The answer is positive, but nevertheless discouraging. Members of language pairs in the sample that are 81.5% or more similar are all related. But only 12 pairs of languages are similar to such an extent, in spite of the fact that there are 1,099 pairs of related languages in the sample! On the other hand, if there are less than 25% shared feature values all language pairs will be unrelated, and this goes for 545 pairs in the sample. If one

Table 2. The 20 most divergent language pairs in the sample

Language A	Language B	Features compared	% similarity
Tümpisa Shoshone	Wari'	48	10.4
Archi	Tukang Besi	46	13.0
Maybrat	Limbu	45	13.3
Italian	Nivkh	51	13.7
Burushaski	Samoan	49	14.3
Tzutujil	Burmese	49	14.3
Ju 'hoan	Yup'ik (Central)	56	14.3
Maybrat	Tamil	55	14.5
Nubian (Dongolese)	Acehnese	48	14.6
Swedish	Koasati	47	14.9
Klamath	Wari'	47	14.9
Kongo	Ladakhi	46	15.2
Bashkir	Māori	46	15.2
Berber (Middle Atlas)	Waorani	45	15.6
Lango	Archi	45	15.6
Archi	Thai	45	15.6
Thai	Retuarā	45	15.6
Ijo (Kolokuma)	Kutenai	50	16.0
Kongo	Evenki	56	16.1
Arabic (Egyptian)	Tümpisa Shoshone	48	16.7

allows for a very small margin of error, it can be predicted that less than 40% shared feature values implies unrelatedness. Only 41 out of the 12,034 pairs that have 40% or less shared feature values fail to meet the prediction (the actual pairs are listed in Section 4 below). Thus, lack of similarity is a good predictor of unrelatedness, but presence of similarity is a bad predictor of relatedness.

### 3 Are there ways of improving the results?

We next consider the question of whether the prediction of relatedness could be improved somehow. In other studies (Wichmann et al. *forthc.*; Holman et al. 2007) we have made exact quantitative explorations of the relationship between typological similarity and geographical distance among languages. Not surprisingly, the greater the geographical proximity is between languages, the

Table 3. The 20 most similar language pairs in the sample

Language A	Language B	Relatedness	Features compared	% similarity
Lango	Luo	same genus	46	80.4
Luvale	Zulu	same genus	97	80.4
Khmer	Vietnamese	same family, diff. genera	89	80.9
Vietnamese	Thai	diff. families	110	80.9
Khalkha	Tuvan	same family, diff. genera	48	81.3
Lithuanian	Russian	same family, diff. genera	64	81.3
Greek (Modern)	Bulgarian	same family, diff. genera	64	81.3
Khmer	Thai	diff. families	91	81.3
Polish	Russian	same genus	71	81.7
Russian	Serbian-Croatian	same genus	45	82.2
Swahili	Zulu	same genus	107	82.2
Dagur	Turkish	same family, diff. genera	46	82.6
Telugu	Kannada	same family, diff. genera	47	83.0
Kongo	Nkore-Kiga	same genus	48	83.3
Dutch	German	same genus	56	83.9
Italian	Spanish	same genus	63	84.1
Drehu	Iaai	same genus	46	84.8
English	Swedish	same genus	60	85.0
French	Italian	same genus	64	85.9
Hindi	Panjabi	same genus	49	91.8

more similar they tend to be (this goes for both related and unrelated languages). If one takes into account the areal factor, this might move the cut-off point to allow more accurate predictions of relatedness. To test this strategy, we found the average similarity between pairs of unrelated languages as a function of the geographical distance between them; we then adjusted the similarity between each pair of languages by dividing the similarity by the average for unrelated languages the same distance apart. The correlation between adjusted and unadjusted similarities was 0.96. The reason for this is probably that the distance measure, as given in the WALS database, identifies

the location of a given language (roughly) with its center of extension. This means that some neighbouring languages, such as German and Dutch, are treated as having a certain geographical distance between them when in reality they don't have any. The more widespread the languages compared are, the bigger this problem gets. Since it is impossible to provide adequate measure of geographical distances for 29,810 language pairs, and not just take recourse to a mechanical measure of distance from one WALS dot to another, it is not viable to improve on the cut-off point in such a way.

Also, the 134 features differ appreciably in the distribution of rarity and commonness among their values. Taking into account the relative rarity of feature values might improve the predictions. To test this strategy, we found the average similarity between pairs of unrelated languages separately for each feature; we then adjusted the similarity between each pair of languages by dividing the similarity by the average for unrelated languages across the same features attested in the given pair. The correlation between adjusted and unadjusted similarities was 0.98. The probable reason is that *rara*, at least as concerns rare WALS feature values, are as prone to diffusion as are more frequent feature values.

Another strategy to improve the results would be to take into account the areality of features. The linguistic typological literature abounds with statements concerning the susceptibility to diffusion of certain features as opposed to others. Wichmann and Holman (2009), however, show that “areality” is not amenable to quantification in any straightforward way since the diffusibility of features varies in different parts of the world.<sup>1</sup> Thus, this strategy is not viable.

A final strategy to try to improve the power of prediction concerning relatedness would be to weight different features or values of features according to their stability. We have explored ways of measuring stability and have established a ranked order of stability for WALS features (Wichmann and Holman, 2009, cf. also Holman et al. 2007: 417–418 for a summary of the method). Conceivably, if the features shared among languages were weighted for their stability the cut-off point could be pushed. On the other hand, Holman et al. (2008: 345–346) report on results suggesting that such a weighting would have little if any effect. They show (in their Figure 4) that correlations between typological distances among languages and distances as defined in traditional classifications are similar whether one uses all 134 WALS features, only the 85 most stable ones or anything in between; and using less than 85 features has a negative effect on the correlations. If excluding unstable

features does nothing to improve correlations between typological distances and expert classifications, then a weighting scheme cannot be expected to improve the predictive power concerning relatedness between two languages held by their amount of similarity. The ultimate reason for this has already been stated in the previous paragraph: the predictive power of the similarity measure is upset by diffusibility, and, unlike stability, diffusibility is not something inherent in features — stable features are as likely to diffuse as unstable ones.

#### **4 Deviant language pairs**

The results reported on in Figure 1 and Table 1 show that there are a few pairs of languages which are related even though showing less than 40% similarities, which is the point where pairs tend overwhelmingly not to be related. It serves the record to provide a list of the pairs of related languages that are deviant in the sense that they show less similarity than related languages normally do. This list is provided in Table 4 on the next page.

There are three general explanations for the small amount of similarities among members of the language pairs in Table 4 in addition to possible explanations of lesser generality and therefore lesser interest.

One explanation is that the languages in question tend to belong to very large families where there is more room for variation to arise. It is hardly a coincidence that all language pairs, with the exception of the two Penutian ones, belong to the 10 largest families in the world (counted in terms of numbers of languages according to Gordon 2005).

A second factor is genealogical separation. All of the language pairs belong to different genera according to the classification in WALS. We have also looked at the more detailed genealogical partitions of Gordon (2005); the column in Table 4 headed “Eth. level” indicates the level of closeness according to this classification. A “1” means that the languages are separated by the root of the family tree, i. e., that they are maximally genealogically distinct; a “2” means that they are separated by a node which is one step down from the root; and so on. 71% of the pairs are in maximally distinct subgroups (“1”s), another 27% are in maximally distinct subgroups within one and the same highest coordinate branch (“2”s), and only one pair belongs to a more deeply embedded subgroup, namely Zulu and Yoruba. This particular case, however, is taken care of by the third general explanation, which is geographical separation.

Table 4. Related languages that have unusually different typological profiles (less than 40% similarities)

Language A	Language B	Language family	Features compared	% sim.	Eth. level	Dist. (km)
Luvale	Ijo (Kolokuma)	Niger-Congo	52	28.8	2	2,606
Zulu	Ijo (Kolokuma)	Niger-Congo	52	28.8	2	4,666
Maidu (Northeast)	Tsimshian (Coast)	Penutian	48	29.2	1	1,527
Ngiti	Koyra Chiini	Nilo-Saharan	47	29.8	1	4,030
Yoruba	Ijo (Kolokuma)	Niger-Congo	51	31.4	2	373
Mundari	Semelai	Austro-Asiatic	66	31.8	1	2,972
Swahili	Ijo (Kolokuma)	Niger-Congo	50	32.0	2	3,909
Maung	Yidiny	Australian	81	32.1	1	1,433
Mundari	Khmer	Austro-Asiatic	78	32.1	1	2,443
Koyraboro Senni	Murle	Nilo-Saharan	65	32.3	1	3,794
Koromfe	Ijo (Kolokuma)	Niger-Congo	49	32.7	2	1,263
Beja	Margi	Afro-Asiatic	45	33.3	1	2,591
Sango	Ijo (Kolokuma)	Niger-Congo	51	33.3	2	1,365
Nandi	Koyraboro Senni	Nilo-Saharan	47	34.0	1	4,217
Nandi	Koyra Chiini	Nilo-Saharan	52	34.6	1	4,555
Marathi	Spanish	Indo-European	52	34.6	1	7,826
Margi	Amharic	Afro-Asiatic	49	34.7	1	2,733
Mundari	Vietnamese	Austro-Asiatic	88	35.2	1	2,701
Garó	Cantonese	Sino-Tibetan	51	35.3	1	2,295
Berber (Middle Atlas)	Kera	Afro-Asiatic	65	35.4	1	3,294
Irish	Marathi	Indo-European	45	35.6	1	7,931
Paamese	Acehnese	Austronesian	45	35.6	2	8,355
Limbu	Mandarin	Sino-Tibetan	45	35.6	1	2,258
Mandarin	Bawm	Sino-Tibetan	76	36.8	1	2,151
Ijo (Kolokuma)	Diola-Fogny	Niger-Congo	46	37.0	2	2,540
Ngiti	Nubian (Dongolese)	Nilo-Saharan	54	37.0	1	1,881
Miwok (S. Sierra)	Tsimshian (Coast)	Penutian	62	37.1	1	1,807
Mundari	Khmu'	Austro-Asiatic	70	37.1	1	1,798
Bagirmi	Nubian (Dongolese)	Nilo-Saharan	64	37.5	1	1,743
Beja	Hausa	Afro-Asiatic	82	37.8	1	3,180
Koromfe	Kisi	Niger-Congo	45	37.8	2	1,173
Yidiny	Tiwi	Australian	90	37.8	1	1,701
Limbu	Meithei	Sino-Tibetan	45	37.8	2	677
Kera	Amharic	Afro-Asiatic	50	38.0	1	2,509
Zulu	Yoruba	Niger-Congo	104	38.5	4	5,035
Beja	Kera	Afro-Asiatic	57	38.6	1	2,430
Ngiyambaa	Maranungku	Australian	74	39.2	1	2,555
Malagasy	Acehnese	Austronesian	56	39.3	2	6,007
Ngiti	Nandi	Nilo-Saharan	48	39.6	1	541
Lugbara	Lango	Nilo-Saharan	53	39.6	1	252
Fur	Ngiti	Nilo-Saharan	58	39.7	1	1,471

Holman et al. (2007: Figure 1) show that beyond around 4,000km there is no effect of geographical proximity on the similarity among related languages, suggesting that this is the limit of diffusion (including diffusion operating chainwise). The average distance among the language pairs in Table 4 is as high as 2.892 km, and Zulu and Yoruba have a particularly great distance (5,035 km).

In order to single out cases that need special explanations we apply the following strategy. A special case is defined as a language pair that does not satisfy at least two of the following three criteria:

- 1) the pair belongs to one of the world's 10 largest families;
- 2) the pair is genealogically separated by the root of the family tree;
- 3) the pair is separated by more than 4,000 km.

The special cases that fall out are: Luvale–Ijo (Kolokuma), Maidu (North-east)–Tsimshian (Coast), Yoruba–Ijo (Kolokuma), Swahili–Ijo (Kolokuma), Koromfe–Ijo (Kolokuma), Sango–Ijo (Kolokuma), Ijo (Kolokuma)–Diola-Fogny, Miwok–Tsimshian (Coast), Koromfe–Kisi, Limbu–Meithei. We immediately notice that Ijo (Kolokuma) appears in six of the pairs. This language belongs to Ijoid, the typologically most divergent branch of Niger-Congo, among other things characterized by the absence of the otherwise characteristic noun class system. We do not have an explanation for the special behavior of Ijo (Kolokuma), but note that it does not really come as a surprise given that it belongs to Ijoid. Two Penutian pairs figure in the list, both including Tsimshian (Coast). Given that Tsimshian (Coast) has never been demonstrated to belong to Penutian to the satisfaction of all experts we may be facing a case where the languages in question are actually not related at all. As for Koromfe–Kisi and Limbu–Meithei we shall not venture to offer any explanation.

We may summarize the results of this section as follows. If two languages have 40% or less features in common we can be almost certain that they are not generally considered to be related. Barring a few special cases, only languages that conform to at least two of the following three criteria fail to meet the prediction: they belong to one of the world's 10 largest families, they are maximally separated genealogically within their given family, and they are geographically very remote (>4,000km) from one another.

## 5 Conclusions

The results reported on in this note were, in part, unsurprising and, in part, unexpected. Figure 1 showed a close correlation between relatedness and typological similarity. This is what we had expected. But we also expected to find some minimal amount of typological similarity among language pairs which would suffice to predict that two languages are related. It turned out to be the case, however, that the amount of similarity required to make this prediction is so high (81.5%) that only few language pairs qualify. In practice, this means that typological features such as those of WALS are not useful for identifying relatedness among languages when it comes to comparisons of single pairs. When groups of languages are compared the situation may be different, but this issue is beyond the scope of this paper (cf. the thread of discussion in Dunn et al. 2005, Donohue and Musgrave 2007, Dunn et al. 2007, and Donohue, Wichmann, and Albu 2008 for an empirical example of the difficulties arising from establishing genealogical relations on the basis of abstract typological features). At the other end of the scale we found that typological *dissimilarity* is a good predictor of *unrelatedness*: with only a small margin of error one can predict that languages which have 40% or less similarity are not related according to the WALS classification. Our finding that a certain amount of typological differences can be used to predict that languages are not commonly believed to be related means that typological differences are a yardstick for gauging the limits of the traditional comparative method.

Returning to the issue of *rara*, raised in the introduction to this paper, let us recall the hypothesis we stated, namely that relatedness and propinquity are the two major, systematic contributors to making languages similar. The reverse of this hypothesis is that unrelatedness and geographical separation are the major contributors to making languages *dissimilar*, and therefore to promote the appearance of rare features or “exotic” language profiles. We may now consider how each of the two factors contribute to languages being dissimilar. The fact that nearly all language pairs that have 40% or less similarities are unrelated can be generalized to a statement that unrelatedness is a necessary condition for languages to be perceived as being mutually exotic. But for languages to be maximally different from one another this condition is not always sufficient — to ensure this, they must normally also be geographically distinct. To judge from Holman et al. (2007: Figure 1), there is an overall tendency for unrelated languages to increase their similarities

by an average of some 12% when one moves from those that are maximally separate (i. e., by 8,000 km or more) to those that are in mutual proximity. We would claim, then, that unrelatedness probably is the primary reason for the appearance of rarity in language and that geographical separation probably is the secondary reason. At the end of the day, however, these two factors will be difficult to tease apart since what we know about genealogical relations among the world's languages is severely limited by the historical-comparative methods at our disposal.

A host of other reasons would be necessary to explain why a particular language, family or language area ends up as a candidate for being particularly exotic, including universal tendencies and chance, as well as factors extrinsic to language evolution as such, i. e. issues of language sampling, choice and formulation of features, as well as variability in data attestation. Since we do not deal with such issues in the present paper our contribution is limited. Nevertheless, we believe that the issues we have treated are important and fundamental to the study of linguistic diversity and, by implication, to that of rarity in language.

### Acknowledgments

We would like to thank Michael Cysouw, Bernard Comrie, Cecil Brown, and Dietrich Stauffer for comments on this manuscript.

### Notes

1. In order to prevent misunderstandings let us stress that by claiming that it is not possible to quantify a certain "areality" (= diffusibility) of features we do not mean to imply that it is not possible to define linguistic areas by the occurrence of certain shared features using a quantitative approach (cf. Bickel and Nichols, in press).

### References

- Bickel, Balthasar, and Johanna Nichols  
In press Oceania, the Pacific Rim, and the theory of linguistic areas. In: *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society*.
- Brown, Cecil H., Eric W. Holman, Christian Schulze, Dietrich Stauffer, and Søren Wichmann  
2006 Are similarities among languages of the Americas due to diffusion or inheritance? An exploration of the WALS evidence. Paper presented at the

- conference “Genes and Languages”, University of California Santa Barbara, September 8–10, 2006.
- Cysouw, Michael  
 In press Quantitative explorations of the world-wide distribution of rare characteristics, or: the exceptionality of north-western European languages. In *Exception in Language*, Horst Simon and Heike Wiese (eds.). Berlin/New York: Mouton de Gruyter.
- Donohue, Mark and Simon Musgrave  
 2007 Typology and the linguistic macrohistory of Island Melanesia. *Oceanic Linguistics* 46 (2): 325–364.
- Donohue, Mark, Søren Wichmann, and Mihai Albu  
 2008 Typology, areality and diffusion. *Oceanic Linguistics* 47 (1): 223–232.
- Dunn, Michael, Robert Foley, Stephen Levinson, Ger Reesink, and Angela Terrill  
 2007 Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46 (2): 388–403.
- Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson  
 2005 Structural phylogenetics and the reconstruction of ancient language history. *Science* 309: 2072–2075.
- Dryer, Matthew S.  
 2005 Genealogical language list. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 584–643. Oxford: Oxford University Press.
- Gordon, Raymond G., Jr. (ed.)  
 2005 *Ethnologue: Languages of the World*. 15th ed. Dallas: SIL International. Online version: <http://www.ethnologue.com/>.
- Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.)  
 2005 *The World Atlas of Language Structures*. Oxford: Oxford University Press.
- Holman, Eric W., Dietrich Stauffer, Christian Schulze, and Søren Wichmann  
 2007 On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguistic Typology* 11 (2): 395–423.
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker  
 2008 Explorations in automated language classification. *Folia Linguistica* 42 (2): 331–354.
- Wichmann, Søren and Eric W. Holman  
 2009 *Temporal Stability of Linguistic Typological Features*. München: LINCOM Europa.
- Wichmann, Søren, Eric W. Holman, Dietrich Stauffer, and Cecil H. Brown  
 forthc. Similarities among languages of the Americas: An exploration of the WAL evidence. to appear in a volume edited by Bernard Comrie.