Linguistic Phylogenetics of the Austronesian Family:
A Performance Review of Methods Adapted from Biology

Arpiar Saunders
B.A. Thesis
Department of Linguistics • Swarthmore College
2005

Dedication

I don't know whether it is appropriate to dedicate a B.A Thesis. If it is, I dedicate this thesis to David Harrison and Robbie Hart, my friends and mentors. Thank you both for teaching me so much about language; I have enjoyed our teamwork immensely.

Table of Contents

0. Abstract

Four methods for inferring biological phylogenies were applied to lexical and structural data of a representative sample of the Austronesian Family of languages. After introducing individual languages and the Family as a whole, each combination of method and data type is performance reviewed through topological comparison with a 'known' tree. The results suggest a two-step method which is described in detail. First, NeighborNet analysis is used to qualitatively assess how "phylogenetic" the data are and thus if tree building is justified. Next, Bayesian analysis is used to construct a tree. Under the proposed method, a combined lexical and structural data set produced a fully historically accurate tree, thus supporting past research through an alternative method. The increase in accuracy with combined data suggests that inferring the natural history of the whole language depends on reconciling the phylogenetic signals from component parts; a tension between the lexicon and structures with traceable correlates in both methods. Lastly, the evolutionary association of structural features is assessed. This result highlights the potential productivity of using biological methods to pursue previously untenable questions about language evolution.

I. Linguistic Phylogenetics: An Introduction

Languages and organisms both diversify over time through an evolutionary process. And while the most salient changes occur at the level of the whole language or species, inferring the natural history of the whole depends on comparing representative parts. Until the late 1950's, biological ancestral relationships were established through comparisons of morphology and behavior. Post-molecular revolution however, the near limitless amount of comparative DNA, RNA and protein data were assumed to establish, with great resolution, the evolutionary history of life. Surprisingly, for many species the trees of relatedness defined by molecules and morphology disagreed. Since then, biologists have struggled to develop strategies to reconcile incongruent natural histories. Their progress has led to a robust literature, well-oiled computational machinery, and many strongly supported phylogenies.

Linguistics, however, has had no revolution of data type to catalyze novel approaches; "the comparative method" is still used to establish language family sub-groupings through shared innovation (Pawley and Ross, 1993). But like organisms, languages are component systems, comprised of phonology (sounds), lexicon (words),

and morpho-syntax (structures)(figure 1). As such, the methods used to compare and resolve the component histories of an organism can be applied to those of language.

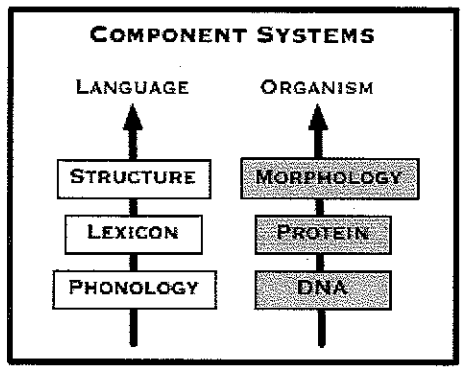


Figure 1. Components of Languages and Organisms

But why does an accurate natural history of the whole depend on reconciling the histories of the component parts? It should be possible to rely on the single most informative sub-component of either system to tell the entire story. While it is true that some categories of comparison are better than others due to resilience against 'borrowing,' the necessity for component congruence stems from inter- and intra-differences on the *constraints* of evolution[i]. For example, areas of the genome coding for proteins that perform the most basic and essential functions of life (e.g. metabolism, cell replication) are relatively intolerant of change, while non-coding areas can sustain large quantities of mutations without affecting the success of the organism. Similarly in language, words for the most basic and essential descriptions of a certain lifestyle are in general less likely to be borrowed than words for new or peripheral items (Atkinson *et al.*, 2005). In terms of phylogenies, the more constrained sub-components are helpful for inferring ancient relationships "deep" in time, while unconstrained sub-components help resolve more "shallow" recent histories. Put together, natural histories based on consensus allow individual components parts to resolve their respective time-scales while leaving areas of contradiction unsupported. This approach has been called 'Total Evidence' in the phylogenetics literature (De Queiroz *et al.*, 2005).

Historically, hypotheses of "deep" language relationships have been controversial. While intra-family groupings can be successfully established by comparing sound change and lexical similarity, both data types are dependent on the presence of cognates or "homologous" words. The lexical component of language, like an unconstrained gene, may lose and acquire forms at a rate that masks similarity after a certain time period - approximately 6,000 to 8,000 years – a depth that is "shallower" than most families (Gray, 2005;Warnow, 1997)(figure 2).

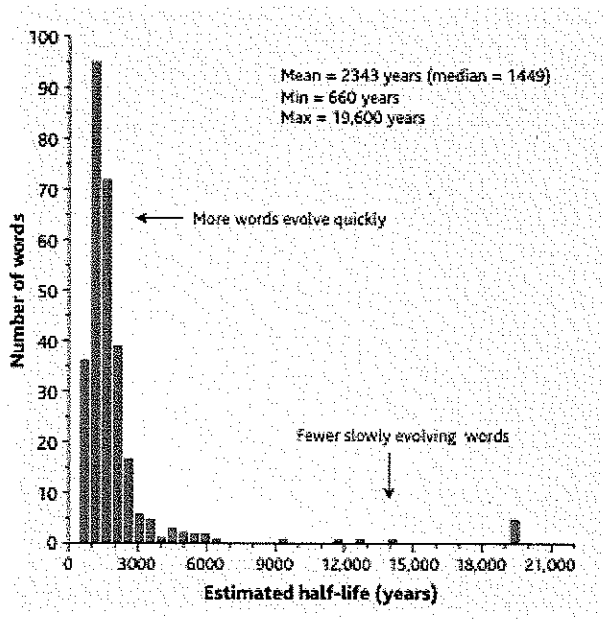


Figure 2. A graph showing the projected half-life for 200 lexical meanings. Dates were generated from a model based on distribution of rates for lexical evolution (Pagel, 2000)(taken from Gray, 2005).[ii]

Inferring "deep" relationships of language must then depend on a more slowly evolving data type; unfortunately, there is only one candidate component left – structural data – which though promising, has a host of theoretical and practical caveats.

Structural data can be roughly defined as the grammar of the language. Unlike words that form an inexhaustible or "open" class, language structures form a finite or "closed" class. And while the boundaries of a word are self-evident, defining structural features is more artificial; that is, categories must be constructed which accurately compare and contrast strategies from a diversity of grammatical systems. Luckily, the relevant grouping of language structures is the mandate of language typology and over

years of research, typologists have developed robust and productive parameters of comparison.

The most major confound of structurally driven phylogenies is 'borrowing', when shared similarity arises from contact instead of ancestry; this distinction is often difficult to reconcile however, since both processes of change are driven by social diffusion (Enfield, 2005). With this similarity in mind, the psychological and functional quality of the characters of comparison must be clearly stated. In specific, while lexical items are also susceptible to borrowing, the 'closed class' nature structural of data (in other words, the grammatical limits of human language) necessitates comparing a small, discreet number of possibilities for a given feature, instead of a much larger number of possibilities for the realization of a word. For example, there are only six possible orders of the Subject, Verb and Object, while there are thousands of words for "head." Thus structural data produces fewer data points, which means and 'borrowing' becomes harder to detect. Features that have been substantially borrowed are called 'areal', and despite the difficulty, their identification and/or exclusion is necessary to maintain the integrity of the phylogenetic signal.

The challenge of detecting 'areal' effects has a non-trivial parallel in biology: for bacteria, which constitute the largest and most abundant kingdom on Earth, borrowing (technically called horizontal transfer) of genetic material between species is thought to be a major player in evolution (Ochman, Lawrence and Groisman, 2000). Biological strategies, in conjunction with extra-linguistic information about the prehistoric interactions of humans (e.g. archaeological and genetic evidence), may help historically locate possible contact situations and those features involved (Curnow, 2001). While difficult to determine, areal influences do not undermine the possibility of structural phylogenies: since the spread of individual features differs, areal effects across a diverse set of features should only make the phylogenetic signal more "noisy" rather than inaccurate (Wichmann, personal communication).

This report attempts to introduce the power of biological phylogenetics for inferring the ancestral relationships of languages. By drawing appropriate parallels between evolutions, the computationally driven approaches used by biologists can be productively adapted to linguistic data. The methods outlined are not intended to replace

traditional methods, but rather to provide an additional tool to verify established groupings and generate new hypotheses. I present these tools through their application to a well-understood sample of 26 Austronesian languages in hopes of satisfying three goals: 1. To replicate the agreed upon grouping of these languages; 2. To demonstrate qualitatively and statistically that lexical and structural data should be used in compliment; and 3. To use the diversity of biological software to generate wholly new - and potentially meaningful – data concerning language evolution. The Austronesian Family, due to its large size, structural diversity, data availability, established history, and nature of dispersal presents an ideal situation to pursue the stated goals.

This report is structured in five sections. Following the introduction (I), the Austronesian language family will be introduced through its history and geographic distribution, along with the experimental sample of languages (II). Next I will describe how the data were collected and encoded (III). Thirdly I will describe the basic phylogenetic methods and their resulting applications to the lexical and structural data sets (IV). Based on the results in IV, I suggest a basic but detailed methodology for phylogenetic analysis and interpretation of linguistic data (V). Lastly, I demonstrate how the questions and applied solutions of evolutionary biology can be used to tackle new questions and produce new types of results for historical linguistics (VI). The final section is a conclusion (VII).

II. The Austronesian Language Family and Experimental Sample

The Austronesian Family (AF) represents a substantial portion of the world's linguistic heritage in terms of number and space. The Ethnologue defines 1268 languages in the family, comprising 20% of the world's total, manifesting a speaker population of approximately 270 million. In terms of surface area, the AF falls only behind Indo-European, with a latitudinal expanse from Madagascar to Easter Island (2/3 of the world's circumference!) and a longitudinal distance from New Zealand to Hawaii (Adelaar, 2005)(Pawley and Ross 1993). How has one language family come to inhabit such a vast and oceanic habitat?

The story of the AF is the audible reflection of a prehistoric expansion in search of land suitable for agriculture. The original Austronesian speakers (AS) were agriculturalists who migrated out of modern-day Taiwan around 6,000 years ago. The original migrants splintered, following two different paths. One was south-west, through Borneo, Indochina and eventually the Malay Peninsula; this group spoke what would become the Western-Malayo-Polynesian (WMP) languages. The second movement was more directly southern through the Philippines, Indonesia, and Melanesia; this group was responsible for those languages classified as Eastern-Central Malayo-Polynesian (CEMP).

These migrations brought the AS to both previously inaccessible new areas and to old areas, inhabited in some cases for more than 40,000 years (Pawley and Ross, 1993). The indigenous people encountered in Melanesia were taro farmers, living predominantly in the mountains and speaking languages wholly unrelated to Austronesian. While the AS never penetrated far inland, contact between the two cultures had linguistic repercussions, as features diffused between the speaking communities (Ross, 2001). Other offshoots of the AS developed technology for long distance navigation and spread predominantly eastwards through the Solomons toward the virgin islands of Oceania – and into, for the most part, linguistic isolation.

While the archaeological and genetic evidence continues to clarify the murky details of Austronesian expansion, the need to understand Austronesian history from a linguistic perspective is clear. Not only for evaluating the historical claims of archaeology and genetics, but as a model system for understanding language evolution, both contact induced and stochastically (randomly) driven.

i. History of Austronesian Linguistics

The identification of the AF, originally coined as Malayo-Polynesian, most probably originates with the Dutchmen Hadrian Reland in 1706. The first printed work referencing the Malayo-Polynesian Family was in Wilhelm von Humboldt's *Über die Kawi-Sprache auf der Insel Java* (1836-1839) (Tryon, 1995). Comparative reconstructions were first undertaken by H.N. van der Tuuk in the 1860's, but not until

the fieldwork of Otto Dempwolff in the early 1900's were major strides achieved. Dempwolff not only reconstructed 2215 Proto-Austronesian lexical items, but also assembled the constituent sounds of the Proto-Austronesian, Proto-Oceanic and Proto-Polynesian phonologies, establishing these new groupings in the face of older geographic ones (Tryon, 1995).

Comparative research on Austronesian exploded in the 1960's, with large additions and modifications to Dempwolff' original work by Dyen, Grace and Dahl. Recently, Robert Blust (1978) has hypothesized the most detailed sub-grouping scheme, which, though hotly debated, is the most accepted amongst scholars (Tryon, 1995). In this analysis, the groupings presented in Tryon (1995), which follow Blust (1978,1990) for the highest order grouping and Ruhlen (1987), Grimes (1990), Ross (1988) and Lynch and Tryon (1985) for the more specific groupings of the WMP, CMP, Oceanic, and Central-Eastern Oceanic respectively, are used as a 'known' reference for the evaluating the accuracy of the experimental phylogenies.

ii. Blust's Sub-Groupings and the Dynamics of Dispersal

Blust's highest order distinction seperates three Formosan sub-groups (Atayalic, Tsouic, and Paiwanic), spoken on Taiwan, from a group with everything else, called Malayo-Polynesian. The Malayo-Polynesian group is divided into Western and Central-Eastern. Central-Eastern is further divided into its Central and Eastern components, with the Eastern group again split in two: the South Halmahera-West New Guinea group and the Oceanic group. Figure 3 is a simplification, only showing those groupings from which sample languages were taken.

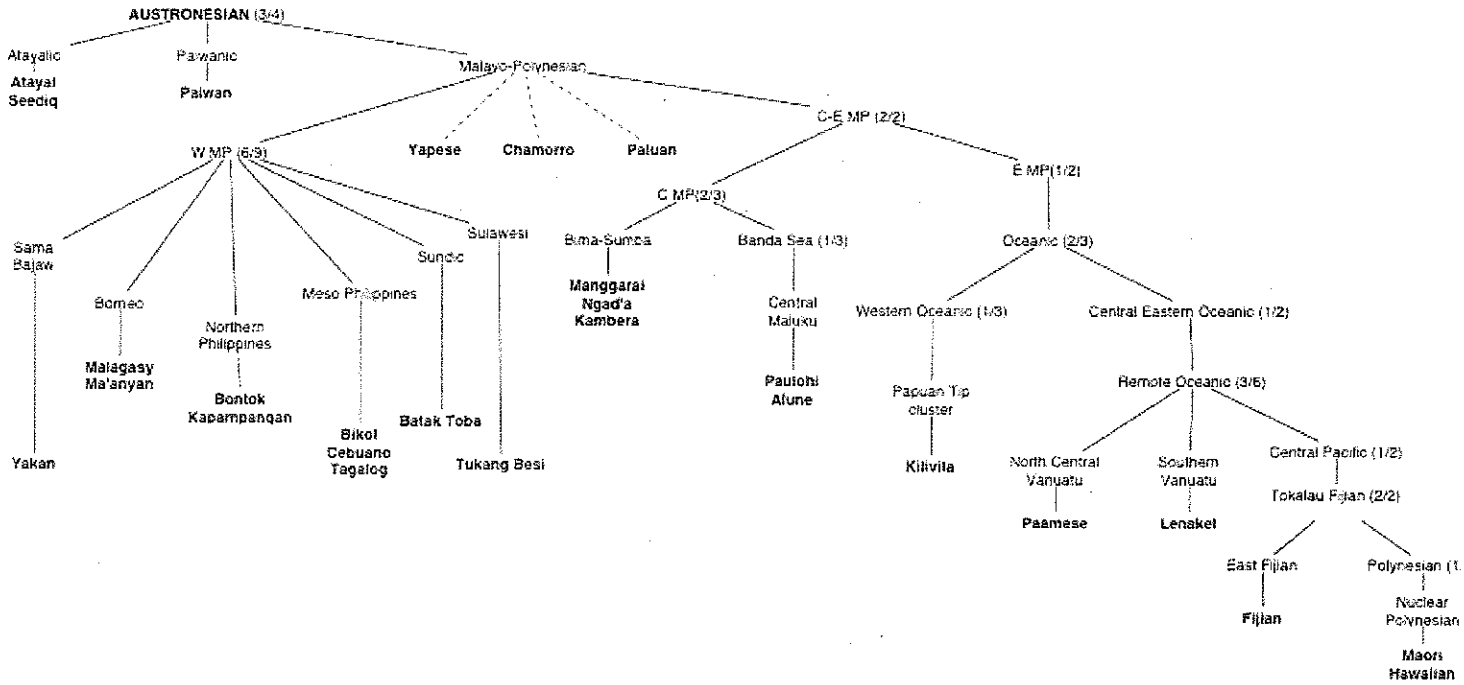


Figure 3. Blust's (1983) Sub-Groupings tailored to represent the language sample (in red). The numbers in parentheses represent (the number of branches shown/the total number under the node) (adapted from Tryon, 2005).

In his review of Austronesian sub-groupings, Adelaar (2005) makes it clear that support – and scholarly acceptance – is highly variable. In fact, the term 'sub-grouping' as a catch-all is itself called into question by Ross (1995), who suggests we reserve the term for instances of speaker group separation, as opposed to gradual dialect differentiation, which Ross (1995) coins as 'linkage.' In the natural history of Austronesian, differentiation through 'sub-grouping' versus 'linkage' follows a regular pattern driven by a continuous migration. In the model Ross (1995) puts forward, each hypothesized a node consists of a "stay-at-home" group, which as the name implies, remained in the area settled while a second group, which I will call the "movers," continued on their path; the languages of the "stay-at-home" group would theoretically evolve through linkage, while the "movers" would evolve through separation (figure 4).

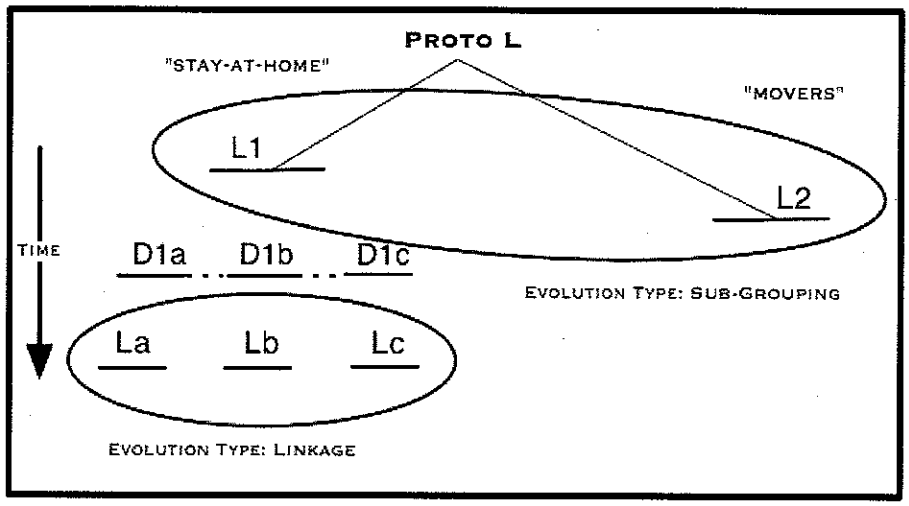


Figure 4. A schematic of Ross's (1995) model for the evolution the Austronesian Family

The differential effect of these two evolutionary scenarios has an impact on the eligibility of sub-grouping. The main difference concerns the presence or absence of an internal hierarchy, since "stay-at-home" languages may not have a reconstructible higher-order (Ross 1995). In the case of the most primal Austronesian node, Ross (1995) suggests that a "stay-at-home" situation occurred in Taiwan before the speakers of what would be Proto Malayo-Polynesian left; it is therefore justifiable to ask from which Formosan language Proto Malayo-Polynesian sprung. Preliminary evidence suggests Amis and Bunun as the most likely candidates, because of a unique sound merger (*C and *t) shared with Proto Malayo-Polenesian.

iii. Austronesian Language Groups and Sample Language Descriptions (figure 5)

Formosan Languages

Both Proto-Austronesian (pAN) and Proto-Malayo-Polynesian (pMP) have been substantially reconstructed, especially in terms of lexicon and phonology. The existence of credible reconstructions supports strongly the division between the Formosan and Malayo-Polynesian. The internal groupings, as stated above, are much less accepted and may not have higher-order resolution (Ross 1995). There is a relatively accepted difference between the Atayalic languages and all others, with another more tentative division of non-Atayalic languages into Tsouic and Paiwanic groups. One of the most pressing questions surrounding the internal hierarchy of the Formosan languages has to do with potential relationships to Proto Malayo-Polynesian; more explicitly, is the language that evolved into pMP still spoken in some form on mainland Taiwan or did the entire speaker community migrate?

Other attempts have been made to associate the Formosan languages with other groupings. For example, some scholars (Dyen and Tsuchida, 1991)(Wolff, 1995) have suggested that the Formosan and "Philippine languages" should be grouped together, on account of similarities in lexicon and morphosyntax. This suggestion is highly controversial, as are the nebulous boundaries of the "Philippine languages" in general (Adelaar, 2005).

The Formosan languages included in the sample are the following, listed by tentative sub-group: (This structure for presenting descriptions of the language sample is generally followed for all groups.)

Atayalic: (Atayal and Seediq)

Atayal

Atayal is spoken by the second largest indigenous group in Taiwan. They populate the mountainous northern region. The Ethnologue (2005) reports ~84,000 total speakers and two dialects, Sqolyeg and Ts'ole'. Atayal and Seedig are very closely related and together comprise two of the ten Formosan languages with over 1,000 speakers (Rau, 1992).

Seedig

Seediq is spoken slightly south of Atayal, in the valley regions running from central Taiwan to the Pacific. The speaker community consists of two main dialects, Eastern (Toda-Truku) and Western (Paran-Tongan); these dialects have recognized differences in terms of phonology, syntax and lexicon, though Paran is the standard dialect for the standard Seediq orthography. The Ethnologue (2005) identifies 4,750 speakers as of 2002 (Holmer, 1996).

Paiwanic: (Paiwan)

Paiwan

Paiwan is a language of Taiwan, spoken by some 67,000 people in the Southern area of the country (Ethnologue, 2005). There are five ethnic groups identified within the Paiwan people, each with their own distinguishing cultural traits, but no information is reported on dialect variation (http://edu.ocac.gov.tw/local/tour_aboriginal /english/a/07.htm).

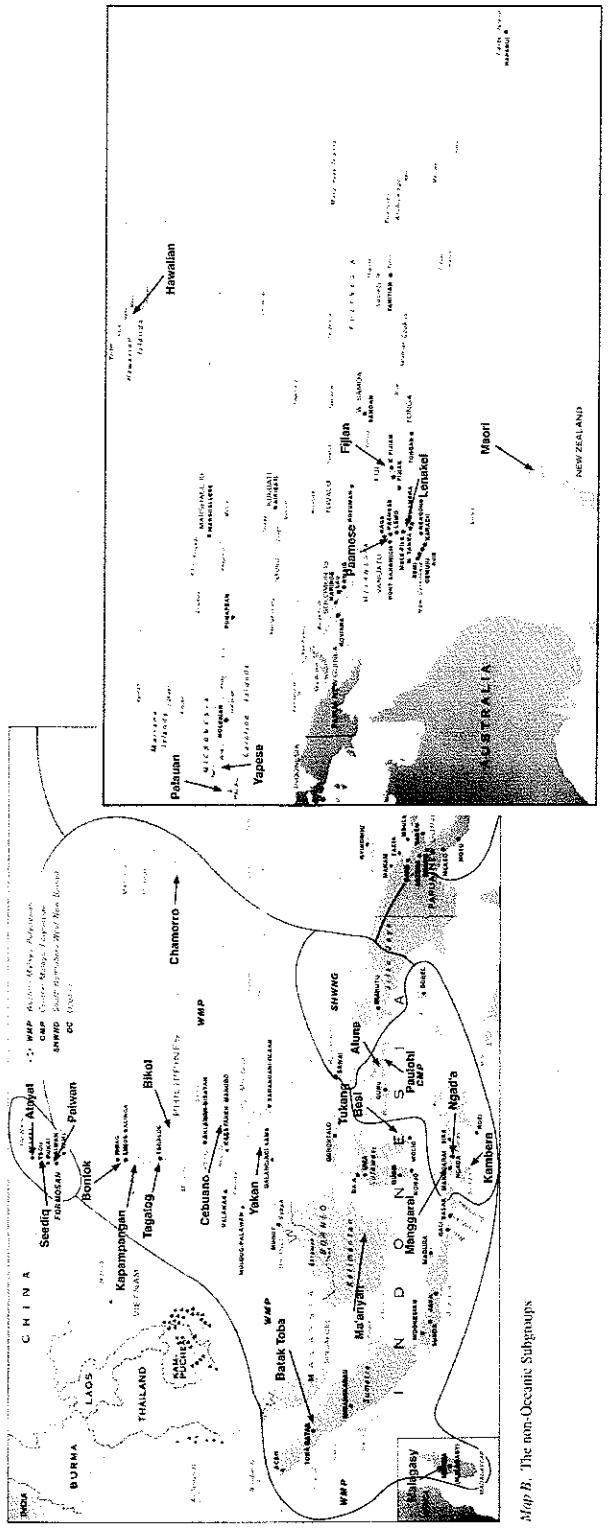


Figure 5. A map of the geographic domain of the Austronesian Family. Languages used in this analysis are listed in red (adapted from Tryon, 1995).

MP Languages

As mentioned above, pMP is a reconstructible and accepted sub-group. Ross (1995) mentions two important MP innovations that are absent from the Formosan languages: 1. a pronoun shift to reflect the polite possessive and 2. a derivational prefix which pivots the actor and undergoer (Blust, 1977; Dahl, 1976 and Reid, 1982; from Ross, 1995). Phonologically, pMP experiences several characteristic phonemic mergers from pAN, including *C and *t → *t and *L and *n → *n, as well as the development of a palatal nasal *ñ (Blust, 1990; from Ross, 1995).

The MP outlier languages, which lack support for a more detailed grouping, are as follows:

MP Outliers: (Chammorro, Palauan and Yapese)

Chamorro

Chamorro is the national langue of the Guam and is in active use. It is spoken by some 62,500 speakers on the island with an additional 1,400 speakers off-island (Ethnologue, 2005). There are dialect differences, but all mutually comprehensible. Spanish has had considerable influence on the Chamorro lexicon and sound system (Topping, 1973).

Palauan

Palauan is the language of the Palauan Islands in Micronesia. Palauan has a speaker community of ~ 15,000, with little dialect variation (Ethnologue, 2005). There are also speakers of Palauan on Guam.

Yapese

Yapese is spoken by ~6,600 people on the four reef-enclosed islands of Yap (Ethnologue, 2005). Yap is situated in the Western Caroline Islands and is one of the four Micronesian states. Yapese has loan words from Spanish, Japanese, German and English (Hsu, 1969).

WMP Languages

Little is known about the genetic affiliations within the WMP subgroup. Ross (1995) chooses to further differentiate the 9 sub-groups of Blust (figure 3) to avoid presumptions of higher order relationships. I will only refer to those sub-groups of Ross (1995) which contain sample languages (Appendix Table 1). I will use the following format:

General Geographic Position: Sub-Groups of Ross (1995) (sampled language/s)

Philippines: Meso-Philippines (Bikol, Cebuano, and Tagalog) **Northern Philippines** (Bontok, Kapampangan)

In attempts to define the members of this sub-group, Ross (1995) states that, "students of Philippine languages have traditionally spoken of a "Philippine" sub-group that includes all languages of the Philippine archipelago (except Sama-Bajaw "sea gypsy" languages), the languages of the Batan Islands between the Philippines and Taiwan, and several groups of languages spoken in the northern arm of Sulawesi, namely Sangiric, Minahasan and Gorontalo-Mongondic" (p.73). Attempts to reconstruct a Proto-Philippine have met with much scrutiny, and some analyses negate this possibility by separating the Southern and Northern Philippine group from the Meso Philippine group (Reid,1982; from Ross,1995).

Bikol

There are around 3.5 million speakers of Bikol languages in the Philippines (Ethnologue, 2005). The Ethnologue (2005) defines 8 distinct languages situated in three sub-groups: Coastal, Inland and Pandan. These languages are spoken on the southern peninsula of Luzon island, along with areas in the provinces Catanduanes and Masbate; the mainland provinces with speaker communities are Camarines Norte, Camarines Sur, Albay and Sorsogon (Mintz, 1971b). The Bikol language chosen for this study is Central Bicolano, which alone has ~2.5 million speakers and part of the Coastal sub-group (Ethnologue, 2005).

Cebuano

There are approximately 20 million speakers of Cebuano, making it one of the two largest languages of the Philippines (Ethnologue, 2005). The Ethnologe (2005) identifies 5 separate dialects, one of which, Boholano, may be considered a distinct language. Cebuano is spoken mainly in Central and Southern Philippines, with communities on the islands of Negros, Cebu, Bohol, Visayas, as well as in parts of southern Mindanao (Valkama, 2000).

Tagalog

In the Philippines, Tagalog is spoken by around 24% (~17 million) of people as a first language and some 40 million as a second language. Tagalog is the lingua franca in Manilla and the most dominant language on the main island of Luzon. Tagalog also forms the base from which Filipino, the national language of the Philippines, has been crafted. Tagaloga speakers can be found throughout the world, including the U.S, Canada, Saudi Arabia and the U.K. (Ethnologue, 2005).

Bontok

Bontok is a set of dialects spoken by ~ 40,000 in the central mountain province of Luzon, Philippines (Ethnologue, 2005).

Kapampangan

Kapampangan is spoken by ~ 2 million people living on the central plain of Luzon, Philippines (Ethnologue, 2005). Though the center of the language is the Pampangan province, it also has speech communities in the Tarlac, Nueva Ecija, Bulacan and Bataan (Forman, 1971). Bilingualism with Tagalog is the norm.

Borneo and Madagascar: East Barito (Malagasy, Ma'anyan)

Evidence suggesting that the East Barito group forms a genetically distinct sub-group from the other languages of Borneo comes from phonology (Dahl, 1977; from Ross, 1995). Malagasy is spoken on Madagascar and is most probably the result of a migration in the 7th century AD (Adelaar, 1991; from Ross, 1995).

Malagasy

Malagasy is the sole Austronesian language spoken on Madagascar, with a speaker community of ~ 14 million people. There are a number of similar dialects, divided into eastern, western and intermediate, defined primarily through phonemic distinctions. Malagasy has adopted loan words from contact with Swahili, Sanskrit, Bantu, English and French, among others. One of the southeastern dialects was written in Arabic script by at least the 15th century, but in 1820 the Malagasy king Radama I choose to adopt the Roman script (Rasoloson and Rubino, 2005.)

Ma'anyan

Ma'anyan is spoken by ~150,000 people in Southern Indonesia, in the area of the Patai River drainage (Ethnologue, 2005). Many Ma'anyan speakers are bilingual with Banjarese, a dialect of Malay (Gudai, 1988).

Sulawesi: Muna-Buton (Tukang Besi)

Muna-Buton sub-group is spoken on the islands South-East of Sulawesi. The Muna-Buton and Central Sulawesi sub-groups are tenuous, while the Central Sulawesi sub-group has been well established (van den Berg, 1991; from Ross, 1995).

Tukang Besi

Tukang Besi is spoken by ~130,000 people in the Tukang Besi Archipelago, located off of southeast Sulawesi, Indonesia. The Ethnologue (2005) recognizes two dialects. Some speakers are also bilingual in Wolio Ethnologue (2005).

Sumatra: North-West Sumatra/Barrier Islands (Batak Toba)

Despite apparent structural diversity, strong comparative evidence has demonstrated the genetic affiliation of the languages spoken off the South West coast of Sumatra (Adelaar, 1981; from Ross, 1995).

Batak Toba

Batak Toba is one of the five dialects/languages of the Batak group. It is spoken by ~2,000,000 in Northern Sumatra. As of the late 1950's, Batak Toba speakers had exposure to Dutch, English, and Indonesian, but all were considered 'foreign' languages.

The original Batak Toba script, a version of the Devanagari alphabet, which was used for works on mythology, astrology and magic, has subsequently been abandoned for the Roman alphabet (Nababan, 1981).

Central and Southern Philippines: Sama Bajaw (Yakan)

The sub-group Sama-Bajaw contains the languages of "sea gypsies;" Proto Sama Bajaw has been reconstructed (Pallesen, 1985; from Ross, 1995).

Yakan

Yakan is spoken by some 105,000 people in the Southern Philippines, around Basilan Island in the Sulu Archipelago and around the coastal areas of the Zamboanga peninsula. The Yakan are muslim with heavy influence from the Qur'an (Brainard and Behrens, 2002; Ross, 1995). Yakan is well established and in substantial use.

Central Eastern MP Languages (CEMP)

Ross (1995) suggests 3 paths of migration through which the pMP speakers penetrated the Indo-Malaysian archipelago: through Borneo, through Sulawesi and through Halmahera. The pMP speakers whose language would eventually evolve into pCEMP probably took the most directly southern of these routes, through Halmahera between 3,000 and 2,500 B.C. The split between Central and Eastern language communities took place as CMP speakers headed further south around 2,000 B.C.

CMP Languages

All three of the CMP sample languages belong to the Bima-Sumba sub-group. In total 7 sub-groups are currently recognized (Appendix Table 2).

Eastern Tip of Java: Bima-Sumba (Kambera, Manggarai, Ngad'a)

All three languages are spoken on or around Sumba Island, which sits at the linguistic border between the CMP and WMP languages.

Kambera

Kambera is a language of Eastern Indonesia, spoken on the Eastern part of the island of Sumba by ~ 235,000 people (Ethnologue, 2005). Klamer (1998) suggests there is no standard dialect, but does suggest a dialect or language relationship with a number of other tongues spoken on Samba. Samba is a rural island consisting mostly of farmers. While most towns have primary schools, the nearest university is on the island of Timor, some 350 km away (Klamer, 1998).

Manggarai

Manggarai is spoken in Eastern Indonesia, in the north-central and western parts of Flores Island by ~500,000 people (Ethnologue, 2005.) Five dialect groupings are recogonized, consisting of forty-three subdialects. The Central dialect is the largest, with a speaker population of ~ 300,000. Manggarai has only been slightly impacted by Dutch. A much larger influence was the Makassak language of South Sulawesi, whose speakers exerted political control over the Manggarai population until the middle of the 18th century (Verheijen and Grimes; in Tryon, 1995). Verheijen and Grimes (1995) note that the Manggarai people are predominantly inland oriented and agricultural.

Ngad'a

Like Manggarai, Ngad'a is also spoken on Flores Island, though by a smaller population of 60,000 – 70,000 living on the south-west coast. There are six recognized dialects for Ngad'a.

Central Maluku: Paolohi and Alune

The Central Maluku grouping, located in eastern Indonesia, is comparatively well-established, although debate continues to surround whether the identifying features of the group are shared by more languages in the area (Collins, 1983)

Paulohi

As of 2005, there are only 50 speakers of Paulohi left. The majority of the population was killed by a severe earthquake and tidal wave (Ethnologue, 2005.)

Alune

Alune speakers, located on the western side of Seram island, number around 17,000 (Ethnologue, 2005). Alune is the largest language on the western side of the island and has approximately 5 distinct dialects.

EMP Languages

The two sub-groups of the EMP are the South Halmahera/West New Guinea (SHWNG) group and the Oceanic group. The Oceanic group was first recognized by Blust (1978) and justified by 56 lexical innovations. While no languages were sampled from the SHWNG group, SHWNG and EMP sub-groups deserve mention because of the fact that both have reconstructible proto-languages; a situation that goes against the Ross

(1995) model for Austronesian dispersal, where "stay-at-homes" (the left hand nodes , figure 4) evolve through dialect differentiation.

To explain, Ross (1995) suggests that the pEMP speaker community lived in an area too small for dialect differentiation to occur. And if the logic that the majority of languages (especially the conservative ones) should still be spoken around the original homeland is followed, then Ross (1995) suggests that either Halmahera or the Cenderawasih Bay was the departure point for speakers of Oceanic languages.

As Oceanic speakers colonized the Pacific, they hopped from island to island, leaving communities behind to evolve in linguistic isolation; this situation which has allowed linguists to accurately define the internal relations of Oceanic despite its tremendous geographic span (Pawley and Ross, 1993).

Oceanic Languages

The original homeland of pOC was most probably the Bismark Archipelago, reached by migrants who followed a path of small islands off of the Northern shores of Irian Jaya and Papua New Guinea (Ross, 1995). There were most likely multiple waves of migration, and pOC speakers had linguistically detectable interaction with coastal Papuan populations along the way. To this day, pockets of Oceanic languages are still spoken on and around Papua New Guinea. Archeological and linguistic evidence suggest 1250 B.C. as approximate date for the massive migrations which spread Oceanic speaking populations throughout Melanesia and into the Western parts of both Micronesia and Polynesia (Ross, 1995).

Ross (1995) states that the groupings of Oceanic are reasonably well-understood, and can be organized into twelve sub-groups (Appendix Table 3). The sub-groups from which contain sample languages are listed below:

Vanuatu: North/Central Vanuatu (Paamese)

The Ethnologue lists 109 Oceanic languages spoken on Vanuatu alone. Clark (1985; from Ross, 1995) explains the North/Central sub-group is the result of dialect

differentiation, with a defined boundary between those languages of the North and Central Vanuatu.

Paamese

Paamese is a language of Vanuatu with two dialects Northern and Southern manifesting a total speaker population of ~ 6,000 (Ethnologue, 2005). Most Paamese speakers are also fluent in Bislama, an English-based pidgin used as *lingua franca* across Vanuatu (Ethnologue, 2005).

Vanuatu: South Vanuatu (Lenakel)

A Proto Language has been successfully reconstructed for this sub-group (Lynch, 1978).

Lenakel

Lenakel is a language of Vanuatu, spoken by ~ 6,500 people in the central and western areas of the island of Tanna (Ethnologue, 2005). Lenakel has a number of dialects; It has also become the go-to language for missionary work involving the other three languages of Tanna (Lynch, 1978).

Central Pacific: Central Pacific (Fijian, Maori, Hawaii)

Fijian

Fijian is spoken as a first language by 331,000 people on the islands of Fiji alone and an additional 5,000 people in communities elsewhere, like Vanuatu and New Zealand. Fijian has 320,000 second language speakers (Ethnologue, 2005). While Fijian is a co-national language with Hindustani and English, there is political pressure to

acknowledge only Fijian as the national language. Fijian has a large number of dialects, with at least 9 recognized dialect-groups.

Hawaiian

Hawaiian is language of the Hawaiian Islands. Hawaiian is currently spoken by ~ 1,000 people as a first language while around 8,000 have some command (Haugen, 1993; from Ethnologue, 2005). At the turn of the 19th century, some 37,000 people spoke Hawaiian. Social and political moves have been taken to revitalize the language through language immersion schooling and possibilities for higher degrees in Hawaiian (Ethnologue, 2005).

III. Methods of Data Collection and Selection

i. Choosing the Languages, Features and Words

The initial language sample was assembled to support phylogenetic analysis both theoretically and practically. Theoretically, I wanted to represent the linguistic diversity of the family. Practically, I needed languages that were well represented in both the World Atlas of Linguistic Structures (WALS)(Haspelmath *et al.*, 2005) and Austronesian Basic Vocabulary (ABV)(Gray and Greenhill, 2005) databases and would sum to a manageable sample size.

Of the 1268 Austronesian languages listed in the Ethnologue, WALS has listings for 311. For each language listed, the number of attested WALS features was calculated. The languages and number of attestations were then sorted into genera, in order to identify the most well-attested exemplars.

WALS identifies 17 Austronesian genera with considerable disparity in size (Appendix Figure1): 2 genera have only single members, 4 have five or fewer members and one has 140 members. In order to make the results more easily comparable with the established groupings from historical linguistics, at least two languages from each genus were selected. Selected languages had 1. high attestations of WALS features and 2. a

listing on the ABV database. Unfortunately, some genera could not be represented by paired exemplars. Naturally this was the case for outliers, but in other situations candidate languages were excluded because they lacked an ABV listings and/or had low WALS feature attestation. In total, the initial sample set consisted of 28 languages representing 14 of the 17 Austronesian genera. Of the 14 included genera, 6 were represented by two or more languages.

While ideally the 28 language sample should be compared with respect to all 142 features in WALS, the patchy attestation of certain features and languages forced me to severely restrict that number to ensure a robust and thorough typological sample for comparison. In selecting WALS features to include, two goals were set: 1. to represent the typological diversity of the WALS database (Appendix figure 2) and 2. to select those features which are resistant to borrowing. In an effort to meet these goals, three sets of information were compared: 1. The number of attestations per feature in the 28 language set, 2. the percent representation of each typological "theme" of WALS and 3. features unlikely to be borrowed, assessed through a ranking of a "p-value[iii]" (Wichmann and Kalmholtz, 2005)(figure 6).

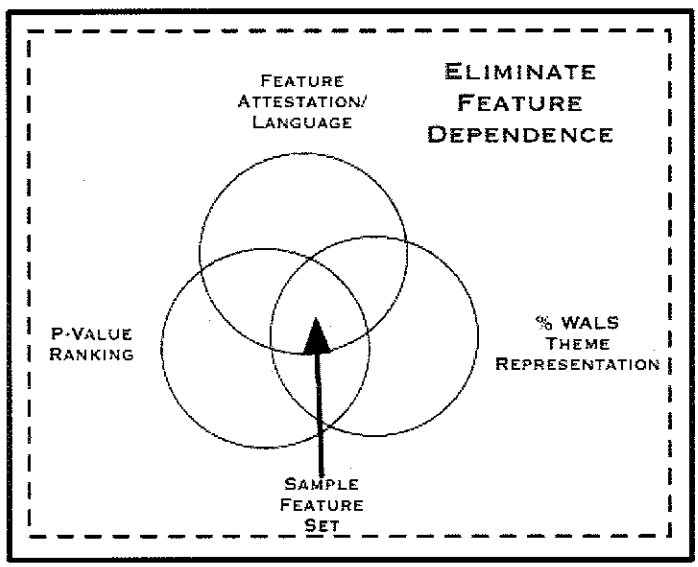


Figure 6. Schematic of WALS feature choosing for phylogenetic analysis

From this set, redundant features were eliminated to avoid "linked" data (i.e. data which is known *a priori* to be dependent). After the filtering, 35 features were selected to comprise the initial sample (Appendix Table 4)

The initial language/feature matrix contained 980 (35 x 28 = 980) possible data points of which 418 were recorded in WALS. I added 258 data points bringing the matrix to 84 % completion. Unfortunately, due to lack of in-depth and/or available grammars many data points were still unknown after substantial searching; these were coded with a "?" (Data Available on Supplementary CD).

After data entry, the matrix was pruned a final time in preparation for phylogenetic analysis. One language (Timugon) and five features were excluded because of sparse attestation (Feature 46 and 123), being uninformative (Features 11 and 30), or because of dependence (Feature 81). The final matrix was 88 % complete, with 27 languages and 31 features in total. For each of the final 27 languages, a modified 200 Swadesh word list was taken from the ABV database.

ii. Coding the Data

Lexical items were coded by cognate class following the format of Gray and Greenhill (2005) where the cognate class serves as the unit of selection. To accomplish this encoding, two matrices were assembled for the 27 languages and 200 lexical items. The first matrix consisted of the words transcribed in the International Phonetic Alphabet (IPA). The second matrix consisted of the preliminary cognate judgments of the ABV website. To differentiate between those words without a cognate class and those which had not yet been annotated by the ABV curators, both matrices were compared. The words with cognate judgments present were color-coded by cognate class. After initial judgments had been marked, the words themselves were compared. Using my best judgment I added words to each cognate class (most of these were obvious members). In no case did I reject an ABV judgment. Afterwards, in a third matrix, the cognate classes were separated into independent columns; the presence of a cognate was scored as "1" and the absence of a cognate as "0" (figure 7B). Note that this encoding strategy neatly

deals with lexical polymorphism. The 27 language sample generated 451 cognate classes from 200 lexical items. The number of cognate classes belonging to a language – and thus the amount of information available for phylogenetic inference - differed markedly, between 33 (Yapese) and 158 (Tagalog) (Appendix figure 3).

Structural data were encoded in two ways for comparison. The first encoding established the WALS feature as the unit of selection (multi-state), while the second encoding established the WALS feature *unit* as the unit of selection (binary). The strategy for adapting multi-state encoding into binary encoding parallels the separation of a lexical item into constituent cognate classes. The members of features with "mixed" or "other" groupings (the "trashcan" problem; see below) were analyzed independently to verify the similarity of feature state. In cases with more than one type of "mixed" or "other category" a new feature state was created. In the binary encoding scheme, languages with "mixed" state were considered polymorphic and scored with multiple "1's".

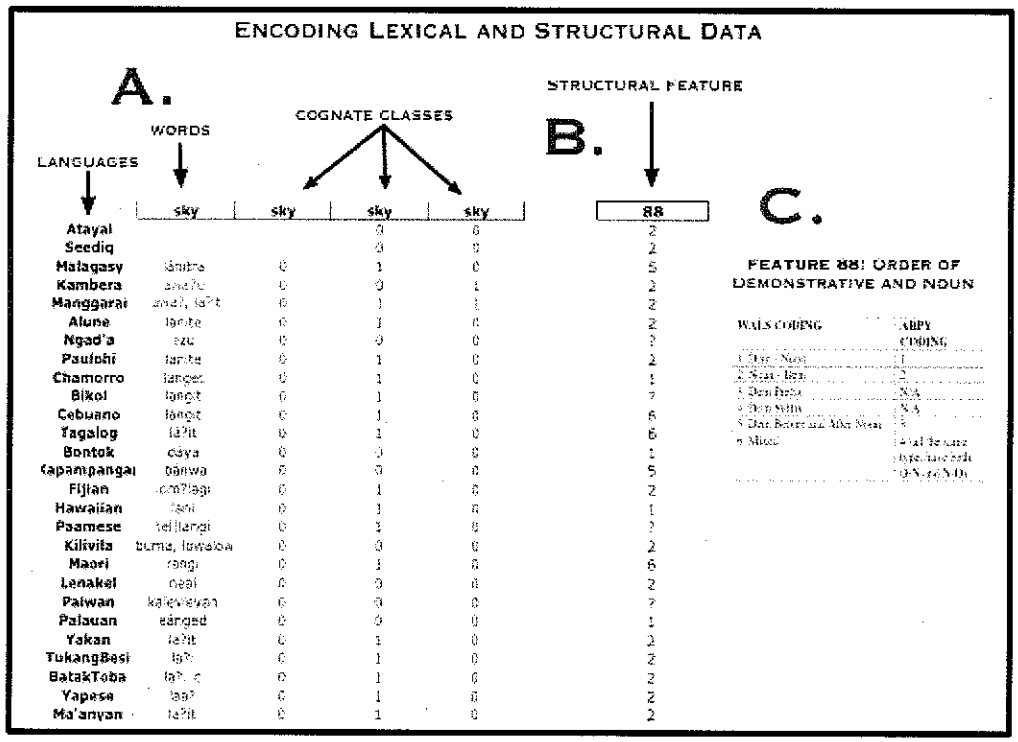


Figure 7. The coding scheme used for A. lexical items (by cognate class) and B. features (multi-state). C. Feature states were then recoded to take into account absent values and to parse "mixed" or "other" categories.

iii. Issues of WALS-based Phylogenetics

The WALS database was created primarily as a visual tool for assessing the geographic distribution of the structural diversity of world's languages. As such, there are a number of issues surrounding the use of WALS features as characters for phylogenetic inference.

Firstly, there is significant inter-feature heterogeneity in how the feature states represent a given feature. In terms of tree building, this amounts to comparing "uneven" units. For example, Feature 26: Prefixing vs. Suffixing in Inflectional Morphology is composed of feature states on a continuum between 'strongly prefixing' and 'strongly suffixing' with an additional category of 'little affixation'. The feature state value is determined by a score based on the behavior of several inflectional parameters of the language (*e.g.* case affixes on nouns, pronominal affixes on verbs, etc). The feature states of 26 thus indicate a general property of the language. Oppositely, the feature states for Feature 7: Glottalized Consonants, represent specific language properties (*e.g* presence of ejectives, glottalized resonants, etc.). This issue of irregular units of comparison can be called the 'apples and oranges problem.'

Secondly, many WALS features are nested or obviously linked. Since independence of characters is a fundamental for non-skewed phylogenetic analysis, this issue arises as a major drawback of WALS-based phylogeny. Linked features must be excluded by hand. As an example, take Feature 87: Order of Adjective and Noun and Feature 97: Relationship between the Order of the Object and Verb and the Order of Adjective and Noun. In this case, the state value for Feature 87 has a direct impact on the value of Feature 97, thus an analysis which included both Features as characters of comparison would be increasing the weight of the language parameter Adjective/Noun order. This can be called the 'independence problem.'

Thirdly, WALS uses "other" feature states to group languages with properties outside of the main feature classifications. This is referred to as the "trashcan problem." If these problematic features are to be used in phylogenetic analysis, members of the "other" category must be addressed individually and re-grouped in terms of the quality of

the feature. In some cases, features are linked through "other" categories, thus compounding the 'trashcan' and 'independence' problems. For example, take features 14 and 15, Fixed Stress Locations and Weigh-Sensitive Stress, respectively: all the languages to which feature 14 applies are lumped in the "other" category for 15 (and vice versa).

In sum, the strategies used to make WALS more readable and straightforward detract from the ease and quality of using the database for phylogenetics. In this report, efforts were made to address some of the problems outlined above: The 'apples and orange' problem was unattended in the multi-state encoding and partially dealt with in the binary encoding. Feature independence was also inspected, although the nature of the WALS categories suggests that fulfilling a requirement for strict independence would necessitate building a new database from the bottom-up. Many cases of feature linkage were thrown out; some, however, remain (including features 14 and 15). Features with "mixed" or "other" categories were examined individually and WALS features were slightly recoded to take the similarities and differences of languages in these categories into account. Recoding also eliminated those feature state values which were not expressed by any language in the sample (figure 7C).

IV. Evaluating Phylogenetic Methods for Linguistic Data

i. Introducing the Methods

In order to assess the most appropriate phylogenetic methods for modeling language evolution, representatives from 4 of the 5 major approaches were used to evaluate the lexical and structural data presented in III. The experimental trees based on lexical, structural, and combined data sets were compared to each other and the 'known' tree, providing a measure of the accuracy and robustness for a given method (figure 8). Before presenting these results, each method is explained in terms of basic assumption and role within the phylogenetics discipline. By complimenting results with a transparent description of mechanism, the suitability of each method can be evaluated for modeling language evolution.

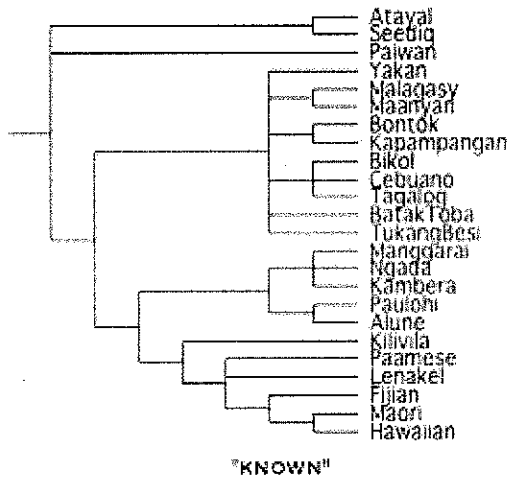


Figure 8. A cladogram of the 'known' tree conducive to quantitative comparison

The most significant methodological split in phylogenetics is between *phenetic* and *cladisitic* approaches (Makarenkov, 2005). The *phenetic* approaches operate on a pair-wise distance matrix calculated by comparing every character for each taxon. Because individual character states only function to generate this initial matrix, these methods make no reference to ancestral relationships. The most widely employed

phenetic approaches are called the *distance methods*. One such method, *Neighbor-Joining* (NJ), is evaluated here.

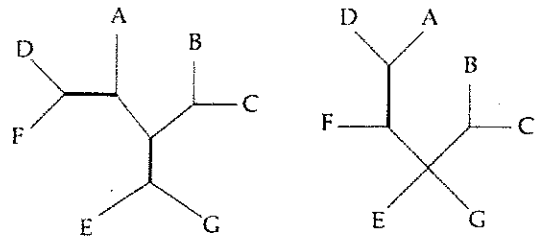
The *cladisitic* approaches do refer to ancestral relationships through the use of specific evolutionary models which function with respect to individual changes in character states. The *cladistics* methods have a wide number of incarnations. Here *Maximum Parsimony* (MP) and *Bayesian Analysis* are evaluated.

Also evaluated is *NeighborNet*, an example of a *network* approach. *Network* approaches may employ both *cladistic* and *phenetic* methods, but differ from those previously described in that they do not force the data into a tree. Instead, multiple relationships between taxa are simultaneously represented. While the format of the "trees" generated from this method disallows quantitative comparison to the 'known', the underlying methods of assembly are nonetheless described.

NJ and MP trees were generated with PAUP* (Swofford, 2003). Bayesian analysis was accomplished with MrBayes (Huelsenbeck and Ronquist, 2001;Ronquist and Huelsenbeck, 2003). NeighborNet analysis was done through SplitsTree4 (Huson and Bryant, 2005.)

ii. Comparing the Known and Experimental Trees

Trees were compared with respect to topology. The topology was assessed with the *symmetric differences test* based on Robinson and Fould (1981) and employed through the PHYLIP program TreeDist (Felsenstein, 2005). In this framework, a series of partitions is generated for each tree branch. These partitions are defined simply by the taxa on either side of a given branch (figure 9). The distance score for any two trees is simply the number of unshared partitions, meaning the larger the distance score, the greater the difference between trees.



Partitions	Partitions
{ADF | BCEG}	{ADF | BCEG}
{DF | ABCEG}	**{AD | BCEFG}**
{BC | ADEFG}	{BC | ADEFG}
{EG | ABCDF}	

Figure 9. An example of the *symmetric length difference test* at work. Partitions for each unrooted tree are shown in tables and unshared branches are in bold. These two trees have a distance score of 3 (from Felsenstein, 2004).

For each method, experimental trees were generated for the lexical, structural, and combined data sets. The structural data were encoded in two different fashions, multi-state and binary-state, thus giving five experimental data sets in total.

Since the performance of each method was judged solely by tree topology, other information including branch lengths and measures of nodal support were ignored. Additionally, the WMP outlier languages (Chamorro, Palauan, and Yapese), lacking a consensus position on the 'known' tree, were excluded from this initial analysis.

iv. The Neighbor-Joining Distance Method

Neighbor-Joining (Saitou and Nei, 1987) is one of the most popular distance methods. NJ is a type of clustering algorithm that generates a tree to match the initial pair-wise distances between taxa. While ignoring specific changes in character states may seem like a devastating loss of information, computational studies suggest

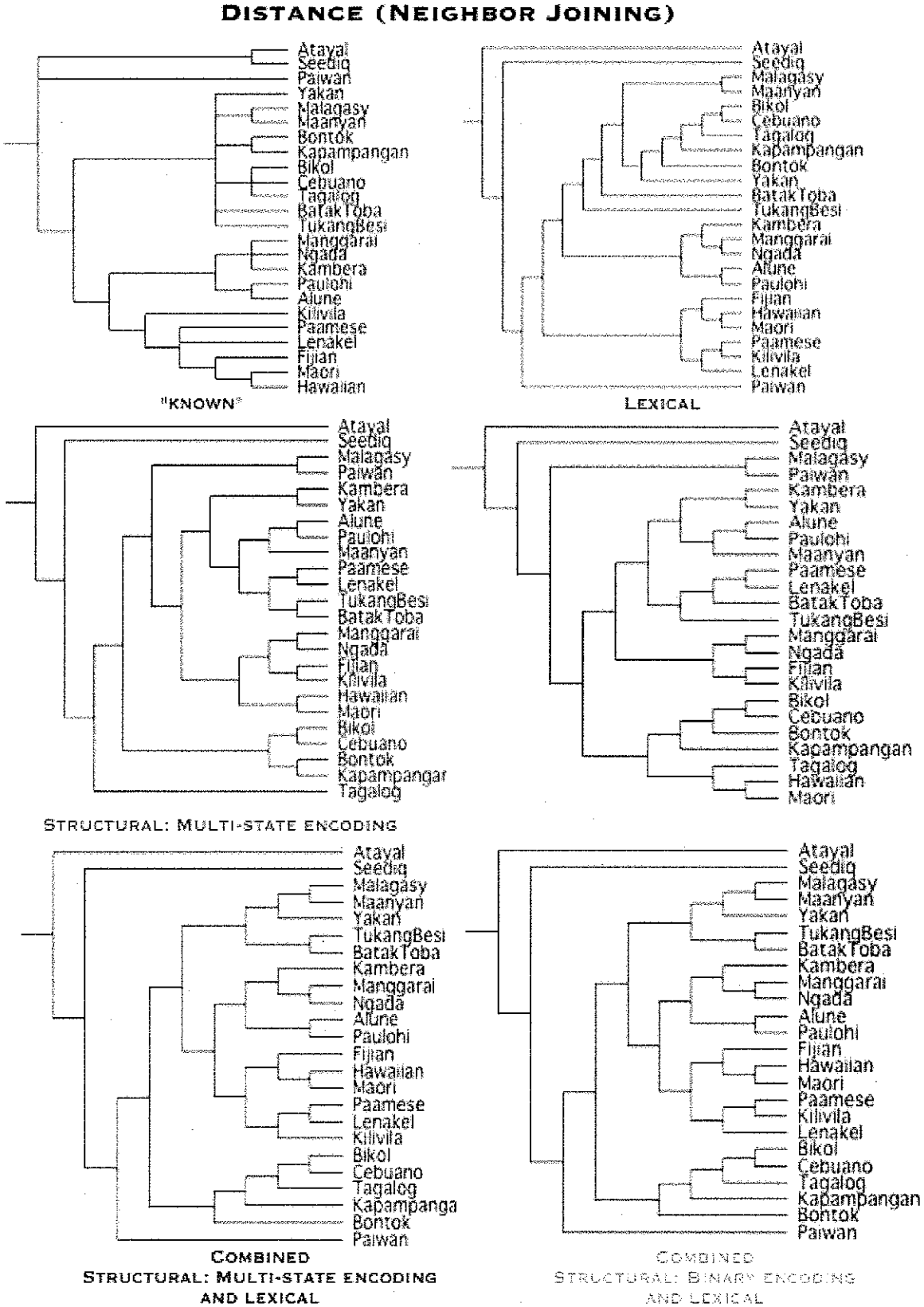


Figure 10. The results of NJ analysis on the five experimental data sets

surprisingly little is lost (Felsenstein, 2004). Felsenstein (2004) articulates this phenomenon by suggesting higher-order character state information is consistently retained in the distance matrix itself. Furthermore, Atteson (1999) demonstrated that if the distance matrix is sufficiently close to the evolutionary distance, NJ generates the true phylogeny (Makarenkov, 2005). Unlike some other distance methods (i.e. UPGMA), the NJ algorithm allows for variation in evolutionary rates over different branches of the tree.

NJ operation can be summarized through a series of basic steps from which a "bush" is transformed into a binary tree. At each iteration of the algorithm, two distances from the pair-wise matrix are joined to form the shortest possible tree. Once joined, the two distances are collapsed under a composite node and the matrix is reformed. The process continues until the step-wise shortest tree has been resolved.

The NJ method offers a number of advantages for the analysis of linguistic data. Perhaps most basically, NJ makes no assumptions of evolutionary model and is thus applicable to evolutionary processes in general. The calculation of a tree from raw distances allows for a straightforward interpretation of the tree produced and avoids possible inaccuracies as a result of over-parameterization. It remains to be seen however, if language evolution, like biological evolution, consistently leaves a trace equivalent to the comparison of higher-order character states in the distance matrix. Finally, NJ is computationally efficient and can therefore offer a quick analysis of the data.

v. The Maximum Parsimony Method

Maximum Parsimony (MP) methods generate the tree (or trees) with the smallest number of evolutionary changes. This larger aim of minimizing change is accomplished by the individual comparison of characters. The various incarnations of MP methods involve different constraints on how the states of these characters are allowed to change. For example, Fitch parsimony uses a minimum of constraints: all character state changes are equally probable and reversible. In Wagner parsimony however, character states are assumed to change in a defined order (Makarenkov, 2005).

Since MP methods calculate the number of total changes by summing the individual changes of each character, affecting the role of each character may be a

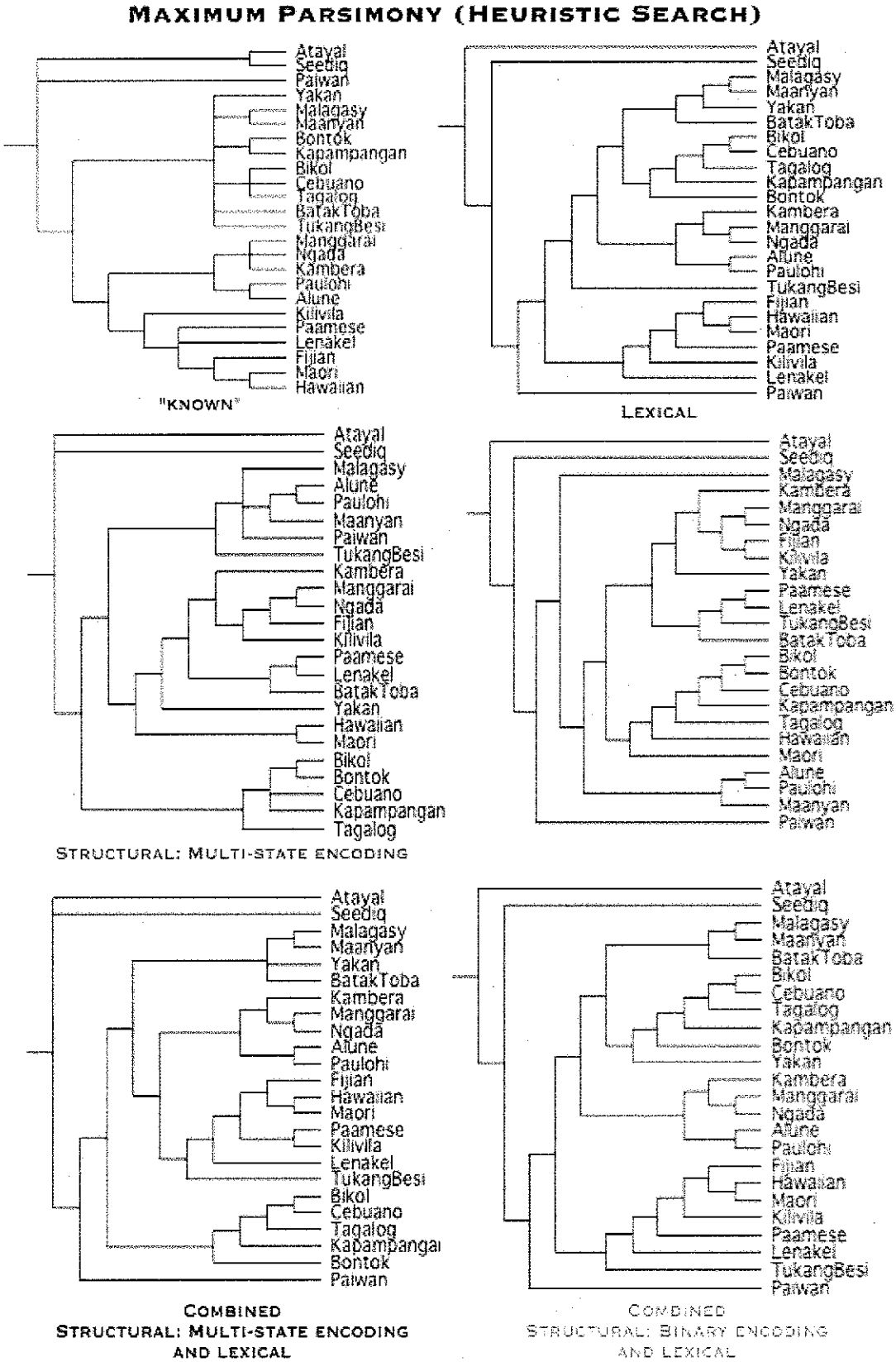


Figure 11. The results of MP analysis on the five experimental data sets

productive way to construct straightforward and accurate models of language change. For example, characters can be differentially weighted to have a greater or lesser effect on the outcome of a tree. Similarly, intra-character state changes can be selectively penalized to make some transitions more evolutionarily "likely". By incorporating informed models of how language characters evolve, MP may be effectively tweaked specifically for language evolution. Of course, varying these parameters does not change the larger assumption that the tree with the fewest changes is the best.

Another aspect of the MP method is the algorithm used to evaluate the possible trees. Since the number of possible trees increases exponentially with number of taxa involved, searching a full set of possible trees is rarely feasible. Instead, *heuristic searches* are employed. Generally *heuristic searches* operate by slightly modifying an initial tree and evaluating whether the new tree reduces the overall number of changes. The process repeats until slight changes fail to produce a more optimal tree. While these algorithms do make MP methods available to large data sets, they are prone to getting "trapped" in local optima (Felsenstein, 2004). In this report, the tree bisection-reconnection (TBR) algorithm, one of three stepwise addition heuristics offered by the PAUP* (Swofford, 2003) software was used.

vi. The Bayesian Analysis Method

While relatively new in their application to phylogenetics, Bayesian methods date back to1790. Bayesian inference of phylogeny is executed through an algorithm which searches through a space of possible trees, preferably steering toward those trees which maximize a value called the *posterior probability*. The *posterior probability* is numerical evaluation of the probability that a given tree is the correct one for the data, and is formulated from the Bayes' Theorem (Huelsenbeck, 2001).

As implemented through MrBayes, a number of these algorithms work simultaneously and in a coordinated fashion to sample trees from their path at a constant rate. Some of these algorithms, called "chains" in the terminology of MrBayes, are free to make large jumps in tree space to find neighbors for comparison. These are called 'hot' chains. One chain, however, is always 'cold' and is constrained to make only local

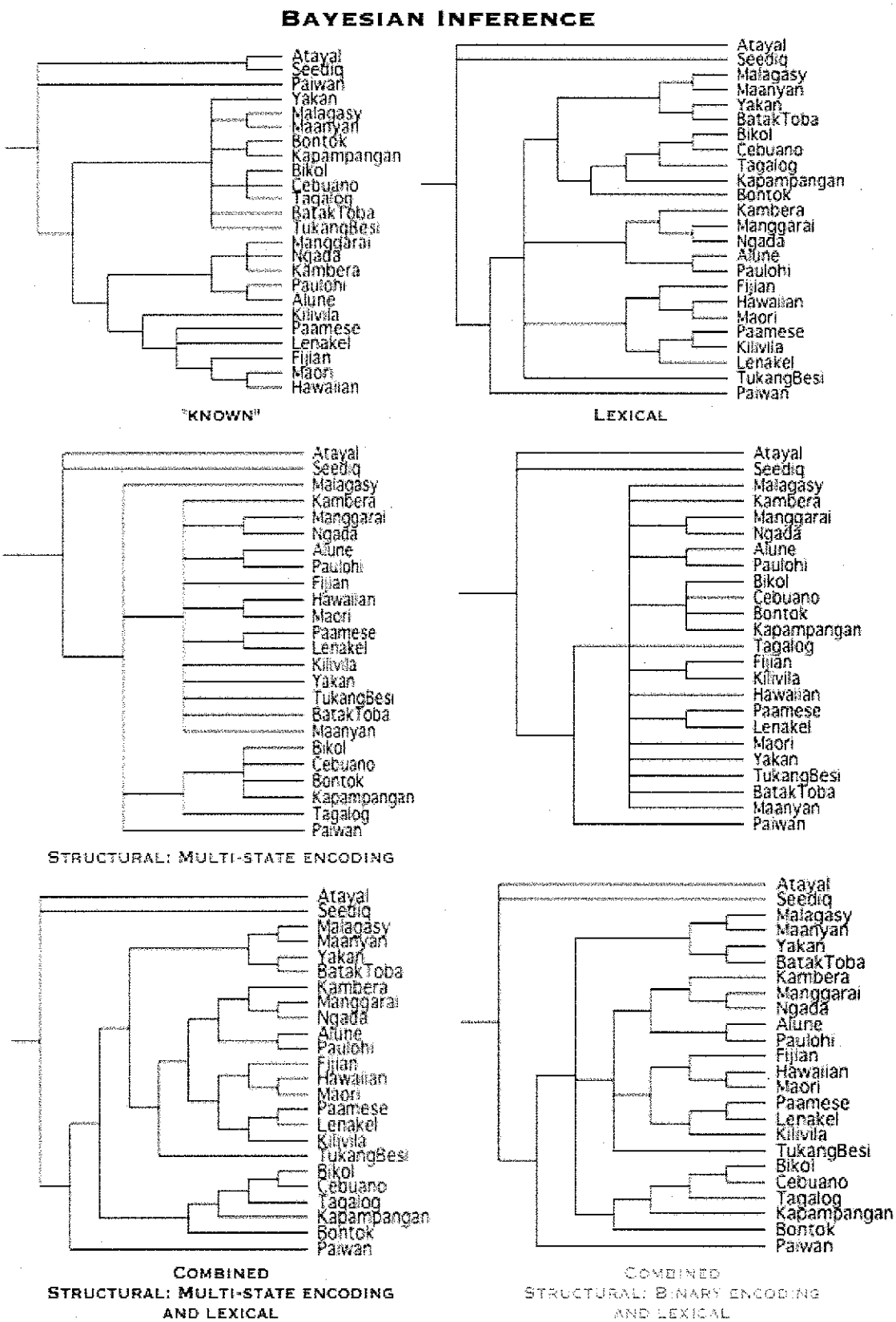


Figure 12. The results of Bayesian analysis on the five experimental data sets

comparisons. After every generation of the program, the chain with the highest posterior probability becomes the 'cold' chain. In concert, these chains avoid locally optimal areas of *posterior probability* and zero in on those areas of tree space with the best trees. A point of convergence is reached when the posterior probabilities of subsequent trees fails to improve; after this point, the algorithms are collecting tree samples with near equal probabilities. The trees sampled before the point of convergence are thrown out in what is called a 'burnin'. From the remaining sample of optimal trees, a consensus tree is assembled following a basic rule: only include those nodes which exist in more than a certain threshold percentage (usually 50%) of optimal trees. Each included node is given a number representing its strength, which is calculated by the probability of finding that node in the set of optimal trees.

Bayesian inference offers a number of advantages for the analysis of linguistic data. Most notably, by driving tree selection by a measure of probability rather than the minimal change framework of MP, no assumption needs to be made about the conservatism of language evolution. However, assumptions must still be drawn about the structure of the evolutionary model and the prior probability distributions of the parameters of the model. These specific aspects of Bayesian phylogenetics will be covered in greater depth in section V. MrBayes also allows for rate variation for the evolution of individual characters; a model which may more realistically describe how specific features of language change over time.

vii. The Network Analysis Method

Phylogenetic Network Analysis (PNA) refers to a group of methods that preserve and describe alternative phylogenetic relationships within a data set. As opposed to traditional methods which force a single tree, PNA provides a qualitative analysis of just how "phylogenetic" the data are: In most implementations, perfectly phylogenetic data appear as a tree, while data lacking a clear phylogenetic signal appear like a web. In biological phylogenetics the inclusion of these relationships can help identify cases of

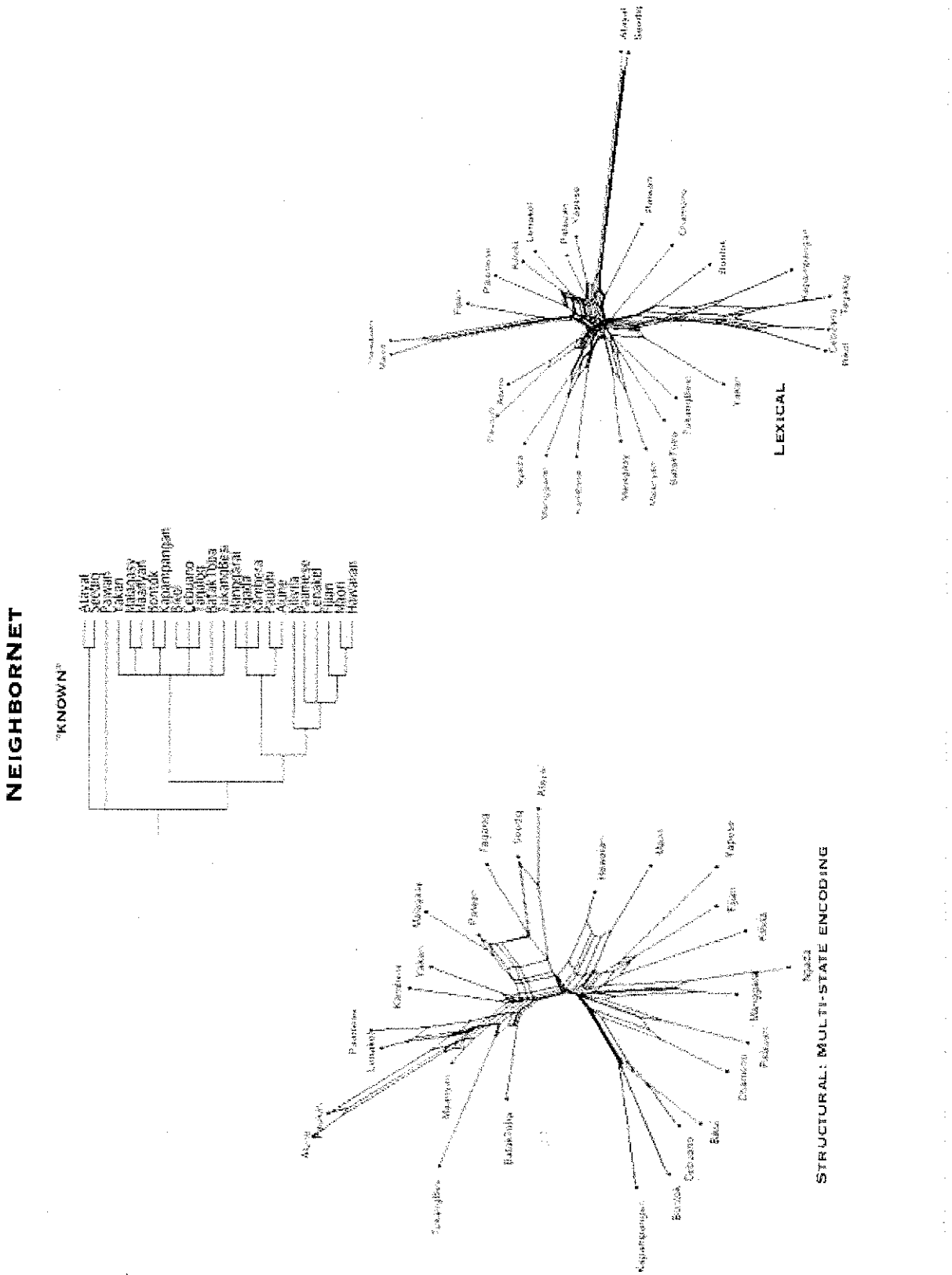
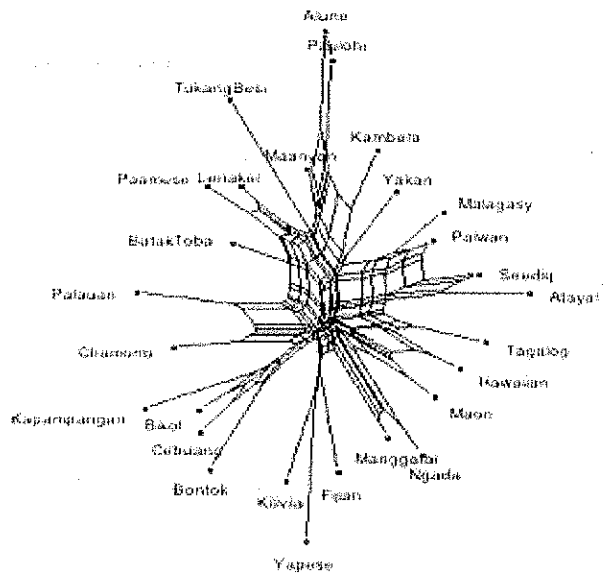
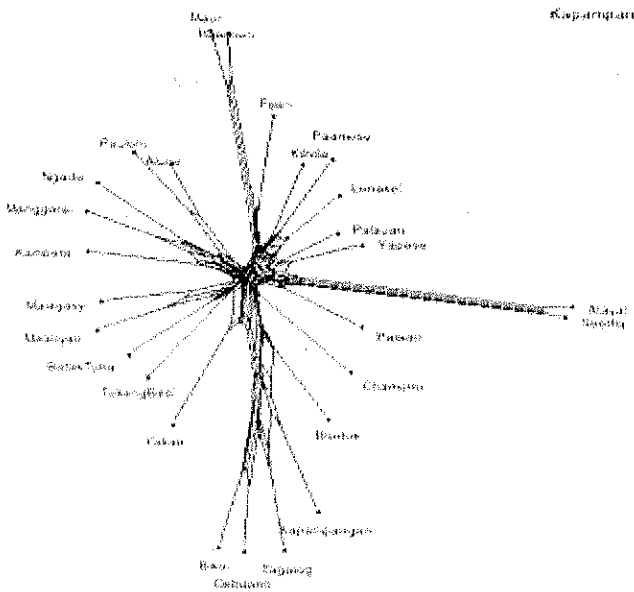


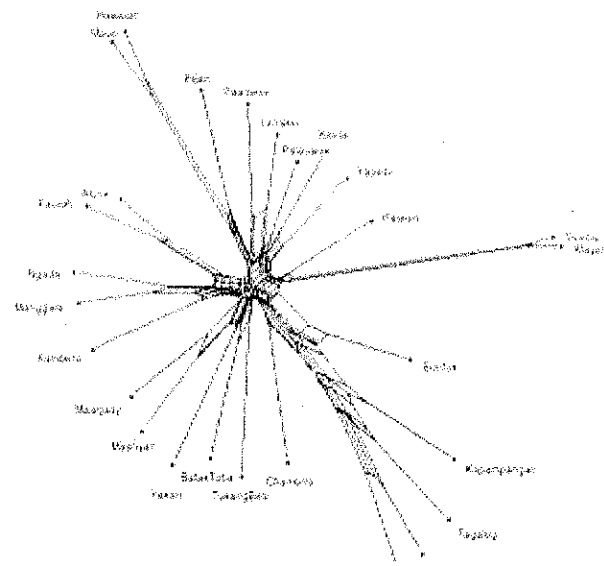
Figure 13. The results of NeighborNet analysis on the five experimental data sets (continued on next page)

NeighborNet (continued)





Combined
Structural: Multi-state encoding
and lexical



Combined
Structural: Binary encoding
and lexical

horizontal gene transfer (HGT), hybridization, and homoplasy; confounds which have non-trivial parallels in language evolution.

As mentioned previously, linguistic borrowing may leave a phylogenetic trace similar to HGT. Events of borrowing are most often preliminarily identified by contradictions between the a gene tree and the consensus tree of the organism or language tree as a whole (Hallet and Lagergren, 2001). Despite the limitations of this approach for language[iv], PNA can nevertheless identify the strength and nature of potential cases of language contact.

Hybridization and creolization may theoretically confound phylogenetic analysis similarly, since both processes create new lineages through the combined interaction of existing organisms or languages. In biology, homoplasy describes a process of convergent evolution where similar phenotypes arise from unrelated species. Methods of detecting homoplasy may conceivably help linguists identify those areas of language not evolving independently, either through collective adaptation to a specific "communication niche" and/or through contingency in the human mind.

PNA share a diversity of algorithmic machinery with traditional phylogenetic approaches (Makarenkov *et al.*, 2005). The method used here is NeighborNet, a variation of the NJ method, implemented through the SplitsTree package.

viii. Results

The phylogenies generated above were compared to the 'known' tree to evaluate which combination of method, data type and coding scheme produced the most historically accurate phylogeny (Table 1, figure 14). Underlying the question of the best method is a subset of questions concerning the role of data type and encoding scheme, the other two variables manipulated. These questions are as follows: 1. Which method was the most precise or internally consistent? 2. What was the effect of multi-state or binary-state encoding for the structural data? 3. What data type, lexical or structural, produced the most accurate results across methods? 4. What was the effect of combining lexical and structural data? 5. Did lexical and structural data represent phylogenetic signals that

were statistically significantly different? Attending to these questions appropriately requires a careful explanation of both the results and the analyses used to interpret them.

Before the specifics, I have two general points. Firstly, the trees are compared both visually and in terms of their *symmetric difference*. The *symmetric difference*, as explained above, is only a rough guide of assessing similarity. Each distance score simply refers to the number of unshared partitions and therefore falls short of giving a statistical measurement of difference. Additionally, all partition differences are scored equally, a bias not shared in visual assessment where the presence or absence of some nodes carries interpretable weight.

Secondly, the 'known' tree is a controversial consensus adapted from a schematic diagram and fit into tree notation. For these reasons, the topology of the 'known' tree was certainly influenced by its role as an informative diagram. In other words, the need for simplicity and aesthetic quality may have detracted from its accuracy. No literature exists on the direct comparison of hand-built and program-built trees, so the 'known' tree should be considered an approximation.

Together these points suggest that a *symmetric difference* comparison to the 'known' tree is a quick and comparative heuristic rather than a stand-alone score of fit. Qualifications withstanding, the *symmetric difference* measurement is far from meaningless: the distance from the 'known' provides helpful preliminary judgments which serve as a skeleton for addressing larger trends in data about the effect of method, data type and encoding scheme.

ix. Precision of Method

Internal consistency is essential for robust and trustworthy results. Since the five data sets analyzed represent the same evolutionary process, a perfectly precise method would reconstruct identical trees. To evaluate precision, the *symmetric difference* test was used to calculate pair-wise distances between each of the five trees generated by a method. The resulting distances were averaged for an average distance (AD) score (Table 2a-c).

Table 1. Symmetric Difference Distances between experimental and "known" trees

Data Type	Maximum Parsimony	Distance	Bayesian
Lexical	15	13	13
Str. Multi	27	27	16
Str. Binary	29	29	15
Lexical + Multi	14	15	13
Lexical + Binary	15	13	13

Tables 2a-c. Symmetric Difference distances between experimental trees built from different data sets within a method

A. Maximum Parsimony	Lexical	Str. Multi	Str. Binary	Lexical + Multi	Lexical + Binary
Lexical	0				
Str. Multi	30	0			
Str. Binary	32	20	0		
Lexical + Multi	9	29	31	0	
Lexical + Binary	8	30	32	9	0

MP Average Distance = 23.0

B. Distance	Lexical	Str. Multi	Str. Binary	Lexical + Multi	Lexical + Binary
Lexical	0				
Str. Multi	32	0			
Str. Binary	32	14	0		
Lexical + Multi	30	30	32	0	
Lexical + Binary	12	28	30	2	0

NJ Average Distance= 24.3

C. Bayesian	Lexical	Str. Multi	Str. Binary	Lexical + Multi	Lexical + Binary
Lexical	0				
Str. Multi	17	0			
Str. Binary	18	5	0		
Lexical + Multi	6	17	18	0	
Lexical + Binary	4	15	16	2	0

Bayesian Average Distance = 11.8

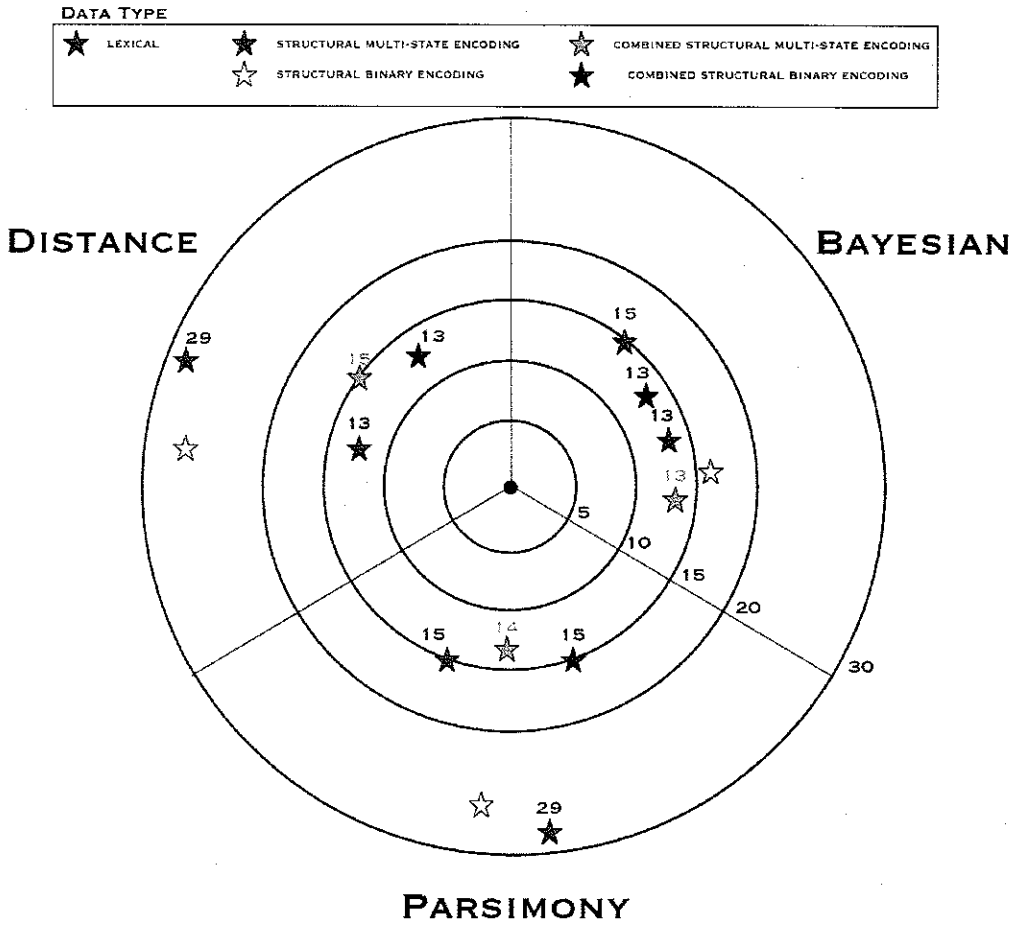


Figure 14. A mapping of symmetric difference distances from experimental trees to the "known," organized by method and data type.

Bayesian analysis, with an AD of 11.8 was substantially lower then both the Distance method (AD = 24.3) and the MP method (AD = 23) (Tables 2a-c). The low AD score of the Bayesian method indicates a comparatively strong internal consistency, meaning Bayesian methods are less prone to fluctuation in tree topology due to data type and coding scheme. This result is unsurprising since the tree produced by MrBayes is inherently a consensus tree: from the 25,000 most optimal trees, only nodes found in 50 % or more of those trees were expressed in the final tree. Distance methods on the other hand generate a single tree with no consensus. With MP, consensus trees are only used when more than one most-parsimonious tree are found. Here, the only data sets to return more than one tree were the multi-state structural data, returning 16 trees, and the

combined structural multi-state and lexical set, which returned 2 trees. In these cases, the same 50% majority consensus method was used.

x. Binary vs. Multi-State Encoding

Binary and multi-state structural encodings produced trees that were equally distant from the "known" under both Distance and MP methods. With SD scores of 27 and 29 for binary and multi-state encodings respectively, these trees were clearly the least accurate (figure 14). Additionally, these trees were inaccurate in different ways, judging by the large SD scores between encodings for a single method (SD = 20 for MP and SD = 14 for Distance).

Under these two methods, visual inspection suggests that multi-state encoding performed slightly better despite being 2 SD points higher. Most strikingly, one of the strongest and most frequently observed nodes uniting the Meso/Northern Philippines group remains intact with multi-state encoding but is broken up with binary encoding. Further support for the multi-state encoding comes from NeighborNet analysis. Here, binary encoding appears to cloud the phylogenetic signal as evidenced by the increased "webbing" as compared to the multi-state network (figure 13).

Under Bayesian analysis, both structural encodings produced relatively unresolved trees. Nevertheless, in both cases the Formosan/non-Formosan split was included; in terms of migration, this is arguably the most salient grouping and was only included in one of the four trees produced by the Distance and MP methods (the MP binary encoding). Excluding the Northern/Meso Philippines group, both binary and multi-state encodings managed to pair 8 languages together appropriately. While far from fully resolved, Bayesian analysis of both encodings avoids making improbable assumptions and thus generates trees which are relatively similar (SD = 5).

xi. Lexical vs. Structural Data and Combined Analysis

Due to the different number of characters included in the lexical and structural data sets, no final judgment can be made about which type of data is inherently better for

phylogenetic analysis. Undoubtedly lexical data generated trees closer to the known across methods (figures 10-13). It is unclear, however, how much of this improved accuracy was due to the substantially greater number of characters available for comparison: lexical data had 451 binary characters compared to the two encodings of the structural data, with 31 multi-state characters or 97 binary characters respectively. While it might be feasible to generate lexical trees based on random groups of 97 cognate sets and compare these to the binary structural trees , this analysis was not undertaken. Instead, rather than excluding characters, the merits of the data type were evaluated with respect to their ability to compliment each other. With these combined data sets, the lexical characters are primarily driving the topology and branch lengths of the tree. Nevertheless, the results suggest that the structural data has an important role in "tweaking" the outcome. Case studies from biological phylogenetics suggest the addition of even small amounts of morphological data to molecular data sets may significantly alter the structure of a tree (Baker and Gatesy, 2002). In other words, this "tweaking" is far from trivial and should be examined across methods.

For the combined Distance analysis, both combined encodings of the structural data produced nearly identical trees (SD = 2), allowing us to consider both trees as virtually identical and evaluate a singular set of "tweakings" from the lexical tree.

The most evident change is the break-up of the WMP group, as the Northern and Meso Phillippines languages are extracted from the Sama Bajaw, Sundic and Sulawesi languages and placed "higher-up" in the tree. This new placement creates a node between the Northern and Meso Philippines languages and all other non-Formosan languages. A second change occurs with this other non-Formosan group, where the Oceanic languages become a "sister" branch to the CMP languages, thus creating an accurate C-E MP node, save the remaining Philippines languages (Yakan, Batak Toba and Tukang Besi). In terms of migration, this "tweaked" topology makes more sense: the lexical tree has the major branches in the wrong order outside of the Formosan group, an issue which is largely resolved in the combined tree when the Northern and Meso Phillippines languages, instead of the Oceanic, are the first to branch off. The remaining conflict with the known occurs because of the placement of the other Philippines languages (Yakan, Batak Toba, and Tukang Besi) inside the C-E MP lineage.

A similar improvement in resolution occurs in the combined Bayesian analysis. Once again, the binary and multi-state encodings of the structural data were largely the same (SD = 2) but the two combined encodings resolve the four-way node connecting Tukang Besi, the Oceanic languages, the CMP languages, and the Phillippines languages differently. While the binary encoding only slightly alters this shared node by "lowering" Tukang Besi, the multi-state encoding fully resolves this and all shared nodes so that never more than two branches branch from a single node. In terms of Ross's dynamics of migration, this is an expected property. In fact, the multi-state combined encoding mirrors the order of branching in the 'known,' where the Philippines languages branch off first, leaving an intact C-E MP node which itself shows a split between the CMP and Oceanic languages.

Under MP, the combined multi-state and binary structural encodings "tweak" the lexical tree differently. The combined binary encoding separates two Philippines languages from their "usual" positions across methods: Yakan assumes the position of being the first branch off of the Northern and Meso Philippines group, while Tukang Besi assumes a similar position with the Oceanic groups. Both of these changes are inaccurate, and the latter change also occurs with the combined multi-state encoding. The combined multi-state encoding does produce a substantial improvement with respect to the placement of the Oceanic languages: they are moved from being the earliest branch off of the non-Formosan group to a position consistent with the 'known' C-E MP node. This improvement over the lexical tree does not appear in the combined binary encoding.

xii. Statistical Difference Between Lexical and Structural Data

To address if two data sets produce significantly different trees, the *incongruence length difference* (ILD) test of Farris (1994) is often employed (Allard and Carpenter, 1996). The ILD operates in the Maximum Parsimony framework: pairs of trees are built from a mixed sample of characters drawn randomly from both data types and are evaluated by summing the number of total character changes in each tree. If the sum of the two trees based on the original data types falls into the "shortest" 5% of pairs of trees sampled, the data represent a significantly different phylogenetic signal.

The ILD test was run for lexical and multi-state structural data types in PAUP*. 200 pairs of reordered trees (replicates) were calculated. The original data set produced a combined tree length of 1545. The shortest combined length was 1541. Eight of the 200 replicates had combined lengths of 1545 or shorter, giving a P-Value of 0.04; thus rejecting the null hypothesis that the two data types represent the same phylogenetic signal was rejected. Put positively, the lexical and multi-state structural data sets were found to express significantly different phylogenetic signals in terms of MP.

xiii. Discussion

Leaving Neighbor-Net analysis temporarily aside, the results presented here suggest that Bayesian analysis is the most appropriate phylogenetic method for linguistic data. In terms of data type and encoding scheme, the combined lexical and multi-state structural data produced the most historically accurate tree; a phylogeny which is not only relatively impressive, but which reconstructs the major divisions and minor pairings of the 'known' tree.

While "tweaking" effects have been observed with Bayesian analysis in biological contexts (Nylander *et al.*, 2004), NJ also demonstrated greater accuracy when "tweaked." Increased accuracy due to data inclusion from different components of language suggests a "total evidence" approach is appropriate for modeling the evolution of whole languages. The relative and objective imprecision of MP, along with a worsening of topology with combined data, suggests that a "minimal number of changes" approach is inappropriate for language evolution. Further support for Bayesian over MP analysis was also found in a different structurally based data set by the author and colleagues (unpublished data[v]).

Analyzing the data types separately, lexical data outperformed structural data in resolving phylogenies across all methods. It remains unclear however, how much of this performance should be attributed to data type versus number of character comparisons.

Encoding data as binary or multi-state had differential effects under different methods and when analyzed alone or in combination with lexical data. Evidence from less web-like NeighborNet representation of multi-state structural data suggests this

encoding produces a more "phylogenetic" signal. Importantly, under Bayesian analysis combined multi-state data provided more accurate "tweakings" then binary data.

Overall, the complimentary approach of NeighborNet and Bayesian analyses allows for informed tree building for languages. Network representation displays relatedness without forcing a tree, thus providing a means to assess whether subsequent tree building is justified. If it is, the empirical evidence presented here suggests that Bayesian analysis operates under assumptions appropriate for modeling language evolution. And since only nodes above a certain threshold probability are reported, the degree of resolution of a Bayesian result depends on the strength of the phylogenetic signal within the data, thus avoiding phylogenetic "leaps of faith." The similarity between the 'known' tree, generated from over a100 years of research, and the Bayesian tree, generated from a 36 hour computer run, supports both the accuracy of the current consensus on the history of the Austronesian languages and potential productivity of a computational approach in quickly developing hypotheses of language evolution.

V. A Basic Method for Inferring Phylogenies from Linguistic Data

In this section, the results reported above are synthesized into a basic method for analyzing linguistic data with biological software. This method consists of two parts: 1. Using the Neighbor-Net analysis of the SplitsTree program to check the data for a coherent "phylogenetic signal" and, if found, 2. using MrBayes to infer the single phylogeny which most probably represents the data. While only two steps are involved, there are a number of methodological and software-based decisions that need to be made appropriately and consistently for meaningful comparative results. My goal is to identify those methodological issues and provide suggested solutions when possible. In walking through the application of this method, the whole data set, including outliers, will be used.

i. Step 1: NeighborNet Analysis

What is NeighborNet, how are its graphs interpreted and when should data be considered "phylogenetic" enough to justify tree building? The NeighborNet algorithm was introduced by Bryant and Moulton in 2004 and implemented in the SplitsTree software package (Hulton and Bryant, 2005). NeighborNet is a network variation of the Neighbor Joining Distance method and operates similarly from an initial distance matrix. The output of Neighbor-Net is called a *splits graph*. To interpret how a *splits graph* represents alternative relationships within the data the mechanisms of the Neighbor-Net algorithm are briefly explained.

Recall from the discussion of the *symmetrical difference test* that a tree can be defined in terms of a series of partitions for each tree branch which divide the taxa into two non-empty parts. When building a tree, it is vital that only compatible partitions (or *splits*) are included; with networks however, the condition for the inclusion of *splits* is weaker than compatibility (Bryant and Moulton, 2004). In NeighborNet, this condition is defined by a weight-measure equal to the length of the branch. For both compatible and non-compatible splits, it is then possible to calculate a revised distance matrix for a modified NJ algorithm to operate on. These distances are the sum of all of the splits connecting any two taxa (x and y). In a *splits graph*, this summed distance is equal to the shortest possible path connecting x and y.

To illustrate how these incompatible and non-incompatible distances are used, recall the basic mechanics of NJ: 1. create a taxon for each node. 2. collapse the two closest nodes into one. 3. re-adjust the pair-wise distance matrix to account for the new node and reiterate the process. NeighborNet differs slightly in agglomerative framework. Instead of collapsing a pair of nodes, the algorithm waits for a second pairing of one of the nodes (a third "neighbor"), at which point the three nodes are collapsed into two. Those two nodes are replaced in the distance matrix and the 3-way NJ process continues (figure 15). When only three nodes remain, the process is reversed and the nodes are fully expanded. This is identical to NJ, except for the number of nodes being replaced: in NJ, a single node is expanded in to two; in Neighbor-Net, two nodes are expanded to three (figure 16).

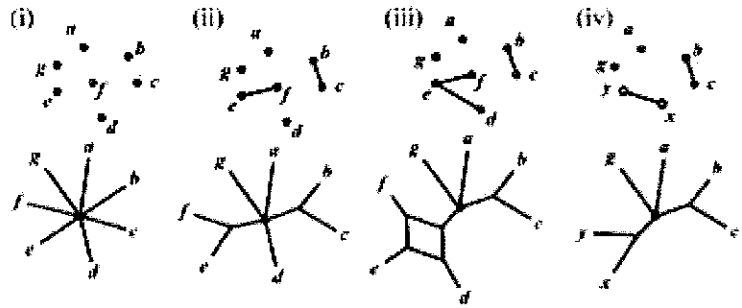


Figure 15. A schematic of the agglomerative process used by Neighbor-Net. Once two nodes have been identified as neighbors (ii), the are not immediately collapsed. After finding a third "neighbor" (iii) the three nodes are collapsed into two (x and y) which are reincorporated into the distance matrix (taken from Bryant and Moulton, 2004).

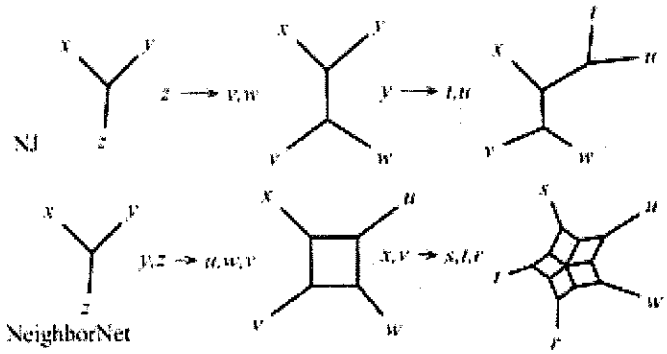


Figure 16. NJ and Neighbor-Net expansion after agglomeration (taken from Bryant and Moulton, 2004).

The final product (a *splits graph*) represents the collection of weighted splits. A clear way to interpret these *splits graphs* is articulated by Bryant and Moulton (2004): In the case where a sub-set of this collection is compatible, it corresponds precisely to a tree, since each edge matches a *split* with a length equal to the weight of the split. Incompatible *splits* are represented by boxes, where a single *split* corresponds to a collection of parallel edges all with the same length. These correspondences can be simplified with the following heuristic: the more boxes, the more incompatible splits, the less tree-like the data.

Bryant and Moulton (2004) impress that Neighbor-Net is an exploratory tool for data representation. As such, there are no quantitative statistical tests for determining whether data are significantly tree-like to justify using traditional tree building methods.

Perhaps such a test, amenable to language, will be available at some point. Until then, linguists must trace signal conflicts back to the original data (Bryant *et al.*, 2004). If non-phylogenetic trends appear between groups of characters and taxa, this may suggest contact events resulting in borrowing. If no trends appear, the incompatible splits are most likely the result of noise inherent in the data or sampling error (Bryant and Moulton, 2004).

Splits graphs are useful for comparing the phylogenetic signals from different data types in detail. Here a *splits graphs* comparison between lexical and structural data reveals an abundance of potentially informative differences (figure 17 and 18). Take for example two conspicuous deviations: 1. The Philippines languages appear highly conserved both lexically and structurally, except for Tagalog, which appears structurally like the Formosan languages (figure 17b and 18b) and 2. Structurally, Paamese and Lenakel, Oceanic languages of Vanuatu, are paired together under a large node containing the Central Malaku languages (Alune and Paulohi) and Maanyan, a Borneo WMP language (figure 17c and 18c). In terms of lexicon, Paamese and Lenakel and Maanyan and Malagasy are grouped together with both the Oceanic languages and the non-Philippines WMP languages respectively; a more "traditionally" accurate grouping. SplitsTree provides tools to visualize and highlight the *splits* underlying these differences, which can then be traced back to the original data to identify the characters responsible for certain *splits* (Bryant *et al.*, 2004).

A similar comparison of *splits graphs* can help pinpoint the "tweaking " effect caused by combining data of different types. Neighbor-Net analysis clearly demonstrates that lexical data (451 binary characters) as opposed to the structural data (27 multi-state characters) is driving the structure of the network. Careful examination reveals the inclusion of the structural data has only a subtle effect: only when the most detailed interior splits are examined can some potential "tweaking" effects be observed (figur19a and b).

To summarize, Neighbor-Net provides a valuable first-perspective on modeling language evolution by allowing the researcher to qualitatively assess how "phylogenetic"

the data are. Additionally, *splits graphs* can be used generate hypotheses about language contact situations since trends based on *incompatible* splits can be pursued at the level of character comparisons.

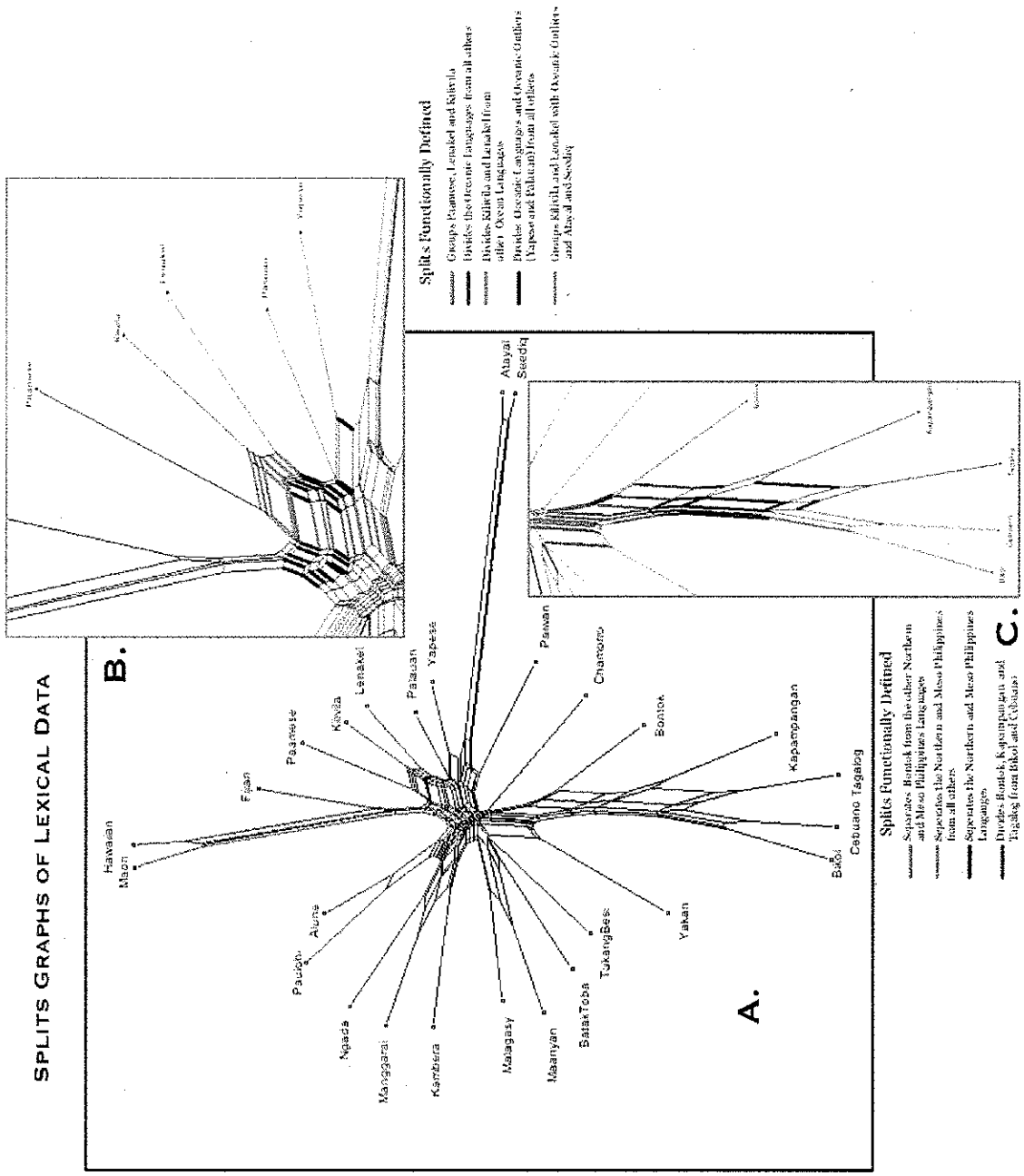


Figure 17A. A NeighborNet *splits graph* based on lexical data. Close-up views of the *splits* contributing to a Philippines group (B) and to the relationships of Paamese and Lenakel (C).

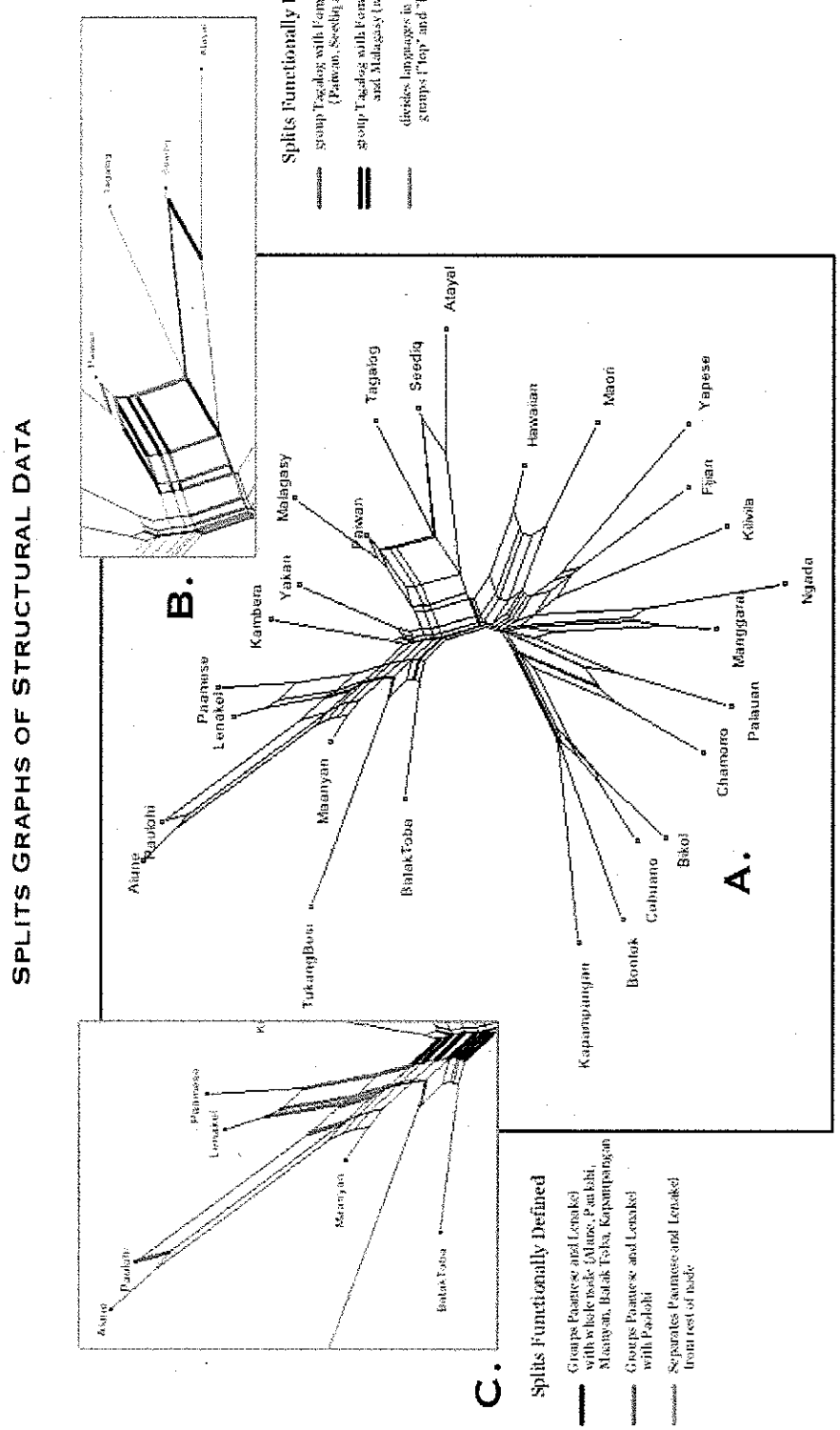


Figure 18A. A NeighborNet *splits graph* based on multi-state structural data. Close-up views of the *splits* contributing to a Philippines group (B) and to the relationships of Paamese and Lenakel (C).

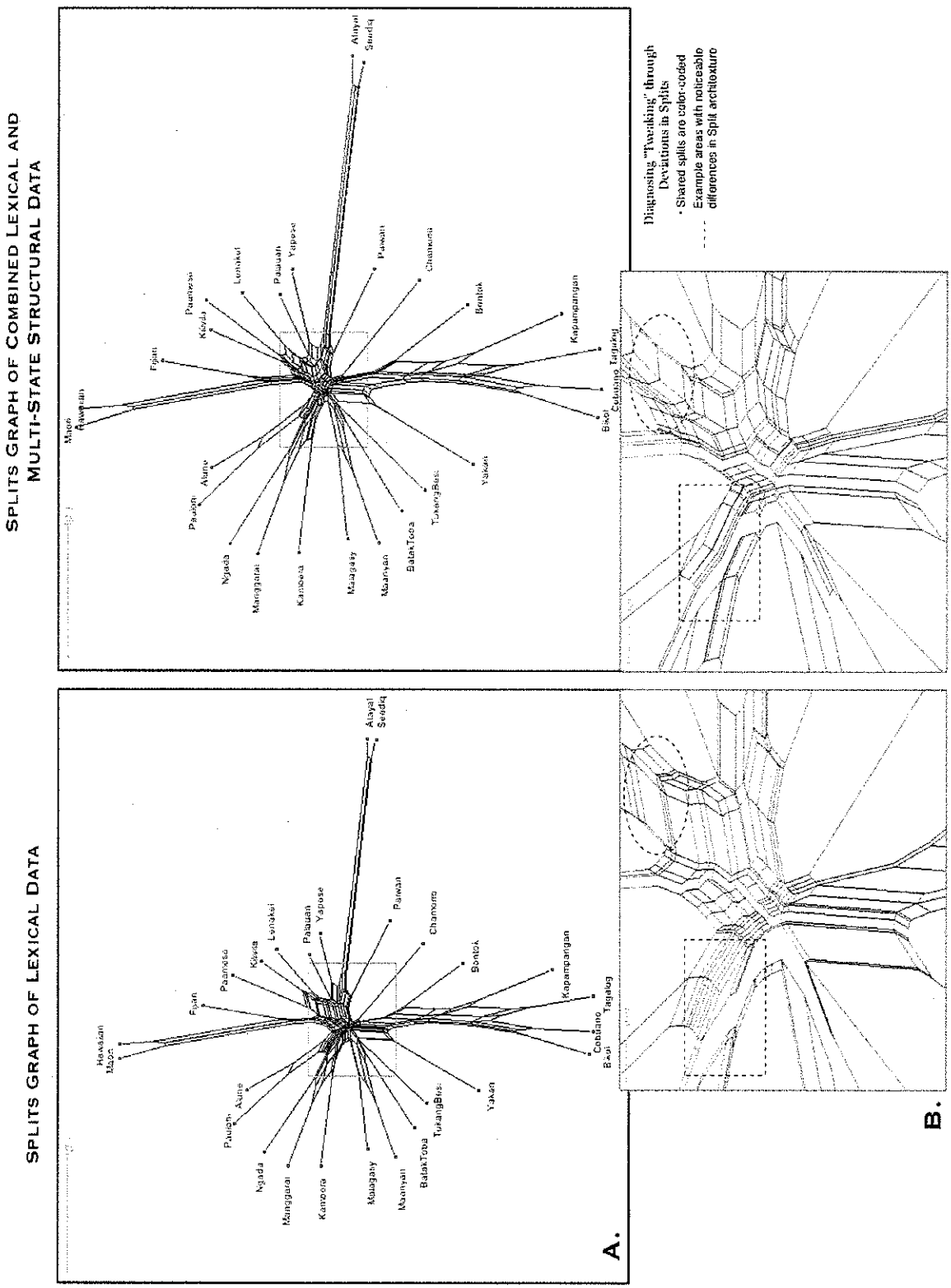


Figure 19. Tracing the "tweaking" effects of adding multi-state structural data to lexical data alone through changes in *splits*. *Splits* differences are undetectable while looking at the *splits graph* as a whole (A); rather, a close-up view is required (B).

ii. Step 2: Bayesian Inference of Phylogeny

Bayesian inference of phylogeny, as implemented through MrBayes, was empirically demonstrated in section IV to be both accurate and precise for modeling the evolution of a small sample of the Austronesian Family. Not only did Bayesian inference outperform both Distance and MP methods, but the tree produced was astonishingly historically accurate. Furthermore, the consensus approach inherent in MrBayes proved less prone to generating unsupported trees; a faculty which is especially important when comparing data of different types. Though MrBayes was run primarily on default settings, in this section the parameters for the evolutionary model used in MrBayes will be discussed. Understanding the role of these parameters is essential, since the Bayesian approach rests critically on these assumptions. Before tackling the details of MrBayes, a general explanation of how the Bayes' Theorem is applied to phylogenetic inference is provided (adapted from Felsenstein, 2004; www.egg.isu.edu/biocourses/bios599/projects/Walter_html). Note that Maximum Likelihood, a method statistically related to Bayesian analysis is not reviewed here due to its restrictive computation cost (Lewis, 2001).

iii. The Nuts and Bolts of Bayesian Inference of Phylogenies

Understanding the advantage of Bayesian inference in tree building requires some familiarity with basic terminology:

1. P (A) = probability of A occurring
2. P (A, B) = probability of A and B occurring (joint probability)
3.a P (A|B) = probability of A occurring given that B has occurred (conditional probability)
   b. P(A|B) = the likelihood of B given A

In 3a, the probability of an event A is based on the assumption of B. In 3b, the same conditional probability function is present, but the first argument (A) is fixed. The relation between B to A is now one of likelihood.

Extending this simple example to phylogenetics, take now the basic formulation of Maximum Likelihood:

1. P(data| model + tree)

While this statement can be read as the likelihood of the model+tree given the data, the meaning in terms of probability has a problematic flipside: If we express the statement in probalisitc terms, we uncover an unsuitable dependency: "the probability of the data occurring given the tree + model," To avoid having the data dependent on the tree we can reformulate the statement to ensure independence:

2. P(model+tree|data)

To understand the role of Bayes' Theorem in calculating this value, the above formulation will once again be simplified as P(A|B), where A = model + tree and B = data. We are back to the following:

3. P(A|B)

We can then use the above statement and an intuitive knowledge of conditional probability to formulate the following equation using the product rule:

4. P(AB) = P(A|B) P(B)

Since in the construction of trees, the two "events" of A and B have no explicit temporal order, we can write the following:

5. $P(AB) = P(BA) = P(A|B)\ P(A)$

Rearranging the above equation brings us to a common form of Bayes' Theorem, where the probability of one event (A) can be computed from the observations of another event and the knowledge of the joint distributions:

6. $P(A|B) = P(B|A) \times P(A) / P(B)$

Now let us once again replace our simplified variables. This replacement shows the phylogenetic incarnation of Bayes' Theorem:

7. $P(\text{tree+model}|\text{data}) = P(\text{data}|\text{tree+model}) \times P(\text{tree+model}) / P(\text{data})$

We can calculate P (data|tree + model) using the established ML methods. The term P (tree+model) has to be assumed, as it is the prior probability of the tree and model. In most cases, all trees are equally probable. The difficult term is P(data), which through rearrangement can be shown to equal the sum of the likelihood x prior probability of the tree+model for all possible trees:

8. $P(\text{data}) = \text{Sum } P(\text{data}|\text{tree+model}) \times P(\text{tree+model})$

To estimate this term, which is impossible to directly calculate, phylogeneticists have adopted the Metropolis-Hastings algorithm, a type of Markov chain Monte Carlo (MCMC). As a class, these algorithms are commonly used to calculate multi-dimensional integrals. This feat is accomplished by a set of "walkers" ("chains" in MrBayes terms) that move throughout tree space and sample from those areas with high posterior probabilities; a value equal to the term P(tree+model|data) in equation 7. This process is accomplished by the Metropolis-Hastings algorithm in 7 steps:

1. Start at some tree T_i
2. Pick a neighbor of this tree, T_J

3. Calculate the ratio of likelihood for T_i and T_j

(This ratio is taken from equation 7 above, where the P(data) term is identical for every sample and thus cancels out)

R = likelihood (T_j) / likelihood (T_i)

4. If $R > 1$, accept new tree
5. If $R < 1$, pick random number between 0 and 1; if random number is $< R$, accept new tree
6. If not, reject new tree and continue with T_i
7. Return to step 2

MrBayes uses multiple algorithmic searching "chains" simultaneously (the default is four). The basic strategy for finding optimal tree space is to have one locally constrained 'cold' chain, which picks only its close neighbors, and multiple 'hot' chains which are free to make larger jumps in tree space. After every generation, the posterior probability scores are compared so that the chain with the highest score becomes the 'cold' chain.

Tree samples are taken from the 'cold' chain after a number of iterated generations. The distance between samples should ensure sample independence. As the program runs, the four chains start to converge on areas of high posterior probability; as a result, the standard deviations between them begin to close. Once the chains have converged on the area of tree space with the highest posterior probability, they continue to sample, compiling a set of optimal trees. The pre-optimal trees, sampled before chain convergence, are thrown out through an operation called the 'burnin'.

From that set of optimal trees, support indices are calculated. These values are easy to interpret. They simply correspond to the probability of a specific clad given the set of optimal trees. From the optimal tree set, MrBayes produces a Majority

Rule Consensus Tree, where only those nodes that appear in 50% or more of the optimal trees are realized.

iv. MrBayes and Language Data

Bayesian methods have been employed by a large number of researchers to tackle a diverse set of issues related to inferring phylogeny (Felsenstein, 2004). In the face of this complexity, perhaps the clearest way to illustrate the Bayesian approach is to walk through the most essential settings of MrBayes, thus introducing the program in specific as well as the method in general. Additionally, the statistical assumptions underlying the analysis of discreet standard data, such as that of language, will be discussed.

MrBayes has two main groups of settings, one for specifying the structure of the model (*lset*) and a second for defining the prior probability distributions (*prset*) necessary for Bayesian inference. The model determines how character states change for a given character. For discreet standard data the model is quite simple since all characters of comparison occur only once. But when the same characters occur repeatedly in a given data set, as is the case for molecular (DNA or protein) data, more complicated models are available to account for the behavior of these dynamic, "building-block" type characters. This difference underlies the discreet versus continuous distinction of data type and correspondingly determines whether a symbol representing a character state is or is not equivalent to the same symbol for a different character. For discreet standard data, the latter is assumed under a condition called *arbitrary state labels*.

Prior probability distributions (*priors*) are equivalently necessary no matter the data type. It is in determining the *priors* where Bayesian methods become controversial. Take this example adapted from Felsenstein (2004): Imagine we send a scout to MIT to find nerds. No nerds are found. Assume our scout was not perfectly diligent so we have only a 1/3 chance of finding them if they were there. If my prior belief was 4:1 that nerds do not exist at MIT, then the posterior odds ratio for nerds at MIT is 1/4 x 1/3 = 1/12. If my prior belief was different, say 4:1 in favor of nerds at MIT, then the posterior odds ratio would change dramatically: 4/1 x 1/3 = 4/3. In terms of the phylogenetic incarnation

of Bayes Theorem presented above, *priors* need to be estimated for the trees and model term (shown in bold below):

7. P(tree+model|data) = P(data|tree+model) x **P(tree+model)** / P(data)

In effect, the *priors* of the model determine the how the MCMC algorithms move through tree space and the *priors* of the tree determine the parameters of tree space in terms of possible topologies and branch lengths (Huelsenbeck, *et al.*, 2001).

v. The Structure of the Model

For standard discreet data, MrBayes uses a model borrowed from Lewis (2001) in which all substitution rates between characters states are equal. This assumption neatly skirts the need for *arbitrary state labels*, since the rates for individual character state changes represented by the same numeric will not differ depending on the character. Figure (20a) shows an example of this model for 3 character states:

$$Q = \begin{bmatrix} & [0] & [1] & [2] \\ [0] & - & 1 & 1 \\ [1] & 1 & - & 1 \\ [2] & 1 & 1 & - \end{bmatrix}. \qquad Q = \begin{bmatrix} & [0] & [1] & [2] \\ [0] & - & 1 & 0 \\ [1] & 1 & - & 1 \\ [2] & 0 & 1 & - \end{bmatrix}.$$

Figure 20. A. The model used by MrBayes for substitution rates of unordered character states. B. The model for ordered character states (0 -> 1 -> 2->1 ->0). All rates must be equal to satisfy the *arbitrary state labels assumption* (taken from Rohnquist, 2005).

Forcing the characters through an order is also possible (figure 20b) though this option is not implemented here.

With continuous data, MrBayes can estimate a number of different model parameters which allow for greater flexibility and therefore a higher potential for accuracy. Two such parameters are *unequal state frequencies* and *substitution rates* between character states; the latter is dependent on the former and both are unavailable for discreet data due to the *arbitrary state labels* condition (Rohnquist *et al.*, 2005). MrBayes does however allow for inter-character variation, meaning that the overall rate

at which a given character is allowed to change its character states can vary. This parameter seems especially appropriate for structural language data where there is little expectation that (for example) phonological and word-order characters would evolve at the same rate. In this study, the inter-character rates of change were allowed to vary by Dirichlet distribution for multi-state or a beta-distribution for binary-state data (*lset rates = gamma*) (figure 21).

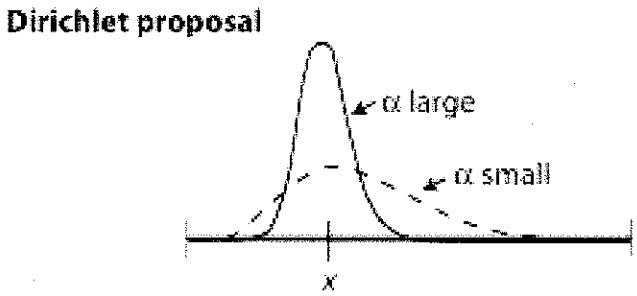


Figure 21. The Dirichlet or analogous Beta-distribution curves. Values are centered around "x "and allowed to vary to a lesser (α large) or greater (α small) extent (from Rohnquist *et al.*, 2005).

Distributions are one way to provide mathematical structure for the estimation of a parameter.

vi. Setting the Model's Priors

The *priors* available for the discreet data model are only a sub-set of those available for continuous data models. If inter-character rate variation is activated, there are four relevant parameters. Each is described with their default *priors* in parentheses: the state frequencies (all are equal), the shape of the state frequency distribution (uniform gamma or beta distribution), the topology (all equally probable) and the branch lengths (unconstrained). The role of the first two parameters was described above so only the latter two are described below.

When defining the tree topologies available for sampling by the MCMC "chains," the normal procedure is to assume all topologies are equally probable *a priori*. This follows similarly for language data. However, in some cases one may wish to force certain nodes together based on other evidence; this option is available in MrBayes, but

only in an all-or-none fashion; future versions of MrBayes should incorporate a probability measure for pairing nodes *a priori* (Rohnquist *et al.*, 2005).

Branch length are also estimated as part of the model. The two basic options are unconstrained or constrained; the latter of which is a setting appropriate for a molecular clock or lexicostatistics type approach. In this analysis the unconstrained option is used, along with the default setting for how much variability to allow in branch length determination (exponential: 10).

In Bayesian Analysis branch lengths have a different meaning then in either NJ or MP. In NJ, branch lengths directly reflect the pair-wise distance scores. In MP they represent discreet number of character state-changes. In Bayesian analysis however, branch lengths are just another aspect of the *posterior probability* value, maximized by the MCMC algorithms as they search through tree space.

vii. The Results: Bayesian Phylogenies for the Full Data Sets

The approach outlined above was implemented in MrBayes for three data sets containing all 27 languages: Lexical, multi-state structural, and lexical and multi-state structural combined. In each case the run lasted for 10 million generations with a sample taken every 100 trees. Of the 100,000 trees collected, only the 25,000 most optimal were kept (a 'burnin' of 75,000 trees). Convergence of the chains toward a tree space of equal probability was judged by the lack of trends associated with the posterior probability values displayed after the completion of the run. Consensus trees were then compiled from those 25,000 trees; only those nodes occurring in at least 50% of the optimal set were included. The results are presented in Figure 22, 23 and 24.

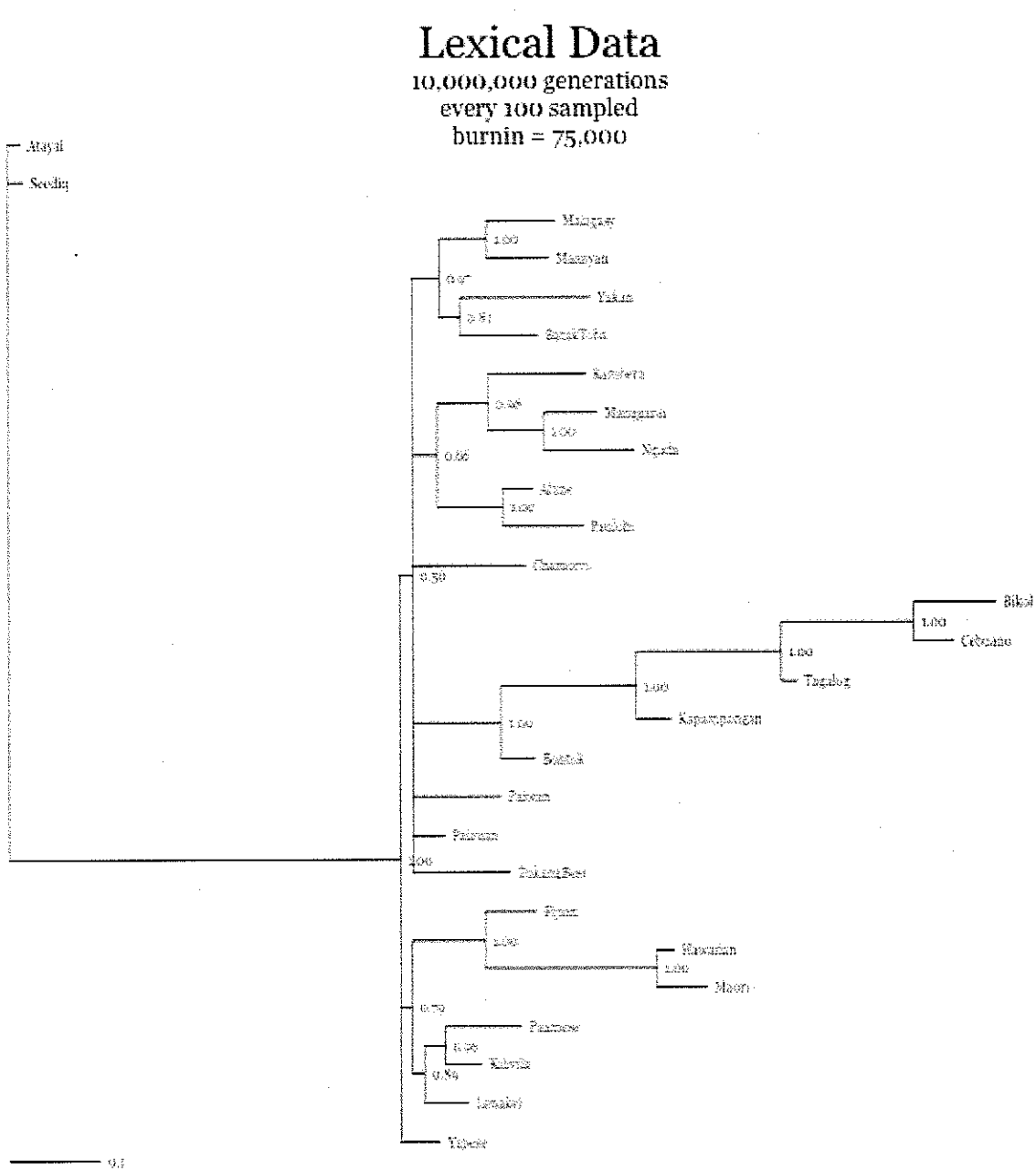


Figure 22. A Bayesian phylogeny inferred from lexical data. The run lasted 10 million generations and samples were taken every 100. A majority rule consensus tree was then created from the 25,000 most probable trees. Numbers at each node (from1.0 to 0.50) represent the probability of finding that node in the set of optimal trees and are thus a measure of nodal support.

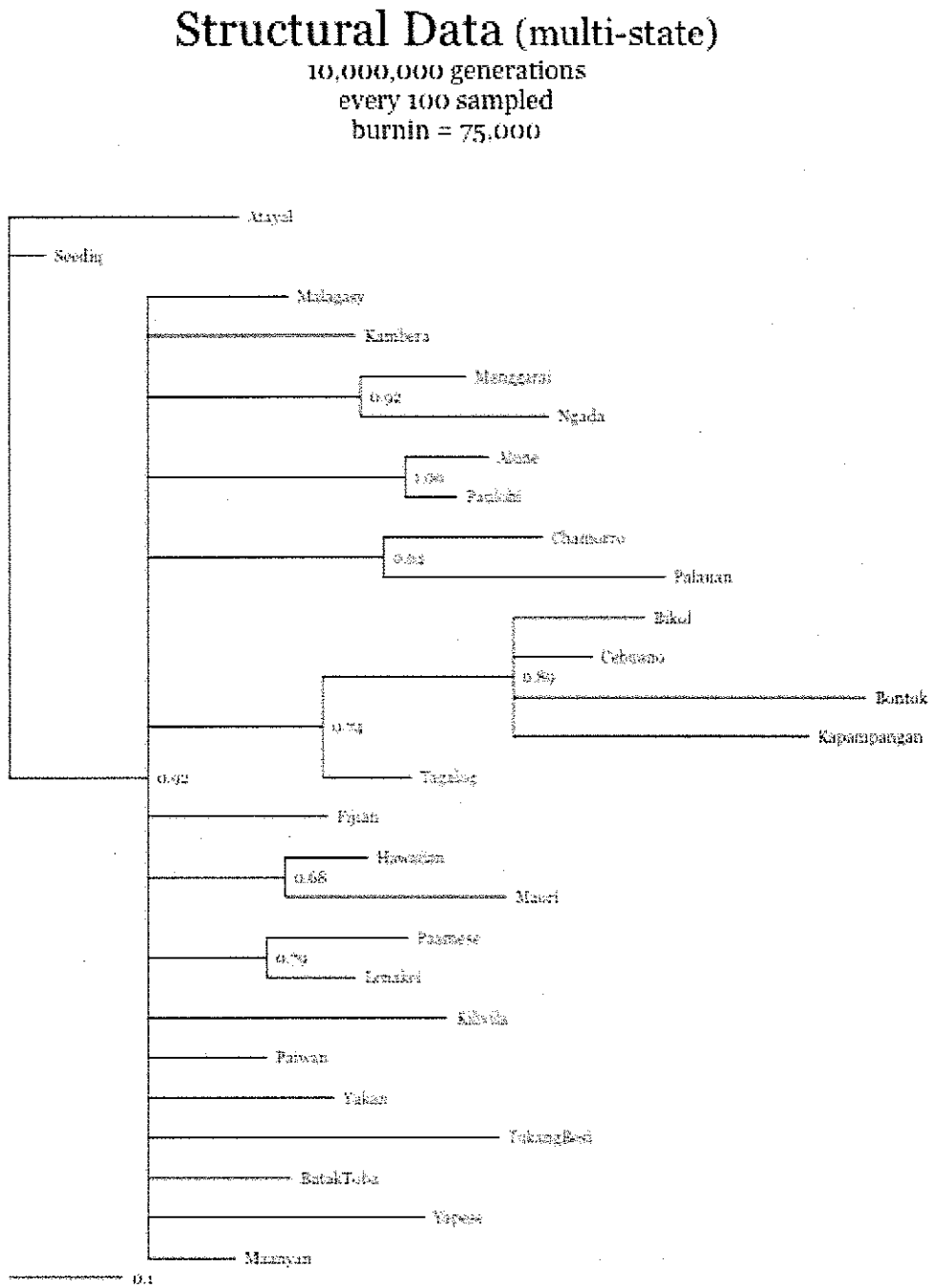


Figure 23. A Bayesian phylogeny inferred from multi-state structural data. The run lasted 10 million generations and samples were taken every 100. A majority rule consensus tree was then created from the 25,000 most probable trees. Numbers at each node (from1.0 to 0.50) represent the probability of finding that node in the set of optimal trees and are thus a measure of nodal support.

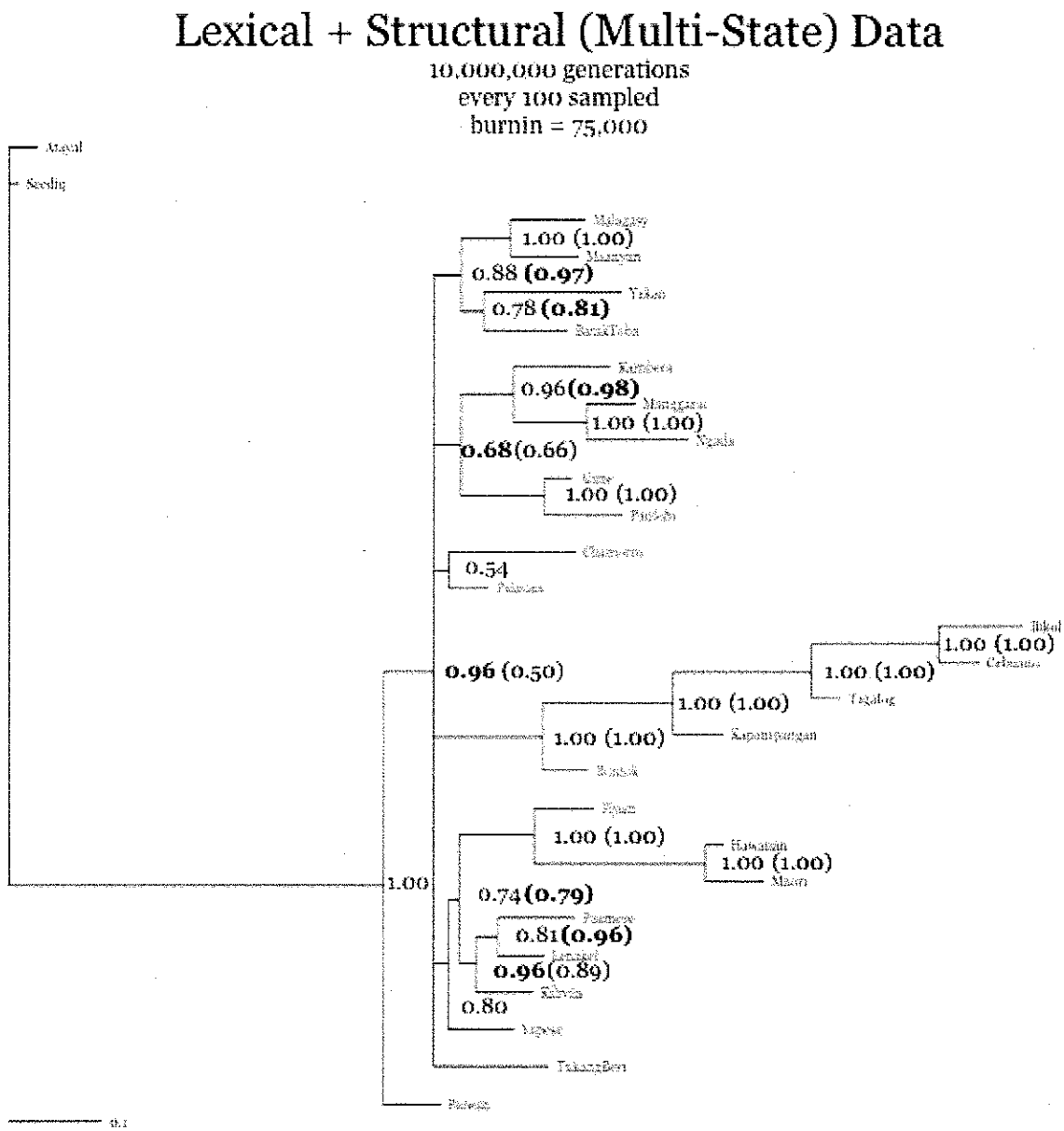


Figure 24. A Bayesian phylogeny inferred from a combined lexical and multi-state structural data set. The run lasted 10 million generations and samples were taken every 100. A majority rule consensus tree was then created from the 25,000 most probable trees. Numbers at each node (from1.0 to 0.50) represent the probability of finding that node in the set of optimal trees and are thus a measure of nodal support. Purple values represent the support for the node with the combined data set. Blue values represent the support for the identical node in the lexical phylogeny. Red values indicate new nodes induced by the presence of the structural data.

The comparative results between full data sets are similar with and without the WMP outliers (Palauan, Chamorro, and Yapese). The format used for displaying the phylogenetic results of the full data sets is different, however. When methods were compared earlier, only *cladograms* were presented. Here branch lengths are included and thus the trees are *phylograms*. Additionally, support values are reported for each node. As described previously, these values correspond to the probability of finding this node in the optimal set. This means that the higher the value (the closer to 1.0), the more a given node is supported.

The structural tree is obviously not as resolved as the lexical tree. In terms of mechanism, this means that many of the nodes suggested by the lexical data were not found in 50% or more of the 25,000 optimal structural trees. Nevertheless, the effect of combining data sets produced some importance differences, detectable by both changes in topology (the creation of new nodes) and the changes in support values for nodes present in both lexical and combined trees. An examination of these differences highlighted in figure 24 demonstrates some of the "tweaking" effects induced by combining data which lead to a more historically accurate tree. In this figure, new nodes induced by the presence of structural data are colored in red. Nodes that are present in both lexical and combined trees are represented in blue and purple respectively. The nodes in bold have the higher value.

Most conspicuously, the combined data set creates a ubiquitous node found in all optimal trees which accurately splits the Formosan languages from everything else. Thus a wholly accurate higher-order WMP grouping is present only in the combined data set. Secondly, in the combined data set Yapese, a known Oceanic outlier, is placed as the first branching member off of a node corresponding explicitly to the Oceanic languages. With the lexical data, Yapese is less resolved, branching off of the tree alongside (but not within) the other Oceanic languages. This lexical-tree position incorrectly places Yapese and the other Oceanic languages as the first two splits from the Formosan languages. This is historically inaccurate, since the Oceanic languages are known to be the last group to diverge (Ross, 1995). In other words, the lexical tree suggests that the Oceanic group

is the "oldest" (after the Formosan languages), when in fact they are the youngest. The last topological change groups two outlier languages, Chamorro and Palauan under a single node. I am unaware about any hypotheses purporting relatedness between these two languages. But if experts do believe they are related, the data presented here suggest that relationship is supported by structural rather than lexical similarity.

In terms of support values, combining data had the effect of weakening 5 nodes while strengthening 3. In terms of total nodal absolute differences, combining data resulted in an increase of 55 probability units (46 came stemming from a single node) and a decrease in 34 probability units with a much smaller distribution (from 2-15 points per node). These grouped values for nodal support are hard to interpret without referring to the nodes themselves, since increased support is not always a sign of increased accuracy.

The most drastic change in nodal support was seen for the approximate WMP node. With lexical data, this node, which separates the Formosan languages from everything else, was at the threshold probability of inclusion (0.50). In the combined data, when Paiwan was kicked out and thus an accurate WMP node was formed, the support value sky-rocketed by 46 points to 0.96. This increase in probability is one manifestation of increased accuracy due to "tweaking."

The most drastic decrease in support (15 probability units) was seen at the node connecting Lenakel and Paamese. This suggests that Lenakel and Paamese are more closely related in terms of lexicon than structure; a support node correlate to those same differences detected in the SplitsTree analysis above (see *Step 1: NeighborNet Analysis*).

In comparing support values for the nodes of a given family, a speculative set of hypotheses can be drawn about the lexical vs. structural nature of the evolution of certain groups. For instance, the high resolution of the Northern/Meso Philippines languages in both data sets suggest the group evolved relatively equally in terms of both lexicon and structure. Somewhat differently, the Philippines languages of Borneo, Sulawesi and Sumatra (Malagasy, Maanyan, Yakan, and Batak Toba) seem slightly more lexically conservative in their evolution.

In conclusion, despite the fact that there were far fewer structural versus lexical characters of comparison, the inclusion of structural data resulted in a "tweaking" which increased the accuracy of the tree. This improvement is reflected in both changes in topology (the inclusion of new nodes) and changes in the support for certain nodes, as evidenced by shifts in node support values. Additionally, the results from NeighborNet analysis have distinguishable correlates in the Bayesian tree, establishing a critical link between interpreting both representations of relatedness. And perhaps most importantly, a run of MrBayes for 10 million generations produced a tree which resolved all the major nodes of the 'known' tree. Interestingly, the inclusion of the outliers did in fact lead to a less resolved structure for the higher order nodes: a comparison of the Bayesian trees generated with a combined data set with and without outliers (figures 24 and 12) shows that without outliers, there are no *multifurcations*; that is, all nodes are binary, with the Meso/Northern Philippines languages splitting off from the Formosan languages first, followed by the other Philippine languages, with the last major split between the CMP languages and the Oceanic family. This non-outlier tree reflects the dynamics of the migration proposed by Ross (1995).

VI. Mapping Characters to Trees: the Association of Structural Features through Evolution

The basic method introduced in the last section demonstrated one productive application of how biological methods and software can be applied to model language evolution. The results reported that under appropriate conditions of analysis, a data set that included character comparisons for two different components of language produced a tree which mirrored the established tree remarkably well. This convergence of results from radically different methods suggests strong support for the current account of the natural history of Austronesian Family. In this section, I want to introduce how questions and methods previously restricted to biological evolution may provide new types of

insights into language evolution. This goal will be pursued through a specific example: how characters of comparison can be mapped to trees.

For an overview of the type of questions now being entertained about biological evolution, and the resulting software used to address them, interested linguists should visit Dr. Joseph Felsenstein's website: http://evolution.genetics.washington.edu/phylip/software.html. Reviewing some of the descriptions of the huge number of programs available should alert linguists to the vastness of the field and the variety of methods which may have applications to language data. The results of one such program SIMMAP (Bollback, 2005) are reported here.

SIMMAP provides a Bayesian framework for mapping the evolution of independent characters. In general, character mapping involves modeling the changes of individual characters onto an established tree. While originally done with MP, the advent of a Bayesian approach allows uncertainty to be taken account both in terms of the trees to which characters are mapped and the mapping process itself (Ronquist, 2004). This type of analysis has recently driven much progress in biology; by modeling the ancestral states of certain characters, a wide range of questions can be entertained, including the identification of ancient behaviors, the structure of proto-hormone receptors, and inference of past dispersal patterns (Ronquist, 2004). In terms of language evolution, inferring ancestral character states could provide insights into the approximate lexicon grammatical structures of Proto-languages; data which could then be compared to more conservative accounts provided by the comparative method.

SIMMAP was chosen because of both the solid performance of Bayesian methods in general and the fact that the program conveniently works hand in hand with MrBayes. In fact, the optimal set of trees which MrBayes outputs serves as one half of the input for SIMMAP; the second half of the input is the data matrix itself. Once SIMMAP has loaded the trees and the data a variety of analyses can be performed. Most basically, SIMMAP uses a process called *posterior mapping* (Neilsen, 2002) to evaluate the how characters change over the given set of most probable trees. For each tree in the set, the

character of interest is mapped. For example, take the mapping of WALS Feature 88: the Order of the Demonstrative and Noun, to 1 of the 25,000 most optimal trees (figure 25).

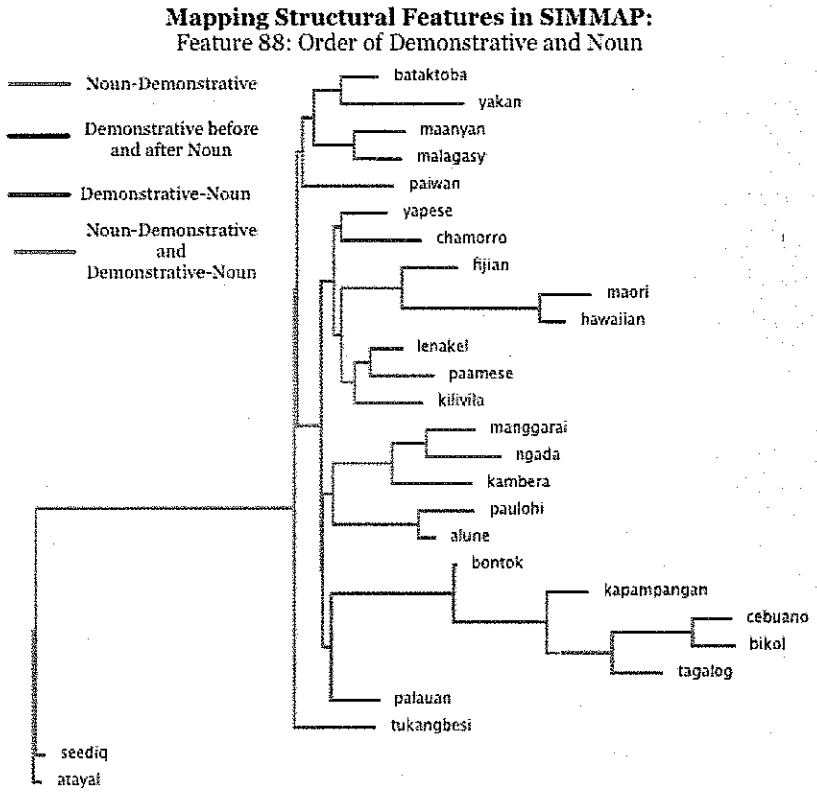


Figure 25. An example of a WALS feature mapped to one of a set of most probable trees. The character states of the feature are color coded to show where on the tree the changes take place.

The degree to which characters associate can be thought of as how linked those characters are in the evolutionary process. In this report, the linkage between structural characters is evaluated for the combined data set. In terms of language evolution, linkage can occur for three non-mutually exclusive reasons: 1. The biological basis of language results in psychological constraints which force elements of language to change in concert. 2. The social, political and geographical context of the development of the Austronesian Family resulted in unique character associations unrelated to language universals. 3. The characters associated are not independent of each other in a methodological sense. While the third reason is identifiable, the first two cannot be teased apart without comparison to other data sets comprised of the same structural characters.

SIMMAP calculates the association between two characters at both the character state and whole-character level. In the latter case, the equation used to calculate the pair-wise character state associations is modified to resolve the character as a whole[vi]. In this report, characters associations were evaluated at the whole character level. For each pair of the 26 structural characters, SIMMAP analysis produced a numerical value representing the degree of association. (Currently SIMMAP can only handle up to seven character states; for this reason, five characters were excluded). The distances were then compiled in a matrix. To visualize the associations, the 35 dimensional association space was lowered to two dimensions using a multi-dimensional scale (figure 26).

The results show a single area of clustering involving approximately fifteen features. What does this group have in common? Firstly, two pairs of features are obviously methodologically linked: Features 14 and 15, which corresponding to fixed and weight-sensitive stress categories respectively, each use "other" categories to code the languages detailed in the complimentary feature. Similarly, features 87 (order of adjective and noun) and 97 (relationship between the order of object and verb and order of adjective and noun) also suffer from this previously described "trashcan" problem. For this set of pairs, evolutionary association is methodological in origin and thus trivial.

Of the remaining ten features, seven pertain to order in either the noun or the verb phrase. The presence of this "ordering" cluster suggests that the changes in the relative positions of words or affixes may exert an evolutionary influence on each other. But if "order" is the theme of the cluster, what should be made of the association of the stress (features 14 and 15) and case (feature 28: case syncretism) features? Unfortunately pursuing this question theoretically is outside the scope of this report. Nevertheless, theories relating order to stress have been put forth (Strauss, 1983; Inkelas, in press).

While in isolation these associations remain difficult to interpret, they do present a new type of result much in the need of comparison. If data from a wide variety of language families were assembled into sets of similar size and content, and if structural characters of comparison were identical, then feature association results may provide the

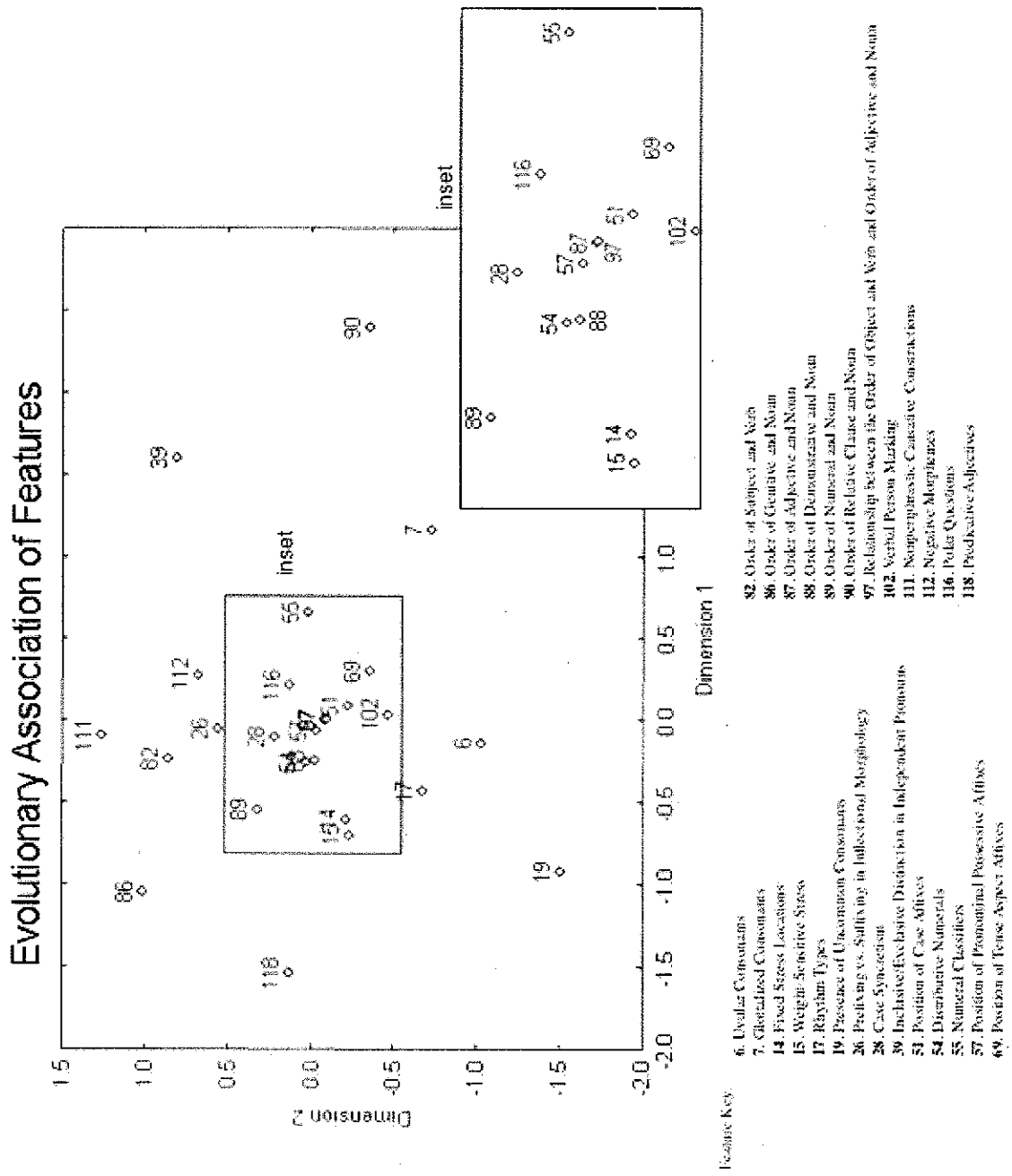


Figure 26. The evolutionary association of structural features in the combined data set, visualized through a multi-dimensional scale.

basis for a typology of language evolution. Such comparisons may suggest preliminary hypotheses for inferring the differences and similarities of how languages change over time.

Character mapping is only one of many methodological inspirations borrowed from biology. Others include the identification of characters likely attained through horizontal transfer, the ability to combine partially overlapping phylogenies into 'supertrees' and the use of host-parasite models to account for language change with respect to other processes of divergence (e.g. population genetics data). In sum, I hope the mention of a few of these possibilities will catalyze an interest in establishing standardized methodologies, since only through comparison will these new types of analyses gain meaning.

VII. Conclusion

The body of this report was devoted to outlining a method for the phylogenetic inference of language evolution. After first introducing the concept of linguistic phylogenetics and the sample languages, the processes of data collection and encoding were described. Next, four methods, two data types, and two encoding schemes were performance-evaluated against a 'known' tree. The results were synthesized into a two-step method which was then described in detail. Lastly, the association of structural features through evolution was modeled using software built for biology, thus demonstrating one of many possible productive intersections between biological methods and linguistic data.

The method proposed produced a tree that was strikingly historically accurate. While an important proof of concept, the topology of the final tree should not distract from other informative trends about the data type and robustness of method.

One such trend was that the most accurate data sets were those with combined lexical and structural data. In these cases, despite contributing only a small number of

total characters of comparison, the structural data had a traceable "tweaking" effect toward greater accuracy, especially noticeable in the two-step method put forth. With NeighborNet analysis, this "tweaking" was defined by comparing the architecture of *splits* between separate and combined data sets. With MrBayes, these changes were detected by comparing the topologies and node support values for the lexical and combined data trees. This ability to associated changes in input data across methods of representation helps NeighborNet and Bayesian analyses work in compliment to evaluate the details of a phylogenetic signal. At its best, this type of inter-data inter-method comparison may be able to distinguish groups of characters based on evolutionary behavior.

One important parameter of behavior is likely to be evolutionary stability. Though not attempted in this report, if structural data is to be used to deepen the time barrier of reliable phylogenetic inference, it should be theoretically possible to use only those structural characters whose phylogenetic signal most closely matches that of the lexical data. While the results from the *incongruence length difference* test suggest that two data sets do express significantly different phylogenetic signals, this difference is based on the principal of minimum change inherent in MP and thus may have limited applicability, especially considering the poor performance of MP presented in this analysis. Nevertheless, other measures of assessing differences in phylogenetic signal more conducive to language data will undoubtedly be developed and those may help identify sub-sets of structural features which share the same phylogenetic signal as the "shallower" and more numerous lexical characters of comparison.

In this study, structural characters were chosen from WALS based largely on their "P-Value," a measure of world-wide stability within a genus. Despite the empirical merit of the P-Value (Wichmann and Kamholtz, unpublished), this measure has obvious drawbacks and is likely to be controversial. While it will be helpful to have some criteria for establishing the most conserved structural characters *a priori*, I hope that the methodological options presented here demonstrate that there is more information in modeling structural and lexical data than just a simple tree.

Taken together, navigating the interaction of language as a component system is as vital to understanding evolution as it is to studying cognition. So while biological data destined for phylogeny may be understood in reductionist fashion, by guiding principals like natural selection and the rules of chemical interaction, linguistic data cannot escape the complexities of cognition and social living. As Enfield (2005) put it, "although integrated into a structured system in the cognition of individual speakers, a language's constituent items are separable, each with their own careers across the community of minds" (p.195). Thus studying language history straddles disciplines, a fact which reiterates the importance for the integration of multiple perspectives. After all, more than organisms, the evolutions of languages are grounded in places and events with other detectable correlates. For this reason, piecing together the natural history of a language should not only involve resolving component histories, but should also incorporate data from meta-language studies involving archaeology and population genetics.

These differences between evolutions suggest that not all questions pursued by biologists have direct or equivalent application to language. Nevertheless, linguists should draw inspiration from the statistical, computational, and infrastructural developments of biological phylogenetics, since many of the mathematical, philosophical, and practical considerations of inferring phylogeny are shared.

For example, the careful coordination of comparative analyses has been greatly aided through biologists establishing an unrivaled system of databases. Since only an Internet connection is required to access data which is both abundant and exceedingly complex, the presence of these databases has fostered a haven for interdisciplinary research, most notably for computer scientists and mathematicians. The results of this open-market, interdisciplinary approach are self-evident: profound progress and immense popularity.

In linguistics however, progress is primarily made by those who have invested a lifetime in research. Their knowledge should not be devalued, just recorded, reformatted and made accessible for all. This amounts to setting standards for databases and software

programs. Before concluding, I want to make several suggestions toward this end which stem from my own experience trying to compile the data for this study.

For lexical information, databases should be systematic both in terms of word selection and format. Conserved word lists allow for large-scale comparisons and avoid error due to patchy sampling. Words should be formatted on multiple levels including constituent sounds and morphemes, to facilitate phonological, structural (morpheme-level) and cognate comparisons. If done consistently, this approach may allow computational approaches to bridge the gap with the comparative method.

Structural databases should be similarly organized on a series of different levels. This most likely means coding data in terms of both language specific morphemes and typologically categories. Importantly, the construction of these typological parameters should fit with the mandate of the type of analysis in mind. In terms of phylogenetics, this translates to structural characters which avoid the 'independence', 'apples and oranges', and 'trashcan' problems; in other words, homogenous and independent characters of comparison.

Phylogenetics software should also be designed explicitly for language data. Ideally this software would support a variety of data types: from the IPA for phonological comparisons to data coded as numbers of structural comparisons. For cognate comparisons, the software should be able to align words, assign cognate classes and produce summaries of types of sound change. Functionally, the software package should be general enough to incorporate a number of different analyses, all of which should be tailored for language. As an example from the results presented here, one might conceivably run a network analysis, phylogenetic analysis and association of feature analysis at the touch of a button after setting several general parameters.

In the long term, I hope the promising results of this report convince historical linguists that the direction taken by biologists over the last 30 years is worth following. In the short term, I hope these results demonstrate that current methods of phylogenetic biology have productive applications for the modeling of language evolution. I also hope biologists, computer scientists and mathematicians interested in evolutionary processes

may see language evolution as viable system worth exploring. After all, compared to organisms, the natural histories of languages are thoroughly untold.

Acknowledgments

The bulk of this research was undertaken at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany during the summer of 2005 and I must graciously thank Dr. Bernard Comrie for making my stay possible. Many of the ideas expressed in this paper grew from extended conversations with Dr. Søren Wichmann, Dr. Michael Cysouw, and Mihai Albu; without their patient assistance, I'd still be looking for trees outside! I also must thank Hans-Jörg Bibiko for the technical assistance and Dr. Russel Gray and Simon Greenhill for graciously providing the lexical data from the ABV database. I am also indebted to Dr. Malcolm Ross who gave me an inspirational afternoon of his time, lending a truly expert opinion on the Austronesian languages. Lastly, I must thank Dr. David Harrison for his constant support over the last three years and for introducing me to his friends at the Max Planck. And despite all the wonderful help, I am solely responsible for any errors, untruths, or misinterpretations presented in this report.

---

[i] Comparisons of language borrowing situations suggest that all aspects of language can be borrowed (Curnow, 2003). However, not all language components may be borrowed readily, thus some may prove more helpful for phylogenetic inference.

[ii] As Gray (2005) points out, the purpose of this graphic is to show that given a model which takes different rates of lexical evolution into account, it is clear that while the majority of shared words between related languages will quickly disappear, there are some which may still be shared after 20,000 years. Hence, different words are constrained differently through the evolutionary process.

[iii] The P-Value is a measure of the stability of a feature within its WALS Genus. Calculations took all of the language data into account, thus the P-Value is not specific to the Austronesian Family.

The P-Value is represented by the three formulae below. For a detailed explanation of the logic and justification of the P-Value, see Wichmann and Kamholtz, 2005.

$$p(n,k,r) = \frac{C(n,k,r)}{k^n}$$

$$C(n,k,r) = k\binom{n}{r}Q(n-r,k-1,r+1) - \sum_{i=\max(2,n-k(r-1))}^{\lfloor n/r \rfloor} \left((i-1)\binom{k}{i}\binom{n}{ir}\prod_{j=2}^{i}\binom{jr}{r}Q(n-ir,k-i,r) \right)$$

$$Q(n,k,r) = k^n - \sum_{i=r}^{n} C(n,k,i)$$

[iv] Only continuous data can be individually compared to a consensus tree. Fore example, the large number of characters (in the form of nucleotide bases) which make up a gene make it feasible to compare the phylogeny of an individual gene versus the phylogeny of the organism. This type of comparison is impossible with discreet data however, since each coherent comparable unit is comprised of only a single character.

[v] Dunn *et al.* (2005) used MP to build trees for Oceanic and Papuan languages based solely on structural data. They justify using their method and data type on the Papuan data, about which little is known, by first attempting to demonstrate that their experimental Oceanic tree is topologically close to the known tree. Using the *symmetrical differences test*, Wichmann and colleagues (the author included) showed that a tree produced with Bayesian analysis on the same data set was closer to the known than the Dunn *et al.*'s MP tree (SD of 6 vs. 8). One thousand random trees were then generated by a model which accounts for language change (Wichmann, unpublished) to assay whether these results occurred by chance. These random trees were then compared to each other using the *symmetric difference test*. The results show that while Dunn *et al*'s MP tree was significantly unlikely to be due to chance, the tree based on Bayesian analysis more accurately reflects the known; thus in another context MP was outperformed by Bayesian analysis for modeling language evolution.

[vivi] The equation below is used by SIMMAP to calculate the association between any two discreet characters states *i* and *j* where the terms *e* and *o* correspond to "expected" and "observed" association respectively. The amount of association is defined by the proportion of time along the tree which is shared by the two states, where more association than expected gives a positive value, and less association than expected gives a negative value (Bollback, 2005).

$$d_{ij} = a_{ij}^{(o)} - a_{ij}^{(e)}$$

The association of two characters in total is thus the summation of the association for each individual character state. This summation is represented by the equation below:

$$D = \sum_{i}^{n} \sum_{j}^{m} |a_{ij}^{(o)} - a_{ij}^{(c)}|$$

Appendix

Table 1. The complete sub-groups for WMP by Ross (1995)

WMP Subgroup	Location		WMP Subgroup	Location
1. Batanic	(not explicitly listed)		13. North-West Borneo	North-west Borneo
2. Northern Philippines	Northern Philippines		14. Land Dayak	Inland south-west Borneo
3. Meso-Philippines	(not explicitly listed)		15. East Barito	South-east Borneo and Madagascar
4. Southern Philippines	Southern Philippines		16. Barito-Mahakam	South-east Borneo
5. South Mindanao	(not explicitly listed)		17. West Barito	Southern Borneo
6. Chamorro and Palauan	Mariana Islands and Belau		18. Lampung	South-east Sumatra
7. Sangiric	(not explicitly listed)		19. North-west Sumatra/Barrier Islands	North-west Sumatra/Barrier Islands
8. Minahasan	North-eastern Sulawesi		20. Java-Bali-Sasak	Java and Bali
9. Gorontalo-Mongondic	Northern Sulawesi		21. Central Sulawesi	Central Sulawesi
10. Sama-Bajaw	Sulu Archipelago and other locations in Philippines		22. South Sulawesi	South Sulawesi
11. Malayo-Chamic	(not explicitly listed)		23. Muna-Buton	Islands off of south-east Sulawesi
12. Moken and Moklen	Islands off the coast of Thailand and Myanmar		24. Tamanic	Central Borneo

Table 2. The complete sub-groups for WMP by Ross (1995)

Subgroup	Location
1. Bima-Sumba	Eastern part of Sumbawa, Sumba, Flores
2. Timor	Timor
3. South-East Maluku	South-East Maluku
4. Aru	(not explicitly listed)
5. Central Maluku	Seram, Buru and their offshore islands
6. North Bomberai	South coast of MacCluer Gulf, Irian Jaya
7. Koiwai	South coast of Bird's Neck, Irian Jaya

Table 3. The complete sub-groups for Oceanic by Ross (1995)

Subgroup	Location		Subgroup	Location
1. Admiralty Islands	Admiralty Islands		7. North/Central Vanuatu	North/Central Vanuatu
2. St. Matthias Islands	St. Matthias Islands		8. South Vanuatu	South Vanuatu
3. Western Oceanic	Papua New Guinea and the western Solomon Islands		9. Southern Oceanic	New Caledonia and the Loyalty Islands
4. Sarmi/Jayapura Bay	Sarmi/Jayapura Bay		10. Nuclear Micronesian	Micronesia
5. Southeast Solomonic	Southeast Solomon Islands		11. Central Pacific	Rotuma, Fiji, Polynesia, New Zealand
6. Utupua and Vanikoro	Te Motu Province, Solomon Islands		12. Yapese	Yap

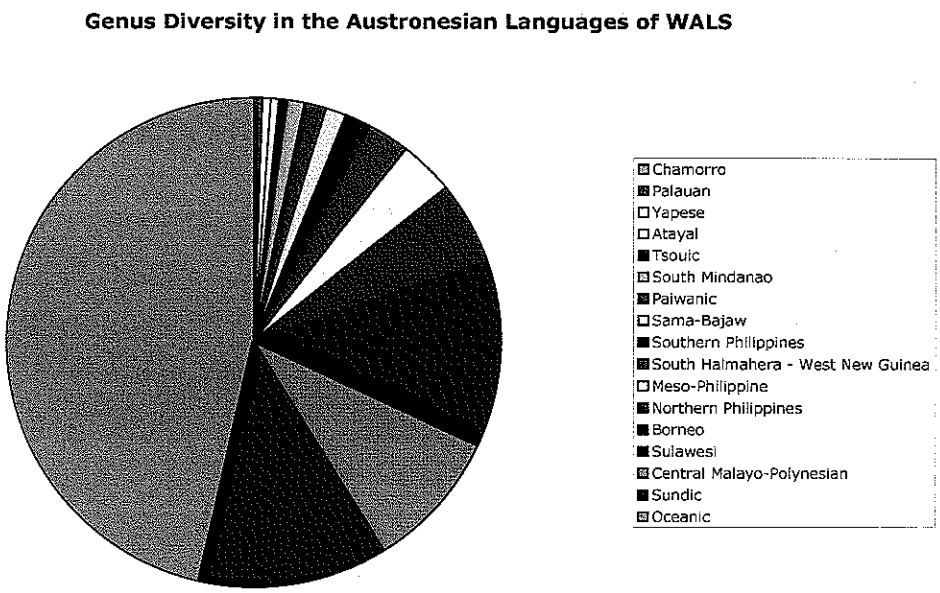


Figure 1. The Number of Languages in Each of the 17 WALS Designated Genera

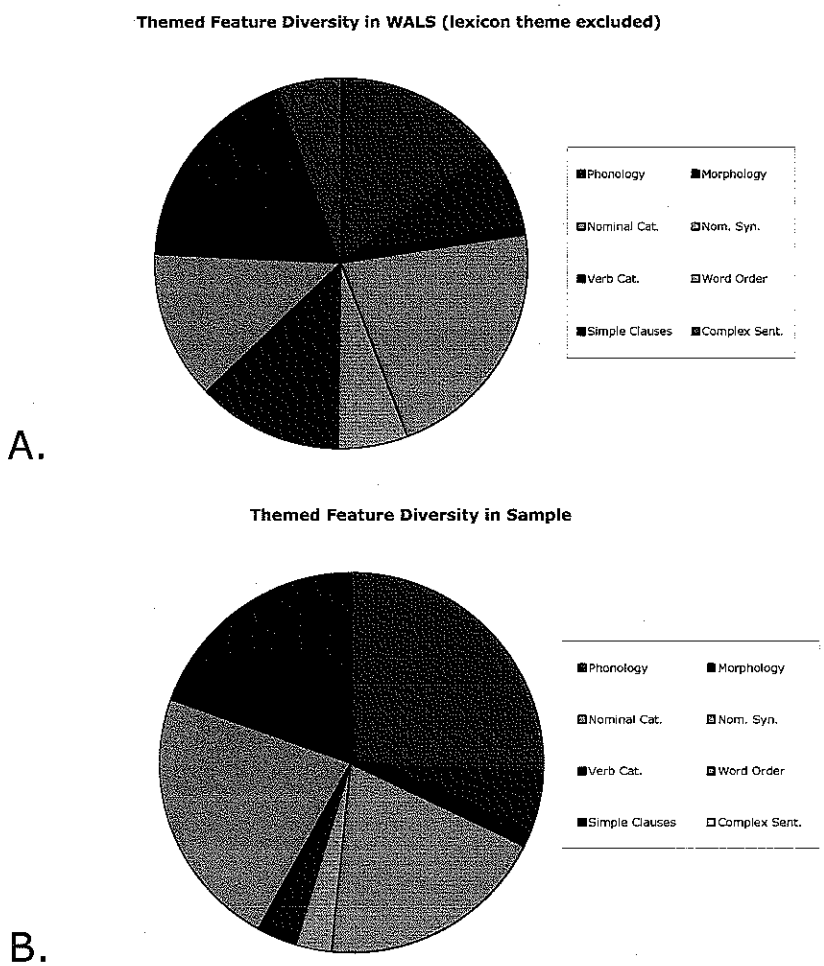


Figure 2. The Themed Diversity of Features in the WALS Database (A.) and the Language Sample (B.)

Table 4. The WALS Features Selected for Analysis.

Features are ranked by P-Value and color coded by WALS "theme." Exluded features are also indicated.
(* = uninformative, # = dependent, % = sparse attestation)

Theme	WALS Feature ID Number	Excluded from Final Set?	Feature	P-Value
Phonology	18		Absence of Common Consonants	13.4
Word Order	90		Order of Relative Clause and Noun	16.8
Word Order	88		Order of Demonstrative and Noun	17.5
Phonology	11	*	Front Rounded Vowels	18.5
Nominal Categories	51		Position of Case Affixes	18.7
Word Order	89		Order of Numeral and Noun	19.7
Nominal Categories	33		Coding of Nominal Plurality	20.5
Word Order	87		Order of Adjective and Noun	20.9
Word Order	86		Order of Genitive and Noun	21.6
Word Order	81	#	Order of Subject, Object and Verb	22.2
Nominal Categories	30	*	Number of Genders	22.5
Nominal Categories	54		Distributive Numerals	22.6
Simple Clauses	118		Predicative Adjectives	22.6
Phonology	7		Glottalized Consonants	22.9
Phonology	19		Presence of Uncommon Consonants	23.3
Nominal Categories	39		Inclusive/Exclusive Distinction in Independent Pronouns	23.4
Word Order	82		Order of Subject and Verb	23.4
Simple Clauses	116		Polar Questions	23.6
Nominal Categories	46	%	Indefinite Pronouns	23.9
Nominal Categories	57		Position of Pronominal Possessive Affixes	23.9
Phonology	6		Uvular Consonants	24
Verb Categories	69		Position of Tense-Aspect Affixes	24.1
Word Order	97		Relationship between the Order of Object and Verb and the Order of Adjective and Noun	25.4
Simple Clauses	111		Nonperiphrastic Causative Constructions	25.7
Morphology	28		Case Syncretism	26.8
Simple Clauses	101		Expression of Pronominal Subjects	26.8
Nominal Categories	55		Numeral Classifiers	27.6
Phonology	17		Rhythm Types	27.7
Morphology	26		Prefixing vs. Suffixing in Inflectional Morphology	28
Nominal Syntax	61		Adjectives without Nouns	28.1
Simple Clauses	112		Negative Morphemes	28.5
Complex Sentences	123	%	Relativization on Obliques	29.6
Phonology	14		Fixed Stress Locations	30.8
Phonology	15		Weight-Sensitive Stress	31.5
Simple Clauses	102		Verbal Person Marking	36.4
Phonology	16		Weight Factors in Weight-Sensitive Stress Systems	36.6

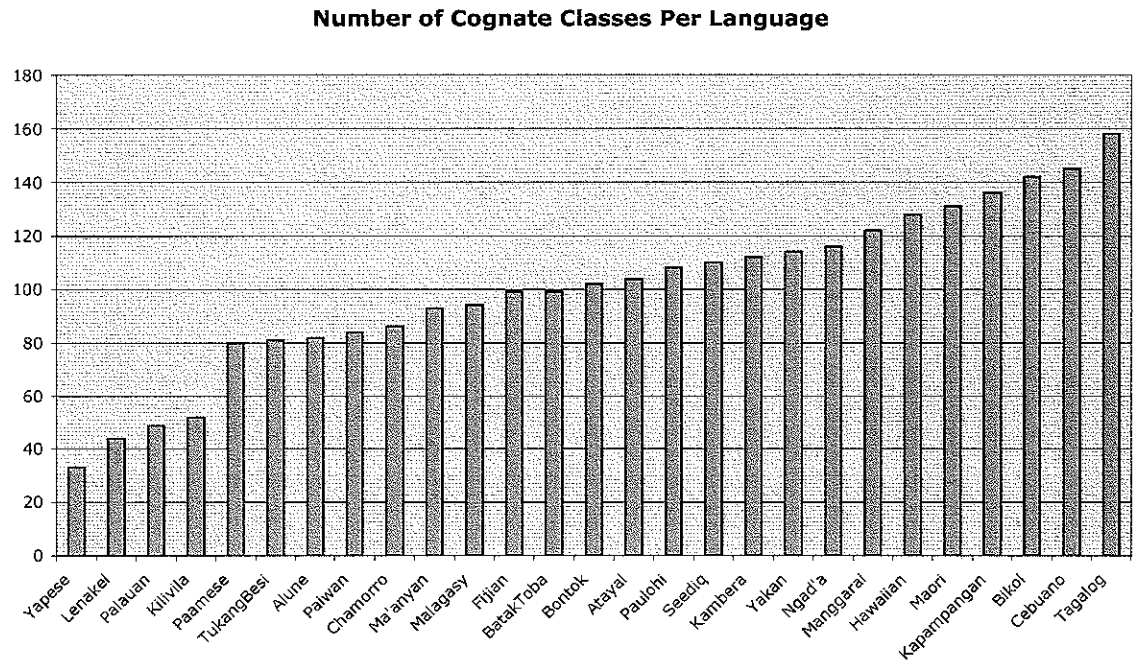


Figure 3. The number of cognate classes belonging to each language in the sample

References

Adelaar, N. 1991. New Ideas on the early history of Malagasy. In: Steinhauer. Ed. *Papers in Austronesian Linguistics No.1*. Pacific Linguistics, A-81.

Allard and Carpenter, 1996. On weighting and congruence. *Cladistics* **12**: 183-198.

Atkinson, Q., Nicholls, G., Welch, D. and Gray, R. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *TransPhilSoc*. **103**(2): 193-219.

Attesson, K.1999. The performance of Neighbor-Joining methods of phylogenetic reconstruction. *Algorithmica* **25**: 251-278.

Baker, R. and Gatesy, J. 2002. Is morphology still relevant? In *Molecular Systematics: Theory and Practice*. DeSalle, R., Giribet, G. and Wheeler, W. Eds. Switzerland: Birkhaüser Verlag.

Bauer, Winfred. 1993. *Maori*. London: Routledge

Berg, van den, R. 1991. Muna dialects and the Munic languages: towards a reconstruction: In: Harlow and Clark. Eds. *VICAL 2: Western Austronesian languages: Papers from the fifth International Conference on Austronesian Linguistics*. Auckland: Linguistic Society of New Zealand.

Bledsoe, A and Raikow, R. 1990. A quantitative assessment of congruence between molecular and nonmolecular estimates of phylogeny. *J. of Mol. Evol.* **30**: 247-259

Blust, R. 1978. Eastern Malayo-Polynesian: A sub-grouping argument. In: Wurm and Carrington. Eds. pp. 181-234. (Summary of the 130pp paper presented at the conference).

Blust, R. 1977. The Proto-Austronesian pronouns and Austronesian sub-grouping: a preliminary report. *WPLUH* **9**(2): 1-15.

Blust, R. 1987. The linguistic study of Indonesia. *Archipel.* **34**: 27-47.

Blust, R. 1990. Central and Central-Eastern Malayo-Polynesian (Paper presented at Conference on Moluccan Linguistics, University of Hawaii, March 1990).

Blust, R. 1998. Beyond the Austronesian homeland: the Austric hypothesis and its implications ofr archaeology. In: Goodenough, W. Ed. *Prehistoric Settlement of the Pacific*. Philadelphia: American Philosophical Society.

Bollback J. 2005. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. Software available from http://brahms.ucsd.edu/simmap.html. Verision 1.0 Beta 2.0.

Brainard, S. and Behrins, D. 2002. *A Grammar of Yakan.* Manilla: Linguistic Society of the Philippines.

Bryant, D., Filimon, F. and Gray, R. (in press). Untangling our past: Languages, trees, splits and networks. in R. Mace, C. Holden, S. Shennan. Eds. The Evolution of Cultural Diversity: Phylogenetic Approaches. UCL Press.

Bull, J., Huelsenbeck, J, Cunningham, C., Swofford, D., and Waddell, P. 1993. Partitioning and combining data in phylogenetic analyses. *Syst. Biol* 42: 384-397.

Cavalli-Sforza, L. and Feldman, M. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics Supplement.* **33**:266-275.

Cavalli-Sforza, L., Piazza, A., Menozzi, P. and Mountain, J. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Science USA* 85:6002-6.

Chippindale P. and Wiens, J. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. *Systematic Biology.* **43:** 278-287

Clark, R. 1985. Languages of north and central Vanuatu: groups, chains, clusters and waves. In: Pawley and Carrington. Eds. *Austronesian linguistics at the 15th Pacific Science Congress.* Pacific Linguistics, C-88. pp. 199-236.

Collins, J. 1983. The historical relationships of the languages of Central Maluku, Indonesia. Pacific Linguistics, D-47. Canberra: Australian National University Press.

Crowley, T. 1982. *The Paamese Language of Vanuatu*. Canberra: The Australian National University Press.

Cunningham, C.W. 1997. Can three incongruence tests predict when data should be combined? *Mol Biol Evol* **14**: 733-740.

Curnow, T. 2001. What Language Features Can be 'Borrowed'?. In: Aikhenvald, A. and Dixon, R. Eds. *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*. New York: Oxford University Press, 2001. pp. 412-435.

Dahl, O. 1977. La subdivision de la famille barito et la place du malgache. *Acta Orientalia* (Copenhagen) **38**: 77-134.

De Queiroz, A., Donoghue, M., and Kim, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics*. **26**: 657-681

Djawanai, S and Grimes, C. 1995. *Ngada*. In: Darryl Tryon. Ed. *Austronesian Comparative Dictionary*. New York: Mouton de Gruyter. pp. 593-99.

Dobson, A. 1969. Lexicostatistical grouping. *Anthropological Linguistics* **11**: 216-221.

Dunn, M., Terill, A., Reesink, G., Foley, R., and Levinson, S. 2005. Structural Phylogenetics and the reconstruction of ancient language history. *Science* **309**: 2072-5

Elbert, S. and Pukui, M. 1979. *Hawaiian Grammar*. Honolulu: University of Hawaii Press

Egerod, S.1980. *Atayal-English Dictionary*. London: Curzon Press.

Enfield, N. 2005. Areal Linguistics and Mainland Southeast Asia. *Annu.Rev.Anthropol* **34**:181-206.

Ethnologue Online. 2005. www.ethnologue.com

Farias, Izeni P., Orti, G., and Meyer, A. 2000. Total Evidence: Molecules, Morphology and the Phylogenetics of Cichlid Fishes.*J.Exp.Zoo* **288**:76-92.

Farris, J., Källersjö, M., Kluge, A. and Bult, C. 1995. Constructing a significance test for incongruence. *Syst Biol* 44: 570-572.

Farris, J., Källersjö, M., Kluge, A. and Bult, C. 1994. Testing significance of incongruence. *Cladistics* **10**:315-319.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Forman, Michael L. 1971. *Kapampangan Grammar Notes*. Honolulu: University of Hawaii Press.

Forster, P., Toth, A. and Bandelt, H. 1998. Evolutionary network analysis of word lists: Visualising the relationships between Alpine Romance languages. *J.Quant. Linguist.* **3**: 174-187.

Forster, P and Toth, A. 2003. Towards a Phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proc.Natl.Acad.Sci.* **100**: 9079-9084.

Foundation for Endangered Languages. www.ogmios.org/143.htm

Fukuda, Takashi. 1997. A Discourse-Oriented Grammar of Eastern Bontoc. *Studies in Philippine Linguistics*. **10**(1).

Global Chinese Language and Culture. http://edu.ocac.gov.tw/local/tour_aboriginal/english/a/07.htm

Gray, R. 2005. Pushing the Time Barrier in the Quest for Language Roots. *Science*. **309**: 2007-8.

Gray, R. and Atkinson, Q. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*. **426**: 435-439.

Gray, R. and Jordan, F. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature*. **405**: 1052-1055.

Gray, R. and Greenhill, S. (in press) Ch. 3: Testing Population Dispersal Hypotheses: Pacific Settlement, Phylogenetic Trees and Austronesian Languages.

Greenhill, S. and Gray, R. 2005. Austronesian Basic Vocabulary Database. http://language.psy.auckland.ac.nz/index.php.

Griffiths, Carole. 1999. Phylogeny of the Falconidae Inferred from Molecular and Morphological Data. *The Auk*. **116** (1): 116-130.

Grimes, C. 1990. Notes on Central Malayo-Polynesian (Mimeo)

Gudai, D. 1988. *A Grammar of Maanyan: A Language of Central Kalimantan*. Canberra: Australian National University.

Hallet, M. and Lagergren, J. 2001. Efficient algorithms for lateral gene transfer problems. In: *Proceedings of the 5th Ann. Int.Conf.Compt.Mol.Biol.(RECOMB 01)*. New York: ASM Press, pp. 149-156.

Hall, B. *Phylogenetic Trees Made Easy: A How-To Manual.* Sunderland, MA: Sinauer Inc., 2004.

Haspelmath, M., Dryer, M., and Comrie, B. Eds. 2005. WALS: World Atlas of Linguistic Structures. Oxford: Oxford University Press.

Hoenigswald, H. 1987. Language family tree, topological and metrical. In: Hoenigswald, Henry M. and Linda F. Wiener (eds.), *Bioloigcal metaphor and cladistic classification: an interdisciplinary perspective*, pp. 257-267. Philadelphia: University of Pennsylvania Press.

Holden, C. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proceedings of the Royal Society: Biological Sciences.* **269**(493): 793-799.

Holmer, A. 1996. A Parametric Grammar of Seediq. Lund: Lund University Press.

Hooker, B., Behrens, B., and Hartung P. 1975. *Papers in Philippine Linguistics No. 7.* Pacific Linguistics, Series A: 44

Hsu, R. 1969. *Phonology and Morphophonemics of Yapese.* Ann Arbor: University of Michigan Dissertation Services.

Huelsenbeck, J., Bull, J. and Cunningham, C. 1996. Combining data in phylogenetic analysis. *Trends in Ecology and Evolution.* **11**: 152-158.

Huelsenbeck, J., Nielsen, R., and Bollpack, J.2003. Stochastic Mapping of Morphological Characters. *Syst. Biol.* **52**(2): 131-158.

Huelsenbeck, J., Ronquist, F., Nielsen, R., and Bollback, J. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science.* **294**: 2310-2314

Huelsenbeck, J. and Ronquist, F. N.d. *MrBayes: Bayesian inference of phylogeny.* http://morphbank.ebc.uu.se/mrbayes/.

Huson, D. and Bryant, D. Application of Phylogenetic Networks in Evolutionary Studies, to appear in: *Molecular Biology and Evolution*, 2005.

Huson, D.1998. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics*, **14**(10): 68–73.

Inkelas, Sharon (forthcoming). Exceptional stress attracting suffixes in Turkish: Representations vs. the grammar. to appear in R. Kager, H. van der Hulst & W. Zonneveld (eds.), *The Prosody Morphology Interface*, Cambridge University Press.

Josephs, L. 1984. *Palauan Reference Grammar*. Honolulu: The University of Hawaii Press.

Klamer, Marian. 1998. *A Grammar of Kambera*. New York: Mouton de Gruyter

Kroeger, Paul. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. Stanford: CSLI Publications.

Larget, B. and Simon, D. 1999. Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Mol.Bio.Evol.* **16**(6): 750-9.

Larson, A. 1994. The comparison of morphological and molecular data in phylogenetic systematics. In: B. Schierwater, B. Streit, G.P. Wagner and R. DeSalle (eds.), *Molecular Ecology and Evolution: Approaches and Applications*. Basel: Birkhäuser, pp. 371-390.

Lawrence, J. and Ochman, H. 2002. Reconciling the Many Faces of Lateral Gene Transfer. *Trends in Microbiology*. **10** (1): 1-4.

Lawrence, J. and Hartl, D. 1992. Inference of horizontal genetic transfer from molecular data: An approach using the bootstrap. *Genetics* 131: 753-760.

Lewis, P. 2001. A likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Syst. Biol.* **50**(6):913-925.

Lohr, M. 1999. Methods for the genetic classification of languages. PhD thesis, University of Cambridge.

Losos, J. 1999. Uncertainty in the reconstruction of ancestral character states and limitations on the use of phylogenetic comparative methods. *Animal Behaviour*. **58**: 1319-1324.

Lutzoni, F. and Vilgalys, R. 1995. Integration of morphological and molecular data sets in estimating fungal phylogenies. *Canadian Journal of Botany*. **73** (Supplement 1): S649-659.

Lynch, J. 1978. *A Grammar of Lenakel*. Canberra: Pacific Linguistics

Lynch, J. and Tryon, D. 1985. Central-Eastern Oceanic: a subgrouping hypothesis. In: Pawley and Carrington. Eds. *Austronesian Linguistics at the 15th Pacific Science Congress*. Pacific Linguistics, C-88.

Maddison, W. 2000. Testing Character Correlation Using Pairwise Comparisons on Phylogeny. *J.Theor.Bio.* **202**: 195-204.

Makarenkov, V., Kevorkov, D. and Legendre, P. 2005. Phylogenetic Network Reconstruction Approaches, to appear in Applied Mycology and Biotechnology, v. 6, Genes, Genomics and Bioinformatics, Elsevier Science.

McMahon, R. 2004. Genes and Languages. *Community Genetics.* **7**:2-13.

McMahon, A. and McMahon, R. 2003. Finding Families: Quantitative methods in language classification. *Trans. Philol. Soc.* **101**: 7-55.

McMahon, A. and McMahon, R. Climbing down from the trees: Network representation for language families, in preparation.

Mintz, M. 1971a. Bikol Grammar Notes. Honolulu: University of Hawaii Press.

Mintz, M. 1971b. Bikol Dictionary. Honolulu: University of Hawaii Press.

Mintz, M. 1973. Case and Semantic Affixes of Bikol Verbs. Ann Arbor: University of Michigan Dissertation Services.

Nababan, P. 1981. *A Grammar of Toba-Batak.* Canberra: Pacific Linguistics.

Nakhleh, L. 2004. *Phylogenetic Networks in Biology and Historical Linguistics.* Ph.D. dissertation, The University of Texas at Austin.

Nakhleh, L., Ringe, D. and Warnow, T. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* **81**(2):382-420.

Nakhleh, L, Wanow, T., Ringe, D. and Evans, S. 2005. A comparison of phylogenetic reconstruction methods on an IE data set. *TransPhilSoc* **3**(2): 171-192.

Nei, M. and Kumar, S. *Molecular Evolution and Phylogenetics.* New York: Oxford University Press, 2000.

Nielsen, R. 2002. Mapping Mutations on Phylogenies. *Syst.Biol* **51**(5): 729-739.

Nylander, J., Ronquist, F., Huelsenbeck, J., and Nieves-Aldrey, J. 2004. Bayesian Phylogenetic Analysis of Combined Data. *SystBio* **53**(1): 47-67.

Ochmann, H., Lawrence, J., and Groisman, E. 2000. Lateral gene transfer among. genomes. *Nature* 405: 299-304.

Pagel, M., Meade, A. and Barker, D. 2004. Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst.Bio.***53**(5):673-684.

Pagel, M. 2000. New approaches to lexicostatistics and glottochronology. In: Renfrew, C., McMahon, A., Trask, L. Eds. *Time Depth in Historical Linguistics.* Cambridge: McDonald Institute for Archaelogical Research, pp. 189-207.

Pallesen, A. 1985. Culture contact and language convergence. Linguistic Society of the Philippines Monograph 24. 1977 PhD diss., University of California, Berkeley. Manila: SIL.

Pawley, A. and Ross, M. 1993. Austronesian Historical Linguistics and Culture History. *Ann.Rev.Anthr.* 22: 425-59.

Rasoloson, J. and Rubino, C. 2005. Malagasy. In: Himmelmann, N. and Adelaar, K. Eds. *The Austronesian Languages of Asia and Madagascar.*London: Curzon Press

Rau, Der-Hwa V.1992. *A Grammar of Atayal.* Ann Arbor: University of Michigan Dissertation Services.

Reid, L. 1976. *Bontok-English Dictionary.* Canberra: Pacific Linguistics

Reid, L. 1982. The demise of Proto-Philippines. In: Halim, Carrington, and Wurm. Eds. *Papers from the 3rd International Conference on Austronesian Linguistics, vol. 2: Tracking the travellers.* Pacific Linguistics, C-75.

Ringe, D., Warnow, T., and Taylor, A. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100: 59-129.

Ringe, D., Warnow, T., and Taylor, A Michailov, A. and Levison, L. 1998. Computational cladistics and the position of Tocharian. In: Mair, V. *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia*, pp. 391-414. Washington: Institute for the Study of Man.

Robinson, D. and Foulds, L. 1981. Comparison of Phylogenetic Trees. *Mathematical Biosciences* **53**: 131-147.

Ronquist, F. 2004. Bayesian Inference of Character Evolution. *Trends.Eco.Evol.* **19**(9):475-481.

Ronquist, F. 2005. MrBayes User Manual. http://mrbayes.csit.fsu.edu/manual.php.

Ross, M. 1988. Proto-Oceanic and the Austronesian languages of western Melanesia. Pacific Linguistics, C-98. Canberra: Australian National University Press.

Ross, M. 1995. Some Current Issues in Austronesian Linguistics. In: Tryon, D. Ed. *Austronesian Comparative Dictionary*. New York: Mouton de Gruyter. pp. 45-120.

Ross, M. Contact-Induced Change in Oceanic Languages in North-West Melanesia. In: Aikhenvald, A. and Dixon, R. Eds. *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*. New York: Oxford University Press, 2001. pp. 134-163.

Ruhlen, M. 1987. A guide to the world's languages. London: Stanford University Press.

Ruvolo, Maryellen. 1987. Reconstructing genetic and linguistic trees: phonetic and cladistic approaches. In: Hoenigswald, H. and Wiener, L. Eds. *Biological metaphor and cladistic classification: an interdisciplinary yperspective.*. Philadelphia: University of Pennsylvania Press.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic tree. *Mol. Biol. Evol.* **4**: 406-425.

Schachter, P. and Otanes, F. 1972. *Tagalog Reference Grammar*. Los Angeles: University of California Press

Schütz, A. 1985. *The Fijian Language*. Honolulu: University of Hawaii Press.

Semuin, A.1993. The basic grammar of Manggarai: Kempo Subdialect M.A. Thesis, La Trobe University Press

Senft, G. 1986. *Kilivila: The language of the Trobriand Islanders*. New York: Mouton de Gruyter.

Sokal, R., Oden, N. and Thomsen, B. 1988. Genetic changes across language boundaries in Europe. *American Journal of Physical Anthropology* **76**: 337-61.

Sokal, R., Oden, N. and Thomsen, B. 1992. Origins of the Indo-Europeans: genetic evidence. *Proceedings of the National Academy of Science USA* **89**: 7669-73.

Strauss, S. 1983. Stress assignment as morphological adjustment in English. *Linguistic Analysis* **11**, 419-427.

Swofford, D.L. *PAUP*: Phylogenetic Analysis under Parsimony (and Other Methods).* Version 4.0. Sinauer Associates, Sunderland, Mass.

Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic inference. In: Hillis, D.M., B.K. Mable, and C. Moritz (eds.), *Molecular Systematics*, pp. 407-514. Sunderland, Mass.: Sinauer Assoc.

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Topping, D. 1973. *Chamorro Reference Grammar*. Honolulu: The University of Hawaii Press.

Topping, D. 1975. *Chamrro-English Dictionary*. Honolulu: The University of Hawaii Press.

Tryon, D. 1995. Introduction. In: Tryon, D. Ed. *Austronesian Comparative Dictionary*. New York: Mouton de Gruyter. pp. 1-44.

Valkama, K. 2000. *Grammatical Relations in Cebuano*. Helsinki: University of Helsinki Publications.

Verheijen, J. and Grimes, C. 1995. *Manggarai*. In: Tryon, D. Ed. *Austronesian Comparative Dictionary*. New York: Mouton de Gruyter. pp. 585-95.

Warnow, T., Ringe, D. and Taylor, A. 1995. Reconstructing the evolutionary history of natural languages. *IRCS Report 95-16*. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania.

Warnow, T., Evans, S., Ringe, D. and Nakhleh, L. 2004. A stochastic model of language evolution and incorporates homoplasy and borrowing. In: *Phylogenetic Methods and the Prehistory of Languages*.

Warnow, T. 1997. Mathematical Approaches to Comparative Linguistics. *Proc. Natl.Acad.Sci.* **94**: 6585-6590.

Wiens, J. Ed. *Phylogenetic Analysis of Morphological Data*. Washington: Smithsonian Institution Press, 2000.

Wiens, John J. 1998. Combining Data Sets with Different Phylogenetic Histories. *Syst.Bio.* **47**(4): 568-581.

Wichmann, S. 2005. On the power-law distribution of language family sizes. *Journal of Linguistics* 41.2.

Wichmann, S. and Kamholz, D. unpublished. *Evaluating the strength of typological features for phylogenetic analyses*.

Wolff, J. 1966. *Beginning Cebuano*. Ann Arbor: UMI Books on Demand.

Other Websites

Explanation of Application of Bayesian Theorem to Phylogenetics:

www.egg.isu.edu/biocourses/bios599/projects/Walter_html