13 The languages of South America: deep families, areal relationships, and language contact

Pieter Muysken, Harald Hammarström, Joshua Birchall, Swintha Danielsen, Love Eriksen, Ana Vilacy Galucio, Rik van Gijn, Simon van de Kerke, Vishnupraya Kolipakam, Olga Krasnoukhova, Neele Müller, and Loretta O'Connor

After summarizing the earlier chapters, we sketch a general overview of the different phases in the development of South America. We then explore the possibility of a continental bias for typological features characteristic of South America, which may point to the early entry of a limited set of features into the continent. Subsequently we analyze possible deep families or macro-groups in the continent, and their regional distribution. We then turn to the issue of whether different subsets of structural features yield different distance matrices for the language families studied. To further explore contact possibilities, the results for language contact in our book are charted. Finally, we conclude and take stock of what has been achieved and how further research should proceed.

1 Introduction

In the contributions assembled in this book we have explored a number of specific cases of language expansion and contact, as well as four sub-domains in which the genealogical and geographic distributions of features in different domains of the grammar were charted.

In this chapter we further reflect on how we can relate these contributions to the general questions posed at the beginning of this book:

(A) Why are there around 108 genealogical units in the continent? Why so many language families, and why so many isolates? What is the distribution of both larger families and isolates?

The present chapter has resulted from the work in our group over the last few years. We also acknowledge the input of the various researchers listed in our acknowledgments, notably also Helder Perri Fereira, at different points on the ideas presented here.

(B) Given the apparent genealogical diversity, why are there so many shared specific areal typological patterns, some characterizing most of the continent as a whole, and some individual parts of the continent?
(C) What can we learn about the relation between the issues in (A)–(B) from the perspective of language history (vertical transmission) and language contact (horizontal transmission)?

After summarizing the chapters in Section 2, we sketch a general overview of the different phases in the development of South America in 3. Section 4 deals with the possibility that there is a continental bias for typological features characteristic of South America, which may point to an entrance of a limited set of features into the continent in the early stages of its peopling. In Section 5 we turn to possible deep families or macro-groups in the continent, and their regional distribution. Section 6 raises the issue of whether different subsets of structural features could yield different distance matrices for the language families studied, and in Section 7 the results for language contact in our book are explored. In Section 8 we conclude and take stock: what has been achieved and how ought we to proceed in further research?

2 Summary of the contributions in the book

In the first chapter Muysken and O'Connor presented the main issues raised in this book, against the background of the **genealogy**, **typology**, and **language contact situation** of the South American indigenous languages. All three areas are underexplored so far, and particularly the relationship between them raised many unresolved questions.

In the subsequent chapter O'Connor and Kolipakam developed a portrait of **population movements** and **contacts** in South America, from initial migrations some 15,000 years ago through millennia of dispersal and interaction, which resulted in localized pockets of population growth and cultural development. Current genetics research supports separate patterns of population density and interaction between East and West, and various types of evidence point to localized social complexity and down-the-line contact without major population dispersals until roughly 4,000 years ago.

Hammarström examined the role of **basic vocabulary comparison** in the classification of South American languages with two empirical results emerging. First, the classification of South American languages by Loukotka (1968), based on basic vocabulary inspection, closely mirrors the classification presented by Campbell (2012a) for which far more extensive lexical and grammatical data had become available. Second, results of automated lexical comparison (ASJP) have a high degree of correspondence to those of traditional methods, despite the simplistic assumptions of the former and question marks on systematicity and objectivity of the latter. Thus shallow groups are robustly

recognizable in basic lexicon, and provide the foundation both for tracing earlier connections between shallow groups and for tracing contact that occurred within the time frame of the shallow groups.

In a regional case study of the **Isthmo-Colombian** area, O'Connor devised a metric of feature categorization that incorporates sensitivity to properties of human interaction. Results indicate that analyses of both contact and genealogical relations are enhanced by categorization that reflects the impact of social constraints on linguistic change as well as conventional notions of stability in linguistic systems. Reflections of social scenarios need to be combined with simple frequency of contact.

Van Gijn's survey of the distribution of Andean and Amazonian features in the **upper Amazon** area shows that the transition from the Andean to the Amazonian area is gradual and complex. This is consistent with the intricate history of contact between the different ethnic groups of the area, and it presents a strong argument for connecting the research traditions associated with these areas. Morphosyntactic influence generally seems to represent older contact situations than phonological influence.

In their chapter on the **Andean matrix**, Van de Kerke and Muysken argued that the traditional division of the Quechuan family into two main branches can be maintained for structural features. However, Aymaran is structurally closer to Central Peruvian Quechua than innovative Ecuador Quechua. Other Andean languages differ much more than previously assumed.

Eriksen and Danielsen sketched the birth, expansion, and fragmentation of the **Arawakan** culture and languages across Amazonia. This ethnolinguistic complex is characterized by a robust uniformity that was sustained until late prehistory, resulting from an intensive exchange system that – despite expansion in a multidirectional and irregular fashion – managed to keep the system together across vast distances.

Eriksen and Galucio showed that one out of five expansive **Tupian** branches, Tupí-Guaraní, expanded through a hybridizing culture that spread across vast geographic distances through the absorption of cultural and linguistic elements from neighboring populations. The linguistic analysis shows that lexical features were better preserved than structural ones, and that the expansion process likely continued into the historical period.

With respect to **Tense/Aspect/Mood/Evidentiality** (TAME) systems, Müller presented evidence that grammatical **desiderative** markers occur more frequently in South American languages than in other parts of the world. Desideratives in the sample stem from proto-forms, but they also developed due to language-internal pressure and contact-induced grammaticalization.

Birchall examined the diverse array of **verbal argument marking** patterns encountered across the continent and tested for regional distributions of certain often-discussed features. Statistical tests showed that many areal proposals in

the literature are in fact not significant, and that an East–West division was often more significant than the classic Andean–Amazonian division.

Krasnoukhova showed that in **Noun Phrase** structure there is a split between languages spoken in the western part vs. the eastern part of the continent, and not between the Andes and the Amazon as has been traditionally assumed. While the western part corresponds to the Andean sphere, the eastern part includes languages spoken far beyond the Amazon region. Furthermore, in a case-study on semantic features encoded by demonstratives, Krasnoukhova has shown that the Chaco and the Southwest Amazon region stand out on the continent for encoding verbal categories with demonstratives.

And finally, Van Gijn showed that **nominalization** as a **subordination** strategy is significantly more pervasive in South America than would be predicted on the basis of global patterns. The patterns found within South America are most consistent with a scenario of several smaller spreads, possibly promoted by a few language families with major extensions (e.g. Quechuan, Tupian, Cariban).

3 Phases in the development of the South American languages

To organize our answers to these questions, we will use a framework in terms of four phases in the history of the continent, building on O'Connor and Kolipakam (this volume). It is impossible to look into the past as far back as 12,000 BCE, but the most likely scenario for the history of the languages of South America that we can infer from the current evidence involves the following:[1]

I 11,000–6000 BCE Initial settlement and dispersal

A small (<10) number of groups moved into the continent and quickly dispersed. Other groups may have followed at later dates with less speed. These groups settled in different parts of the continent and then fractured into small bands. The bands developed separate identities, strengthened by separate lexical systems, but kept interacting on a local level, through the exchange of goods and sexual partners.

The evidence for this early phase includes archaeological data, which support settlements across the continent dated around 9000 BCE. Genetic data suggest a relatively uniform, possibly quite small, initial population (O'Connor and

[1] The time span for developing the linguistic diversity in current models is short. The date of 33,000 BCE for the Monteverde site in Chile has been proposed, and this would open up an alternative scenario of social and linguistic development of the continent. Although the Monteverde dates have not been repeated yet there are strong indications from archaeology that the traditional time span of the human occupation of the Americas is much longer than previously thought.

Kolipakam, this volume). At the same time it is evident that some groups (e.g. Chibchan) obviously must have come to South America at a later date (Constenla 2012). Phase I linguistic features, if they exist, can be assumed to be characteristic of large areas or possibly all of the continent, and should be highly stable. The internal linguistic development of South America would have taken place between 12,000 and 2000 BCE, with the most intensive linguistic diversification probably dated 11,000 to 6000 BCE. Already at 6000 BCE, there were long-distance connections between groups from Colombia to the mouth of the Amazon – social contacts that already at this point in time would have served to equalize some of the linguistic differences of the continent.

Note that Phase I falls outside of the "lexical horizon," the date where two cognate forms would no longer be likely identifiable without advanced reconstruction of proto-phonologies of the respective language families. We don't have these data, so it would be impossible to evaluate any lexical relationship or identify any borrowings beyond the 8 K horizon. Any identifiable cognacy or borrowing will most likely have emerged after this horizon.

The scenario in Phase I is compatible with the low rates of lexical borrowing in hunter-gatherer societies (Bowern et al. 2011), coupled with the wider geographic distribution of specific features, as we will try to show below. It is clear that the bands cannot have been completely isolated, since small groups cannot sustain themselves without exchange with other groups.

II 6000 BCE–2000 BCE Pre-formative

As technology developed, and plants were domesticated and developed into agricultural crops, different groups started expanding and invading territories previously occupied by other groups. Sometimes there was population displacement, but some cultural expansions also took place without large groups of people moving. This is also the period in which ceramic techniques were developed, a real coup for Amazonia, with some of the earliest known ceramics in the Americas.

Evidence for this phase comes from the appearance of domesticated food cultivars in the archaeological record. The spread of these cultivars would also correspond to the same social relationships that allow for the spread of language, genes and other technology characteristic of Phase II. Dunn et al. (2005) argue that structural features may delineate a deeper level of genealogical time depth, as lexicon is easier and more obvious to manipulate as a badge of social identity. We should be able to trace genealogical relationships back to this period.

The expansions of specific larger genealogical units such as Chibchan and Macro-Jê can be documented, with estimated starting dates. Specific cultural

practices, words, and grammatical features can be documented and traced to dispersal languages.

III 2000 BCE–1500 CE Formative

A period with a marked increase in intensive food production and thus sedentism. Typical for this period are the Huari expansion linked to Quechua II (Van de Kerke and Muysken, this volume), the Arawakan expansion (Eriksen and Danielsen, this volume), and the Tupian expansion (Eriksen and Galucio, this volume). In the later stages, population density increased and more complex and larger networks were created.

Evidence comes from the spread of ceramic traditions, landscape "domestication," and anthropogenic soils. Sedentism, population growth, and the resulting areas of dense population would lead to different social dynamics than those involving hunter-gather groups in contact. All of the large families other than Chibchan and Macro-Jê migrated on a large scale only during Phase III, and their general membership can be identified through comparison of basic lexicon. Phase III features are associated with particular expansions and their influence on the surrounding languages, and these can be reconstructed for each particular language family. Their spread may be accompanied by lexical borrowings from the expansion language.

At this point there is traceable evidence of specific borrowings associated with cultural elements. Multilingual complex networks in the Rio Negro and the Xingú regions emerged during the last part of this phase.

IV 1500 CE – European invasion and colonization

The Spanish and Portuguese conquest and colonization of the continent had the effect of decimation and fracturing of populations, and the disappearance of entire groups. Populations and languages were displaced, while certain languages were promoted as *língua geral* or *lengua general*, and subsequently expanded further. New contact zones were created through *reducciones* or reserves.

For this period there is of course the historical record, coupled with anthropological observations, travelers' accounts, and so forth.

It is important to consider these phases not as solid and mutually exclusive blocks disposed in a line (with only one direction), but as bubbles often coexisting in the same time span. For example, while agriculture was profoundly changing the social dynamics in the eastern Amazon, large parts of the western Amazon may have been still experiencing a scenario much more akin to Phase I, perhaps influenced by factors of physical geography inhibiting fast expansion

and growth (Nichols 1992). The large regional differences were also linked to climatic changes.

This cumulative perspective could help to account for part of the diversity we encounter today, in terms of different levels of integration of language systems.

4 A possible continental bias

Since we assume that only a limited number of populations entered the South American continent, a relatively limited set of linguistic features was part of the original linguistic base that helped shape the languages of South America. A first set of research questions then, related to Phase I, would be whether the typological features of the South American languages show a continental bias, i.e. are significantly more present in South America than elsewhere. We cannot study this on the basis of the data we gathered, since our questionnaire was not used outside of South America. However, the WALS data allow us to answer the question of continental bias. Are there specific feature specifications which are significantly different for South America than for other continents? To find feature values which are significantly more common in South America, we checked all 565 feature values in the WALS (http://wals.info accessed 1 June 2012). For an individual feature value a 2×2 contingency-table is obtained by taking the number of South American versus non-South American languages with and without the feature value in question. We can then apply a one-tailed Fisher Exact Test to test for significance of the overrepresentation of the value in South America. A number of features remain significantly overrepresented in South America even after correcting for multiple testing (by Bonferroni correction).

There turn out to be a number of such features. They can be organized as in Table 13.1, with some examples provided per domain.

Our goal here is merely to affirm that, where we have access to non-South American data, there turn out to exist characteristics that are overrepresented in South America. We refrain from posting further details on the precise nature of these characteristics from the WALS, focusing instead on our database of much more fine-grained features for South American languages. In a similar manner, Dediu and Levinson (2012) argue that the structural stability profiles of South American language families form a significant cluster.

Nonetheless, it is clear that in a broad range of areas South America presents special features. In Müller (this volume) desideratives are argued to be a feature significantly more present in South America than elsewhere, and Van Gijn (this volume) has a similar result in his subordination chapter for nominalizations. Krasnoukhova (2012: 75) has shown that a fully grammaticalized category of possessive pronouns is rare in South America compared to other parts of the world.

Table 13.1 *WALS features for which South American languages show a significantly distinct profile as a group*

Feature and domain	Value
Postverbal negative morphemes	Negative suffix
Negative morphemes	Negative affix
Order of negative morpheme and verb	[V-Neg]
Order of object and verb	OV
Order of adposition and noun phrase	Postpositions
Order of numeral and noun	Numeral-Noun
Coding of nominal plurality	Plural suffix
Position of tense-aspect affixes	Tense-aspect suffixes
The velar nasal	No velar nasal
Presence of uncommon consonants	None
Order of adverbial subordinator and clause	Subordinating suffix
'Want' complement subjects	Desiderative verbal affix
Coding of evidentiality78A	Separate particle
The perfect 68A	No perfect
Hand and arm 129A	Different
Numeral bases 131A	Restricted

It is tempting to try to relate various typological properties of the South American languages, such as the high levels of evidentiality marking, elaborate modality systems, and possibly the elaborate systems of demonstratives noted by Krasnoukhova (this volume) to a shared ethos of a heightened awareness of one's social place, but this requires a more comprehensive semantic study of the properties of the languages of the continent.

5 Deep families, macro-groups, and their regional distribution

The languages in our database belong to known families or are isolates. To look for deeper relations between these families, we do not compare modern languages directly, but compare typological profiles projected to the "proto"-language of every family. For every lineage, feature values were reconstructed for the proto-language as follows. First, the received tree sub-grouping (Hammarström, this volume) was used for every lineage with two or more languages. (In all cases, this sub-grouping is not based on typological characteristics, but on lexicon, sound shifts and/or morphology.) Next, for each feature and interior node, the most parsimonious value was chosen, i.e., the value which required the fewest changes to the observed values at the leaves and the tree topology. Thus, the feature values at the root represent the reconstructed typological profile of the proto-language of a family. For lineages

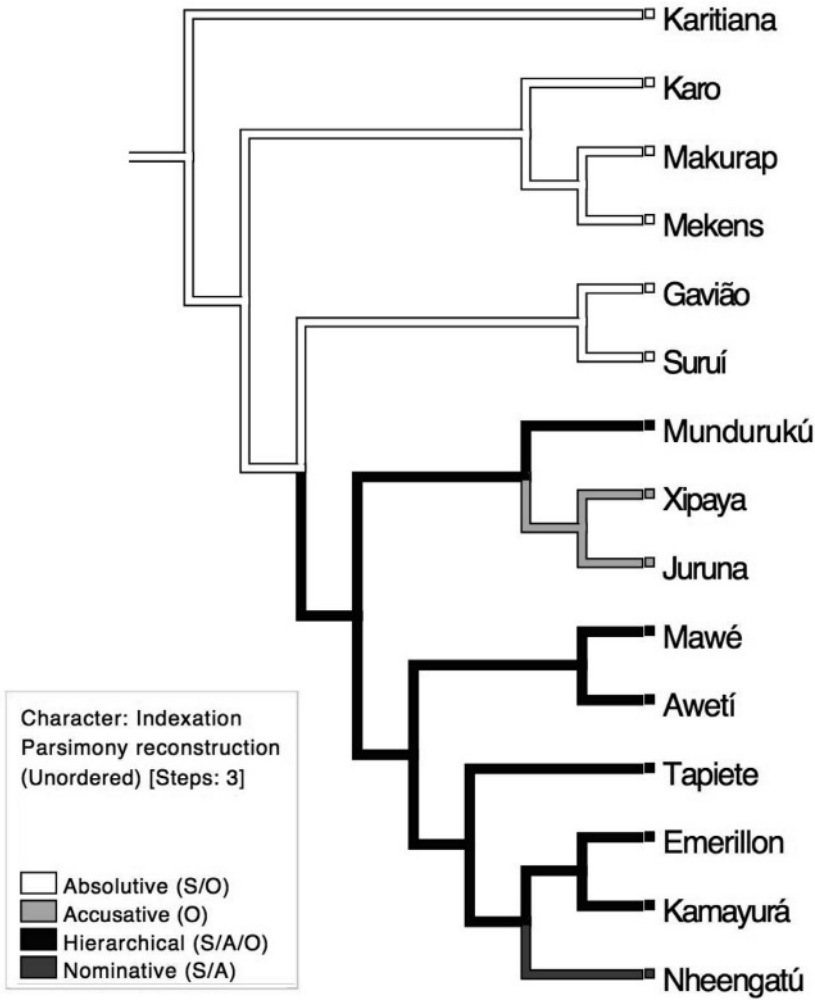

Figure 13.1 Parsimony reconstruction for alignment in Tupian (Birchall 2014, based on the tree typology of Walker et al. 2012)

with only one member, the profile of that member will represent the typological profile of the lineage; of course, this is not ideal, but there are no obvious alternatives. An example is given from a simplified tree for Tupian alignment, in Figure 13.1.

The number of genealogical units that we have data for (including the isolates) is fifty-four. The typological profiles of the proto-languages can be

pair-wise compared by a straightforward relative Hamming distance, i.e., the proportion of differing feature values. Not every pair of the fifty-four units could be compared due to differences in the sets of features coded. For example, Cuna (coded only in the Constenla dataset) could not be compared to Kwaza (coded in the South American languages dataset) because the feature sets are different.

It is important to realize that taking structural features rather than lexical elements or sound patterns as the basis of comparison does not suddenly make all differences between the language families vanish. Dunn et al. (2005) have argued that structural features may reveal greater time depths than lexical features, but it is also evident that they may be sensitive to contact. In any case, we find a blurring of sharp contours between families in parts of western South America, as will be seen below. Both Birchall (this volume) and Krasnoukhova (2012, this volume) propose a west–east split for Argument Marking and the Noun Phrase. We therefore proceed to look at this split more closely.

In the top fifty (with shortest Hamming distance) there are twenty-five pairs from the western region (WW, ranked internally also in terms of distance). **Bold** are languages from the Andean matrix; in SMALL CAPS we find languages from the Isthmo-Colombian region; languages that are also part of a highly ranked pair across the east–west divide are underlined. Thus, in the twenty-five top ranked western (WW) pairs, nine involve languages from the Andean matrix, and nine languages from the Isthmo-Colombian region. The closest pair in the whole sample is Aymaran-Quechuan.

Aymaran-Quechuan
CHIBCHAN-MISUMALPAN
Leko-**Quechuan**
Aymaran-Uru-Chipaya
Jivaroan-Leko
Barbacoan-JICAQUEAN
CHIBCHAN-JICAQUEAN
Kallawaya-Quechuan
CHIBCHAN-Jivaroan
Aymaran-Kallawaya
Jivaroan-Panoan
Quechuan-Uru-Chipaya
Leko-Panoan
Aymaran-Hibito-Cholon
CHIBCHAN-**Leko**
Jivaroan-Paez
Leko-Tucanoan
Barbacoan-MISUMALPAN

Boran-CHOCOAN
Nadahup-Tucanoan
Araucanian-Jivaroan
Leko-Nadahup
CHIBCHAN-CHOCOAN
CHOCOAN-JICAQUEAN
Hibito-Cholon-Quechuan

These links show also that the Andean area and the foothills are historically rather intimately connected – see also Van Gijn (this volume) on the Andean foothills. The fact that these connections also come to the surface when based on reconstructed typological profiles suggests moreover that this connection is old, or that contact was intense and sustained.

The eastern group (EE) involves fourteen pairs out of the top fifty. The languages belonging to the postulated Tupian-Cariban-Jê group (Rodrigues 1985) are marked **bold**; languages that also form a highly ranked pair across the east–west divide are underlined.[2]

Bororoan-Tupian
Arawakan-Guaicuruan
Arawakan-Itonama
Itonama-**Tupian**
Arawakan-Tupian
Kanoe-Kwaza
Kwaza-Nambikwaran
Arawakan-Matacoan
Chapacuran-Itonama
Kanoe-Muniche
Kanoe-Matacoan
Tupian-Urarina
Kanoe-Yanomamic
Chonan-**Tupian**

Finally, there are eleven pairs in the top fifty that cut across the east–west divide (EW), as defined in terms of the projected homeland.

Boran-Chonan
Puinave-Tupian
Jivaroan-Kanoe
Kwaza-Leko
Chibchan-Kanoe
Kwaza-Nadahup

[2] It is not always obvious how to classify a language family. Arawakan is a case in point, since it had a relatively "western" origin and members of the family are spoken in both eastern and western regions.

Table 13.2 *Comparing the top 200 language family pairs in the sample*

	Total # of pairs in top 200	Total # of pairs in sample	Total # of pairs in top 100	Number of language families in top 100	Top 100 language families divided by pairs
WW	74	128	39	20	0.51
EE	51	227	25	17	0.68
EW	75	330	36	24	0.67

Kwaza-Tucanoan
Jivaroan-Tupian
Nadahup-Tupian
Kwaza-Jivaroan
Arawakan-Boran

It is unclear why a language in certain pairs may cut across the east–west divide such as in the case of Kwaza. Is this because of the Guaporé-Mamoré linguistic area described by Crevels and van der Voort (2008), does it reflect earlier population movements, or is it chance?

These arrays of language pairs may or may not be interesting by themselves, but when compared to the total number of language family pairs in the sample, a result emerges. So in fact, even though there are more EW pairs compared (330), a lower proportion turns up in the top group (75), as shown in Table 13.2.

It is also interesting to see that WW pairs (in the western region) are more often linked to another family in the same group in the top 100 of pairs than the EE and EW pairs (0.51 linkage versus 0.67/0.68).

Similarly, the average Hamming distance, i.e., the fraction of differing values, (measuring structural dissimilarity) between all the pairs in the western region is shorter than in the eastern region or than in the east–west connections:

West West 0.391
East East 0.439
East West 0.453

Of the pairs of projected proto-languages with a sufficient number (at least sixty-seven) of the same features defined, the closest pair is Quechua-Aymara with a Hamming distance of only 0.18, followed by a web of other potential relations.

The location for a proto-language of a family is inferred using the observed locations of the daughter languages and the tree sub-classification of the family. This procedure uses the same intuition as other manual and automated procedures in the past (Wichmann et al. 2010), namely the principle of maximal diversity. In our case the diversity differences are directly determined by the tree topology. The location of each interior node is projected to be the average x-coordinate and the average y-coordinate of its immediate children. This is done recursively until the root is reached.

It has often been argued that structural features are more revealing of geographic relationships than of genealogical affiliation (Donohue et al. 2011). For this reason, we tested the relation between Hamming distance and geographic distance for all language pairs in our sample. How much of the Hamming distance is predictable by the geographic distance? A plot (Figure 13.2) with geographic distance (x-axis) between each pair of projected homelands and typological distance (y-axis) if the pair had at least forty features defined shows that there is a tiny correlation: 0.09. We can take this to mean that geographic distance does not explain most of the structural distance found.[3] A similar result is obtained if we take the four grammatical domains separately.

Nichols (2003) attempts to relate different kinds of stability (genealogical, typological, areal stability) to different types of language change scenarios (inheritance, borrowing, substratum, and selection). It becomes clear from this chapter that the behavior of inherently stable and unstable features may be overshadowed by the sociolinguistic situation the speakers of a language are in. We often do not really know how the interaction between the different kinds of stability and scenarios of language contact is expected to play out. What is needed, therefore, is a historically informed application of stability measures, so that we can assess the potentiality for change of each typological feature in a given situation. For South America, we are not yet in a situation where we can do this, but the major developments in the field are encouraging. Hopefully the results presented in this volume can contribute to this debate.

6 Comparing different sets of features

In our research we included various kinds of features. Our original hypothesis was that there are asymmetries between different components, with TAME features the least stable, Argument Marking intermediate in stability (internally not a homogeneous set), and Noun Phrase and Subordination strategies fairly

[3] This conclusion could indicate that structural similarities between geographically distant languages possibly are the remains of a state when these (proto-)languages were spoken close to each other (11,000–6000 BCE).

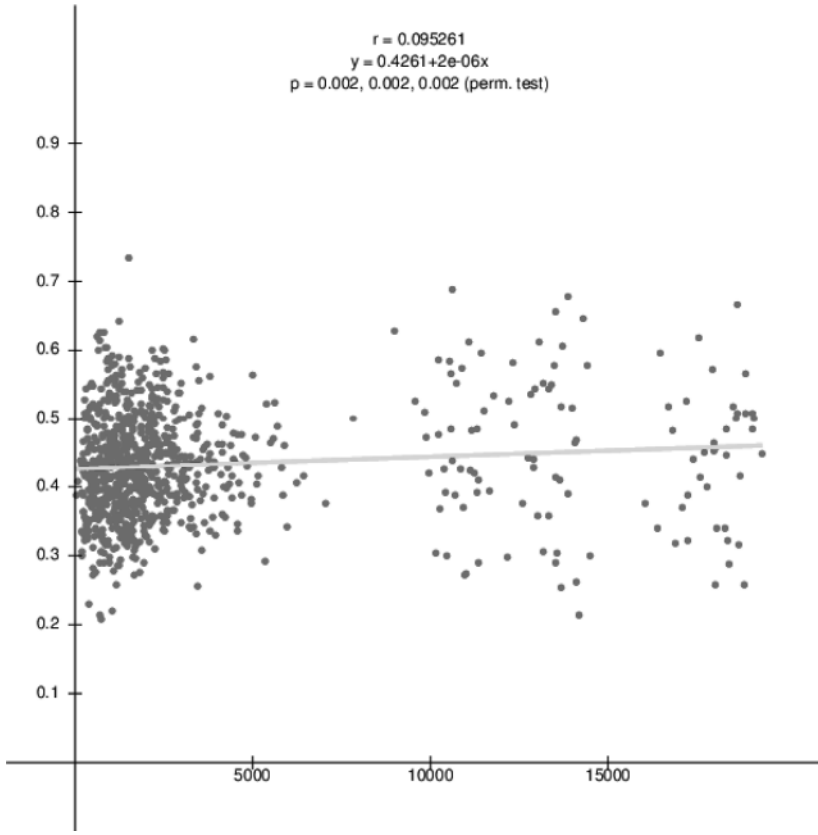

Figure 13.2 Correlation between Hamming distances (y-axis) and geographic distances for the pairs of families in the sample (x-axis, in km)

stable. However, this needs further testing and also there is the possibility that individual features may show particular stability, as demonstrated in earlier research.

To date, there is no consensus on how to measure stability for structural features. Wichmann et al. (2010b) evaluate different methods of measuring stability, and apply the method most predictive of genealogical relationships to the data assembled in Haspelmath et al. (2005). The philosophy behind their metric is that if one given feature more often tends to have the same value for languages that are related than does another given feature, then the first of the two may be considered to be more stable. The Wichmann–Holman list can be used as an index for relative stability, but it is a de-contextualized summary of many different regional realities,

Table 13.3 *Rank order correlation between the language pairs for TAME, SubOrd, ArgMar, and NP*

X	Y	rho	p	# items
ArgMar	TAME	0.23	0.000	1444
NP	SubOrd	0.22	0.000	841
NP	TAME	0.11	0.000	1369
SubOrd	TAME	0.14	0.000	841
ArgMar	NP	0.17	0.000	1444
ArgMar	SubOrd	0.20	0.000	841

so that it may underestimate the effects of lineage-specific or area-specific influences.

Following the general approach of Wichmann et al. (2010b), but limiting it to the South American context, we can evaluate which feature values have remained stable within the individual families in our sample. The parsimony reconstruction described in Section 5 also allows estimates of the stability of features to be gauged. When feature values for interior nodes have been projected, we have the result that every transition from an interior node to its daughter either changed the feature value or retained it. The proportion of retain-transitions represents a measure of stability. Actually, it is an estimate of maximal stability, since the inferred value for interior nodes assumed no or few changes (as far as this was possible under the leaf node values and the tree topology). Thus, essentially, the stability measure says how stable a feature is if every time it *can* be stable it also *is* stable. For good estimates on stability we need a large number of transitions. With shallow families and incomplete information about the languages in them, we have few transitions to gather evidence from.

Other measures of stability that use essentially the same heuristic, i.e., a feature value is stable to the extent that it is homogeneous in known families, are discussed and compared in Dediu and Cysouw (2013).

Given these measures of stability of individual features, we can make global comparisons of the distance matrices for clusters of features. Do these change if we select "stable" features as described in the literature? Do some components show greater diversity in feature specifications than others? In our study four large feature sets can be contrasted, as noted above.

For our own dataset, a first question is whether the values for the variables in the four domains correlate. We calculate a simple rank order correlation between the language pairs for TAME, SubOrd, ArgMar, and NP, leaving aside the pairs for which only one language has values; see Table 13.3.

Table 13.4 *The variance in the four domains*

Domain	Degree of variance
Noun Phrase	0.105
Argument Marking	0.097
TAME	0.097
Subordination	0.075

Table 13.5 *Average stability of our different sets of features for the language families*

	avg stability	# features	avg # transitions
Subordination	0.89	78	29.0
Argument Marking	0.86	67	47.5
Noun Phrase	0.84	67	33.7
Constenla Features	0.83	90	13.0
TAME	0.78	33	38.9

The correlations are positive and significant but relatively small for all domain pairs.

We may also compute a measure of diversity within the four domains. The variance, i.e., average distance to the mean, of all language pairs within each domain is shown in Table 13.4.

The highest degree of variance, and hence diversity, is found in the Noun Phrase structure domain, with the domains TAME and ArgMar ranked equally, and Subordination lowest.

In Table 13.5 the average stability of our different sets of features for the language families is presented.

Subordination ranks highest on stability, and as shown in Table 13.4, lowest on variance. TAME ranks relatively low on both counts, Argument Marking ranks relatively high on both, and Noun Phrase ranks highest on variance and high on stability.

In Table 13.6 the highest-ranked thirty language pairs in terms of their distance on the Noun Phrase domain are contrasted with those same highest-ranked pairs in the TAME domain, the Subordination domain, and the Argument Marking domain.

Among the highest-ranked pairs in the NP domain we find the postulated relatives Bororoan and Jê-Jabutí, and a large group of western languages

Table 13.6 *Language pairs ranked highest on the domains of NP, TAME, SubOrd, and ArgMar (languages from the western region bold, Tupian-Cariban-Macro-Jê languages italic)*

NP		TAME		SubOrd		ArgMar	
Bororoan Je-Jabutí	0.16	Chonan Itonama	0.21	Arawakan Itonama	0.06	*Bororoan Tupian*	0.07
Chibchan Jivaroan	0.19	**Boran Chocoan**	0.22	**Boran** Itonama	0.06	**Aymaran Quechuan**	0.13
Barbacoan Jivaroan	0.20	Chapacuran Matacoan	0.22	Kwaza Nadahup	0.06	**Aymaran Hibito-Cholon**	0.14
Chocoan *Tupian*	0.20	Kanoe Muniche	0.23	Nadahup **Tucanoan**	0.06	**Hibito-Cholon Uru-Chipaya**	0.15
Jivaroan Quechuan	0.20	Arawakan Guaicuruan	0.24	**Quechuan Boran**	0.08	*Rikbaktsa Bororoan*	0.17
Jivaroan Panoan	0.21	Arawan Chonan	0.24	**Chibchan Nadahup**	0.08	*Cariban Bororoan*	0.17
Aymaran Quechuan	0.21	Chapacuran Chonan	0.24	Chapacuran Itonama	0.09	**Uru-Chipaya Aymaran**	0.19
Jivaroan Leko	0.22	**Chocoan** Kanoe	0.25	Itonama **Quechuan**	0.09	**Jivaroan Leko**	0.20
Araucanian Jivaroan	0.22	**Chocoan** Warao	0.25	Kwaza **Tucanoan**	0.10	**Quechuan Hibito-Cholon**	0.21
Chocoan *Je-Jabutí*	0.23	Guaicuruan Kanoe	0.25	Arawakan **Boran**	0.10	Guaicuruan Arawakan	0.23
Chocoan Paez	0.23	**Jivaroan** Kanoe	0.25	Itonama *Tupian*	0.12	**Uru-Chipaya Quechuan**	0.23
Chibchan Panoan	0.23	Kanoe Matacoan	0.25	Arawakan **Chibchan**	0.12	Warao **Boran**	0.23
Barbacoan Quechuan	0.23	*Tupian* Itonama	0.25	Arawan **Nadahup**	0.12	*Tupian Rikbaktsa*	0.24
Kanoe Puinave	0.24	**Araucanian** Urarina	0.27	Arawakan **Nadahup**	0.12	Guaicuruan Matacoan	0.24
Chocoan Leko	0.24	Arawakan *Tupian*	0.27	Kwaza Arawakan	0.12	*Cariban Tupian*	0.24
Chocoan Panoan	0.24	Arawakan **Urarina**	0.27	Arawakan **Tucanoan**	0.12	Warao **Urarina**	0.25
Boran Chonan	0.25	**Barbacoan** Muniche	0.27	*Tupian* Arawakan	0.13	Arawan **Leko**	0.25
Bororoan Chocoan	0.25	Guaicuruan Muniche	0.27	Chapacuran Arawakan	0.13	**Mochica Kallawaya**	0.26
Kanoe Yanomamic	0.25	**Tucanoan Leko**	0.27	Chapacuran *Tupian*	0.13	**Kallawaya Quechuan**	0.26
Leko Quechuan	0.25	Arawakan *Bororoan*	0.28	Boran Chapacuran	0.13	Kwaza **Jivaroan**	0.26
Jivaroan Tacanan	0.25	Arawan Warao	0.28	**Araucanian Tucanoan**	0.13	**Boran** Nadahup	0.26
Bororoan Tupian	0.26	Bororoan Tupian	0.28	Kwaza **Quechuan**	0.13	**Kallawaya Aymaran**	0.27
Barbacoan Chibchan	0.26	Chapacuran Bororoan	0.28	**Chibchan Tucanoan**	0.13	Arawakan Matacoan	0.27
Leko Panoan	0.26	Guaicuruan Itonama	0.28	Arawakan **Quechuan**	0.13	Arawakan Kwaza	0.27
Panoan Quechuan	0.26	Warao Arawan	0.28	Yurakaré **Quechuan**	0.13	**Boran Chocoan**	0.28
Chibchan Chocoan	0.26	Chapacuran Arawakan	0.30	**Leko Quechuan**	0.14	**Leko Tucanoan**	0.28
Quechuan Urarina	0.26	Chibchan Muniche	0.30	Kwaza Arawan	0.14	Itonama Guaicuruan	0.28
Kanoe Kwaza	0.27	Chonan Muniche	0.30	**Araucanian** Kwaza	0.14	Kwaza **Leko**	0.28
Barbacoan Paez	0.27	Guaicuruan Tupian	0.30	**Araucanian** Chapacuran	0.14	Muniche **Leko**	0.28
Leko Tacanan	0.27	Guaicuruan **Urarina**	0.30	**Tucanoan Tacanan**	0.14	*Cariban* Guaicuruan	0.28

(43 out of 60, 21 pairs) already identified as showing much structural similarity overall. In the pairs ranked highest on TAME there are far fewer (11 out of 60, 1 pair) from the western languages, and in the pairs highest on SubOrd somewhat more (23 out of 60, 6 pairs). In the pairs ranked highest on ArgMar we have 31 languages out of 60 (11 pairs); in this domain Bororoan, Tupian, Rikbaktsa, and Cariban rank highly together.

If we assume that Noun Phrase and Argument Marking are the most reliable pointers to deep time relations, many of the western language families in our sample may be ultimately related. This requires much further research. The same assumption would suggest that Tupian, Cariban, and postulated Macro-Jê language families form a grouping, as has been assumed by Rodrigues (1985) on different grounds.

TAME appears to give a weaker signal and the domain of Subordination needs to be explored in terms of more critically differentiating features before it can give sharp insights in this area, for which it certainly has the potential.

The impression that the western languages may show older ties is confirmed by the fact that they pair best on Noun Phrase and Argument Marking. There is evidence too for a structural grouping of the Macro-Jê languages, together with Tupian and possibly also Cariban. There are also pointers to other possible groupings in the data, which need further exploration. As with any hypothesis-generating exercise, there are bound to be spurious groupings in the data as well.

7 Language contact

In the introductory chapter, a number of language contact scenarios were identified as potentially relevant to the South American languages. We will disregard prestige and trading partner borrowing here, since we did not do any lexical studies. However, the other five contact scenarios listed – Substrate and shift, Bilingual convergence due to prolonged coexistence, Metatypy, Koineization and expansion languages, and Intertwining and mixed languages – are very relevant to our findings.

A number of language clusters involving intensive mixture between different Amerindian language varieties will be considered here. Notably, structured varieties belonging to large families are discussed. This is mostly a methodological requirement, since only when several members of the same family can be compared can we talk with confidence about processes of contact and restructuring. For this reason we concentrate on languages involving Tupian, Cariban, Arawakan, and Quechuan.

7.1 Substrate and shift

Undoubtedly there have been many more cases of shift and substrate formation in South America, but two cases were mentioned in this book:

(1) The shift from Aymara to Quechua in the Southern Andes (Cuzco and Puno). In this case several Aymaran phonological, morphological, and discursive features can be identified.
(2) The shift of the Arawakan Chané to Chiriguano (Tupí-Guaranian) in Eastern Bolivia. Further work is needed on internal variation in Chiriguano and a possible Arawakan substrate.

Further cases also involving koineization are discussed below.

7.2 Bilingual convergence due to prolonged coexistence and metatypy

Given the large number of genealogical units, there are many situations in which languages belonging to different families have coexisted for a considerable time period. In a number of cases, this has had structural effects.

A case in point are the Kakua, Nadahup, and Puinave languages in the Colombia-Brazil border area, which do not ostensibly belong to the neighboring large Arawakan and Tucanoan families. Kakua and Nadahup share morphosyntactic features, marriage partners, and cultural vocabulary with the neighbouring Tucanoan languages (Bolaños and Epps 2009) while there is little shared basic vocabulary between any or all of Tucanoan, Kakua, Nadahup, and Puinave (Bolaños 2011; Bolaños and Epps 2009).

Our data confirm the striking convergence between Quechuan and Aymaran in the Central Andes, even to the extent that some variants of Quechua have become structurally closer to Aymara than to other Quechua variants (see Van de Kerke & Muysken, this volume). We find structural similarities with Hibito-Cholón and Uru-Chipaya, although here a specific set of borrowed elements can be identified.

In the Upper Amazon and Andean foothills (van Gijn, this volume) and the Guaporé-Mamoré zone (Crevels & van der Voort 2008), numerous isolates and small families have interacted. By and large they have maintained separate and distinct typological profiles, although a number of more abstract structural traits seem to have diffused to different extents. This may point towards a policy of identity maintenance under contact (see Eriksen 2011).

If our data are correct and the Guaycuruan and Zamucoan families in the Chaco as such are not particularly close, individual members of these families appear to show striking convergence.

A special case of convergence is *metatypy*: the drastic restructuring of a language profile on the model of another language, under sociolinguistic conditions of asymmetric bilingualism (Ross 1999, 2006). There may very likely have been other cases of metatypy but the ones discussed in Van de Kerke and Muysken (this volume) are Puquina in Northern Bolivia, which has undergone influence from Quechua in the early twentieth century, ultimately resulting in Kallawaya, and nearby Uchumataqu on the Bolivian Altiplano, which has undergone influence from Quechua and Aymara.

Table 13.7 *Some well-known expansion varieties in South America*

Family	Languages	Substrate languages	Location
Tupian	Nheengatú	Arawakan, Tucanoan	Currently Rio Negro, Brazil, formerly much wider distribution
Tupian	Cocama-Cocamilla-Omagua	Arawakan, Quechua (pidgin), . . .	Amazon in the border region of Brazil and Peru
Quechuan	Ecuadorean Quechua or Quichua	Barbacoan, Jivaroan	Ecuadorean highlands and lowlands

7.3 Koineization and expansion languages: are there South American indigenous Creoles?

In this book, three major language expansions were surveyed: the expansion of Arawakan roughly between 1000 BCE and 1200 CE, the subsequent expansion of Tupian 1–1600 CE, and finally that of Quechuan out of south central Peru, roughly in the period 500–1600 CE.

In all three cases, varieties were brought far away from the original homeland of the language family. Thus the *social scenario* of these expansions resembles the expansion of the European languages in the colonial period and their transformation into pidgins and Creoles in the setting of the slave trade and plantation economies.

In Eriksen and Galucio (this volume) and Van de Kerke and Muysken (this volume) several language varieties are mentioned which have undergone drastic restructuring, with consequences for their typological profile and position in the language family, as in Table 13.7.

Thus in terms of their *structural features* several expansion languages in the Tupian and Quechuan families may be relevant for the study of Creoles. Can we include the South American expansion varieties in the list of Creole languages?[4] If we could, this would expand and broaden the typological and regional database for Creole studies.

Two caveats are in order. First, we should note that not all language expansion is accompanied by the kind of major restructuring associated with Creole genesis. First of all, numerous varieties of the Tupí-Guaraní family and numerous

[4] There are undoubtedly other cases we could have discussed here, such as the Akuntsu-Kanoê Pidgin reported for Rondônia by Van der Voort (p.c.), but there the documentation is very limited. We also do not discuss possible pidginized or mixed varieties in which Portuguese and Spanish play a major role; these are very important, and particularly the Amazonian varieties need to be studied much more (e.g. as in Adelaar with Muysken 2004).

Quechuan varieties such as those in Bolivia and Argentina have spread without reduction and restructuring. There is evidence of reduction and regularization, but on a smaller scale (Kusters 2003).

Second, the spread of the Arawakan family has often been associated with ethnogenesis, as in Eriksen and Danielsen (this volume). However, this ethnogenesis, a cultural development parallel to creolization, is not accompanied by the reduction in Arawakan characteristic of known Creole languages. In earlier work (Danielsen et al. 2011), and the studies by Aikhenvald (2002) and Seifart (2011), language contact involving Arawakan is treated in more detail, but much of it involves borrowing rather than intensive reduction. However, it may be that we are not yet looking at the Arawakan family with the right analytic glasses on. Its situation may reflect a scenario of sustained contact, in contrast with the rapid expansion of Tupian.[5]

However, the cases listed in Table 13.7, involving a social history of expansion and a structural history of reduction, are highly relevant for Creole studies, because this field has often made universalist claims about language and its essential properties. The import of most of these claims is limited by the fact that they are based on a typologically skewed set of languages, the canonical European-lexifier Creoles of the Atlantic and the Pacific. These Creoles are related to their western European lexifier languages, and have as substrates some Kwa and Bantu languages, and a few languages from the Pacific. Most of the contributing languages have little morphology. If we could expand the database of languages that have undergone creolization to include languages with more morphology, this would strengthen the field of Creole studies immensely.

Making these comparisons explicit could also help to elucidate processes of language mixing and contact in South American Indian languages. Many scholars working on the languages of South America are keenly aware that some of the languages they study do not directly fit into classical genealogical trees or that they show unusual patterns of change from their putative ancestors (cf. e.g. Cabral 1995, 2007 with respect to the Tupian expansion variety Cocama). However, they often feel the need for a more developed inventory of concepts to describe and analyze these special cases. Placing the languages involved in the framework of Creole studies at least helps elucidate and systematize some of the features characterizing these languages. However, to study the varieties in Table 13.7 as Creoles, a specific definition of "Creole" is needed, neither purely sociohistorical, like Mufwene (2002), nor purely structural, like McWhorter (2005).

[5] Thus there are many differences between the Arawakan and Tupian expansions, particularly in terms of the way these two language families traditionally interacted with their neighbors, but there are also similarities in terms of the way these language families were able to adapt and adopt in new scenarios of contact.

Mufwene (2002: 11440) takes the position that pidgins and Creoles (PCs) should be defined strictly historically in terms of the European expansions:

> Strictly speaking, PCs are new language varieties, which developed out of contacts between colonial nonstandard varieties of a European language and several non-European languages around the Atlantic and in the Indian and Pacific Oceans during the seventeenth to nineteenth centuries.

Mufwene (2001: 178) justifies this limited definition by arguing that other expansion varieties, in Africa, such as Kituba, Lingala, and Sango, are "restructured varieties" rather than Creoles, and cites Mazrui and Mazrui (1998), who claim that Swahili and Lingala enable "horizontal integration" of their speakers, in contrast to the colonial European languages, which have putatively established "vertical integration," i.e. more social stratification. However, it is likely that in South America there were cases of asymmetric imposition as well, as in the Inca Empire. This invalidates Mufwene's limited definition on the social dimension.

One alternative to Mufwene's purely historical perspective is to adopt a purely typological definition (McWhorter 2005), where a Creole is defined through a specific set of structural features characterizing the Creole Prototype: inflectional affixation is extremely rare or nonexistent, tones are not used to encode morphosyntactic distinctions, and all derivation is compositional.

The problem with this definition is that it requires an intuitive list of Creole languages to start with and hence is circular: the original class of Creoles is delineated in terms of the perception by linguists of specific typological features in a class of languages. A language such as Cocama will automatically fall outside of the definition because of the implicit criteria scholars have used to label languages as "Creole." Even though it shows signs of "frozen" and reduced morphology (Cabral 1995), it also has productive inflectional and derivational affixes, even if less varied than its Tupian ancestors.

Thus both a purely historical approach and a structural approach are problematic. We advocate a relational approach, the scenario approach sketched in Muysken and O'Connor (this volume). A scenario is a specific set of circumstances in which languages in contact are modified in specific ways. The definition of Creole should involve the relation between an initial linguistic state and a final state, as well as with the circumstances responsible for the transition. A Creole results from the modification of the typological properties of a specific language when confronted with other languages under specified circumstances. A Creole with French lexicon can then be a very different language typologically from a Creole with Quechua lexicon, even though the processes of restructuring involved can be defined in universal terms.

7.4 Intertwining and mixed languages

Possibly every language in the world is mixed in the sense that it contains elements from more than one genealogical source, as pointed out by scholars like Hugo Schuchardt and later Givón (1979). Nonetheless, in many languages the number of words demonstrably not inherited from a direct ancestor is limited (Van Hout and Muysken 1994). Although it is hard to find exact figures on this, a proportion of more than 40 percent of core vocabulary would be exceptional, and even 20 percent would be noteworthy (Greenhill and Gray 2012: 528). Furthermore, cross-linguistic influence in the domain of morphology ("affix borrowing") or syntax ("borrowing of syntactic patterns") has been argued to be much rarer (e.g. Muysken and O'Connor, this volume). There have been several attempts to separate languages showing heavy borrowing from "mixed languages" defined in terms of notions such as intertwining or relexification.

"Mixed language" for us is a heuristic term rather than a theoretical construct. There may well be other cases, but the only example of an intertwined truly "mixed language" discussed in this book is Kallawaya, the ritual language from northern Bolivia with both Quechuan and Puquina components (Van de Kerke and Muysken, this volume).[6]

8 General conclusions and suggestions for further research

Overall, the use of grammatical features as a way of charting possible relationships between the families of South America was fruitful. We were able to replicate already established families, and show larger patterns in the data. The question remains, of course, how these large patterns can be explained. Do they reflect early large genealogical units or areal effects?

Drawbacks in trying to answer this question are that there are not enough comparable data from outside of South America, and that our language sample was rather broad, and not dense either geographically or genealogically, with a few exceptions. Although our database allows this, we have not yet pursued the study of the regional spread of individual features or smaller clusters of features.

We hope at least this volume and the database associated with it open the way to much new research. Altogether a multidisciplinary approach to changes in the languages of the continent needs to be adopted, taking into account variable rates of change, and differentiation between pulse periods (leading to more tree-like configurations) and pause periods, leading to complex networks. Moreover,

[6] We are disregarding here, as elsewhere, the numerous cases of language mixing involving European and possibly also African languages.

a systematic comparison with geographic, archaeological, ethnohistorical, and genetic data is required to shed light on how the diversity patterns of the continent came to be and how human behavior and linguistic change are tied together.

In the previous sections some of the global findings of our research were presented. There is an overall east/west division, typologically, which may reflect settlement patterns and deep genealogical relations, and possibly linked to genetic profiles associated with high population densities. Our database allows much further research where individual language pairs are explored in more detail for which little structural distance was found, as e.g. for Kwaza and Nambikwaran. Just like the ASJP database, our database can function as a hypothesis-generating device for exploring further relationships.