Contrasting linguistics and archaeology in the matrix model: GIS and cluster analysis of the Arawakan languages

Gerd Carling, Love Eriksen, Arthur Holmer and Joost van de Weijer

1. Language contact, linguistic distance and the emergence of socio-cultural identity: The matrix model

Scientific research on the pre-Columbian cultures of lowland South America has taken a great leap forward during the first decade of the new millennium. In terms of material culture, the excavations of large-scale settlements with Amazonian Dark Earths (ADE), more commonly known as terra preta soils, indicating the remains of large-scale systems of indigenous intensive agriculture (Lehmann et al. 2003; Glaser and Woods 2004; Woods et al. 2009), and the recent discovery of the huge earthwork complex in the Brazilian state of Acre where so far more than 250 geometric figures, many of them measuring 200–300 meters across (Schaan et al. 2007; Mann 2008; Saunaluoma 2010), are just two examples of archaeological discoveries of great significance for our view of pre-Columbian Amazonian societies. At the same time, linguistic research, aided by new developments in software technology and computer capacity, has moved toward large-scale comparison projects where new ways of measuring linguistic distance are supplying new evidence of genetic relationships as well as contact scenarios (Muysken 2011; Danielsen et al. 2011; Walker and Ribeiro 2011; O'Connor and Muysken 2012). Seen as a whole (a perspective strongly advocated by the current investigators), these discoveries calls for methodological advance in order to integrate the results from different academic disciplines into a more comprehensive picture of pre-Columbian Amazonian societies and their mutual interaction.

The great development in the sciences devoted to the study of the material cultures and languages of indigenous Amazonians has also demanded new theoretical models for the socio-cultural expansion of the large-scale, contact-seeking, pre-Columbian societies of the region. Several attempts at synthesizing this knowledge have also appeared (Hornborg 2005; Eriksen 2011; Hornborg and Hill 2011), all advocating the need for an integrated
perspective, joining together expertise from archaeology, anthropology, and linguistics in order to understand the development of material culture, socio-cultural organization, and language as a whole. It may sound like an understatement to say that the need for an integrated perspective on these research questions stems from the fact that language is always used in a cultural context and that the study of a particular language therefore always should include the study of the cultural context in which the language is used, but since this is not always the case, we wish to emphasize that this is exactly what the present study is devoted to doing. We therefore constantly seek to contrast the results of our linguistic investigations with those of archaeology and anthropology in order to acquire a more comprehensive picture of the cultural and linguistic development in pre-Columbian Amazonia.

As for Arawak, the language family to which this study is mainly devoted, one attempt to synthesize the current knowledge across scientific disciplines appeared in the publication Comparative Arawakan histories: Rethinking language family and culture area in Amazonia (Hill and Santos-Granero 2002), in which a cultural matrix model in which socio-cultural and linguistic features of Arawak-speaking societies in space and time were joined together into a cultural matrix composed of the following elements:

- suppression of endo-warfare,
- a tendency to establish socio-political alliances with linguistically related groups,
- a focus on descent and consanguinity as the basis of social life,
- the use of ancestry and inherited rank as the foundation for political leadership, and
- an elaborate set of ritual ceremonies that characterizes personal, social, as well as political life (Santos-Granero 2002: 42ff).

Furthermore, Eriksen (2011: 9) defined a complementary set of features related to the material culture typical of Arawak-speaking societies:

- the use of various types of high-intensity landscape management strategies as the basis of subsistence (Hill 2011),
- a tendency to situate their communities in the local and regional landscapes through the use of such techniques as "topographic writing", extensive systems of place-naming, and rock art (Santos-Granero 1998),
- an elaborate set of rituals including a repertoire of sacred musical instruments and extensive sequences of chanting, often performed as part of place-naming rituals (Hill 2007),
- a proclivity to establish settlements along major rivers and to establish trade and other social relations through river transportation (Hornborg 2005).

In a cultural matrix model, the matrix feeds information to its users by acting as a cultural backbone and the features of the matrix are displayed by the users in order to communicate the ethnic identity and distinctiveness of the group vis-à-vis other ethnic groups when in contact with them. Viewed as a whole, the Arawak cultural complex formed an integrated unit in which material as well as non-material culture and language helped construct and spread a cultural package across Amazonia with the present-day distribution of Arawak-speaking societies as the persisting result.

Since a majority of the features of the Arawak matrix can be related precisely to situations of social (and linguistic) contact (e.g. trade, ceremonies, political leadership), it is impossible to overlook the importance of the connection between language, material culture and ethnic identity when analyzing the spread of linguistic features among Arawak languages. We therefore view language as an integrated part of the ethnic identity of Arawak-speakers and carefully take the importance of ethnic identity into consideration when evaluating the mechanisms of language contact involving Arawak languages. Finally, we also believe that the powerful cultural matrix of the Arawaks and the strong ethnic identity derived from it was an important element in the vast expansion of Arawak languages across South America and the Antilles during the pre-Columbian period.

2. Implementing the model — general problems

The basic aim of the current project is to develop tools for the implementation and interpretation of the culture matrix model into an existing set of data. The data from the region in question — both in terms of cultural and linguistic data — is indeed huge, which requires sophisticated tools for handling and organizing it.

Since the matrix model relates both to material and non-material culture as well as language as important markers of ethnic identity, the data collected and contrasted consists of material and non-material culture, as well as language. At present, a selected number of relevant items of both types will be contrasted (cf. sections 4–5).

First of all, the implementation of a general, theoretical framework on a set of data like this is typically connected to problems of defining boundaries.
This is in particular relevant when it comes to defining or narrowing down data into standardized data formats, which is a necessary prerequisite for applications of the type that is being done in the current project. The problems appear at all levels: on the one hand when a complex setup is simplified and narrowed down to a simple Y/N alternative or when complex patterns have to be simplified or generalized in order to be squeezed into a single formula, and on the other hand when the outcome of the computer generated applications is open to a number of alternative interpretations.

In the current project, the main focus has been on the tools and methods of implementing the questions arising from the culture matrix model described in the previous section. When data is being collected, analyzed and projected onto applications two basic parameters are taken into fundamental consideration: expected stability and expected contact sensitiveness. With stability we mean the probability of the system remaining stable without changing, with contact sensitiveness we mean the probability of being substitutions arising in a contact situation (see further 2.1–2.2.). Prerequisites for the applications are that data are a first analyzed and grouped according to these parameters and thereupon projected against each other, using a layering technique, which will be described further in section 3.

A necessary basic tool for investigating and understanding contact-related features as well as the spread of features over large areas is constituted by GIS-technology, which will be described further under 3.1. Cladograms, NeighborNets, and cluster analysis can give information about a number of correspondences, depending on the nature of the input data and the methods of applying them. Cladograms in the form of trees basically give information on linguistic distance and sub-grouping of dialects as well as dialectal nodes (cf. 3.2.). Unrooted trees (NeighborNets), which are favored in a number of current studies of the Arawakan family (e.g., Walker and Ribeiro 2011; Danielsen et al 2011), both group languages into subgroups and show linguistic distance, as do (rooted) trees, but they are also able to give information on shared similarities, depending on either retention or contact.

2.1. Calculating stability in language – how could it be done?

For any linguistic property shared across two or more languages, be it a lexical item, a morphological form, a construction or a general principle of structure, there are logically speaking four possible reasons for the similarity: (a) shared inheritance; (b) loans under contact; (c) shared language-internal typological pressure (as expressed in universals or universal tendencies); and (d) coincidence. Depending on the factors underlying the shared property and the type of property involved, we speak variously of areal features (loans of structural properties), calques (loans of constructions), lexical loans, universals (e.g. grammatical properties resulting from communication requirements), core vocabulary etc.?

Under the assumption that all similarities derive from one of these four factors, it remains an empirical issue to determine which types of property are more likely to derive from which factor. Which properties are more likely to be diffused by areal contact, which tend to survive unchanged across time (i.e. to be inherited diachronically from earlier stages of a language), and which are susceptible to typological pressure? We will discount the fourth alternative, coincidence, here, partly because it is theoretically uninteresting, and partly because we assume that it will behave like random noise across the various categories, and will therefore not be expected to be a major confounding factor either.

The concept of diachronic stability is therefore the propensity to be inherited, i.e. to resist both contact-induced change and typologically induced change, while the concept of contact sensitivity is the propensity to be loaned across language boundaries. That these two concepts can be addressed separately and are not in a directly inverse relation is due to the fact that a feature which is neither contact sensitive nor diachronically stable might be particularly sensitive to typological pressure. Examples of typological pressure include the cross-linguistic correlation between SOV and overt case morphology (Greenberg 1963: 96).

As for vocabulary the situation is somewhat different. Here typological pressure is not directly applicable, so stability and borrowability are more or less in an inverse relationship: The core vocabulary is a feature with typically high stability and low borrowability, while cultural vocabulary is typically more unstable and contact sensitive (cf. Tadmor and Haspelmath 2009). The whole concept of glottochronology of lexicostatistics is based on this assumption, namely, that certain parts of the lexicon are more stable for classification (Swadesh 1952; Greenhill et al. 2010).

2.2. Calculating stability in material culture

The four different categories used for explaining linguistic similarity that were outlined in section 2.1 can also be used for describing the occurrence of similar forms of material culture occurring at two or more geographical locations. In terms of “translating” the scenarios of linguistic contact into those
of material culture, categories a, b, and d (shared inheritance, loans under contact, and coincidence) are relatively self-explanatory also in the context of material culture, while category c, shared language-internal typological pressure, is translated into an artifact’s use-value or cultural affordance; i.e., its ability to retain its position in a given society despite constant technological development occurring through time (e.g., an artifact such as the scythe has greatly decreased in terms of numbers and importance in Europe due to the mechanization of agriculture during the 19th and 20th century; its ability to fit into the new conditions is low, and its use will therefore decrease through time).

Although material culture may behave differently from language in terms of its tendency to be transformed in contact scenarios, different types of material culture can nevertheless be measured according to its tendency to show stability vs. contact sensitivity in different social contexts. An important point to mention here is that although the material aspects of artifacts may be preserved for almost indefinite time (archaeologists may recover stone tools dating back millions of years back in time), their cultural context regulating e.g., meaning, techniques, functions, etc. can easily be lost if the artifact is separated from its cultural context or if the cultural context censes to exist (as in the example with the scythe above). The use of a particular artifact and the raw material from which it is made may also change significantly over time, e.g., in the case of the large amounts of gold that was stolen by the Spanish from the Incas, taken to Europe and transformed into coins and objects with new social functions that today are being used to manufacture computers, items of high significance in Western society, but nevertheless very different from the role of the Inca goldwork in pre-Columbian South America.

Apart from the extremely abrupt changes caused by the European colonization of the New World, Amazonian material culture displays many interesting examples of stability as well as contact sensitivity among its material cultures, some of which will be discussed in greater detail in this section. Overall, the ceramic material of pre-Columbian Amazonia (ceramic items generally being the most well-preserved artifacts of Amazonian archaeological contexts) show long temporal continuities, stretching millennia back in time. The oldest pottery of the New World is found along the lower Amazon River (Roosevelt et al 1991) and it spread rather extensively in northern South America already during 5000–6000 BC (Eriksen 2011: 228f). It is likely that once the knowledge of manufacturing solid ceramic vessels of good quality had appeared and been diffused across the continent, the technique was not reinvented again, but passed on to new potters from generation to generation. This would indicate a high degree of stability in the ceramic technology itself, and traces of contact sensitivity should therefore be sought for in such features as the design and composition of the ceramic material.

Interestingly, while several Amazonian ceramic traditions, e.g., Zoned-Hachured, Barrancoid, Saladoid, and Amazonian Polychrome, spread across vast areas during timespans of up to 2000 years, some ceramic traditions, e.g., Camani, Paredão, and Cedeñoïd, continued to be locally produced during almost equally extensive time spans (Eriksen 2011). The expansion of the major ceramic traditions in Amazonia was most likely conducted through indigenous exchange systems (as opposed to migratory spread) (Hornborg 2005; Eriksen 2011), resulting in the adoption of particular ceramic traditions by new ethno-linguistic groups through contact (e.g., trade, ritual exchange, etc.). This indicates a high degree of inherent stability in the material culture, but also, to a certain degree, contact sensitivity, because the major ceramic traditions were split up into many subtraditions (so called ceramic phases) that were locally or regionally manufactured, but with a certain set of features still in common with the overarching ceramic tradition to which they were associated.

In contrast to the widely distributed ceramic traditions, characteristic locally produced ceramic styles were maintained by some groups for millennia without adapting the style of their pottery despite long periods of contact and outer pressure from major ceramic tradition. One such example is the Camani phase of the Araracuara area of Southeastern Colombia, which despite close geographical contact with groups manufacturing Barrancoid and Amazonian Polychrome pottery for at least 600 years maintained their distinctive ceramics, until around 1000 AD when they finally shifted their ceramics (Eriksen 2011: 175). Another example is the Paredão phase (AD 700–1200 (Eriksen 2011: 99)) of the middle Amazon in which locally produced ceramic style was maintained for more than 1500 years despite close geographical proximity to the Barrancoid and Amazonian Polychrome traditions. A third case of this type of scenario is the conservatism of the makers of the Cedeñoïd ceramics of the middle Orinoco Valley, who started to produce a distinctive kind of ceramics dated by C14-sampling to 1050 BC (Zucchi et al. 1984) and kept on doing so up until the time of European colonization (for further discussion, see Eriksen 2011: 232).

These three cases may serve as examples of stability in the archaeological material, while the spread of the major ceramic traditions indicate a high degree of contact sensitivity among the material cultures of the ethno-linguistic groups that shifted their traditional wares to the new ceramics of
the major traditions. However, it should be pointed out that the spread of major ceramic traditions took place through processes of transformation, whereby new ethno-linguistic groups were incorporated into the major cultural complexes through the adaption of their material culture (and in many instances also their non-material culture and language) to those of the major traditions. Such shifts also affected the style and composition of the major ceramic traditions through a recursive process feeding back traits from the newly incorporated ceramic styles into the previously established tradition, a phenomena indicating a certain degree of contact sensitivity among the major ceramic traditions of Amazonia. This ability to transform and to be transformed also indicates that a certain degree of contact sensitivity was advantageous in terms of successfully navigating in the socio-cultural landscape of pre-Columbian Amazonia.

3. Methods and materials – a description

In this section, methods and materials used in the present project will be described. First of all we use GIS-technology, then cluster analysis and finally a combination of the techniques, which is described under 3.3. The organization of the database and problems connected with the organization is described in 3.4. Applications and evaluations are given in section 4.

3.1. Geographical Information Systems

Although GIS, Geographical Information Systems, trace their background to physical geography and the natural sciences, its applicability in the humanistic and social sciences is great, and its potential in linguistic research has perhaps been overlooked so far. However, new internet-based applications such as The World Atlas of Language Structures (WALS) Online (http://wals.info/), allowing anyone from interested laymen to academic scholars to search and display the distribution of linguistic features around the world, are currently breaking new ground in the spatial dimensions of linguistic research. While WALS Online effectively shows the great potential of what the results of a spatial analysis of linguistic features have to offer to linguists, another motive for using GIS in linguistic research would be in order to open the field up for transdisciplinary comparisons with spatially distributed data derived from other scientific disciplines. Since spoken languages are always situated in a socio-cultural and ecological context, spatially distributed data from archaeology, ethnography, economy, and ecology, may tell us much about the context in which the language is spoken, and hence about the external factors influencing language use and language contact.

GIS offers a tool for transdisciplinary investigations of the interplay between language, culture, and materiality. In the present investigation, GIS is used for researching the socio-cultural context in which Arawak languages have expanded across Amazonia, and to investigate how linguistic distance has been affected by contact and isolation between groups of speakers in different parts of lowland South America during the past 2,000 years. More concretely, we are using spatially distributed data from linguistics, archaeology, ethnography, etnohistory, and physical geography in order to investigate the relationship between languages of the Arawak family and to map out the features of the Arawak cultural matrix (see section 1). Since a given language is used in a particular socio-cultural context and at defined spatial locations, GIS either allows us to investigate other types of data related to the socio-cultural context of the language (e.g. a specific tribe and its material culture) or the distribution of other types of data at the specific location in which the language is currently spoken (changes in the historical/archaeological data or alterations in the physical geography of a language setting). Such analyses have been made possible through large-scale data collection in the project “Nature and Culture in Prehistoric Amazonia” (Eriksen 2011), in which large amounts of transdisciplinary data with a particular focus on features relating to the Arawak language family and the Arawak matrix have been collected and analyzed, allowing for broad-scale comparison with linguistic features in space and time.

Technically speaking, the interplay between the linguistic and the socio-cultural/ecological data stored in a GIS takes place through the linking of our SQL-database (3.4) containing linguistic features, to the GIS database containing socio-cultural features for the same groups, and geographical coordinates for them. Due to the nature of the construction of these databases, using a common denominator, e.g., an identity code (in this case the names of the ethno-linguistic groups), the two data sets can easily be linked together through a so called “join operation”, resulting in the so called combined applications (see 3.3.). This operation immediately allows the two data sets to share all features – linguistic as well as non-linguistic – with each other in a new, spatially distributed data set. The operation results in a new data set in which the spatial distribution of linguistic features can be examined and compared to the spatial distribution of the socio-cultural features of the Arawak matrix, thereby allowing for true transdisciplinary work to take place.
3.2. Cluster analysis of cladograms

In the present study, we have performed a cluster analysis of 100 basic lexical concepts (including pronouns, adjectives, nouns, verbs). The data has been provided by Walker and Ribeiro (2011). For these concepts, the number of different lexical variants range from 1 to 16, yielding a total of 694 variants. Cognates are identified in 60 languages. These are entered into a data set that indicates whether a variant occurs in a language or not. We have created a distance matrix for binary data and performed the cluster analysis (hierarchical cluster analysis, divisive approach) on this matrix. The results are represented by the tree diagram in figure 1. On the whole, the results of the analysis are rather similar to those reported by Walker and Ribeiro (2011), that is, that languages which shared many variants in our study were also similar in theirs.

We have looked for a reasonable division of the clusters which was not too general on the one hand, and not too detailed (e.g., yielding clusters with only one language) on the other hand. Trying out different numbers of groups, we finally decided on dividing the clusters into 10 groups as indicated by the grey rectangles in figure 1. The reason for selecting 10 groups is described in more detail in section 5: most important would be that the difference between clusters A and B becomes apparent first from 8 clusters (i.e., not yet at 7 clusters). For the application, see figure 2.

3.3. Combined applications and layering technique

In the current project, the main focus is laid on combined and layered applications (cf. 3.1). The organization of the data into the Amazonian database (see 3.4) is adjusted towards maximum applicability of these applications. The following combined applications are made possible:

- Cultural/archaeological data are projected onto:
  - Maps, showing geographic distribution of features in time and space
  - Maps with linguistically derived information
- Linguistic data (typological/lexical) are projected onto:
  - Maps, showing geographic distribution
  - Cladograms, showing linguistic distance and subgrouping
  - Combinations of maps and cladograms by means of cluster analysis, showing linguistic distance/subgrouping in relation to geographical distance.

Figure 1. The language tree with the ten clusters (A–J).
The layered applications (see section 5) make possible a further step of combining features: here, combinations derived from language data could be projected onto maps with cultural/archaeological data, either general or more specific. Individual features of interest could be projected onto groups or clusters of features. The possible combinations and projections of general and individual data in the database are in principle endless.

3.4. The Amazonian database – a short description

The "database" used in the current description consists in fact of two physically separated databases, based on the same identification system: one database with archaeological/cultural data (basis for Eriksen 2011) and one with linguistic data. The underlying identification system is based on labels of "ethno-linguistic" groups, given a certain identification number, which is complemented by a series of GIS-coordinates (see 3.1). These ethno-linguistic groups are based primarily on linguistic descriptions (see Eriksen 2011). Since the Amazonian situation is particularly complex with alternative names of groups, mismatches between descriptions (anthropological/linguistic) and identical or similar names of different groups can occur, the database distinguishes several columns with alternative names beside the standardized name, used in applications. Reference columns are also inserted. The database contains a huge number of groups from 20 different families and a number of isolates. After the id-number and name columns follows a column with linguistic family affiliation, e.g., Arawak, Tupi, Carib, based either on resources like Ethnologue or individual descriptions (Payne 1991; Aikhenvald 1999). Thereupon follow the columns with the various data, lexical or typological. Each feature (lexical/typological) normally requires three or four columns, e.g., for typological data: (1) actual data (including internal variations), (2) standardization for the purpose of applications and (3) references, or, for lexical data: (1) actual forms (including variation), (2) cognate identification or reconstruction of forms, (3) references.

4. The applicability of language data: lexical and typological features compared

4.1. Lexical data

The lexical data in the project are selected and applied in accordance with two basic parameters: (1) expected stability versus contact sensitivity, (2) correlation with the features distinguished in the matrix (see section 1). Both these parameters might be problematic, in particular if seen as oppositional. At a general level, studies of typology of borrowability like Haspelmath and Tadmor (2009) are indeed useful, but at a regional level, a number of separate factors have to be taken into consideration before stability or contact sensitivity is calculated. If the definition of a matrix is regarded as a set of features typical for the Common or Proto-Arawakan linguistic and cultural area, then several matrix-related items could count as belonging to a "stable" part of language. An item reconstructed as belonging to the basic vocabulary of one language family could be an outsider in another language family, depending on the ecological habitats of speakers. Further, the probability of borrowing, even if the item belongs to the reconstructable proto-vocabulary, is there in any case, in particular in the Amazonian region, where intra-linguistic cultural borrowing seems to be very frequent (Payne 1991; Danielsen et al. 2011) and the huge impact of colonizing languages (Spanish, Portuguese) blurs the patterns (cf. Brown 1999). A further problem (which counts for Arawak as well as other Amazonian languages with no old written sources) is constituted by the great difficulties in defining relative chronologies, a prerequisite for distinguishing true cognates from early loans in cultural vocabulary.

However, the layering technique, enabled by the GIS-technology, opens a number of possibilities. Here, individual items of interest, be it crops, kinships terms, flutes or names for various animals, could be either lumped with other items of the basic vocabulary and used in clusters, or projected (against maps) either individually or in groups of relevance.

In this study, two different methods of reconstruction are used. For the purpose of creating cladograms and clusters, we identify cognates in order to enable Y/N or numerical marking. For the purpose of GIS applications of individual items, we produce ad hoc reconstructions, in which not only cognates are identified, but the actual reconstructed forms are also presented (cf. the method by Payne 1991). This kind of reconstruction can be done within a well established language family by looking at the actual forms of ascertained semantic notions (e.g., maize, manioc) and excluding speculative
etymologies. Expected paths of sound change, both general (cf. Campbell 2004: 40ff) and language-specific (cf. Payne 1991) are then taken into consideration by comparing forms of languages and setting up a tentative tree of forms. This type of reconstruction can, to a certain extent, distinguish more preserved forms from more changed, as well as to a certain extent - identify sub-nodes, but it can hardly distinguish true cognates from early loans. For this a greater amount of data for comparison would be required. In figure 2, an ad hoc-reconstruction of words for “maize” in a number of languages of various families (data basically from Birket-Schmidt 1943, see also Brown 2006) has been worked out (fig. 2). The various forms in these languages can be boiled down to 19 reconstructed forms plus a number of isolated forms (not included). Of these forms, several can be reconstructed into sub-nodes (*mahi ki, e.g., Pallikur mahiki, versus *makana si, e.g., Mandahuasha maka na zi and *awa ri, e.g., Manao aw aty, and *a'na ti, e.g., Chamicuro n'atši).

Figure 2. The spatial distribution of reconstructed maize roots in Amazonia.

4.2. Typological data

What has been described in the previous section for the lexicon holds equally for typological data: the basic dichotomy concerns the distinction between stability and contact sensitivity. However, there is an important difference between lexicon and typology: while the lexicon of a language is to a large extent arbitrary, and can display almost limitless cross-linguistic variation, the structural properties of language are (a) much more restricted and (b) often functionally motivated. This has important consequences for our investigation. For one thing, given that the number of possible options for any typological feature is highly limited, chance similarities become a potential problem: if two languages share SVO word order this is very possibly coincidental. Even highly marked combinations such as VSO word order and postpositions could surface independently as a matter of chance (as indeed appears to have happened: the combination is found in both Guajajara, cf. Harrison 1986, and Yagua, which are spoken at more or less opposing ends of the Amazonian region, and there is no evidence of genetic relationship between them: Yagua is a Peba-Yaguan language and Guajajara is a Tupi-Guarani language).

Further, structural properties are more likely to be affected by internal typological pressure: if a language displays one property along one dimension, it is also more likely than otherwise to display another, seemingly unrelated property. Going back all the way to Greenberg (1963), one such correlation is that between VSO word order and the existence of prepositions, another is the tendency for SVO languages not to have a case system (and even stronger, the near-universal tendency for SVO languages not to have an ergative case alignment). We are hesitant to use the term “universals” coined by Greenberg, since few, if any, of the generalizations are actually universal. However, the idea of typological pressure expresses the same intuitions. In situations dealing strictly with word order, it corresponds to Hawkins’ (1983) concept of cross-categorial harmony, and otherwise more closely to Nichols’ (1992: 100ff) discussion of typological correlations. We assume that consistent correlation patterns which are statistically significant and attested cross-linguistically reflect a tendency towards maximal efficiency and economy in the language system. We also assume that the tendency itself is universally valid but may be overshadowed locally or regionally by other factors.

The typological issue is particularly a concern if a set of languages shares a number of unusual features, in which case it might be tempting to posit areal contact based on only typological evidence. However, the set of features might equally well be the result of one feature being shared by coincidence,
and a whole cluster of concomitant features arising due to typological pressure exerted by this one chance feature.\textsuperscript{8}

Given typological properties shared across languages, and again assuming that coincidence will at most represent troublesome noise in the data, we are faced with the issue of whether these properties are the result of inheritance, areal contact or typological pressure, and which properties are most likely to be the result of which factor.

Previous studies have focused primarily on the borrowability of typological features (cf. Aikhenvald 2002: 12 and works cited therein) and diachronic stability (cf. in particular Nichols 1992: 167ff.). Both concepts are highly relevant to our study, but the contribution of the present approach will be to focus on teasing apart typologically motivated characteristics and areal features by systematic comparisons of linguistic data with attested contacts (both contemporary and historical), as evidenced by the spread of cultural artefacts and culture words, supplemented by archaeological data.

Among typological features, some appear to be genetically more stable than others (Nichols 1992: 167ff.). In particular, morphological complexity and case alignment are properties which are shared across members of the same genetic phylum (and thus presumably diachronically relatively stable), head- and dependent marking are less so, whereas, at the other extreme, word order is more likely to be an areal feature.\textsuperscript{9} There are differences in this respect between different parts of the world, but unfortunately, for sampling reasons, Nichols (1992) does not include Amazonian data in the relevant section. However, Aikhenvald (2002: 102) observes that in Tariana, an Arawakan language areally influenced by Tucanoan languages, case alignment to a certain extent follows the inherited Arawakan split-ergative / active pattern, whereas Tariana shares with Tucano a mixture of head and dependent marking (other Arawakan languages are more consistently head marking).\textsuperscript{10} Further, Tariana is more typically Tucano-like in being strictly postpositional and displaying a tendency towards verb-final structure (ibid. 167).

Note that the relative genetic stability of case alignment does not imply total stability: from various language families, we have evidence of languages switching case alignment: e.g. from the Kartvelian split-ergative alignment found in Georgian, Laz has developed into a fully ergative language, while Megrelian has developed into a fully (marked) accusative language (Harris 1985: 26ff); in Indo-European, it is a matter of debate whether the proto-language was active or accusative (cf. Drinka 1999), but the family as such contains (partially) ergative, active and accusative languages.

Thus, word order and case alignment appear to be the features which are maximally distinct when it comes to genetic stability versus contact sensitivity. It is therefore important to point out that these two features are also typologically very closely connected. Here we should expect to find cases where word order may have changed due to areal influence, while case alignment is genetically preserved, leading to a mismatch between word order and case-alignment. How this tension is resolved in various languages is an interesting issue.\textsuperscript{11}

In the present database we include word order at various levels in the clause (adpositions, noun-modifier and clause-level) and case-alignment of different types,\textsuperscript{12} various kinds of subordination construction, polar and content interrogativity as well as various properties of the verbal paradigm. One aim is to chart in greater detail the various degrees of genetic stability and contact-sensitivity displayed by different typological features, as well as to examine how well the general tendencies shown by Nichols carry over to a specifically Amazonian setting, with its very special characteristics.

As a principle, we expect typological pressure to be universally valid.\textsuperscript{13} Insofar as certain geographical areas seem on the surface to be partially exempt from this, i.e. to be rife with exceptions to generally valid cross-linguistic generalizations (impressionistically, Amazonia seems to be a case in point, although this remains to be confirmed), we must conclude that some other factor may be overriding typological pressure.

One factor which might be involved is general instability in the language systems caused by widespread language attrition in various populations, partly today, but quite probably also in connection with the sudden demographic changes caused by European contact: between disease and genocide, European contact led to the untimely death of more than half of the population of the American continent, even according to cautious and conservative estimates (Cook 1998 and sources cited therein), and the aftermath led to a diaspora of various peoples scattered across the whole of Amazonia.

Further, given the principle of linguistic exogamy found in various Amazonian populations (Hornborg 1988; Aikhenvald 2002: 22), multilingualism is widespread, and structural loans across typologically diverse languages could possibly lead to unusual combinations of features surfacing within a single language.

Whatever constellations of features have appeared as a result of this, it is possible that the time elapsed since then has not been sufficient for typological pressure to act on the languages. To evaluate this account would require some kind of knowledge of the rate of change of the typological features involved.

A final, more speculative, hypothesis might derive inspiration from Barth's (1969) seminal study on the construction of ethnic identity and ethnic
boundaries, and propose that linguistic properties can be actively exaggerated by speakers to emphasize ethnic distinctions in a multilingual setting. If this is the case, it should be possible to show independently that certain linguistic properties are consciously accessible to speakers.

5. Results of combined applications

The interpretations proposed by the current investigation take as their point of departure the discussion of expected stability versus expected contact sensitivity among the linguistic and cultural features. To start with, Arawak languages display a certain degree of overall stability in the sense that they still share some overarching genetic similarities despite a very large degree of separation in space and time. Furthermore, Arawak cultures also share similarities through the Arawak matrix, indicating that certain cultural features (both material and non-material) also show considerable stability over time. This section seeks to outline which features are subject to a high degree of stability, and which features are to a greater degree contact sensitive.

In terms of the language tree (fig. 1.) and its clusters, it is in itself indicative of a certain kind of macro-stability given that the basic vocabulary of the languages are still similar to a certain degree despite considerable spatial and temporal division between many of the languages of the tree. The fact that the languages do cluster may indicate that at the micro level, e.g. in a given spatial area, a certain kind of contact sensitivity contributed to making the languages more similar through areal contacts among the languages within the respective clusters. However, the purpose of using the most stable part of language, i.e., basic vocabulary, for clustering, is precisely to being able to contrast contact-sensitivity to stability at a micro level.

The parallels between linguistic and geographical clustering are confirmed by figure 3, showing that the clusters in the language tree display themselves in particular geographical settings according to their position in the language tree (this is not in any way implied in the cladistics analysis, which only indicates the degree of separateness in terms of basic vocabulary and thus contains no spatial component whatsoever).

Figure 3 shows, (a) that lexical similarities reflect spatial closeness within the clusters, but (b), that spatial separation is not a direct reflection of lexical distance (cluster B is located far from cluster A and C, the clusters with which it shares basic vocabulary most closely). This finding is indicative of the fact that other factors besides geographical distance are influential in the distribution of basic vocabulary among members of the Arawak family.

The fact that cluster B is geographically located far from cluster A and C is noteworthy, particularly in the light of the proposed north-south split in the Arawak languages suggested by Aikhenvald (1999). This finding is confirmed by the interpretation of Walker and Ribeiro (2011: 3), who provided the original data set on which this conclusion is based.

Furthermore, the spatial analysis of the language tree identifies cluster I and J at two opposite ends of the continent. It is interesting to see that these two clusters share close lexical similarities despite no indications of socio-cultural contacts between speakers of these languages for at least 1.5 millennia (Eriksen 2011). Regarding the spatial distribution of clusters A, C, and D, their positions are located at reasonably expected distances from each other given some kind of geographical diffusion process from a common proto-language. This is also the case for clusters E and F, and in the case of Apurinã, Machinere, Itapari, and Piru of cluster E, it indicates that these languages rather share a common history with Marawά and Warakú, the two other languages of cluster E, and ultimately with the Arawak languages of the northwest Amazon, than with the languages of cluster I; Asháninka,
Caquinte, Machiguenga, Nomatsiguenga. This finding is of course noteworthy given that these groups are located in relative geographical proximity to each other, but not surprising since the historical amnesia among the Arawak-speaking groups of cluster E regarding their common ancestry with the languages of cluster I has been attested by ethnographers for quite some time (Gow 2002: 153).

In terms of non-linguistic features, many of the cultural features of the Arawak matrix show remarkable temporal stability, a fact that has allowed previous research to reconstruct an Arawak-mediated regional exchange system as far back as 900 BC in the northern part of the continent (Eriksen 2011). An interesting observation that can be made in this context is that the present-day pottery of the upper Xingú region (a multi-ethnic area heavily acculturized by its Arawak-speaking groups) shows direct analogies to the ceramics of the Orinoco Valley dated back to 900 BC (the Barrancooid tradition) at the locations of the early nodes of the Arawak regional exchange system (Heckenberger 2005). This is a strong indication of stability in terms of the material culture of the Arawak matrix. Also worth noting is that due to the historical ethno-linguistic production specialization of the upper Xingú region (Eriksen 2011: 88), Arawak-speaking female potters have been responsible for the region-wide pottery production, meaning that other ethno-linguistic groups have been subjected to both Arawak material culture and language when in trade contacts with the Arawaks, a process that has been ongoing for a considerable time depth. This may be interpreted as indicative of a greater degree of contact sensitivity among the languages of the non-Arawak-speaking groups of the upper Xingú region, but it is also indicative of the strong cultural influence conducted by the Arawak-speaking groups.

The fact that the reconstructed words for maize (fig. 2) are borrowed throughout the area is in itself an indication of the close socio-linguistic contact of the region. As seen in figure 5, the distribution of root 1a, *mahiki, is clearly an Arawak-dominated root spread through the exchange system and to some non-Arawak-speaking groups through socio-cultural contact. Thus, the distribution of root 1a is clearly a stable feature that is borrowed only in certain types of social contexts like the close interaction sphere of the upper Xingú area. Overall, we see a clear effect of social interaction through socio-cultural exchange on the distribution of lexical features (fig. 3), but an even more detailed picture can be painted by examining the distribution of individual lexical features and their relationship to the exchange system. As for the distribution of the proto-forms for maize, a close examination indicates that socio-cultural contact through the Arawak exchange system was a much more influential force in the distribution of the proto-words than the genetic relationship between the languages, indicating that certain types of language contact situations occurred after the initial geographic dispersal of the languages of the Arawak family.

Nichols (1992: 166ff) notes that word order is a feature which readily spreads areally between unrelated languages. The distribution of verb-initial ordering (VSO/VOS) in the Arawakan languages is a case in point. While verb-initial order is found in several Arawakan languages, it is largely concentrated to the western areas of Amazonia (cf. fig 6), in areas where there are several other non-Arawakan languages which display verb-initial ordering. In other regions of Amazonia, Arawakan languages tend to display other word order patterns (particularly SVO in northern areas, and SOV further south). In contrast, the active verb agreement pattern (with prefixes cross-referencing agentive arguments and suffixes cross-referencing patientive arguments) is found throughout Amazonia.

In summary, the Arawak matrix developed and spread through a regional exchange system initiated around 900 BC in the Orinoco area. In terms of
the present study, the most important time period in the system is between AD 200 and 600, a time period when the system displays its greatest internal similarities in terms of material culture. At this point in time, the materiality of the Arawak matrix displays similarities so great that the same type of ceramic adornos (a decorative feature of the Arawak pottery) was used in an area stretching from the Antilles to Central Amazonia. Comparing the extension of the regional exchange system at this point in time to the distribution of clusters A, B, C, and D in the language tree may explain the high degree of lexical similarities between cluster A and B despite great geographical separation of the speakers; the distribution of the lexical features is the result of language contact up to AD 600. After this period, the internal homogenization of the system was less powerful, perhaps as a consequence of the constant geographical expansion of the trade networks that kept on linking more and more distant areas of Amazonia together. By AD 1000 (fig. 3) the system experienced its greatest territorial extent, but, as mentioned above, the internal similarities in material culture and language had constantly decreased after AD 600. After AD 1000, a slow internal fragmentation process continued up until AD 1200, when the expansion of the Tupian languages (Lathrap 1970: 150f; Meggers 1971: 122–130; Neves et al 2004: 133; Rebello et al 2009: 22, 27) finally broke the Arawak regional exchange system apart once and for all. What remained of the system after AD 1200 were the current Arawak-speaking groups, and small sections of the formerly great regional exchange system still upheld by these local Arawak-speaking communities.15

Notes

1. We thank Alf Hornborg and Chris Sinha for valuable comments. We also thank Judith Josephson for correcting our English. The research has been funded by Centre for Cognitive Semiotics/Bank of Sweden Tercentenary Foundation and the Human Ecology Division, Lund University.

2. The assumption made here is that the same factors hold for lexical items, grammatical constructions and basic language structure alike, albeit in different degrees
(which is part of the topic of this paper). Thus, as pointed out by an anonymous reviewer, e.g. calques are simply a special (construction-based) subset of the more general phenomenon of areal features (or, conversely, areal features can be described as clusters of calques). The label of “typological pressure” can be taken to mean non-coincidental features which are neither inherited nor loaned, but are in some way functionally or cognitively motivated. Under this reading, a lexical parallel to typological pressure can possibly be found, namely iconicity (e.g. sound-symbolism, onomatopoeia). However, we will not deal further with the issue of iconicity, since its effect in creating surface similarities across languages is much lower than that of inheritance or loans.

3. Here, an IPA-based system of writing reconstructed forms is used, which also includes a more traditional historical-linguistic method (e.g., Szemerényi 1990): in case of the occurrence of palatalization and/or fricativization in a number of languages, the reconstructed form is marked by a non-palatalized plosive + palatalization marker /j/. In case of reflexes in languages that yield several possibilities, e.g., bilabial plosives /p/ or /b/ as being the possible reconstructions, capitals (referring to the most likely form) are being used, e.g., */b/ = bilabial plosives (/b/ /p/), */T/ = dental/alveolar/postalveolar plosives (/t/ /d/), */N/ = nasals (/n/ /m/; more probably /n/), */M/ = nasals (/m/ /n/; more probably /m/). For problematic vocalic reconstructions, the classical “schwa” */a/ is used.

4. The Yagua data is open to various interpretations, but one possible interpretation is that Yagua adpositions are postpositions, either suffixed directly to nouns, or suffixed to agreement clitics which precede the noun, which gives the appearance of agreeing prepositions (Payne and Payne 1990: 363ff).

5. Here WALS cites a single exception, not surprisingly, perhaps, a language in Amazonia: the Arauan language Paumaari, spoken in western Brazil, on the Purus river.

6. Indeed, Evans and Levinson (2009) refer to the “myth of language universals”, capitalizing on the fact that they are tendencies rather than exceptionless universals, although they do admit that certain strong tendencies are due to functional economy (op. cit. 445). We are aware of the non-universality of Greenberian “universals”, of which there is perhaps more evidence in Amazonia than elsewhere. However, our basic assumption (to be empirically tested) is that several tendencies are so strong that they cannot be the product of mere chance.

7. Efficiency and economy can be measured in terms of the degree to which structures can be re-used for different types of constructions in the language (e.g. head-final structures appearing both in OV verb phrases and in postpositional phrases), and the degree to which unnecessary redundancy is done away with.

8. As is argued in Holmer (2006) to be the explanation of the striking similarities at the syntactic and morphological levels between Basque, Georgian, Burushaski and Chukchi (including polypersonal agreement and prefixed morphology), namely that they all derive from typological pressure exerted by the one coincidentally shared property of ergativity.

9. An anonymous reviewer points out that gender is another feature which correlates highly with language family, i.e. is, in our parlance, genetically stable.

10. One reason is that Tariana has borrowed overt case marking from Tucano (although the alignment itself is not fully accusative as it is in Tucano).

11. Or if it is resolved at all: strictly speaking, it is an ergative alignment of the case marking system which is more or less excluded in SVO languages, whereas the split-ergative / active alignment found in Amazonia concerns verb agreement paradigms. WALS does, indeed, list 6 languages which combine SVO word order with an active verb agreement alignment, 4 of them in Amazonia, and three of these (Apurinã, Arawak [Lokono] and Warekena) actually belonging to the Arawakan family.

12. e.g. whether the alignment pattern is instantiated by case-marking, verb agreement or syntactic relations.

13. pace Dunn et al. 2011, who argue that many instances of typological correlations are lineage-specific, i.e. represent shared inheritance.

14. Heckenberger (2005: 31) estimates that the Arawak population of the upper Xingu dates back to AD 500.

15. The scattered remains of the Arawak exchange system were often transformed into multi-ethnic interaction spheres around the centuries of European colonization (Eriksen 2011: 197). This process was intensified as a response to the pressures of the colonial powers on indigenous communities across Amazonia, and it is likely that there were also substantial effects on language use accompanying this process.

References

Aikhenvald, Alexandra

Aikhenvald, Alexandra

Barth, Fredrik

Birket-Schmidt, Kaj
1943 The Origin of Maize Cultivation. Copenhagen: Ejnar Munksgaard.

Brown, Cecil H.


Predicting language-learning difficulty

Michael Cysouw

1. Introduction

The difficulty people have in learning a foreign language strongly depends on how different this language is from their native tongue (Kellerman 1979). Although this statement seems uncontroversial in the general form as it is formulated here, the devil lies in the detail, namely in the problem how to define differences between languages. In this paper, I investigate various factors that quantify differences between languages, and explore to which extend these factors predict language learning difficulty. This investigation results in concrete predictive formulas that derive the learning difficulty for native English speakers depending on a small selection of linguistic factors of the language to be learned.

Section 2 presents the data for language learning difficulty that will be used in this paper. This data originates at the Foreign Services Institute (FSI) of the US Department of State and it includes only approximate average learning times of foreign languages for English speakers. The data is rather rough, but it is highly interesting because it gives comparable estimates for language learning difficulty for a large number of strongly different languages from all over the world. Section 3 investigates the relation of these estimates for language learning difficulty to very general predictors like geographical distance and genealogical affiliation. In both cases, the further away a language is from English, both geographically and genealogically, the more difficult a language is expected to be. All empirical effects point in the expected direction, though the factor Germanic vs. non-Germanic turns out to be the strongest predictor for language-learning difficulty.

Section 4 takes up the differences in writing systems as used for the various languages in the current sample. Using the Universal Declaration of Human Rights, the orthographic similarity between English and other languages is established. For languages with a Latin script, there is a strong correlation between language learning difficulty and the similarity in frequency distribution of orthographic symbols. Section 5 investigates structural grammatical properties of languages using data from the World Atlas of Language Structures. I establish which structural differences from