

# Similarities among languages of the Americas: An exploration of the WALs evidence

*Søren Wichmann*

(Max Planck Institute for Evolutionary Anthropology & Leiden University)

*Eric W. Holman*

(University of California, Los Angeles)

*Dietrich Stauffer*

(University of Cologne)

*Cecil H. Brown*

(Northern Illinois University)

## Summary introduction

An exploration of WALs (i.e. the *World Atlas of Language Structures*, edited by Haspelmath et al., 2005) has brought out a number of linguistic features that are significantly better represented in the New World than anywhere else. 24 such features remain when features that are restricted to certain subareas of the Americas are excluded. Two possible explanations are consistent with the available data. The features in question may be inherited from some language ancestral to all or most of the languages of the Americas. Alternatively, the features may be diffused across distances much greater than typically encountered in language areas. More data will be necessary to decide between these explanations.

## Methods and results

As a source for comparing languages of the Americas, we have been using the *World Atlas of Language Structures* (Haspelmath et al. eds., 2005, henceforth WALs). It contains data from 2560 languages and shows the distribution of various structural language traits pertaining to domains such as phonology, morphology, and syntax, which are plotted on approximately 140 world maps.

Initially we made a statistical test on the totality of the data in WALS to determine whether there was a significant difference between Native American languages and languages of the rest of the world with respect to typological features. Using a randomization technique we were able to determine that the languages of the Americas indeed do stand out, and that the probability that this could be due to chance is less than 0.0001. We next looked at the individual traits in New World languages, mostly grammatical in nature, contributing to the total difference. We uncovered a number of traits that are significantly more widespread in Americas as opposed to the rest of the world. When adding the criterion that these traits must not be found in only one part of the Americas or in adjacent parts only, the number of traits was reduced to 24. All of these 24 traits are more-or-less uniformly distributed through all regions of the Western Hemisphere. The list of traits is given in Table 1 (here “f” stands for ‘feature’ and “v” for ‘value number’ in WALS).

Table 1. 24 traits that are significantly better represented in the New World than elsewhere. OW = % occurrence in Old World sample; NW: % occurrence in New World sample; V: % variance explained

<b>Trait</b>	<b>OW</b>	<b>NW</b>	<b>V</b>
Intensifiers and reflexive pronouns are differentiated f(47), v(2)	16	81	40.8
Initial interrogative phrase f(93), v(1)	9	64	31.2
Conjunctions and universal quantifiers are formally different f(56), v(1)	30	88	22.6
Non-reduction in relativization on subjects f(122), v(2)	9	48	17.4
Epistemic possibility realized by affixes on verbs f(75), v(2)	27	67	14.1
No dominant order of object, oblique, and verb f(84), v(6)	10	45	11.9
Reciprocal construction identical to reflexive f(106), v(4)	14	46	10.8
Verb-subject order f(82), v(2)	3	29	10.0
Contrast between nasalized and non-nasalized vowels f(10), v(1)	9	37	9.2
Verbal encoding of predicative adjectives f(118), v(1)	28	59	9.1
Small vowel quality inventory f(2), v(1)	8	34	8.2
No overlap between situational and epistemic modal marking f(76), v(3)	43	72	6.8
Pronominal subjects expressed as affixes on verbs f(101), v(2) <sup>a</sup>	42	69	5.9
The verb “give” involves a secondary object construction f(105), v(3)	12	35	5.6
No voicing contrast in plosives and fricatives f(4), v(1)	27	53	5.5
Nominal plurality encoded by plural word f(33), v(7)	3	21	5.1
Case expressed by postpositional clitics f(51), v(6)	6	25	5.0
Coding of evidentiality by separate particle f(78), v(4)	5	23	4.4
Subject clitics on variable host f(101), v(3) <sup>a</sup>	0	12	4.3
No front rounded vowels f(11), v(1)	86	98	4.3
Polar questions involve interrogative verb morphology f(116), v(2)	15	36	4.1
Only inclusive distinction in verbal inflection f(40), v(4)	0	10	3.7
Prohibitive expressed by imperative plus declarative negation f(71), v(1)	12	30	3.6

<sup>a</sup> Two of the traits refer to the same feature: values 2 (subject affixes) and 3 (subject clitics) for feature 101. Cliticization may be interpreted as the original state and affixation simply as a further development.

Having ruled out chance as the origin of the New World/Old World difference, two viable explanations remained. The first of these we call the founder-effect model. According to this model the higher frequency of occurrence of certain structural traits in New World languages compared to Old World languages is due to the inheritance of those traits from a founder language or from a very small number of structurally similar founder languages ancestral to all contemporary languages of the Americas. Such a model is also becoming increasingly better supported from genetics. Recent studies show support for a single wave of migration into the New World (e.g., Stone and Stoneking 1998, Silva et al. 2002, Zegura et al. 2004). So while geneticists are increasingly moving away from the Greenbergian tripartite model they are now supporting an even more radical hypothesis entailing a single origin of all Native Americans.

The second explanation is the diffusional model. Old and New World trait frequency differences may be explained by borrowing. For example, at some point in the past a single language trait may have occurred at approximately the same frequency in both Old and New World languages. For whatever reason, in the Americas the trait could have been widely borrowed by languages lacking it from languages possessing it. If such a diffusion event did not also occur for languages of the Old World, then a situation would develop in which the percentage of New World languages in which the trait is present would become substantially greater than the percentage of Old World languages that possess it.

Diffusion is rarely studied quantitatively, so there is no available yardstick to tell us exactly how many traits it takes which are generally shared in an area and are infrequent elsewhere to argue that some common ancestor is involved. What we did, then, was to make binary comparisons involving groups of languages of several different world regions in order to look at the behavior of other areas where language families are known or at least believed not to be related. It turned out that when we compared any

particular group of languages of our WALS sample that were geographically conjunctive, e.g., languages of Asia or languages of Europe, with any other group of geographically conjunctive languages, total-data randomization tests almost always showed a statistically significant difference in the percentage of occurrence of structural traits between groups. This was true of groups of languages of both small and large geographic regions. For instance, Europe turned out to have more traits that are significantly better represented in this region than elsewhere than do the Americas. So it became apparent to us that the New World versus Old World finding was hardly exceptional, but, rather, more like the norm.

This result clearly indicates that structural traits readily diffuse across geographically conjunctive languages, even if the geographic areas are enormous regions such as the whole Western Hemisphere. Thus, we could not rule out the possibility that significant differences between Old and New Worlds in the occurrence of structural traits of language are due to diffusion of different sets of traits across the two respective gigantic regions of the world rather than to a New World founder effect.

Given the ubiquity of diffusion, the next question is whether quantitative study might produce a baseline for diffusion against which a founder effect might be discerned. In order to address this question we introduce the strategy of analyzing spatial autocorrelations (Holman et al. 2007). Spatial autocorrelation refers to a systematic decrease in the similarity among the members of a set of entities as the geographical distance increases. It is a phenomenon well known in many branches of science, including genetics (e. g., Wright 1943) and geography (e. g., Tobler 1970). Nevertheless, within linguistics at large the phenomenon has previously been studied only in dialects (e. g., Séguy 1971, Nerbonne and Kleiweg 2007) and a few closely related languages (Cavalli-Sforza and Wang 1986). In order to acquire an understanding of the factors that have an effect on the spatial autocorrelation of languages we have employed computer simulations where we varied parameters such as the rate of internal language change, the amount of diffusion, the speed of migration, and the amount of language shift. The simulations indicate that at relatively large distances, the only relevant parameter that can substantially affect the curve for spatial autocorrelation is descent from a common ancestor vs. non-relatedness. In the WALS data, empirical spatial autocorrelation curves

show much more similarity among related languages (in the same family) than among unrelated languages (in different families) at all distances. The simulations combined with the empirical results lead to the prediction that if the languages in different New World families are descended from a common founder, then their curve should show more similarity at large distances than the curve for different Old World families. A difference in the predicted direction does indeed appear in the WALS data, but it is not statistically significant.

This inconclusive result raises the question whether the WALS features are stable enough through time to preserve ancestral information since the peopling of the Americas. Our investigations of stability (Wichmann and Holman 2009) suggest that only the most stable features have much chance of retaining the same value for as long as 12,000 years. This finding implies two further predictions about New World languages from the hypothesis of common ancestry. First, the most stable features should show greater similarity between languages in different families than do the other features. A difference in the predicted direction is again observed, and again it is not statistically significant. Second, Table 1 should contain a disproportionate share of very stable features. In fact, the distribution of stability in Table 1 is practically the same as in WALS as a whole.

## **Conclusions**

Although the present results do not allow a decision about whether the similarities among New World languages reflect ancestry or diffusion, they do suggest that the question may be decidable given enough data. The computer simulations show that ancestry can be disentangled from other parameters under the right conditions. The stability data show further that the relevant time depths may be within reach of the most stable features.

In any case, the features in Table 1 are worthy of attention. If the languages of the New World share a common ancestor, then the features give the best available typological description of this ancestor. Otherwise, the features are evidence for a language area spanning two continents.

Perhaps even more important is the progress that we believe we have made towards understanding how to deal with the distribution of typological features. Various

attempts to build actual phylogenetic trees cutting across the established language families of the Americas using WALS data have been inconclusive. We believe that this is mainly due to the scarcity of data. The WALS database is very heterogeneous and only around 100 languages have been investigated for all or nearly all features. This means that the sample of Native American languages for which any sensible comparisons can be made is very small. Adding more data is crucial if we are to make further progress. Geneticists operate with a concept called ‘total evidence’, that is an approach that draws upon all available data, whether morphological or molecular. The ‘total evidence’ approach is generally considered preferable. Similarly, in comparative linguistics we should use all the evidence we have, combining lexical and typological comparisons.

Doing this in a systematic may eventually help us to unravel the mystery that the Americas are supposed to have been settled in one or just a few migrations yet they are home to more language families than any other region in the world.

## References

- Cavalli-Sforza, Luigi Luca, William S.-Y. Wang. 1986. Spatial distance of lexical replacement. *Language* 62: 38-55.
- Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.
- Holman, Eric W., Christian Schulze, Dietrich Stauffer, 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.
- Nerbonne, John and Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14: 148-167.
- Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane* 35: 335-357.

- Silva Jr., Wilson A., Sandro L. Bonatto, Adriano J. Holanda, Andrea K. Ribeiro-dos-Santos, Beatriz M. Paixão, Gustavo H. Goldman, Kiyoko Abe-Sandes, Luis Rodriguez-Delfin, Marcela Barbosa, Maria Luiza Paçcô-Larson, Maria Luiza Petzl-Erler, Valeria Valente, Sidney E. B. Santos, and Marco A. Zago. 2002. Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *American Journal of Human Genetics* 71: 187-192.
- Stone, Anne C. and Mark Stoneking. 1998. mtDNA analysis of a prehistoric Oneota population: implications for the peopling of the New World. *The American Journal of Human Genetics* 62: 1152-1170.
- Tobler, Waldo. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234-240.
- Wichmann, Søren and Eric W. Holman. 2009. *Assessing Temporal Stability for Linguistic Typological Features*. München: LINCOM Europa.
- Wright, Sewall. 1943. Isolation by distance. *Genetics* 28: 114-138.
- Zegura, Stephen L., Tatiana M. Karafet, Lev A. Zhivotovsky, and Michael F. Hammer. 2004. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Molecular Biology and Evolution* 21: 164-175.