Diffusion and Inheritance of Language and Culture: A Comparative Perspective

Eric W. Holman  
*University of California, Los Angeles*

Søren Wichmann  
*Max Planck Institute for Evolutionary Anthropology and Kazan Federal University*

Cecil H. Brown  
*Northern Illinois University and University of West Florida*

E. Anthon Eff  
*Middle Tennessee State University*

**ABSTRACT**

Using three worldwide databases, we investigate how average similarity between pairs of languages and cultures are influenced by geographic distance and time of common ancestry. Generally, the similarity between languages or cultures decreases as the geographic distance increases. This occurs even for languages and cultures without a known common ancestor, suggesting the influence of diffusion. At any given distance, related languages are more similar than unrelated languages. However, remotely related cultures are no more similar than entirely unrelated ones, indicating that inherited cultural features tend to be lost more readily over time than inherited linguistic features.

In general, the further apart languages or cultures are located geographically from one another, the fewer traits they have in common and, hence, the less similar they are to each other. This relationship has been confirmed empirically in a wide range of studies by Milke (1949), Jorgensen (1969), Séguy (1971), Cavalli-Sforza and Wang (1986), Goebel (2001), Holman *et al.* (2007), Holman *et al.* (2008),
Spruit et al. (2009), and Nerbonne (2010). Milke, Séguy, Cavalli-Sforza and Wang, and Nerbonne all fit different mathematical functions to their data, and Nerbonne observes that whatever the specific function, the relationship is generally sublinear in the sense that a given change in distance has less effect on similarity at longer distances than at shorter ones.

Three principal factors contribute to the observed relationship between distance and similarity. One is borrowing or diffusion between languages or cultures in contact, a process of horizontal transmission in which people of different languages or cultures acquire information from each other. Since people are most likely to interact with others who live nearby, diffusion and thus similarity will be negatively related to distance.

Another factor is inheritance of attributes from a common ancestral language or culture, a process of vertical transmission in which people of earlier generations pass on information to people of later ones. To show how inheritance is related to distance, Cavalli-Sforza and Wang (1986) point out that after two linguistic or cultural groups diverge from a common ancestor, they become less similar to each other with the passage of time and also more likely to have migrated away from each other.

A third factor, also indirectly related to distance, is adaptation to the local environment. Adaptation promotes similarity between cultures located in similar environments, which in turn are likely to be in geographic proximity. Although adaptation has traditionally been considered relevant to culture but not to language, evidence for an effect of adaptation on phonology has been assembled by Munroe et al. (1996), Fought et al. (2004), Ember and Ember (2007a, 2007b), Munroe and Fought (2007), and Munroe et al. (2009). Moreover, as Borgerhoff Mulder et al. (2006) mention, adaptation can influence both diffusion and inheritance, because well-adapted traits are more likely to be borrowed or inherited.

Another factor influencing similarity is random variation. Chance matches between features of some languages or cultures can contribute to the overall level of similarity found in the data, and chance can also produce random fluctuations in the curve relating similarity to distance. However, chance cannot account for the fact that similarity tends to decrease as distance increases. This
tendency is instead attributable to some combination of diffusion, inheritance, and local adaptation.

The relative contributions of these three factors to the relationship between distance and similarity can be estimated. Most research attempting to do so has applied correlational methods to cultural data. A series of studies have analyzed the material culture of villages located on the north coast of New Guinea inhabited by speakers of languages from several different (unrelated) language families; the data were originally analyzed by Welsch et al. (1992) and subsequently reanalyzed by Moore and Romney (1994, 1996), Roberts et al. (1995), Welsch (1996), and Shennan and Collard (2005). All show an effect of geographic proximity with language held constant, an effect attributed to diffusion. Depending on how cultural similarity is defined, most of the studies also find an effect of language with geographic proximity held constant, this attributed to inheritance.

Studies by Guglielmino et al. (1995) and Hewlett et al. (2002) analyze cultural data from sub-Saharan Africa in Murdock's (1967a) Ethnographic Atlas. As in the New Guinea studies, spatial proximity is used as a measure of diffusion and phylogenetic proximity in language classifications is used as a measure of inheritance. In addition, environmental similarity as inferred from vegetation maps is used as a measure of local adaptation. Each of these measures is correlated with the similarity between pairs of African societies with respect to each cultural trait. Differences among the correlations suggest differences among cultural traits in the relative importance of inheritance, diffusion, and adaptation.

Following Mace and Pagel (1994), a large number of investigations trace the histories of cultural features by mapping their evolution onto linguistic trees, mainly in order to test for correlated evolution (Pagel 1994) between cultural features. In particular, the relationships among systems of kinship, marriage, inheritance, and residence have drawn much attention (Cowlishaw and Mace 1996; Borgerhoff Mulder et al. 2001; Fortunato et al. 2005; Jordan et al. 2009; Fortunato and Jordan 2010; Jordan 2011). Mace and Jordan (2011) provide a review of this literature. Other studies have constructed trees based on aspects of material culture, such as basketry in native societies of California (Jordan and Shennan 2003) or textiles in Iran (Matthews et al. 2011). Collard et al. (2006) com-
pare phylogenetic trees constructed with the same methods for cultures and for animal species. The trees fit the data about as well for cultures as for species, suggesting that inheritance plays about as strong a role in cultural evolution as in organic evolution.

Concern has been raised that the speed of cultural change and the effects of diffusion may pose problems for cultural phylogenies (Boyd et al. 1997; Nunn et al. 2006, 2010). This issue has been addressed in an optimistic fashion by Gray et al. (2007), Greenhill et al. (2009), and Currie et al. (2010), who point to the availability of appropriate phylogenetic methods that take reticulation into account, to methods of error estimation, and to the relative robustness of Bayesian phylogenetic methods against horizontal transmission. Less discussion has been devoted to the speed of cultural change. Gray et al. (2007) describe several theoretical reasons to expect less stability in culture than in language, but they note a lack of sufficient empirical data to test this expectation. One purpose of the present study is to provide additional data on the issue.

Our study combines aspects of previous lines of research. Like Milke (1949) and others, we construct empirical functions relating similarity in linguistic and cultural traits to spatial distance; like the correlational studies, we infer inheritance from linguistic classifications; and like Collard et al. (2006), we conduct parallel analyses in two different domains, in this case language and culture.

The recent availability of two large linguistic databases allows worldwide comparisons with an older ethnographic database. To distinguish between causal factors, separate functions relating similarity to distance are constructed at different taxonomic levels of linguistic relationship. A difference between taxonomic levels at the same spatial distance can then be attributed more to inheritance or adaptation than to diffusion, and an effect of spatial distance at the same taxonomic level can be attributed more to diffusion or adaptation than to inheritance. Given that adaptation can influence both inheritance and diffusion, our methods do not distinguish adaptation from either alternative, and we will therefore not attempt to draw inferences about adaptation.

DATA AND METHODS

The linguistic data are from two different sources. One is the Automated Similarity Judgment Program (Wichmann et al. 2012, hence-
forth ASJP), which provides lexical information. The ASJP database consists of 40-item word lists transcribed into a standard orthography described by Brown et al. (2008). The 40 items are those estimated by Holman et al. (2008) to be the most stable items in Swadesh's (1955) 100-item list. The present study excludes pidgins, creoles, protolanguages, constructed languages (such as Esperanto), lists based on materials earlier than 1700 CE, and lists with fewer than 28 of the 40 items attested, leaving lists for 5,088 languages and dialects, which represent 3,581 distinct languages with different ISO639-3 designations in the 16th edition of Ethnologue (Lewis 2009).

The other linguistic database is the World Atlas of Language Structures (Dryer and Haspelmath 2011, henceforth WALS), which provides structural information on living languages. WALS contains 138 maps showing the distribution of different phonological, grammatical, and lexical features. Each feature has anywhere from two to nine discrete values or traits. For example, Map 75 refers to the feature ‘epistemic possibility’, which has three values, ‘verbal constructions’, ‘affixes on verbs’, and ‘other’. The present study draws on 134 of the 138 features, excluding those with redundant data. This study also excludes pidgins, creoles, and sign languages, leaving 2,445 languages attested for at least one of the 134 features. Languages are extremely variable in their degree of attestation: 219 languages are attested for only a single feature, one (English) is attested for all 134, and the rest are scattered across the interval between these extremes. Features are less variable, ranging from 111 to 1,369 in the number of languages attested.

Cultural data are from the Ethnographic Atlas (henceforth EA), originally published by Murdock (1967a) and subsequently augmented by Murdock (1967b, 1968a, 1968b, 1968c, 1971), Barry (1980a, 1980b), Korotayev et al. (2004), and Bondarenko et al. (2005). EA summarizes the published literature on 80 cultural features related mainly to subsistence economy, family and kinship, social stratification, division of labor, and house construction. Each feature has from two to fourteen discrete values. For example, the sixth feature is ‘prevailing mode of obtaining a wife’, which has seven values, ‘bride-price or bride-wealth’, ‘bride-service’, ‘token bride-price’, ‘gift exchange’, ‘exchange’, ‘absence of any significant consideration’, and ‘dowry’. In order for the linguistic and cultural samples to be comparable, the present study excludes societies dated before 1700 CE, leaving 1,271 societies. The degree of attesta-
tion ranges from 17 to 80 features across societies, and from 367 to 1,271 societies across features.

Both WALS and EA include language classifications; ASJP uses the same classification as WALS, which is also used in the present study because it is more recent than the one in EA. The WALS classification, described by Dryer (2011), groups languages into non-overlapping genera and genera into non-overlapping families. Genera are defined as the most inclusive groups descended from a common ancestral language spoken within the last 3,500 to 4,000 years. Families are defined as the most inclusive groups widely considered to be descended from a common ancestor, a definition that corresponds closely to Nichols's (1990) definition of a ‘stock’; Nichols estimates most stocks to be about 6,000 years old and few if any are estimated to be more than 10,000 years old. As examples of a family and genera, Dryer offers respectively Indo-European and its major subgroups such as Germanic and Celtic.

The same classification is used both for the languages in ASJP and WALS and for those spoken by peoples of the societies in EA. The 5,088 lists from ASJP are classified into 512 genera and 217 families. The 2,445 languages from WALS are classified into 458 genera and 205 families. The 1,271 societies from EA are classified into 276 genera and 117 families.

Geographic distances between languages and between societies are calculated from latitudes and longitudes provided in each database. These distances follow the shortest path on the surface of a sphere, ignoring obstacles such as mountains, deserts, and oceans.

The data are analyzed by calculating the degree of similarity between languages or societies as a joint function of taxonomic level and spatial distance. In some published studies, the results are presented in scatter plots with each point denoting a pair of languages or societies. This procedure applied to the present data would produce on the order of a million points per scatter plot. The usual solution to this excess, which we adopt from earlier studies such as those of Milke (1949) and Séguy (1971), is to divide distance into intervals and average the points within each interval. We divide distances up to 1,000 kilometers into intervals of 100 kilometers, such as 0–100 kilometers, 100–200 kilometers, and so on; and we divide distances above 1000 kilometers into intervals of 500 kilometers, such as 1000–1500 kilometers, and so on. In each interval, pairs are then sorted into one of three taxonomic levels, members of the pair being either (1) in the same genus, (2) in dif-
ferent genera but the same family, or (3) in different families. Only
combinations of intervals and levels represented by at least ten
families are analyzed further, in order to avoid unreliable results
from unrepresentative samples.

In the ASJP data, the similarity between each pair of lists is
calculated as described in Bakker et al. (2009). For any pair of
words, the Levenshtein distance (LD) is defined as the minimum
total number of additions, deletions, and substitutions of symbols
necessary to transform one word into the other. LD is then normal-
ized by dividing it by the length of the longer of the two words. For
the pair of lists, normalized LD is averaged across all pairs of
words with the same meaning shared by the two lists. To correct
for chance similarity, this average is divided by the average nor-
malized LD of all pairs of words on the lists with different mean-
ings. The result is subtracted from 100 per cent to produce the simi-
arity of the pair. Similarity of list pairs is averaged across taxo-
nomic groups at the given level in the given distance interval,
weighted by the square root of the number of pairs in each group.
The square-root weighting compensates for the fact that the num-
ber of pairs in a group increases approximately as the square of the
size of the group, while the amount of data in a group increases
only as the size of the group. In a pair of completely unrelated lan-
guages, the average normalized LD is expected to be slightly high-
er for words with the same meaning than for words with different
meanings, because of sound symbolism (Wichmann et al. 2010).
Therefore, the baseline similarity between unrelated languages is
expected to be slightly above zero.

In the WALS and EA data, similarity is based on features. For
each feature, the similarity between the pairs in a taxonomic group
at a given level is defined as the number of pairs with the same
value of the feature, divided by the number of pairs for which the
feature is attested in both members. Similarity is averaged across
taxonomic groups at the given level, again weighted by the square
root of the number of pairs in each group. Weighted average simi-
larities are averaged across all attested features and expressed as
a percentage in order to represent the overall similarity of the pairs
at a taxonomic level in a distance interval. No correction for chance
similarity is employed. The procedure used for the ASJP data would
not be effective for WALS or EA, since the frequency distribution
across feature values differs strongly between features (unlike the
frequency distribution across sounds for different word meanings).
Consequently, the baseline similarity between unrelated languages or cultures is expected to be considerably above zero.

RESULTS

Fig. 1 plots similarity between languages and between cultures as a function of spatial distance at each of three taxonomic levels, that is, within genera, across genera of the same family, and across different families. Panels A and B refer to lexical similarity between languages, with the curves for all three taxonomic levels in A and the curve for different families repeated in B on an expanded scale for better visibility. These graphs improve on Fig. 3 of Holman et al. (2008) by applying a better measure of similarity to a much larger version of the ASJP database. Panel C refers to structural similarity between languages as a function of spatial distance at each taxonomic level. This graph improves on Fig. 1 of Holman et al. (2007)

Fig. 1. Percentage similarity between languages and cultures (vertical axes) as a function of the spatial distance between them (in kilometers, horizontal axes). A and B: lexical similarity between languages. C: structural similarity between languages. D: similarity between cultures. Solid lines: different families. Dashed lines: same family but different genera. Dotted lines: same genus.
by adding the level of genera and using the latest version of the *WALS* database. Panel D refers to similarity between societies in the *EA* database.

Fig. 1 shows distance on a logarithmic scale. Nerbonne (2010) fits his data with logarithmic functions as examples of the sort of sublinear function that he proposes to approximate the effect of distance on similarity. A logarithmic function of distance corresponds to a linear function of the logarithm of distance. The functions in Fig. 1 are indeed more or less linear out to about 10,000 kilometers, although the curve for lexical similarity between languages in different families is concave-up, and all three curves for similarity between societies are concave-down.

Beyond about 10,000 kilometers, the curves for languages and cultures in different families fluctuate with no clear trend. The fluctuations at long distances probably reflect mainly chance similarity, given that lexical similarity, which is corrected for chance, is close to zero, while structural and cultural similarity, which are not corrected, fluctuate around 40 per cent. Other possible contributors to similarity at long distances are sound symbolism, loanwords from world languages such as Arabic and Spanish, cultural effects of world religions, and universal preferences for some feature values over others.

The large differences attributable to different baselines of chance similarity would confound any comparison involving the absolute levels of lexical, structural, or cultural similarity. Comparisons between curves in the same graph do not suffer from this ambiguity because they are all relative to the same baseline. One such comparison is the prediction from inheritance that the curves for lower taxonomic levels will be above the curves for higher levels. The curves for lexical and structural similarity are ordered as predicted: in each distance interval, similarity within genera is highest, similarity between genera within families is next, and similarity between families is lowest. Cultural similarity shows different results: similarity within genera is again higher than similarity between genera within families in each distance interval, but the curve for genera within families is practically superimposed on the curve for different families. The entire pattern of results suggests that linguistic inheritance reaches back to families but cultural inheritance does not extend beyond genera.
DISCUSSION

Fig. 1 shows that similarity between languages and cultures in different families decreases as distance increases out to about 10,000 kilometers. The most direct explanation for this effect is diffusion or possibly local adaptation, assuming that different families are not related by inheritance. An alternative explanation is that the present measures of similarity are sensitive to inherited relationships beyond the reach of conventional methods. For similarity between cultures, this alternative is inconsistent with the evidence in Fig. 1 that cultural similarity is not inherited beyond genera. For lexical and structural similarity between languages, Fig. 1 indicates inheritance beyond genera within families and thus does not exclude inheritance beyond families. However, Wichmann (2013) finds that the present measure of lexical similarity (ASJP) generally becomes unreliable at time depths greater than about 5000 years; thus, the lexical similarities are not sensitive to inheritance between families unless the families are relatively young. As for structural similarity, Greenhill et al. (2010) and Wichmann and Holman (2009) find that the rates of change in the structural features of WALS are on the same order as rates of lexical change. These results therefore favor diffusion over inheritance as an explanation for the effect of distance on linguistic similarity between families in Fig. 1. This explanation is consistent with Dryer’s (1992) suggestion of language areas as large as continents, and also with the large culture areas inferred from EA data by Burton et al. (1996) and further discussed by Korotayev and Kazankov (2000) and Jones (2003). Moreover, the absence of an effect of distance beyond about 10,000 kilometers is consistent with size of the proposed language and culture areas.

The present evidence for linguistic inheritance within genera and also between genera within families is consistent with the fact that language genera and families are themselves inferred in part from lexical and structural information of the sort contained in the ASJP and WALS databases. For cultures, the present results confirm, with completely different methods of analysis, the finding of most correlational studies that cultural similarity depends on linguistic affiliation as well as geographic proximity. In particular, there is evidence for inheritance of cultural traits within genera, but little or
none beyond genera even within families. Inherited cultural features, then, tend to be lost more readily over time than linguistic features.

The finding that cultural inheritance is limited in time provides empirical support for the theoretical expectation of Gray et al. (2007) that culture is less stable than language. This also suggests that the technique of mapping cultural features onto linguistic trees will be the most informative about cultural history within language genera, rather than at greater time depths. Although cultural inheritance is weaker than linguistic inheritance in the present study, it is about as strong as biological inheritance in the study of Collard et al. (2006). Two reasons can be offered for this contrast. First, most of the cultural data sets analyzed by Collard et al. include representatives of only one or a few genera; cultures this closely related can be expected to show a substantial effect of inheritance according to Fig. 1. Second, the conventional identification of cultural inheritance with language classifications may be an oversimplification; the cultural phylogenies inferred by Collard et al. may include cases where language and culture are inherited along separate pathways.

In summary, for both languages and cultures similarity decreases as distance increases. This is so even for languages and cultures without known common ancestors, suggesting the influence of widespread diffusion or possibly local adaptation. At any given geographic distance, similarity decreases with increasing remoteness of common ancestry up to the time depth of families for languages, but only up to the time depth of genera for cultures. The present findings, then, suggest that languages are more conservative than cultures.

ACKNOWLEDGMENTS

We thank Bernard Comrie, Malcolm M. Dow, Dietrich Stauffer, and anonymous referees for helpful suggestions on earlier versions of this paper. Wichmann’s research was funded by an ERC Advanced Grant (MesAndLin(g)k, Proj. No. 295918) and by a subsidy of the Russian Government to support the Program of Competitive Development of Kazan Federal University.

REFERENCES


