Title

From Spiders to Languages: a Phylogenetic Journey

Author

Chousou-Polydouri, Natalia

Publication Date

2014

Peer reviewed|Thesis/dissertation

From Spiders to Languages: a Phylogenetic Journey

by

Natalia Chousou-Polydouri

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Science, Policy, and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rosemary Gillespie, Chair
Adjunct Professor Charles Griswold
Associate Professor Kipling Will
Assistant Professor Lev Michael

Fall 2014

Abstract

From Spiders to Languages: a Phylogenetic Journey

by

Natalia Chousou-Polydouri

Doctor of Philosophy in Environmental Science, Policy, and Management

University of California, Berkeley

Professor Rosemary Gillespie, Chair

The parallels between biological and linguistic evolution have been recognized for a long time. In this work, contemporary phylogenetic methods are applied to address empirical questions in both fields. Additionally, the parallel histories and methodologies between biological systematics and historical linguistics are reviewed and novel approaches for coding linguistic data for contemporary phylogenetic methods are suggested.

Biogeography of Orsolobidae The biogeography of the spider family Orsolobidae was used as a case study to test the 'drowned' New Zealand hypothesis and to assess the importance of the breakup of Gondwana in its diversification. A molecular phylogeny of Orsolobidae and numerous outgroups was dated using both fossil node and substitution rate calibrations. The resulting divergence age estimates are consistent with the Gondwana breakup, while they reject the 'drowned' New Zealand hypothesis.

Linguistic characters and phylogenetics The history and methods of biological systematics and historical linguistics are briefly reviewed while underlining their similarities and differences. Novel coding methods are proposed for lexical, phonological and morphological linguistic characters and how they can be analyzed with computational phylogenetic methods.

Classification of Tupí-Guaraní A new internal classification is suggested for the Tupí-Guaraní language family based on lexical data analyzed with Bayesian phylogenetic methods. The dataset is coded using a novel cognate presence-absence method which includes items that have undergone semantic shift. The results largely agree with the lower subgroups of previous classifications, but suggest a significantly different higher structure of the family, with Kamaiurá sister to all other languages. Our results also suggest a Tupí-Guaraní homeland between the Xingu and Tocantins rivers and new hypotheses for the spread of the family in South America.

To Diamantis,
Without whom I wouldn't have even started a PhD, let alone finish it.

Contents

List of Figures

List of Tables

Acknowledgments

This work would not have been possible without the guidance and support of the members of my PhD committee, Rosemary Gillespie, Charles Griswold, Kipling Will, and Lev Michael, as well as without the financial support provided by a scholarship for graduate studies abroad from the Greek National Scholarship Foundation and by the PBI Oonopidae project. When I started my PhD, I had very little idea of what a journey it would be. I began to work on spider systematics and biogeography without any experience in spiders or phylogenetics and in the middle of my PhD I found myself learning linguistics and applying phylogenetic methods to the evolution of languages!

I would like to thank Rosie for being enthusiastic and positive through the ups and downs of my projects and for encouraging me to pursue my work in historical linguistics. Rosie taught me to think broadly and to follow my intuitions and was my first teacher on spider biology and biogeography. She also provided much needed practical advice throughout my years in graduate school.

I am grateful to Charles who introduced me to the fascinating world of spider evolution and phylogeny, and taught me everything I know about spider morphology. He also gave me the opportunity to share his experience observing and collecting spiders. Finally, I would like to thank him for the unforgettable moments we spent together in the field.

I always thought I didn't care about algorithms and philosophy, and yet Kip, along with Brent Mishler and David Lindberg, gained me with their passion about phylogenetic principles, methods, and history. To the point that one of the chapters of my PhD is a historical review and comparison of phylogenetic thought in biological systematics and historical linguistics! I would like to especially thank Kip for his always helpful advice and the stimulating discussions we had on the philosophy and application of phylogenetic methods.

When I became interested in historical linguistics, I could not have met a better mentor than Lev. Not only he had confidence in my ideas, but he involved me in his team as an equal member and encouraged my participation in all aspects of our project. He gave me the opportunity to deepen my knowledge of linguistics and his support was decisive for me to venture into this field for my postdoc.

Apart from my committee, I would like to thank many other people that made my PhD possible and my time in Berkeley interesting, fruitful and fun. Sarah Crews, Joel Ledford, and Darrell Ubick for teaching me spider identification. Patrick Kirch, for sparking my interest in historical linguistics through his excellent class on South Pacific Archaeology. Will Chang, my first linguistics teacher, for being patient and insistent. Peter Oboyski for accompanying me to much challenging fieldwork and his invaluable advice. Vivian Wauters and Zach O'Hagan for the endless work hours they shared with me and their patient explanations in all my questions. Athena Lam, Anna Sellas and Anthea Carmichael for teaching me molecular techniques and how to not despair when they fail.

I would like to thank all the members of the evolab, past and present, for sharing the good and bad times, the frustrations, failures and successes, for creating such a supportive and

comradely atmosphere. Also, to my friends and family for keeping me sane and motivated through the difficult times.

Last but not least, I am grateful to my husband, Diamantis Sellis, for his advice and ideas, for his help and support, for loving and trusting me. I would also like to thank my daughter, Leila, for keeping things in perspective during the last year of my PhD, and Diamantis, Rosa, Nora, Artemis, and Kostas for being with her when I couldn't.

Chapter 1

The biogeography of the spider family Orsolobidae

1.1 Introduction

The biogeography of the southern continents has been at the center of the debate between the relative importance of vicariance and long distance dispersal for a long time (Platnick and Nelson, 1978; Queiroz, 2005; Rosen, 1978).[1] South America, Africa, Australia, India, Madagascar, and Antarctica were all connected in the supercontinent Gondwana from 500Ma (with Gondwana's formation) until 170Ma (when Gondwana started breaking up) (Blakey, 2008). The formation of Gondwana coincided with the start of the Phanerozoic, and thus its subsequent breakup has had a profound influence on the biogeography of land plants and animals (see Sanmartín and Ronquist (2004) and references therein). However, extinction, long distance dispersal and incomplete fossil records can obscure or erase previous vicariant biogeographical patterns (Upchurch, 2008), while in some cases, it seems that the divergence events are simply too recent to be attributable to the Gondwanan breakup (Cook and Crisp, 2005; Raxworthy, Forstner, and Nussbaum, 2002; Renner, 2004).

Another, more recent, debate has been around a smaller fragment of Gondwana with a unique flora and fauna, New Zealand. It has been long regarded that New Zealand has a rare assemblage of animals and plants, relicts of its Gondwanan past. However, Campbell and Landis (2003) questioned the continuous emergence of New Zealand, suggesting that it could have been completely submerged during the Oligocene (the 'drowned' New Zealand hypothesis). As a result, all modern flora and fauna on the islands must be explained by long distance dispersal in the last 22Ma (Waters and Craw, 2006).

In this study, we present evidence for these biogeographic questions from a family of small spiders (Araneae, Orsolobidae), which has a Gondwanan distribution and its greatest taxonomic diversity on New Zealand (Platnick 2014). A molecular phylogenetic analysis

[1] This chapter is based on an unpublished manuscript by Natalia Chousou-Polydouri, Anthea Carmichael, Tamás Szüts, Alma Saucedo, Rosemary Gillespie, and Charles Griswold.

of Orsolobidae calibrated using fossils and molecular substitution rates yields a topology and divergence time estimates consistent with the Gondwanan breakup, while rejecting the 'drowned' New Zealand hypothesis. Our analysis also is the first molecular phylogeny regarding relationships within the family Orsolobidae and the superfamily Dysderoidea.

The breakup of Gondwana

The southern supercontinent Gondwana started to break up around 170Ma. The timeline of the Gondwana breakup in terms of plate tectonics is shown in Figure 1.1 (Blakey, 2008). The pattern of separation of the Gondwana continents does not provide an idealized area cladogram. Each rifting event takes tens of millions of years to be completed and many of them overlap. Also, not all organisms are affected the same way or at the same time by continental drift. To complicate things even further, there is the possibility for the existence of various land bridges connecting pieces of Gondwana up to 90Ma (Sereno, Wilson, and Conrad, 2004). More specifically, there are proposed landbridges connecting South America to Africa and to Antarctica, as well as Antarctica to India+Madagascar (see Figure 1.2). It is not surprising therefore that there are a few predominant tree topologies linked with the Gondwanan breakup. Two typical area cladograms for Gondwanan taxa are shown in Figure 1.3. One scenario has Africa separating first (Africa-first scenario), while the other has Africa and South America as sisters (West Gondwana scenario) (Upchurch, 2008).

The 'drowned' New Zealand hypothesis

The continental plate that includes New Zealand rifted away from Australia with the opening of the Tasman sea between 85 and 55Ma (Li and Powell, 2001). Much of the continental plate is now submerged. Campbell and Landis (2003) and Landis et al. (2008) have suggested that there is no geological evidence for the continuous presence of land, so it is possible that no part of New Zealand was above sea level during the Oligocene. At the same time though, Landis et al. (2008) admit that there is no direct geological evidence for complete submergence. The Oligocene 'drowning' of New Zealand had a profound impact among biologists (Waters and Craw, 2006), with studies presenting evidence for (Baker et al., 2005; Pole, 1994) or against it (Cooper et al., 2001; Ericson et al., 2002; Knapp et al., 2007; Lee, Bannister, and Lindqvist, 2007).

The family Orsolobidae

Orsolobidae are small haplogyne spiders primarily living in the leaf litter (Forster and Platnick, 1985). The current members of the family Orsolobidae had been previously included in the families Dysderidae and Oonopidae. Forster and Platnick (1985) elevated Orsolobids to the family rank, included in them genera previously placed in other families and described numerous new genera and species. Since their revision of Orsolobidae, new genera and species have been described (Baehr and Smith, 2008; Griswold and Platnick, 1987; Izquierdo and

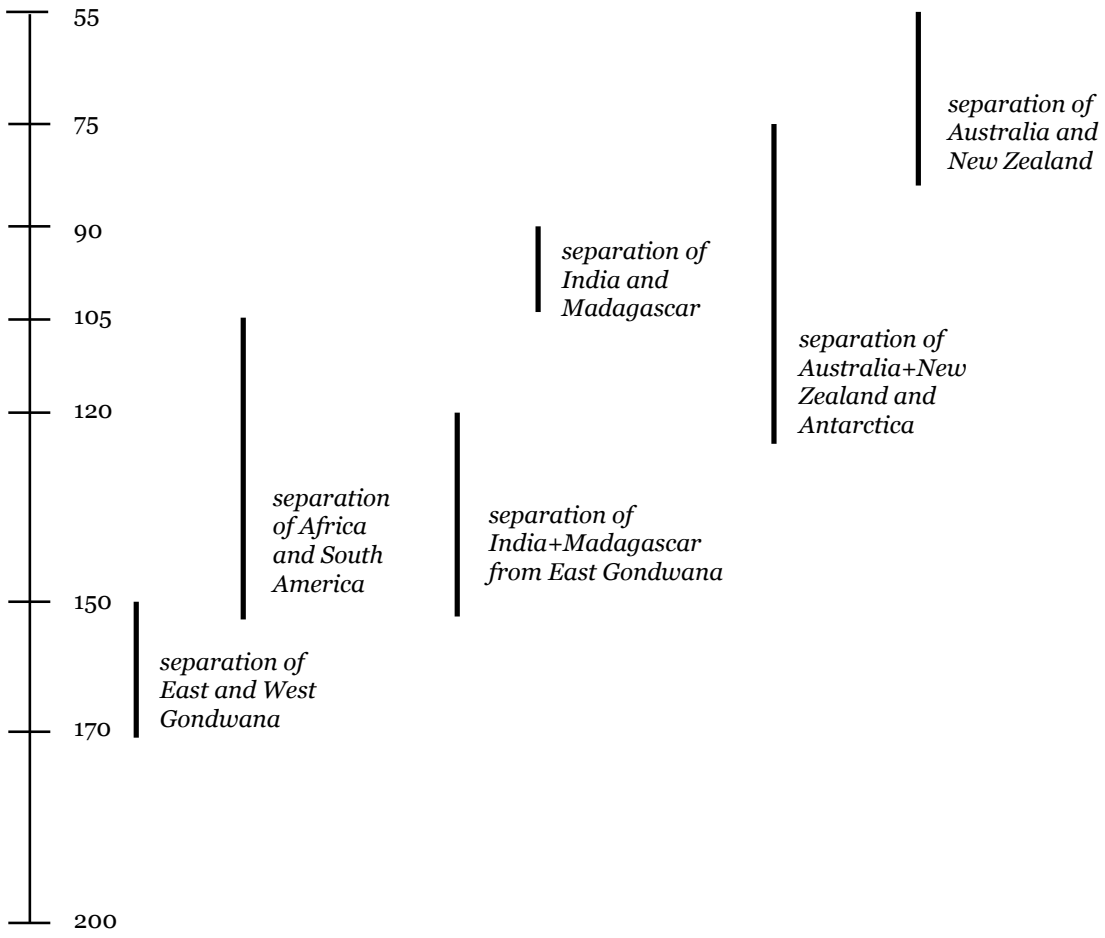

Figure 1.1: Timeline for the breakup of Gondwana based on Blakey (2008) and Li and Powell (2001). The vertical axis is in million years before present. Vertical lines represent the duration of each rifting event resulting in the complete separation of the continental fragments involved.

Labarque, 2010; Lise and Almeida, 2006; Ott and Lise, 2004; Ott, Platnick, et al., 2013; Platnick and Brescovit, 1994) and currently the family includes 30 genera and 187 species (Platnick, 2014). All Orsolobidae are united by a unique elevated tarsal organ (Forster and Platnick, 1985). The family has a Gondwanan distribution and is present in South America, Southern Africa, New Zealand and Australia. It is remarkable that New Zealand and nearby

Figure 1.2: Proposed land bridges between Gondwanan fragments at about 90Ma. The landbridge between Africa and South America may have ceased to exist 30 to 50Ma earlier. Adapted from Sereno, Wilson, and Conrad (2004).

islands harbor more than half of the generic and specific diversity of Orsolobidae.

The monophyly of Orsolobidae genera has not been tested and phylogenetic relationships among Orsolobidae genera are largely unknown. Forster and Platnick (1985) speculated

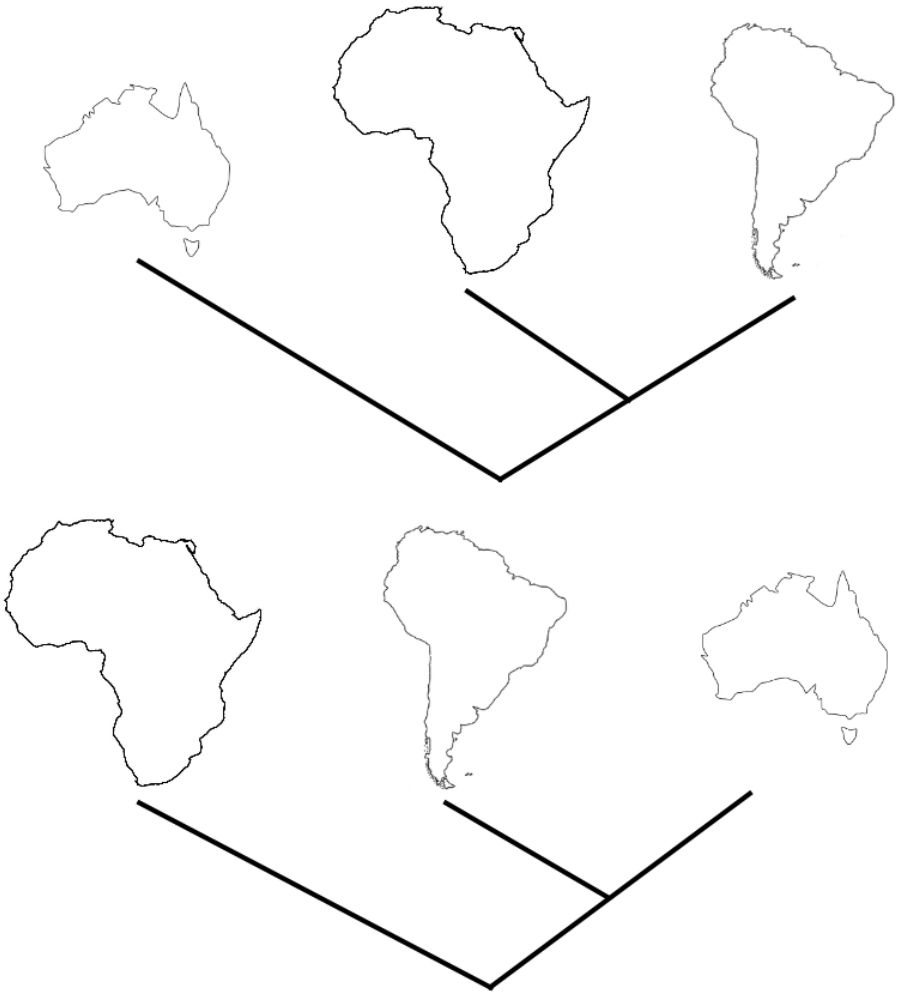

Figure 1.3: Common area cladograms of Gondwanan taxa (adapted from Upchurch (2008)). Australia represents areas of East Gondwana (Australia, New Zealand, Madagascar). West Gondwana scenario above and Africa-first scenario below.

about Orsolobidae relationships noting at the same time that different characters imply conflicting groupings. Although the taxon sampling of the present study is not designed to resolve the relationships among Orsolobidae genera, or even to confirm their monophyly, our results provide the first test of these hypotheses.

The superfamily Dysderoidea

After Forster's and Platnick's revision of Orsolobidae, there has been consensus that the superfamily Dysderoidea includes 4 families: Dysderidae, Segestriidae, Orsolobidae and Oonopidae (Forster and Platnick, 1985; Platnick, Coddington, et al., 1991; Ramírez, 2000). However, the relationships within the superfamily remain unclear with different characters suggesting different affinities. One of the most widespread hypotheses is that Orsolobidae and Oonopidae are sister taxa, based on the presence of tarsal proprioreceptor bristles and bipectinate tarsal claws (Forster and Platnick, 1985; Platnick, Coddington, et al., 1991; Ramírez, 2000). Lipke, Ramírez, and Michalik (2014) suggested that Orsolobidae and Dysderidae may be sister taxa based on spermatozoic characters.

The sister group to Dysderoidea is not well established either. Recently, the family Caponiidae has been proposed as sister to Dysderoidea (Ramírez, 2000), while Griswold, Audisio, and Ledford (2012) suggested the newly discovered family Trogloraptoridae. Previously hypothesized members of Dysderoidea, such as Tetrablemmidae, Telemidae and Leptonetidae, are also possible candidates (Forster and Platnick, 1985). In our analysis we have included representatives of most of these families in an effort to find the closest relatives of Dysderoidea using molecular data.

1.2 Materials and Methods

Taxonomic Sampling

A total of 99 specimens were included in the present study. Of these, 61 are in the family Orsolobidae, 25 in the other families of the superfamily Dysderoidea (18 Oonopidae, 4 Dysderidae, and 3 Segestriidae) and 13 are outgroup taxa belonging to other haplogyne families (Caponiidae, Filistatidae, Leptonetidae, Scytodidae, Tetrablemmidae, Trogloraptoridae). We rooted the tree between *Hypochilus* (Hypochilidae) and all other OTUs (Forster, Platnick, and Gray, 1987; Platnick, 1977). All the taxa included in the study, along with their voucher information are shown in Tables 1.1 and 1.2.

Within the family Orsolobidae, our specimens represent all 4 landmasses where Orsolobids have been found (South America, Africa, New Zealand and Australia) and cover approximately half of the generic diversity of the family. More specifically, we included representatives of 2 African genera (out of 3), 2 Australian genera (out of 4), 4 South American genera (out of 7) and 6 New Zealand genera (out of 16) (Platnick, 2014). All included Orsolobidae genera were represented by 2-11 individuals. Many of the genera not represented are monotypic and/or known from very few specimens. To our knowledge, we have included almost every specimen preserved in a way permitting DNA extraction and amplification.

[2]Sequences from Bidegaray-Batista and Arnedo (2011).

[3]Sequences from Wood, Griswold, and Gillespie (2012).

[4]This OTU is chimaeric. 18S and 28S from *H. pococki* (Wheeler and Hayashi, 1998). Actin from *H. thorelli* (Vink et al., 2008).

OTU ID	Genus/Species	Family	Voucher
9023772_Caponia	*Caponia*	Caponiidae	9023772
9024987_Caponia	*Caponia*	Caponiidae	9024987
9035527_Calponia	*Calponia*	Caponiidae	9035527
9037955_Dcrocata	*Dysdera crocata*	Dysderidae	9037955
Dgomerensis	*Dysdera gomerensis*[2]	Dysderidae	crba001393
Hapennicola	*Harpactocrates apennicola*[2]	Dysderidae	crba000715
Htroglophilus	*Harpactocrates troglophilus*[2]	Dysderidae	crba000992
79_Kukulcania	*Kukulcania hibernalis* [3]	Filistatidae	9034219
Hypochilus	*Hypochilus* [4]	Hypochilidae	
9023925_Archoleptoneta	*Archoleptoneta*	Leptonetidae	9023925
9024327_OonoGen1	*cf. Trilacuna*	Oonopidae	9024327
9024968_OonoGen2	gen. nov.	Oonopidae	9024968
9024170_Opopaea	*Opopaea*	Oonopidae	9024170
9024216a_Opopaea	*Opopaea simoni*	Oonopidae	9024216
9022325_Trilacuna	*Trilacuna*	Oonopidae	9022325
9024161_Noideatella	*Noideatella*	Oonopidae	9024161
9024169_OonoGen3	gen. nov.	Oonopidae	9024169
9024216b_Farqua	*Farqua*	Oonopidae	9024216
9031354_Orchestina	*Orchestina*	Oonopidae	9031354
9035059_Ischnothyreus	*Ischnothyreus*	Oonopidae	9035059
9023800_Scytodes	*Scytodes socialis*	Scytodidae	9023800
9029852_Scytodes	*Scytodes socialis*	Scytodidae	9029852
9037957_Segestria	*Segestria*	Segestriidae	9037957
9027749_Ariadna	*Ariadna*	Segestriidae	9027749
9019233_Tetrablemma	*Tetrablemma thamin*	Tetrablemmidae	9019233
9029929_Paculla	*Paculla*	Tetrablemmidae	9029929
9040069_Trogloraptor	*Trogloraptor*	Trogloraptoridae	9040069
9040051_Trogloraptor	*Trogloraptor*	Trogloraptoridae	9040051

Table 1.1: Outgroup taxa (including Dysderoidea) included in the study. Voucher numbers are CASENT (California Academy of Sciences) numbers unless otherwise noted.

OTU ID	Genus/Species	Voucher	Landmass
9037121_Afrilobus	*Afrilobus* sp.	9037121	Africa
9037034_Afrilobus?	*cf. Afrilobus*	9037034	Africa
9031013_Ascuta	*Ascuta parornata*	9031013	New Zealand
9024171_Azanialobus	*Azanialobus* sp.	9024171	Africa
9026037_Azanialobus	*Azanialobus* sp.	9026037	Africa
9037119_Azanialobus	*Azanialobus* sp.	9037119	Africa
36006_Duripelta	*Duripelta* sp.	OONO36006	New Zealand
9023587_Hickmanolobus?	*cf. Hickmanolobus*	9023587	Australia
9035010_Hickmanolobus?	*cf. Hickmanolobus*	9035010	Australia
9027539M_Losdolobus	*Losdolobus* sp.	9027539M	South America
9027538_Losdolobus	*Losdolobus* sp.	9027538	South America
9027681_Mallecolobus	*Mallecolobus* sp.	9027681	South America
9023469_Mallecolobus	*Mallecolobus maullin*	9023469	South America
9023473_Mallecolobus?	*cf. Mallecolobus*	9023473	South America
9031019_Maoriata	*Maoriata magna*	9031019	New Zealand
9037958_Maoriata	*Maoriata magna*	9037958	New Zealand
9035070_Orsolobus	*Orsolobus* sp.	9035070	South America
9027819_Orsolobus	*Orsolobus* sp.	9027819	South America
9035068_Orsolobus	*Orsolobus*	9035068	South America
9035284_Orsolobus	*Orsolobus*	9035284	South America
9035122_Osornolobus	*Osornolobus* sp.	9035122	South America
9028179_Osornolobus	*Osornolobus thayerae*	9028179	South America
9035123_Osornolobus	*Osornolobus* sp.	9035123	South America
9035124_Osornolobus	*Osornolobus* sp.	9035124	South America
9031020_Pounamuella	*Pounamuella complexa*	9031020	New Zealand
9031030_Pounamuella	*Pounamuella complexa*	9031030	New Zealand
9035065_Subantarctia	*Subantarctia fiordensis*	9035065	New Zealand
9035061_Subantarctia	*Subantarctia* sp.	9035061	New Zealand
36007_Tasmanoonops	*Tasmanoonops daviesae*	OONO36007	Australia
9035003_Tasmanoonops	*Tasmanoonops* sp.	9035003	Australia
9035008_Tasmanoonops	*Tasmanoonops* sp.	9035008	Australia
9023532_Tasmanoonops	*Tasmanoonops cf. ripus*	9023532	Australia
9031022_Subantarctia	*Subantarctia fiordensis*	9031022	New Zealand
9035028_Tasmanoonops	*Tasmanoonops*	9035028	Australia
9037956_Wiltonia	*Wiltonia rotoiti*	9037956	New Zealand
9037960_Wiltonia	*Wiltonia rotoiti*	9037960	New Zealand
9028181_OrsoGen1	gen. nov.	9028181	South America
9027772_Orsolobidae	gen. nov.	9027772	South America

Table 1.2: Orsolobidae included in the study (not including pruned sequences, see Table 1.4).

Specimens were collected by the authors and kindly provided by Hannah Wood and other collaborators. They were preserved in 95% ethanol and vouchers stored at -20°C at the California Academy of Sciences.

DNA extraction, PCR amplification and sequencing

DNA extractions were done from tissue of 4 legs with a modified DNeasy Quiagen kit protocol and extracted DNA is stored at -80°C at the California Academy of Sciences. We amplified with PCR partial fragments of the mitochondrial genes cytochrome c oxidase subunit I (CO1), 16S rRNA (16S) and of the nuclear genes 28S rRNA (28S), 18S rRNA (18S), Histone H3 (H3) and Actin (Act). The sequences of the primers used and the PCR conditions for each gene are shown in Table 1.2.

Amplifications were performed according to the PCR primary conditions and reaction protocols shown in Table 1.2. In cases where the primary conditions were not successful, we used the alternative conditions shown in parentheses in Table 1.2 instead.

PCR products were purified using ExoSAP-IT and cycle-sequenced for both directions using one of the PCR primers and BigDye with an ABI automated sequencer at the California Academy of Sciences Center for Comparative Genomics. DNA sequences were edited using Sequencher 4.7 and Geneious 6.1.6.

For some genes, multiple primer pairs were used in order to obtain the full fragment. Also, in some cases, we were able to obtain the sequence of only part of the initially targeted fragment.

Gene	Primers	PCR conditions	Reaction protocol (in µl)
CO1	**Forward:** 1628: 5'-ATAATGTAATTGTTACTGCTCATGC-3' (Vandergast, Gillespie, and Roderick, 2004) 1490OONO: 5'-CWACAAAYCATARRGATATTGG-3' (mod. from Folmer et al. (1994)) **Reverse:** 2198: 5'-TAAACTTCAGGGTGACCAAAAAATCA-3' (Folmer et al., 1994) 2191: 5'-CCCGGTAAAATTAAAATATAAACTTC-3' (Simon et al., 1994)	95 2min 95 30s **50 (46)** 30s 72 1min 72 7min 35 cycles	Buffer 5 $MgCl_2$ 2.4 (3) Primer 1.3 dNTPs 0.42 BSA 1 Taq 0.13 (0.25) DNA 2
H3	**Forward:** H3nF: 5'-ATGGCTCGTACCAAGCAGAC-3' (Colgan et al., 1998) H3a / H3aF: 5'-ATGGCTCGTACCAAGCAGACVGC-3' (Colgan et al., 1998) **Reverse:** H3nR: 5'-ATRTCCTTGGGCATGATTGTTAC-3' H3b / H3aR: 5'-ATATCCTTRGGCATRATRGTGAC-3' (Colgan et al., 1998)	95 2min 95 30s **52 (50)** 30s 72 30s 72 7min 35 (45) cycles	Buffer 5 $MgCl_2$ 2.4 (3) Primer 1.3 dNTPs 0.42 BSA 1 Taq 0.13 DNA 2
16S	whole fragment **Forward:** N1-J-12261mod: 5'-TCATAWGARATYATTTGGGC-3' **Reverse:** LR-N-13398mod: 5'-TGACTGTTTAYCAAAAACAT-3' A fragment **Forward:** N1-J-12261mod: 5'-TCATAWGARATYATTTGGGC-3' **Reverse:** LR-N-13300: 5'-TGTRCTAAGGTAGCATAATCAATTG-3' B fragment **Forward:** N1-J-12866mod: 5'-ACCGGTCTGAACTCAAATCATGT-3' **Reverse:** LR-N-13398mod: 5'-TGACTGTTTAYCAAAAACAT-3' C fragment **Forward:** N1-J-12261mod: 5'-TCATAWGARATYATTTGGGC-3' **Reverse:** LR-N-12866mod: 5'-ACATGATTTGAGTTCAGACCGGT-3'	95 2min 95 30s **46 (43)** 45s 72 1min 30s 72 10min 35 (45) cycles	Buffer 5 $MgCl_2$ 2.4 (3, 5) Primer 1.3 dNTPs 0.42 Betaine 4 BSA 1 Taq 0.13 (0.25, 0.35) DNA 2

Act	**Forward:** ActinF1: 5'-GTCGCCCTGGACTTCGAGCA-3' (this study) ActinF2: 5'-CGCCCTGGACTTCGAGCAGG-3' (this study) **Reverse:** ActinR: 5'-TCCACATCTGCTGGAAGGTGGACA-3' (this study)	95 2min 95 30s **55 (57)** 30s 72 1min 30s 72 7min 35 cycles	Buffer 5 $MgCl_2$ 2.4 Primer 1.3 dNTPs 0.42 BSA 2 Taq 0.13 DNA 1
28S	**Foward:** 28S0cs: 5'-CGTGAAACTGCTCAGAGG-3' (Hedin and Maddison, 2001) L0264_Coelshort: 5'-CGGGTTGCTTGGGAGTGC-3' (Maddison, Moore, et al., 2009) ITS2Spr: 5'-CCCGCTGAATTTAAGCATAT-3' 28SmidF: 5'-CTGGCGGCGAGTAGGTCG-3' 28SA: 5'-GACCCGTCTTGAAACACGGA-3' (Whiting et al., 1997) **Reverse:** 28Sc: 5'-GGTTCGATTAGTCTTTCGCC-3' (Hedin and Maddison, 2001) 28SR: 5'-CCGTGTTTCAAGACGGGTCG-3' (mod. from Whiting et al. (1997) 28SB: 5'-TAGTAGCTGGTTCCTTCCGA-3' 28midrev: 5'-ACTCGCGCACATGTTAGAC-3'	95 2min 95 30s **50 (46)** 45s 72 2min 72 10min 35 cycles	Buffer 5 $MgCl_2$ 2.4 (3) Primer 1 dNTPs 0.42 BSA 2.5 Taq 0.13 (0.25) DNA 2

18S	**Forward:** 18S_1F: 5'-TACCTGGTTGATCCTGCCAGTAG-3' (Giribet et al., 1996) 18S_5F: 5'-GCGAAAGCATTTGCCAAGAA-3' (Giribet et al., 1996) 18S3Fl: 5'- GTTCGATTCCGGAGAGGGAGC-3' (mod. from Giribet et al. (1996)) 18S_5_9_intF: 5'- ATTCCGWTAACGADCGAG-3' (Miller et al., 2010) **Reverse:** 18S_5-9intR: 5'-CTCGHTCGTTAWCGGAAT-3' (Miller et al., 2010) 18S_9R: 5'-GATCCTTCCGCAGGTTCACCTAC-3' (Giribet et al., 1996) 18S_5R: 5'-CTTGGCAAATGCTTTCGC-3' (Giribet et al., 1996) 18S3Rs: 5'-GCTCCCTCTCCGGAATCGAAC-3' (mod. from Giribet et al. (1996)	95 2min 95 30s **50 (46)** 45s 72 2min 72 10min 35 cycles	Buffer 5 $MgCl_2$ 2.4 (3) Primer 1.3 dNTPs 0.42 BSA 2 Taq 0.13 (0.25) DNA 2

Table 1.3: Primers and PCR conditions and protocols. Reagent concentrations: $MgCl_2$ 25mM, primers 10µM, BSA 10mg/ml, GO Taq 5u/µl GO Taq, green buffer 5x, dNTPs 10mM, Betaine 5M. All temperatures are in °C.

Alignment

The alignment of the protein coding genes CO1, H3, and Act was done manually in Geneious using the aminoacid translation as a guide. For the ribosomal genes, we used MAFFT v7.017 (Katoh and Standley, 2013) as implemented in Geneious6 (`http://www.geneious.com/`) using the E-INS-i strategy with default options. The alignment of 18S had very few and short gaps making their placement obvious, while 28S and 16S had more and longer gaps. Gappy regions of 28S and 16S were manually removed to create a 'non-gappy' alignment in order to compare the effect of these regions on the phylogenetic analyses. We also performed tests of substitution saturation (Xia and Lemey, 2009; Xia, Xie, et al., 2003) for the third codon position of protein coding genes. Gene matrices were concatenated using Mesquite v. 2.75 (Maddison and Maddison, 2007). Non-sequenced fragments were scored as missing data. Two different concatenated alignments (with and without 16S) were created in order to compare the effects of its inclusion on the phylogenetic results, as we were unable to obtain 16S sequences from two thirds of our specimens.

Phylogenetic Analyses

Model choice and partitioning

Model choice and partitioning scheme was done with PartitionFinder (Lanfear, Calcott, Ho, et al., 2012; Lanfear, Calcott, Kainer, et al., 2014) using the BIC criterion. We did not use evolutionary models combining proportion of invariable sites and Gamma-distributed rate variation, as the two parameters are strongly correlated and cannot be optimized independently of each other (Yang, 2006). Also, we included the JC model as an option when running PartitionFinder to choose models for BEAST runs. We followed PartitionFinder's recommendation in terms of partitioning scheme and model choice in all phylogenetic analyses, except where otherwise noted.

Preliminary analyses - MrBayes

We performed preliminary analyses with MrBayes3.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) of the full and the 'non-gappy' alignments of 28S and 16S, as well as the concatenated alignments with or without 16S in order to compare the results. All analyses were done with default priors and 4 MCMC chains of 10 million generations each sampling every 1000 generations and discarding the initial 25% of the samples as burn-in.

Dataset Pruning

In our dataset, we had groups of 2-11 specimens for each genus, and many of these individuals were conspecific and/or turned out to have almost identical sequences. We included them in the preliminary analyses as some of them were only identified to the level of genus or belonged to yet undescribed genera or species in order to verify our tentative identifications. However,

in order to use BEAST for divergence time estimation, our taxon sampling presented a problem regarding tree prior choice. While for most of our terminals a speciation prior would be appropriate, there were groups corresponding to population level sampling, which would better fit a coalescent prior. Therefore, we removed from the dataset individuals with almost identical sequences (pruned dataset), so we could use a speciation prior. The selection of the specimen to be included from each group of almost identical sequences was based on maximizing available data (number of genes covered).

BEAST analyses

All analyses were performed with BEAST1.8 (Drummond, Suchard, et al., 2012; Drummond and Rambaut, 2007) at the California Academy of Sciences CCG PhyloCluster. For each analysis we ran 4 independent chains of 100 million generations each, sampling every 10000 generations. Stationarity, convergence and burn-in were assessed with Tracer v1.6.0 (Rambaut et al., 2013) and AWTY (Nylander et al., 2008).

Models and Priors For the BEAST analysis, we simplified the models of evolution selected by PartitionFinder in some cases, in order to achieve good mixing for the MCMC chains (see §BEAST results). In order to choose an appropriate prior for the tree and for the molecular clock, we performed iterative comparison with Bayes Factors, using the AICM procedure (Baele et al., 2012) as implemented in Tracer1.6. First, we compared runs with a strict, uncorrelated lognormal and exponential clock using a Yule tree prior for all of them. Then, we compared runs with a Yule, birth-death and birth-death with incomplete sampling tree prior using a lognormal clock for all of them.

Divergence times estimation

Orsolobidae unfortunately do not have a known fossil record. So, in order to estimate divergence times we followed two alternative approaches: using gene-specific substitution rates from the literature and using fossil node calibrations from outgroup taxa.

Rate calibration For the rate calibrated phylogenenies, we used the 28S substitution rate from Bidegaray-Batista and Arnedo (2011), who used a combination of fossil and biogeographical calibrations to estimate substitution rates for spiders in the family Dysderidae. This study represents the most rigorously estimated substitution rates for spiders closely related to Orsolobidae, as Dysderidae are also in the superfamily Dysderoidea. Unfortunately, their CO1 rate estimate could not be used because we excluded the third codon position (due to saturation, see §Saturation Tests), thus rendering CO1 rates across the two studies non-comparable.

We used the lognormal clock estimates as prior distributions for our overall lognormal clock. The prior for the mean value of the lognormal distribution (ucld.mean) was set as a normal distribution with mean=0.0011 and standard deviation=0.0003 and for the

standard deviation of the lognormal distribution (ucld.stdev) as a normal distribution with mean=1.249 and standard deviation=0.18 (Bidegaray-Batista and Arnedo, 2011). Since these were the priors for the overall clock rate, the relative rate of 28S in our analysis was fixed to 1, while all other relative rates for each gene were estimated.

Fossil calibrations For the fossil calibrated phylogenies, we used the oldest known fossils from two Dysderoid families, Segestriidae and Oonopidae, since they were the closest ones to the ingroup. The family Dysderidae also has a fossil record, but it is much more recent. The oldest fossils for both Segestriidae and Oonopidae are found in Lebanese amber of Valanginian age (120-130Ma) (Penney, 2000). The priors were gamma distributions with shape=69, scale=2 in [120, $+\infty$) (see Figure 1.4). We used the fossil dates to calibrate the most recent common ancestor of Segestriidae and Oonopidae respectively (crown calibration) or its parent node (stem calibration), in order to compare the estimates.

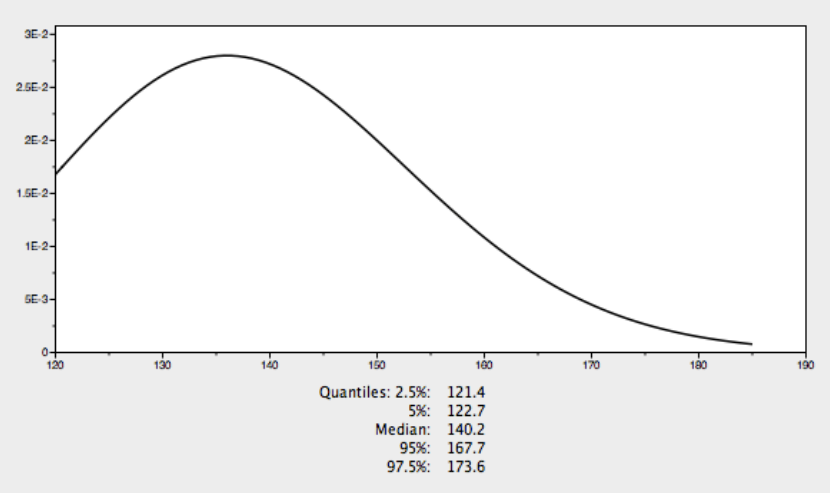

Figure 1.4: Prior distribution for the node calibration of Oonopidae and Segestriidae.

1.3 Results

Saturation Tests

The saturation tests indicated that the third codon position for CO1 was completely saturated, so we excluded this position from all subsequent analyses. All other third codon positions were not substantially saturated, so they were included.

Full dataset results

Majority-rule consensus trees resulting from the default alignments of 28S and 16S were compatible with the trees resulting from the 'non-gappy' alignments: the same nodes were present and well-supported (above 90% posterior probability), while the default alignments resulted in additional well-supported nodes. Therefore, the default alignments of 28S and 18S were used for all subsequent analyses. The majority-rule consensus tree from the concatenated alignment of the full dataset is shown in Figure 1.5. All families were recovered as monophyletic in the full analysis except Tetrablemmidae. The superfamily Dysderoidea was recovered with very high support and its internal structure was also very highly supported with Segestriidae sister to all other families and Orsolobidae sister to a Dysderidae-Oonopidae clade. As for Orsolobidae genera, most of them were monophyletic, while others polyphyletic or paraphyletic (see Discussion).

Dataset Pruning

In order to be able to use a speciation prior for our BEAST analyses, we removed identical or near-identical sequences from our dataset (which would correspond to population-level sampling). We identified 20 clusters of redundant sequences (highlighted in Figure 1.5 and shown in Table 1.4) and we selected one representative from each cluster (noted with an asterisk in Table 1.4). As a result, in the BEAST analysis a total of 66 specimens were included. Of them, 38 are in the family Orsolobidae, 16 in the other families of the superfamily Dysderoidea (10 Oonopidae, 4 Dysderidae, and 2 Segestriidae) and 12 are outgroup taxa belonging to other families (Caponiidae, Filistatidae, Hypochilidae, Leptonetidae, Scytodidae, Tetrablemmidae, Trogloraptoridae).

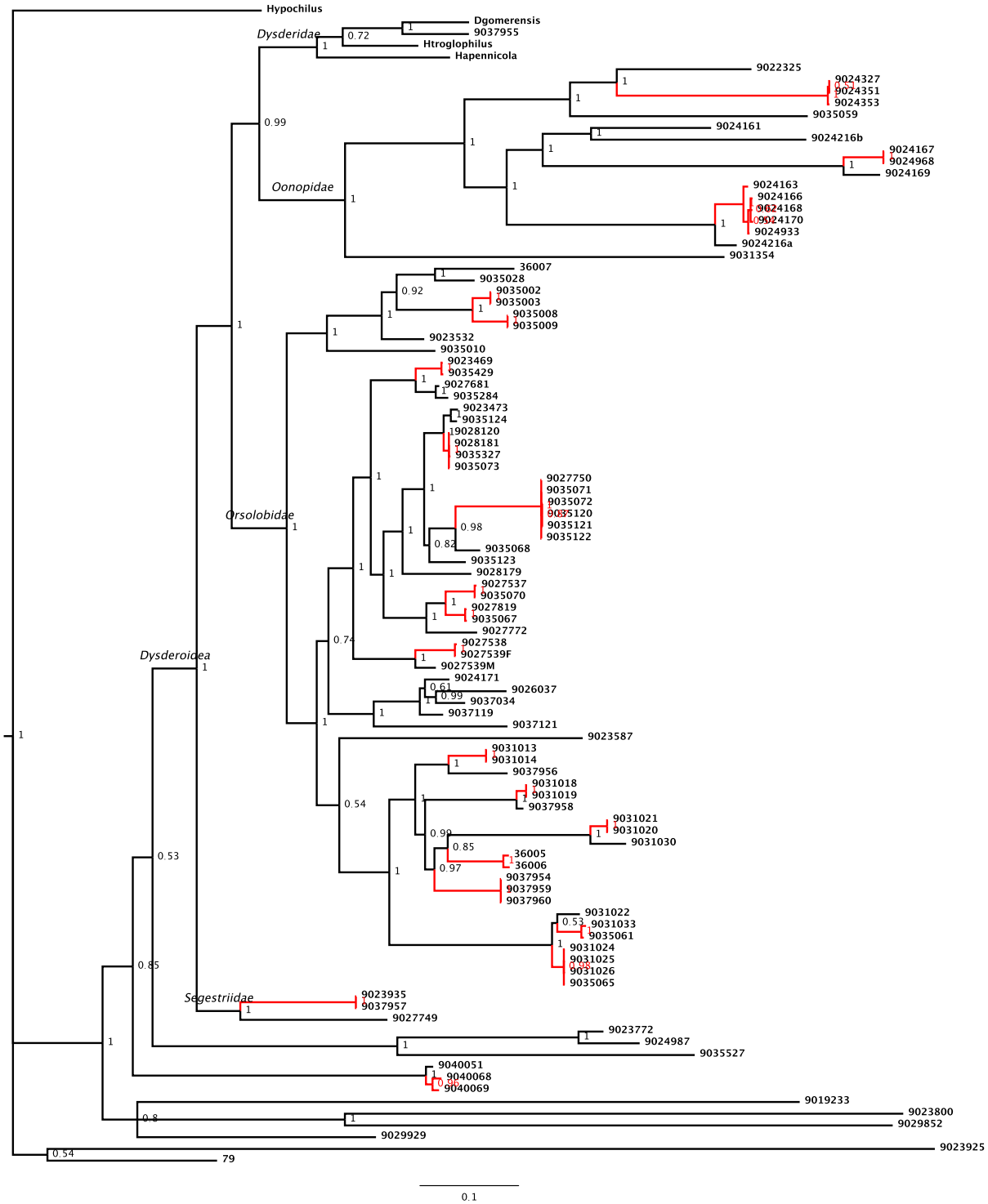

Figure 1.5: Majority-rule consensus tree of the full dataset using MrBayes. Node values are posterior probabilities. Clusters of almost identical sequences are highlighted in red.

Voucher	Genus/Species	Family	Voucher	Genus/Species	Family
9040068	*Trogloraptor* sp.	Trogloraptoridae	9035002	*Tasmanoonops* sp.	Orsolobidae
*9040069	*Trogloraptor* sp.	Trogloraptoridae	*9035003	*Tasmanoonops* sp.	Orsolobidae
OONO36005	*Duripelta* sp.	Orsolobidae	9027537	*Orsolobus* sp.	Orsolobidae
*OONO36006	*Duripelta* sp.	Orsolobidae	*9035070	*Orsolobus* sp.	Orsolobidae
9037954	*Wiltonia* sp.	Orsolobidae	*9024327	gen. nov. 1	Oonopidae
9037959	*Wiltonia rotoiti*	Orsolobidae	9024351	gen. nov. 1	Oonopidae
*9037960	*Wiltonia rotoiti*	Orsolobidae	9024353	gen. nov. 1	Oonopidae
*9027538	*Losdolobus* sp.	Orsolobidae	*9031013	*Ascuta parornata*	Orsolobidae
9027539F	*Losdolobus* sp.	Orsolobidae	9031014	*Ascuta parornata*	Orsolobidae
9031018	*Maoriata magna*	Orsolobidae	*9031020	*Pounamuella complexa*	Orsolobidae
*9031019	*Maoriata magna*	Orsolobidae	9031021	*Pounamuella complexa*	Orsolobidae
9031024	*Subantarctia fiordensis*	Orsolobidae	9028120	*Osornolobus anticura*	Orsolobidae
9031025	*Subantarctia fiordensis*	Orsolobidae	*9028181	gen. nov.	Orsolobidae
9031026	*Subantarctia fiordensis*	Orsolobidae	9035073	*Osornolobus* sp.	Orsolobidae
*9035065	*Subantarctia fiordensis*	Orsolobidae	9035327	*Orsolobus* sp.	Orsolobidae
*9035008	*Tasmanoonops* sp.	Orsolobidae	9024167	gen. nov. 2	Oonopidae
9035009	*Tasmanoonops* sp.	Orsolobidae	*9024968	gen. nov. 2	Oonopidae
9023935	*Segestria* sp.	Segestriidae	*9023469	*Mallecolobus maullin*	Orsolobidae
*9037957	*Segestria* sp.	Segestriidae	9035429	*Orsolobus* sp.	Orsolobidae
*9027819	*Orsolobus* sp.	Orsolobidae	9031033	*Subantarctia fiordensis*	Orsolobidae
9035067	*Orsolobus pucara*	Orsolobidae	*9035061	*Subantarctia* sp.	Orsolobidae
9027750	*Osornolobus* sp.	Orsolobidae	9024163	*Opopaea* sp.	Oonopidae
9035071	*Osornolobus* sp.	Orsolobidae	9024166	*Opopaea* sp.	Oonopidae
9035072	*Osornolobus* sp.	Orsolobidae	9024168	*Opopaea* sp.	Oonopidae
9035120	*Osornolobus* sp.	Orsolobidae	*9024170	*Opopaea* sp.	Oonopidae
9035121	*Osornolobus* sp.	Orsolobidae	9024933	*Opopaea* sp.	Oonopidae
*9035122	*Osornolobus* sp.	Orsolobidae			

Table 1.4: Clusters of almost identical sequences. For BEAST analyses, only one specimen per cluster was included (marked with an asterisk).

BEAST results

The PartitionFinder recommendations regarding model of evolution and partitioning scheme for the BEAST analysis are shown in Table 1.5. In order to achieve good mixing of the chains, models were slightly simplified for some partitions. More specifically, some GTR models (6 rate parameters) were substituted by TN93 models (3 rate parameters). Based on the Bayes Factors comparison, there was decisive evidence against the strict clock, while there was not much difference between a lognormal and an exponential clock (Bayes Factor difference in favor of the lognormal clock was 2.33). We have used a lognormal clock in all subsequent analyses. In terms of the tree prior, all three tested priors yielded very similar Bayes Factors (the differences were <1). Birth-death tree priors with incomplete sampling were used for all subsequent analyses.

partition	rec. model	impl. model (if different)	gene fragment(s)
1	GTR+G	TN93+G	CO1pos1
2	GTR+G	TN93+G	CO1pos2
3	GTR+G	TN93+G	16S
4	K80+G		H3pos3
5	SYM+G	TN93(ef)+G	Actinpos1, H3pos1
6	JC+I		Actinpos2, H3pos2
7	SYM+G		18S
8	SYM+G		Actinpos3
9	GTR+G		28S

Table 1.5: PartitionFinder model recommendations and implemented models for BEAST analyses

Divergence time estimation

The maximum clade credibility tree using the fossils to calibrate nodes including stems is shown in Figure 1.6. The topology of the tree is, as expected, identical to that of the full analysis (except some very low support relationships among outgroup taxa). Orsolobids from each landmass except Australia form well-supported monophyletic groups in our analysis. Australian specimens form two clades at the base of Orsolobidae. The African and South American clades are sister to each other. The relative position of the New Zealand clade remains unresolved (the posterior probability of the branch uniting the New Zealand clade with the African-South American clade is only 66%). The divergence time estimates for key nodes are shown in Table 1.6.

When using the fossils to calibrate crown groups, the resulting divergence times were in general slightly older. Finally, when using the substitution rate of 28S, the median ages

are even older and the 95% credibility interval wider and skewed towards older ages (see Table 1.6).

node	fossils (crown)	fossils (stem)	28S rate
Orsolobidae	130 (92-167)	106 (81-132)	178 (97-314)
Africa - South America split	93 (65-125)	76 (55-99)	129 (68-227)
New Zealand	73 (51-103)	60 (42-82)	101 (53-185)

Table 1.6: Divergence date estimates for certain nodes of interest according to different calibration methods. Median values in million years ago with 95% credibility intervals.

1.4 Discussion

Biogeographic implications

Tree Topology In our analyses we recovered all landmasses monophyletic with the exception of Australia. The Australian Orsolobids form a main monophyletic group with one enigmatic specimen placed outside. This pattern is indicative of very low dispersal if any, since we have no instances of lineages from one continent nested within clades from another continent. Also, our topology, and especially the sister group relationship between the African and South American clades, matches one of the common 'Gondwanan' area cladograms (see Fig. 1.3) and supports the West Gondwanan scenario.

The New Zealand taxa form a highly supported monophyletic group, which makes a recent dispersal explanation improbable. If New Zealand was colonized through dispersal more recently than the last 23Ma, we would expect multiple colonizations to account for the huge diversity of endemic taxa. Even if all the present day taxa diversified somewhere else and then dispersed to New Zealand, in order for them to form a monophyletic group, we would have to assume that all their relatives went extinct everywhere else. Furthermore, the New Zealand clade is not nested within taxa from another continent, which would presumably be the source of colonists.

Divergence Date Estimates We used Dysderoid fossils and substitution rates to calibrate the Orsolobid tree. Throughout the tree of life, it has been observed that different calibration methods can give very different results, as fossil calibrations tend to underestimate ages, while molecular rate calibrations tend to overestimate them (Benton and Ayala, 2003). Indeed, in our analyses, the ages obtained by the 28S rate calibration were significantly older and their distributions were skewed towards older dates. As for fossil calibrations, stem calibrations give younger and therefore more conservative ages than crown calibrations.

Despite the large range in our estimates, the obtained ages are broadly consistent with the timeframe of the Gondwanan breakup. The family Orsolobidae is estimated to be at least

100Ma old and it diversified during the breakup of Gondwana. The age of the split between Africa and South America ranges from 129Ma according to the rate calibration (during the opening of the south Atlantic ocean) to 76Ma according to the fossil stem calibration (after the landbridge connecting the two continents is suggested to have been broken), and does not allow us to differentiate between these two possibilities.

The picture is much clearer though, concerning the New Zealand node. In all our analyses, the 95% credibility interval of the node age does not include the Oligocene (34-23Ma), thus rejecting the possibility that the New Zealand taxa diversified after the proposed Oligocene 'drowning'. The common ancestor of all New Zealand taxa included in the analysis has an estimated age of at least 60Ma, an age consistent with the separation of New Zealand from Australia. As explained above, the assumption that New Zealand taxa diversified elsewhere and subsequently dispersed, would require the additional assumption that all of their relatives went extinct at the source. Given that the geology is equivocal in regards to New Zealand being fully or partially submerged (Landis et al., 2008), we argue that biogeographical evidence suggesting continuous land presence, such as this study, should be of decisive importance.

Taxonomic implications

The superfamily Dysderoidea The superfamily Dysderoidea was recovered as monophyletic with very high support in all analyses. The sister group of Orsolobidae is a highly supported clade uniting Oonopidae and Dysderidae, while Segestriidae is sister to all other Dysderoid families. This internal topology of Dysderoid families has not been suggested before to our knowledge. The proposed synapomorphies uniting Orsolobidae and Oonopidae (proprioreceptor bristles, bipectinate tarsal claws) and Orsolobidae and Dysderidae (spermatozoic characters) need a reversal in order to be optimized on our tree. At the same time, no putative morphological synapomorphy has been suggested uniting Oonopidae and Dysderidae to our knowledge.

As for the sister taxon of the superfamily Dysderoidea, our results cannot provide a definite answer. The support values outside Dysderoidea are extremely low in the BEAST analysis (Figure 1.6, essentially corresponding to a polytomy. In the MrBayes analysis however, there is a moderately supported clade (0.85 posterior probability) consisting of Dysderoidea, Caponiidae and Trogloraptoridae (Figure 1.5). This arrangement agrees with recent results and hypotheses based on morphological characters (Griswold, Audisio, and Ledford, 2012; Ramírez, 2000). This clade is completely absent from the BEAST analysis, most probably because the topology is parameterized very differently in the two programs. Another clade, also absent from the BEAST analysis but present in the MrBayes analysis with high support is uniting all families except Filistatidae and Leptonetidae. This clade, with the exception of Leptonetidae corresponds to the Sunspermiata proposed by Michalik and Ramírez (2014).

The Orsolobidae genera Although our taxon sampling is not adequate to test the monophyly of all Orsolobid genera included in our analysis, our results provide preliminary val-

idation for some genera, while suggesting that others need to be re-evaluated as they are possibly polyphyletic or not clearly defined.

All New Zealand genera were recovered as monophyletic with the exception of *Wiltonia*. Both specimens included in the BEAST analysis were identified as *Wiltonia rotoiti*, but didn't group together. The genus *Subantarctia* is sister to all other New Zealand genera in our analysis. The African genera were not recovered monophyletic, as a representative of *Afrilobus* was nested within *Azanialobus*. However, this may reflect our inadequate taxonomic knowledge, as the specimen in question was only provisionally placed in *Afrilobus* and it could be a new genus. As for the Australian taxa, they do not all form a monophyletic group. The only specimen placed outside the main Australian clade is provisionally placed in *Hickmanolobus*, but could also represent a new undescribed genus.

In New Zealand, Africa and Australia, the discrepancies mentioned above are relatively minor and could correspond to new undescribed taxa, rather than errors in the definition of described taxa. The situation in South America, however, is much more complicated. *Losdolobus* is the only South American genus that was recovered as monophyletic in our analysis, and it is sister to all other South American specimens. *Mallecolobus*, *Orsolobus*, and *Osornolobus* are all polyphyletic in our analysis, with specimens assigned to different genera having very similar or even identical sequences. Indeed the placement of many specimens to genus was difficult, as the genera are defined as a combination of characters (Forster and Platnick, 1985) and the specimens often exhibited some characters and not others for more than one genera.

In regards to the relationships among Orsolobidae genera, Forster and Platnick (1985) suggested that *Subantarctia* and *Orsolobus* exhibited plesiomorphic characters and could be basal within the family. Our analysis does not support such a placement for these genera. They also suggested *Chileolobus* as a potentially early divergent taxon, but we were unfortunately unable to include a representative in our analysis.

1.5 Conclusion

Our results show that the estimated divergence times of Orsolobidae are consistent with the timeframe of the Gondwanan breakup. Also, Orsolobidae exhibit one of the commonly attested Gondwanan area cladograms, with African and South American taxa sister to each other. New Zealand Orsolobids form a highly supported monophyletic group with its common ancestor estimated to be at least 60Ma old, therefore rejecting the 'drowned' New Zealand hypothesis.

This study also represents the first molecular phylogenetic analysis for the family Orsolobidae and the superfamily Dysderoidea. Our results suggest a new hypothesis for relationships between Dysderoid families, with Segestriidae sister to all other families and Orsolobidae sister to a clade consisting of Oonopidae and Dysderidae. Also, some Orsolobid genera were polyphyletic in our analysis, especially the ones occurring in South America,

indicating the need for a more complete phylogenetic analysis of Orsolobidae using both molecular and morphological data and a new taxonomic revision.

1.6 Acknowledgements

This work was funded from the NSF spider AToL project (C.G.) and the PBI Oonopidae project (N. C.-P. and A.S.). Also from the Hagey Research Venture Fund (A. C. and C. G.), the G. Lindsay Field Research Fund (T. S. and C. G.), the Exline-Frizzell Fund for Arachnological Research (N.C.-P., A.C., T.S., A.S., and C.G.), and the Schlinger Foundation (A.C. and C.G.).

The authors would like to thank Diana Silva, Elizabeth Arias, Elizabeth Morrill, Lina Almeida-Silva, Jeremy Miller, and especially Hannah Wood for providing specimens for this study.

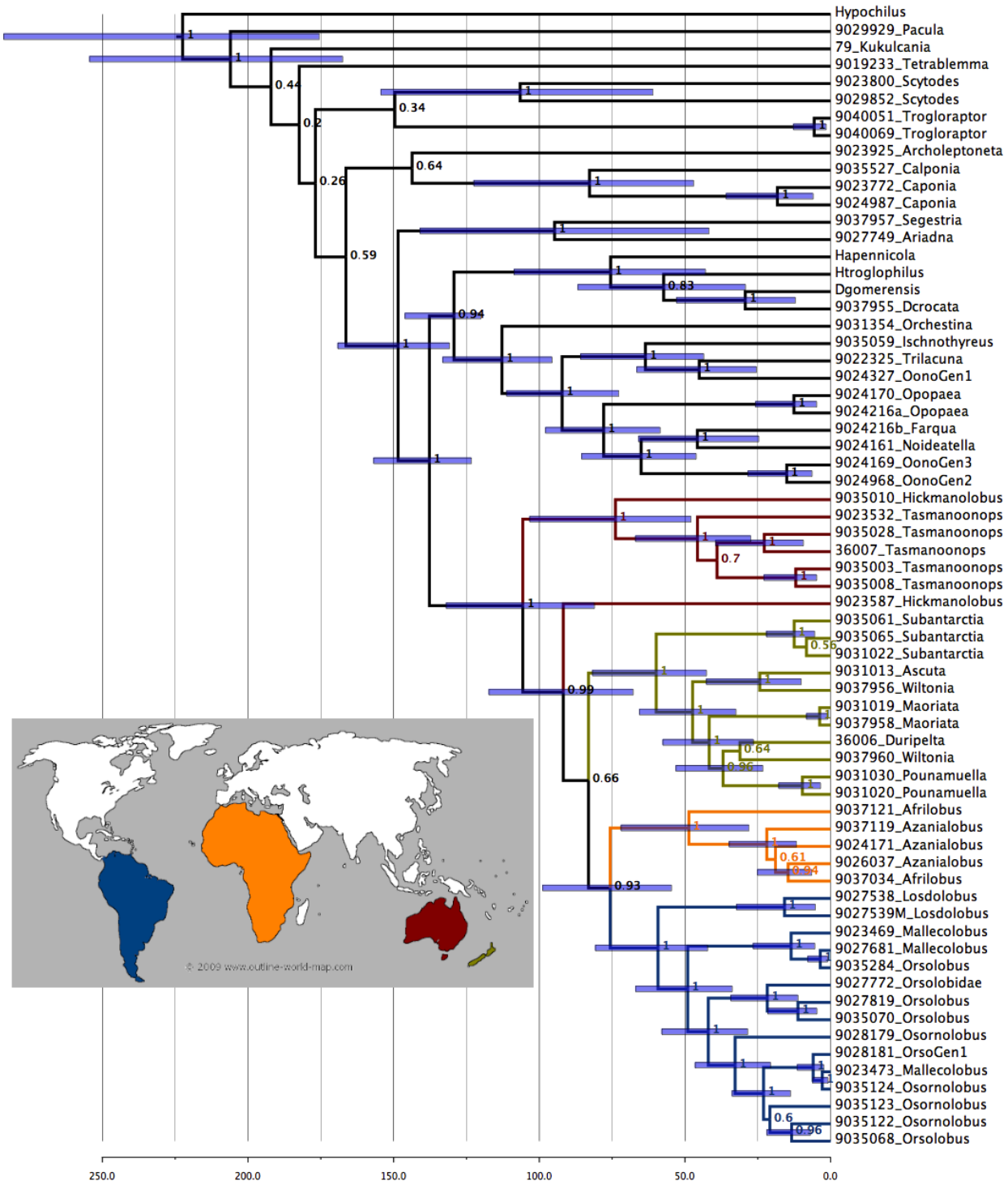

Figure 1.6: BEAST results: maximum clade credibility tree with median heights using fossil stem calibrations. Ages in million years. Node values are posterior probabilities. Clades are colored according to biogeographical regions.

Bibliography

Baehr, BC and HM Smith (2008). "Three new species of the Australian orsolobid spider genus *Hickmanolobus* (Araneae: Orsolobidae)". In: *Records of the Western Australian Museum* 24, pp. 325–336.

Baele, Guy et al. (2012). "Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty." In: *Molecular Biology and Evolution* 29.9, pp. 2157–67.

Baker, Allan J et al. (June 2005). "Reconstructing the tempo and mode of evolution in an extinct clade of birds with ancient DNA: the giant moas of New Zealand." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.23, pp. 8257–62.

Benton, MJ and FJ Ayala (2003). "Dating the tree of life". In: *Science* 300.5626, pp. 1698–1700.

Bidegaray-Batista, Leticia and Miquel A Arnedo (2011). "Gone with the plate: the opening of the Western Mediterranean basin drove the diversification of ground-dweller spiders." In: *BMC Evolutionary Biology* 11.1, p. 317.

Blakey, Ronald C (2008). "Gondwana paleogeography from assembly to breakup—A 500 m.y. odyssey". In: *Geological Society of America Special Papers* 441, pp. 1–28.

Campbell, Hamish and Chuck Landis (2003). "New Zealand awash". In: *New Zealand Geographic*, pp. 6–7.

Colgan, D. J. et al. (1998). "Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution". In: *Australian Journal of Zoology* 46.5, p. 419.

Cook, L. G and M. D Crisp (2005). "Not so ancient: the extant crown group of *Nothofagus* represents a post-Gondwanan radiation". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1580, pp. 2535–2544.

Cooper, A et al. (Feb. 2001). "Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution." In: *Nature* 409.6821, pp. 704–7.

Drummond, A. J., M. A. Suchard, et al. (2012). "Bayesian Phylogenetics with BEAUti and the BEAST 1.7". In: *Molecular Biology and Evolution* 29.8, pp. 1969–1973.

Drummond, AJ and A Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC evolutionary biology* 7.1, p. 214.

Ericson, Per G P et al. (Feb. 2002). "A Gondwanan origin of passerine birds supported by DNA sequences of the endemic New Zealand wrens." In: *Proceedings. Biological sciences / The Royal Society* 269.1488, pp. 235–41.

Folmer, O et al. (1994). "DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates". In: *Mol. Mar. Biol. Biotechnol* 3.5, pp. 294–299.

Forster, Raymond R and Norman I Platnick (1985). "A review of the austral spider family Orsolobidae (Arachnida, Araneae), with notes on the superfamily Dysderoidea." In: *Bulletin of the American Museum of Natural History* 181, pp. 1–230.

Forster, RR, NI Platnick, and MR Gray (1987). "A review of the spider superfamilies Hypochiloidea and Austrochiloidea (Araneae, Araneomorphae)." In: *Bulletin of the American Museum of Natural History* 185.

Giribet, Gonzalo et al. (Jan. 1996). "First molecular evidence for the existence of a Tardigrada+ Arthropoda clade." In: *Molecular Biology and Evolution* 13.1, pp. 76–84.

Griswold, Charles E, Tracy Audisio, and Joel M Ledford (Jan. 2012). "An extraordinary new family of spiders from caves in the Pacific Northwest (Araneae, Trogloraptoridae, new family)." In: *ZooKeys* 215, pp. 77–102.

Griswold, Charles E. and Norman I. Platnick (1987). "On the first African spiders of the family Orsolobidae (Araneae, Dysderoidea)." In: *American Museum Novitates* 2892.

Hedin, M C and W P Maddison (Mar. 2001). "A combined molecular approach to phylogeny of the jumping spider subfamily dendryphantinae (araneae: salticidae)." In: *Molecular phylogenetics and evolution* 18.3, pp. 386–403.

Huelsenbeck, JP and F Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees". In: *Bioinformatics* 17.8, pp. 754–755.

Izquierdo, Matías Andrés and Facundo Martín Labarque (Dec. 2010). "Description of the female of Orsolobus pucara Forster & Platnick 1985, with comments on the functional morphology of the female genitalia in Dysderoidea (Araneae: Dysderoidea: Orsolobidae)". In: *Journal of Arachnology* 38.3, pp. 511–520.

Katoh, Kazutaka and Daron M Standley (Apr. 2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." In: *Molecular biology and evolution* 30.4, pp. 772–80.

Knapp, Michael et al. (Oct. 2007). "The drowning of New Zealand and the problem of Agathis." In: *Systematic Biology* 56.5, pp. 862–70.

Landis, C. A. et al. (Mar. 2008). "The Waipounamu Erosion Surface: questioning the antiquity of the New Zealand land surface and terrestrial fauna and flora". In: *Geological Magazine* 145.02, pp. 173–197.

Lanfear, Robert, Brett Calcott, Simon Y W Ho, et al. (June 2012). "Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses." In: *Molecular biology and evolution* 29.6, pp. 1695–701.

Lanfear, Robert, Brett Calcott, David Kainer, et al. (Jan. 2014). "Selecting optimal partitioning schemes for phylogenomic datasets." In: *BMC evolutionary biology* 14.1, p. 82.

Lee, Daphne E., Jennifer M. Bannister, and Jon K. Lindqvist (Dec. 2007). "Late Oligocene-Early Miocene leaf macrofossils confirm a long history of Agathis in New Zealand". In: *New Zealand Journal of Botany* 45.4, pp. 565–578.

Li, ZX and CMA Powell (2001). "An outline of the palaeogeographic evolution of the Australasian region since the beginning of the Neoproterozoic". In: *Earth-Science Reviews* 53.3, pp. 237–277.

Lipke, Elisabeth, Martín J Ramírez, and Peter Michalik (June 2014). "Ultrastructure of spermatozoa of orsolobidae (Haplogynae, Araneae) with implications on the evolution of sperm transfer forms in Dysderoidea." In: *Journal of morphology*. In press.

Lise, AA and L Almeida (2006). "A new species of *Losdolobus* Platnick & Brescovit, 1994 (Araneae: Dysderoidea: Orsolobidae) from southern Brazil". In: *Zootaxa* 9, p. 249.

Maddison, D R, W Moore, et al. (Jan. 2009). "Monophyly of terrestrial adephagan beetles as indicated by three nuclear genes (Coleoptera: Carabidae and Trachypachidae)." In: *Zoologica scripta* 38.1, pp. 43–62.

Maddison, WP and DR Maddison (2007). *Mesquite: a modular system for evolutionary analysis. Version 2.75. 2011*. URL: `http://mesquiteproject.org`.

Michalik, Peter and Martín J Ramírez (June 2014). "Evolutionary morphology of the male reproductive system, spermatozoa and seminal fluid of spiders (Araneae, Arachnida) - Current knowledge and future directions." In: *Arthropod structure & development* 43.4, pp. 291–322.

Miller, Jeremy A et al. (June 2010). "Phylogeny of entelegyne spiders: affinities of the family Penestomidae (NEW RANK), generic phylogeny of Eresidae, and asymmetric rates of change in spinning organ evolution (Araneae, Araneoidea, Entelegynae)." In: *Molecular phylogenetics and evolution* 55.3, pp. 786–804.

Nylander, Johan A.A. et al. (2008). "AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics". In: *Bioinformatics* 24.4, pp. 581–583.

Ott, R and AA Lise (2004). "Description and ecology of two new species of the Brazilian spider genus Losdolobus Platnick and Brescovit (Araneae, Dysderoidea, Orsolobidae)". In: *Revista ibérica de Aracnología* 9, pp. 249–257.

Ott, R, NI Platnick, et al. (2013). "*Basibulbus*, a hard-bodied, haplogyne spider genus from Chile (Araneae, Dysderoidea)". In: *American Museum Novitates* 3775, pp. 1–20.

Penney, D (2000). "Miocene spiders in Dominican amber (Oonopidae, Mysmenidae)". In: *Palaeontology* 43.2, pp. 343–357.

Platnick, NI (1977). "The hypochiloid spiders: a cladistic analysis, with notes on the Atypoidea (Arachnida, Araneae)." In: *American Museum novitates* 2627.

— (2014). *The world spider catalog, version 15.* URL: `http://research.amnh.org/iz/spiders/catalog/`.

Platnick, Norman I. and Antonio D. Brescovit (1994). "A new genus of the spider family Orsolobidae (Araneae, Dysderoidea) from Brazil." In: *American Museum novitates* 3112.

Platnick, Norman I., Jonathan A. Coddington, et al. (1991). "Spinneret morphology and the phylogeny of haplogyne spiders (Araneae, Araneomorphae)." In: *American Museum novitates* 3016.

Platnick, Norman I. and Gareth Nelson (1978). "A Method of Analysis for Historical Biogeography". In: *Systematic Zoology* 27.1, p. 1.

Pole, Mike (Oct. 1994). "The New Zealand Flora-Entirely Long-Distance Dispersal?" In: *Journal of Biogeography* 21, pp. 625–635.

Queiroz, Alan de (2005). "The resurrection of oceanic dispersal in historical biogeography." In: *Trends in ecology & evolution* 20.2, pp. 68–73.

Rambaut, A. et al. (2013). *Tracer v.1.6*. URL: `http://beast.bio.ed.ac.uk/Tracer`.

Ramírez, MJ (2000). "Respiratory system morphology and the phylogeny of haplogyne spiders (Araneae, Araneomorphae)". In: *Journal of Arachnology* 28.2, pp. 149–157.

Raxworthy, C. J., M. R. J. Forstner, and R. A. Nussbaum (2002). "Chameleon radiation by oceanic dispersal". In: *Nature* 415.6873, pp. 784–787.

Renner, S. S. (2004). "Multiple Miocene Melastomataceae dispersal between Madagascar, Africa and India". In: *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 359.1450, pp. 1485–1494.

Ronquist, F and JP Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models". In: *Bioinformatics* 19.12, pp. 1572–1574.

Rosen, Donn E. (1978). "Vicariant Patterns and Historical Explanation in Biogeography". In: *Systematic Zoology* 27.2, p. 159.

Sanmartín, Isabel and Fredrik Ronquist (Apr. 2004). "Southern hemisphere biogeography inferred by event-based models: plant versus animal patterns." In: *Systematic biology* 53.2, pp. 216–43.

Sereno, Paul C, Jeffrey A Wilson, and Jack L Conrad (July 2004). "New dinosaurs link southern landmasses in the Mid-Cretaceous." In: *Proceedings. Biological sciences / The Royal Society* 271.1546, pp. 1325–30.

Simon, Chris et al. (1994). "Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers". In: *Annals of the entomological Society of America* 87.6, pp. 651–701.

Upchurch, Paul (Apr. 2008). "Gondwanan break-up: legacies of a lost world?" In: *Trends in ecology & evolution* 23.4, pp. 229–36.

Vandergast, Amy G, Rosemary G Gillespie, and George K Roderick (July 2004). "Influence of volcanic activity on the population genetic structure of Hawaiian Tetragnatha spiders: fragmentation, rapid population growth and the potential for accelerated evolution." In: *Molecular ecology* 13.7, pp. 1729–43.

Vink, Cor J et al. (2008). "Actin 5C, a promising nuclear gene for spider phylogenetics". In: *Molecular phylogenetics and evolution* 48.1, pp. 377–382.

Waters, Jonathan M and Dave Craw (Apr. 2006). "Goodbye Gondwana? New Zealand biogeography, geology, and the problem of circularity." In: *Systematic Biology* 55.2, pp. 351–6.

Wheeler, Ward C and Cheryl Y Hayashi (June 1998). "The Phylogeny of the Extant Chelicerate Orders". In: *Cladistics* 14.2, pp. 173–192.

Whiting, M. F. et al. (Mar. 1997). "The Strepsiptera Problem: Phylogeny of the Holometabolous Insect Orders Inferred from 18S and 28S Ribosomal DNA Sequences and Morphology". In: *Systematic Biology* 46.1, pp. 1–68.

Wood, Hannah Marie, Charles E. Griswold, and Rosemary G. Gillespie (Dec. 2012). "Phylogenetic placement of pelican spiders (Archaeidae, Araneae), with insight into evolution of the "neck" and predatory behaviours of the superfamily Palpimanoidea". In: *Cladistics* 28.6, pp. 598–626.

Xia, Xuhua and Philippe Lemey (2009). "Assessing substitution saturation with DAMBE." In: *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*, pp. 615–630.

Xia, Xuhua, Zheng Xie, et al. (Jan. 2003). "An index of substitution saturation and its application". In: *Molecular Phylogenetics and Evolution* 26.1, pp. 1–7.

Yang, Ziheng (2006). *Computational molecular evolution*. Oxford New York: Oxford University Press.

Chapter 2

Linguistic Characters and Phylogenetics

2.1 Introduction

The analogies between biological and linguistic evolution were recognized already by Darwin (Darwin, 1871), and, as it will be argued below, systematic biology and historical linguistics have strikingly similar histories and methods. Even though the two fields have certainly inspired each other in the past (Atkinson and Gray, 2005), they had rarely been in direct communication (Hoenigswald and Wiener, 1987), until recently with the increasing use of phylogenetic algorithms developed in evolutionary biology to study language relationships (Bouckaert et al., 2012; Bowern and Atkinson, 2012; Gray, Drummond, and Greenhill, 2009; Gray and Jordan, 2000). These recent developments are a very hopeful sign for fertile collaborations between biologists and linguists and for deepening our understanding of both linguistic and biological evolution. However, until now most studies use only lexical data and a coding method that produces non-independent characters (see §The Gray and Atkinson method). It is therefore important to develop coding methods that take advantage of all the phylogenetic information that languages can provide and at the same time comply with the assumptions of phylogenetic algorithms.

In the biological literature there has been lively debate on what is a character and how to code its states (Lipscomb, 1992; Pimentel and Riggins, 1987; Pleijel, 2005), as well as on inference methods (Hennig, 1965; Mayr, 1965; Sneath, 1971) . Linguistic characters share a lot of characteristics with morphological characters in biology, and the inference methods in biological systematics and historical linguistics are strikingly similar, as it will be explained below. Since contemporary phylogenetic methods were developed with biological characters in mind, understanding these analogies and debates is vital for the correct application of such methods in linguistics.

Similar to biological organisms, languages have a number of different levels of organization and in principle each one can provide informative characters for phylogenetic reconstruction. The main organization levels of a language are:

- Phonology: the sounds of a language and the rules that govern them, how they are

combined to form roots and functional items, such as affixes

- Lexicon: the vocabulary of a language
- Morphology: the grammatical-functional items of the language, such as suffixes, prepositions, pronouns and the rules that govern how to combine them with lexical roots in order to produce words
- Syntax: how to combine words in sentences
- Semantics: what the words mean and how meanings change over time
- Pragmatics: how is the language used in real-life situations, phenomena that are outside of the sentence and make their appearance in discourse

Historical linguists have traditionally relied on phonology and morphology for reconstructing language relationships, and to a lesser degree on the lexicon, which is often considered more susceptible to borrowing. Therefore, it is important that phylogenetic methods be adapted to handle phonological and morphological characters.

In this chapter, I will first briefly review of the parallel histories of the fields of biological systematics and historical linguistics in order to establish the correspondences in concepts and methods. I will also touch upon some key differences in the development of the two fields. Then, I will describe how linguistic characters can be coded and analyzed in a phylogenetic framework, while respecting the methods' assumptions as well as taking into account linguists' intuitions at the same time.

2.2 Biological Systematics and Historical Linguistics: Parallel Lives

The following account is far from a complete review of the history of biological systematics and historical linguistics. My goal is to underline key concepts and methods in the two fields and how they are related. For more detailed information, see Atkinson and Gray (2005) and references therein.

Hennigian cladistics and the comparative method

Willi Hennig, a German entomologist, revolutionized the field of systematic biology in the middle of the 20th century (Hennig, 1950, 1966) and laid the foundations of the cladistic method (Queiroz and Gauthier, 1992). Hennig's most well-known book, *Phylogenetic Systematics* (1966), presented for the first time an organized and coherent method of classification based on 'the degree of phylogenetic kinship', and defined formally many now commonly-used terms, such as *synapomorphy*, *symplesiomorphy*, *monophyletic* and *paraphyletic*. Below are the main steps of the cladistic method, as outlined in Hennig's book:

1. identify characters that belong to the same 'phylogenetic transformation series' (i.e. *homologous* characters)

2. determine which character (or character state according to modern terminology) within each transformation series is plesiomorphous (ancestral) and which apomorphous (derived)

3. use mutual possession of apomorphous characters (*synapomorphy*) as the basis for grouping organisms in monophyletic groups.

Historical linguists will easily recognize in the above procedure the comparative method, the main method for classifying languages. The only difference is the terminology: synapomorphies are shared innovations and symplesiomorphies shared retentions. Indeed, the essence of the cladistic method was known and used by historical linguists long before Hennig arrived at the same principles. The German linguist Karl Brugmann had already distinguished innovations from retentions and argued that only shared innovations are relevant for inferring relationships in 1884 (Atkinson and Gray, 2005). Another field where essentially cladistic methods were widely in use before Hennig's work is *stemmatics*, the study of ancient manuscripts with the goal of reconstructing the original. In this context, the innovations are mistakes introduced to the text through repeated copying.

But the similarity between the comparative method and Hennigian cladistics is striking even beyond the basic principles. The first step in both methods is to establish what is "the same" across the languages or organisms under study, i.e. to establish *cognacy* or *homology*. As Hennig put it, we cannot directly observe transformation series, so we need to employ criteria in order to posit hypotheses of homology between characters. Some of the criteria for homology of morphological features in biology is similarity in position and fine structure (Hennig, 1966; Remane, 1971). In molecular characters, multiple sequence alignment uses position to create columns of homologous nucleotides or aminoacids. Similarly, to find cognate words in historical linguistics one searches for words similar in phonological form and meaning. True cognate words, when "aligned" give rise to regular *sound correspondences*, a somewhat more abstract kind of linguistic data, which can form the basis of phonological characters (Crowley and Bowern, 2010).

The second step in Hennig's method is the character analysis, where characters are polarized by identifying the ancestral state. The criteria used for finding the ancestral state are very similar in historical linguistics and biology. Through experience, researchers acquire a sense of which direction of change is more "natural" or "common". In both fields there are conditions that are considered irreversible or dead ends (e.g. flightlessness in birds or a phonemic merger, where two originally distinct sounds start being pronounced the same). External evidence is taken into account in many cases, such as the fossil record or early written forms of a language, which can shed light on what the earliest known state of a character is. In biology, ontogenetic evidence is another kind of external evidence. In historical linguistics, acoustic phonetics or articulation arguments have been employed to favor certain directions of change (e.g. nasalization of oral vowels before nasal coda segments). In cases

where polarity cannot be established otherwise, both fields have made use of the unreliable principle "common equals primitive", according to which the most common state is taken as ancestral (Campbell, 2004; Crowley and Bowern, 2010; Hennig, 1966).

The Rise and Fall of Numerical Taxonomy and Lexicostatistics

In the 1950s, the first computational classification methods were introduced in both systematic biology and historical linguistics. They were both distance-based methods very similar in their philosophy: first, pairwise "distances" are calculated between entities to be classified (e.g. number of common cognates for certain meanings or percent identity in DNA sequences) and then a clustering algorithm is used to construct a tree based on the distance matrix . In systematic biology, these methods are known as *phenetics* or *numerical taxonomy* and were first introduced by Sneath and Sokal (Michener and Sokal, 1957; Sneath, 1957). In historical linguistics, the method was introduced in 1952 by Swadesh and is known as *lexicostatistics* (Swadesh, 1952). Interestingly, in both fields these methods were heavily criticized and are now largely discredited, although variations of them are still in use.

Despite the similarity in principles and procedure between numerical taxonomy and lexicostatistics, biologists and linguists arrived there from different paths. Sneath and Sokal emphasized that in order to classify organisms a great number of features needs to be taken into account and they should be weighted equally. These views came as a reaction to the already established school of 'evolutionary' systematists (Mayr, 1965), who focused on few "important" features (Sneath and Sokal, 1962). They also favored *overall similarity* as the only criterion for classification. They were fully aware that their classifications did not necessarily mirror phylogenetic relationships, but they argued that taxonomy is primarily a practical endeavor and it should not be based on evolutionary history, since overall similarity is more interesting and useful to biologists and knowledge of true phylogeny is unattainable (Sneath and Sokal, 1973; Sokal and Sneath, 1963). Indeed, Sneath and Sokal argued that classifications should not be based only on homologous features (i.e. they should also include features that have converged to the same state independently), since it is impractical to exclude those cases (Sokal and Sneath, 1963). At the same time, they claimed that phenetic groups are more likely than not to be monophyletic.

Swadesh on the other hand, developed lexicostatistics as an analogy of radiocarbon dating, employing a principle of lexical "decay" leading vocabularies of related languages to become increasingly dissimilar (Swadesh, 1952). Based on the amount of retained basic vocabulary between ancient written languages and their modern descendants he calculated a similar constant rate of lexical replacement, i.e. a glottoclock. Swadesh created a list of 200 words (later reduced to 100) of basic vocabulary, which he claimed was resistant to borrowing and evolved at approximately the same rate universally (Swadesh, 1955). Using this fixed rate he could calculate when languages diversified from each other, a method he termed *glottochronology*. Swadesh's emphasis was on dating rather than classification, and

he doesn't comment on whether his method produces genetic[1] groupings or not. In historical linguistics lexicostatistics and glottochronology were directly linked from their inception. The analogue of glottochronology in biology is the *molecular clock*, first reported in the evolution of proteins in 1962 and then in DNA sequences as well (Bromham and Penny, 2003).

Critiques of numerical taxonomy in biology and lexicostatistics in historical linguistics have some common aspects but largely focused on different issues. The primary criticism of numerical taxonomy is that it frequently results in classifications that do not reflect genealogical relationships among the taxa being classified. This is a result of the fact that it is based on overall similarity, which includes similarity due to convergence, to symplesiomorphy and to synapomorphy (Hennig, 1966; Simpson, 1964) . As it became more and more established that biological classifications should reflect phylogeny and with the development of contemporary computational phylogenetic methods, phenetic classification methods fell out of favor. However, they remain still in use in population genetics and in microbiology.

Lexicostatistics and glottochronology on the other hand, were chiefly criticized on practical rather than philosophical grounds and in connection to the estimation of divergence times. Among the first observations, most of them accepted by Swadesh himself, was that the wordlist was not universal as it was not easily translatable in various unrelated languages. Also, it soon became evident that the distinction between "stable" - basic and "unstable"-cultural vocabulary was language dependent (Hoijer, 1956; Swadesh, 1955). The other main point of criticism is that it is in fact false that languages have a steady decay rate in their basic vocabulary, and that this rate is common crosslinguistically.[2] As a result lexicostatistics and glottochronology have become largely discredited in the linguistic literature (Campbell, 2004; Crowley and Bowern, 2010). However, Lexicostatistics has not been criticized for its central premise as a distance-based method: grouping according to overall similarity. When counting cognates in common in the basic vocabulary, even if we exclude borrowings, each of them could be either a shared innovation or a shared retention. Yet both of them are counted as evidence for grouping the languages together. Given that linguists have been using an essentially cladistic method for more than a century, it is striking how this fundamental difference has not been pointed out. As a result, linguists are not familiar in general with this fundamental difference between distance-based and character-based methods and consequently tend to either accept both as equally legitimate or to distrust any computational method altogether because it is considered a form of lexicostatistics.

[1] In historical linguistics the term genetic is equivalent to phylogenetic or "related through descent from a common ancestor" in biology, as opposed to groupings produced through borrowing and convergence.

[2] These criticisms are very similar to the debate in biology about the use of molecular clocks. Current uses of the molecular clock rarely assume a strict clock, but instead 'relax' the clock in various ways and allow multiple points of calibration using the fossil record or biogeography.

Contemporary phylogenetic methods

The phylogenetic methods used today in systematic biology are computational character-based methods (as opposed to distance-based) and their philosophy is closest to cladistics.

2.3 Linguistic characters in a contemporary phylogenetic context

Characters useful for phylogenetics can come from many different aspects of a language. Different types of characters (lexical, morphological, phonological etc) represent independent lines of evidence and can be used in combination to resolve relationships among languages. They also can reveal different histories and pressures for different aspects of a language. Modern phylogenetic methods give the possibility to combine different kinds of characters in a single analysis and they to ask explicit evolutionary questions about rates, correlated features, and divergence time estimates. Below, I describe methods for the definition and coding of various kinds of linguistic characters and their analysis within a phylogenetic framework.

Lexicon

The lexicon (the vocabulary of a language) is by far the most easily accessible and describable aspect of a language. Many times, especially for extinct languages, the only available information is a wordlist. Every lexical item is associated with two things: the *root* it is a descendent from and the *meaning* it bears. In a protolanguage (the most recent common ancestor of a monophyletic group of languages) there is a protoform associated with a protomeaning. The protoform is inherited to daughter languages with modification or it can occasionally be lost. The descendants of the same protoform are called *cognates*, they are instances of a common root in different languages and they form a *cognate set*. As the protoform passes to the daughter languages, it remains associated with a meaning. However this association is not always stable. In many cases the descendent word in a language does not bear the protomeaning but a different, more or less removed meaning. For example, English for the meaning '*Canis lupus familiaris*', has the word 'dog', while German (a close relative) has the word 'hund'. English has a cognate to the German 'hund', namely 'hound', but it no longer means '*Canis lupus familiaris*' as a cover term, rather it signifies a subcategory of dogs.

Languages can gain new roots either by borrowing them from other languages or by inventing them. If the language they borrow the root from is unrelated[3], then the result is the same as if they invented it. If however the languages are related, then we have the potential of mistaking the borrowed word for a genetic cognate, a situation analogous to

[3]In historical linguistics it is not proven or widely accepted that all extant languages are ultimately related.

horizontal transfer in biology. There are also cases, where a word is borrowed widely in many unrelated languages at the same time period, such as the word for 'telephone', which is more or less the same all over the world.

Below, I will describe three different coding approaches that can be employed to code lexical data for phylogenetic analysis and discuss briefly their advantages and disadvantages. Neither of these methods has been applied yet to a linguistic dataset as far as I know. The last chapter of the present work is a first attempt to implement a method similar to cognate presence-absence coding for the study of Tupí-Guaraní languages (see Chapter 3).

Cognate presence-absence

Based on the evolution of roots described above, an obvious coding method emerges: root presence/absence in a language. Basically, every character corresponds to a root and languages are coded based on the presence or absence of a cognate of this root. The resulting characters are irreversible if we exclude borrowing and sound symbolism,[4] because it is extremely improbable that the same root is invented twice.

The cognate presence-absence method has the advantage that it is philosophically sound and intuitive for linguists, as roots are the most obvious inherited entity in linguistics. It is also attractive because it can be easily modeled since the states are comparable across characters (0s mean absent and 1s means present). There are evolutionary models appropriate for such characters in standard phylogenetic software: generalized binary model or restriction site model in MrBayes and BEAST, and stochastic Dollo model in BEAST.

However, this coding method is really difficult to implement for the following reason: it is almost impossible to say with certainty that a root is absent in a language, unless someone has knowledge of its whole lexicon. In the above dog example, if we hadn't written down or couldn't ask someone how specific breeds of dogs are called in English, or if they have a word roughly sounding like 'hund' and meaning something related to dogs, we would never know that 'hound' existed in the language. All we would know is that by asking the word for '*Canis lupus familiaris*', we got the response 'dog', which is clearly not a cognate with German 'hund'. Data collection in linguistics typically starts with a wordlist, which is essentially a list of meaning slots. Therefore, in most cases, roots that are seemingly absent, can exist in the language but in meanings outside the wordlist used to gather lexical items. If we are too cautious with all such cases and we code them with question marks in our character matrix, we will end up with a largely uninformative matrix of 1s and ?s (and extremely few 0s).

It must be noted here that over the past decade a seemingly similar coding method has been used in a series of influential publications by R.Gray and Q. Atkinson and collaborators (Bouckaert et al., 2012; Bowern and Atkinson, 2012; Gray, Drummond, and Greenhill, 2009; Gray and Atkinson, 2003). Although the method results in similar irreversible characters, the criterion is not only cognacy (i.e. descent from the same root) but also identity in

[4]Sound symbolism can result in the independent innovation of very similar words in different languages, such as words for 'cuckoo' that imitate the bird's sound.

meaning. The characters are in fact binary recoded semantic-based multistate characters and are not independent, thus violating a key assumption of all phylogenetic methods. For more information on semantic-based characters, the Gray and Atkinson method and its disadvantages, see the following paragraph and chapter 3, §Cognate Sets and Character Coding.

Semantic-based cognate coding

The nature of the data collection procedure for lexical items, i.e. the use of a basic wordlist (usually of 100 or 200 words), as well as the organization of dictionaries by meaning point to an alternative coding method: multistate semantic (i.e. meaning-based) characters. For this approach, the inherited element is the meaning and the assumption is that this meaning is constant throughout the evolution of the group in question, i.e. every language ancestral and current should have a particular item for that semantic slot.[5] The states for this character would be the different cognate sets used for the particular meaning. In the dog example again, the semantic slot/character is '*Canis lupus familiaris*': If we consider the Indo-European language family, there is a root **kwon* which has cognates in most Indo-European languages ('canis' in Latin, 'chien' in French, 'hund' in German, etc), but English has the unrelated root 'dog' and Spanish the unrelated root 'perro' for the same meaning. So, in this case there will be 3 different states for this character, one for each root or cognate set. This approach results in multistate characters, much like typical morphological characters in biology.

Such characters are ideal for parsimony analyses, while modeling can lead to overparameterization because the states are not comparable across characters. However, recent advances in phylogenetic methods allow the analysis of generic multistate characters in a Bayesian framework (MrBayes). Semantic-based coding could probably be more informative for recent divergences than cognate presence-absence coding, since the meaning of a form can change, without the form going extinct.

Cognate-based semantic coding

A third possibility for coding lexical data is to use the cognate sets as the basis of the characters and code the meanings as the states. This coding would also yield multistate characters as the semantic-based coding mentioned before. However, in this case the states could be at least partially ordered, as there are semantic shifts that are more plausible than others. Such characters could be analyzed with parsimony, if partially ordered, or also with Bayesian methods if left unordered. This coding method presents additional challenges though, because meanings can be overlapping rather than distinct, making such characters similar to continuous characters in biology. Finally, some cognate sets are going to be absent from some languages, thus posing the problem of how these languages should be coded.

[5]It must be noted here, that it is necessary for these meanings to be stable and applicable to all languages included in the study, but not universally, as the semantic space can be partitioned differently in different languages and language families.

This is a case of inapplicable characters, which has no problem-free solution at least in a parsimony framework (Lee and Bryant, 1999; Maddison, 1993; Strong and Lipscomb, 1999).

Morphology

The smallest sound sequence that bears a meaning or has a function (such as the -ed suffix marking the past in English) is called a morpheme in linguistics. Morphology refers to how morphemes (roots and functional items) are combined to form words. Morphology has been a major source of characters traditionally in historical linguistics and reconstruction of morphological paradigms (such as noun declension suffixes) in the proto-language is common in well-studied families.

Functional or grammatical morphemes are similar to lexical items, since they too form cognate sets stemming from a protoform and are associated with a meaning or rather a function. Therefore, all coding methods that can be used for lexical characters can also be applied to grammatical morphemes. Another area of morphology that bears phylogenetic information is the patterning of morphemes within a morphological paradigm. For example, in modern Greek the nominative and accusative suffixes have become identical (or have merged) for neuter and feminine nouns, but not for masculine ones.[6] A different pattern of mergers is present in other Indo-European languages. Morphological characters of this nature are complex and idiosyncratic exactly like morphological characters in biology. In this case, the researcher needs to tailor the characters based on the study system with the goal to code as much of the variation as possible. Finally, another potential source for phylogenetically informative characters on the interface of morphology and the lexicon are cases of grammaticalization, the process by which a lexical item loses progressively its meaning and acquires a grammatical function (e.g. 'let us' becoming 'let's', a suggestion marker, in English).

Phonology

The sound systems of languages is another area traditionally used for inferring language relationships. Each distinct sound in a language's inventory is a phoneme. When a phoneme changes (such as [b] of ancient Greek becoming [v] in modern Greek), this change is typically without exceptions and affects every instance of the phoneme in the language, i.e. in all words that contain it. This way, like a protoform being inherited with modification and creating a cognate set, a protophoneme is inherited with modification creating a sound correspondence set. Sound correspondence sets are inferred by aligning cognate sets and isolating recurring 'columns'. Each sound correspondence set can be the basis of a multistate character. Also, the states can be ordered at least partially in most cases, as certain sound changes are more 'natural' than others. Given that sounds are largely shared across languages, it is possible

[6]Modern Greek has nouns in three grammatical genders, masculine, feminine and neuter.

to develop more sophisticated probabilistic models to simulate sound change that would be widely applicable.

2.4 Conclusion

Computational phylogenetic methods have revolutionized biological systematics and have a lot to offer in the field of historical linguistics, given the close analogies and the parallel history of the two fields. Even though there has been progress in this direction over the last decade, mainly with the work of Gray and Atkinson (Atkinson and Gray, 2005; Bouckaert et al., 2012; Gray and Atkinson, 2003), both the data and the methods used can and should be improved. In terms of data, the reliance only on lexical data is equivalent to 'scratching the surface', when languages can provide a wealth of other kinds of data, which incidentally are regarded even more trustworthy by historical linguists (Crowley and Bowern, 2010). In terms of coding methods, the commonly used binary recoding method of Gray and Atkinson (Gray and Jordan, 2000) violates the independence assumption of all phylogenetic methods. Finally, in terms of models of evolution, linguists have been reusing models devised for evolutionary biology, with the exception of the stochastic Dollo model (Alekseyenko, Lee, and Suchard, 2008). The development of coding techniques and evolutionary models for other kinds of linguistic characters beyond the lexicon is an exciting new direction for historical linguistics and for phylogeneticists towards a meaningful integration of the two fields.

Bibliography

Alekseyenko, Alexander V., Christopher J. Lee, and Marc A. Suchard (2008). "Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos". In: *Systematic Biology* 57.5, pp. 772–784.

Atkinson, Quentin D and Russell D Gray (Aug. 2005). "Curious parallels and curious connections–phylogenetic thinking in biology and historical linguistics." In: *Systematic Biology* 54.4, pp. 513–26.

Bouckaert, Remco et al. (2012). "Mapping the origins and expansion of the Indo-European language family." In: *Science* 337.6097, pp. 957–60.

Bowern, Claire and Quentin Atkinson (2012). "Computational phylogenetics and the internal structure of Pama-Nyungan". In: *Language* 88.4, pp. 817–845.

Bromham, L and D Penny (2003). "The modern molecular clock". In: *Nature Reviews Genetics* 4, pp. 216–224.

Campbell, L (2004). *Historical linguistics: An introduction*. Cambridge, Mass: MIT Press.

Crowley, T and C Bowern (2010). *An introduction to historical linguistics*. Oxford New York: Oxford University Press.

Darwin, Charles (1871). *The Descent of Man, and Selection in Relation to Sex*. New York: Appleton.

Gray, R D, A J Drummond, and S J Greenhill (2009). "Language phylogenies reveal expansion pulses and pauses in Pacific settlement." In: *Science* 323.5913, pp. 479–83.

Gray, R D and F M Jordan (2000). "Language trees support the express-train sequence of Austronesian expansion." In: *Nature* 405.6790, pp. 1052–5.

Gray, Russell D and Quentin D Atkinson (2003). "Language-tree divergence times support the Anatolian theory of Indo-European origin." In: *Nature* 426.6965, pp. 435–9.

Hennig, Willi (1950). *Grundzüge einer Theorie der phylogenetischen Systematik*. Berlin: Deutscher Zentralverlag.

— (1965). "Phylogenetic systematics". In: *Annual review of entomology* 10, pp. 97–116.

— (1966). *Phylogenetic systematics*. Urbana: University of Ilinois Press.

Hoenigswald, Henry and Linda Wiener (1987). *Biological metaphor and cladistic classification : an interdisciplinary perspective*. Philadelphia: University of Pennsylvania Press.

Hoijer, H (1956). "Lexicostatistics: a critique". In: *Language* 32.1, pp. 49–60.

Lee, DC and HN Bryant (1999). "A reconsideration of the coding of inapplicable characters: assumptions and problems". In: *Cladistics* 15.4, pp. 373–378.

Lipscomb, Diana L. (1992). "Parsimony, Homology and the analysis of multistate characters". In: *Cladistics* 8.1, pp. 45–65.

Maddison, WP (1993). "Missing data versus missing characters in phylogenetic analysis". In: *Systematic Biology* 42.4, pp. 576–571.

Mayr, E (1965). "Numerical phenetics and taxonomic theory". In: *Systematic Zoology* 14.2, pp. 73–97.

Michener, CD and RR Sokal (1957). "A quantitative approach to a problem in classification". In: *Evolution* 11.2, pp. 130–162.

Pimentel, Richard A. and Rhonda Riggins (1987). "The nature of Cladistic data". In: *Cladistics* 3.3, pp. 201–209.

Pleijel, Fredrik (2005). "On Character Coding for Phylogeny Reconstruction". In: *Cladistics* 11.3, pp. 309–315.

Queiroz, K De and J Gauthier (1992). "Phylogenetic taxonomy". In: *Annual review of ecology and systematics* 23, pp. 449–480.

Remane, Adolf (1971). *Die Grundlagen des natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik*. Koenigstein-Taunus: Otto Koeltz.

Simpson, G G (1964). "Numerical Taxonomy and Biological Classification." In: *Science* 144.3619, pp. 712–713.

Sneath, P. H. A. (1957). "The Application of Computers to Taxonomy". In: *Journal of General Microbiology* 17.1, pp. 201–226.

— (June 1971). "Numerical taxonomy: criticisms and critiques: Presidential address to the Systematics Association, November 1970". In: *Biological Journal of the Linnean Society* 3.2, pp. 147–157.

Sneath, P. H. A. and R. R. Sokal (1973). *Numerical taxonomy. The principles and practice of numerical classification.* San Francisco: W.H. Freeman.

Sneath, P. H. A. and Robert R. Sokal (Mar. 1962). "Numerical Taxonomy". In: *Nature* 193.4818, pp. 855–860.

Sokal, Robert R. and P. H. A. Sneath (1963). *Principles of numerical taxonomy*. San Francisco: W.H. Freeman.

Strong, EE and D Lipscomb (1999). "Character coding and inapplicable data". In: *Cladistics* 15, pp. 363–371.

Swadesh, M (1952). "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos". In: *Proceedings of the American philosophical society* 96.4, pp. 452–463.

— (1955). "Towards greater accuracy in lexicostatistic dating". In: *International journal of American linguistics* 21.2, pp. 121–137.

Chapter 3

A Bayesian phylogenetic classification of Tupí-Guaraní

3.1 Introduction

This paper presents an internal classification of the Tupí-Guaraní family based on the application of computational phylogenetic methods to lexical data from 29 languages of this geographically extended family, as well as from two non-Tupí-Guaraní Tupían languages, Mawé and Awetí, which serve as outgroups for the analysis.[1] This classification confirms aspects of previous influential classifications of the family, such as Rodrigues and Cabral (2002) and Rodrigues (1984/1985), but suggests a different organization for certain low-level subgroups. The greatest difference between the classification we propose and previous ones, however, lies in the structure at the base of Tupí-Guaraní (particulary the position of Kamaiurá and the Southern subgroup) as well as in its considerably more articulated subgrouping, resulting in a much less 'flat' tree.

The results in this paper are based on computational phylogenetic methods originally developed to infer evolutionary trees in evolutionary biology. In recent years, phylogenetic methods have been extended to the study of classification of natural human languages, and have been employed in the study of Austronesian (Gray, Drummond, and Greenhill, 2009b; Greenhill, Drummond, and Gray, 2010; Greenhill and Gray, 2005, 2009), Indo-European (Bouckaert et al., 2012a; Forster and Toth, 2003; Gray and Atkinson, 2003a; Nakhleh, Ringe, and Warnow, 2005; Ringe, Warnow, and Taylor, 2002; Warnow et al., 2004), and Pama-Nyungan (Bowern and Atkinson, 2012b) language families. However, with the exception of Walker and Ribeiro (2011) and Walker, Wichmann, et al. (2012), these methods have not been applied to any language family of South America, and this paper presents the first analysis of South American languages that relies on a character-based method, which are considerably more reliable than distance-based methods (see §Distance-based methods).

[1] This chapter is based on an unpublished manuscript by Natalia Chousou-Polydouri, Zachary O'Hagan, Keith Bartolomei, Erin Donnelly, Vivian Wauters, Sérgio Meira, and Lev Michael.

Also, in this study we employ an innovative coding scheme for lexical data that we argue better reflects cognate presence and absence (see §Cognate Sets and Character Coding).

The present paper results from work carried out by members the UC Berkeley Comparative Tupí-Guaraní Project, which began in 2010 with the aim of better understanding the position of Omagua – a Peruvian Amazonian language on which O'Hagan, Wauters, and Michael have conducted primary fieldwork – within Tupí-Guaraní.[2] It became apparent that a broader examination of relationships among Tupí-Guaraní languages would be necessary to adequately assess the position of Omagua in the family, leading us to develop a large comparative lexical database for comparative research. The phylogenetic analysis presented in this paper is intended as the first step in this broader project.

Tupí-Guaraní Languages

The Tupí-Guaraní family is a major subgroup of the Tupían stock (Dietrich, 2010), and constitutes one of the most geographically widespread genetic[3] groupings in South America (see Figure 3.1).

The family includes over forty recognized varieties, with the majority of Tupí-Guaraní languages found in Brazil, but with members of the family also found in Argentina, Bolivia, French Guiana, Paraguay, and Peru. Despite the geographical extent of the family and its large number of members, the Tupí-Guaraní family is generally believed to have relatively shallow time depth (2,000-3,000 years ago) (Noelli, 2008), with high degrees of lexical and grammatical similarity between varieties, especially in comparison to the larger Tupían stock of which it is a part. Some named varieties even exhibit significant mutual intelligibility, such as Tembé and Guajajara, and the 'Guaranian' varieties such as Mbyá, Kaiowá, and Ñandéva. Against the background of considerable similarity, it has been argued that several of these languages have undergone significant structural reorganization due to language contact, as in the case of Kokama-Kokamilla and Omagua (Cabral, 1995). Language contact has also been raised as a possible explanation for the divergent grammatical features of languages such as Aché (Rößler, 2008) and Xetá (Rodrigues, 1978).

Tupí-Guaraní peoples played a significant historical role in interactions between indigenous Amazonian peoples and European colonizers, being among the earliest and most numerous South American indigenous populations to make contact with Europeans (Métraux, 1927; Reeve, 1993). This early role led to Tupí-Guaraní languages being the focus of some of the earliest substantive linguistic descriptions and analyses of South American languages, es-

[2]Omagua, along with its sister Kokama-Kokamilla (Vallejos, 2010), has been significantly affected by intense pre-Columbian language contact in ways that are not attested in other Tupí-Guaraní languages (Cabral, 1995, 2007, 2011; Cabral and Rodrigues, 2003b; Michael, to appear; O'Hagan, 2011; O'Hagan, Michael, and Vallejos Yopán, 2013).

[3]In historical linguistics, the term 'genetic' refers to language groupings due to shared ancestry, also known as 'subgroups', as opposed to geographical or typological groups, which do not necessarily share a common ancestor.

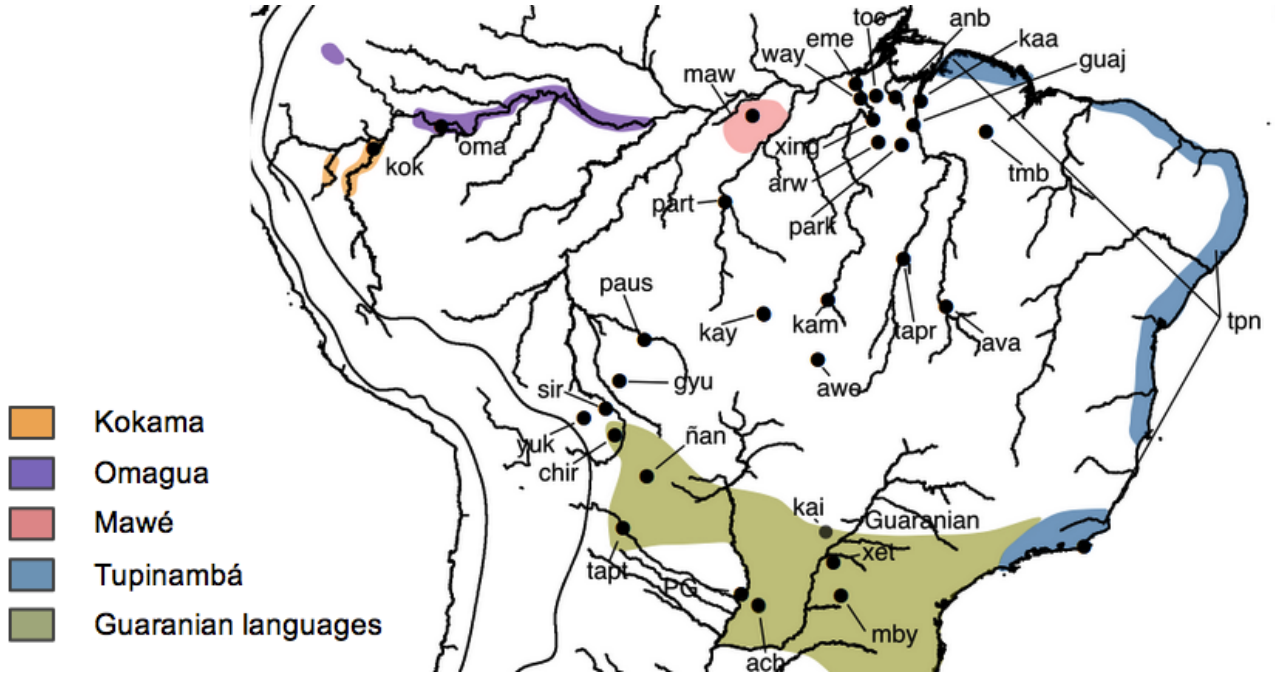

Figure 3.1: Earliest known distributions of Tupí-Guaraní languages. Shaded areas correspond to widespread languages. For language names abbreviations, see Table 3.3.

pecially in the case of Tupinambá (Anchieta, 1595; Faria, 1903; Figueira, 1687) and Guaraní (Ruíz de Montoya, 1640).

Tupinambá, the designation for a number of Tupí-Guaraní varieties spoken on and near the Atlantic coast of Brazil, went on to become an important language in colonial Brazil, especially in its creolized form, Língua Geral or Nheengatú. Tupinambá and Língua Geral spread west into the Amazon basin, becoming the principal medium of interethnic communication among indigenous populations, European colonists, and Catholic missionaries in the 18th and early 19th centuries, until its use was suppressed by the colonial government in the 19th century. Although use of Nheengatú has been dramatically curtailed, it is still spoken as a first language by some peoples of the upper Rio Negro (Cruz, 2011).

Guaraní, originally spoken in the greater Paraná River basin, was similarly instrumental in the work of Catholic missionaries beginning in the 17th century (Meliá, 1992), and the language has gone on to become one of the official languages of Paraguay.

In contrast to these politically and historically prominent Tupí-Guaraní varieties, other varieties remained undocumented well into the latter half of the 20th century, principally due to the (voluntary) isolation of their speakers. For example, the Yuki first entered into sustained contact with outsiders in 1965 (Stearman, 1986), the Araweté in 1976 (Viveiros de Castro, 1992), and the Zo'é in 1987 (Gallois, 1997).

Previous Classifications

Rodrigues (1958) provided the first modern classification of the Tupí-Guaraní family based on lexicostatistics, while Lemle (1971), using data from ten languages, reconstructed 221 protoforms and proposed a classification based on six sound changes. Rodrigues (1984/1985) proposed a classification of the family into eight subgroups, based principally on geographical criteria and proposed sound changes. This very influential classification did not propose any higher subgrouping, yielding a flat structure for the family. The eight subgroups proposed by Rodrigues, have, however, served as a reference point for most subsequent comparative work on the family.

Dietrich (1990), making use of quantitative distance measures based on grammatical features, was the first to argue for a more articulated structure, proposing two main branches, a southern and an Amazonian one. Rodrigues and Cabral (2002) revised Rodrigues' earlier classification, adding languages for which new data had become available, changing the position of certain languages in the classification, and proposing a major subgroup corresponding roughly to Dietrich's Amazonian branch. Rodrigues and Cabral (2002) constitutes the most recent internal classification of the family. We present a detailed comparison of our proposed classification with selected previous classifications in §Comparison with previous classifications.

Further historical work includes Jensen (1989) and Rodrigues and Dietrich (1997), which present additional reconstructed forms and Jensen (1998, 1999) and Schleicher (1998), which reconstruct aspects of Proto-Tupí-Guaraní morphosyntax. Mello (2000) reconstructs 761 protoforms, and provides maps of thirty lexical isoglosses, but does not argue for a classification based on the apparent sound changes he enumerates.

Tupí homeland and expansion

Historical linguistics have played an important role in the effort to uncover the history and culture of Tupí-Guaraní peoples, since most of their material culture is based on organic material which disintegrates rapidly in the tropical rainforest, limiting archaeologists to pottery. Since the nineteenth century there have been various speculations and hypotheses by ethnologists, archaeologists and historical linguists about the location of the Tupí homeland and the routes of expansion across South America (reviewed in Noelli (1998, 2008)). Some of them only refer to the homeland of the Tupí stock, and not to the Tupí-Guaraní family, while others to both. No hypothesis identifies a different homeland for Tupí than for Tupí-Guaraní. Below are the most recent and influential hypotheses:

- Lathrap's (1970) and Brochado's (1984) hypothesis that the Tupí stock originated along the Amazon river and its families radiated in Amazonia following tributaries of the Amazon River and especially the Madeira and its tributaries (Urban, 1996)
- Brochado's hypothesis about the common ancestor of Guaraní and Tupinambá originating along the main course of the Amazon and then Guaraní moving south along

the Madeira river and Tupinambá moving first east along the Amazon river and then south along the eastern coast of Brazil (Urban, 1996)

- Urban's hypothesis of successive waves of expansion from a homeland between the Madeira and Xingú Rivers. The first wave corresponds to the Tupí stock, the second to the Tupí-Guaraní family (Urban, 1996)

- Rodrigues's hypothesis of Tupí-Guaraní originating in Rondônia, between the Juruena and Arinos rivers and subsequently expanding in successive waves (Rodrigues, 2000)

The hypotheses mentioned above can be grouped in two categories: the Amazonian ones of Lathrap and Brochado based on archaeological data and the Rondônian ones of Urban and Rodrigues based on linguistic data (see Figure 3.2). Our results support an Amazonian homeland, between the lower Xingú and Tocantins rivers (see §Implications about the Tupí-Guaraní expansion).

3.2 Phylogenetic Inference

The internal classification of Tupí-Guaraní in this paper draws not on traditional comparative methods in linguistics, but on techniques of phylogenetic inference originally developed to infer evolutionary trees for biological organisms (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). In this section we present an overview of the theory and methods of phylogenetic inference and discuss how phylogenetic inference relates to standard methods in historical linguistics.

The term 'phylogenetic inference' covers a variety of methods used to infer phylogenetic trees, i.e. trees that represent evolutionary relationships among taxonomic units, such as species and biological families in biology, or linguistic varieties in linguistics. Phylogenetic methods and evolutionary trees can in principle be used for any phenomenon that is insightfully understood as centrally involving descent with modification from a common ancestor, as in the cases of linguistic and cultural evolution (Rogers and Ehrlich, 2008), as well as, of course, biological evolution.

Most phylogenetic methods are based on the comparison of *homologous* features of the entities studied, i.e. features found in those entities that are descended from an ancestral protofeature. Trees are then inferred by an algorithm that evaluates different possible tree structures usually with respect to an optimality criterion that relates the distribution of the homologous features of the entities studied to the structure of the trees under consideration. Trees that maximize or minimize the optimality criterion are selected by the algorithm as the optimal trees based on the information provided from the features included in the study. One common optimality criterion, parsimony, minimizes the number of innovations that must be posited to account for the attested distribution of homologous features.[4]

[4]For an in depth review of the different methods and algorithms, see Felsenstein (2004), Wiley and Lieberman (2011), and Lemey, Salemi, and Vandamme (2009).

Figure 3.2: Tupí-Guaraní homeland hypotheses: Lathrap's (1970) and Brochado's (1984) Amazonian hypothesis in dark orange, Urban's (1992) in light orange, Rodrigues's (2000) in red.

Characters, states and assumptions

The empirical basis of phylogenetic inference is the set of homologous features in the entities under study. These features are typically coded in a *character matrix*, a table in which

rows correspond to the entities being classified (e.g. biological species or languages), often called Operational Taxonomic Units (*OTUs*), while columns represent the *characters*, i.e. the homologous features compared across all the OTUs. Each character can take two or more values, such as presence or absence of a particular lexical root, or the reflexes of a protosound as, say, [k], [p], or [ʔ]. These different values are called character states or simply *states*.

Perhaps the most important step in a phylogenetic analysis, and usually the least discussed, is the choice of characters and how to code them. Good phylogenetic characters have the following characteristics:

They are convincing potential homologies. As explained above, features are considered homologous when they are instantiations of the same protofeature of a common ancestor. In practice, it is difficult to determine whether any two features are homologous, since we usually lack the historical records necessary to unambiguously establish the homology. Instead, we use intersecting criteria to establish *potential* homologies (Hennig, 1966), such as phonological similarity between morphemes, sound correspondences, and semantic relatedness. Each character in a character matrix is thus ultimately a hypothesis for homology between its states. Until evidence against homology emerges (e.g. through the results of phylogenetic inference, or through a more refined linguistic understanding of the forms between which homology is posited), it is accepted by default that all OTUs bearing the same state have a common ancestry, and that the common ancestor had a protofeature from which all observed states were derived.

They are heritable. A feature is heritable if it is transmitted vertically (at least primarily), from ancestor to offspring. The vast majority of phylogenetic methods infer trees and as such do not deal well with reticulated structures that result from the horizontal transmission of features (e.g. via linguistic borrowing). Features that can be shown to be transmitted very easily horizontally, or suspected of being so, should not be used as the basis for phylogenetic characters, since they would be misleading if they are interpreted as stemming from a common ancestor. That said, phylogenetic methods can tolerate horizontal transmittance or borrowing to a degree before being misled by it and are capable of discovering instances of horizontal transmittance, since the evolutionary history of a borrowed feature does not match the tree structure of the preponderance of other features.

They are independent. A set of characters is said to be independent when the states of any one character are not causally related to the states of any other characters, i.e. it is not possible to predict the states of one character on the basis of the states of others. The independence assumption is a key assumption of all phylogenetic methods and lies at the heart of evolutionary biology and historical linguistics. It is the 'coincidence' of many independent changes that points to languages having a common source and partially common history. If there is another causal link, different from shared history, that unites two or more

characters, then they are in fact *one* historical event, and they should count as one unit of evidence for evolutionary relationships.

They have discrete states. The computational implementation of most phylogenetic methods requires that states be discrete. Thus, characters with continuous-valued states that cannot be binned in a principled manner in discrete states must be avoided, even if they may be otherwise acceptable characters.

They are moderately variable. In order for a character to contribute usefully to the inference of a resolved tree, its states should vary across the OTUs in the dataset. At the same time, the states should not be different for all OTUs. Constant or invariable characters (i.e. characters for which all the OTUs bear the same state) or autapomorphic characters (i.e. characters for which only one OTU or all OTUs have a different state) are completely uninformative for parsimony methods and minimally inform the topology of the tree when using model-based methods.[5]

The relatedness assumption

All phylogenetic methods are based on the assumption that all the OTUs in the analysis are in fact related by descent from a common ancestor. In other words, phylogenetic methods cannot tell us *if* taxonomic units are related to each other, but *how* they are related. While this is not typically a problematic assumption for the inference of evolutionary trees in biology, where there is consensus that life originated only once, the situation is not as simple in historical linguistics. Once languages are included in an analysis and features are coded across all of them, the researcher is willing to accept common ancestry as the default explanation.

Overview of Phylogenetic Methods

If the character matrix is the empirical basis of computational phylogenetic methods, the phylogenetic inference method is its analytical basis. There are two broad types of phylogenetic methods: distance-based methods and character-based methods. We will describe the main characteristics of each type and discuss their most important differences, and conclude by arguing that character-based methods are the only suitable ones to infer evolutionary trees.

Distance-based methods

Distance-based methods start with a character matrix and compute pairwise 'distances' between OTUs on the basis of the matrix. Every pair of OTUs is thus associated with a

[5]Parsimony-uninformative characters are informative for model-based methods regarding the estimation of model parameters and branch lengths and should not be excluded in such analyses.

number, which is intended to correspond to the similarity between them. These distances are then used to group the languages in a rooted or unrooted tree, or in a network, according to some criterion (e.g. minimum evolution) that serves as the basis of the clustering algorithm. There are a variety of ways to compute the distances in question and many different clustering algorithms for generating trees, but all distance-based methods share the following characteristics:

Grouping according to overall similarity Distance-based methods by design group OTUs according to their overall similarity and not necessarily according to their evolutionary relationships. In particular, distance-based methods neutralize the difference between two different types of similarity found in branching evolutionary processes: shared innovations and shared retentions. Only the former are relevant for reconstructing the evolutionary relationships between entities. However, the distinction between shared innovations and shared retentions cannot be made with distance-based methods because both types of similarity 'count' towards the computed distances.

And in terms of the utility of overall similarity as a measure of relatedness, we can note that although we expect sister languages to be similar, the reverse is not always true (i.e. when languages share a great number of features this doesn't mean necessarily that they are sisters).

Data Reduction The first step in all distance-based methods is the reduction and simplification of collected data in a few numbers, the pairwise distances. This way the distribution of character states for each character is not maintained. This procedure obscures the complexity of, and potentially conflicting signals in, the data, and arguably even discards much of the researchers' hard work in collecting the data.

Speed The last common characteristic of distance-based methods, and maybe their only advantage, is their speed. Due to the data reduction procedure and their one-pass algorithms (without cycles of permutations and optimizations), distance-based methods are extremely fast, which made them popular in the early days of computers and now when dealing with huge datasets.

At this point, one may rightfully wonder why anybody interested in evolution would use distance-based methods at all, and whether they should even be called phylogenetic methods if they do not group according to phylogeny. It turns out that in many cases their results are reasonable (i.e. when they are not misled by shared retentions), making them the 'quick and dirty' solution. Unfortunately, it is not possible to tell if they are misled or not without running a character-based analysis (more on this below), making their results untrustworthy. Of course, distance-based methods have a place in other fields of study as clustering methods. They are also useful as starting points for character-based methods (e.g. a Neighbor-Joining tree as a starting tree in a Bayesian MCMC search). Distance-based methods include Unweighted Pair Group Method with Arithmetic Mean

(UPGMA)(Sokal and Sneath, 1963), Neighbor-Joining (NJ) (Saitou and Nei, 1987), and lexicostatistics (Swadesh, 1952). Most of the popular phylogenetic network methods, such as Split Decomposition (Bandelt and Dress, 1992) and NeighborNet (Bryant and Moulton, 2004), are also distance-based methods, but they infer networks instead of trees.

Character-based Methods

The common characteristic of all character-based methods is that they take as input the character matrix itself, without calculating pairwise distances, and evaluate how well a range of possible trees fit the character states of the character matrix according to an optimality criterion, such parsimony or likelihood. A total score for each evaluated tree is computed on the basis of how well the tree accounts for the distribution of states for each character. Depending on the method, the algorithm may search for the most optimal tree(s) (parsimony and maximum likelihood), or collect a sample of trees according to their posterior probability (Bayesian inference). Among character-based methods, parsimony doesn't need an explicit model of evolution, while maximum likelihood and Bayesian inference do. Character-based methods include parsimony, maximum likelihood, Bayesian inference, and the standard linguistic comparative method (although the latter is not a computational method).[6]

Rooting trees Actual evolutionary trees are intrinsically rooted, in the sense that the direction in which time flows along branches between nodes is determined by the position of the branch relative to the root node. In order for trees resulting from phylogenetic methods to be interpretable, they likewise need to be rooted. When evolutionary trees are created by hand, such as through the comparative method, a common way to root them is by determining for each character which state is the ancestral (Crowley and Bowern, 2010), a process known as 'character polarization' in the phylogenetic literature. In sound change, for example, this is accomplished by accumulated knowledge regarding tendencies in the directionality of sound change. However, character polarization may not be feasible in many cases, e.g. in instances of lexical replacement, and moreover, most phylogenetic algorithms operate on unrooted trees, which increases the speed of the algorithms, but requires an *a posteriori* rooting procedure.

The most common *a posteriori* rooting technique is the *outgroup method*, which relies on *a priori* knowledge of an 'outgroup' OTU, namely an OTU that is known to be closely related to the group of OTUs being studied (the 'ingroup'), but is nevertheless outside that group. The outgroup is included in the phylogenetic analysis and the inferred tree(s) are rooted where the outgroup joins the rest of the tree. Even though it is possible to root a tree with only one outgroup OTU, ideally multiple outgroup OTUs should be included and the most distantly related one be used for rooting. This ensures that the monophyly of the ingroup is tested and also makes the ancestral state optimized for the root more reliable.

[6]Hennig's cladistic method (Hennig, 1966) is surprisingly similar to the comparative method and was used in systematic biology before the current computational methods. Although the comparative method predates Hennig by at least 70 years, it is unlikely that Hennig knew about it (Atkinson and Gray, 2005).

Choosing an appropriate outgroup is not trivial and is crucial for the analysis. An ideal outgroup is as closely related to the ingroup as possible, but without being part of it.

Parsimony Methods Parsimony methods seek to minimize the number of innovations necessary to explain the distribution of character states.[7] In a parsimony phylogenetic analysis, all characters are optimized individually on a given tree and the number of innovations is counted. A search in the tree space is then conducted for the tree that accounts for the entirety of the character matrix with the least number of evolutionary steps. Due to the vast number of possible trees once the number of OTUs exceed a small number (around 25 now), exhaustive searching of the tree space is often computationally unfeasible, and heuristic algorithms are instead employed to effectively sample the space to find the most parsimonious tree(s).

Parsimony methods are not entirely unproblematic, however, leading to development of model-based phylogenetic approaches, discussed in the next section. Felsenstein (1978) showed that under certain circumstances, parsimony can fail to recover the true tree due to *long branch attraction*. This phenomenon occurs when the true tree exhibits long branches which can accumulate a large number of parallel changes (i.e. changes that are not shared innovations). Parsimony favors reconstructing such changes as shared innovations, erroneously joining the long branches together. Long branch attraction is exacerbated when the rate of evolution is high and the number of possible character states small (e.g. binary characters or DNA sequences, which have only 4 possible states). The influx of molecular sequence data in evolutionary biology motivated the development of explicit models of evolution and likelihood-based methods that are not as susceptible to long branch attraction.

Model-based approaches There are two types of model-based phylogenetic approaches: maximum likelihood and Bayesian inference. The main difference between these two approaches is that maximum likelihood provides a 'point estimate' of the phylogeny and the model parameters (i.e. the single value that maximized the likelihood of the data, an optimality criterion), while Bayesian inference produces a posterior distribution of values for each model parameter and the topology (i.e. it accounts for uncertainty in the estimated values). The result of a Bayesian analysis is thus not a single tree, but a sample of the posterior distribution of trees, in which each tree topology is represented in proportion to its posterior probability. Bayesian inference also allows the incorporation of previous beliefs and estimates for the model parameters as prior distributions in the analysis. Below we focus on Bayesian inference, which has emerged as the method of choice for phylogenetic analyses of lexical data.

Bayes Theorem Bayesian phylogenetic inference is a special case of the more general Bayesian approach to model evaluation, which seeks to answer the question: *what is the*

[7] The comparative method is also a parsimony method which uses character polarization for rooting.

probability of a model, given the data. Bayesian analyses provide a value for the probability – the *posterior probability* – that the model is correct, after the data have been taken into account.[8] The posterior probability (i.e. $P(model|data)$) is calculated using the Bayes theorem, given in (3.1).

$$P(model|data) = P(data|model)\frac{P(model)}{P(data)} \tag{3.1}$$

The likelihood of the data under the model (i.e. $P(data|model)$), or simply 'likelihood', is the probability that the observed data would be produced by the hypothesized model. The prior probability of the model (i.e. $P(model)$) are the various prior distributions we furnish the analysis with, which we discuss in §Models of Evolution: Parameters and Priors, below. The denominator of the formula is the probability of the data integrated over all possible parameter values, and is impossible to calculate for a phylogenetic analysis. This is why we use the Metropolis-Hastings algorithm, a Markov Chain Monte Carlo (MCMC) method, to approximate the posterior probability distribution.[9]

Models of Evolution: Parameters and Priors Models of evolution simulate how character states change, and thus necessarily presuppose particular properties of the character whose evolution they seek to model. The most commonly used models for linguistic cognate evolution are the generalized binary model[10] and the stochastic Dollo model (Alekseyenko, Lee, and Suchard, 2008).[11] Both of these models simulate the evolution of *binary* characters and are thus suitable for modeling, for example, characters based on the presence or absence of a particular cognate. With the latter application in mind, the main difference between the two models is that the Dollo model assumes that each cognate originates only once on a tree, while the generalized binary model allows cognates to originate more than once.

For the assumption that the Dollo model is based on to be valid, it would be necessary to identify and remove all intra-family loans, identify and remove all inter-family loans that occur more than once in the family, and also identify all reflexes of a given protoform. Especially in understudied families like Tupí-Guaraní, the Dollo model represents an extremely strong assumption, leading us to adopt a generalized binary model instead.

The generalized binary model simulates characters with two states – present (represented by 1) and absent (represented by 0) – and permits different rates of change between the two

[8] In phylogenetic inference, the model includes a variety of parameters: the tree topology, the branch lengths, the transition rate matrix, the stationary probabilities, rate variation among sites parameters, etc. All these parameters at the end of the analysis have an associated *posterior probability distribution*, i.e. a distribution which shows the posterior probability over a range of values.

[9] The Metropolis-Hastings algorithm avoids the calculation of the probability of the data because it is based on ratios of posterior probabilities of the model. The probability of the data is the same for any permutation of the model parameters because it is integrated over all possible parameter values.

[10] In MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), the phylogenetic software we use in this study, this type of model is called a *restriction site model*.

[11] The latter is implemented in BEAST (Drummond, Suchard, et al., 2012; Drummond and Rambaut, 2007), a phylogenetic application that has been used in several phylogenetic linguistic studies.

states. The latter is very important for modeling cognate evolution, as independent gains of the same cognate are expected to be very rare,[12] while independent losses of a certain cognate are comparatively common. The binary model has only one free parameter, the stationary probabilities of the states.[13]

In addition to the basic parameter of the binary model, one can model rate variation across cognate sets to allow some cognate sets to evolve faster than others (usually this is modeled as gamma distributed rates). Finally, every phylogenetic analysis includes the parameters of the tree, which are the topology and branch lengths.

Once a suitable model of evolution is chosen, the other necessary step in a Bayesian analysis is the choice of prior distributions (or simply 'priors') for all the parameters of the model. The priors represent our expectations for the values of each parameter *before* we look at any of the data that will be included in the analysis, and can be based on intuitions and prior analyses (but of different datasets, so as to avoid circularity). In cases where there is not much prior information or strong expectations, one can use uninformative or 'flat' priors that place more or less the same probability across many different values for each parameter. Bayesian inference allows the data to 'swamp the prior' if there is overwhelming evidence in a different direction.

Metropolis-Hastings algorithm (MCMC) The Metropolis-Hastings algorithm is an approximation technique used in Bayesian inference in order to estimate a parameter's posterior probability distribution, in our case the tree topology, the branch lengths and the model parameters. The approximation is based on the posterior sample of trees and parameters (typically the sample includes tens of thousands of values in order for the estimation to be trustworthy), in which every tree is sampled in proportion to its posterior probability. The samples are collected by MCMC 'chains' which are run typically for tens of millions of iterations, called 'generations'. The steps of each generation as well as the initialization of a chain are described below:

1. Initialization: All parameters of the model are initiated at some value (from their prior distributions), i.e. a tree topology is picked, values for the branch lengths, etc. Then the likelihood of the data and the prior probability of the model is calculated based on this combination of parameters. This is the current state of the chain, $\mathrm{S}_{current}$.

2. Proposal: A single parameter, such as the tree topology or the value of the cognate gain parameter, is chosen to be modified and the likelihood and the prior probability of the model are re-calculated. This is the new state of the chain, S_{new}.

[12] In practice, independent gains would be instances of borrowing or mistaken cognancy decisions.

[13] The stationary probabilities are the proportion of each state in a number of cognate sets when they have evolved for an infinite amount of time. In practical terms, the stationary probabilities are proportional to the rates of cognate gain (stationary probability of state 1) and cognate loss (stationary probability of state 0).

3. Acceptance or Rejection of S_{new}: The ratio of the two values, $S_{new}/S_{current}$, is calculated. If $S_{new}/S_{current} > 1$, i.e. if the posterior probability of S_{new} is greater than that of $S_{current}$, the chain accepts the S_{new} and it becomes the new $S_{current}$. This behavior ensures that the chain always moves towards areas of higher posterior probability. If $S_{new}/S_{current} < 1$, the chain *may* accept the new state, with $p = S_{new}/S_{current}$. This behavior allows the chain to avoid being trapped at local maxima, permitting it to move across areas of low posterior probability.[14] The bigger the drop in probability, the more likely it is that the chain will reject the new state, meaning that the chain will tend to stay in (and return large numbers of samples from) regions of high posterior probability, while only occasionally visiting regions of low posterior probability. If the new state is rejected, then the current state of the chain remains unchanged. In any case, the chain returns to step 2 for a new generation.

As the chain is running, we collect 'samples' at regular intervals, e.g every 1,000 generations[15], where each sample is a snapshot of the current state of the chain, i.e. a tree with particular branch lengths and parameter values.

Stationarity and Convergence In order for a Bayesian analysis to be trustworthy, it is necessary that the MCMC has run long enough that the posterior probability distribution is adequately approximated with the collected samples, i.e. that the chain has made it into the region of the tree space in which $P(model|data)$ is high in comparison to other regions. We infer that MCMC has run long enough when the chains exhibit both stationarity and convergence.

An MCMC chain has reached *stationarity* when the sampled distribution has stabilized (i.e. when the chain mostly remains in a given region of tree space). Because chains start at arbitrary points, the initial portion of the chain trajectory and the associated samples are not representative of the posterior probability distribution. These samples are discarded as 'burn-in'. The issue of *convergence* arises from the fact that it is typical to run multiple independent chains to explore the often very complex tree spaces associated with phylogenetic problems. If the chains converge to the same region of tree space, i.e. if they are all sampling from the same distribution, we infer that the algorithm has been successful, and we are able to combine and summarize the samples from the different chains.

[14] An additional technique to avoid being trapped in local maxima, Metropolis Coupled MCMC (or MC3), involves two different kinds of simultaneous chains: cold chains and hot chains. In this approach, parallel to the sampling chain (= cold chain), the algorithm runs a number of exploratory hot chains, which have higher acceptance probabilities (i.e. they accept steps towards regions of lower posterior probability more often), and thus traverse the tree space more rapidly. At regular intervals, the current states across cold and hot chains are compared, and if a hot chain is at a region of higher posterior probability than the cold chain, then the hot chain in question becomes the cold chain.

[15] This is done in order for the samples to be independent, because consecutive samples tend to be autocorrelated.

Summary of results The usual way to summarize the results of an MCMC run is to integrate the posterior probability of every parameter over all the possible values of the other parameters. For parameters with numerical values, we usually calculate estimated values and 95% credibility intervals. In terms of tree topologies, common ways include the construction of a majority-rule consensus tree from the sample (i.e. integrating how often a certain subgroup is found in the sampled trees) or a maximum clade credibility tree (i.e. the tree from the sample that has the maximum sum of subgroup posterior probabilities).

Conclusion

As explained in the preceding section, character-based phylogenetic methods differ fundamentally from lexicostatistics and are philosophically closer to the comparative method, despite their computational nature. We are convinced that phylogenetic methods can be very useful for historical linguistics, provided that both characters and methods of analysis are chosen properly.

3.3 Dataset

Languages

The analysis and classification of the Tupí-Guaraní family presented in this paper is based on data from 30 Tupí-Guaraní languages and 2 non-Tupí-Guaraní Tupían languages, Awetí and Mawé, which serve as outgroups for the phylogenetic analysis (see §Rooting trees). Our dataset includes almost all languages with relatively extensive lexical sources and spans all eight traditionally recognized subgroups (Rodrigues, 1984/1985; Rodrigues and Cabral, 2002). The sources from which lexical data were drawn include published grammars and dictionaries, wordlists, and articles, and unpublished fieldwork data of the authors and generous colleagues. As might be expected, the sources for different languages differed in their total comprehensiveness, yielding different degrees of lexical coverage (i.e. percentage of comparative wordlist meanings for which we could find a corresponding item). The mean lexical coverage for the entire dataset was 71%, with a maximum of 98% (Tembé) and a minumum of 20% (Ñandéva). Our dataset initially included two more languages, Apiaká and Turiwará with lexical coverage below 20%, which are excluded for the main analyses presented here (see §full dataset results and exclusion of 'wildcards'). The languages included in our dataset, the corresponding lexical sources, and the lexical coverage percentages are given in Table 3.3.

Table 3.1: Languages included in the dataset.

Language	abbr.	% Coverage	Source(s)

Aché	ach	89.4	Heckart and Hill (2007); Hill and Hawkes (1983); Rößler (2008); Rößler (pers. comm.)
Paraguayan Guaraní	PG	97.0	Guasch (2003)
Kaiowá	kai	43.6	Bridgeman (1961); Cardoso (2008); Harrison (1971); Taylor (1984a,b)
Mbyá	mby	85.1	Dooley (1998, 2006)
Ñandéva	ñan	22.7	Costa (2002, 2007); Dooley (1991)
Xetá	xet	32.1	Vasconcelos (2008)
Tapiete	tapt	89.8	González (2005, 2008)
Chiriguano	chir	84.7	Dietrich (2007)
Guarayú	gyu	88.7	Armoye Urarepia (2009); Höller (1932)
Siriono	sir	87.8	Priest and Priest (1985)
Yuki	yuk	86.1	Garland (1978); Villafañe (2004)
Pauserna	paus	60.3	Horn Fitz Gibbon (1955); Riester (1972)
Omagua	oma	93.0	Gilii (1782); Girard (1958); Hervás y Panduro (1787); Marcoy (1866); Martius (1867); Uriarte ([1776]1986); O'Hagan (pers. comm.)
Kokama-Kokamilla	kok	92.6	Espinosa Pérez (1989); Faust (1959, 1972); Vallejos (2010); Vallejos (pers. comm.); O'Hagan (pers. comm.)
Tupinambá	tpn	96.4	Lemos Barbosa (1951, 1970)
Tapirapé	tapr	75.7	Almeida, Jesus, and Paula (1983); Praça (2007)
Asuriní do Tocantins	toc	87.2	Cabral and Rodrigues (2003a); Harrison (1963, 1975); Nicholson (1978, 1982)
Parakanã	park	80.6	Silva (2003)
Avá-Canoeiro	ava	52.3	Borges (2006, 2007)
Tembé	tmb	99.3	Adam (1896); Boudin (1978); Caetano (n.d.); Montoya (1876); Restivo (1893); Meira (pers. comm.)
Turiwará	twt	11.3	Nimuendajú (1914)
Anambé	anb	34.1	Silva Julião (2005)
Araweté	arw	54.5	Solano (2009); Viveiros de Castro (1992)
Asuriní do Xingú	xing	54.4	Nicholson (1978, 1982); Pereira (2009)

Kayabí	kay	64.7	Borges e Souza (2004); Dobson (1973, 1988, 1997)
Apiaká	api	16.2	Padua (2007)
Parintintin	part	88.8	Betts (1981); Pease (1968); Sampaio (1997)
Kamaiurá	kam	78.5	Drude (2011); Seki (1982, 1983, 1987, 1990, 2000, 2007, 2010)
Wayampí	way	91.2	Grenand (1989); Olson (1978)
Emérillon	eme	81.3	Couchili, Maurel, and Queixalós (2001); Gordon and Rose (2006); Queixalós (2001); Rose (2002, 2003, 2008, 2009); Rose (pers. comm.)
Ka'ápor	kaa	88.4	Caldas (2009); Kakumasu and Kakumasu (1988); Lopes (2009)
Guajá	guaj	48.6	Cunha (1987); Magalhães (2006, 2007); Nascimento (2008)
Awetí	awe	79.3	Corrêa da Silva (2010); Drude (2006, 2008, 2011); Drude (pers. comm.)
Mawé	maw	79.8	Corrêa da Silva (2010); Drude (2006); Franceschini (1999); Meira (pers. comm.)

Low-coverage Languages

Due to the uneven documentation of the different languages, our dataset varies considerably in terms of lexical coverage. Some researchers have excluded OTUs with large percentages of missing data using 50% as a cutoff (Bowern and Atkinson, 2012a). However, simulations and empirical studies including fossils and other highly incomplete OTUs have shown that, although highly incomplete taxa can be associated sometimes with poorly resolved consensus trees and decreased phylogenetic accuracy, this is not always the case. On the contrary, in many cases highly incomplete OTUs can be placed on phylogenetic trees with accuracy and they even have the potential to increase tree resolution in some cases (Wiens, 2003; Wiens and Morrill, 2011). Therefore, we chose not to exclude a priori any language based solely on the percentage of missing data.

Comparative List

This study was based on an extensive comparative wordlist that includes numerals up to five, body parts, animals and plants common to the Amazon basin, kinship terms, natural features and phenomena, manmade items, and basic adjectives and verbs, keeping in mind culturally and areally appropriate practices likely to be represented in lexical sources (see Appendix

A). After certain meanings were excluded due to pervasive problems in extracting them from published sources, 500 meanings remained for purposes of initial cognate set construction. The comparative wordlist was subsequently increased to 543 meanings in order to represent in an unbiased way cognate sets whose 'central' meaning was outside the initial wordlist (see §Additional Meanings).

In an effort to maximize the probability of finding cognate words in the lexical sources, synonymous and very near synonymous terms were also kept in mind when searching lexical sources, and forms bearing these (near) synonymous meanings were also added to the dataset (with annotations of the meanings attributed to them). In addition, we searched for forms based on the expected form of the item in the given language, making use of published sound correspondences, and our own understandings of sound correspondences in the dataset, to predict the approximate form we expected for the cognate. In the cases of forms demonstrably borrowed from Portuguese, French, Quechua and other non Tupí-Guaraní languages, the words were marked as a loan and were coded accordingly (see §Loans).

Lexical Sources

We used a variety of sources in order to collect as many lexical items as possible for each language. Among them are dictionaries, grammars, phonological descriptions, and unpublished material from many collaborators. The sources used for each language are listed in Table 3.3.

3.4 Cognate Sets and Character Coding

We employed an innovative coding method, which more closely approximates cognate presence and absence in the languages studied, than the commonly used method of Gray and Atkinson (2003b). Unlike the latter, where in order for a cognate to be considered present it has to bear precisely the meaning or gloss given in the comparative lexical list used to collect the data, we include in our cognate sets all reflexes we can find, irrespective of meaning. Our motivation for this methodological difference is twofold: first, the Gray and Atkinson method does not accurately represent the presence and absence of particular cognates, with the result that the binary recoded characters are not independent, a violation of a key assumption of all phylogenetic methods; second, due to the size of our wordlist and our data collection method we were able to detect many more cognates that had undergone semantic shift, than if we were collecting words based on a typical Swadesh list. We will explain below our coding method in detail, as well as its relationship with and differences from the Gray and Atkinson method and cognate presence-absence coding.

The Gray and Atkinson method

The Gray and Atkinson method of coding lexical characters for phylogenetic analyses was first introduced in 2003 (Gray and Atkinson, 2003b) and it has been widely used in subsequent studies (Bouckaert et al., 2012b; Bowern and Atkinson, 2012a; Gray, Drummond, and Greenhill, 2009a). The method consists of 3 steps:

1. Collection of lexical items using a wordlist. Typically a Swadesh list of 100 or 200 items is used, or occasionally a longer list of 'basic' meanings. Typically one item per meaning per language is collected.

2. Organization of items in cognate sets *within* each meaning. Items bearing the same meaning across all languages studied are examined and potential cognates are identified, resulting in a number of different cognate sets for each meaning. This step results essentially in multistate semantic characters (every meaning is a character and the cognate sets within each meaning are its states).

3. Binary Recoding. In the last step, every state of the previously constructed multistate characters becomes itself a binary presence-absence character. Crucially each characters codes the presence (coded with 1) or absence (coded with 0) of a particular cognate in a particular meaning, i.e. a form-meaning association.

The Gray and Atkinson method has the advantage of producing characters that can be easily modeled. Model-based methods require states to be comparable across characters in order to apply the same evolutionary model parameters to all of them. At the time the Gray and Atkinson method was introduced, the only way to analyze generic multistate characters (as opposed to binary characters and molecular DNA characters) was with parsimony.[16] By converting multistate into binary characters, the Gray and Atkinson method allowed their analysis by standard Bayesian Inference phylogenetic software.

However, the Gray and Atkinson method has a series of drawbacks. First of all, at the data collection step, the criteria with which only one form per meaning is chosen among synonyms or near-synonyms are not clear. But perhaps the most problematic point in the procedure is the binary recoding step. As mentioned before, one of the key assumptions of all phylogenetic methods, including the comparative method, is that the characters used are independent. This assumption is violated with the binary recoding step, as all binary 'cognate' characters produced from one multistate 'semantic' character are interdependent (i.e. when a languge has 1 for a binary character it necessarily has 0s for the rest of the characters stemming from the same initial multistate character). In the simple case of two cognate sets per meaning, we have complete dependence of the resulting binary characters. In cases where there are more than two cognate sets per meaning, we have more complex patterns of dependence. Binary recoding has been extensively criticized in evolutionary

[16]Currently, there are Bayesian phylogenetic methods that can handle generic multistate characters even if their states are not strictly comparable across characters (Lewis, 2001; Nylander et al., 2004).

biology literature (Buth, 1984; Farris, 1983; Meier, 1994; Murphy, 1993). Unfortunately, the effects of large scale non-independence on the results of an analysis are not known through simulation studies, when using methods other than parsimony. One logical consequence is the artificial inflation of support values, since the same change is 'counted' twice. For example, in the case of 2 cognate sets for meaningA, the subgroup receives support because of the shared loss of the association meaningA-cognate1 and the shared gain of the association meaningA-cognate2, while this is in fact one substitution event and not two independent events.

Cognate presence-absence coding

Cognate presence-absence coding, i.e. the concept of basing one's characters on a protoform and tracking the presence of its reflexes in daughter languages, is very appealing and intuitive for linguists. Indeed, vertical passage of forms from parent to daughter languages is the closest analogue to genetic inheritance in biology. Real cognate presence-absence characters (i.e. not form-meaning associations) are also largely independent, as a cognate can be present in a language but associated with a different meaning than in the other languages (e.g. English has a cognate to the German 'hund', namely 'hound', it just happens that it doesn't mean 'dog' as in German, but rather a specific kind of dog). The biggest obstacle in implementing cognate presence-absence coding is knowing when a cognate is really absent.[17] To be reasonably sure, a researcher needs access to the full lexicon of a language, which is difficult for most non-intensively studied language families, and impossible in most cases of extinct languages.

Although superficially similar, the binary characters resulting from the Gray and Atkinson method do *not* represent cognate presence or absence because they have as a prerequisite that the meaning remains the same (i.e. they by definition exclude any instance of a root that bears a different meaning and consider that root absent). In other words, the 0s of the binary matrix conflate real absences and instances of semantic shift. Our coding method, described below, is an attempt to minimize this conflation by uncovering as many instances of semantic shift as possible in order to better approximate cognate presence-absence coding.

Our Coding Method: Quasi cognate coding

After the initial data collection employing the 500-long wordlist, we constructed cognate sets within each meaning and we performed binary recoding of each cognate set according to the Gray and Atkinson method. This step resulted in cognate sets organized in semantic groups. If a meaning was not found in some language's sources, all cognate sets in that semantic group were coded as unknown (?). As explained above, at this stage, many of the absences (coded with 0) in our matrix are not real absences, but the consequence of semantic shift. In

[17] Coding all apparent absences as unknowns is not an option, as this will yield phylogenetically uninformative characters.

order to minimize the amount of false absences and therefore the amout of non-independence of our characters, we inspected all cognate sets in many iterations and made the following adjustments:

Consolidation We consolidated cognate sets based on the same root but originally present in two or more semantic groups (due to semantic shift). The cognate set resulting from the consolidation was included in the semantic group of its 'central' meaning.[18]

Compounds During the construction and subsequent inspection of cognate sets, we found a large number of compound or otherwise complex words (e.g. nominalized or otherwise morphologically complex forms). We will henceforth refer to both these categories of words as compounds, as we treated them in the same way.

We distinguished two types of compounds, 'genetic' and 'potentially independent' compounds. Genetic compounds include cases where the compounding or derivation could be argued to have happened once in the past and then been inherited as a unit in the daughter languages. Potentially independent compounds include cases where the compounding or derivation cound have happened multiple times independently at least in some of the languages. We used the following criteria to distinguish genetic compounds:

1. The meaning of the compound is unpredictable (e.g. 'land' + 'white' = 'cloud').
2. The meaning of the compound is predictable, but the reflexes of the compound show evidence of erosion (e.g. the *wɨrapar* compound for bow, instead of the uneroded form *ɨwɨrapar*).
3. The meanings of the compound constituents cannot be identified and the compound is widely distributed in many languages of our dataset.
4. Singleton compounds (i.e. found in only one language) are by definition genetic.

We formed cognate sets based on genetic compounds and we included them as characters in the analysis. For example, the word for 'star' in most languages is a compound of the words for 'moon' and 'fire' (e.g. Tupinambá (ton) *jasɨ* 'moon', *(t)atá* 'fire', *jasɨtatá* 'star'). As such compounding is not likely to have happened many times independently, we included a compound character in the semantic group 'star' based on this compound.[19]

[18] Among all the meanings present in a cognate set, we call 'central' the meaning from which all other meanings could be more easily derived. The 'central' meaning is not necessarily the most common meaning, nor is it a claim for the root's protomeaning. In cases where no central meaning could be identified, we used the most common meaning instead.

[19] In order for an item to be included in a compound cognate set, all its constituents need to be cognate and in the same order, reflecting that the compound was formed only once and then inherited as a unit. In other words, our compound cognate sets are *not* coding a common compounding strategy.

Conversely, the word for 'broom' is very often a complex form of the verb 'sweep' (ton *p*eʔiɾ) and an instrumental nominalizer (ton -*s*aβ), i.e., *t*upeʔisaβ). As this innovation could happen multiple times independently, we did not create a compound cognate set based on this complex form.

Both genetic and potentially independent compounds were taken as evidence for the presence of their constituent roots in the corresponding language. For example, if a language didn't have a reflex of the root *p*eʔiɾ for 'sweep' but the word for 'broom' was derived from the root *p*eʔiɾ, the word for broom was taken into account as evidence for the presence of the root *p*eʔiɾ in this language.

Loans We were able to identify loans from other languages into Tupí-Guaraní (Quechua, Portuguese, Spanish, French), as well as a few from Tupí-Guaraní languages into the outgroup language Mawé. Loanwords were coded as 'singleton' (apomorphic) characters, i.e., characters that were coded as present only for one language. For example, Ka'ápor, Tapiete, Omagua, and Kokama have all borrowed the word for 'mother' from Spanish or Portuguese. As all these borrowings represent independent events, there are 4 different apomorphic characters in the semantic group 'mother', one for each language. Although apomorphic characters are uninformative in parsimony analyses, in a likelihood and Bayesian framework, apomorphic characters are informative for the estimation of evolutionary rates and branch lengths and should not be excluded.

Additional Meanings Due to semantic shift, our dataset included reflexes of roots whose 'central' meaning was not present in our initial wordlist. For example, in our original dataset, cognates of the root *aʔaŋ* were present in various languages but for seemingly disparate meanings: in Tapiete and Kayabí for 'sing', in Guarayú and Tembé for 'draw', in Tocantins Asuriní for both 'sing' and 'draw'. Nominalized forms of the same root were present in Tupinambá and Paraguayan Guaraní for 'spirit'. Even after the recognition that all these forms belonged to the same cognate set, it was unclear what the 'central' meaning of the root is. Therefore, for many languages apparently lacking a cognate, this absence could be misleading. The picture became clear with the discovery of the 'central' meaning for the root in question, namely 'imitate'. After expanding our wordlist to include the meaning 'imitate', we found cognates in most languages. To resolve cases like the example above, where we could relatively easily identify the 'central' meaning of a certain root, we expanded our initial wordlist to include another 43 meanings and repeated the coding process.

Searching for cognates based on expected form We searched for cognates for certain roots in languages where they were lacking, based on the expected form of the cognates and not based on meaning. We employed this search method in cases where under a certain 'central' meaning there were two or more cognate sets that were widespread, but for which we could not distinguish a difference in meaning. We searched for cognates only in well-

documented languages where we could reasonably expect to be able to find them (Paraguayan Guaraní, Wayampí, Tupinambá, Omagua, Kokama, Parintintin, Tembé).

Absence Coding In all the above procedures, whenever a cognate form was found it replaced a misleading zero, thus resulting in a matrix representing more accurately cognate presence and absence. In other words, a cognate was coded as absent for a particular language in our dataset if all the following conditions were met: no cognate was found when searching for the 'central' meaning of the cognate set in question, no cognate was found when searching for meaning synonymous or near-synonymous to the 'central' meaning, there was no compound or otherwise complex word in our dataset that involved the cognate for that particular language, and, in the cases of well-documented languages and of relatively vague 'central' meanings, no cognate was found when searching based on the expected form.

If we imagine a continuum between the Gray and Atkinson method on the one end and cognate coding on the other, our coding method is somewhere in the middle. With the Gray and Atkinson method we have maximum levels of false absences and consequently of non-independence among characters, while with cognate coding we have no false absences and truly independent characters. Our method cannot eliminate false absences, but it enriches the matrix with real absences decreasing the overall amount of non-independence in the dataset and approximating cognate coding to the degree possible with available resources.

3.5 Phylogenetic Analysis

Model and Priors

Given that there are no well-established sound correspondences across the Tupí-Guaraní family, we were not able to identify internal loans, except a few cases involving outgroup languages. Therefore, we used a generalized binary model implemented in MrBayes3.2 (also known as restriction site model), which, while allowing for asymmetric rates of gain and loss of cognates, does not impose only one origin per cognate set. Apart from cases of borrowing, other cases where the algorithm may infer multiple gains of the same cognate set in our dataset are errors in cognacy judgement, parallel semantic shift from a meaning outside our dataset into a meaning included in our dataset, and cases where inadequate language documentation leads us to mistakenly code a cognate as absent (i.e. if a certain cognate set arises twice on our tree, it may be that some of the apparent absences are in reality false that should be coded as 'unkown', so with the full lexicon we would have only one gain of the cognate set).

The only parameter for the cognate gain and loss model is the stationary probabilities (which are proportional to the rates of loss and gain of cognate sets). We used a flat Dirichlet prior, which places equal probability to all possible ratios of cognate gain and loss. This prior is totally uninformative regarding the relative rates of cognate gain and loss, so any asymmetry in these rates is generated by the data.

Our dataset includes varying degrees of 'basic' vocabulary, and it is possible that some areas have a higher rate of cognate loss and gain. In order to allow for variability in the rates of evolution across different cognate sets, we compared a model including gamma-distributed rates and a model without rate heterogeneity using Bayes Factors. To estimate the Bayes Factors we used the AICM procedure, as implemented in Tracer v1.6 (Baele et al., 2012). For the shape parameter of the gamma distribution we used a uniform prior in the interval (0, 200).

Phylogenetic Analyses and Ancestral State Reconstructions

All analyses were performed with MrBayes3.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) at the California Academy of Sciences CCG PhyloCluster. For every analysis, we ran 2 independent chains of 10 million generations each, logging results every 1000 generations. We also ran a total of 6 'hot' chains (3 for every 'cold' one) with swaps being proposed every 50 generations. We used a conservative 25% burn-in for MCMC diagnostics and our results. In all cases, stationarity and convergence for all parameters were verified using Tracer (Rambaut and AJ, 2007), while topology convergence was assessed with the average standard deviation of split frequencies, which in all cases fell below 0.01.

Majority-rule consensus trees of the posterior sample were made with MrBayes3.2 and annotated using FigTree (`http://tree.bio.ed.ac.uk/software/figtree/`). The posterior sample was visualized with DensiTree (Bouckaert, 2010).

All characters were reconstructed using maximum likelihood and the estimated model parameters of cognate gain and loss on the majority-rule consensus tree using Mesquite (Maddison and Maddison, 2007).

3.6 Results

Cognate Sets

We constructed a total of 4205 cognate sets, of which 1113 were parsimony-informative and 2989 were singleton cognate sets (i.e. attested only in one language).

Full dataset results and exclusion of 'wildcards'

The analysis of our full dataset of 34 languages recovered the Tupí-Guaraní family as monophyletic (with a posterior probability of 1), the Nuclear Tupí-Guaraní Subgroup (Nuclear TG, see Discussion section), as well as some low-level subgroups (see Figure 3.3).

However, many areas of the majority-rule consensus tree are largely unresolved with quite low posterior probability values ranging from 0.5 to 0.6. A largely unresolved consensus tree like this can be a result of one, or a few, languages that attach with more or less equal probability in different areas of the tree, while otherwise the trees summarized may share

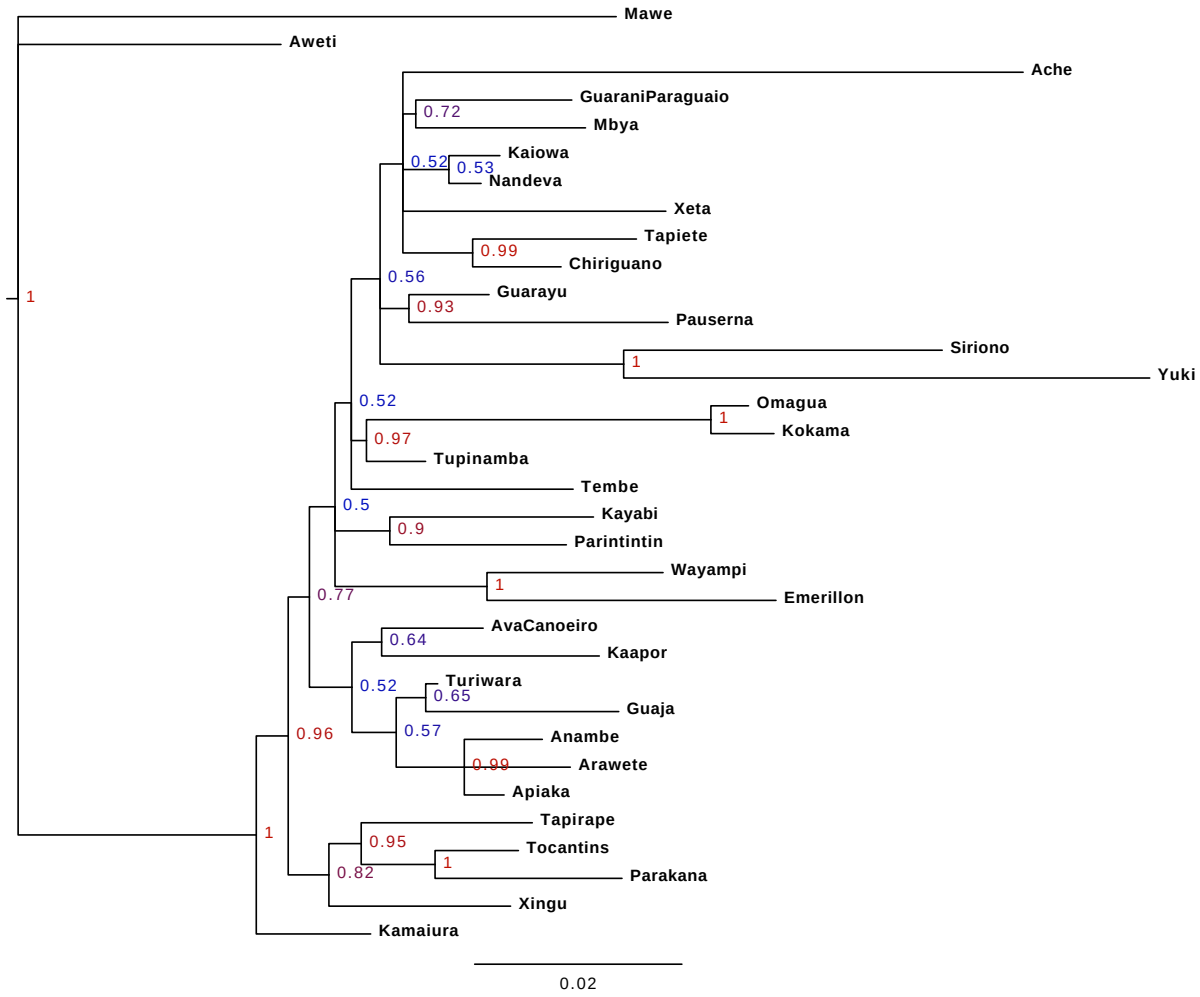

Figure 3.3: Majority-rule consensus tree of full dataset. Node values are posterior probabilities.

very similar topology with the exception of the placement of these languages. Such OTUs have been termed 'wildcards' in the biological literature (Nixon and Wheeler, 1992). A common characteristic of wildcards is that they often lack sufficient information to be placed with confidence at any particular point in the tree (Wiens, 2003). Thus, a usual 'suspect' for wildcard behavior are languages with a high percentage of missing data, although not all languages with high percentage of missing data are going to exhibit this kind of behavior.

Removing Turiwará from our dataset (the language with the lowest coverage) and rerunning the analysis resulted in a more resolved and better-supported tree (Figure 3.4) with the exception of the base of Nuclear TG. Removing Apiaká (the language with the second

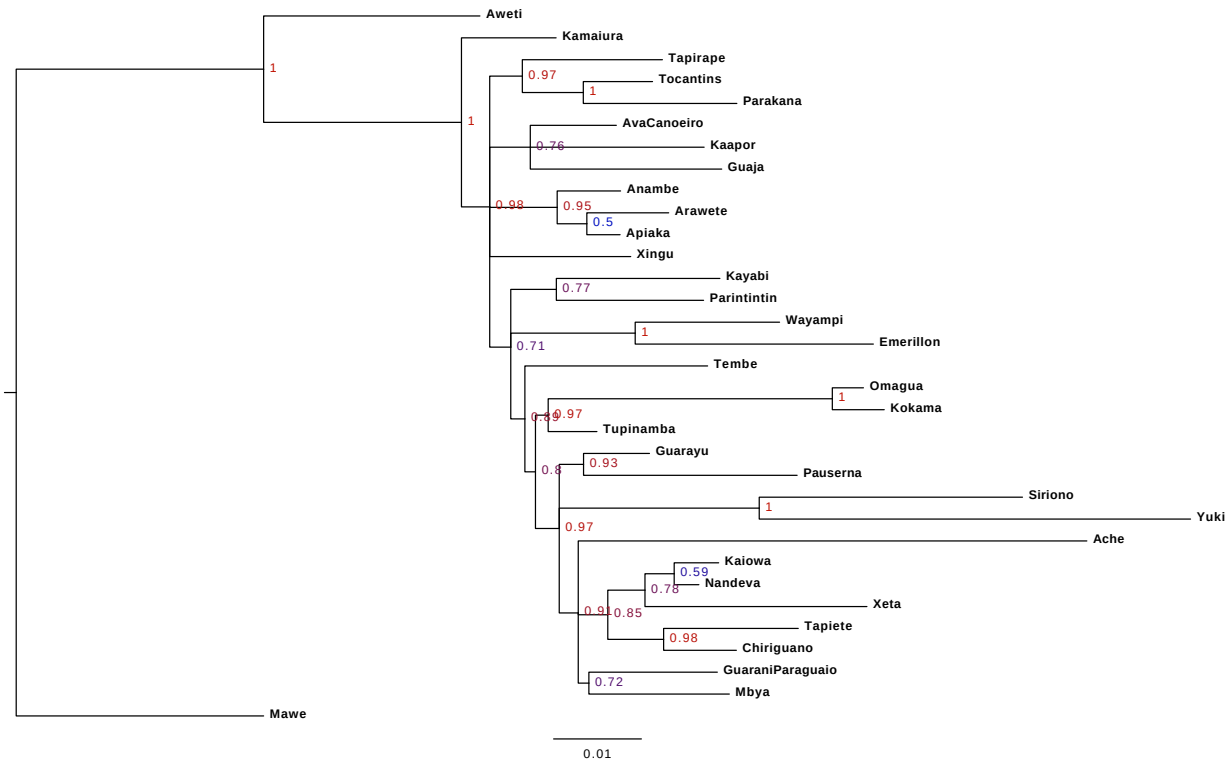

Figure 3.4: Majority-rule consensus tree without Turiwará.

lowest coverage) and rerunning the analysis, yielded a consensus tree with the opposite characteristics of the minus Turiwará tree: a resolved base of Nuclear TG with all other higher structure unresolved (Figure 3.5). In this latter tree, the'wildcard' effect of Turiwará is obvious as it attaches at a completely different point compared to the full tree, collapsing all the nodes in between. From these analyses it seems that both Turiwará and Apiaká are affecting the resolution of the consensus trees in different areas, so we removed both of them for our subsequent analyses (therefore referred to as 'main' analysis).

As expected, the consensus tree of our main analysis was overall more resolved. Further removal of languages with low coverage, such as Ñandéva and Xetá, or languages that were attaching at the base of Nuclear TG with low posterior probability, such as Guajá, in various combinations did not have qualitative differences in terms of topology from our main analysis, but further increased the posterior probability values of the main clades (data not shown). We chose to not remove any more languages, as they didn't seem to have an effect on our ability to effectively summarize the posterior sample. There is one topology difference that must be noted between the full tree and our main analysis presented below and that is the position of Anambé and Araweté. In the full analysis, Anambé and Araweté group

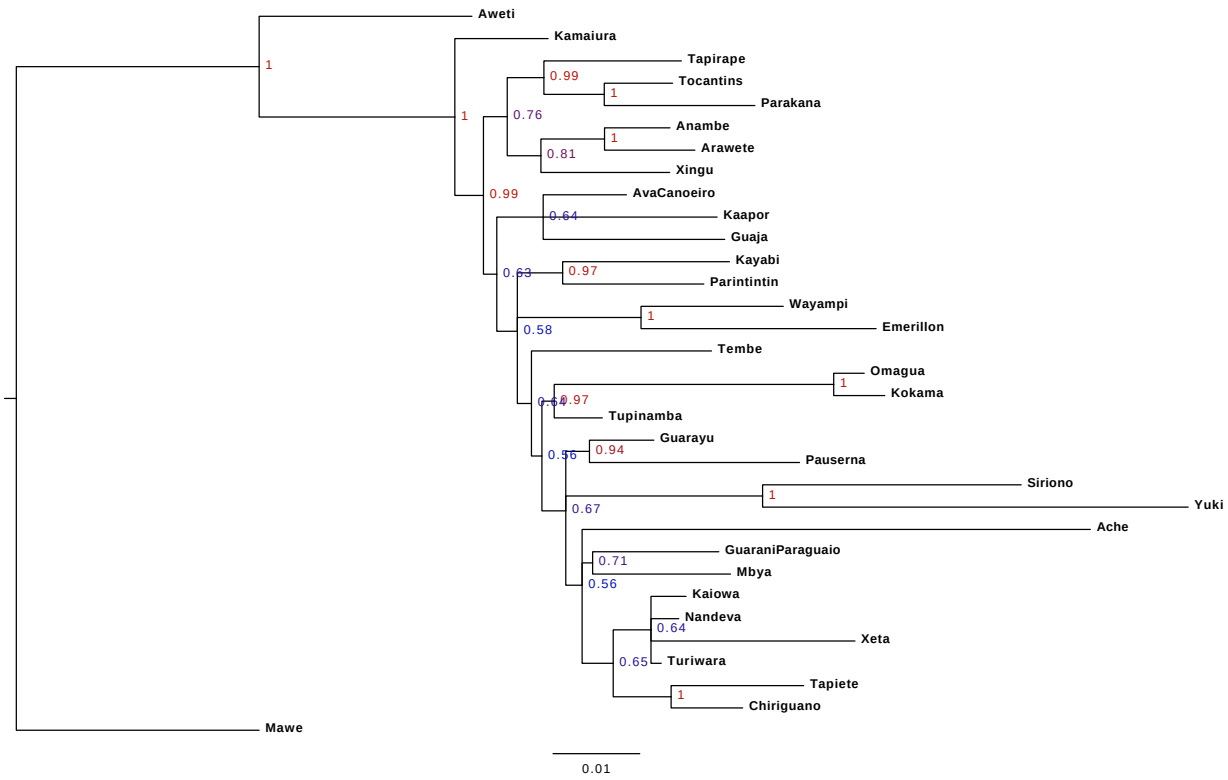

Figure 3.5: Majority-rule consensus tree without Apiaká.

with Apiaká, while in our main analysis they group with Xingú Asuriní. Among the various language exclusion permutations we did, Apiaká did group strongly with Anambé and Araweté, but the placement of the whole clade was not stable. In any analysis where Apiaká was absent, Anambé and Araweté grouped with Xingú Asuriní, as in our main results. We conclude therefore that this was another result of the destabilizing effect of Apiaká at the base of the tree.

Results of Main Analysis

Bayes factors comparison between runs with and without gamma-distributed rate variation across cognate sets, yielded decisive support for the inclusion of rate heterogeneity (BF difference 1316 in favor of gamma-distributed rates) (Kass and Raftery, 1995). The majority-rule consensus tree of our main analysis (excluding Turiwará and Apiaká) is shown in Figure 3.6. The asymmetry between cognate loss and gain is 31:1, which suggests Dollo-like behavior and low level of borrowing within the family. A list of some cognate gains and losses reconstructed at each of the nodes can be found at Appendix 2.

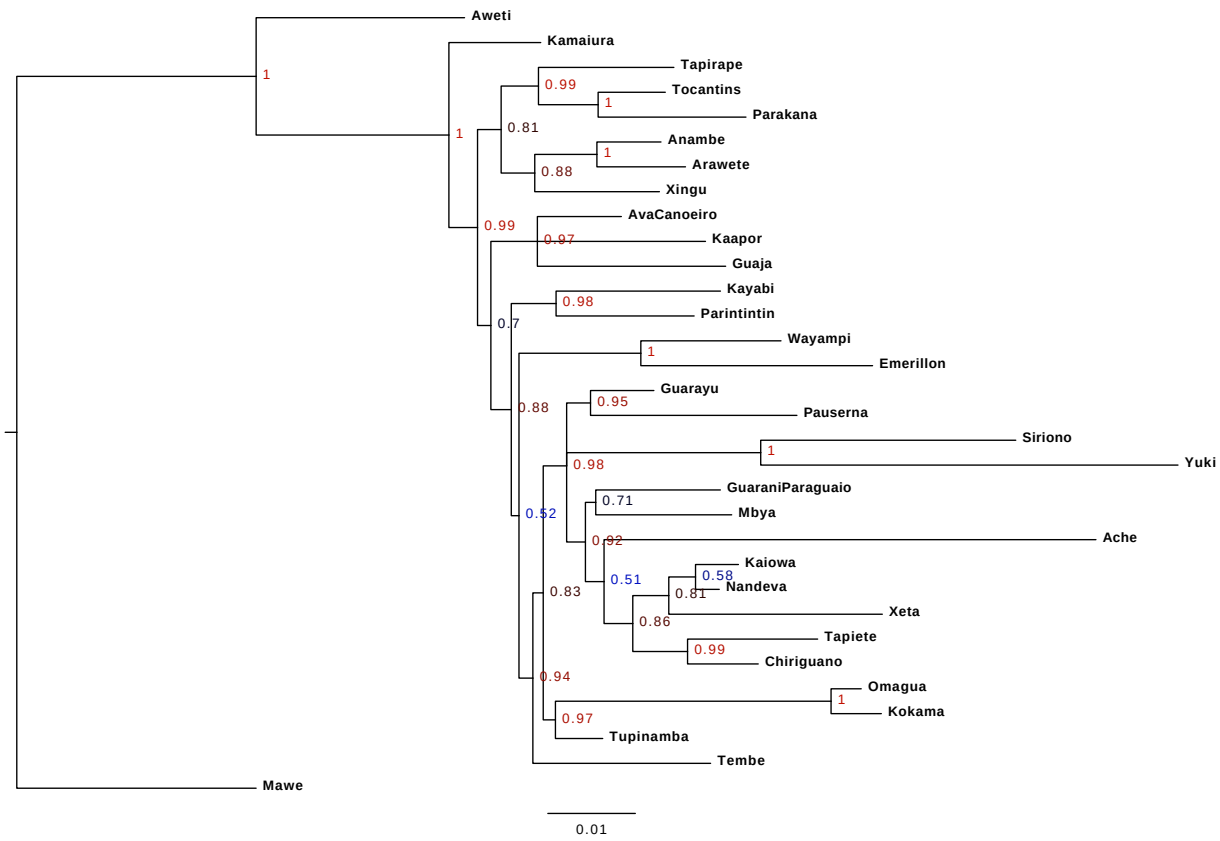

Figure 3.6: Main Results: Majority-rule consensus tree without Apiaká and Turiwará.

3.7 Discussion

Proposed classification of Tupí-Guaraní

Our new classification of Tupí-Guaraní is shown in Figure 3.7. For this conservative proposal we show only the nodes with posterior probability equal to or above 80%. Nodes with lower posterior probability are collapsed in the classification and shown as polytomies. We also provide names for major subgroups.

Nuclear Tupí-Guaraní and the position of Kamaiurá According to our analysis, Kamaiurá is the sister language to all other Tupí-Guaraní languages, which comprise the Nuclear Tupí-Guaraní subgroup. This relationship has not been suggested before, although Kamaiurá has been classified alone as Group VII by Rodrigues (1984/1985) suggesting it doesn't have close relatives. However, in Lemle's (1971) and Rodrigues and Cabral's (2002)

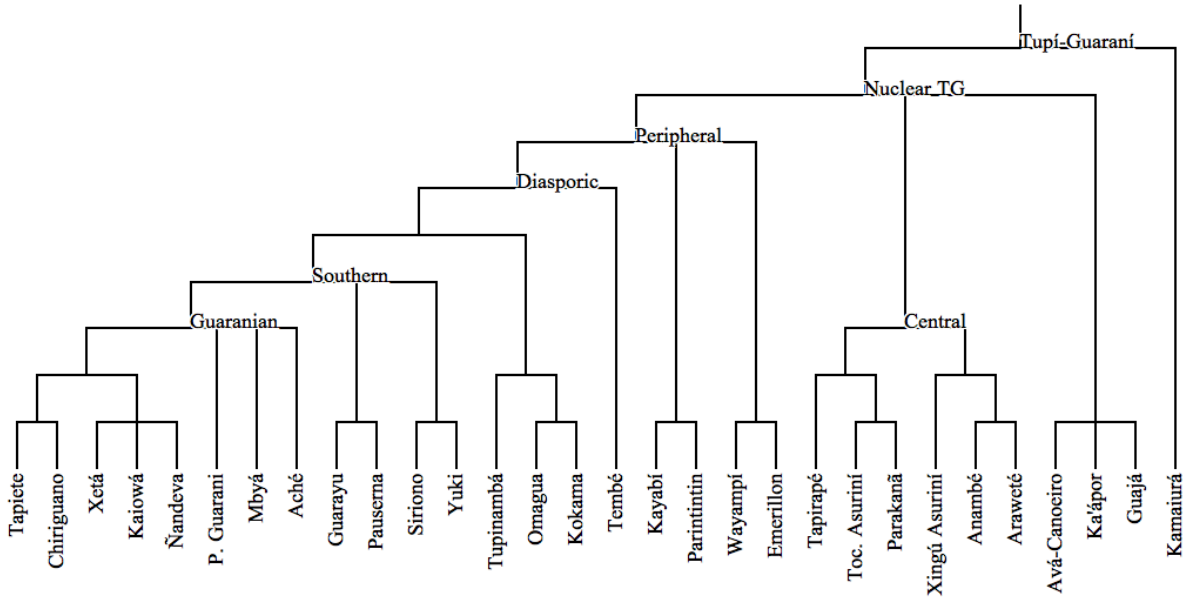

Figure 3.7: Proposed classification of Tupí-Guaraní. Simplified from the consensus tree in Figure 3.6 with an 80% posterior probability cutoff.

classifications, Kamaiurá is part of other subgroups. It must be noted here that in our analysis this position of Kamaiurá could be driven by loans from Awetí, one of the outgroup languages, as the two languages are spoken close to each other in the same national park. However, inspection of the cognate sets that can be reconstructed at the Nuclear TG node showed a large number of cognate gains (at least 13) and a relatively small number of cognate losses (3). If indeed the position of Kamaiurá was the effect of borrowing, we would expect many more cognate losses reconstructed at the Nuclear TG node (since the "loss" of the loans would be optimized at the Nuclear TG node) and correspondingly few cognate gains. More Kamaiurá data, and especially verification that these apparently missing Nuclear TG cognates are indeed absent, would strengthen our hypothesis and rule out the possibility that this pattern is driven by borrowing.

The Central subgroup The Central subgroup is relatively weakly supported in our analysis (0.81 posterior probability). It corresponds to Group V plus a subset of Group VI of Rodrigues (1984/1985). Most of the languages of this subgroup are located in the area between the Xingú and Tocantins rivers. Curiously, almost all the changes reconstructed for the Central subgroup are cognate losses.

The Diasporic subgroup The Diasporic subgroup includes Tembé, the Tupinambá subgroup and the Southern subgroup. All languages of the Diasporic subgroup are at the edges or outside the Amazon basin. Tembé is the only language located close to the main concentration of Tupí-Guaraní languages close to the mouth of the Amazon. The Diasporic subgroup is supported by at least 9 cognate gains (see Appendix 2).

The Southern subgroup The Southern subgroup includes the Siriono subgroup, the Guarayú subgroup and the Guaranian subgroup. The basal structure of the Southern subgroup is not resolved and the three branches mentioned above create a trichotomy in our consensus tree. This trichotomy may be resolved with more or different kinds of data, or it may be an indication of rapid differentiation of Proto-Southern in these three branches. The Siriono and the Guarayú branches are located in the headwaters of the Madeira river, while the Guaranian subgroup is spread across the Paraná river basin. The Southern subgroup is one of the better supported subgroups in our analysis, with 7 cognate gains and 11 cognate losses reconstructed at this level. Among them are animal and plant terms, such as howler monkey, pineapple, bat, gourd and deer. Of course none of these word meanings are lacking in other Tupí-Guaraní languages, but they are expressed with reflexes of different roots than in the Southern subgroup. These lexical innovations and losses indicate moving to a different environment and/or contact with other language families. One of the most conspicuous lexical gains is the word for howler monkey. The common term for howler monkey in most other Tupí-Guaraní languages (and the two Tupian languages, Awetí and Mawé) is cognate with [akɨkɨ] in Tupinambá. This root is lost in the Southern subgroup and a new root is gained for howler monkey, [karaja] in Paraguayan Guaraní. Interestingly, the howler monkey in the Paraná basin (*Alouatta caraya*, the specific epithet is its Tupí-Guaraní common name) is a different species than the ones living in the Amazon basin.[20]

The Guaranian subgroup The Guaranian languages form a well-supported subgroup but its internal structure is not well-resolved. This is not surprising in light of the fact that many of these languages are mutually intelligible to a large degree and some are considered dialects of the same language (Michael, pers.comm. ethnologue website). Two notable languages grouping within the Guaranian subgroup are Aché and Xetá. They both exhibit divergent grammatical features which presumably developed under contact situations (Rodrigues, 1978; Rößler, 2008). Even though our analysis is not based on grammatical features, both Aché and Xetá show instability in terms of where they attach on the tree and are at the tips of some of the longest branches. Long branches indicate a large amount of changes due to high rates and/or long time. These results show that there is a correlation between the amount of grammatical and lexical change. The Guaranian subgroup is supported by 6

[20]The classification of howler monkeys (genus *Alouatta*) has been in a lot of flux. Howler monkeys of the Amazon basin were formerly classified as two species, *A. seniculus* north of the Amazon river and west of the Madeira river and *A. belzebul* east of the Madeira and south of the Amazon, each with a number of subspecies, which are now raised to species level (IUCN website).

cognate gains and 4 losses. Among the gains, there are two animal terms for anteater and tapir.

Comparison with previous classifications

Lemle (1971) Lemle (1971) proposed a classification of Tupí-Guaraní based on 10 languages (see Figure 3.8). In our analysis, we recover one of the higher subgroups of Lemle's classification, which includes the Southern subgroup and the Tupinambá subgroup as sisters. However, the other subgroups in Lemle's classification are not recovered in our analysis.

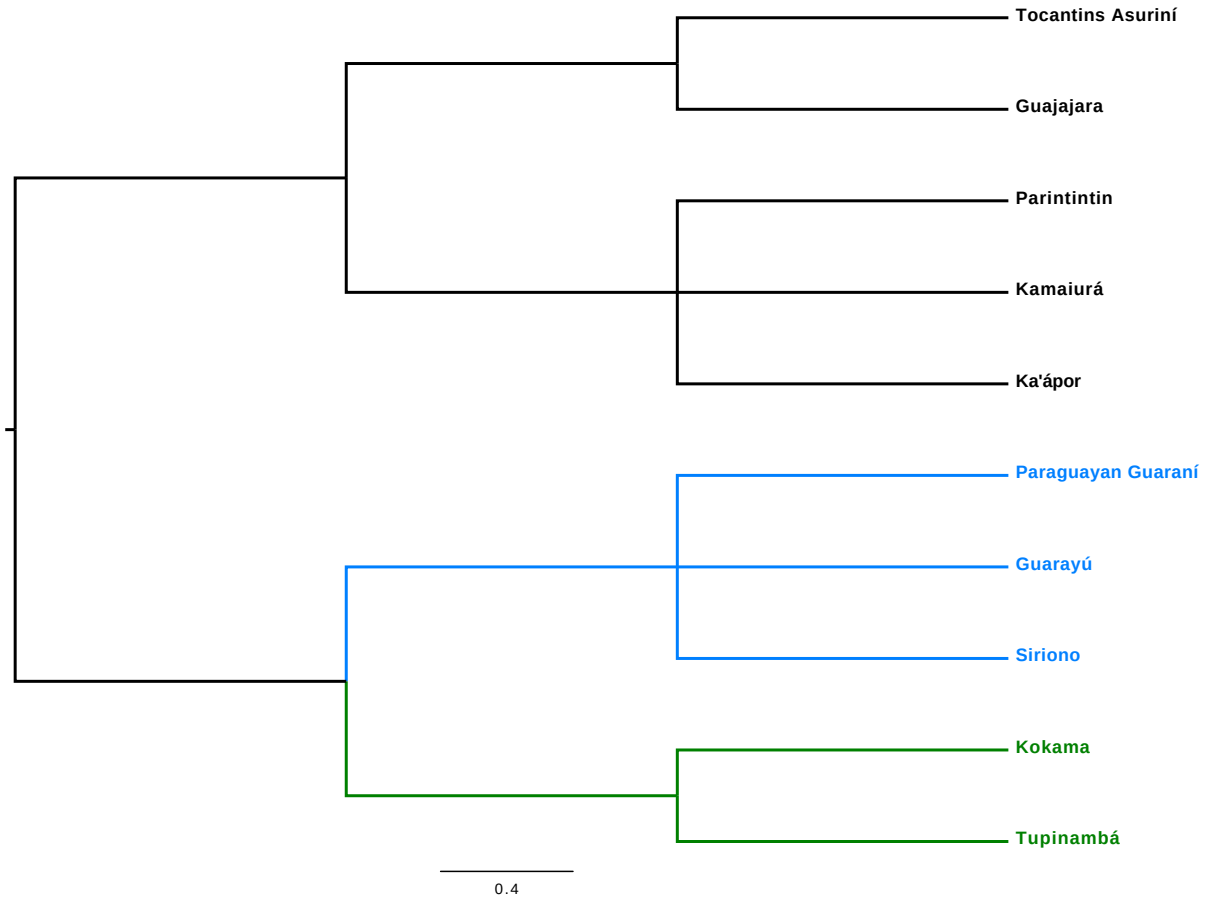

Figure 3.8: Tupí-Guaraní classification according to Lemle (1971). Southern subgroup in blue and Tupinambá subgroup in green.

Rodrigues and Cabral (2002) Overall, our lower level subgroups are more or less in agreement with Rodrigues and Cabral (2002). In our analysis we recovered 5 of the 8 groups proposed by Rodrigues (1984/1985) and Rodrigues and Cabral (2002) (see Figure 3.9):

Groups I, III, V, VI, and VII. Also, Group II is paraphyletic in our analysis. Groups IV and VIII are not recovered in our analysis.

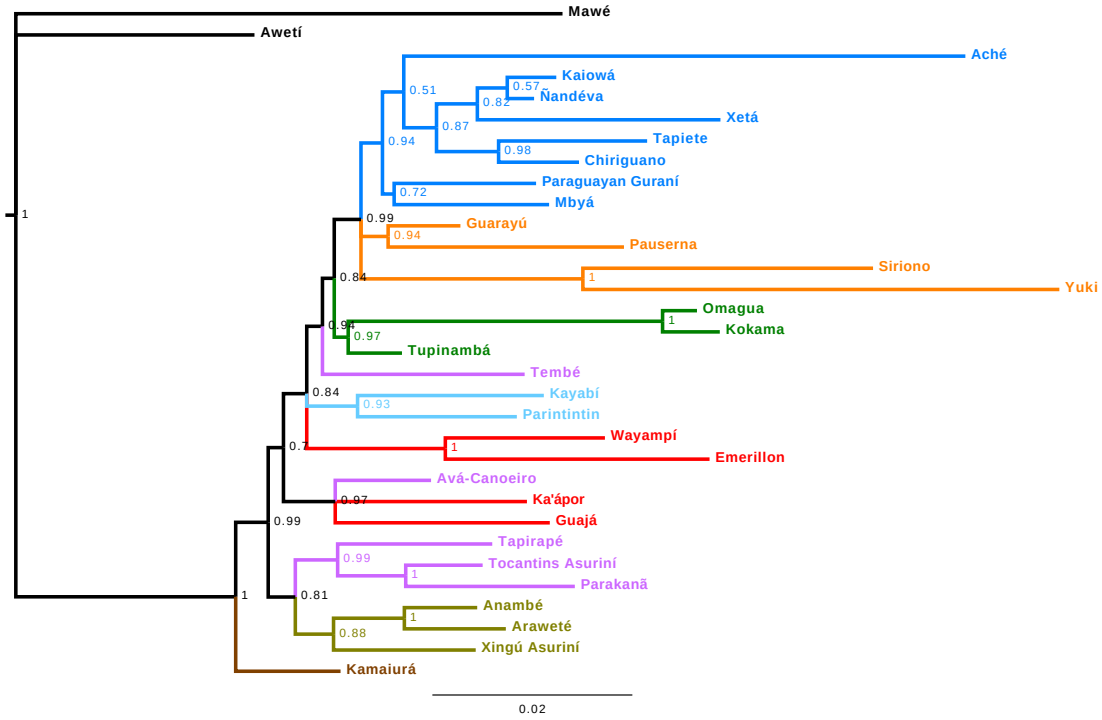

Figure 3.9: Majority-rule consensus tree. Colors according to groups of Rodrigues and Cabral (2002). Group I in dark blue, Group II in orange, Group III in dark green, Group IV in purple, Group V in olive green, Group VI in light blue, Group VII in brown, Group VIII in red.

However, the higher level structure is significantly different between the two classifications (see Figure 3.10). Rodrigues and Cabral proposed 3 first-order subgroups, of which only one, Guaranian or Group I, is recovered as monophyletic in our analysis. The others are two successive paraphyletic grades at the base of the Guaranian subgroup. Paraphyletic grades can be misinterpreted as monophyletic groups because of shared retentions. Rodrigues and Cabral used phonological and morphological characters for their classification, but again no specific changes are reconstructed at each node, making it difficult to replicate their analysis. It seems that the same changes are repeated independently on multiple branches of their tree and in their discussion they mention both shared innovations and shared retentions as criteria for subgrouping. The phonological and morphological characters used by Rodrigues and Cabral need to be reevaluated and re-optimized on a tree, as it is possible that the evidence from these characters is not contradicting the lexical characters that we used in our analysis.

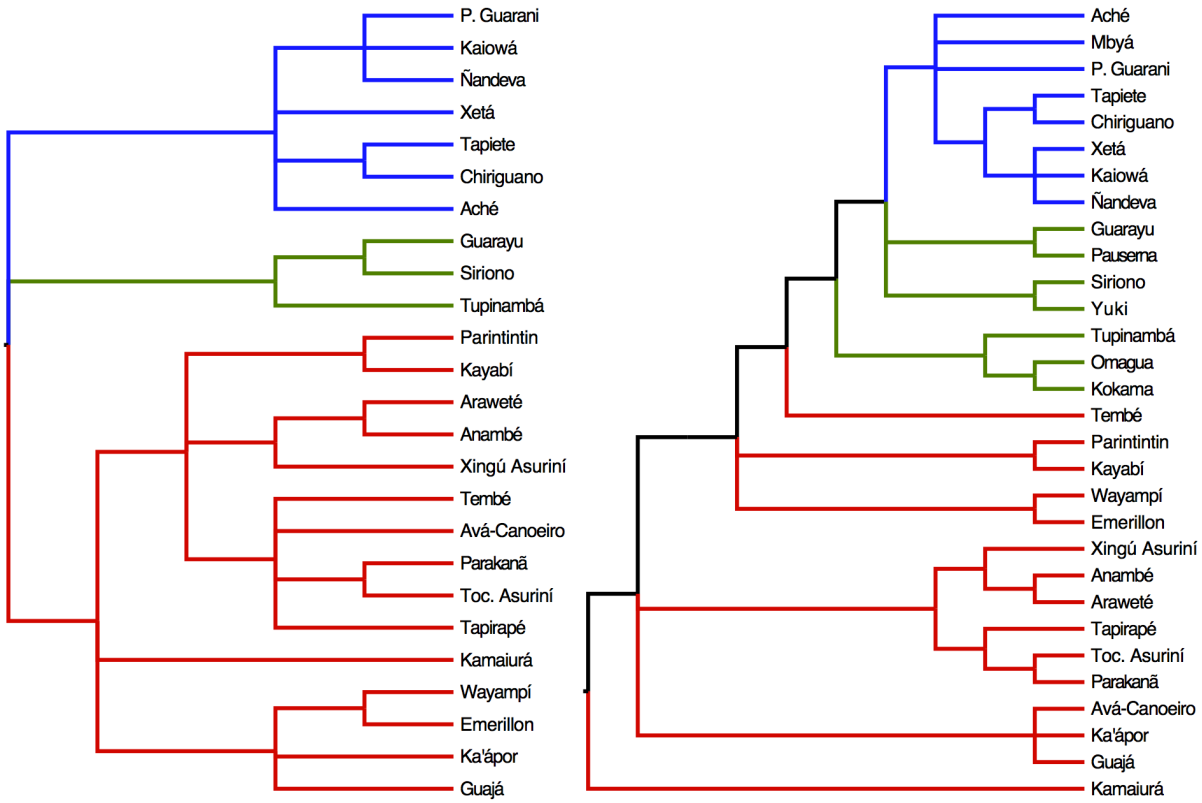

Figure 3.10: Comparison of higher structure between our classification (on the right) and Rodrigues and Cabral (2002) (on the left). Corresponding areas of the trees are colored the same.

Mello (2002) Mello (2002) reorganized the 8 groups of Rodrigues (1984/1985) into 9 groups, by splitting some of the groups and by changing the subgroup membership for some languages, such as Kamaiurá, Parintintin, Guajá, and Xingú Asuriní. None of these changes are supported in our analysis.

Walker, Wichmann, et al. (2012) Walker, Wichmann, et al. (2012) in a paper on Tupí cultural evolution and spread present a Neighbor-Joining tree of the whole Tupí stock based on 40 words. Tupí-Guaraní is recovered as monophyletic in their analysis, but its internal structure is very different from other classifications and our results, except for some low-level subgroups that they all have in common. These differences are not surprising given the very small dataset used and the tree-building method, which is distance-based and therefore doesn't distinguish shared innovations from shared retentions.

Implications for the Tupí-Guaraní homeland and expansion

Regarding the Tupí-Guaraní homeland, all the first order branches of Nuclear TG are entirely within or have members in the area between the Xingú and Tocantins rivers, pointing to this area as the center of diversification of the Nuclear TG subgroup. Kamaiurá, the sister language to Nuclear TG according to our analysis, is further upriver on the course of the Xingú river, making it tempting to suggest that it is the result of an early upriver expansion. Although it cannot be ruled out that the reverse is the case (i.e. the Proto-Nuclear TG moved downriver the Xingú) or some other scenario, we favor the former explanation, as Tupí-Guaraní groups seem to have carried out a number of upriver movements along various Amazonian tributaries (see below).

If, as proposed by Rodrigues (2000), the Tupí homeland is indeed in the state of Rondônia at the headwaters of the Madeira river, the location of Mawé, our first outgroup, at the point where the Madeira joins the Amazon, is suggestive of movement down the Madeira river and subsequently the Amazon by the Mawetí-Tupí-Guaraní, followed by two migrations up the Xingú by Awetí and Kamaiurá.

In conclusion, our results suggest an Amazonian homeland at least for Nuclear TG, if not for Tupí-Guaraní itself, between the Xingú and Tocantins rivers, just south of the Amazon. This location does not agree with any of the previous hypotheses for the TG homeland, and has important implications for the Tupí-Guaraní expansion, as explained below.

Taking the location of Nuclear TG in between the Xingú and Tocantins rivers as a starting point, we have a series of hypothesized expansion routes for Tupí-Guaraní languages and presumably their speakers. First, we have a series of relatively small movements:

- Avá-Canoeiro migrates upriver the Tocantins.
- Tapirapé, the only member of the Central subgroup to move significantly away from the homeland, migrates upriver the Araguaia, a tributary of Tocantins.
- The Proto-Parintintin-Kayabí moved up the Tapajós river, a tributary of the Amazon.

Unlike these relatively small migrations, the Diasporic subgroup has members across the South American continent: Tupinambá was spoken historically all along the eastern coast of Brazil, Omagua and Kokama are relatively widespread at the headwaters of the Amazon, while the Southern subgroup is spread mostly in the Paraná basin. Tembé is the only member of the Diasporic subgroup that has remained relatively close to the homeland. It is evident that the Amazon river and the Atlantic coast played a vital role in the Proto-Omagua-Kokama and Tupinambá expansion respectively. Such an expansion possibly points to some cultural innovation (such as a new crop, advanced canoe technology or a sociopolitical breakthrough) that allowed both for a population expansion as well as for the geographical spread itself. The position of Tembé as the most basal branch of Diasporic in our tree and its geographical position east of the Tocantins is an indication that these changes likely happened at the mouth of the Amazon.

The most impressive act of the Tupí-Guaraní expansion though, and perhaps the least well understood, at least with the data at hand, is the spread of the Southern subgroup over the Paraná river basin. It is unclear what route the Proto-Southern followed to reach the Paraná basin: possibilities include various tributaries of the Amazon, such as the Madeira, the Tapajós or the Tocantins-Araguaia, and the Atlantic coast. If Southern moved south along the Atlantic coast, this migration should have happened before the Tupinambá spread, as it doesn't seem that Tupinambá is sister to the Guaranian languages. The internal topology of the Southern subgroup lends some credibility to the Madeira route, as both the Guarayú and Siriono branches are located in the headwaters of the Madeira river. Thus, we can imagine that they were 'left behind' as the Proto-Guaranian moved on to enter the Paraná river basin.

3.8 Conclusion

This study represents one of the largest efforts to date to clarify the relationships of Tupí-Guaraní languages both in terms of number of languages included, as well as the dataset used. It also represents the first attempt to apply phylogenetic methods to the study of Tupí-Guaraní. Based on a dataset of 543 lexical meanings, we propose a new internal classification of Tupí-Guaraní, which, although broadly compatible at the lower level subgroups with previous classifications, differs significantly in the higher-level topology. One of the most important differences of our results is that the widely recognized Southern subgroup is not a first-order subgroup as in previous classifications, but a deeply nested group. Also, other previously suggested higher-level groups are paraphyletic grades in our analysis. The position of the highly dispersed languages, such as the Tupinambá subgroup and the Southern subgroup, deeply nested within the Tupí-Guaraní phylogeny, suggests an Amazonian origin for the Tupí-Guaraní languages.

Future Directions

In order to further resolve the relationships and illuminate our hypotheses regarding the homeland and expansion of Tupí-Guaraní languages, there are several avenues of research. First, other kinds of linguistic data, such as morphology and sound correspondences, need to be incorporated and analyzed jointly and independent from the lexicon. Second, divergence time estimation based on historical literature sources (e.g. old grammars from the early colonial times, Old Guaraní) and contemporary phylogenetic methods is a natural step forward to get a reliable time-frame for the Tupí-Guaraní diversification and possibly to associate the languages with archaeological evidence, such as pottery traditions. Last but not least, in order to uncover the history of Tupí-Guaraní languages and cultures, an integrative approach is necessary using all the available evidence: linguistics, archaeology, ethnography, genetics and the knowledge and intuitions of the Tupí-Guaraní peoples themselves.

3.9 Acknowledgements

We are indebted to Sebastian Drude, Françoise Rose, Eva-Maria Röβler, and Rosa Vallejos, who kindly shared unpublished lexical data from Awetí, Emérillon, Aché, and Kokama-Kokamilla, respectively. Noé Gasparini provided access to data on Yuki and Anambé. We thank audiences at **dhworom*, an occasional historical linguistics working group at the University of California, Berkeley, and at the 2013 Workshop on American Indigenous Languages at the University of California, Santa Barbara, for helpful comments on earlier versions of this work. Diamantis Sellis facilitated the automated binary coding of the data set and developed scripts to verify consistency between comparative and cognate lists. This work was partially supported by an NSF DEL award #0966499: *Collaborative Research: Kokama-Kokamilla (cod) and Omagua (omg): Documentation, Description, and (Non-)Genetic Relationships* to Lev Michael, and a UC Berkeley Social Science Matrix seminar grant to the same author.

Bibliography

Adam, Lucien (1896). *Matériaux pour servir à l'établissement d'une grammaire comparée des dialectes de la famille tupi*. Paris: J. Maisonneuve.

Alekseyenko, Alexander V., Christopher J. Lee, and Marc A. Suchard (2008). "Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos". In: *Systematic Biology* 57.5, pp. 772–784.

Almeida, Antonio, Irmãzinhas de Jesus, and Luiz Gouvea de Paula (1983). *A língua tapirapé*. Rio de Janeiro: Biblioteca Repográfica Xerox.

Anchieta, Joseph (1595). *Arte de grammatica da lingoa mais usada na costa do Brasil*. Coimbra: Antonio Mariz.

Armoye Urarepia, Celso (2009). "Análisis de la lengua guarayo (tesina)". Santa Cruz de la Sierra.

Atkinson, Quentin D and Russell D Gray (Aug. 2005). "Curious parallels and curious connections–phylogenetic thinking in biology and historical linguistics." In: *Systematic Biology* 54.4, pp. 513–26.

Baele, Guy et al. (2012). "Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty." In: *Molecular Biology and Evolution* 29.9, pp. 2157–67.

Bandelt, HJ and AWM Dress (1992). "Split decomposition: a new and useful approach to phylogenetic analysis of distance data". In: *Molecular Phylogenetics and Evolution*.

Betts, La Vera (1981). *Dicionário parintintin-portugues portugues-parintintin*. Cuiabá: Summer Institute of Linguistics (SIL).

Borges e Souza, Patrícia de Oliveira (2004). "Estudos de aspectos da língua kaiabi (tupi)". MA thesis. Campinas: Universidade Estadual de Campinas.

Borges, Monica Veloso (2006). "Aspectos fonológicos e morfossintáticos da língua avá-canoeiro (tupí-guaraní)". PhD dissertation. Universidade Estadual de Campinas.

— (2007). "Posposições da língua avá-canoeiro (tupí-guaraní)". In: *Línguas e Culturas Tupí*. Ed. by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues. Campinas: Editora Curt Nimuendajú, pp. 385–389.

Bouckaert, Remco R (2010). "DensiTree: making sense of sets of phylogenetic trees." In: *Bioinformatics* 26.10, pp. 1372–3.

Bouckaert, Remco et al. (2012a). "Mapping the Origins and Expansion of the Indo-European Language Family". In: *Science* 337.6097, pp. 957–960.

Bouckaert, Remco et al. (2012b). "Mapping the origins and expansion of the Indo-European language family." In: *Science* 337.6097, pp. 957–60.

Boudin, Max H (1978). *Dicionário de tupi moderno: Dialeto tembé-ténetéhara do alto do rio Gurupi*. São Paulo: Conselho Estadual de Artes e Ciencias Humanas.

Bowern, Claire and Quentin Atkinson (2012a). "Computational phylogenetics and the internal structure of Pama-Nyungan". In: *Language* 88.4, pp. 817–845.

Bowern, Claire and Quentin D. Atkinson (2012b). "Computational Phylogenetics and the Internal Structure of Pama-Nyungan". In: *Language* 88.4, pp. 817–845.

Bridgeman, Loraine I. (1961). "Kaiwa (Guarani) Phonology". In: *International Journal of American Linguistics* 27.4, pp. 329–334.

Bryant, D and V Moulton (2004). "Neighbor-Net: an agglomerative method for the construction of phylogenetic networks". In: *Molecular Biology and Evolution*.

Buth, DG (1984). "The application of electrophoretic data in systematic studies". In: *Annual Review of Ecology and Systematics* 15, pp. 501–522.

Cabral, Ana Suelly Arruda Câmara (1995). "Contact-Induced Language Change in the Western Amazon: The Non-Genetic Origin of the Kokama Language". PhD dissertation. University of Pittsburgh.

— (2007). "New observations on the structure of Kokáma/Omágwa". In: *Language Endangerment and Endangered Languages*. Ed. by Leo Wetzels. Indigenous Languages of Latin America (ILLA). Leiden: Research School CNWS, Leiden University, pp. 365–379.

— (2011). "Different Histories, Different Results: The Origin and Development of Two Amazonian Languages". In: *PAPIA: Revista Brasileira de Estudos Crioules e Similares* 21.1, pp. 9–22.

Cabral, Ana Suelly Arruda Câmara and Aryon Dall'Igna Rodrigues (2003a). *Dicionário asuriní do tocantins-português*. Belém: Universidade Federal do Pará.

— (2003b). "Evidencias de crioulização abrupta em kokama?" In: *PAPIA: Revista Brasileira de Estudos Crioules e Similares* 13, pp. 180–186.

Caetano, B (n.d.). "Vocabulário das palavras guaranis usadas pelo tradutor da" Conquista Espiritual" do Pe". In: *A. Ruiz de Montóia*.

Caldas, Raimunda Benedita Cristina (2009). "Uma proposta de dicionário para a língua ka'apór". PhD dissertation. Universidade de Brasília.

Cardoso, Valéria Faria (2008). "Aspectos morfossintáticos da língua kaiowá (guaraní)". PhD dissertation. Universidade Estadual de Campinas.

Corrêa da Silva, Beatriz Carretta (2010). "Mawé/awetí/tupí-guaraní: Relações lingüísticas e implicações históricas". PhD dissertation. Universidade de Brasília, p. 448.

Costa, Consuelo De Paiva Godinho (2002). "A nasalização em Nhandewa-Guarani". In: *Línguas indígenas brasileiras: Fonologia, gramática e história*. Ed. by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues. Vol. 1. Belém: Editora Universitária, Universidade Federal do Pará, pp. 403–412.

— (2007). "Apyngwa rupigwa: Nasalização em nhandewa-guaraní". PhD dissertation. Universidade Estadual de Campinas.

Couchili, T., D. Maurel, and Francisco Queixalós (2001). "Classes de lexemes en emerillon". In: *Amerindia* 26/27, pp. 173–208.
Crowley, T and C Bowern (2010). *An introduction to historical linguistics*. Oxford New York: Oxford University Press.
Cruz, Aline da (2011). *Fonologia e gramática do nheengatú*. Utrecht: Landelijke Onderzoekschool Taalwetenschap (LOT).
Cunha, Péricles (1987). "Análise fonemica preliminar da língua guajá". MA thesis. Universidade Estadual de Campinas.
Dietrich, Wolf (1990). *More Evidence for an Internal Classification of Tupí-Guaraní Languages*. Berlin: Gebr. Mann Verlag.
— (2007). "Nuevos aspectos de la posición del conjunto chiriguano (guaraní del chaco boliviano) dentro de las lenguas tupí-guaraníes bolivianos". In: *Lenguas indígenas de América del Sur: estudios descriptivo-tipológicos y sus contribuciones para la lingüística teórica*. Ed. by Andrés Romero-Figueroa, Ana Fernández Garay, and Ángel Corbera Mori. Caracas: Universidad Católica Andrés Bello, pp. 9–18.
— (2010). "O tronco tupi e as suas famílias de línguas: Classificação e esboço tipológico". In: *O português e o tupi no Brasil*. Ed. by Wolf Dietrich and Volker Noll. São Paulo: Editora Contexto, pp. 9–26.
Dobson, Rose M. (1973). "Notas sobre substantivos do kayabí". In: *Serie Lingüística* 1, pp. 30–56.
— (1988). *Aspectos da língua kayabí*. Serie Lingüística 12. Cuiabá: Summer Institute of Linguistics (SIL).
— (1997). *Gramática prática com exercícios da língua kayabí*. Cuiabá: Summer Institute of Linguistics (SIL).
Dooley, Robert A. (1991). "Apontamentos Preliminares sobre Ñandéva Guaraní Contemporaneo". Brasília.
— (1998). *Léxico Guaraní, dialeto Mbyá: versão para fins acadêmicos*. Porto Velho: Sociedade Internacional de Linguística.
— (2006). *Léxico Guaraní, dialeto Mbyá com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística*. Cuiabá: Summer Institute of Linguistics (SIL).
Drude, Sebastian (2006). "On the Position of the Awetí Language in the Tupí Family". In: *Guaraní y Mawetí-Tupí-Guaraní: Estudos históricos y descriptivos sobre una familia lingüística de América del Sur*. Ed. by Wolf Dietrich and Haralambos Symeonidis. Berlin: LIT Verlag, pp. 11–45.
— (2008). "Nasal Harmony in Awetí and the Mawetí-Guaraní Family (Tupí)". In: *Amerindia* 32, pp. 239–267.
— (2011). "Awetí in Relation with Kamayurá: The Two Tupian Languages of the Upper Xingu". In: *Alto Xingu: Uma Sociedade Multilíngue*. Ed. by Bruna Franchetto. Rio de Janeiro: Museu do Índio; Fundação Nacional do Índio (FUNAI), pp. 155–191.
Drummond, A. J., M. A. Suchard, et al. (2012). "Bayesian Phylogenetics with BEAUti and the BEAST 1.7". In: *Molecular Biology and Evolution* 29.8, pp. 1969–1973.

Drummond, AJ and A Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC evolutionary biology* 7.1, p. 214.

Espinosa Pérez, Lucas (1989). *Breve diccionario analítico castellano-tupí del Perú: Sección cocama*. Iquitos: Instituto de Investigaciones de la Amazonía Peruana (IIAP); Centro de Estudios Teológicos de la Amazonía (CETA).

Faria, Francisco Raimundo Correia de (1903). "Compêndio da língua brazílica para uzo dos que a ella se quizerem dedicar". In: *Annaes da Bibliotheca e Archivo Publico do Pará*. Vol. 2. Rio de Janeiro: Typographia Pinheiro & C., pp. 293–333.

Farris, J (1983). "The logical basis of phylogenetic analysis". In: *Proceedings of the II meeting of the Willi Hennig Society*.

Faust, Norma (1959). "Vocabulario breve del idioma cocama (tupí)". In: *Perú Indígena* 8.18-19, pp. 150–158.

— (1972). *Gramática cocama: Lecciones para el aprendizaje del idioma cocama*. Serie Lingüística Peruana 6. Lima: Summer Institute of Linguistics (SIL).

Felsenstein, J. (1978). "Cases in which Parsimony or Compatibility Methods will be Positively Misleading". In: *Systematic Biology* 27.4, pp. 401–410.

Felsenstein, Joseph (2004). *Inferring phylogenies*. Sunderland, Mass: Sinauer Associates.

Figueira, Luiz (1687). *Arte de grammática da língua brasílica*. Lisbon: Officina de Miguel Deslandes.

Forster, Peter and Alfred Toth (2003). "Toward a Phylogenetic Chronology of Ancient Gaulish, Celtic, and Indo-European". In: *Proceedings of the National Academy of Sciences*. Vol. 100, pp. 9079–9084.

Franceschini, D (1999). "La Langue Sateré-Mawé-description et analyse morphosyntaxique". PhD dissertation. Université Paris VII.

Gallois, Dominique T. (1997). *Zo'é: History of contact*. URL: `http://pib.socioambiental.org/en/povo/zoe/1965` (visited on 08/09/2014).

Garland, Mary (1978). *Diccionario yuki-inglés*. Chimore: m.s.

Gilii, Filippo Salvatore (1782). *Saggio di storia americana; o sia, storia naturale, civile e sacra de'regni, e delle provincie sapgnuole di Terra-Ferma nell'America meridionale*. Vol. 3. Rome: L. Perego erede Salvioni.

Girard, Rafael (1958). *Indios selváticos de la Amazonía peruana*. Mexico City: Libro Mex.

González, Hebe Alicia (2005). "A Grammar of Tapiete (Tupí-Guaraní)". PhD dissertation. University of Pittsburgh.

— (2008). "Una aproximación a la fonología del tapiete (Tupí-Guaraní)". In: *LIAMES - Línguas Indígenas Americanas* 8, pp. 7–43.

Gordon, Matthew and Françoise Rose (2006). "Émérillon Stress: A Phonetic and Phonological Study". In: *Anthropological Linguistics* 48.2, pp. 132–168.

Gray, R D, A J Drummond, and S J Greenhill (2009a). "Language phylogenies reveal expansion pulses and pauses in Pacific settlement." In: *Science* 323.5913, pp. 479–83.

Gray, Russell D. and Quentin D. Atkinson (2003a). "Language-tree divergence time support the Anatolian theory of Indo-European origin". In: *Nature* 426, pp. 435–439.

Gray, Russell D and Quentin D Atkinson (2003b). "Language-tree divergence times support the Anatolian theory of Indo-European origin." In: *Nature* 426.6965, pp. 435–9.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill (2009b). "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement". In: *Science* 323.5913, pp. 479–483.

Greenhill, Simon J., Alexei J. Drummond, and Russell D. Gray (2010). "How accurate and robust are the phylogenetic estimates of Austronesian language relationships?" In: *PLoS ONE* 5.3, e9573.

Greenhill, Simon J. and Russell D. Gray (2005). "Testing Population Dispersal Hypotheses: Pacific Settlement, Phylogenetic Trees and Austronesian Languages". In: *The Evolution of Cultural Diversity: Phylogenetic Approaches*. Ed. by Ruth Mace, Clare J. Holden, and Stephen Shennan. London: UCL Press, pp. 31–52.

— (2009). "Austronesian Language Phylogenies: Myths and Misconceptions about Bayesian Computational Methods". In: *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*. Ed. by Alexander Adelaar and Andrew Pawley. Canberra: Pacific Linguistics, pp. 1–23.

Grenand, Françoise (1989). *Dictionnaire Wayãpi/Français*. Paris: Peeters/Selaf.

Guasch, Antonio (2003). *Diccionario básico guaraní-castellano, castellano guaraní*. Ed. by Bartomeu Meliá. Asunción: CEPAG.

Harrison, Carl H. (1963). "Pedagogical Information and Drills for the Asuriní Language". Brasília.

— (1971). "The Morphophonology of Asuriní Words". In: *Tupí Studies I*. Ed. by David Bendor-Samuel. Summer Institute of Linguistics Publications in Linguistics and Related Fields. Summer Institute of Linguistics (SIL), pp. 21–71.

— (1975). "Gramática Asuriní: Aspectos de una gramática transformacional e discursos monologados da língua asuriní, família tupí guaraní". In: *Serie Lingüística*. Serie Lingüística 4.

Heckart, Claudia J. and Kim Hill (2007). *Aché. Intercontinental Dictionary Series*. URL: `http://lingweb.eva.mpg.de/cgi-bin/ids/ids.pl?com=simple_browse&lg_id=288` (visited on 2012).

Hennig, Willi (1966). *Phylogenetic systematics*. Urbana: University of Ilinois Press.

Hervás y Panduro, Lorenzo (1787). *Saggio pratico delle lingue*. Cesena: Insegna di Pallade.

Hill, K and K Hawkes (1983). "Neotropical hunting among the Ache of eastern Paraguay". In: *Adaptive responses of native Amazonians*. Ed. by William T. Vickers Raymond B. Hames. New York: Academic Press, pp. 139–188.

Höller, Alfredo (1932). *Guarayo-Deutsches Wörterbuch*. Guarayos: Verlag der Missionsprokura der P.P. Franziskaner, Hall in Tirol.

Horn Fitz Gibbon, Friedrich von (1955). *Breves notas sobre la lengua de los indios pausernas: El üaradu-ñé-e (un dialecto tupí-guaraní en el oriente de Bolivia*. Publicaciones de la Sociedad de Estudios Geográficos e Históricos. Santa Cruz de la Sierra: Imprenta Emilia.

Huelsenbeck, JP and F Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees". In: *Bioinformatics* 17.8, pp. 754–755.

Jensen, Cheryl (1989). *O desenvolvimento histórico da língua wayampí*. Campinas: Editora da UNICAMP.

— (1998). "Comparative Tupí-Guaraní Morphosyntax". In: *Handbook of Amazonian Languages*. Ed. by Desmond C. Derbyshire and Geoffrey K. Pullum. Vol. 4. New York: Mouton de Gruyter, pp. 489–618.

— (1999). "Tupí-Guaraní". In: *The Amazonian Languages*. Ed. by R. M. W. Dixon and Alexandra Y. Aikhenvald. Cambridge: Cambridge University Press, pp. 125–163.

Kakumasu, James Y. and Kiyoko Kakumasu (1988). *Dicionário por Tópicos Kaapor-Portugues*. Brasília: Summer Institute of Linguistics (SIL).

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795.

Lemey, Philippe, Marco Salemi, and Anne-Mieke Vandamme (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge New York: Cambridge University Press.

Lemle, Miriam (1971). "Internal Classification of the Tupí-Guaraní Linguistic Family". In: *Tupí Studies I*. Ed. by David Bendor-Samuel. Summer Institute of Linguistics Publications in Linguistics and Related Fields. Summer Institute of Linguistics (SIL), pp. 107–129.

Lemos Barbosa, Antônio (1951). *Pequeno vocabulário tupí-portugues*. Rio de Janeiro: Livraria São José.

— (1970). *Pequeno vocabulário portugues-tupí*. Rio de Janeiro: Livraria São José.

Lewis, Paul O. (Nov. 2001). "A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data". In: *Systematic Biology* 50.6, pp. 913–925.

Lopes, Mário Alexandre Garcia (2009). "Aspectos gramaticais da língua ka'apor". PhD dissertation. Universidade Federal e Minas Gerais.

Maddison, WP and DR Maddison (2007). *Mesquite: a modular system for evolutionary analysis. Version 2.75. 2011*. URL: `http://mesquiteproject.org`.

Magalhães, Marina Maria Silva (2006). "Harmonia Vocálica como Processo Desencadeador de Mudanças Estruturais na Língua Guajá". In: *Estudos da Língua(gem)* 4.2, pp. 67–75.

— (2007). "Sobre a morfologia e a sintaxe da língua guajá (família tupí-guaraní)". PhD dissertation. Universidade de Brasília.

Marcoy, Paul (1866). *Voyage a travers l'Amerique du Sud, de l'Ocean Pacifique a l'Ocean Atlantique*. Paris: Librairie de L. Hachette et cie.

Martius, Carl Friedrich Philipp von (1867). *Beiträge zur Ethnographie und Sprachenkunde Amerikas*. Vol. 2. Leipzig: Friedrich Fleischer.

Meier, R. (1994). "On the inappropriateness of presence/absence recoding for non-additive multistate characters in computerized cladistic analyses". In: *Zoologischer Anzeiger* 232.5-6, pp. 201–212.

Meliá, Bartomeu (1992). *La lengua guaraní del Paraguay: Historia, sociedad y literatura*. Madrid: Editorial Mapfre.

Mello, Antônio Augusto Souza (2000). "Estudo histórico da família lingüística tupí-guaraní: Aspectos fonológicos e lexicais". PhD dissertation. Universidade Federal de Santa Catarina.
— (2002). "Evidencias fonológicas e lexicais para o sub-agrupamento interno tupí-guaraní". In: *Línguas Indígenas Brasileiras: Fonologia, Gramática e História*. Ed. by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues. Vol. 1. Belém: Editora Universitária, Universidade Federal do Pará, pp. 338–342.
Métraux, Alfred (1927). "Migrations historiques des tupi-guarani". In: *Journal de la Société des Américanistes* 19, pp. 1–45.
Michael, Lev (to appear). "On the Pre-Columbian Origin of Proto-Omagua-Kokama". In: *Journal of Language Contact* 7.2.
Montoya, AR de (1876). *Arte de la lengua guaraní, ó mas bien tupi.*
Murphy, RW (1993). "The phylogenetic analysis of allozyme data: invalidity of coding alleles by presence/absence and recommended procedures". In: *Biochemical Systematics and Ecology* 21.1, pp. 25–38.
Nakhleh, Luay, Don Ringe, and Tandy Warnow (2005). "Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages". In: *Language* 81.2, pp. 382–420.
Nascimento, Ana Paula Lion Mamede (2008). "Estudo fonético e fonológico da língua guajá". MA thesis. Universidade de Brasília.
Nicholson, Velda C. (1978). "Aspectos da Língua Assuriní". Brasília.
— (1982). "Breve Estudo da Língua do Xingú". In: *Ensaios Lingüísticos* 5.
Nimuendajú, Curt (1914). "Vocabularios da língua geral do Brazil nos dialectos dos manajé do Rio Ararandéua, tembé do Rio Acará Pequeno e turiwará do Rio Acará Grand, Est. do Pará". In: *Zeitschrift für Ethnologie* 46, pp. 615–618.
Nixon, Kevin C and Quentin D Wheeler (1992). "Extinction and the Origin of Species". In: *Extinction and Phylogeny*. Ed. by Michael J Novacek and Quentin D Wheeler. New York: Columbia University Press, pp. 119–143.
Noelli, Francisco S. (1998). "The Tupi: Explaining origin and expansions in terms of archaeology and of historical linguistics". In: *Antiquity* 72.277, pp. 648–663.
— (2008). "The Tupi Expansion". In: *The Handbook of South American Archeology*. Ed. by Helaine Silverman and William H. Isbell. New York: Springer, pp. 659–670.
Nylander, Johan et al. (2004). "Bayesian Phylogenetic Analysis of Combined Data". In: *Systematic Biology* 53.1, pp. 47–67.
O'Hagan, Zachary (2011). "Proto-Omagua-Kokama: Grammatical Sketch and Prehistory". BA thesis. University of California, Berkeley.
O'Hagan, Zachary, Lev Michael, and Rosa Vallejos Yopán (2013). *Hacia la reconstrucción morfológica del proto-omagua-kokama*. Austin: Talk to at CILLA VI, October 24-26.
Olson, Gary Paul (1978). "Descrição preliminar de orações Wayãpi". In: *Ensaios Lingüísticos* 3.
Padua, Alexandre Jorge (2007). "Contribuição para a fonologia da língua apiaká (tupí-guaraní)". MA thesis. Universidade de Brasília.

Pease, Helen (1968). "Parintintin Grammar". Porto Velho.
Pereira, Antônia Alves (2009). "Estudo morfossintático do asuriní do xingú". PhD dissertation. Universidade Estadual de Campinas.
Praça, Walkíria Neiva (2007). "Morfossintaxe da língua tapirapé". PhD dissertation. Universidade de Brasília.
Priest, Perry N. and Anne M. Priest (1985). *Diccionario sirionó y castellano*. Cochabamba: Summer Institute of Linguistics (SIL).
Queixalós, Francisco (2001). "Le suffixe referentiant en emerillon". In: *Des noms et des verbes en tupi-guarani: etat de la question*. Ed. by Francisco Queixalós. Munich: LINCOLM Europa, pp. 115–132.
Rambaut, A and Drummond AJ (2007). *Tracer v1.4*. URL: `http://beast.bio.ed.ac.uk/Tracer`.
Reeve, Mary-Elizabeth (1993). "Regional Interaction in the Western Amazon: The Early Colonial Encounter and the Jesuit Years: 1538-1767". In: *Ethnohistory* 41.1, pp. 106–138.
Restivo, Paulo (1893). *Lexicon Hispano-Guaranicum."Vocabulario de la lengua Gauraní"(secundum Vocabularium Antonii Ruiz de Montoya)*. Stuttgardiae: Guilielmi Kohlhammer.
Riester, Jürgen (1972). *Die Pauserna-Guaraŝug'wä: Monographie eines Tupí-Guaraní-Volkes in Ostbolivien*. St. Augustin bei Bonn: Verlag des Anthropos-Instituts.
Ringe, Don, Tandy Warnow, and Ann Taylor (2002). "Indo-European and Computational Cladistics". In: *Transactions of the Philological Society* 100.1, pp. 59–129.
Rodrigues, AD (2000). "Hipótese sobre as migrações dos três subconjuntos meridionais da família Tupi-Guarani". In: *Atas do II Congresso Nacional da ABRALIN*.
Rodrigues, Aryon Dall'Igna (1958). "Classification of Tupí-Guaraní". In: *International Journal of American Linguistics* 24.3, pp. 231–234.
— (1978). "A língua dos índios Xetá como dialeto guaraní". In: *Cadernos de Estudos Lingüísticos* 1, pp. 7–11.
— (1984/1985). "Relações internas na família lingüística tupí-guaraní". In: *Revista de Antropologia* 27/28, pp. 33–53.
Rodrigues, Aryon Dall'Igna and Ana Suelly Arruda Câmara Cabral (2002). "Revendo a classificação interna da família tupí-guaraní". In: *Línguas Indígenas Brasileiras: Fonologia, Gramática e História*. Ed. by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues. Belém: Editora Universitária, Universidade Federal do Pará, pp. 327–337.
Rodrigues, Aryon Dall'Igna and Wolf Dietrich (1997). "On the Linguistic Relationship Between Mawé and Tupí-Guaraní". In: *Diachronica* 14.2, pp. 265–304.
Rogers, Deborah S and Paul R Ehrlich (2008). "Natural selection and cultural rates of change." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.9, pp. 3416–20.
Ronquist, F and JP Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models". In: *Bioinformatics* 19.12, pp. 1572–1574.
Rose, Françoise (2002). "Le probleme de la nasalité dans l'inventaire phonologique de l'émérillon". In: *Amerindia* 26/27, pp. 147–172.

Rose, Françoise (2003). "Le marquage des personnes en Emérillon (Tupí-Guaraní): Un systeme d'accord hiérarchique". In: *Faits de Langues* 21.2, pp. 107–120.
— (2008). "A Typological Overview of Emerillon, a Tupí-Guaraní Language from French Guiana". In: *Linguistic Typology* 12, pp. 431–460.
— (2009). "A Hierarchical Indexation System: The Example of Emerillon (Teko)". In: *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*. Ed. by Patricia Epps and A. Arkhipov. Berlin: Mouton de Gruyter, pp. 63–84.
Rößler, Eva-Maria (2008). "Aspectos da gramática achê: Descrição e reflexão sobre uma hipótese de contato". MA thesis. Universidade Estadual de Campinas.
Ruíz de Montoya, Antonio (1640). *Arte y bocabulario de la lengua guarani*. Madrid: Iuan Sanchez.
Saitou, N and M Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular Biology and Evolution* 4.4, pp. 406–425.
Sampaio, Wany Bernardete de Araujo (1997). "Estudo comparativo sincronico entre o parintintin (tenharim) e o uru-eu-uau-uau (amondava): Contribuições para uma revisão na classificação das línguas tupí-kawahib". MA thesis. Universidade Estadual de Campinas.
Schleicher, Charles Owen (1998). "Comparative and Internal Reconstruction of Proto-Tupí-Guaraní". PhD dissertation. University of Wisconsin, Madison.
Seki, Lucy (1982). "Marcadores de pessoa no verbo kamaiurá". In: *Cadernos de Estudos Lingüísticos* 3, pp. 22–40.
— (1983). "Observações sobre variacão sociolingüística kamayurá". In: *Cadernos de Estudos Lingüísticos* 4, pp. 73–87.
— (1987). "Para uma caracterização tipológica do kamaiurá". In: *Cadernos de Estudos Lingüísticos* 12, pp. 15–24.
— (1990). "Kamaiurá (Tupí-Guaraní) as an Active-Stative Language". In: *Amazonian Linguistics: Studies in Lowland South American Languages*. Ed. by Doris L. Payne. Austin: University of Texas Press, pp. 367–391.
— (2000). *Gramática do kamaiurá: Língua tupí-guaraní do alto Xingu*. Campinas: Editora da UNICAMP.
— (2007). "Partículas e tipos de discurso em kamaiurá (tupí-guaraní)". In: *Lenguas indígenas de América del Sur: Estudios descriptivo-tipológicos y sus contribuciones para la lingüística teórica*. Ed. by Andrés Romero-Figueroa, Ana Fernández Garay, and Ángel Corbera Mori. Caracas: Universidad Católica Andrés Bello, pp. 145–157.
— (2010). *Jene ramỹjjwena juru pytsaret = O que habitava a boca de nossos ancestrais*. Rio de Janeiro: Museu do Índio - Funai.
Silva Julião, Maria Risolêta (2005). "Aspects morphosyntaxiques de l'anambe". PhD dissertation. Université de Toulouse, Le Mirail.
Silva, Gino Ferreira da (2003). "Construindo um dicionário parakanã-português". MA thesis. Universidade Federal do Pará.
Sokal, Robert R. and P. H. A. Sneath (1963). *Principles of numerical taxonomy*. San Francisco: W.H. Freeman.

Solano, Eliete De Jesus Bararuá (2009). "Descrição gramatical da língua araweté". PhD dissertation. Universidade de Brasília.

Stearman, Allyn M. (1986). "The Yuquí Connection: Another Look at Sirionó Deculturation". In: *American Anthropologist* 86.3, pp. 630–650.

Swadesh, M (1952). "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos". In: *Proceedings of the American philosophical society* 96.4, pp. 452–463.

Taylor, John M. (1984a). "A interrogação na língua kaiwá". In: *Estudos sobre línguas tupí do Brasil*. Ed. by Robert A. Dooley. Serie Lingüística. Brasília: Summer Institute of Linguistics (SIL).

— (1984b). "Marcação temporal na língua kaiwá". In: *Estudos sobre línguas tupí do Brasil*. Ed. by Robert A. Dooley. Serie Lingüística. Brasília: Summer Institute of Linguistics (SIL).

Urban, Greg (1996). "On the geographical origins and dispersions of tupian languages". In: *Revista de Antropologia* 39.2, pp. 61–104.

Uriarte, Manuel J. ([1776]1986). *Diario de un misionero de Maynas*. Ed. by Joaquín García. Iquitos: Instituto de Investigaciones de la Amazonía Peruana (IIAP); Centro de Estudios Teológicos de la Amazonía (CETA).

Vallejos, Rosa (2010). "A Grammar of Kokama-Kokamilla". PhD dissertation. University of Oregon.

Vasconcelos, Eduardo Alves (2008). "Aspectos fonológicos da língua xetá". MA thesis. Universidade de Brasília.

Villafañe, Lucrecia (2004). *Gramática yuki: Lengua tupí-guaraní de Bolivia*. Tucumán: Ediciones del Rectorado, Universidad Nacional de Tucumán.

Viveiros de Castro, Eduardo (1992). *From the Enemy's Point of View: Humanity and Divinity in an Amazonian Society*. Chicago: University of Chicago Press.

Walker, Robert S. and Lincoln A. Ribeiro (2011). "Bayesian Phylogeography of the Arawak Expansion in Lowland South America". In: *Proceedings of the Royal Society* 278.1718, pp. 2562–2567.

Walker, Robert S., Søren Wichmann, et al. (2012). "Cultural Phylogenetics of the Tupi Language Family in Lowland South America". In: *PLoS ONE* 7.4, e35025.

Warnow, Tandy et al. (2004). "Stochastic Models of Language Evolution and an Application to the Indo-European Family of Languages". In: *Technical Report, Department of Statistics, University of California, Berkeley*.

Wiens, John J. (Aug. 2003). "Missing Data, Incomplete Taxa, and Phylogenetic Accuracy". In: *Systematic Biology* 52.4, pp. 528–538.

Wiens, John J and Matthew C Morrill (Oct. 2011). "Missing data in phylogenetic analysis: reconciling results from simulations and empirical data." In: *Systematic Biology* 60.5, pp. 719–31.

Wiley, EO and BS Lieberman (2011). *Phylogenetics: theory and practice of phylogenetic systematics*. Hoboken, N.J: Wiley-Blackwell.

Appendix A

Tupí-Guaraní wordlist

Below is the full wordlist that was used in the study with English, Spanish, Portuguese, and French glosses. Some Spanish glosses reflect Peruvian or other varieties of South American Spanish. Similarly, some Portuguese glosses are Brazilian Portuguese and a few French glosses are in the Creole language of French Guyana.

English	Spanish	Portuguese	French
one	uno	um, uma	un, une
two	dos	dois, duas	deux
three	tres	três	trois
four	cuatro	quatro	quatre
five	cinco	cinco	cinq
head	cabeza	cabeça	tête
brain	cerebro	cérebro	cerveau, cervelle
hair	cabello (pelo)	cabelos	cheveux, poil
face	cara (rastro)	face, cara, rosto, rastro	visage
cheek	mejilla	bochecha	joue
mouth	boca	boca	bouche
teeth, tooth	diente (muela)	dente	dent, dents
lip	labio	lábio	lèvre
tongue	lengua	língua	langue
throat	garganta	garganta	gorge
neck	cuello	pescoço	cou
eye	ojo	olho	oeil
eyebrow	ceja	sobrancelha	sourcil
nose	nariz	nariz	nez
ear	oreja (oído)	orelha, ouvido	oreille
beard, moustache, facial hair	barba, bigote, vello facial	barba, bigode, pelo facial	barbe, moustache
jaw	mandíbula	mandíbula/maxilar	mâchoire
chin	barbilla (cachete)	queixo	menton
breast	mama (seno)	seio	sein
chest	pecho	peito	poitrine
lung	pulmón	pulmão	poumon
heart	corazón	coração	coeur
belly	vientre (barriga)	barriga (ventre)	ventre
stomach	estómago	estômago	estomac
intestines, guts	intestino (entrañas)	intestinos, entranhas	entrailles, intestins
kidney	riñón	rins	rein
liver	hígado	fígado	foie
spleen	bazo	baço	rate
rib	costilla	costela	côte
back	espalda	costas/dorso	dos
nape, back of neck	nuca	nuca, cogote	nuque
belly button	ombligo	umbigo	nombril

English	Spanish	Portuguese	French
testicles	testiculo	testículos	testicule
penis	pene	pênis	pénis
vagina	vagina	vagina	vagin, vulve
groin	ingle, entrepierna	virilha, entrepernas	aine
buttocks	nalga, culo	nádega, trazeiro, anca	fesse
anus	ano	ânus	anus
shoulder	hombro	ombro	épaule
arm	brazo	braço	bras
elbow	codo	cotovelo	coude
hand	mano	mão	main
finger	dedo	dedo	doigt
fingernail	uña	unha	ongle
foot	pie	pé	pied
heel	talón	calcanhar	talon
toe	dedo de pie	dedo do pé	orteil, doigt de pied
toenail	uña de pie	unha do pé	ongle de pied
leg	pierna	perna	jambe
thigh	muslo	coxa	cuisse
knee	rodilla	joelho	genou
ankle	tobillo	tornozelo	cheville
blood	sangre	sangue	sang
vein, artery	vena, arteria	veia, artéria	veine, artère
skin	piel	pele	peau
bone	hueso	osso	os
corn	maíz	milho	maïs
manioc	yuca	mandioca	manioc
plantains	plátano	banana-da-terra, tachagem	plantain
banana	banana	banana	banane
pineapple	piña	abacaxí (ananás)	ananas
tobacco	tabaco	tabaco	tabac
ají, pepper	ají, pimiento, chile	pimenta	piment
fruit	fruta (huayo)	fruta	fruit
gourd, pumpkin, squash	calabaza (zapallo)	abóbora, cabaça	gourde, citrouille
food	comida	comida	nourriture
fat (lard)	manteca	banha	matière grasse, graisse, gras
manioc beer	masato (chicha)	cauim	cachiri

English	**Spanish**	**Portuguese**	**French**
honey	miel	mel	miel
salt	sal	sal	sel
meat	carne	carne	viande
corn drink	chicha de maíz	bebida de milho, chicha	cachiri de mais
saliva	saliva	saliva/baba	salive
Mauritia flexuosa	aguaje	buriti	*Mauritia flexuosa*
flour/manioc flour	harina	farinha	farine
sweet potato	camote (batata)	batata doce	patate douce
achiote	achiote (urucu)	urucum	*Bixa orellana*, roucou
peanut	maní (cacahuete)	amendoim	*Arachis hypogeaea*, arachide
passion fruit	granadilla (maracuyá)	maracujá	*Passiflora edulis*, fruit de la Passion
sugar cane	caña de azucar	cana de açúcar	*Saccharum officinarum*, canne à sucre
snake	culebra, serpiente (víbora)	cobra (víbora, serpente)	serpent
agouti	añuje	cutia, cotia	Dasyprocta, Myoprocta, *Cuniculus* sp., agouti
monkey	mono	macaco	singe
howler monkey	mono aullador, mono coto	bugio, macaco-uivador, guariba	singe hurleur
spider monkey	mono araña, maquisapa	macaco-aranha, coatá	singe atèle, atèle
sloth	perezoso, pelejo, perico	preguiça	paresseux, ais (Bradypodidae), unaus (Megalonychidae)
tapir	tapira (sachavaca)	anta	*Tapirus terrestris*, tapir
jaguar	tigre, jaguar	onça	*Panthera onca*
majás (paca)	paca, majás	paca	*Cuniculus paca*
bird	pájaro (ave)	pássaro (ave)	oiseau
ant	hormiga	formiga	fourmi

English	Spanish	Portuguese	French
caiman	caimán, lagarto	caimão/jacaré (alligator)	genera *Caiman*, *Melanosuchus*, *Paleosuchus*
armadillo	armadillo, carachupa	tatu	tatou, family Dasypodidae
vulture	buitre, gallinazo	urubu	vautour, family Cathartidae
termite	comején	cupim	termite
deer	venado, ciervo	veado	cerf, family Cervidae
spider	araña	aranha	araignée
wasp	avispa	vespa	guêpe
bee	abeja	abelha	abeille
louse	piojo	piolho	pou
mosquito	zancudo	pernilongo, carapanã	moustique
gnat	jején	mosquito, mosquitinho, pium	(no gloss)
fly	mosca	mosca	mouche
sandfly	manta blanca	maruim, pium	phlébotome
horsefly	tábano	mutuca, butuca, moscardo	taon
tick	garrapata	carrapato	Acari, Ixodidae
chigger	nigua	bicho-do-pé	aoûtat, rouget, vendangeon
flea	pulga	pulga	puce
cockroach	cucaracha	barata	cafard, blatte, family Blattidae
grasshopper	saltamontes	gafanhoto	criquet, sauterelle
cricket	grillo	grilo	grillon
cicada	cigarra, chicharra	cigarra, cigarrinha	cigale
fish	pez (pescado)	peixe	poisson
piranha	piraña, paña, palometa	piranha	pirais
earthworm	lombriz	minhoca	ver de terre
botfly larva		berne	
caterpillar	oruga	lagarta	chenille
centipede	ciempiés (escolopendra)	lacraia, centopéia	scolopendre, mille pattes, chilopodes, centipèdes

English	Spanish	Portuguese	French
millipede	ciempiés, milpiés	centopéia, piolho-de-cobra, milípede	mille-pattes, diplopode
firefly	luciérnaga	vaga-lume, pirilampo	luciole, lampyre
coati	coati (achuni)	quati	coati
mouse, rat	ratón, rata	camundongo/ camondongo, rato	souris, rat
anteater	oso hormiguero	tamanduá	fourmilier
bat	murciélago	morcego	chauve-souris
butterfly	mariposa	borboleta	papillon
collared peccary (Pecari tajacu)	sajino	caititu	pécari à collier
white-lipped peccary (Tayassu pecari)	huangana	queixada, taiaçu	pécari à lèvres blanches
tail	cola, rabo	cauda, rabo	queue
feather	pluma	pena (pluma)	plume
horn	cuerno (cacho)	chifre	corne
egg	huevo	ovo	oeuf
wing	ala	asa	aile
animal	animal	animal	animal
domesticated animal, pet	animal domesticado, mascota	animal domesticado	animal domestique, animal familier
beak	pico	bico	bec
mother	madre	mãe	mère
father	padre	pai	père
son	hijo	filho	fils
daughter	hija	filha	fille
husband	marido, esposo	marido, esposo	mari, époux
wife	esposa	esposa	épouse, femme
grandchild	nieto, nieta	neto, neta	petit-fils, petite-fille
grandmother	abuela	avó	grande-mère
grandfather	abuelo	avô	grand-père
man	hombre	homem	homme
woman	mujer	mulher	femme
old woman	vieja, mujer vieja	velha, mulher velha	vieille femme
person, human	persona, humano	pessoa, ser humano, humano	personne, être humain

English	Spanish	Portuguese	French
white person, non-Indian	persona blanca, blanco, extranjero, foráneo	pessoa branca, branco, estrangeiro, forâneo	personne de race blanche
young man	joven	rapaz	jeune homme
young woman	muchacha, chica	moça	jeune femme
boy, male child	muchacho, niño	moço, garoto, menino, criança	garçon
girl, female child	chica, niña (muchacha)	moça, menina, criança	fille
minga	minga	mutirão	
orphan	huérfano	orfão	orphelin
owner, boss (master)	dueño, propietario	dono (propietário)	propriétaire, maître
spirit (soul)	espíritu (alma)	espírito (alma)	esprit
friend	amigo	amigo	ami
companion	compañero	companheiro	
witch, shaman, healer	brujo, bruja, shamán, hechicero, hechicera, curandero, médico	bruxo, bruxa, pajé, feiticero, feitcera, curandeiro	sorcier, sorcière, chamane
chief (leader)	jefe (cacique, apu)	chefe; cacique	chef
red	rojo	vermelho (m.)	rouge
ripe	maduro	maduro	mûr
white	blanco	branco	blanc
black	negro	preto	noir
dark	oscuro	escuro	sombre
unripe	no maduro, verde, flaco	não maduro, verde	vert, immature, pas mûr
green	verde	verde	vert
blue	azul	azul	bleu
grue			
yellow	amarillo	amarelo	jaune
yesterday	ayer	ontem	hier
tomorrow	mañana	amanhã	demain
day	día	dia	jour
today	hoy	hoje	aujourd' hui
long time ago	hace mucho tiempo, antiguamente	há muito tempo, antigamente	il y a longtemps
year	año	ano	an, année

English	Spanish	Portuguese	French
night (evening)	noche	noite	nuit
dawn	amanacer	amanhecer	aube, aurore
morning	mañana	manhã	matin
afternoon	tarde	tarde	après-midi
later	más tarde	mais tarde	plus tard
river	rio	rio	rivière, fleuve
water	agua	água	eau
rain	lluvia	chuva	pluie
lake	lago (cocha)	lago	lac, mare
stream	quebrada, arroyo (riachuelo)	riacho, ribeiro	ruisseau
shore, bank	orilla (banda, margen)	costa, margem (ribanceira)	bord, berge
forest (jungle)	selva (bosque)	mato	forêt
root	raíz	raiz	racine
tree	árbol	árvore	arbre
branch	rama	ramo, galho	branche
bark	corteza	casca de ávore	écorce
trunk	tronco	tronco	tronc
flower	flor	flor	fleur
leaf	hoja	folha	feuille
bamboo	bambú, caña	bambu	bambou
grass	hierba	capim, grama, gramado	herbe
weed	maleza	erva daninha	mauvaise herbe
vine	soga	cipó	liane, plante grimpante
land, ground, earth	tierra	terra	terre
island	isla	ilha	île
shadow, shade	sombra	sombra	ombre
mud	barro (barreal, fango)	lama	boue, vase
dust	polvo	poeira	poussière
sand	arena	areia	sable
cotton	algodón	algodão	coton
cloud	nube	nuvem	nuage
thunder	trueno, tronar	trovão, trovejar	tonnerre, tonner
sky	cielo	céu	ciel
sun	sol	sol	soleil
moon	luna	lua	lune

English	Spanish	Portuguese	French
star	estrella	estrela	étoile
Pleiades	Pléyades	Pleiades	Pléiades
seed	semilla	semente	graine
stone, rock	piedra (roca)	pedra (rocha)	pierre
wind	viento	vento	vent
hill	colina (altura)	morro, colina	colline
cutbank, cliff, bluff	barranco, acantilado, precipicio	barranco, penhasco, falésia	falaise, escarpement
hole	hueco	buraco	trou
candela	fuego (candela)	fogo	feu
ashes	cenizas	cinza	cendre
smoke (n.)	humo	fumaça	fumée
sap (resin)	resina (brea)	seiva (resina)	sève
canoe	canoa	canoa	canot
prow	proa	proa	proue
hammock	hamaca	rede	hamac
grill, barbeque	parilla, barbacoa	grelha/churrasco (jirau)	gril, boucan
wall, fence	pared, cerca	parede, cerca	mur, clôture, barrière
door	puerta	porta	porte
house	casa	casa	maison
tambo, shelter	tambo	abrigo, tambo	abri
patio (porch)	patio	patio	patio, cour
beach	playa	praia	plage
village	pueblo	aldeia	village
port	puerto	porto	port
path, road	camino	caminho (trilha)	chemin
chacra	chacra	roça	champ, potager, jardin, abbatis (plantation)
cure, medicine	remedio	remédio	remède
wound	herida	ferida	blessure, plaie
flute	flauta, pífano	flauta	flûte
fan	ventilador (abanico)	leque, ventilador	éventail
bead	abalario, mostacilla	miçanga	perle (de verre)
broom	escoba	vassoura	balai
digging stick	azada	enxada	bâton à creuser, houe

English	Spanish	Portuguese	French
paddle (n.)	remo	remo	pagaie
cloth	paño (tela)	pano	tissu
garbage	basura	lixo	ordures
firewood	leña	lenha	bois de chauffage, bois à brûler
grindstone	piedra de moler	pedra de amolar	meule
mortar	mortero	pilão, almofariz	mortier (Pilon in Créole)
pestle	mano de mortero, mazo	mão de pilão,	pilon (Bâton-pilon in Créole)
thread	hilo	linha (fio)	fil
rattle, maraca	maraca, sonaja, sonajero, shacapa	maracá, chocalho	hochet
rope	cuerda	corda	corde
fishing net	red de pescar, tarrafa	rede de pescar	filet de pêche
poison	veneno	veneno	poison
fish poison	barbasco	timbó	poison à pêcher
fish hook	anzuelo	anzol	hameçon
spear	lanza	lança	lance (sagaie, javelot)
axe	hacha	machado	hache
bow	arco	arco	arc
arrow	flecha	flecha	flèche
knife	cuchillo	faca	couteau
cup	copa (huingo, cuya)	copo (cuia)	tasse
crock (earthenware)	cántaro, tinaja	cântaro, pote	cruche, pichet, pot
pot	olla	panela	marmite
sieve (strainer)	tamiz, cernidor, colador	peneira	tamis
basket	cesto, canasta	cesto (canasta)	panier
thorn	espina	espinho	épine
pain (n.)	dolor	dor	douleur
work (n)	trabajo	trabalho	travail
amount	cantidad, porción	porção	quantité
beginning	comienzo	começo	début
thing	cosa	coisa	chose
name (n.)	nombre	nome	nom
story	cuento, historia	conta, estória	histoire

English	Spanish	Portuguese	French
good	bueno	bom, boa	bon
ugly	feo	feio	laid
bad	mal	ruim (mau)	mauvais, méchant
big, large	grande	grande	grand, gros
small	pequeño	pequeno	petit
fat	gordo	gordo	gros
narrow	estrecho, angosto	estreito	étroit
skinny, thin	flaco	magro	maigre, mince, fin
dirty	sucio	sujo	sale
clean	limpio	limpo	propre
wet	mojado	molhado	mouillé
dry	seco	seco	sec
slow	lento	lento, devagar	lent
fast	rápido	rápido	rapide
cold	frío	frio	froid
hot	caliente	quente	chaud
new	nuevo	novo	nouveau, neuf
old	viejo (anciano)	velho	vieux
short	corto, bajo	curto, baixo	court, petit
long	largo	comprido	long
heavy	pesado	pesado	lourd
spicy	picante	picante	épicé, piquant, piménté
sour	agrio	azedo	aigre
bitter	amargo	amargo	amer
sweet	dulce	doce	doux
tasty	sabroso	saboroso	savoureux
rotten	podrido	podre	pourri
full	lleno	cheio	plein
hungry	tener hambre (hambriento)	faminto, estar com fome/ter fome	avoir faim
tired	cansado	cansado	fatigué
quiet, calm	tranquilo, calmado (tranquilo, callado)	quieto, calmo	calme
sick	enfermo	doente	malade
lazy	perezoso, flojo	preguiçoso	paresseux
poor	pobre	pobre, coitado	pauvre
drunk	borracho	bêbado	ivre
naked	desnudo	nu, despido	nu

English	Spanish	Portuguese	French
angry, mad, be angry (intr)	enojado (rabiado, fastidiado), enojarse (rabiarse, fastidiarse)	zangado, com haiva, furioso, ficar com raiva (estar irritado)	fâché, furieux, être en colère, être fâché
happy	alegre, feliz, contento	feliz, alegre, contente	heureux, gai
weak (feeble)	débil	fraco	faible
strong	fuerte	forte	fort
hard	duro	robusto, duro	dur
thick	grueso	grosso	épais
round	redondo	redondo	rond
straight	recto, ser recto (derecho)	reto, direito	droit, être droit
right	derecha	(a direita)	(droite)
left	izquierda	esquerda	gauche
bent, twisted	torcido	torcido	courbé, tordu
far	lejos	longe	loin
near (nearby)	cerca	perto	près
open	abierto	aberto	ouvert
wide	amplio	largo, ampio	large
painted (spotted)	pintado	pintado	peint
dull	desafilado, romo, embotado	sem corte, cego (blunt)	émoussé
sharp	afilado	afiado	affûté, tranchant, effilé, pointu
smooth	suave, liso	liso, macio	lisse, doux
murky	turbio	turvo	boueux
correct	correcta	correto	correct
high up (adv.)	arriba, en el alto	no alto	en haut
high, tall	alto	alto	grand, haut
deep	hondo	fundo	profond
deaf	sordo	surdo	sourd
blind	ciego	cegos	aveugle
again	otra vez (de nuevo, nuevamente)	de novo (outra vez, novamente)	encore, de nouveau
still	todavía	ainda	encore
also	también	também	aussi
more	más	mais	plus
other	otro	outro	autre
many (much)	muchos	muito	beaucoup de

English	Spanish	Portuguese	French
all	todo	todo	tous, tout
some	algun(os)	alguns	quelques uns, certains
few (a little)	poco(s)	pouco(s)	un peu
first	primero	primeiro	premier
no	no	não	non
side	lado	lado	côté
go	ir	ir	aller
go (IMP)	ir (IMP)	ir (IMP) = vá	vas (IMP)
come (intr)	venir	vir	venir
arrive (intr)	llegar	chegar	arriver
leave from (intr)	salir	sair	sortir, partir
return (intr)	regresar, volver	voltar, retornar	retourner, revenir
stop (intr, trans)	parar	parar	arrêter, s' ârreter
stay (intr)	quedar	ficar	rester
walk (intr)	caminar, andar	andar, caminhar	marcher
run (intr)	correr	correr	courir
swim (intr)	nadar	nadar	nager
float	flotar, boyar	boiar, flutuar	flotter, faire la planche
plant (trans)	plantar, sembrar	plantar	planter
bury (trans)	enterrar	enterrar	
steal (trans)	robar	roubar	voler
hide (intr, trans)	esconder (ocultar)	esconder (ocultar)	cacher, se cacher
cultivate (trans)	cultivar	cultivar	cultiver
dig (trans)	excavar (cavar, huequear)	cavar, escavar	creuser
drip (intr)	gotear	gotejar	goutter
flow, run off (liquids)	fluír, escurrir	fluir, escorrer	s' écouler, couler
slip (intr)	resbalarse	escorregar	glisser
sweat (intr)	sudar	suar	suer, transpirer
borrow	tomar/pedir prestado	tomar/pegar/pedir emprestado	emprunter
live (intr)	vivir	viver	vivre
die (intr)	morir	morrer	mourir
kill (trans)	matar	matar	tuer
fly (intr)	volar	voar	voler
climb (trans)	subir	subir	monter

English	Spanish	Portuguese	French
cut, cut down, fell tree (trans)	cortar (rebanar, tumbar, derrumbar)	cortar, derrumbar	couper, abattre, couper
break (trans)	romper, quebrar	quebrar	casser
split (trans)	partir(se)	partir, rachar	fendre
divide, separate	dividir, separar	dividir, separar, repartir	partager, diviser, séparer
burst, break open	reventar, estallar	rebentar, arrebentar, estourar, estalar	éclater
fish (trans)	pescar	pescar	pêcher
hunt (intr, trans)	cazar (mitayar)	caçar	chasser
sweep (intr, trans)	barrer	varrer	balayer
shake (intr, trans)	sacudir, agitar	sacudir, agitar	agiter, secouer
rain (v.) (intr)	llover	chover	pleuvoir
flood (intr, trans)	inundar	inundar, transbordar	inonder
drown (intr)	ahogarse	afogar-se	noyer
sink (intr)	hundirse	afundar	couler (un canot)
breathe (intr)	respirar	respirar	respirer
yawn (intr)	bostezar	bocejar	bâiller
vomit (intr)	vomitar	vomitar	vomir
lie on a surface (intr)	echarse, acostarse	deitar(-se)	se coucher, s' allonger
put (ditr)	poner (colocar)	pôr, colocar	mettre
turn (intr, trans)	girar	virar	tourner, retourner
twist (intr, trans)	torcer	torcer	tordre
escape (intr)	escapar(se)	escapar	s' échapper
leave (trans)	dejar	deixar	quitter
enter (intr)	entrar	entrar	entrer
cough(intr)	toser	tossir	tousser
swallow	tragar	engolir	avaler
capture (trans)	capturar	capturar	capturer, prendre
appear (intr)	aparecer	aparecer	apparaître
disappear (intr)	desaparecer	sumir	disparaître
lose way	errar	errar o caminho	
lose (oneself) (intr)	perderse (errar)	perder-se, errar	être perdu, se perdre
help (trans)	ayudar	ajudar	aider

English	Spanish	Portuguese	French
fear (n.), be scared (v.)	miedo, tender miedo	medo, temer, ter medo	peur, avoir peur
scare (trans)	asustar (espantar)	assustar, espantar	faire peur
be able to (trans)	poder	poder	pouvoir
bring (trans)	traer	trazer	apporter, amener
pull (trans)	jalar	puxar	tirer
shoot (trans)	disparar (flechar)	disparar, flechar (atirar)	tirer (au fusil)
grab (trans)	agarrar, arrancar	agarrar, arrancar	attraper, saisir
take out, remove (trans)	sacar	tirar, sacar	enlever, arracher,
beat, hit (trans)	pegar, golpear	bater, acertar	battre, frapper
reach (trans)	alcanzar	alcançar	atteindre
haber	haber	ter	il y a
have (trans)	tener	ter	avoir
give (ditr)	dar	dar	donner
throw, throw away (trans)	botar, tirar, arrojar (echar)	jogar, jogar fora	jeter, lancer
get used to (trans)	acostumbrarse	acostumar-se	s' habituer
lie to, deceive (trans)	mentir (engañar)	mentir (enganar)	mentir
forget (trans)	olvidar	esquecer, olvidar	oublier
know (trans)	saber (conocer)	saber	savoir
look for (trans)	buscar	procurar, buscar	chercher
look (intr)	mirar	olhar	regarder
see (intr, trans)	ver	ver	voir
order, command	ordenar, (co)mandar	ordenar, (co)mandar	commander, donner ordres
meet, encounter, come across, find	encontrar(se), hallar	encontrar(-se), deparar(-se), achar	rencontrer, trouver
remember (trans)	acordarse de, recordar(se) (de)	lembrar	se souvenir
think (intr, trans)	pensar	achar, pensar	penser
want/like/love	querer/gustar/amar	querer/gostar/amar	vouloir/aimer
copulate, have sex with	copular, tener relaciones sexuales, aparear	copular, ter relações sexuais	copuler, faire l' amour, coucher avec qqn
believe (trans, comp)	creer	acreditar	croire
obey (trans)	obedecer	obedecer	obéir

English	Spanish	Portuguese	French
hear (trans)	escuchar, oír	escutar, ouvir	entendre
fight (trans)	luchar, pelear	lutar, brigar	combattre
follow (trans)	seguir	seguir	
respond (trans)	responder	responder	répondre
sing (intr)	cantar	cantar	chanter
ask for (ditr?)	pedir	pedir	demander
ask (trans)	preguntar	perguntar	demander
call (trans)	llamar a	chamar	appeler
converse (trans)	conversar	conversar	discuter, parler avec quelqu'un
tell (a story) (trans)	contar (narrar)	contar (narrar)	raconter
say (trans)	decir	dizer	dire
speak (utter) (intrans)	hablar	falar	parler
emit noise, sound	sonar	soar	sonner, faire du bruit
imitate, represent, signal	imitar, representar, ser un señal	imitar, representar, sinalizar	imiter, representer, symbolizer
dance (intr)	bailar	dançar	danser
wake up (intr)	despertarse	despertar-se	se réveiller
sleep (intr)	dormir	dormir	dormir
laugh (intr)	reir	rir	rire
smile (intr)	sonreír	sorrir	sourire
cry (intr)	llorar	chorar	pleurer
mourn	lamentar, lloriquear, sollozar	prantear, lamentar, chorar	lamenter, faire un deuil, pleurer pour un deuil
be sad (intr)	estar triste	estar triste	être triste
be thirsty (intr)	tener sed	estar com sede	avoir soif
defecate (intr)	defecar	defecar	déféquer, chier
urinate (intr)	orinar	urinar	uriner
cook (intr, trans)	cocinar (cocer)	cozinhar	cuisiner, cuire
eat (intr, trans)	comer	comer	manger
drink (intr, trans)	beber	beber	boire
drink alcohol	tomar/beber alcohol	tomar/beber alcohol	boire du cachiri, boire de boisson alcoolisée, boire d' alcool
roast (trans)	asar	assar	griller, rôtir

English	Spanish	Portuguese	French
roast fariña (trans)	turrar	torrar	griller
smoke food	ahumar	defumar, moquear, afumar	fumer, boucaner
burn	quemar(se)	queimar(-se)	brûler
flame (v.), shine (as in giving off light)	arder	arder	brûler, éclairer
light (v.)	encender	acender	allumer
shine	brillar	brilhar	briller, eclairer
illuminate, shine light on	iluminar	iluminar	illuminer, éclairer
fry (trans)	freír	fritar	frire
boil (trans)	hervir	ferver	bouillir
bubble	borbotar, borbotear, burbujar	borbotar, borbulhar	bouilloner
swell (intr)	hincharse	inchar	gonfler, enfler
strain (trans)	colar (cernir)	coar, peneirar	égoutter
mix (trans)	mezclar, batir	misturar	mélanger
draw (intr, trans)	dibujar	desenhar	dessiner
paint (intr, trans)	pintar	pintar	peindre
dye	teñir	tingir	teindre, colorer
adorn, decorate	adornar, decorar, engalanar	adornar, decorar, enfeitar	décorer, orner
grow (intr, trans)	crecer	crescer	grandir
heal (intr, trans)	sanar	sarar	guérir
pass (trans)	pasar	passar	passer
cross, go across	cruzar, atravesar, vadear	cruzar, atravessar, vadear	traverser
pierce (trans)	perforar	furar	percer, transpercer
sew (trans)	coser	costurar	coudre
weave	tejer, trenzar	tecer, trançar	tisser
play (intr, trans)	jugar	brincar, jogar (game), folgar	jouer
pour (trans)	derramar (echar aguar)	derramar	verser
jump (intr)	saltar (brincar)	pular	sauter
push (trans)	empujar	empurrar	pousser
rub (trans)	sobar	esfregar	frotter
scrape, scratch (trans)	raspar, arañar	arranhar, coçar	gratter, racler
grind (trans)	moler	moer, trituar	moudre

English	Spanish	Portuguese	French
spin (thread) (trans)	hilar	fiar	filer
sit (intr)	sentarse	sentar-se	s' asseoir
stand, stand up (intr)	estar de pie (pararse), levantarse	ficar de pé, estar de pé, levantar-se	être debout
smell (intr, trans)	oler	farejar, cheirar	sentir
be odorous (fragrant) (intr)	estar oloroso	estar cheiroso	sentir bon
be stinky (smell bad) (intr)	apestar (heder)	feder	puer
squeeze (trans)	comprimir (estrujar)	comprimir, espremer	presser
tighten (trans)	apretar (asegurar)	apertar	serrer, resserrer
stretch	estirar, alargar	espichar, esticar, alongar, distender	s' étirer, tendre, s' étendre, distendre
paddle (intr)	remar (bogar)	remar	pagayer
go upriver (intr)	surcar (ir aguas arriba)	ir na contracorrente/sembem o rio	remonter le fleuve, monter la riviere
go downriver (intr)	bajar (ir aguas abajo)	ir rio abaixo	aller en aval, descendre le courant
go down (intr)	bajar (descender)	descer	descendre
be pregnant (intr)	estar embarazada, estar preñada	estar grávida	être enceinte
fall from a height (intr)	caer del alto	cair do alto	tomber de haut
fall over, fall down (intr)	caer	cair	tomber
give birth to (trans)	dar a luz a (parir)	dar á luz a	donner naissance, accoucher
be born (intr)	nacer	nascer	naître
breastfeed (trans)	mamar	mamar	allaiter
suck (trans)	chupar	chupar	sucer
bite (trans)	morder	morder	mordre
spit (intr)	escupir	cuspir	cracher
blow (trans)	soplar	soprar	souffler
wet, dampen (trans)	mojar	molhar	mouiller
dry (intr, trans)	secar	secar	sécher

English	Spanish	Portuguese	French
await (trans)	esperar	esperar	attendre
hurt, wound	herir(se), lastimar(se)	ferir(-se), machucar(-se)	blesser
dock (V trans, intr)	atracar (a tierra)	atracar	mettre a quai,
attach, adhere, stick to	sujetar(se), pegar(se), adherir(se), arrimar(se)	aderir(-se), colar, grudar	attacher, lier, adhérer, coller
join, (put together, attach) (trans)	juntar, unir	juntar	joindre, unir, raccorder, lier, attacher
tie (trans)	amarrar, atar	amarrar (atar)	nouer, lier, attacher
untie (trans)	desatar	desatar, desamarrar	défaire, dénouer
gather, harvest (trans)	cosechar (recoger)	recolher, juntar	ramasser, cueillir, moissonner
finish (trans)	acabar, terminar (completar)	acabar	finir
begin (v.) (trans, comp)	empezar, comenzar	começar	commencer
feed, offer food (trans)	alimentar	alimentar	nourrir, donner à manger
weed (intr, trans)	desyerbar	capinar	désherber
rest (intr)	descansar	descançar	se reposer
carry (trans)	cargar, amarcar (acarrear)	carregar	porter

Appendix B

Ancestral State Reconstructions

Below are the cognate sets that are reconstructed as lost or gained for selected nodes on the majority-rule consensus tree of our main analysis (without Turiwará and Apiaká). The list is not exhaustive for each node, but it includes all cognates that can be reconstructed as lost or gained on that node with high likelihood ($> 90\%$). Cognate gains not followed by subsequent losses in any language and unique cognate losses (i.e. not lost independently in other subgroups) are indicated with an asterisk. The forms in the tables below are given as examples of the respective cognate set and in the orthography of the source.

Table B.1: Ancestral state reconstructions for the Guaranian subgroup

character number	cognate set	reconstruction	form (language)
493	tapir1	gain	mborevi (PG)
668	anteater6	gain	kaguare (Mbya)
1869	dry2	gain	ypi (PG)
2358	open1	gain	ojei (PG)
3119	deceive1	gain	japu (PG)
910	chief8	loss	morerecoara (tpn)
958	dark2	loss	pihun (Tembe)
1873	dry6	loss	tubyra (tpn)[1]
4111	clean(v)2	loss	quyt$\tilde{i}$(ng)oca (tpn)

[1] it means dust in Tupinambá

Table B.2: Ancestral state reconstructions for the Southern subgroup.

character number	cognate set	reconstruction	form (language)
341	pineapple3	gain	karaguata (PG)
680	bat1	gain	mbopi (PG)
1450	digging stick1	gain	sype (PG)
3275	follow1	gain	(a)moña (PG)
4167	embrace3	gain	/kwãwa/ (Chiriguano)
480	howler monkey1	gain*	karaja (PG)
368	gourd3	loss	cuia (tpn)
389	lard4	loss	caba (tpn)
483	howler monkey4	loss	aquyquy (tpn)
533	deer1	loss	/ɨtɨ:/ (maw)
982	yellow5	loss	taguá (tpn)
2214	weak6	loss	membeca (ton)
2322	far5	loss	amõ (tpn)
2497	stop6	loss	pyca (tpn)
3105	throw4	loss	eityca (tpn)
3493	light(v)3	loss	mondyca (ton)
3967	finish6	loss	syca (tpn)

Table B.3: Ancestral state reconstructions for the Diasporic subgroup.

character number	cognate set	reconstruction	form (language)
1414	flute1	gain	mimby (tpn)
2303	bent, twisted3	gain	banga (tpn)
2318	far1	gain	mombyry (PG)
2601	flow2	gain	sururu (Chiriguano)
3397	mourn1	gain	apirõ (tpn)
3600	pass2	gain	puana (tpn)
3745	be stinky1	gain	timbora (ton)
4132	touch3	gain	atõia (tpn)
4143	shake2	gain	myia (tpn)

Table B.4: Ancestral state reconstructions for the Nuclear TG subgroup.

character number	cognate set	reconstruction	form (language)
52	cheek3	gain	atypy (tpn)
449	peanut1	gain	mandubi (tpn)
808	old woman1	gain	waiw (kay)
871	spirit2	gain	anhanga (tpn)
1155	island1	gain	ypaũ (tpn)
1677	thing2	gain	marã (tpn)
2003	tasty1	gain	é (tpn)
2043	full4	gain	ynyssema (the compound) (tpn)
2044	full5	gain	(ynys)sema (tpn)
2559	bury1	gain	atyba (tpn)
2790	breathe1	gain	pytuẽ (tpn)
3330	say1	gain	mombeú (tpn)
3617	sew1	gain	mobybyca (tpn)
153	stomach8	loss*	ty?a (awe)
1158	island4	loss*	i-'apema (kam)
1602	pot10	loss*	ja?apehẽ (kam)