Michael Cysouw

Disentangling geography from genealogy

1. Introduction

There are various reasons why two languages might be typologically similar. Basically, there are four different possible causes of similarity that can be classified into two groups of two, as shown in (1).

(1) Languages can be typologically similar because of:
   A. Historical factors, being either:
      i. Genealogical descent ("vertical transfer")
      ii. Borrowing ("horizontal transfer")
   B. A-historical factors, being either:
      i. Inherent characteristics of human language ("universals")
      ii. Coincidence ("chance")

One of the major challenges for current typological-comparative linguistics is to find methods to disentangle which of these reasons apply in any given situation of typological similarity. Traditionally, there has been great interest in developing methods to separate historical factors from a-historical factors through various kinds of sampling. The basic idea behind such methods is that it is possible to control for the influence of historical factors by carefully selecting languages across known genealogical and areal groupings. The remaining question then is how to distinguish universals from chance. Both questions – how to remove historical influences from the data, and how to subsequently distinguish universals of human language from chance effects – are much-debated questions in the field of linguistic typology (cf. Cysouw 2005 for a survey), and much more could be said about these topics. Yet, the current paper will not concentrate on such a-historical characteristics of human language, but focus on historical factors that result in typological similarity.

When investigating historical factors leading to typological similarity, it might seem as if there are likewise two questions to be dealt with. First, how to factor out a-historical factors, and, second, how to distinguish similarities caused by genealogical descent from similarities caused by borrowing. However, on closer inspection the first question turns out to be easily answerable.

A-historical factors should apply equally to all languages, so their effects should be statistically equal for all languages. The influence of universals and chance thus amounts to a constant factor in the diversity of languages, which can simply be ignored when investigating historical factors. Still, there is a potential problem when including many different characteristics in the comparison (as will be the case in this paper). Ideally, to investigate historical factors throughout many characteristics, these characteristics should be independent of each other. A group of, say, ten characteristics that are all definitionally similar in a collection of, say, hundred characteristics in total will introduce an a-historical bias into the comparison, favoring the linguistic similarities as found in the cluster of those ten definitionally similar characteristics. The most glaring of such dependencies will have to be removed (e.g. by weighting the characteristics). The remaining question, which will be the main topic of this paper, is how to disentangle typological similarities caused by genealogical descent from similarities caused by borrowing.

In this paper I will not seek to settle this question for individual cases of shared characteristics between two specific languages (e.g. why do French and German have no distance contrast in demonstratives?; see Diessel 2008; Cysouw 2011), because individual historical developments cannot be predicted by a general theory of human language. Specific historical events can only be reconstructed by an in-depth investigation of the actual history of a specific situation. However, I propose that the influence of borrowing vis-à-vis genealogical descent can be investigated in the aggregate (cf. Nerbonne and Siedle 2005; Nerbonne 2009 on the notion "aggregate").

To investigate the relationship between typological structure, genealogical descent, and borrowing, I will use data from the *World Atlas of Language Structures* (WALS, Haspelmath et al. 2005). This resource provides information about typological structure and genealogical descent, but not about possible contact or the probability of borrowing. To approach the probability of borrowing, I will use the present-day geographical distribution of languages, assuming that the probability of borrowing is inversely correlated with geographical distance. Specifically, geographically close languages will have a higher probability of contact, and likewise a higher probability of borrowing.

2. The Eurasian data set

As a concrete example to discuss the current proposals I will use a dataset, drawn from WALS, featuring a selection of Eurasian languages. The Eurasian macro-area is chosen because most languages and their approximate geographical location will be familiar to most readers. WALS includes data

on 391 different Eurasian languages, but far from all are covered in any detail (missing data is a general problem when dealing with WALS quantitatively; cf. Cysouw 2008). To obtain a data set with sufficient coverage, only languages that appear in at least 70 WALS maps are included here, resulting in a sample of 32 Eurasian languages (see Appendix A).

Given the data from WALS, I will define a notion of pairwise structural similarity. There are various different aspects that can be included in such a definition of similarity, only a selection of which I will be using here. Also note that in practice I will define a notion of distance, which is of course just a trivial transformation of any notion of similarity. To define a notion of (dis)similarity, the following principles can be used (see Appendix B for the details):

- Basically, the distance between two languages is related to the number of characteristics that are different between the two languages.
- Because of the many missing data points, this value has to be normalized to the number of comparisons made for each language pairing, i.e. the sum of the number of similarities and the number of differences (cf. the "relativer Identitätswert" [RIW] presented in Goebl 1984).
- Weightings can be used to balance the impact of the characteristics, making some characteristics more important than others. This mechanism will be used here to remove some of the most glaring definitional redundancies in the WALS data.
- Further, similarities are not necessarily all equal. Languages that share a rare characteristic can be seen as more similar than two languages that share a common characteristic (because the sharing of a rare characteristic is a more telling similarity, cf. the "gewichteter Identitätswert" [GIW] in Goebl 1984).
- Finally, differences are not necessarily all equally different. For example, a language A with a small vowel inventory is less different from a language B with an average vowel inventory than from a language C with a large vowel inventory. Both the pairing (A, B) and (A, C) are different, but to different degrees. Such specification of internal structure of WALS characteristics will not be used here, because it is far from obvious how such specification of differences should be determined, and how they should be combined with specifications of similarities. Therefore, exploring this issue is reserved for another occasion.

For this paper I will start with the basic fraction of the number of differences divided by the number of comparisons made as a measure of dissimilarity be-

tween two languages (this yields Goebl's "relativer Identitätswert"). In addition, each feature (i.e. each 'map' in WALS) can be weighted to remove definitional redundancies in the WALS data. This is necessary because, on closer inspection, the features included in WALS are not independent of each other. The relationship between the various features in WALS turns out to be a highly complex topic, with various overt and covert dependencies between them (cf. Cysouw 2008). For the sake of this paper, I adopted the following solution to remove the most glaring redundancies. I grouped the features into sets of (definitionally) related ones (see Appendix B), and every feature in such a set is weighted by the inverse of the number of features in the set. For example, the WALS features 30, 31, 32, and 44 all deal with gender marking, and without correction, two languages without gender marking will be counted as being similar four times on all four features, though it is the same underlying similarity that is counted each time. To correct for this implicit weighting, each of the features will be explicitly weighted as counting only 1/4th (because there are four features in the set). Such weighting could also be used to emphasize typologically stable features.

Further, similarities between two languages can be weighted. Following Goebl's (1984) basic insight that sharing rare features is more telling than sharing common features, a weighting of similarities can be introduced. Such a weighting is specified for each value in each feature. So, for example, there is a difference when two languages both have tone, and when they both lack tone. Goebl proposed to weight each similarity by the fraction of occurrences in the sample. For example, in WALS there are 307 out of 527 languages that do not have tone (Maddieson 2005b). For two languages that both do not have tone, instead of counting one similarity, Goebl proposed to count $0.417 = 1 - (307/527)$ similarities. Languages that share a complex tone system, which is much rarer, are assigned a higher similarity of $0.833 = 1 - (88/527)$. Another variant of this principle would be to interpret the frequency of a characteristic as typological information, and the rarer a characteristic, the more informative it is. From an information-theoretic perspective one would calculate the similarity for not having tone as $0.235 = -\log(307/527)$ and for having complex tone as $0.777 = -\log(88/527)$. In general, the weights based on the logarithm will give similar, but slightly more extreme weights compared to the weights based on the fraction, especially for characteristics that occur in less than 15 % of all languages in WALS.

Based on these different possible notions of dissimilarity, various distance matrices were compiled for the 32 Eurasian languages selected from WALS. These matrices represent different ways to define pairwise aggregate dissimilarities. It turns out that the various ways to define typological dissimi-

larity only differ in detail, and all matrices are strongly correlated (Pearson's *r* ranges between 0.89 and 0.99; see Appendix B). Given any such notion of typological dissimilarity, the main question addressed in this paper is whether it is possible to say something about how much of this dissimilarity is caused by genealogical descent, and how much by geographical proximity.

3. Genealogy

Two languages that are both descendants from one and the same proto-language will share typological characteristics that have not changed since they split from their last common predecessor. Given that changes will accumulate over time, it is to be expected that closely related languages share more similarities than languages that separated earlier. Such a trend is clearly visible in typological data, and this observation has even led to the proposal that typological profiles might be used for the reconstruction of historical descent (cf. Nichols 1992; Dunn et al. 2005). Such an approach is of course only viable when the influence of genealogical descent on the typological profile is stronger than the influence of any subsequent areal convergence.

The impact of genealogical descent is also clearly visible in the current data selection, as shown in Figure 1. Taking the genealogical classifications as specified in WALS, I have separated all language pairings into three groups, being either (1) of the same genus, (2) different genus within the same family, or (3) non-related given currently accepted views (see Appendix A for a survey of the genealogical classification of the current selection of languages). The same-genus-pairings involve only Germanic, Romance and Slavic languages. Pairs of languages from the same family, but not from the same genus include pairings from Altaic (which does not include Korean and Japanese according to WALS), Indo-European, Nakh-Dagestanian and Uralic.

To quantitatively assess the strength of the correlation between typological distance and genealogical distance, I performed Mantel tests (Mantel 1967) to correlate the various typological distance measures with genealogical distance. Genealogical distance was simply defined as an approximate linear scale, being '1' when two languages were from the same genus, as '2' when they came from the same family, but from different genera, and as '3' otherwise (note that these numbers are of course ranks, but I do not know of a Mantel test that can deal with ranks). All Mantel tests for the different kinds of typological distance were highly significant, with only slight differences in the statistics, as shown in Table 1. Logarithmic value weightings (to favor rare similarities) combined with the feature weightings (to reduce the impact

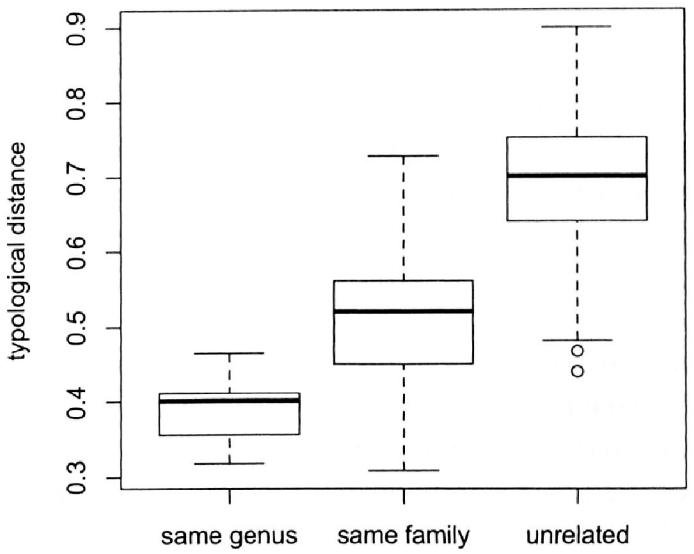

Fig. 1. Languages from the same genus are on average more similar typologically than languages from different genera, but from the same family, which are in turn more similar than unrelated languages.

of groups of definitely similar features) yielded the highest correlation scores. For the remainder of this paper I will restrict attention to this typological distance measurement, as it seems to be the one most closely matching genealogical descent.

	No feature weighting	Feature weighting
No value weighting	0.574	0.603
Value weighting by fraction	0.634	0.616
Value weighting by logarithm	0.616	0.642

Table 1. Correlations between genealogical distance and different definitions of typological distance. All correlations are highly significant at $p < 0.001$ (according to a Mantel test), though there are slight differences in the strength of the correlation scores.

4. Geographical proximity

In line with Tobler's first law of geography that "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970: 237), typological (dis)similarities are strongly related to geographical distance. Geographically close languages are in general typologically similar, while geographically distant languages are generally typological different.

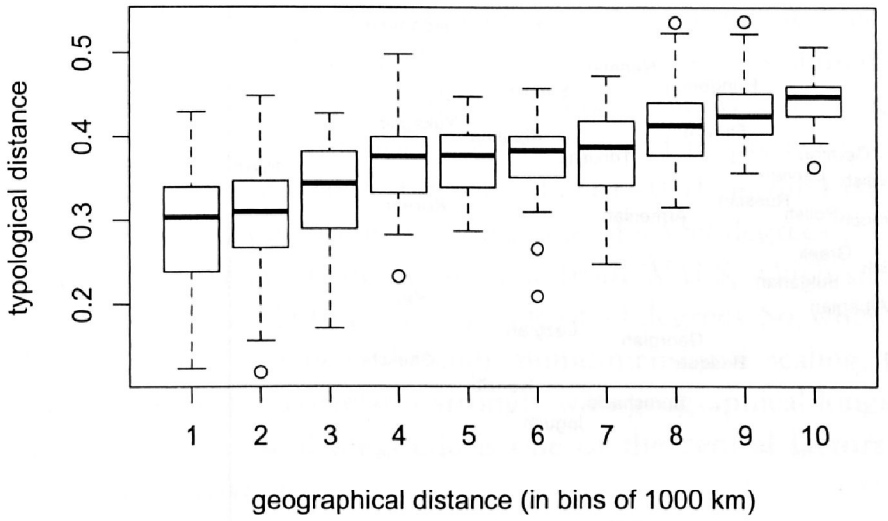

Fig. 2. There is a strong correlation between the geographical distance between two languages and their typological dissimilarity. Every observation in this figure is a language pairing, shown here as a box plot grouping language pairings depending on their geographical distance in bins of 1000 kilometers.

This correlation is immediately obvious for the current selection of 32 Eurasian languages. Shown in Figure 2 are all 496 language pairings (i.e. $32 \times 31/2$ pairings) grouped by geographical distance. Note that as a measure of geographical distance I have here simply taken the linear distance on a perfect sphere between two coordinates on the surface (see Appendix A for the coordinates used). To aid visual interpretation, all language pairings are grouped into ›bins‹ of thousand kilometers, i.e. all language pairings with a distance between 0 and 1000 form one group, and all language with a distance between 1000 and 2000 form another group, and so on. For each of these groups a box is shown in Figure 2, the medians of which show almost a linear relationship between geographic distance and typological distance, surprisingly in this case without any sign of the expected flattening at extreme geographical distances (cf. Nerbonne, this volume, on an in-depth discussion of the nature of this relationship). Using a Mantel test to evaluate this correlation gives again a highly significant result ($r = 0.616, p < 0.001$).

A visually more impressive way to show the strength of the correlation between geographical distance and typological distance is shown in Figure 3. This figure plots the first two dimensions of a non-metric multidimensional scaling (MDS) of the typological distances. Multidimensional scaling is a method to mathematically derive abstract dimensions of variation from a matrix of distances. The first dimension is defined such that as much of the

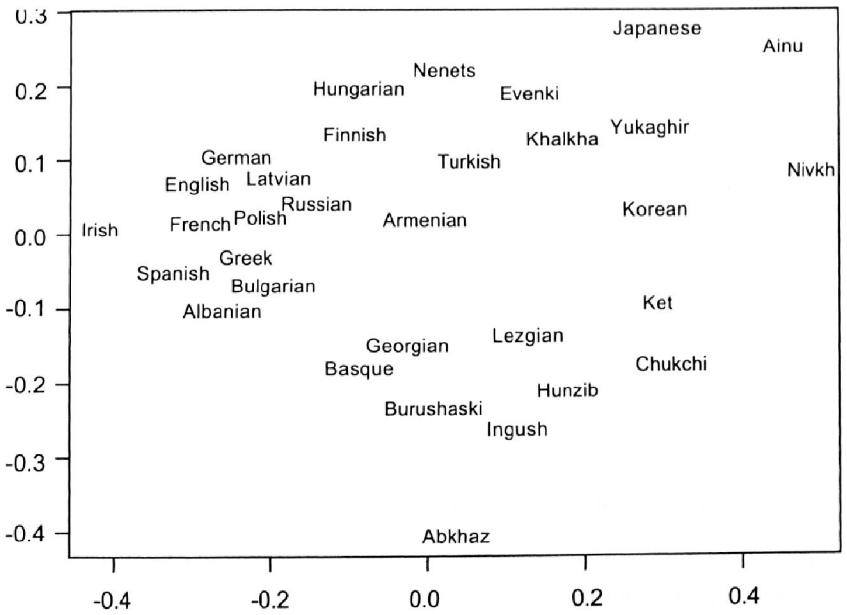

Fig. 3. The first two dimensions of non-metric multidimensional scaling of language distances based on WALS data for 32 Eurasian languages (the precise locations of the names in the MDS has been slightly tweaked manually to prevent overlapping). The first dimension (shown horizontally) strongly correlates with geographical longitude. The second dimension (shown vertically) separates the Caucasian languages with Burushaski, Basque and Chukchi from the rest.

variation as possible is captured. Subsequent dimensions account for the leftover variation in diminishing order. In the current example the first dimension accounts for 67 % of the variance. In Figure 3 this first dimension is shown horizontally and shows an astonishing overlap with geographical longitude. The west European languages are shown to the left, the far Eastern and Siberian language (except Chukchi) to the right. The second dimension, shown vertically, captures another 12 % of the variance. This dimension basically separates the outliers from the Eurasian mainstream, namely the Caucasian languages together with Burushaski, Basque and Chukchi.

The central result of the MDS analysis, though, is that the most important dimension (the first dimension) is almost perfectly related to geographical longitude. Correlating this first dimension (i.e. the left to right order of the languages in Figure 3) with longitude reaches extremely high significance ($r = 0.88$, $p < 0.001$). This implies that the longitudinal location of the current set of languages can be predicted from the value on the first dimension of the MDS, which is in itself just a derivative of typological similarity. It is thus possible to predict typology from geography (or vice versa). A linear in-

terpolation of the first MDS dimension with longitude results in the following formula: Longitude (in degrees) = 57.8 + 199 × dimension 1 of the MDS. The constant 57.8 in this formula is the average longitude of the sampled languages. This formula can be used to predict geographical location. For example, Spanish has a value of –0.31 on the first MDS dimension, so the formula predicts a longitude of –3.89 degrees (= 57.8 – 199 × 0.31). This perfectly matches the data from WALS, which situates Spanish geographically in Madrid at a longitude of –4 degrees. So, when investigating the typological distances through multidimensional scaling, the most important dimension correlates strongly with geographical longitude, indicating that geographical longitude is one of the central factors determining typological variation.

5. Separating genealogy from geography

Typological distance between languages is strongly correlated with genealogical distance, but also with geographical distance. Further, genealogically related languages are in general located geographically close to each other, so one would also expect a correlation between genealogy and geography. And indeed, for the current test case this correlation is statistically significant ($r = 0.367$, $p < 0.001$), though the correlation is clearly less strong than the correlations discussed previously for genealogy and geography. Yet, we are left with a tangle of significant correlations between typology and geography, typology and genealogy, and geography and genealogy. There are various statistical approaches that might help to tear apart the interaction between these factors.

In general, the issue could be approached using basic regression modeling, were it not for the fact that we are dealing with distances here. Different from the normal situation in which one would like to use regression modeling, the 'observations' here are pairwise measures of distance for which it is not clear whether they can simply be treated as 'atomic' observations. The problem is that dissimilarity measures are not independent observations, because each language is compared with all others, so duplication of information is introduced. However, we can of course just pretend that the dissimilarities are real observations and perform a linear regression. The results are shown in Table 2. All factors turn out to be highly significant, so it still remains unclear which factor is more important for the explanation of typological distances (and the significance values are probably meaningless anyway, for the reasons discussed above).

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	0.3481	$1.505 \cdot 10^{-2}$	23.14	< 0.001 ***
Genealogy	0.1316	$8.594 \cdot 10^{-3}$	15.31	< 0.001 ***
Geography	$1.822 \cdot 10^{-5}$	$1.299 \cdot 10^{-6}$	14.03	< 0.001 ***

Table 2. Linear regression of typological distances to genealogical and geographical distance ($r^2 = 0.58$). All factors are highly significant, but this is probably meaningless because the 'observations' are distances.

Assessing significance of correlations between distances is normally done using the Mantel test (as we did in the previous sections). There is a variant of the Mantel test that can assess the significance of the correlation between two distances matrices while keeping a third distance matrix constant. Such a test is called a partial Mantel test (Legendre 2000). This approach seems to be ideal to address the current problem, but unfortunately it does not decide on the question which of the correlations is most important. The correlation between typology with genealogy, while keeping geography constant, is significant (partial Mantel $r = 0.568$, $p < 0.001$). However, the correlation between typology with geography, while keeping genealogy constant, is similarly significant (partial Mantel $r = 0.534$, $p < 0.001$).

There are other recent developments in statistical methods to deal with this problem. Specifically, there is a proposal for multivariate regression modeling with distance matrices as dependent variable and factors or continuous variables as independent variables (Zapala and Schork 2006). In this approach, we can use the typological distances as dependent variable, but we have to use separate categorical predictor variables for family and genus, and continuous predictor variables for longitude and latitude. From the results, as shown in Table 3, it seems that the genealogical factors are the strongest factors and the only significant ones. The results thus indicate that the basic correlation is between typology and genealogy, and that the correlation with geography is only a secondary effect. However, note that the results differ rather radically when the order of the predictor variables is changed, which casts doubt on the proper interpretation of these results.

	Df	F Model	R2	Pr(>F)
Family	14	4.8814	0.69641	0.001 ***
Genus	13	1.9653	0.26035	0.011 *
Latitude	1	1.2761	0.01300	0.237
Longitude	1	0.9670	0.00985	0.512
Residuals	2	0.02038		
Total	31			

Table 3. Regression model using typological distance as dependent variable. Genealogical factors (Family and Genus) are the only significant factors. Geographical factors (Latitude and Longitude) are not significant.

Finally, it is highly informative to look at the typological residuals after regression with genealogy or geography. I will use here statistics calculated in a similar fashion as the linear regression reported in Table 2, though I shall ignore the significance values. The basic idea is to remove the impact of genealogical relatedness from the typological distances, and look at the residuals of the typological distances (and do so likewise for geographical distances). The interpretation of such residuals seems to be linguistically interesting. First, if there is a correlation between genealogical distance and typological distance, then it should be interesting to spot language pairings that are more similar typologically than expected from genealogy. Such excess similarity might be indicative of areal convergence. Second, given that there is a correlation between geographical distance and typological distance, I will look for language pairings that are more similar than expected given their geographical distance. Such language pairings – which are, in a sense, 'too far away' for their typological distance – might be indicative of (relatively) recent population movement.

So I linearly regressed typological distance against genealogical distance, and then ordered all language pairings according to their residuals. The languages pairings with the lowest residual typological distance are all geographically close, and many pairings indeed seem to be readily interpretable as cases of language contact. High on the list are the pairings Korean-Japanese, Khalkha-Japanese and Khalkha-Korean. These are languages that are known to have been in close contact over centuries, up to the point that they are sometimes claimed to be genealogically related. Also, the classic European Sprachbünde are represented in the top of the list: the Baltic Sprachbund (Russian-Finnish, Latvian-Finnish), the Balkan Sprachbund (Bulgarian-Greek, Albanian-Greek), and the Charlemagne Sprachbund (German-French). Also on top of the list are the pairings Armenian-Georgian and Armenian-Turkish, which

are also clear examples of language contact. The remaining top pairings are less clear: Burushaski-Georgian, Burushaski-Lezgian and Nenets-Evenki. Whether the observed surplus of typological similarity in these cases is the result of contact, or caused by other factors, is unclear to me.

The waters are muddier when we attempt to remove geographical influence from typological distance, but there still is some indication of an influence of population movement. When looking at the residuals of typological distance after regression by geographical distance, the language pairing with the lowest residual typological distance is Khalkha-Turkish, with Evenki-Turkish following a bit further down. Turkish is clearly an example of a language subject to a relatively recent population movement over a long distance, leading to a situation in which the Turkish language is still relatively similar typologically to its Altaic kin, though geographically it is too distant. The other language pairings in the top twenty of 'too similar' languages relative to their geographical distance are all pairings of Indo-European languages.

5. Conclusion

Investigating the typological diversity of the world's languages has been an active field of research over the last few decades. However, the basic premise has been that this kind of research is worthwhile because it will help unravel universal properties of human language. The strong correlations with historical factors, both genealogical and geographical, as discussed in this paper cast doubt on the allegedly important role of universal properties on the currently observable typological diversity. The entanglement between typological diversity and genealogical relationship has been acknowledged for a long time in the literature. In contrast, the similarly intricate entanglement between typological diversity and geographical proximity has not sparked similarly in-depth investigations.

The existence of correlations between genealogy and geography should not merely be seen as a nuisance factor in the investigation of universal properties of human languages – it can also be taken as a possible starting point to unravel the dynamics of typological change and language history. Instead of building samples that from the start prevent genealogical or geographical bias, I think we should deliberately collect data from samples with such 'biases'. Only by including many related languages and/or geographically close languages will it be possible to investigate the impact of genealogy and geography on typological diversity. And any correlations attested can then be accounted for statistically. As argued in this paper, it does even seem to be possible to infer situations of contact or population movement from 'biased' typological samples.

Appendices

Appendix A: Languages selected from WALS

Name	Longi-tude	Lati-tude	Genus	Family
Abkhaz	41	43.08	Northwest Caucasian	Northwest Caucasian
Ainu	143	43	Ainu	Ainu
Albanian	20	41	Albanian	Indo-European
Armenian	45	40	Armenian	Indo-European
Basque	-3	43	Basque	Basque
Bulgarian	25	42.5	Slavic	Indo-European
Burushaski	74.5	36.5	Burushaski	Burushaski
Chukchi	187	67	Northern Chukotko-Kamchatkan	Chukotko-Kamchatkan
English	0	52	Germanic	Indo-European
Evenki	125	56	Tungusic	Altaic
Finnish	25	62	Finnic	Uralic
French	2	48	Romance	Indo-European
Georgian	44	42	Kartvelian	Kartvelian
German	10	52	Germanic	Indo-European
Greek	22	39	Greek	Indo-European
Hungarian	20	47	Ugric	Uralic
Hunzib	46.25	42.17	Avar-Andic-Tsezic	Nakh-Daghestanian
Ingush	45.08	43.17	Nakh	Nakh-Daghestanian
Irish	–8	53	Celtic	Indo-European
Japanese	140	37	Japanese	Japanese
Ket	87	64	Yeniseian	Yeniseian
Khalkha	105	47	Mongolic	Altaic
Korean	128	37.5	Korean	Korean
Latvian	24	57	Baltic	Indo-European
Lezgian	47.83	41.67	Lezgic	Nakh-Daghestanian
Nenets	72	69	Samoyedic	Uralic
Nivkh	142	53.33	Nivkh	Nivkh
Polish	20	52	Slavic	Indo-European
Russian	38	56	Slavic	Indo-European
Spanish	–4	40	Romance	Indo-European
Turkish	35	39	Turkic	Altaic
Yukaghir	150.83	65.75	Yukaghir	Yukaghir

Appendix B: Defining typological dissimilarity

The basic formula to establish the typological dissimilarity (or: distance) between two languages L_1 and L_2 is based on the number of similar characteristics s and the number of different characteristics d. When the basic "languages by characteristics" data matrix is complete for n characteristics, then of course $s = n - d$, but in typological data there will probably always be many missing data points, so it will be necessary to establish s and d independently. Given s and d, the basic unweighted dissimilarity D between L_1 and L_2 is defined as:

$$D_{unweighted}(L_1, L_2) = d / (d + s) = 1 - s / (d + s) = 1 - RIW$$

(RIW: Goebl's *Relativer Identitätswert*)

To introduce a stronger influence of rare characteristics, a weighting function v(s) for similar characteristics can be included. Instead of simply counting each similarity as '1', each similarity s_i is counted as $v(s_i)$, and the formula then includes a summation over all these $v(s_i)$:

$$D_{value\text{-}weighted}(L_1, L_2) = d / (d + \Sigma\, v(s_i))$$

The basic idea is that this function v should rate those similarities higher that are only rarely attested. Goebl originally proposed to take for each value s_i the fraction of occurrence p_{si} in the data, and then define $v(s_i) = 1 - p_{si}$. This is referred to in the paper as 'weighting by fraction' and the resulting typological distance between two languages is then identical to 1 – GIW (Goebl's *Gewichteter Identitätswert*). Another possible approach is to define $v(s_i) = -\log(p_{si})$, which can be seen as a measure of information content: the rarer the shared characteristic, the more informative it is for the similarity between the languages. This weighting is referred to in the paper as "weighting by logarithm".

A similar principle of weighting can also be applied to differences. Instead of counting each difference as '1' it is possible to explicitly specify the precise value for each of the various differences. For example, for consonant inventories (Maddieson 2005a) there is a much larger difference between language pairings with a small versus large consonant inventory than between language pairings with a small versus average consonant inventory. There are two practical problems preventing me from adding such a weighting here. First, it is unclear how such weights should be determined, other than by ad-

ding intuitively specified numerical values. Second, it is unclear how such a specification of differences interacts with specification of similarities. Specifically, language pairings with many differences might in special cases become more similar than language pairings with many similarities. However, these problems should be surmountable given more research.

Further, the features as a whole can be weighted, so instead of counting any similarity or difference in a feature F (i.e. in a specific 'map' in WALS) equally as 1, a function w(F) can be defined to selectively change the impact of complete features. The resulting typological distance will then be defined as:

$$D_{\text{feature+value-weighted}}(L_1, L_2) = \Sigma\, w(d_i) \,/\, (\Sigma\, w(d_i) + \Sigma\, v(s_i) \cdot w(s_i))$$

This feature-weighting function has been used in the current paper to remove some obvious definitional dependencies between features in WALS. Specifically, the features 3, 25, 95, 96, and 97 have been weighted as zero (i.e. they have been removed from the data) because they are combinations of other features in WALS. Similarly, the features 139, 140, 141, and 142 have been weighted as zero because the set of languages discussed in these features is incompatible with the other features. Moreover, the following groups of definitionally related features have been weighted by the inverse of the number of features in the group (i.e. a feature in a group of four is weighted as 1/4):

- 14, 15, 16 (stress system)
- 30, 31, 32, 44 (gender marking)
- 37, 38 (articles)
- 39, 40 (clusivity, also known as "inclusive/exclusive distnctions")
- 49, 50 (case marking)
- 26, 51, 69 (affixation)
- 81, 82, 83, 85, 86, 93 (sentence word order)
- 87, 88, 89, 91 (nominal word order)
- 84, 90, 94 (complex sentence order)
- 77, 78 (evidentiality)
- 98, 99 (alignment)
- 40, 29, 100, 101, 102, 103 (verbal person inflection)
- 113, 114 (negation)
- 122, 123 (relativization)
- 125, 126, 127, 128 (clause conjunction)
- 132, 133, 134, 135 (color terms)

The result of all these weightings are six different measures of typological distance. These measures are all strongly correlated, as shown in Table 4, so the effects of using one over the others are minimal. When the paper simply refers to 'the' typological distance, this will imply the distance determined by weighting features and weighting the values by logarithm.

Table 4. Different ways to measure typological dissimilarity are highly correlated (Pearson r values for all measurement pairings).

	A	B	C	D	E	F
A) No weighting	1.000	0.979	0.985	0.892	0.943	0.984
B) Only value weighting (by logarithm)	0.979	1.000	0.994	0.910	0.969	0.988
C) Only value weighting (by fraction)	0.985	0.994	1.000	0.912	0.966	0.992
D) Only feature weighting	0.892	0.910	0.912	1.000	0.942	0.903
E) Weighting feature + value (by logarithm)	0.943	0.969	0.966	0.942	1.000	0.966
F) Weighting feature + value (by fraction)	0.984	0.988	0.992	0.903	0.966	1.000

References

Cysouw, Michael 2005: Quantitative methods in typology. In: Gabriel Altmann, Reinhard Köhler & Rajmund Piotrowski (eds.), *Quantitative Linguistics: An International Handbook*, 554–578. Berlin/New York: Mouton de Gruyter.

Cysouw, Michael 2008: Generalizing scales. In: Marc Richards & Andrej Malchukov (eds.), *Scales*, 379–396. Leipzig: Institut für Linguistik, Universität Leipzig.

Cysouw, Michael 2011: Quantitative explorations of the world-wide distribution of rare characteristics, or: the exceptionality of north-western European languages. In: Horst Simon & Heike Wiese (eds.), *Expecting the Unexpected*, 411–431. Berlin/New York: Mouton de Gruyter.

Diessel, Holger 2008: Distance constrasts in demonstratives. In: Martin Haspelmath, Matthew M. Dryer, David Gil & Bernard Comrie (eds.), *The World Atlas of Language Structures Online*, Chapter 41. Munich: Max Planck Digital Library.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Steve C. Levinson 2005: Structural phylogenetics and the reconstruction of ancient language history. *Science* 309(5743): 2072–2075.

Goebl, Hans 1984: *Dialektometrische Studien: anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und AFL.* (Beihefte zur Zeitschrift für Romanische Philologie 191). Tübingen: Niemeyer.

Legendre, Pierre 2000: Comparison of permutation methods for the partial correlation and partial Mantel tests. *Journal of Statistical Computation and Simulation* 67(1): 37–73.

Maddieson, Ian 2005a: Consonant inventories. In: Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *World Atlas of Language Structures*, 10–13. Oxford: Oxford University Press.

Maddieson, Ian 2005b: Tone. In: Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *World Atlas of Language Structures*, 58–61. Oxford: Oxford University Press.

Mantel, Nathan 1967: The detection of disease clustering and a generalized regression approach. *Cancer Research* 27(2): 209–220.

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

Nerbonne, John 2009: Data-driven dialectology. *Language and Linguistics Compass* 3(1): 175–198.

Nerbonne, John & Christine Siedle 2005: Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72(2): 129–147.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Tobler, Waldo R. 1970: A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–240.

Zapala, Matthew A. & Nicholas J. Schork 2006: Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences of the United States of America* 103(51): 19430–19435.