

Kinbank: A global database of kinship terminology

Sam Passmore^{1,2,3*}, Wolfgang Barth⁴, Simon Greenhill^{4,5}, Kyla Quinn⁴, Catherine Sheard^{2,6}, Paraskevi Argyriou⁷, Joshua Birchall^{8,9}, Claire Bowern¹⁰, Jasmine Calladine^{2,11}, Angarika Deb^{2,12}, Anouk Diederer¹³, Niklas P. Metsäranta¹⁴, Luis Henrique Araujo¹⁵, Rhiannon Schembri¹⁶, Jo Hickey-Hall², Terhi Honkola^{2,14}, Alice Mitchell^{2,17}, Lucy Poole², Péter Rácz^{2,18}, Sean G Roberts^{2,19}, Robert M. Ross²⁰, Ewan Thomas-Colquhoun², Nicholas Evans^{1,4*}, Fiona M. Jordan^{2*}

¹ Evolution of Cultural Diversity Initiative (ECDI), Australian National University, Canberra, Australia

² Department of Anthropology and Archaeology, University of Bristol, Bristol, UK

³ Faculty of Environment and Information Studies, Keio University, Fujisawa, Japan

⁴ ARC Centre of Excellence for the Dynamics of Language (CoEDL), Australian National University, Canberra, ACT, Australia

⁵ Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

⁶ School of Earth Sciences, University of Bristol, Bristol, UK

⁷ School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK

⁸ Museu Paraense Emílio Goeldi, Belém, Pará, Brazil

⁹ The University of New Mexico, NM, USA

¹⁰ Yale University, New Haven, CT, USA

¹¹ Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

¹² Department of Cognitive Science, Central European University, Vienna, Austria

¹³ Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

¹⁴ Department of Finnish, Finno-Ugrian, and Scandinavian Studies, University of Helsinki, Helsinki, Finland

¹⁵ Universidade Federal do Pará, Belém, Pará, Brazil

¹⁶ Research School of Biology, Ecology, and Evolution, Australian National University, Canberra, Australia

¹⁷ Institute for African Studies, University of Cologne, Cologne, Germany

¹⁸ Cognitive Science Department, Budapest University of Technology and Economics, Budapest, Hungary

¹⁹ School of English, Communications and Philosophy, Cardiff University, Cardiff, UK

²⁰ School of Psychological Sciences, Macquarie University, Sydney, Australia

*Corresponding Authors: Sam Passmore, Nicholas Evans, and Fiona Jordan

Email: Samuel.passmore@anu.edu.au (SP); Nicholas.evans@anu.edu.au (NE); Fiona.jordan@bristol.ac.uk (FMJ)

Abstract

For a single species, human kinship organization is both remarkably diverse and strikingly organized. Kinship terminology is the structured vocabulary used to classify, refer to, and address relatives and family. Diversity in kinship terminology has been analyzed by anthropologists for over 150 years, although recurrent patterning across cultures remains incompletely explained. Despite the wealth of kinship data in the anthropological record, comparative studies of kinship terminology are hindered by data accessibility. Here we present Kinbank, a new database of 210,903 kinterms from a

global sample of 1,229 spoken languages. Using open-access and transparent data provenance, Kinbank offers an extensible resource for kinship terminology, enabling researchers to explore the rich diversity of human family organization and to test longstanding hypotheses about the origins and drivers of recurrent patterns. We illustrate our contribution with two examples. We demonstrate strong gender bias in the phonological structure of parent terms across 1,022 languages, and we show that there is no evidence for a coevolutionary relationship between cross-cousin marriage and bifurcate-merging terminology in Bantu languages. Analysing kinship data is notoriously challenging; Kinbank aims to eliminate data accessibility issues from that challenge and provide a platform to build an interdisciplinary understanding of kinship.

Introduction

Human kinship organization is remarkably diverse. Our patterns of caring for each other, finding partners, and cooperating with relatives show more variation than any other species on the planet, despite kinship being anchored in biological and social constraints (1). Kinship relations in human communities are created through both basic reproductive processes and the social making of family ties which, importantly, are transmitted through language and culture (2). Central to social organization, kinship has influenced many aspects of our evolutionary history including the distribution of linguistic and genetic diversity (3), technology (4), the likelihood of external warfare (5,6), and is thought to be the factor behind purported psychological “WEIRD-ness” observed in western societies (7). Anthropological studies of kinship have traditionally encompassed marriage customs, the tracing of descent and community relations, the jurisdiction of rights and responsibilities in offspring, and the variety of residential groupings for family, with existing calls to use quantitative and evolutionary methods to understand kinship diversity, and its role in cultural evolution and linguistic change (8,9). In the last half-century this remit has broadened, including e.g. new reproductive technologies, single-parent families, LGBT+ kinship, and different kinds of relatedness (10,11). While research foci change, one consistent strand has been the linguistic denotation and organization of family members in *kinship terminology*: the patterned vocabulary of words for kin.

This paper introduces Kinbank, a free and open-source database that centralizes and systematizes global cross-cultural data on kinship terms of 1,229 spoken and signed languages, providing an accessible and comprehensive resource for documenting and analyzing kinship diversity. The organization of kinship terminology is remarkably variable across languages. Different languages distinguish relationships by using distinct terms, or they conflate the relationship by using the same term. For example, in Lau Fijian (Fiji) a woman refers to her male sibling using the same word that he refers to her (*weka* i.e. ‘opposite-sex sibling’); Ngiyambaa (Australia) only has one term for children, unlike son and daughter in English; and Dutch (Netherlands) groups parent’s siblings children (cousins), and sibling’s children (nieces and nephews) together under *neef* (m) and *nicht* (f). With Kinbank, we centralise terminology from approximately 15% of known language diversity, at least 3.28 billion speakers (47% of the speakers in Ethnologue), including speakers for 96% of the territories covered in Ethnologue, and at least 20% of every continent (12). Kinbank

contains the largest and broadest collection of terminology available, offering a platform to investigate variation in kinship terminology.

This structured variation in kinship terms has preoccupied anthropologists for 150 years (13). Virtually all ethnographic and descriptive linguistic scholars in the 20th century elicited kin terms from speakers in the communities they studied. As a result, kinship terminology is richly documented. However, these data are scattered. Terms have been collected in comparative surveys (9), documented within ethnographies (14), as word lists within grammars (15), or scattered through dictionaries and specialized articles. Bringing these data together is important, because kinship terminology studies have made valuable contributions to our understanding of social structure. For example, studies have shown how linguistic denotation indicates group marriage norms (16) and how those rules are manipulated for economic (17), or reproductive benefit (18,19). Research has used kin terms to identify cross-cultural differences in parental care roles (20) and revealed logical structures in terminology (21,22). Furthermore, variations in kinship terminology have been used to infer the structures of early human society (23) and societies in prehistory (24–26).

In specialist study, kinship terminology has been a fruitful avenue for understanding contemporary and historical relationships between language and behavior. For example, the adoption of novel algebraic methods to transcribe and interpret kinship terminologies have provided valuable insights to the intersection of genealogical and cultural understandings of kinship (27). Similarly, applying optimality theory has reframed kinship terminology diversity within a domain of rules and constraints that can generate cultural variability (1). Despite the value of these approaches, there is an increasing disconnect between the field's theoretical advancement and the data routinely used by other social sciences. The specialist knowledge needed to engage with cutting-edge kinship scholarship has meant work in fields such as psychology, linguistics, and wider anthropology largely rely on 70-year-old typologies (28), which lack cross-cultural validity (29). This disconnect perpetuates ideas that existing knowledge does not support, such as the relationship between kinship typology and social norms (30). Kinbank has the potential to re-engage diverse disciplines with the study of kinship terminology by offering streamlined access to data and sources, allowing specialists to quickly expand and generalize their findings, and centralizing key new data in a single resource.

Since the turn of the 21st century, fields such as anthropology, psychology, and linguistics have taken a (re)turn to the empirical analysis of kinship systems (1,8,28,31). Cognitive models of kinship terminology highlight strategies to uncover universal processing principles (1,32,33), and social categories of kinship are used to prescribe cultural patterns of behavior (34,35). Such studies rely on kinship terminology, but have been restricted to existing aggregated data (32), or sourcing their own terminology (33), and would be improved or aided by larger datasets. Kinship is also becoming increasingly central to economic historians, who have used kinship terminology structure to infer broader patterns of social structure (7,36). With access to broader data, members of the Kinbank project have: demonstrated links between linguistic and social organization (37), and the regional variation in these relationships (30); shown the modularity of Pama-Nyungan kinship terminology and quantified the connections between modules (38); presented a collection of terms for

languages whose kinship systems had never been fully described alongside new methods for Tupian and Cariban languages (39); used corpus-based studies to show the difference in cultural evolutionary changes between kinship words and basic vocabulary (40); used field studies to revive research on the acquisition of kinship language and concepts by children (41,42); and performed multiple studies showing how kinship finds its way into core grammar (43,44).

With access to this extended, centralised, open access dataset, researchers can leverage Kinbank to examine these and many other questions regarding kinship and kinship terminology. We offer two examples of analyses. First, we show a global analysis on forms of kin terms, testing whether we observe gender bias in the phonology of parental terms. Second, we use phylogenetic techniques to test for a co-evolutionary relationship between kinship terminology structure and patterns of cross-cousin marriage. These examples exemplify the types of methodological and theoretical approaches researchers can address using the Kinbank dataset. For example: is there any generalisable relationship between the behavioural structures and social norms of kinship and kinship terminology? What are the unattested categories of kin and what does that tell us about the constraints of language? How learnable are different terminological systems? The questions we can ask of kinship terminology are plentiful. With a comparative database, they are now also testable.

The Kinbank database

Data, sampling, and database structure

Kinbank provides a digitized, open-access, and global database of kinship terminologies, resulting from the international, multi-funder collaboration of four aligned research projects. They are: Parabank based at the ARC Centre of Excellence for the Dynamics of Language, Australian National University; VariKin at the University of Bristol; MPEGKin at the Museu Paraense Emílio Goeldi; and Kinura at the University of Helsinki. The Kinbank database is freely accessible at <https://zenodo.org/record/7232746>, as are details of how each individual entry or dataset can be cited (S2 Table).

The etic approach

Capturing the global variation in kinship terminology requires a framework from which to compare languages. While much of the recent academic literature about kinship terminologies is written in English, the variety of categorizations found cross-linguistically means that using English kinship terms to describe cross-cultural variation would be scientifically inaccurate and ethnocentric. Historically, there have been two approaches to kinship terminology: emic (language-internal logic) and etic (objective language-independent grid for comparison) (45). The emic approach seeks to unlock the inner logic of a language's kinship terms by taking locally meaningful categories of terms as a fundamental unit of interest. The etic approach relies on a language independent yardstick, often in the form of a genealogical grid of *kin types* (a genealogical position in the etic grid), which *kin terms* (the word used to describe one or more genealogical positions) are laid on top of. For example: the Kayardild (Australia)

kinship terms *kularrint* and *duujint* would be emically described as "opposite-sex sibling" and "younger same-sex sibling", but *kularrint* would be etically described as equating to man's female sibling and woman's male sibling, and *duujint* would be man's younger male sibling, and woman's younger female sibling. The emic approach provides a succinct description of the relationship and how the terms fit together within the language, but is difficult to apply cross-linguistically (e.g. opposite-sex sibling does not apply in an English terminology). The formulations used in an etic approach are less elegant, but we can more easily apply the category of man's younger male sibling to other languages (*brother* in English).

The goal of this database is to facilitate high-level cross-cultural comparison by providing the first open standardized base from which to compare kinship terms across languages, so Kinbank assumes an etic and genealogical approach. The etic approach provides a fine-grained and language-independent yardstick for the set of kin types over which each kin term can range, enabling us to quantify the similarity and differences across languages (46). Here are three examples of how this could apply to kinship terminology: 1) By measuring how far particular terms extend over etically defined referents like father, father's brother, and mother's brother, we can readily measure the extent of referential categories and compare their semantics across societies: e.g. Does *uncle* for parent's male siblings in English cover the same relatives as *oba* in Japanese (yes) or *kakuju* in Kayardilt (no)? 2) We may ask questions about the similarity of referential ranges in different subsets of kinterms: e.g. Do parental terms tend to be extended to parents' siblings more often than sibling terms to parents' siblings' children? And 3) we can also explore the structural similarity of kinship terminologies across languages (29). Using an etic grid presents the opportunity to ask more of these types of questions and to draw macro- and cross-cultural conclusions. This has been done in previous approaches, such as when Nerlove and Romney (47) used an etic grid to determine 12 types of sibling terminologies, representing 98% of the observed languages. These authors established a design space of 4,140 possible sibling terminologies, derived from an etic set of eight sibling categories derived from three rules of distinction (gender of sibling, gender of speaker, and relative age). Using Kinbank's etic structure and advances in statistical methods, we can extend this approach to larger subsets of kin and more complex methods of comparison (e.g., (29)).

The Kinbank sample

At time of publication, Kinbank holds 210,903 different data points across 1,229 languages (Fig 1). The collaborative approach to this project has built a database with a broad global sample, coupled with focused sampling from specific language families. Each language is linked to a stable and unique linguistic identifier "Glottocode" that is commonly used to link language and cultural data to other databases (48). By indexing on source, we can also separate terminologies within a language across sources, which allows analyses to measure concordance across sources, track change over time, or identify within-language variation.

Parabank collected kinship terminologies opportunistically, including the digitisation of the terminologies from Morgan's 1871 landmark global kinship survey, *Systems of Consanguinity and Affinity* (13). The global nature of the Parabank sample means the

database holds terminologies from societies with a range of ecological pressures and kinship structures, allowing us to explore convergent patterns of terminology (e.g. (29) across the global sample.

Varikin focused on sampling languages that are paired both with a dated language phylogeny (e.g. (49) and with existing anthropological databases (e.g. D-PLACE (50)). The most sampled language families in Kinbank are Austronesian (n = 375), Atlantic-Congo (117), Indo-European (105), and Pama-Nyungan (104). This sampling strategy allows for a range of phylogenetic analyses that control for patterns of autocorrelation that might occur through descent (51,52). The focus on societies within the Ethnographic Atlas additionally allows kinship terminology data to be connected to Demographic and Health surveys (53).

MPEGKin compiled data on two large language families of South America, Tupian and Cariban, sampling as densely as possible across the major branches of the families given the available ethnographic and linguistic documentation. Beyond making new data available for these lesser-known languages (39), MPEGKin was designed to support quantitative comparative studies and traditional linguistic reconstruction with these language families (e.g., (54,55)).

Kinura collected kinship terminologies of all the main groups of the Uralic family. The sampling focused on matching the kin term data with the speaker groups included in the genetic study by Tambets et al. in 2018 (56).

Focused sampling for languages attached to existing cultural and linguistic databases (e.g. D-PLACE (50)) and to computational language phylogenies (e.g. (57)), allows evolutionary analytic approaches to this important cultural domain. The combination of focused regional samples and broad global coverage will allow future research to compare the patterns and correlates of diversity across different scopes. Kinbank is citable as a whole and as separate regional projects (see S1 Table).

Fig 1: Locations of languages in Kinbank. Each point indicates a unique language variety and is centered on the geographical center-point of the area where the speakers live when the data were recorded, but may also indicate a historical location, the demographic center-point or some other representative point. Colored points indicate languages from the 7 language families mentioned in text: Austronesian (light blue; n = 377); Pama-Nyungan (red; n = 105); Indo-European (yellow; n = 106), Bantu (orange; n = 113), Uralic (purple; n = 25), Tupian (maroon; n = 29); Cariban (pale blue; n = 7) and other languages (grey).

Concepts in Kinbank

The primary search criteria for Kinbank were a core set of 155 kin types (88 genealogical kin and 27 kin by marriage (affine); available in S2 Table). The core set of kin types encompass parents, siblings, and children; up to grandparents and down to grandchildren, and then from parents to their siblings and parents' siblings' children. We limited the set to two generations above and below the ego (grandparents to grandchildren). This is a commonly discussed range in the kinship literature, but it also

represents the most common set of generations that exist during ego's lifetime, which is four generations with a generation of between 20-30 years. Not all generations necessarily exist at the same time, e.g., ego might have grandparents when they are young, and grandchildren when they are old. It is uncommon for five generations to co-exist. Many languages have separate terms depending on the gender of the speaker, and we collected all such terms where available. We aimed to collect affinal terms for spouses, spouses of siblings, and the spouses' nuclear family. Within ego's generation (ego being the central figure from which relationships stem) and ego's parents' generation, kin types are also distinguished by relative age, and age of linking relative, where appropriate. This set is derived from a genealogical grid of relatives and aims to capture a globally recurrent set of cross-culturally valid kin-members.

The dataset contains terms of reference (the answer to: "who is this person to you"), as opposed to terms of address (the answer to: "what do you call this person") (48). Reference terms are more commonly the focus of anthropological data collection, since these designate formal categories of relationship (e.g. "father" vs "dad"), and are less prone to politeness effects which often extend the range of kin terms (e.g. by employing "uncle" / "aunt" to address non-related elder men and women as a term of respect). Terms of address were recorded when they were easily available but were not the focus of this collection.

The set of individuals designated by kinterms is culturally variable (58). Some communities have a restricted set of kinterms, and others an expanded set extending beyond the established genealogical grid (59); however, since the unit of the database is the kinterm, the database is flexible to account for different emic characterizations of kin relations. In the case of a restricted set, cells that are not required are left empty. This presents the difficult challenge of knowing whether a term is absent because this relative is not considered a kin member or because it was not recorded in the source. The difference between a language lacking a term for a particular category (true absence), and available records of a language lacking particular terms (data uncertainty), is important. Comparisons with closely related languages, the use of multiple sources, and iterative, community-elicited data collection can help to narrow this uncertainty in future.

Collection procedure

Data is entered into Kinbank as either primary or secondary data, and is affiliated with a source reference detailing where the data came from. The data is most commonly secondary, and transcribed by coders into the Kinbank collection template. Secondary sources ranged from ethnographies and grammars, to simpler descriptions like wordlists. In the majority of cases, secondary sources describe kin terms etically following the tradition laid out by Morgan (13). The etic descriptions usually align with the list of 155 core kin types. In these instances, kin term collection is a matter of transcription. It is common for source categories to be less refined than Kinbank categories. In this case, coders would apply the term to all relevant categories. For example: the category father's brother is commonly used in secondary sources, but this aggregates four Kinbank categories (father's elder brother, father's younger brother, and whether the speaker is male or female). A coder would apply the term to all four of these categories in the absence of further information. In the rare

instance that kin terms were not etically described (often using English nomenclature), coders would search the text for context and confirmation that these categories align and record any inferences in the notes column for each kin term. Coders were also encouraged to flag where original sources may have mistakes in them.

Kin terms are transcribed as closely as possible to their source form. Since terminologies are predominantly collected by anthropologists and not linguists, sources are primarily in Roman script (as opposed to other standardized orthographies) and can contain transcription inconsistencies across languages. Kinbank makes no judgments on which transcriptions are correct or incorrect. The database accommodates phonetic transcriptions in the International Phonetic Alphabet if available. Some sources recorded terminology using Cyrillic; in this case kin terms are transliterated into Roman characters, with the original form in an adjacent column. This approach means that transcription is consistent within source but may not be consistent across sources or languages.

Some data was elicited from native speakers or academics actively working with a particular language. In this case the person who offered the terms is listed as the source. When data was collected from native speakers, a Kinbank collector would describe the goal and walk through each of the 155 core kin types with the native speaker, entering the relevant kinterm for the genealogical category of interest. Active researchers were provided with the template sheet and a description of the task and asked to fill in the categories with any necessary comments or adjustments. When there was specific context, meaning, or otherwise interesting information for a particular kinterm it was recorded in the notes. Some sources (speakers or secondary) may offer many more kinterms that apply beyond the 155 core types. In this instance, we expand the parameter list to accommodate the additional kin terms (see S2 Table). In total there are 940 kin types within Kinbank, but many of these are sparsely populated. Some languages also contain multiple terms for a particular category. In this instance all kin terms are entered separately.

Kinbank database format

Data is stored and distributed in the Cross-Linguistic Data Format (CLDF; (60)), which is a flat structured database stored in comma-separated value (CSV) files (Table 1), and built with the affiliated tools CLDFbench (61). This format is easily importable into common analysis tools like Microsoft Excel, and data analysis languages like R or Python.

It is common for languages to have multiple terms recorded for some kin types. Due to the combination of form and kin type, the database can accommodate multiple terms per kin type. Multiple terms for a single kin type may occur for a number of reasons: borrowed and native terms coexisting, dialectal differences, different registers, or simply language flexibility (62). The database offers no judgement as to the specific cause. Again, users can refer to the original source(s) for further information.

Kinbank is built for expansion and correction, and is a key feature of the construction pipeline. Living datasets are stored at github.com/kinbank. By using Github to store the living datasets, users can raise issues with existing data, propose corrections, and upload new

languages. All changes are reviewed by Kinbank authors. New languages can be added by filling in the template document of core kin types found on the homepage of each repository. Since a living database will not be static, the datasets are periodically versioned on Zenodo, ensuring that users of the database can analyze the database at stable points.

Table 1: Descriptions of the tables used in kinbank and the data they contain. An extended description with additional structuring files is given in S3 Table.

File	Description
forms.csv	This contains the forms, or kinterms, for each language. This file links to the languages.csv file by the Languages_ID column, parameters.csv file by the Parameters_ID column, and the sources.bib file by the sources_bibtex column.
languages.csv	This file contains metadata on each language (Name, Glottocode, ISO639P3 code, Macroarea, Latitude, Longitude, and Language Family)
parameters.csv	This file contains descriptions of each kin type used in the dataset.
sources.bib	A bibtex file containing information on the sources used in Kinbank.

Inter-rater reliability

Forty-four languages were collected independently by both ANU and Bristol teams and used to determine the level of intercoder reliability. A major avenue for error is when kinship terms are collected for one language but two different sources disagree on kinterms. We focus on ensuring the structural paradigm of a particular kinship terminology is consistent (i.e. that all parent's female siblings are syncretized to "aunt" in English, not whether one source specifies "aunt" or "aunty"); inter-rater reliability is based on this benchmark. To compare collections, we used a structural similarity kappa value, which compares the pattern of syncretism within each collection, accounting for similarity by chance. Across all compared relationships, we obtained a structural similarity kappa value of 0.80. Calculation and discussion of interrater reliability is available in supplementary material and S4 Table.

Analysis of Kinbank data

The Kinbank collection of 1,229 languages offers a large digitised sample of kinship terminology in an accessible format. The collection has 887 languages that contain all grandparent terms, 864 languages with all sibling terms, 728 languages with all terms for parents and parent's siblings, 604 languages with all sibling terms, and 506 languages with all sibling and cousin terms. The most frequently collected kin terms are all nuclear family terms: Father (n = 1,197), Mother (1,192), younger brother (1,161),

elder brother (1,149), elder sister (1,138), and younger sister (1,134). Closely followed by father's father (1,110) and mother's father (1,119). The least common terms, from within the core set, are a combination of affinal and cousin terms: wife's sister (545), father's brother's daughter (552), husband's sister (552), husband's brother (554), and mother's sister's daughter (556). On average, languages have 23 terms: 2 grandparent terms, 3 sibling terms, 5 terms for parents and parent's siblings, 4 sibling terms, and 6 sibling and cousin terms. Averages only tell us part of the picture however, as languages can have as low as one kin term for each set (with the exception of parents and parent's siblings which has a minimum of two) and as high as a single term for all relatives, commonly known as a descriptive system. Although languages range significantly in the number of kin terms they use, some syncretisms are "socially unthinkable" (Héritier, 1981, Godelier 2012). The law of collaterality states that we cannot observe syncretisms between kin that are linked by a member of the opposite sex, without other parallel syncretisms occurring, a rule which is abided by in all Kinbank languages. For example: Kinbank contains no languages which have a term for father and mother's brother, but not father's brother. Similarly, Kinbank contains no languages that have distinct terms for father's mother, or mother's father, with a single term for all other grandparents.

Below we provide two examples of how researchers can use Kinbank's features and data to provide powerful insights to cultural change. First, we investigate the phonological structure of global terms for "mother" and "father" to show strong gender bias with regard to phonological form. Second, we integrate Kinbank with anthropological data and find no evidence to support a coevolutionary relationship between kin terms and marriage systems in Bantu languages.

Example 1: Are you my mama? Gender bias in the phonology of parental terms.

The global recurrence of certain sounds in parental terms (e.g. [ma], [pa]) among geographically distant and historically unrelated languages is hypothesised to stem from the constraints on early baby babbling (63). A babbling theory for parental term similarity was first identified by Murdock (64) as a statistical regularity, and then subsequently theorized by (63) to recur because of the maximal phonetic contrasts the sounds make. The combination of a stop or nasal, followed by a low vowel, creates the largest contrast amongst the sound's babies are capable of making and are recognizable, distinguishable, and identifiable noises. The bias towards these perceivable signals of early communication then further evolves to become codified as words for parents in spoken language, as opposed to its babbled form.

An extension of this theory suggests that mother terms are more likely to start with [ma] because of the sound relationship to breastfeeding (63,64). The bilabial nasal sound [m] is putatively an anticipatory murmur; the [a] sound is created by the baby's mouth opening preparing to breastfeed. The onomatopoeic link to breast-feeding is hypothesized to produce a phonological gender bias. Although early descriptive statistics support this conclusion (64), recent research on Australian languages has shown that parental terms with initial [ma] sounds frequently refer to father, not mother; furthermore *mama* is a reconstructable proto-form for "father" in Pama-Nyungan

languages (65,66). (65) also show that other nasal consonants such as [ŋ] (velar nasal) are commonly used for mother, which are also compatible with the breastfeeding hypothesis. The regional and historical relationship between [ma] and "father" in Australia highlights the importance of accounting for the interdependence between languages when looking at cross-linguistic patterns. Had the sample been biased towards Australian languages without a historical control, we would conclude that most languages use [ma] for father, but a phylogenetic control accounts for this bias through shared evolutionary history (52). Here, we ask whether there is a gender bias in the phonology of parental terms, using phylogenetic approaches to explicitly model and control for the historical relationships between languages. We code the first syllable of 3,068 parental kinterms from 1,022 languages for a consonant and vowel, following categories established in (67), and model their relationship to the referent using a phylogenetically-controlled repeated measures multilevel Bayesian logistic regression (See methods for more detail).

Example 1: Results

The model shows that consonant use is a strong predictor of parental referent. Vowels offer little predictive value. Since consonant-vowel combinations are the real-world depiction of these sounds, we calculate the probability of theoretically implicated, and contrastive, consonant-vowel pairs from the model posterior: [ma], [na], and [ŋa] (common mother sounds), and [pa], and [ta] (common father sounds; since vowels have little influence we only use [C-a] in the examples; Fig 2). The model only analyses "mother" or "father" words: mother words are coded as 1. Since the only other possibility is for a word to refer to father, probabilities close to zero indicate the likelihood a sound refers to father. There is positive evidence for [ŋa] and [na] more commonly referring to mother, and that [pa] and [ta] more commonly refer to father. Notably, there is no evidence that [ma] sounds refer preferentially to mother or to father. A summary of the model and all effects are available in S5 Table.

Fig 2: Probability of consonant and vowel combinations used in words referring to mother. Each row shows the probability density estimate from the model posterior. The colored sections and annotated numbers show 89% highest probability density intervals. Intervals that contain 0.5 have no statistical effect (since there is a 50:50 chance the sound refers to mother or father). Results show words with a root syllable of [ŋa] and [na] are more likely to refer to mother and [pa] and [ta] to refer to father. [ma] words are predicted to be used equally between mother and father words.

Example 1: Discussion

Our analysis has shown that the bilabial-low vowel combination [ma], which is often thought to be linked to mothers, is actually equally linked to fathers. On the other hand, we present strong evidence in a global sample for [ŋa] and [na] sounds aligning exclusively with mother terms, as reported by (65,66) for Australian languages. Closer inspection of Murdock's tabulations shows very few languages using velar nasal sounds in parental terms, in stark contrast to their relatively high frequency in our broader sample. This difference highlights the importance of wide language documentation, and an understanding of the full range of linguistic diversity before drawing premature "universal" conclusions (68). Kinbank provides us with a broader sample of diversity,

which reveals a new understanding of gender bias in parental term phonology. With a broader sample comes the possibility that languages are similar because of historical relatedness, rather than an external factor. Here, we have used global linguistic relationships to control for the possibility of auto-correlation within our sample. A puzzle yet to be solved is whether there is a causal reason for the relationship between [ta] and [pa] and father terms, or whether the bias is a result of phonological dispersion i.e., maximal differentiation as a phonological driver for parental terms.

Example 2: Does crossness indicate marriage preferences?

Within anthropology, the structure of kinship terminology is thought to reflect patterns of social structure (9). A common example is the linguistic denotation of marriage taboos, indicating which kin are marriageable and which are not (69). The presence of cross-cousin marriage is most often linked to the linguistic structure called *crossness*. Crossness is the presence of a linguistic distinction between opposite-gender siblings; for example: the child of a father's sister is a cross-cousin because a father's sister is a father's opposite-gendered sibling (or crossed-gendered sibling). When crossness occurs within the parental generation it is commonly called a bifurcate-merging terminology, which merges same-sex siblings, and splits (or bifurcates) the opposite sex siblings (59). This often leads to two classes of relatives per gender: for male referents, father and father's brother are grouped under the same term, and mother's brother is a separate term (as mother's opposite-sex sibling), and mother and mother's sister are grouped, and father's sister is a separate term (70). Historical interest in bifurcate-merging terminology surrounds two conditions. In the archetypal society with mandatory first-cousin cross-cousin marriage, a parent's opposite-sex sibling is also a spouse's parent (or a parent-in-law), making a kinterm applied to this relative polysemous, between a consanguineal and affinal relative (call this type A). Under these conditions, linguistic crossness could be seen to represent the limits of a culturally imposed incest taboo. However, bifurcate-merging terminology is often observed in societies lacking mandatory cross-cousin marriage, and thus a parent's opposite-sex sibling and parent-in-law are different people (type B). Determining why type B exists – i.e. why there should be crossness in the absence of mandatory cross-cousin marriage – is an ongoing topic of discussion amongst kinship terminology scholars (69,71).

One suggestion is that all bifurcate-merging terminology occurs with some form of cross-cousin marriage, but does not require it to be mandatory for crossness to arise (69). Here, we test the hypothesis that bifurcate-merging terminology co-evolves with allowable cross-cousin marriage in 56 Bantu societies (allowable being a weaker social norm than mandatory cross-cousin marriage). We define a language as bifurcate-merging if it distinguishes between a parent's same and opposite gender siblings. Specifically, we categorize a terminology as bifurcate-merging under three scenarios: if the term for father is merged with father's brother, and not mother's brother, if the term for mother is merged with mother's sister, but not father's sister, and if there is complete merging in both parent's male and female siblings (see supplementary methods for details). Cross-cousin marriage occurs in 38% of Bantu societies which have been studied, but the archetypal type A terminology is rare in contemporary languages, meaning a strict language-behavioral relationship is unlikely to be the main driver of crossness (50,71). Historical linguists have previously shown that the likely Proto-East-

Bantu parental kinship terminology contained a bifurcate-merging pattern, from which a historical presence of preferential cross-cousin marriage is inferred (26), but it remains to be seen why bifurcate-merging terminology persists throughout the language family.

Cross-cousin marriage data is sourced from the Ethnographic Atlas via D-PLACE, and binary coded for the presence or absence of any cross-cousin marriage, and matched to Kinbank using Glottocodes (50,72). Language and marriage data are then paired to a sample of 100 Bantu phylogenies (73). Using phylogenetic models of coevolution, implemented in BayesTraits v3.0.1, we fitted two different models, one where changes between the linguistic and behavioral traits depend on each other (co-evolve), and one where they are independent. We used Bayes factors to compare the likelihood of each model. A Bayes factor >2 indicates positive support for the dependent model and is evidence for a coevolutionary relationship. All models were run twice to test MCMC convergence (see details in the supplementary materials).

Example 2: Results

We find that there is no positive evidence for the coevolution of cross-cousin marriage and bifurcate merging terminology in any of the three terminology variables. In tests of "complete bifurcate-merging", and "bifurcate-merging in women", there is no evidence either for or against the co-evolution of crosscousin marriage and bifurcate-merging organizations (Complete log Bayes Factor = -0.83; Women: -0.18). Tests of "bifurcate-merging in men" show positive evidence for independent evolution (-2.28), suggesting in this set of data, bifurcate-merging terminologies have no general relationship to cross-cousin marriage. Fig 3 displays the data for the "complete bifurcate-merging" variable on the Bantu phylogeny. The internal nodes, shown as probabilistic pie-charts for the four possible states, show confidence in relatively recent changes, but contain high levels of uncertainty in deeper nodes.

Fig 3: Maximum clade credibility tree of Bantu languages displaying four possible states: no cross-cousin marriage (CM) and no bifurcate-merging (BM) of kin terms (black), no cross-cousin marriage with bifurcate-merging (yellow), cross-cousin marriage and no bifurcate merging (red), and both cross-cousin marriage and bifurcate merging are present (light blue), with their likely probabilities from a dependent model of evolution. Languages marked with grey show where kinterms data were present, but social data were imputed by BayesTraits. Pie-graphs indicate the probability of states at each node, which were calculated using the BayesTrait command "RecNode". Deeper into the phylogeny there is much uncertainty as to the relationship between these two traits, indicated by the almost equal probability of all four states.

Example 2: Discussion

The relationship between bifurcate-merging terminology and cross-cousin marriage is statistically inconclusive from this analysis, adding to an increasing body of evidence disputing purported relationships between kinship terminology and social behaviors (30,38). In particular, it may be the case that societies can arrive at the same structural organization of kinship terminology but with categories conveying different cultural meanings. For example: Bena (Tanzania) speakers use linguistic crossness, allow but do not prescribe cross-cousin marriage, as well as practicing polygyny (74), while

Lumasaaba (Kenya) speakers use crossness but no cross-cousin marriage (75). Amongst Bena, the distinction of cross-cousins reflects the predicted relationships between marriageable and non-marriageable cousins. Cross-cousin marriage is considered a high-status marriage, although other marriages frequently occur. The emic category of cross-cousin in the Bena is broad and includes relatives stemming from a cross-cousin relationship many generations before (e.g. great great fathers' sister's offspring). The closer the genealogical relationship between the marrying couple, or if the relationship is traced through a person of high status, the more status is granted to the children of that marriage. In contrast, Lumasaaba parents' opposite-sex siblings play important roles in a child's life, by holding important ceremonial roles in a child's rite of passage and traditionally being the relative through whom the child will inherit wealth (although this tradition has since changed (75)). These important cultural roles highlight a special relationship between children and their mother's brother or father's sister, which could be the cause of the linguistic distinction. These are examples taken from our sample, but the relationship of kinship terminology structure to various semantic meanings is an avenue for future research.

Our coevolutionary results challenge a long-held belief within kinship terminology research. While some researchers have suggested that the relationship between kinship terminology and behavior is more complex than often assumed (76), the longstanding consensus of the field is that there is coupling between language and norms (1,9,32). However, we test only one example of a behavioural-linguistic link from a sea of proposed hypotheses (e.g. (77,78)), and we hope this database encourages future research on the relationship between kinship norms and language. Using Kinbank will aid in transparent and iterated hypothesis testing: as methods such as phylogenetic analysis become mainstream, and sophisticated statistical methods that incorporate not just shared cultural ancestry but spatial autocorrelation are developed, the granularity of kin term data in Kinbank allows the analyst to move flexibly between typological levels depending on the research question at hand.

Conclusion

Kinbank offers an open and transparent database of kinship terminologies from a global sample of 1,229 languages. The examples presented in this paper illustrate the benefit of systematic and flexible storage of large cultural datasets, as they combine the rich ethnographic record with decades of theoretical debate in linguistics and anthropology. Moreover, the focus on language family sampling allows for the application of computational phylogenetic approaches from evolutionary biology. While the cultural basis of kinship has long been an anthropological focus, interest from such fields as economics (7), psychology (28), and linguistics (32) means that now, more than ever, there is a need for large, diverse cross-cultural datasets that allow us to test old and new hypotheses in one of the founding anthropological domains.

Materials and Methods

Example 1: Are you my mama?

Kinbank contains 3,068 words for mother or father, from 1,022 languages. Some languages have multiple words for mother and father terms. All words are reduced to the first syllable of their root and coded for a consonant and vowel ignoring sound order, following (64). The root syllable is typically the first syllable of the word, with exceptions being the presence of a prefix, or some other feature of the language which might indicate a different syllable. Prefix decisions are at the coders discretion. Syllables are coded as one of 34 consonant types and one of seven vowel types (following Blasi et al. 2016; S6 Table). For example: the Serbian word for mother is *majka*: here the first and root syllable is [ma]. The first syllable [ma] has a consonant coded as [m] and vowel coded as [a]. Coding was performed manually and inter-rater reliability statistics are available in S7 Table.

To control for the relatedness between languages, we use a global phylogenetic tree (79). Though not a perfect measure of relatedness between languages, it is a vast improvement on correlational studies historically employed in kinship terminology studies (e.g. (9)). Pruning and cleaning of phylogenies was performed in *R* using the packages *ape* v5.3 (80), *phangorn* v2.5.5 (81) and *phytools* v0.7 (82).

Model

We built a phylogenetically-controlled repeated measures multilevel Bayesian logistic regression using *brms* v2.16.3 (83). By using a repeated measures approach, multiple terms can be modelled per language (e.g. mother and father), while controlling for the phylogenetic relationships between languages. The response is a binary variable which indicates whether a term is mother (1) or father (0) and is independently predicted by the consonant and vowel sound codes. To control for language relatedness we included an inverse variance-covariance matrix built from the phylogenetic tree, as well as a random effect for language. This controls for both historical relatedness, and other factors that may be explained at the language level. Models were run for four chains, with 5,000 iterations, and 2,000 burn-in iterations. All fixed effects had normal priors with a mean of zero and standard deviation of 10. Probability estimates of consonant-vowel combinations of interest were predicted from the model posterior.

Example 2: Does crossness indicate marriage preferences?

We used terminological data from Kinbank to determine the presence of crossness across the whole parental generation (complete bifurcate-merging; $F = FB \neq MB$ & $M = MZ \neq FZ$), or only within men ($F = FB \neq MB$), or only within women ($M = MZ \neq FZ$). Detailed data description is in S8 Table. Cousin-marriage preference data is taken from the Ethnographic Atlas question EA023 from D-PLACE and coded to indicate the presence or absence of cross-cousin marriage (S9 Table; (50,72)), and linked to languages on the Bantu phylogeny (64). The intersection of Kinbank and D-PLACE results in a set of 56 societies for which 1) Kinbank contains kin terms for all parents and parents' siblings, 2) D-PLACE contains information on the presence of cross-cousin marriage and 3) the language is represented on a dated Bantu phylogeny. More information on data coding decisions is given in the supplementary material.

We implement a Bayesian correlated evolution phylogenetic approach, using *BayesTraits* v3.0.1 (84). By using a Bayesian approach with a sample of phylogenetic

trees, the model does not only control for shared ancestry between societies but also for uncertainty in the phylogenetic relationships of languages. For each statistical test there are two models: one model where correlated evolution is assumed, and one where traits evolve independently. These models are compared to calculate a log Bayes Factor (BF) to determine which model best fits the data. $BF < 2$ indicates weak evidence, > 2 positive evidence, 5-10 strong evidence, and > 10 very strong evidence (85). All models are run for 11,000,000 iterations, sampling every 1,000 iterations, with a burn-in of 1,000,000 iterations, on a posterior sample of 100 trees (approximately 200 samples per tree, but not enforced by using EqualTrees). All parameters have an exponential prior with a mean of 10. Each model was run twice to assess convergence; all MCMC runs show MCMC Gelman and Rubin statistics are < 1.1 (S10 Table). MCMC trace plots are available in S1 Fig.

Author Contributions: NE & FMJ conceptualized the database and obtained funding. SP, CS, WB, KQ, SJG, JB, and TH designed and constructed the database. SP performed the analyses. SP, SJG, CS, JB, CB, TH, RR, NE, and FMJ wrote the paper. All authors contributed data to the database and commented on the final version of the paper.

Acknowledgments: We would like to thank Isobel Clifford, Lieke Hoenselaar, and Maarten van den Heuvel for their assistance with data collection. We would also like to thank two anonymous reviewers for their comments on the manuscript.

References

1. Jones D. Human kinship, from conceptual structure to grammar. *Behavioral and Brain Sciences*. 2010 Oct;33(5):367–81.
2. Chapais B. Primeval kinship: how pair-bonding gave birth to human society. Harvard University Press; 2009. 369 p.
3. Lansing JS, Abundo C, Jacobs GS, Guillot EG, Thurner S, Downey SS, et al. Kinship structures create persistent channels for language transmission. *PNAS*. 2017 Nov 20;114(49):1291–12915.
4. de la Croix D, Doepke M, Mokyr J. Clans, Guilds, and Markets: Apprenticeship Institutions and Growth in the Preindustrial Economy. *Q J Econ*. 2018 Feb 1;133(1):1–70.
5. Divale WT. Migration, External Warfare, and Matrilocal Residence. *Behavior Science Research*. 1974 May 1;9(2):75–133.
6. Macfarlan SJ, Erickson PI, Yost J, Regalado J, Jaramillo L, Beckerman S. Bands of brothers and in-laws: Waorani warfare, marriage and alliance formation. *Proceedings of the Royal Society B: Biological Sciences*. 2018 Nov 7;285(1890):20181859.
7. Schulz JF, Bahrami-Rad D, Beauchamp JP, Henrich J. The Church, intensive kinship, and global psychological variation. *Science*. 2019 Nov 8;366(6466):eaau5141.
8. Shenk MK, Mattison SM. The Rebirth of Kinship. *Hum Nat*. 2011 Jul 1;22(1–2):1–15.

- 683 9. Murdock GP. Social structure. Vol. xvii. Oxford, England: Macmillan; 1949. 387 p.
- 684 10. Stone L. Kinship and Gender. 5th ed. Washington State University; 2014. 674 p.
- 685 11. Carsten J. After kinship. Vol. 2. Cambridge University Press; 2004.
- 686 12. Ethnologue, Languages of the World [Internet]. Library of Congress, Washington, D.C.
687 20540 USA. [cited 2022 Oct 18]. Available from:
688 <https://www.loc.gov/item/lcwaN0021868/>
- 689 13. Morgan LH. Systems of Consanguinity and Affinity of the Human Family. Smithsonian
690 Institution; 1871. 656 p.
- 691 14. Malinowski B. Argonauts of the Western Pacific: An Account of Native Enterprise and
692 Adventure in the Archipelagoes of Melanesian New Guinea. Routledge; 1922. 502 p.
- 693 15. Ekstrom J. Word list for Uto-Aztecan languages. Parker, Arizona; 1959.
- 694 16. Lounsbury FG. A formal account of the Crow-and Omaha-type kinship terminologies.
695 Bobbs-Merrill; 1964.
- 696 17. Irons W. The Yomut Turkmen [Internet]. University of Michigan Press; 1975 [cited 2021
697 May 15]. Available from: <http://www.jstor.org/stable/10.3998/mpub.11394884>
- 698 18. Chagnon NA. Terminological kinship, genealogical relatedness and village fissioning
699 among the Yanomamo Indians. Natural selection and social behavior. 1981;490–508.
- 700 19. Chagnon NA. Manipulating kinship rules: A form of male Yanomamö reproductive
701 competition. In: Adaptation and Human Behavior. Routledge; 2017. p. 115–32.
- 702 20. Danziger E. What Might Mother Mean? The Acquisition of Kinship Vocabulary in Mopan
703 Maya. In: The Proceedings of the Twenty-fifth Annual Child Language Research Forum.
704 Center for the Study of Language and Information Stanford; 1993. p. 227–34.
- 705 21. Read DW, Atkins J, Buchler IR, Fischer M, De Meur G, Lally E, et al. An Algebraic
706 Account of the American Kinship Terminology [and Comments and Reply]. Current
707 Anthropology. 1984;25(4):417–49.
- 708 22. Leaf M. The Recognition of Kinship Terminologies As Formal Systems. Structure and
709 Dynamics [Internet]. 2013 [cited 2021 May 19];6(1). Available from:
710 <https://escholarship.org/uc/item/2b90r031>
- 711 23. Allen NJ, Callan H, Dunbar R, James W. Early Human Kinship: From Sex to Social
712 Reproduction. John Wiley & Sons; 2008. 337 p.
- 713 24. Blust R, Aberle DF, Allen NJ, Barnes RH, Chowning A, Dahl OChr, et al. Early
714 Austronesian Social Organization: The Evidence of Language [and Comments and
715 Reply]. Current Anthropology. 1980;21(2):205–47.
- 716 25. Ehret C. Reconstructing ancient kinship. Practice and theory in an African case study.
717 Kinship, Language, and Prehistory: Per Hage and the Renaissance in Kinship Studies,
718 University of Utah Press, Salt Lake City. 2011;46–74.
- 719 26. Marck J, Hage P, Bostoen K, Muzenga JGK. Kin terms in the East Bantu proto
720 languages: Initial findings. Kinship, language, and prehistory: Per Hage and the
721 renaissance in kinship studies. 2010;79–82.

27. Read DW. A New Approach to Forming a Typology of Kinship Terminology Systems: From Morgan and Murdock to the Present. 2013 [cited 2017 Mar 18]; Available from: <https://papers.ssrn.com/abstract=2624169>
28. Cronk L, Steklis D, Steklis N, van den Akker O, Aktipis A. Kin terms and fitness interdependence. *Evolution and Human Behavior*. 2018 Dec 22;40(3):281–91.
29. Passmore S, Barth W, Quinn K, Greenhill SJ, Evans N, Jordan FM. Kin Against Kin: Internal Co-selection and the Coherence of Kinship Typologies. *Biological Theory* [Internet]. 2021 Jun 7; Available from: <https://doi.org/10.1007/s13752-021-00379-6>
30. Passmore S, Jordan FM. No universals in the cultural evolution of kinship terminology. *Evolutionary Human Sciences* [Internet]. 2020 ed [cited 2020 Sep 9];2. Available from: <https://www.cambridge.org/core/journals/evolutionary-human-sciences/article/no-universals-in-the-cultural-evolution-of-kinship-terminology/0BF406C9CFC182F9142749FDD0442471>
31. Dousset L, Hendery R, Bowern C, Koch H, McConvell P. Developing a Database for Australian Indigenous Kinship Terminology: The Austkin Project. *Australian Aboriginal Studies*. 2010;(1):42.
32. Kemp C, Regier T. Kinship Categories Across Languages Reflect General Communicative Principles. *Science*. 2012 May 25;336(6084):1049–54.
33. Mollica F, Piantadosi ST. Logical word learning: The case of kinship [Internet]. *PsyArXiv*; 2019 [cited 2021 May 19]. Available from: <https://psyarxiv.com/a7tnb/>
34. McNamara RA, Henrich J. Kin and kinship psychology both influence cooperative coordination in Yasawa, Fiji. *Evolution and Human Behavior*. 2017 Mar;38(2):197–207.
35. Chen (陈新仁) X, Ren (任娟娟) J. A memetic cultural practice: The use of generalized kinship terms in a research seminar attended by Chinese graduate students. *Lingua*. 2020 Oct 1;245:102942.
36. Henrich J. The Weirdest People in the World [Internet]. Farrar, Straus and Giroux; 2020 [cited 2021 Jan 29]. Available from: <https://www.penguin.co.nz/books/the-weirdest-people-in-the-world-9781846147968>
37. Racz P, Passmore S, Jordan F. Social Practice and Shared History, Not Social Scale, Structure Cross-Cultural Complexity in Kinship Systems. *TOPICS IN COGNITIVE SCIENCE*. 2020 Apr;12(2):744–65.
38. Sheard C, Bowern C, Dockum R, Jordan FM. Pama–Nyungan grandparent systems change with grandchildren, but not cross-cousin terms or social norms. *Evolutionary Human Sciences* [Internet]. 2020 ed [cited 2021 May 13];2. Available from: <https://www.cambridge.org/core/journals/evolutionary-human-sciences/article/pamanyungan-grandparent-systems-change-with-grandchildren-but-not-crosscousin-terms-or-social-norms/FD3DD720CEA3415016927CE1BBF3A4C2>
39. Birchall J, Jordan FM. New Perspectives on kinship terminology in Tupian and Cariban Languages. *Boletim do Museu Paraense Emílio Goeldi Ciências Humanas*. 2019 Apr;14(1):11–4.

- 762 40. Rácz P, Passmore S, Sheard C, Jordan FM. Usage frequency and lexical class
763 determine the evolution of kinship terms in Indo-European. *Royal Society Open Science*.
764 2019 Oct 30;6(10):191385.
- 765 41. Mitchell A, Jordan FM. Kinship, seniority, and rights to know in Datooga children's
766 everyday interaction. *Journal of Pragmatics*. 2021 Aug 1;181:49–61.
- 767 42. Mitchell A, Jordan F. The Ontogeny of Kinship Categorization. *JOURNAL OF*
768 *COGNITION AND CULTURE*. 2021 Jun;21(1–2):152–77.
- 769 43. Evans N. Context, Culture, and Structuration in the Languages of Australia. *Annual*
770 *Review of Anthropology*. 2003;32(1):13–40.
- 771 44. Blythe J, Tunmuck J, Mitchell A, Rácz P. Acquiring the lexicon and grammar of universal
772 kinship. *Language [Internet]*. 2020 Mar 31 [cited 2020 Aug 15]; Available from:
773 [https://researchers.mq.edu.au/en/publications/acquiring-the-lexicon-and-grammar-of-](https://researchers.mq.edu.au/en/publications/acquiring-the-lexicon-and-grammar-of-universal-kinship)
774 [universal-kinship](https://researchers.mq.edu.au/en/publications/acquiring-the-lexicon-and-grammar-of-universal-kinship)
- 775 45. Evans N. *Semantic typology*. Oxford University Press; 2010.
- 776 46. Majid A, Jordan F, Dunn M. Semantic systems in closely related languages. *Language*
777 *Sciences*. 2015 May 1;49:1–18.
- 778 47. Nerlove S, Romney AK. Sibling terminology and cross-sex behavior. *American*
779 *Anthropologist*. 1967;69(2):176–87.
- 780 48. Hammarström H, Forkel R, Haspelmath M, Bank S. Glottolog database 4.4 [Internet].
781 Zenodo; 2021 [cited 2021 Dec 2]. Available from: <https://zenodo.org/record/4761960>
- 782 49. Gray RD, Drummond AJ, Greenhill SJ. Language Phylogenies Reveal Expansion Pulses
783 and Pauses in Pacific Settlement. *Science*. 2009 Jan 23;323(5913):479–83.
- 784 50. Kirby KR, Gray RD, Greenhill SJ, Jordan FM, Gomes-Ng S, Bibiko HJ, et al. D-PLACE:
785 A Global Database of Cultural, Linguistic and Environmental Diversity. *PLOS ONE*. 2016
786 Jul 8;11(7):e0158391.
- 787 51. Jordan FM, Dunn M. Kin term diversity is the result of multilevel, historical processes.
788 *Behavioral and Brain Sciences*. 2010 Oct;33(5):388–388.
- 789 52. Roberts S, Winters J. Linguistic Diversity and Traffic Accidents: Lessons from Statistical
790 Studies of Cultural Traits. *PLOS ONE*. 2013 Aug 14;8(8):e70902.
- 791 53. Bahrami-Rad D, Becker A, Henrich J. Tabulated nonsense? Testing the validity of the
792 *Ethnographic Atlas*. *Economics Letters*. 2021 Jul 1;204:109880.
- 793 54. Carvalho FO de, Birchall J. A comparative reconstruction of Proto-Tupi-Guarani kinship
794 terminology. *LIAMES: Línguas Indígenas Americanas*. 2022 Mar 9;22:e022001–
795 e022001.
- 796 55. Birchall J, Oliveira LH, Jordan FM. A note on the kinship system in Proto-Tupí-Guaraní.
797 *Boletim do Museu Paraense Emílio Goeldi Ciências Humanas*. 2019;14:79–99.
- 798 56. Tambets K, Yunusbayev B, Hudjashov G, Ilumäe AM, Rootsi S, Honkola T, et al. Genes
799 reveal traces of common recent demographic history for most of the Uralic-speaking
800 populations. *Genome Biology*. 2018 Sep 21;19(1):139.

- 801 57. Honkola T, Vesakoski O, Korhonen K, Lehtinen J, Syrjänen K, Wahlberg N. Cultural and
802 climatic changes shape the evolutionary history of the Uralic languages. *Journal of*
803 *Evolutionary Biology*. 2013;26(6):1244–53.
- 804 58. Barnard A. Universal systems of kin categorization. *African Studies* [Internet]. 1978 [cited
805 2020 Aug 7];37(1). Available from:
806 <https://www.tandfonline.com/doi/pdf/10.1080/00020187808707509?needAccess=true>
- 807 59. Parkin R. *Kinship: an introduction to basic concepts*. Oxford: Blackwell; 1997. xi+208.
- 808 60. Forkel R, List JM, Greenhill SJ, Rzymiski C, Bank S, Cysouw M, et al. Cross-Linguistic
809 Data Formats, advancing data sharing and re-use in comparative linguistics. *Sci Data*.
810 2018 Oct 16;5(1):1–10.
- 811 61. Forkel R, List JM. CLDFBench: Give your cross-linguistic data a lift. In *European*
812 *Language Resources Association (ELRA)*; 2020 [cited 2022 Oct 14]. p. 6995–7002.
813 Available from:
814 https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3259068
- 815 62. Haspelmath M, Tadmor U, editors. *World Loanword Database (WOLD)* [Internet].
816 Leipzig: Max Planck Institute for Evolutionary Anthropology; 2009 [cited 2019 Jun 11].
817 Available from: <https://wold.clld.org/>
- 818 63. Jakobson R. Why “Papa” and “Mama.” Perspectives in psychological theory: Essays in
819 honor of Heinz Werner. 1960;124–34.
- 820 64. Murdock GP. Cross-Language Parallels in Parental Kin Terms. *Anthropological*
821 *Linguistics*. 1959;1(9):1–5.
- 822 65. Hendery R, McConvell P. Mama and papa in Indigenous Australia. In: *Kinship Systems:*
823 *Change and Reconstruction* [Internet]. 2013 [cited 2020 May 28]. p. 217–38. Available
824 from: <https://researchdirect.westernsydney.edu.au/islandora/object/uws%3A28307/>
- 825 66. Bower C. Phylogenetic signal in the lexicon: Are parental terms influenced by baby
826 talk? *ICHL*. 2019;
- 827 67. Blasi DE, Wichmann S, Hammarström H, Stadler PF, Christiansen MH. Sound–meaning
828 association biases evidenced across thousands of languages. *PNAS*. 2016 Sep
829 27;113(39):10818–23.
- 830 68. Evans N, Levinson SC. The myth of language universals: Language diversity and its
831 importance for cognitive science. *Behavioral and Brain Sciences*. 2009 Oct;32(5):429–
832 48.
- 833 69. Viveiros de Castro E. Dravidian and related kinship systems. *Transformations of kinship*.
834 1998;332–85.
- 835 70. Dumont L. 54. The Dravidian Kinship Terminology as an Expression of Marriage. *Man*.
836 1953;53:34–9.
- 837 71. Hage P. Dravidian Kinship Systems in Africa. *L’Homme Revue française*
838 *d’anthropologie*. 2006 Jun 1;(177–178):395–407.
- 839 72. Murdock GP. *Ethnographic Atlas: A Summary*. *Ethnology*. 1967;6(2):109–236.

- 840 73. Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M. Bantu expansion
841 shows that habitat alters the route and pace of human dispersals. PNAS. 2015 Oct
842 27;112(43):13296–301.
- 843 74. Culwick AT (Arthur T, Culwick GM, Kiwanga T. Ubená of the Rivers [Internet]. London:
844 G. Allen & Unwin, Ltd.; 1935. Available from:
845 <https://ehrafworldcultures.yale.edu/document?id=fn31-001>
- 846 75. Heald S. Controlling anger: the sociology of Gisu violence [Internet]. International African
847 library. Manchester [England]: Manchester University Press, for the International African
848 Institute, London ; Distributed in the USA and Canada by St. Martin's Press; 1989. p. xi,
849 296. Available from: <https://ehrafworldcultures.yale.edu/document?id=fk13-003>
- 850 76. Kronenfeld DB. Fanti Kinship: The Structure of Terminology and Behavior¹. American
851 Anthropologist. 1973;75(5):1577–95.
- 852 77. Goody J. Cousin Terms. Southwestern Journal of Anthropology. 1970 Jul 1;26(2):125–
853 42.
- 854 78. Goody J, Irving B, Tahany N. Causal Inferences Concerning Inheritance and Property.
855 Human Relations. 1971 Aug 1;24(4):295–314.
- 856 79. Jäger G. Global-scale phylogenetic linguistic inference from lexical resources. Scientific
857 Data. 2018 Oct 9;5(1):180189.
- 858 80. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R
859 language. Bioinformatics. 2004 Jan 22;20(2):289–90.
- 860 81. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3.
- 861 82. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other
862 things). Methods in Ecology and Evolution. 2012 Apr 1;3(2):217–23.
- 863 83. Bürkner PC. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of
864 Statistical Software. 2017;80(1):1–28.
- 865 84. Pagel M, Meade A. Bayes Traits V3. Reading: University of Reading. 2017;
- 866 85. Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association.
867 1995 Jun 1;90(430):773–95.

868

869 **Supporting Information**

870 **S1 Fig. Bayesian Model convergence plots.** Plots show MCMC results for each of the
871 three co-evolutionary models (male bifurcate merging, female bifurcate merging, and
872 complete bifurcate merging). Each model is a sample of 10,000 iterations over two chains.
873 All models are seen to converge.

874

875 **S1 Table. List of locations and citations for each data collection.**

876

S2 Table. Core kin types used for primary search criteria in Kinbank. If concepts outside this list were available in a language, they were added to the database, but were not the focus of collection. See <https://github.com/kinbank/kinbank/blob/master/kinbank/cldf/parameters.csv> for the full list of concepts.

S3 Table. CLDF table descriptions of Kinbank.

S4 Table. Structural similarity scores between Parabank and VariKin collections in a set of 44 overlapping languages. The average similarity of the total structure of languages is 0.80, and with a random sample 0.74. If languages are coded identically, the structural similarity will have a score of 1, and if they are complete opposites, they will have a score of 0. Italicised rows indicate similarity scores that consider the entire sample.

S5 Table. Table of coefficients for the phylogenetically-controlled repeated measures multi-level Bayesian logistic regression. Significant coefficients are in bold.

S6 Table. Description of vowel and consonant codes. Refer to S1 table in Blasi et al. (2016).

S7 Table. Interrater reliability statistics. Between pairs of coders Cohen's kappa is used, and between all coders, Light's Kappa. Statistics can be interpreted using guidelines from Landis & Koch (1977): <0 No agreement, 0 — .20 Slight, .21 — .40 Fair, .41 — .60 Moderate, .61 — .80 Substantial, .81–1.0 Perfect.

S8 Table. Bifurcate-merging (BM) codes, derived from Kinbank. Codes for the 61 Bantu languages, alongside their taxa codes used in the Grollemund et al. (2015) phylogeny. 1 indicates presence of bifurcate merging, 0 indicates an absence, and -1 is no data.

S9 Table. Coding for Cousin marriage statistics. Presence (1) or absence (0) or permitted cross-cousin marriage using EA023 from the Ethnographic Atlas (Murdock, 1967; Kirby et al., 2016)

S10 Table. Gelman Rubin Diagnostic statistics. Statistics are shown for the independent and dependent model of each analysis. Showing the Estimate and, in brackets, the Upper 95% confidence interval. Gelman and Rubin statistic <1.1 are shown to have converged.