MICHAEL CYSOUW (Leipzig)
MIHAI ALBU (Leipzig)
ANDREAS DRESS (Shanghai)

Analyzing feature consistency using dissimilarity matrices

In this note, we present three methods to discover the most consistent features in the *World Atlas of Languages Structures* (WALS). These methods measure the fit between each individual WALS feature and the overall dataset of all features combined. Features that show a strong fit to the overall dataset are hypothesised to be more central for the structure of human language than those features that show a weak fit. The three techniques we will use are based on (i) MANTEL's congruence test (MANTEL 1967), (ii) the evaluation of feature coherence relative to the overall dataset, and (iii) the comparisons of ranks. All three methods attempt to identify those features that fit best to the dataset in its entirety, though it turns out that they do not identify exactly the same features. Still, we are able to give some indications of the kind of features that appear to be most promising for future research. Finally, we investigate whether such highly consistent features might be suitable to uncover genealogical relationships between languages.

1. Introduction

Besides being a printed atlas, the *World Atlas of Language Structures* (WALS, HASPELMATH *et al.* 2005) is also a database that contains data for 2,560 languages, recording characteristic traits for up to 142 predefined features represented by the various maps in the atlas (we will use the terms 'feature' and 'map' interchangeably). One of the many questions that can be tackled with this tremendous resource is the question as to whether there are some linguistic features that are more "consistent" than others. Typologically, consistent features are such features that are most indicative of the overall structure of a language. In a sense, such consistent features would be most indicative of the typological profile, or 'genius', of a language (if such a typological profile exists at all). On a different level of analysis, highly consistent features might also be more predictive of the genealogical relationships between languages. Both prospects indicate that the search for a proper measurement of feature consistency should be an important goal of research (see also the papers by WICHMANN & KAMHOLZ and PARKVALL in this issue for related issues). In this article, we will propose and compare three different approaches that might help us to establish the relative consistency of linguistic features based on the WALS data. Our basic assumption is that the consistency of a particular linguistic feature can be established by comparing it to the overall structure derived from summarising over all features. More consistent features will show a stronger match to that overall structure.

The WALS data needs to be analyzed with care (cf. CYSOUW *et al.* 2005) because of (i) non-encoded (yet essential and partially presupposed) expert knowledge regarding various specific properties of the features, (ii) non-canonical feature values, like "absence" or "other" (meaning "something, but none-of-the-above") and (iii) lateral dependency relationship between concepts. Further, the WALS data table is rather sparse as many datapoints are just missing. In spite of such shortcomings, WALS represents a huge amount of information that we deem to deserve further exploration. For our present study, we decided to take the data as it is presented in the WALS, and not to do any time-consuming recoding. The only *a priori* correction of the data is that we disregarded the data from map 3 (MADDIESON 2005), map 25 (NICHOLS & BICKEL 2005*a*), maps 95, 96, 97 (DRYER 2005*a,b,c*), maps 139, 140 (ZESHAN 2005*a,b*), and map 141 (COMRIE 2005*a*). The data in maps 3, 25, 95, 96, and 97 are all combinations of data presented individually in other maps. There is no new information in these maps, but only different presentations of the same information already given in other maps. The reason to disregard the maps 139, 140, and 141 is of a completely different nature. The maps 139 and 140 represent data about sign languages. It is not possible to compare these maps to all the other maps because there is no overlap with the maps on spoken languages. There is nothing inherently problematic that would make sign languages incomparable to spoken languages. However, at present we simply do not have the data to make such a comparison. Finally, the map on writing systems (COMRIE 2005*a* = WALS 141) is also not applicable, as there is no information given on individual languages, but only about geographical areas in which particular kinds of writing systems are particularly prominent.

In Section 2, we will first present a few basic definitions regarding dissimilarity matrices that will form the basis of our measurements of consistency. Then, in Sections 3 to 5, we will present our three different approaches to measuring matrix consistency. A few basic tests on the validity of these measurements are discussed in Section 6. We will then compare the results from the different measurements in Section 7. Section 8 contains a few preliminary observations on the usability of highly consistent features for the investigation of genealogical relationships. Finally, Section 9 summarizes our conclusions.

2. Distance matrices

Given any language L and feature F, WALS assigns to L and F a value $F(L)$—an integer between 1 and, at most, 9—in case that feature is "defined" for L. For the many cases where there is no value assigned, we put $F(L) := 0$. Based on this data, we will define a distance matrix describing the pairwise distances between all pairs of languages. Given a finite set **N**, then an **N** x **N** distance matrix D is a two-dimensional array with rows and columns indexed by the elements in **N**, the entry $D(L_1,L_2)$ at row L_1 and column L_2 being assumed to record the *distance* or *dissimilarity* between any two elements L_1,L_2 of **N** computed according to some preconceived scheme regarding the elements in **N**. Thus, an **N** x **N** distance matrix is a symmetric **N** x **N** matrix containing non-negative real numbers as entries and, in general, zeros along its diagonal: $D(L,L) = 0$. Based on the features from WALS and a selection **L** of languages that have good data coverage, we define the following **L** x **L** distance matrices.

First, we define, for every single feature F, a distance matrix D_F whose entry $D_F(L_1,L_2)$ is given, for any two languages L_1 and L_2 in **L**, by the term defined in (1).

$$(1)\qquad D_F(L_1,L_2) = \begin{cases} 2 & \text{if } F(L_1) \neq F(L_2) \text{ and } F(L_1) \neq 0 \text{ and } F(L_2) \neq 0, \\ 1 & \text{if } F(L_1) = 0 \text{ or } F(L_2) = 0, \\ 0 & \text{else (i.e., if } F(L_1) = F(L_2) \text{ and } F(L_1) \neq 0 \text{ and } F(L_2) \neq 0). \end{cases}$$

In words, the “*F*-distance” is equal to two (i.e. the languages are declared to be *really* different relative to *F*) whenever the feature *F* is well-defined for both languages and the two values $F(L_1)$ and $F(L_2)$ are distinct in WALS. The distance is set to be zero when *F* is well-defined for both languages, and the two values $F(L_1)$ and $F(L_2)$ this features attains at L_1 and L_2 in WALS coincide. Most importantly, the distance is defined as one when either one (or both) of the two languages is not coded for this particular feature. We prefer this definition to simpler ones (i.e. considering the cases with unavailable data as either all similar or all different) because we want the “*F*-distance” to reflect the situation that there is missing information in WALS.

Second, we define an overall distance matrix *D* taking account of all features simultaneously. In this matrix, the distance between two languages is computed only on the basis of available information. To achieve this, we denote by **F** the set of all features under consideration and define, for any language *L*, the set **F**(*L*) as the collection of all features *F* for which WALS provides data, cf. (2).

(2) $$\mathbf{F}(L) := \{F \in \mathbf{F} : F(L) \neq 0\}$$

Then, we define the normalized distance between any two languages L_1 and L_2 according to (3).

(3) $$D(L_1, L_2) = \frac{\sum_{F \in (\mathbf{F}(L_1) \cap \mathbf{F}(L_2))} D_F(L_1, L_2)}{\# \left(\mathbf{F}(L_1) \cap \mathbf{F}(L_2)\right)}$$

In words, for quantifying the distance between L_1 and L_2 only those features are considered for which, according to (2), data is available for both languages (i.e. all cases where $D_F(L_1,L_2) = 1$ are ignored). Then, in (3), the distances over all these available features are summarised, and divided by the number of available features. This procedure assures that the available data in WALS is completely used. For every two languages, however, a different set of features might be used, depending on the available information.

This way, 134 distance matrices of the form D_F result, one for each individual feature *F*, and one overall distance matrix *D*, using all 134 features together. Our goal is to identify those features *F* that somehow harmonize with the overall data. The idea behind this goal is to find out which features are the best predictors for the overall similarities between languages. To investigate the relation between an individual feature *F* and the overall dataset, we compare each matrix D_F with the overall *D* matrix. We will discuss three different methods that can be used for such comparisons between distance matrices. First, we will look at MANTEL’s congruence test (Section 3), and then introduce two other methods of our own design: the coherence method (Section 4) and the rank method (Section 5).

3. Mantel’s congruence test

Consider an **N** x **M** matrix X providing records regarding a collection **M** of experiments (measurements) applied to candidates (objects) from a set **N**. Dividing

M into two nonempty disjoint subsets $\mathbf{M_1}$ and $\mathbf{M_2}$, it is natural to ask for the correlation between the results obtained by performing the experiments in $\mathbf{M_1}$ and those obtained performing by the experiments in $\mathbf{M_2}$. To answer this question, the first step is to derive corresponding dissimilarity matrices from the **N** x $\mathbf{M_1}$ data matrix X_1 and the **N** x $\mathbf{M_2}$ data matrix X_2. The main idea is that, if two rectangular matrices contain concordant information, the distances derived from them should be significantly correlated. However, one cannot use a standard correlation coefficient to asses this significance, because the elements in the dissimilarity matrices are not independent of each other. For example, the distance between two objects A and B is not independent of the distance between object A and another object C because A is involved in both.

One method to quantify the "congruence" between two dissimilarity matrices was first proposed by MANTEL (1967). By combining KENDALL's W coefficient of concordance among matrices, FRIEDMAN's χ^2 statistics, and the associated *p*-value, MANTEL suggested to use the following procedure for measuring the correlation between the two similarity matrices X_1 and X_2. He proposed to arbitrarily permute the rows *k* times within one of the two matrices and recalculate the correlation coefficients. If there is some correlation, the disruption caused by the permutations should reduce the correlation coefficient. As a measure of congruence between X_1 and X_2, he therefore proposed to choose the quotient $s(X_1|X_2)$ of the number of times that the original correlation coefficient (R_0) was exceeded by the coefficients obtained for the permuted matrices, and the number *k* of all permutation tests being performed. For example, if these coefficients exceeded R_0 in only one from one thousand permutation tests, this would imply $s(X_1|X_2) = 0.001$. Conversely, if the matrices were uncorrelated, there is no reason to assume that the permutations would decrease the correlation coefficient. They may indeed as well increase it. So, we would assume that $s(X_1|X_2)$ would be close to 0.5.

Here, we use Mantel's test to assess the strength of the correlations between the dissimilarity matrices D_F (for each individual feature from WALS) and the overall dissimilarity matrix *D*. The smaller the value of $s(D_F|D)$, the better the individual feature *F* predicts the overall similarity between the world's languages.[1]

4. The coherence method

Alternatively, we propose to measure, for each of the matrices D_F corresponding to one particular feature *F*, its "coherence" with the overall matrix *D* by calculating the *triangle coherence index* for each feature matrix D_F relative to *D*. To do this, we first define the *excess* of any two elements L_1 and L_2 relative to a third element L_3 with respect to a distance matrix *M* as shown in (4):

(4) $$exc_M(L_1L_2|L_3) = M(L_1,L_3) + M(L_2,L_3) - M(L_1,L_2)$$

The excess is, roughly spoken, the extra distance to be travelled between L_1 and L_2 when the route is taken via L_3, instead of taking the direct path from L_1 to L_2.

[1] To calculate the MANTEL statistics, we used the CADM software as described in LEGENDRE & LAPOINTE (2004), available online at <http://www.bio.umontreal.ca/casgrain/en/labo/cadm.html> .

Based on that concept, the *triangle coherence index* $\Delta_{coh\text{-}index}(F)$ of a feature F is then defined as shown in (5). For every triplet of languages L_1, L_2, L_3, the quotient is taken of the excess relative to the overall matrix D and the excess relative to the single feature matrix D_F (the quotient is set at infinite when exc$_{DF}$(L_1 L_2| L_3) = 0 holds). The triangle coherence index for a feature F then is the average of all these quotients for all possible triplets from the total set of languages **L** under consideration. A larger triangle coherence index will indicate a higher degree of coherence of a feature matrix D_F with the overall matrix D.

(5) $$\triangle_{coh-index}(F) = avg\left(\frac{1+exc_D(L_1L_2|L_3)}{1+exc_{D_F}(L_1L_2|L_3)} \mid L_1,L_2,L_3 \in \mathbf{L}\right)$$

5. The rank method

As a third method to investigate the consistency between a feature and the overall dissimilarity matrix D, we propose a rank-based method. The rank $rk_{L1}(L_2)$ of a language L_2 with respect to a language L_1 (relative to a distance matrix D) is defined as shown in (6):

(6) $$rk_{L_1}(L_2) = rk^D_{L_1}(L_2) := \#\{L \in \mathbf{L} | D(L_1,L) \leq D(L_1,L_2)\}, (L_1,L_2 \in \mathbf{L}).$$

In words, for every language L_1, we are counting the number of languages L whose distance to L_1 that is not larger than the distance between L_1 and L_2. Metaphorically speaking, we are looking for the languages that are at least as good a friend of L_1 as is L_2 (cf. DEVAUCHELLE *et al.* 2005, ALBU *et al.* 2006)

By using this rank value, we can derive a rank matrix R_D from a distance matrix D as shown in (7):

(7) $$R_D(L_1,L_2) := rk^D_{L_1}(L_2), \text{for all } L_1,L_2 \in \mathbf{L}$$

This rank matrix is not necessarily symmetric, because $rk_{L1}(L_2)$ is not necessarily the same as $rk_{L2}(L_1)$. In general, the rank matrix has values of 1 along its diagonal. Higher values on the diagonal will only occur in case there are different languages with completely identical datasets. To illustrate the derivation of such rank matrices from distance matrices, consider the example in (8). The distance matrix (8a) will be transformed into a rank matrix as shown in (8b). For example, from the perspective of L$_3$, all three languages L$_1$, L$_2$, and L$_3$ are at least as similar to L$_3$ as L$_1$, so the rank value $rk_{L3}(L_1)$ is 3. However, from the perspective of L$_1$, only L$_1$ and L$_3$ are as similar to L$_1$ as L$_3$. L$_2$ is less similar to L$_1$ than L$_3$. Thus, $rk_{L1}(L_3)$ is 2.

(8) a. Distance Matrix

	L_1	L_2	L_3
L_1	0	4	3
L_2	4	0	2
L_3	3	2	0

b. Rank Matrix

	L_1	L_2	L_3
L_1	1	3	2
L_2	3	1	2
L_3	3	2	1

Next, we consider, for each feature *F* and language *L*, the set **L**(*L,F*) of all languages that share the same feature value with *L*, as formally defined in (9).

(9) $$\mathbf{L}(L,F) = \{L' \in \mathbf{L} | F(L) = F(L')\}$$

If a feature *F* fits well into the overall data structure encoded by the distance matrix *D*, then the rank rk_L(L′) of the languages L′ in the subset **L**(*L,F*) relative to L should be significantly smaller, for any given L ∈ **L**, than the rank rk_L(L′′) of the languages L′′ in the complement **L** – **L**(*L,F*) relative to L. Metaphorically, the languages with the same feature value as L should be better friends of *L* than the languages with different feature values. Consequently, we propose a measure of fitness for a feature *F* to the overall matrix *D* using the ranking procedure as defined in (10).

(10) $$\rho_D(F) := \frac{1}{\#\mathbf{L}} \sum_{L \in \mathbf{L}} \frac{\sum_{L' \in \mathbf{L}(L,F)} rk_L^D(L')}{\binom{1 + \#\mathbf{L}(L,F)}{2}}$$

In words, for each language *L*, we first take the sum of the ranks, relative to *L*, of all other languages *L′* with the same feature value divided by the smallest possible value of that sum, and then take the average over all languages in **L**(*F*). The lower this average, the better the fit between the feature *F* and the overall matrix *D*.

Note that the meaning of these averages is 'opposite' to that of the numbers resulting from the Mantel congruence test and coherence method. In the Mantel congruence test and the coherence method, higher values indicate better fit. In the rank method, lower values indicate better fit. Technically speaking, if the coherence method and the rank method would agree on which features show a high value, then their correlation would be strongly negative. In contrast, if the coherence method and the Mantel congruence test would agree, then their correlation would be strongly positive. For convenience, we will change the sign in all the correlations concerning the rank measure in the following sections, so that the sign of the correlations will have the same interpretation for all three methods.

6. Evaluation of the methods

To test the three methods, we selected a set of 150 languages from WALS (see Appendix B). The selection was mainly based on data coverage, choosing those languages for which most datapoints are available in WALS. Further, we choose only one language per genus (of course the one with the largest data coverage), in order to get a good survey of the worldwide diversity. Further, we also had to select a subset of the available maps. In WALS, there are 142 different maps. However, not all of these maps could be used for our analyses. As already mentioned in the introduction, we disregarded the maps on sign languages (139 and 140), on writing systems (141) and the maps that replicate data from other maps (3, 25, 95, 96 and 97). However, because of our selection of 150 languages, the data coverage for a few additional maps was also very low. In particular, the maps on the paralinguistic usage of clicks (GIL 2005 = WALS 142) and the maps on colour terms (KAY & MAFFI 2005*a,b,c,d* = WALS 132, 133, 134, and 135) turned out to be only very sparingly represented in our 150-language sample. Removing these maps from our data, we were left with 129 different features to use for the analysis.

To test how strong the methods depend on the choice of languages, we semi-randomly divided the 150 languages in three datasets (50 languages per dataset) in such a way as to obtain a uniform distribution of the number of datapoints over the three datasets. We then ran each of the three methods for each of the three datasets. For all three methods, there was a highly significant correlation between the results from the three subsets (cf. Table 1, first column). In comparison, the strength of the correlations is clearly higher for the coherence method than for the rank methods and MANTEL's congruence test. This indicates that the coherence method is least influenced by the choice of languages. Because of the strongly significant correlations between the three datasets for all methods, we will subsequently only use the average value over all three datasets as a measure of consistency between each feature and the overall dataset (these averages are reported on in Appendix D).

Table 1. Internal consistency of the methods (reported values are Pearson's *r*, significances $p < .001$ are indicated with a star).

Method	Correlations between subsamples	Correlations with data coverage	Correlations with deviation from homogeneous distribution
Mantel	.74* .75* .69*	.52 *	.08
Coherence	.89* .89* .90*	.72 *	.29 *
Rank	.59* .64* .67*	.18	.10

One problem regarding the WALS data is the large amount of missing data. Through our selection of languages and features we already obtained a rather well represented subset of the WALS data. However, while choosing 129 features and 150 languages should ideally result in 129*150 = 19,350 datapoints, there are 'only' 15,589 datapoints available in our selection (80.6 %). Unfortunately, the missing data is not equally distributed throughout the table. The maximum of available data for a feature is the full 150 languages, but the minimum of available data is found for the feature on the word for *tea* (DAHL 2005 = WALS 138) which is available for only 53 languages out of our 150-language sample. Overall, the number of languages per feature has a mean of 120, with a standard deviation of 26.3 (all individual numbers are summarised in Appendix D). With such widely varying frequencies of available data, it is important to check *post-hoc* whether this has any influence on our three methods. We correlated the results for each method with the number of available data, which yielded significant correlations for both the MANTEL congruence test and the coherence method (cf. Table 1, second column). Both methods thus give higher values when more data is available. The strongest effect can be discerned for the coherence method.

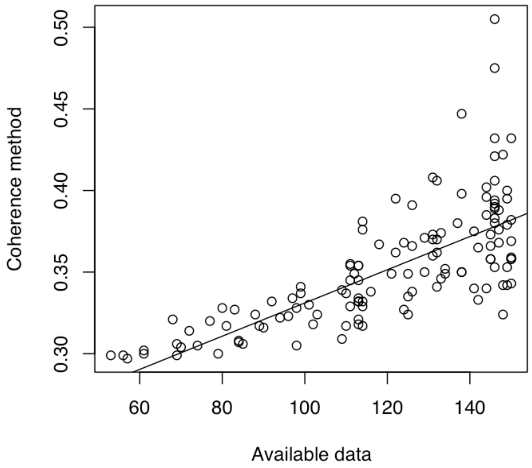
Another parameter that widely differs between the various features is the concentration of the languages over the feature values. Some features have roughly equally distributed numbers of languages over the various values, but other features have some values attained by very many languages and other values attained by only a few languages (cf. MASLOVA, this issue). We used a normalised version of the HERFINDAHL-HIRSCHMAN Index (cf. HIRSCHMAN 1964 on the origin of this index) as shown in (11) to investigate the correlation between our three methods and the homogeneity of the distribution of the feature values (see Appendix A for some notes on the relation between the HERFINDAHL-HIRSCHMAN Index and related mathematical functions, like the entropy).

(11) $$C_F = 1 - \frac{1}{k_F \sum_{i=1}^{k_F} v_i^2}$$

This index C_F gives a measure of the homogeneity or ‘concentration’ of the languages over the various values distinguished by the feature F. For each feature F, there are k_F values (ranging between 2 and 9). In the formula, for each value $i \in \{1, \ldots, \mathrm{k_F}\}$, v_i is the proportion of languages that have this particular value. To compute the HERFINDAHL-HIRSCHMAN Index, the sum of all squares of these proportions is taken. All other transformations in (11) are only added to normalise this raw concentration value to the interval [0,1], with the lower side describing situations in which all values are roughly equally probable (weak concentration) and the higher side indicating situations in which a few values dominate (strong concentration). The correlations between this measure of concentration and the three methods to measure consistency are shown in Table 1 (third column). The correlations are all very low, the highest (and the only significant) correlation being the one with the coherence method. Note that the degree of deviation from homogeneous distribution and the data coverage are not correlated with each other at all, so the facts that the coherence method shows the strongest correlation with both of them are two independent observations.

In principle these significant correlations could be a problem for the validity of our measurements of consistency. For example, when the amount of available data is correlated with a measure of consistency, then is it really consistency that we are measuring, or rather the fact that, for some features, we have higher data coverage than for others? Fortunately, for the present data and measurements, the attested significant correlations are not problematic because the variation of the measures also increases with the amount of available data. This effect is illustrated in Figure 1, showing the correlation between the amount of available data and the measurements of the coherence method. As can be seen, the more data available, the larger the variation in the coherence measurements. The line in this plot is the regression line. Now, because of the large variation on the right side of the plot, the features with high coherence measurements are also high when taken relative to the regression line. This means that high coherence measures are still high when the amount of available data is factored out. Note that this does not hold for low coherence measures. Low values of coherence of a feature could be due to low consistency as well as to the paucity of available data. The same situation as shown in Figure 1 is also found for the other significant correlations in Table 1.

Figure 1. Correlation between the amount of available data and the values from the coherence method.



7. Comparing the methods

By using three different approaches to measure consistency, our hypothesis was that they all would yield comparable results. If that would be the case, then we could be confident that the different measures are indeed an indication of the consistency of a feature. The ideal situation would thus be that, independently of what kind of method we would use, the same features would be considered highly consistent with the overall dataset. The actual results are not as equivocally as we would have hoped. The Pearson correlation coefficients between the three methods are shown in Table 2. The correlations between the coherence and rank methods on the one side, and the MANTEL approach on the other, are both very low and not significant. However, the correlation between the rank and the coherence method is actually rather good, and significant as well. This situation becomes even clearer when the factors as considered in the previous section are corrected for. Considering the significant influences of data coverage and deviation from homogeneous distribution, we did a regression analysis and then considered the residuals after regression. The correlations between these residuals are shown in Table 3. The association between the coherence and the rank method now is slightly higher, and the correlation between the rank method and the MANTEL approach becomes even worse.

Table 2. Correlations between the three methods (reported values are Pearson's *r*, significances *p* < .001 are indicated with a star).

	Coherence	Rank
Mantel	.22	.22
Coherence		.65*

Table 3. Correlations between the three methods, taking the residuals after regression relative to data coverage and deviation from homogeneous distribution (reported values are Pearson's *r*, significances *p* < .001 are indicated with a star).

	Coherence	Rank
Mantel	-.26	.15
Coherence		.75*

Because there is no complete consensus between the different methods, we cannot draw any far-reaching conclusions about general consistency between individual features and the overall data structure of WALS. However, it is important for future linguistic data collection to get at least a rough impression about what kind of features among the WALS data show a good consistency with the overall data structure. The following interpretation is not derived by strictly defined statistical tests, but by a manual inspection of the top ranked features from the three methods.

All three methods put various word order features high on their ranking. Specifically, subject-verb order (DRYER 2005*e* = WALS 82), object-verb order (DRYER 2005*f* = WALS 83), adposition order (DRYER 2005*g* = WALS 85), genitive-noun order (Dryer 2005*h* = WALS 86), and demonstrative-noun order (DRYER 2005i = WALS 88) are ranked high by all methods. This agreement between the methods, however, might be due to the fact that there are various maps in the WALS about word order, and that these are all more or less significantly correlated to each other.

The MANTEL congruence test further argues for the inclusion of features related to the morphological structure, both verbal and nominal. In the realm of verbal morphology, the marking of clusivity on verbs (CYSOUW 2005*a* = WALS 40), the location of tense/aspect affixes (DRYER 2005*j* = WALS 69), and the presence and order of verbal person marking (NICHOLS & BICKEL 2005*b* = WALS 23; SIEWIERSKA 2005*a* = WALS 102) are deemed important. In the realm of nominal morphology, various features related to case marking end up high on the list (BICKEL & NICHOLS 2005 = WALS 21, NICHOLS & BICKEL 2005*b* = WALS 23, IGGESEN 2005*a,b* = WALS 49 and 50, DRYER 2005*k* = WALS 51 and COMRIE 2005*b* = WALS 98).

The coherence method and the rank method argue for the inclusion of different features. Besides word order, these methods suggest to include features about the presence or absence of strictly defined kinds of consonants, like uvular consonants (MADDIESON 2005*b* = WALS 6), glottalised consonants (MADDIESON 2005*c* = WALS 7), the velar nasal (ANDERSON 2005 = WALS 9), and more generally the absence of common consonants (MADDIESON 2005*d,e* = WALS 5 and 18) and the presence of uncommon consonants (MADDIESON 2005*f* = WALS 19). Further, the coherence and rank methods suggest various seemingly disparate features, like the passive (SIEWIERSKA 2005*b* = WALS 107), inflectional optatives (DOBRUSHINA *et al.* 2005 = WALS 73), front rounded vowels (MADDIESON 2005*g* = WALS 11), tone (MADDIESON 2005*h* = WALS 13), and clusivity and gender in pronouns (CYSOUW 2005*b* = WALS 39; SIEWIERSKA 2005*c* = WALS 44) as being important features.

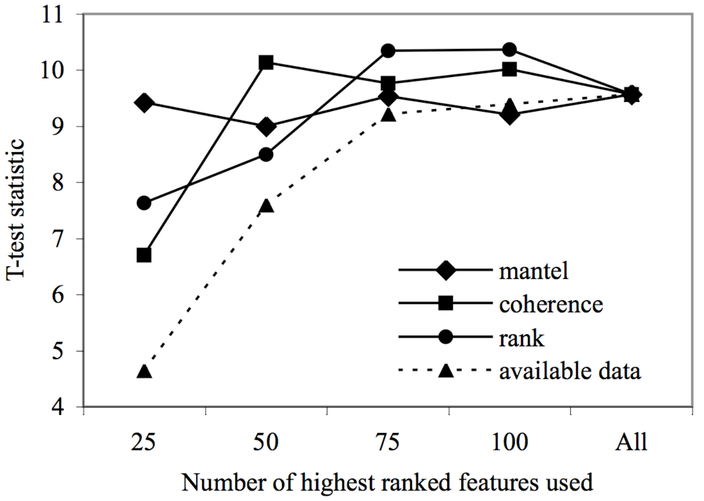
8. Predicting genealogical relationship

In the methods that we have presented to measure consistency we have not used any information about genealogical relationships. Still, we wanted to see whether consistency might be a good predictor for genealogical relationship. To test this hypothesis, we constructed a sample of eleven families from WALS, taking three languages out of each family (see Appendix C). The choice of families and languages was completely driven by data availability. We wanted to know how well a particular selection of features would be able to distinguish pairs of related languages from pairs of unrelated languages in this sample.

To investigate this, we constructed an overall distance matrix for the 33 languages sampled on the basis of all data in WALS. The distance matrices were compiled using the method as described in (3) above. Likewise, we constructed various distance matrices based on a selection of the features. Each selection of features was determined by the ranking of features as given by the various measures of consistency that we have discussed. For every method, we subsequently considered the most consistent 25, 50, 75 and 100 features, and constructed distance matrices on that basis. As a control, we also considered the amount of available data as a ranking, constructing distance matrices on the basis of the best covered 25, 50, 75 and 100 features. In this way, we had sixteen different distance matrices for our test sample of 33 languages.

All distances in such a matrix were then divided into two groups: one group with all distances between pairs of related languages and one group with all distances between pairs of unrelated languages. We then wanted to know, whether the distances between related languages are generally smaller than the distances between unrelated languages. To investigate this, we used a t-test to determine the significance of the difference between the two groups. It turned out that the two groups were significantly different for all sets of features considered. Still, there are clear differences between the various selections. This can be seen by considering the t-test statistics itself (not the significances). These t-test statistics are summarised in Figure 2. Selecting features by available data (the dotted line in Figure 2) gives some sort of baseline to compare our various methods against. The dotted line starts low and rises continuously, though the slope flattens the more features are considered. This indicates that we are able to get better differentiation between related and unrelated language pairs the more features we consider, though there seems to be a level of differentiation that cannot be improved upon. Looking now at the various selections of features, we see that the best 25 features as selected by the three methods all show a clearly stronger differentiation between related and unrelated pairs compared to taking the best covered 25 features. Taking the best 25 features from the MANTEL approach even gives roughly the same differentiation as given by considering all features together. Several of the other selections even improve on this. This indicates that, by selecting a set of consistent features, it is possible to improve the recognition of genealogical relationships compared to simply taking all available data.

Figure 2. t-test statistic for the differentiation between related and unrelated pairs of languages for language distances as established by selected sets of features on the basis of the ranking of consistency.



9. Conclusions

We have presented three methods to measure consistency between a feature and the overall dataset of WALS: the MANTEL congruence test, the coherence and the rank method. All three measures are relatively independent from the choice of languages, though especially the coherence method shows a tendency to give lower values for features with a low data coverage. As unavailable data is a perennial problem in typological databases, care should be taken not to use the coherence method for features with lots of missing data. Comparing the three methods with each other, it turns out that a strong correlation exists between the coherence and the rank method, the more so when features with low data coverage are discarded. The MANTEL congruence test results in strikingly different consistency values for the various features. Apparently, it picks up other distributional patterns compared to the coherence and rank methods, though we have not yet been able to exactly determine how these differences come about. Impressionistically, the MANTEL congruence test favours word order features and features related to morphological structure. The coherence and the rank methods also favour word order features, but besides that, both these methods prefer various phonological features and the structure of pronouns. Finally, we considered whether consistent features were also genealogically stable features, and we found at least some indication in this direction.

References

Albu, Mihai, Claudine Devauchelle, Andreas W. M. Dress, & Alexander Grossman (2006): A rank based approach to phylogenetics. Unpublished manuscript.

Anderson, Gregory D. S. (2005): The Velar Nasal, in: Haspelmath *et al.*, 42-45.

Bickel, Balthasar & Johanna Nichols (2005*a*): Exponence of Selected Inflectional Formatives, in: Haspelmath *et al.*, 90-93.

— &— (2005*b*): Exponence of Selected Inflectional Formatives, in: Haspelmath *et al.*, 90-93.

Comrie, Bernard (2005*a*)*:* Writing Systems, in: Haspelmath *et al.*, 566-569.

— (2005*b*): Alignment of Case Marking of Full Noun Phrases, in: Haspelmath *et al.*, 399-401.

Cysouw, Michael (2005*a*): Inclusive/Exclusive Distinction in Verbal Inflection, in: Haspelmath *et al.*, 166-169.

— (2005*b*): Inclusive/Exclusive Distinction in Independent Pronouns, in: Haspelmath *et al.*, 162-165.

—, Jeff Good, Mihai Albu, & Hans-Jörg Bibiko (2005): Can gold "cope" with wals? retrofitting an ontology onto the world atlas of language structures. *E-MELD Workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Linguistic Resources*.

Devauchelle, Claudine, Andreas W. M. Dress, Alexander Grossman, Stefan Gruenewald, & Alain Henaut (2005): Constructing hierarchical set systems, in: *Annals of Combinatorics* 8.4, 441-456

Dobrushina, Nina, Johan van der Auwera, & Valentin Goussev (2005*b*): The Optative, in: Haspelmath *et al.*, 298-301.

Dryer, Matthew S. (2005*a*): Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase, in: Haspelmath *et al.*, 386-389.

— (2005*b*): Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun, in: Haspelmath *et al.*, 390-393.

— (2005*c*): Relationship between the Order of Object and Verb and the Order of Adjective and Noun, in: Haspelmath *et al.*, 394-398.

— (2005*d*): Order of Subject and Verb, in: Haspelmath *et al.*, 334-337.

— (2005*e*): Order of Object and Verb, in: Haspelmath *et al.*, 338-341.

— (2005*f*): Order of Adposition and Noun Phrase, in: Haspelmath *et al.*, 346-349.

— (2005*g*): Order of Genitive and Noun, in: Haspelmath *et al.*, 350-353.

— (2005*h*): Order of Demonstrative and Noun, in: Haspelmath *et al.*, 358-361.

— (2005*i*): Position of Tense-Aspect Affixes, in: Haspelmath *et al.*, 282-285.

— (2005*j*): Position of Case Affixes, in: Haspelmath *et al.*, 210-213.

Gil, David (2005): Paralinguistic Usages of Clicks, in: Haspelmath *et al.*, 570-573.

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

Hirschman, Albert O. (1964): The paternity of an index, in: *American Economic Review* 54.5, 761-762.

Iggesen, Oliver A. (2005*a*): Number of Cases, in: Haspelmath *et al.*, 202-205.

— (2005*b*): Asymmetrical Case Marking, in: Haspelmath *et al.*, 206-209.

Kay, Paul & Luisa Maffi (2005*a*): Number of Nonderived Basic Colour Categories, in: Haspelmath *et al.*, 535-543.

— (2005*b*): Number of Basic Colour Categories, in: Haspelmath *et al.*, 535-543.

— (2005*c*): Green and Blue, in: Haspelmath *et al.*, 535-543.

— (2005*d*): Red and Yellow, in: Haspelmath *et al.*, 535-543.

Legendre, P. & F.-J. Lapointe (2004): Assessing the congruence among distance matrices: single malt Scotch whiskies revisited, in: *Australian and New Zealand Journal of Statistics* 46, 615-629.

MADDIESON, IAN (2005*a*): Consonant–Vowel Ratio, in: HASPELMATH *et al.*, 18-21.
— (2005*b*): Uvular Consonants, in: HASPELMATH *et al.*, 30-33.
— (2005*c*): Glottalized Consonants, in: HASPELMATH *et al.*, 34-37.
— (2005*d*): Voicing and Gaps in Plosive Systems, in: HASPELMATH *et al.*, 26-29.
— (2005*e*): Absence of Common Consonants, in: HASPELMATH *et al.*, 78-81.
— (2005*f*): Presence of Uncommon Consonants, in: HASPELMATH *et al.*, 82-85.
— (2005*g*): Front Rounded Vowels, in: HASPELMATH *et al.*, 50-53.
— (2005*h*): Tone, in: HASPELMATH *et al.*, 58-61.
MANTEL, N. (1967): The detection of disease clustering and a generalized regression approach, in: *Cancer Research*, 27, 209-220.
MASLOVA, ELENA (this issue): Meta-typological distributions.
NICHOLS, JOHANNA & BALTHASAR BICKEL (2005*a*): Locus of Marking: Whole-language Typology, in: HASPELMATH *et al.*, 106-109.
— &— (2005*b*): Locus of Marking in the Clause, in: HASPELMATH *et al.*, 98-101.
PARVALL, MIKAIL (this issue): [Title unknown].
SIEWIERSKA, ANNA (2005*a*): Verbal Person Marking, in: HASPELMATH *et al.*, 414-417.
— (2005*b*): Passive Constructions, in: HASPELMATH *et al.*, 434-437.
— (2005*c*): Gender Distinctions in Independent Personal Pronouns, in: HASPELMATH *et al.*, 182-185.
WICHMANN, SØREN & DAVID KAMHOLZ (this issue): A stability metric for typological features.
ZESHAN, ULRIKE (2005*a*): Irregular Negatives in Sign Languages, in: HASPELMATH *et al.*, 558-559.
— (2005*b*): Question Particles in Sign Languages, in: HASPELMATH *et al.*, 560-565.

Appendix A: A note on the Herfindahl-Hirschman Index

The HERFINDAHL-HIRSCHMAN index (HHI) as it is used here is a variant of a measure used frequently in economics. However, it is in fact only a special case of a more general set of functions to which also the entropy belongs. Recall that the entropy is never negative, vanishes in case of extreme concentration (only one value is attained), and reaches it maximum $log(k_F)$ in case all v_i coincide. Actually, given any (strictly) convex function H with $H(0) = H(1) = 0$, then putting

$$H_0 := \frac{\sum_{i=1}^{k_F} H(k_F \nu_i)}{k_F},$$

one has

$$0 \leq H_0 \leq \frac{H(k_F)}{k_F},$$

with equality on the left-hand side if (and only if) all v_i coincide, and on the right-hand side if (and only if) all but one of the v_i vanish. Thus, forming, for example, the terms

$$H_1 := \frac{H(k_F)}{k_F} - H_0$$

or

$$H_2 := \frac{H_0}{1+H_0}$$

will yield indices exhibiting the same behaviour as the entropy or the HHI, respectively. Indeed,

$$0 \leq H_1 \leq \frac{H(k_F)}{k_F} \text{ as well as } \frac{H(k_F)/k_F}{1+H(k_F)/k_F} \geq H_2 \geq 0$$

will always hold, with equality on the left-hand side if (and only if) all but one of the v_i vanish, and on the right-hand side if (and only if) all the v_i coincide. Furthermore, for $H(x) := x \cdot log(x)$, this yields the entropy, and for $H(x) := x \cdot (1-x)$, it yields the HERFINDAHL-HIRSCHMAN index usedin this paper.

Appendix B: Sample of 150 languages with maximum data coverage. Maximally one language per genus has been sampled.

Abkhaz, Acoma, Ainu, Alamblak, Amele, Apurin, Arabic (Egyptian), Araona, Arapesh, Armenian (Eastern), Asmat, Awa Pit, Aymara, Bagirmi, Barasano, Basque, Beja, Berber (Middle Atlas), Brahui, Burmese, Burushaski, Cahuilla, Canela-Krah, Cayuvava, Chamorro, Chinantec (Lealao), Chukchi, Comanche, Coos (Hanis), Cree (Plains), Daga, Dani (Lower Grand Valley), Diola-Fogny, English, Epena Pedee, Evenki, Ewe, Finnish, French, Garo, Georgian, Gooniyandi, Grebo, Greek (Modern), Greenlandic (West), Guaraní, Haida, Hausa, Hindi, Hixkaryana, Hmong Njua, Hungarian, Hunzib, Igbo, Ika, Imonda, Indonesian, Ingush, Iraqw, Irish, Jakaltek, Japanese, Ju|'hoan, Kannada, Kanuri, Karok, Kayardild, Ket, Kewa, Khalkha, Khasi, Khmer, Khmu', Khoekhoe, Kiowa, Koasati, Korean, Koyraboro Senni, Krongo, Kunama, Kutenai, Lakhota, Lango, Latvian, Lavukaleve, Lezgian, Makah, Malagasy, Mandarin, Mangarrayi, Maori, Mapudungun, Maranungku, Maricopa, Maung, Maybrat, Meithei, Miwok (Southern Sierra), Mixtec (Chalcatongo), Mundari, Murle, Nahuatl (Tetelcingo), Ndyuka, Nenets, Nez Perce, Ngiyambaa, Nivkh, Nubian (Dongolese), Nunggubuyu, Oneida, Oromo (Harar), Otomi, Paiwan, Paumarí, Persian, Pirahã, Quechua (Imbabura), Rama, Russian, Sango, Sanuma, Semelai, Shipibo-Konibo, Slave, Squamish, Suena, Supyire, Swahili, Taba, Tagalog, Thai, Tiwi, Trumai, Tsimshian (Coast), Tukang Besi, Turkish, Ungarinjin, Vietnamese, Warao, Wardaman, Wari', Wichí, Wichita, Yagua, Yaqui, Yimas, Yoruba, Yukaghir (Kolyma), Yurok, Zoque (Copainalá).

Appendix C: Test set consisting of eleven groups of three languages from the same family with high data coverage .

Afro-Asiatic	Hausa, Egyptian Arabic, Hara Oromo
Altaic	Turkish, Evenki, Khalkha
Austro-Asiatic	Vietnamese, Khasi, Khmer
Austronesian	Indonesian, Maori, Malagasy
Dravidian	Kannada, Brahui, Tamil
Indo-European	English, French, Russian
Nakh-Daghestanian	Lezgian, Hunzib, Ingush
Niger-Congo	Supyire, Swahili, Zulu
Sino-Tibetan	Mandarin, Burmese, Meithei
Trans-New Guinea	Amele, Kobon, Kewa
Uralic	Finnish, Hungarian, Nenets

Appendix D. Results.

No.	*Available datapoints*	*Homo-geneity*	*Mantel statistic*	*Coherence method*	*Rank method*
1	144	0.174	0.073	0.340	3.589
2	146	0.195	0.104	0.380	2.498
4	146	0.172	0.157	0.353	2.886
5	146	0.495	0.148	0.383	2.331
6	146	0.597	0.120	0.421	2.332
7	146	0.748	0.147	0.406	2.206
8	146	0.580	0.118	0.390	2.754
9	146	0.224	0.117	0.392	2.355
10	131	0.257	0.087	0.408	2.511
11	146	0.721	0.067	0.505	1.306
12	141	0.172	0.089	0.375	2.730
13	144	0.374	0.106	0.402	2.266
14	113	0.368	0.077	0.318	3.759
15	113	0.630	0.080	0.332	3.474
16	113	0.513	0.058	0.321	4.180

No.	*Available*	*Homo-*	*Mantel*	*Coherence*	*Rank*
	datapoints	*geneity*	*statistic*	*method*	*method*
17	70	0.470	0.073	0.304	3.066
18	146	0.753	0.134	0.475	1.372
19	146	0.753	0.103	0.432	2.160
20	118	0.734	0.145	0.367	2.495
21	116	0.491	0.202	0.338	2.793
22	114	0.372	0.093	0.317	3.623
23	125	0.215	0.204	0.324	3.346
24	125	0.342	0.151	0.335	3.371
26	142	0.292	0.216	0.333	3.245
27	112	0.362	0.060	0.349	3.053
28	147	0.402	0.174	0.376	2.751
29	147	0.045	0.144	0.368	2.459
30	131	0.549	0.195	0.360	3.362
31	131	0.322	0.205	0.373	2.844
32	131	0.289	0.193	0.370	3.071
33	138	0.620	0.149	0.350	3.533
34	84	0.470	0.079	0.308	3.478
35	150	0.513	0.113	0.343	3.858
36	85	0.055	0.092	0.306	3.519
37	114	0.370	0.085	0.329	3.633
38	102	0.344	0.091	0.318	3.648
39	149	0.550	0.120	0.395	2.343
40	149	0.317	0.201	0.353	3.020
41	109	0.539	0.020	0.339	3.026
42	92	0.399	0.030	0.332	2.858
43	110	0.369	0.023	0.317	4.204
44	149	0.605	0.143	0.400	2.501
45	124	0.554	0.112	0.368	2.820
46	96	0.530	0.129	0.323	3.063
47	97	0.005	0.055	0.334	2.698
48	150	0.379	0.148	0.382	2.553
49	148	0.421	0.263	0.324	3.686
50	148	0.349	0.234	0.342	3.032
51	138	0.561	0.237	0.350	2.779
52	88	0.372	0.113	0.324	2.977
53	109	0.321	0.108	0.309	4.036
54	69	0.395	0.062	0.299	3.301
55	99	0.448	0.024	0.341	2.765
56	61	0.070	0.030	0.300	3.034
57	126	0.218	0.135	0.338	3.099
58	122	0.332	0.104	0.395	2.305
59	122	0.510	0.132	0.362	2.973
60	61	0.592	0.076	0.302	2.724
61	56	0.610	0.050	0.299	2.490
62	98	0.302	0.044	0.305	4.523
63	77	0.188	0.132	0.320	2.296
64	101	0.234	0.084	0.330	2.883
65	113	0.009	0.103	0.354	2.639
66	113	0.351	0.189	0.334	2.983
67	113	0.000	0.155	0.354	2.576
68	113	0.471	0.153	0.345	2.651
69	142	0.501	0.227	0.365	2.757
70	145	0.461	0.160	0.366	2.852
71	132	0.189	0.065	0.341	3.707
72	133	0.468	0.077	0.374	2.705
73	138	0.380	0.054	0.447	2.069
74	126	0.321	0.053	0.366	2.823
75	129	0.089	0.118	0.350	3.092
76	110	0.173	0.090	0.337	3.028

No.	*Available datapoints*	*Homo-geneity*	*Mantel statistic*	*Coherence method*	*Rank method*
77	145	0.141	0.218	0.373	2.338
78	145	0.512	0.223	0.358	2.607
79	132	0.497	0.093	0.370	3.009
80	132	0.716	0.082	0.406	2.417
81	141	0.491	0.343	0.340	2.877
82	148	0.436	0.199	0.422	2.007
83	147	0.219	0.421	0.388	1.637
84	84	0.432	0.188	0.307	2.687
85	146	0.546	0.386	0.389	1.804
86	146	0.287	0.302	0.394	2.128
87	144	0.421	0.177	0.385	2.392
88	144	0.478	0.239	0.396	2.451
89	137	0.454	0.177	0.380	2.617
90	121	0.550	0.126	0.349	2.813
91	81	0.318	0.104	0.317	2.664
92	124	0.388	0.064	0.327	3.513
93	129	0.304	0.116	0.371	2.637
94	113	0.497	0.208	0.332	2.893
98	145	0.514	0.238	0.358	3.014
99	133	0.576	0.189	0.346	2.998
100	150	0.493	0.155	0.359	3.028
101	134	0.558	0.169	0.349	3.488
102	150	0.443	0.231	0.369	2.439
103	150	0.465	0.179	0.358	2.790
104	150	0.381	0.150	0.358	2.860
105	114	0.310	0.089	0.332	3.429
106	89	0.398	0.109	0.317	3.094
107	150	0.002	0.071	0.432	1.835
108	126	0.514	0.092	0.391	2.854
109	125	0.651	0.094	0.349	3.228
110	69	0.219	0.057	0.306	2.662
111	138	0.583	0.071	0.398	2.872
112	134	0.553	0.167	0.352	3.243
113	149	0.106	0.133	0.379	2.488
114	149	0.419	0.106	0.342	3.823
115	80	0.648	0.095	0.328	2.102
116	132	0.627	0.154	0.362	2.894
117	79	0.099	0.064	0.300	3.901
118	111	0.023	0.147	0.329	3.037
119	111	0.081	0.138	0.355	2.477
120	111	0.041	0.088	0.354	2.589
121	57	0.187	0.115	0.297	2.637
122	99	0.541	0.090	0.337	2.557
123	74	0.403	0.045	0.305	2.924
124	90	0.480	0.057	0.316	3.247
125	94	0.170	0.091	0.322	2.948
126	103	0.088	0.150	0.324	3.073
127	98	0.247	0.105	0.328	2.986
128	83	0.479	0.038	0.327	2.831
129	72	0.100	0.014	0.314	2.612
130	68	0.387	0.006	0.321	2.025
131	111	0.575	0.112	0.345	3.319
136	114	0.543	0.083	0.376	2.259
137	114	0.571	0.036	0.381	2.490
138	53	0.148	0.090	0.299	2.381

Correspondence Address

Michael Cysouw
Department of Linguistics
Max Planck Insitute for Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig
Germany
cysouw@eva.mpg.de