More complex societies have more complex kinship lexicons

Sihan Chen[1], David Gil[2], and Antonio Benítez-Burraco[3]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[3]Department of Spanish, Linguistics & Theory of Literature (Linguistics), University of Seville, Seville, Spain

Significance Statement Kinship systems vary drastically in complexity, but what caused such diversity? Previous studies have offered several hypotheses, but they mainly focus on a small sample of languages and on a specific domain of kinship (e.g. cousins). This work presents evidence for the hypothesis that more complex societies tend to have more complex kinship systems. Drawing from four different large-scale, independently constructed databases, we use state-of-the-art statistical methods to support our hypothesis. To explain the correlation between societal and kinship-term complexity, we argue that complex societies might demand more expressive power in order to communicate diverse messages accurately.

Author contributions ABB conceived the paper. SC conducted the data analyses. SC, DG, and ABB analyzed the results. SC, DG, and ABB wrote and approved the final manuscript.

Author declarations The authors declare no conflicts of interest.

Corresponding author To whom correspondence should be addressed. Postal address: Área de Lingüística. Departamento de Lengua Española, Lingüística y Teoría de la Literatura. Facultad de Filología. Universidad de Sevilla. C/ Palos de la Frontera s/n. 41004-Sevilla (España/Spain). E-mail: abenitez8@us.es

keywords linguistic diversity | sociopolitical diversity | language typology | kinship terms | complexity

Abstract

Increasing evidence suggests that language complexity is sensitive to sociopolitical factors. While most quantitative research has focused on morphology and syntax, the complexity of the lexicons of the world languages can be expected to be sensitive to extralinguistic factors too. In this paper, we have implemented a mathematical method for calculating the complexity of kinship term systems. We have furthermore conducted principal component analyses aimed to determine whether this complexity is impacted by core features characterizing sociopolitical complexity, including the status of the language within its society, the size of the language family that a language belongs to, the number of jurisdictional levels above the local community, the size of local communities, population size and density, fixity of residence, and distance moved each year. For this, we have drawn upon independently constructed databases of sociopolitical and linguistic complexity (WALS, D-Place, Ethnologue, Glottolog, and KinBank). We found that social complexity positively correlates with the complexity of kinship terms. We interpret this finding as suggesting that the languages spoken by complex societies develop greater expressive power in order to share decontextualized knowledge and know-hows with strangers. We expect that our algorithm can capture the complexity of other domains of the lexicons of the world's languages.

According to the uniformitarian hypothesis, all human languages are expected to exhibit roughly the same overall complexity (Dixon, 1997; Fromkin et al., 1998). At the same time, it is acknowledged that languages may show more complexity in one particular domain, but they will then be simpler in other domain(s) (Hockett, 1958; Miestamo, 2017). Ongoing quantitative research has provided a more nuanced view of this scenario. First, overall language complexity might differ cross-linguistically (Sampson et al., 2009; McWhorter, 2011; Koplenig et al., 2022). Second, trade-offs between language domains might not be compulsory (Shosted, 2006; Sinnemäki, 2008; Miestamo, 2009; Benítez-Burraco et al., 2024). Third, if they exist, trade-offs might not entail an equal overall complexity (Fenk-Oczlon and Fenk, 2014; Sinnwell et al., 2014; Bentz et al., 2022). Finally, global complexity, trade-offs between language domains, and even specific language features have been proved to be sensitive to extralinguistic factors, instead of just depending on factors internal to language, as previously assumed. Accordingly, the types of morphology or syntax exhibited by the world languages have been found to correlate with diverse social factors, including the type of sociopolitical organization, the tightness or the looseness of social networks, the number of speakers, the degree of bilingualism, or the number of adult learners of a language (Sinnemäki, 2009; Lupyan and Dale, 2010; Trudgill, 2011; Nettle, 2012; Atkinson et al., 2018; Gil, 2021; Chen et al., 2024).

When one considers the language features subject to variation together with the social factors impacting on language structure, an interesting pattern emerges. On the one hand, the languages with larger lexicons, increased compositionality, enhanced semantic transparency, more complex and more layered syntax (with more specialized and obligatory grammaticalized distinctions and a greater reliance on embedding), but with less complex morphology and phonology are found to be spoken by larger and more complex human groups, characterized by widespread but looser social networks, increased inter-group contacts, and generalized cultural exchanges. In contrast, the languages with more complex and more opaque morphology (with more irregularities and morpho-phonological constraints), larger sound inventories and more complex phonotactics, reduced compositionality and semantic transparency (resulting in an abundance of idioms and idiosyncratic constructions), but with simpler and less layered syntaxes, tend to be spoken by

small human groups, forming small but tight social networks, with high proportions of native speakers. Building on seminal characterizations by Bolender (2007) and Wray and Grace (2007), Chen et al. (2024) have called the first type of languages Type X languages, with X standing for eXoteric (or open) societies. Likewise, they coined the term Type S languages for the second type of languages, with S standing for eSoteric (or close-knit) societies. Potentially, one could expect that some of the correlations described above involve some sort of causation, so that specific factors external to language can indeed explain specific structural features of the world languages. More research is needed on this topic, but for instance, it has been hypothesized that the differences between Type X and Type S languages could result from a differential context-dependency. Accordingly, Type X languages would be endowed with increased expressive power (resulting from their more sophisticated syntax and larger lexicons), since they are used to convey decontextualized information to strangers. Conversely, Type S languages would exhibit more features related to group identity (like idioms or irregular items), since they are mostly used by people sharing considerable amounts of knowledge (see Bolender, 2007; Wray and Grace, 2007, for details). Another possibility (compatible with the latter) is that Type X languages are optimized for being learned and used by adults, since these languages have more non-native speakers, whereas Type S languages are easier to learn and use by children. Hence, while complex morphology is problematic for adults, children experience problems with mastering extensive vocabularies or complex syntactic structures because of memory shortages (see Lupyan and Dale, 2016; Benítez-Burraco and Kempe, 2018, for details).

In this paper, we aim to delve, specifically, into the potential effect of sociopolitical diversity, as found among human groups, on the structural diversity of the lexicons of the world's languages. As noted, most quantitative research on the effects of factors external to language on linguistic structure has focused on morphological and syntactic features. Certainly, it is widely acknowledged that languages have words for prominent aspects of their physical and cultural environments, hence the differences observed cross-linguistically between the language's lexicons. However, quantitative studies examining whether the forms of sociopolitical organization impact on the size and the complexity of the languages' vocabularies are less abundant. In their seminal work using dictionary entries for 44 written languages, Witkowski and Burris (1981) concluded that large-scale societies have larger lexicons than small-scale societies, mostly because of their greater diversity of language users and greater elaboration of usage situations. They also pointed out that although the size of the core lexicon (that is, the words known to virtually all members of a society) can be regarded as similar in all languages, differences resulting from sociopolitical factors can still be found from one language to another. Accordingly, complex societies can exhibit a smaller number of specific plant names (Witkowski and Brown, 1978), but larger numbers of color terms (Berlin and Kay, 1969; Ember, 1978) or general names for animals (Brown, 1979). Additionally, we also have studies determining changes in vocabulary sizes in historical times in selected languages (e.g. Goulden et al., 1990, for English) or selected language families (e.g. Bromham et al., 2015, for Polynesian languages). These studies also suggest that as societies grow larger and more complex, the number of content words also increases. Computational simulations reinforce this view. Accordingly, as noted by Reali et al. (2018), ease of diffusion might account for the larger vocabularies exhibited by the languages spoken by open societies, since words are linguistic conventions that are easier to learn than grammatical conventions, which require more frequent interactions between individuals to be fixed, this typically in close-knit

societies. More generally, computational approaches have shown that larger communities tend to develop larger and more expressive (and easier to understand) categorization systems, particularly, for rarely communicated meanings, and that this is due to the greater communicative challenges that larger communities experience due to their greater size and structural complexity (Lev-Ari, 2024).

In our paper, we have focused on one specific semantic field: the kinship lexicon. One practical reason is the recent release of a comprehensive database of kinship terminology in the world languages, KinBank (Passmore et al., 2023, http://www.kinbank.net). Another reason is certainly the existence of a wealthy body of research aimed at characterizing the diversity of kinship lexicons across the world's languages, as well as the ultimate foundations of the attested typologies. Kinship lexicons can vary to highlight specific social distinctions (like age, gender, generation, etc.) and also depending on whether one specific term is used for referring to different kinds of kin (like the use of the word for mother for referring to one's aunt too). Still, there are only a limited number of kinship naming systems in the world, and some of them are more abundant than others (Lévi-Strauss, 1971). Kinship seems to emerge quite early during human evolution, seemingly because it plays an important role in social life, including marriage practices, friendship patterns, or status issues (Fox, 1983; Hughes, 1988). Interestingly, while kinship systems are indeed diverse, none of the world's kinship systems have names for relatives who are more distant than cousins and second-cousins, seemingly because this level of kinship fills up all the slots for a standard clan or community (about 150 people): outside that circle we do not have personalized relationships with individuals, so extra kinship terms are not needed (Dunbar, 2009, 2022).

The diversity of kin terminology has been hypothesized to result from the interaction between the general principles governing how we create conceptual structures and how we communicate (Jones, 2010). More specifically, constraints on kinship systems might derive from a few universal mental schemas of sociality (including genealogical distance, social rank, and group membership) (Jones, 2004) and two domain-general communicative principles: simplicity and informativity, also relevant to other semantic domains (e.g. Kemp and Regier, 2012; Zaslavsky et al., 2018). That said, some research has tried to determine whether, as with other structural aspects of language, the complexity of kinship systems might also correlate with factors external to language. This approach still has limitations (hence our aim in this paper), mostly because studies have examined a narrow sample of languages, and/or have focused on specific societal factors and specific kinship terms/domains. For instance, in their survey of 73 languages, Witkowski and Brown (1978) found that societal complexity results in more collaterality distinctions, but not in more bifurcation distinctions. Likewise, Rácz et al. (2019) found that contrary to the computational findings by Reali et al. (2018), it is not population size (and ultimately, learning constraints), but social practices that mostly shape kinship systems. In our paper, we test the broader possibility that, as with other aspects of the lexicon, exoteric societies speak languages featuring more complex kinship lexicons. For this, we have relied on an ample set of features characterizing human societies, as well as on a rich description of the world's kinship systems.

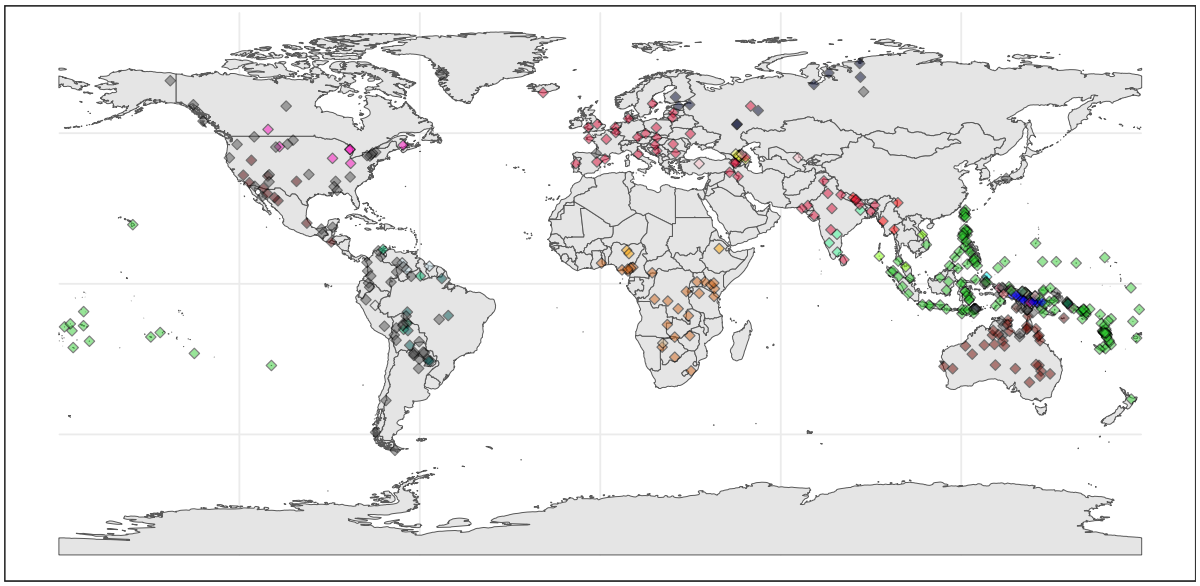

Figure 1: **The geographical distribution of the 440 languages studied in this work.** Each point represents a language, colored by the language family it belongs to, according to Glottolog (Hammarström et al., 2022). The geographical information of each language is taken from the KinBank database (Passmore et al., 2023).

Results

To quantify the relation between societal exotericity and kinship complexity, we drew data from four different, independently constructed databases. Kinship features were drawn from KinBank (Passmore et al., 2023, http://www.kinbank.net/). Meanwhile, societal features were collected from three databases: Ethnologue (Lewis, 2009, https://www.ethnologue.com/), Glottolog (Hammarström et al., 2022, https://glottolog.org) and D-Place (Kirby et al., 2016, https://d-place.org).

A total of 440 languages (Fig. 1) were part of the analysis, each with its sociopolitical complexity index (henceforth SCI), a kinship system complexity score, its phylogeny, and its geographical information available in the four databases. The SCI was calculated following the methods in (Chen et al., 2024): we first selected nine sociopolitical features from the aforementioned three databases and imputed the missing values (Stekhoven and Bühlmann, 2011; R Core Team, 2021). Then, we conducted a principal component analysis and extracted the first principal component as the SCI. Note that the more negative the SCI is, the more complex the society is. The kinship complexity score was calculated based on a minimal description length approach (Juola, 1998; Dahl, 2009; Kemp and Regier, 2012): the more it takes to fully describe a kinship system, the more complex the system is. In particular, from KinBank (Passmore et al., 2023), for each language, we gathered a list of all the concepts that could be referred to by each kinship term in that language, and we compressed the lists using the gzip algorithm (Deutsch, 1996). We operationalized the notion of description length as the length of the output of the gzip algorithm: the longer the output, the more complex the kinship system is. The phylogenetic information for each language is taken from the EDGE tree (Bouckaert et al., 2022), a tree containing the inferred phylogenetical relatedness of languages globally. The geographical information for each language is provided in the KinBank database (Passmore et al., 2023).

We then conducted a Bayesian mixed-effects linear regression between kinship complexity and sociopolitcal complexity, quantified by the kinship complexity score and the SCI, respectively, using the brms package (Bürkner, 2017). In our regression, we also controlled for Galton's problem (Roberts et al., 2015): the confound that languages coming from similar lineages might inherit similar kinship systems from their common ancestor, and that languages geographically close to each other are likely to borrow each other's kinship systems.

The analysis was conducted under an uninformative prior over 4 chains, each with 10000 iterations and a warm-up period of 5000 iterations. We reported the posterior mean, the 2.5% quantile, and the 97.5% quantile. We say the result is significant if the 2.5% quantile and the 97.5% quantile are both positive or both negative. Specifically, because a more negative SCI value corresponds to a more complex society, and a more positive SCI value corresponds to a less complex society, a negative relation between SCI and kinship complexity implies a positive correlation between societal complexity and kinship complexity, and vice versa.

Figure 2 shows the global relation between SCI and kinship complexity, and the results suggest a negative correlation between SCI and kinship complexity and hence a positive correlation between sociopolitical complexity and kinship complexity. A linear regression suggests that this is indeed the case on a global scale (β = -21.083, $p < 0.001$). Results from the Bayesian mixed-effects linear regression indicate that the effects are not just due to language relatedness and geographical proximity [β = -13.43; 95% posterior credible interval (-22.85, -3.94)]. Figure

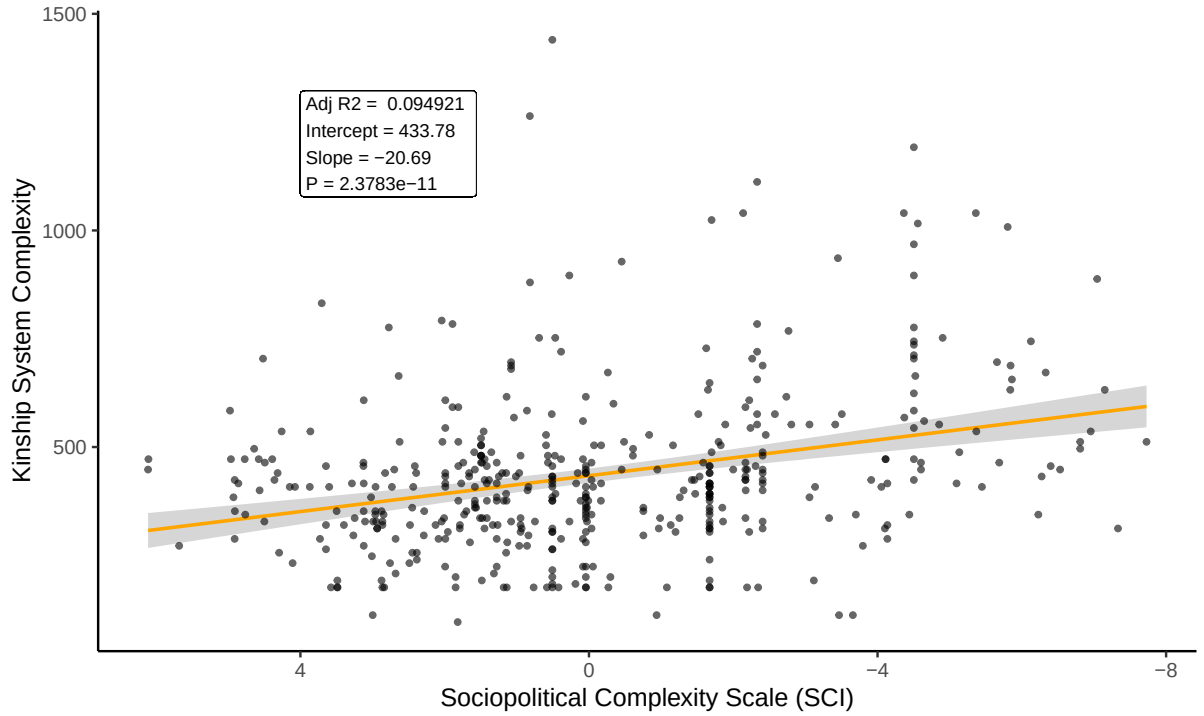

Figure 2: **Kinship system complexity (y-axis) is plotted against the sociopolitical complexity index (SCI), the metric for sociopolitical complexity (x-axis)**. The complexity of kinship system is calculated using the KinBank database (Passmore et al., 2023), following a descriptive complexity approach. The sociopolitical complexity score is calculated from a principal component analysis, drawing data from Ethnologue (Lewis, 2009), Glottolog (Hammarström et al., 2022), and D-Place (Kirby et al., 2016) A more negative SCI value indicates a more complex society, and vice versa. Results from a linear regression suggest that on a global scale, more complex societies tend to have a more complex kinship system (β = -21.083, $p < 0.001$). The trend is further supported by a Bayesian mixed-effects linear regression controlling for language phylogeny and geographical proximity [β = -13.43, 95% posterior credible interval (-22.85, -3.94)].

S1 shows a breakdown of this analysis by language families, which suggests that this positive correlation between sociopolitical complexity and kinship complexity is stronger in some groups, such as Indo-European, Dravidian, and Pama-Nyungan.

Discussion

Previous research has suggested that the complexity of different structural aspects of language are not simply driven internally by other aspects of language. Instead, they are subject to the pressure from various extralinguistic factors, such as environmental and social factors (e.g. Sinnemäki, 2009; Lupyan and Dale, 2010; Trudgill, 2011; Nettle, 2012; Atkinson et al., 2018; Blasi et al., 2019; Gil, 2021; Everett and Chen, 2021; Chen et al., 2024). In this work, we have focused on the relation between the lexical complexity of a language and the societal structure of the population speaking the language. In particular, our hypothesis was that the languages spoken by exoteric societies (Type X languages) exhibit larger vocabularies, and specifically, more complex kinship systems. Our results align with our broad hypothesis. Overall, we found that languages spoken in exoteric societies indeed tend to have more complex kinship systems, and this effect is robust after controlling for language relatedness and language contact.

Whereas the diversity of kinship systems has been known and described for a long time, the reasons why different human groups refer differently to their relatives is less clear. Past studies modeling kinship systems as codes in a communication game suggested that they largely satisfy two constraints related to communication: accuracy and simplicity (Kemp and Regier, 2012). Specifically, Kemp & Regier (Kemp and Regier, 2012) found that given a level of complexity, kinship systems attested in human languages achieved near-optimal communicative accuracy. However, their results still leave one question open: why do different languages have kinship systems of various complexity in the first place? This paper offers a possible factor: the societal complexity where the language is spoken. A more complex society might demand more expressive power to be able to share more diverse and more decontextualized knowledge and know-hows with others and to do so accurately. In terms of communication, this implies that languages spoken in more complex societies tend to be willing to trade-off simplicity in exchange for more accuracy. This is in line with previous computational approaches (e.g. Lev-Ari, 2024) showing that larger communities tend to create more expressive categorization systems seemingly in response to increased communicative challenges due to their size and complexity. But it is also in line with previous research showing that more complex societies also speak languages with more complex syntaxes (e.g. Gil, 2021; Chen et al., 2024).

Our approach to complexity is more fine-grained in that it is able to capture the complexity due to both the number of kinship terms and the number of kinship concepts distinguished by kinship terms. Past studies (e.g. Kemp and Regier, 2012; Rácz et al., 2019) based their analysis on the Murdock data (Murdock, 1970), which only recorded which concepts were referred to by the same word, but not the number of words referring to the same concept. The latter is informative: having multiple words referring to the same concept could potentially imply differentiations in formality or intimacy (e.g. “dad” vs. “father”), which should contribute to the complexity of a kinship system. Hence, it is possible that the reason why Rácz et al. (2019) did not find a significant relationship between kinship system complexity and societal complexity is that

the Murdock data (Murdock, 1970) did not include such nuanced information. Ultimately, our study reinforces the convenience of adopting information-theoretic approaches to the study of complexity in human languages. Future research on this topic would benefit from incorporating the frequency of use of each kinship term, which is unavailable for most of the languages in current kinship databases. For example, across different dialects of Malay/Indonesian, "younger sister" is expressed by a collocation of two words, "younger sibling" and "woman"; however, while in most dialects the collocation is most appropriately analyzed as a kinship term "younger sister" plus a descriptive noun, in the Papuan dialect, the same collocation occurs with substantially greater frequency, suggesting that it may be in the process of becoming lexicalized as a single complex kinship term "younger sister". More generally, approaches to kinship complexity based on Optimality Theory (Jones, 2004) seem promising, since the observed kinship systems are hypothesized to result from the satisfaction of diverse conflicting constraints, although it should be now clear that these constraints are not only "internal" (i.e. cognitive, informational), but also "external" (sociopolitical). Still, our approach has some limitations. Seemingly, the most important is the quality of the data, which can be pretty variable, with some kinship systems being characterized in their entirety and in similar ways by different researchers, but with others being described only partially, in some inconsistent ways, or in a variety of formats. However, this is still a problem common to most studies based on cross-linguistic databases (see Anderson et al., 2018; Forkel et al., 2018; Rzymski et al., 2020).

Data Availability

All the raw data, analysis code, and the figures and tables generated are available at https://github.com/cshnican/kinship_complexity.

Materials and Methods

Database preprocessing

KinBank (Passmore et al., 2023) is a database containing kinship term information on 1235 languages coming from diverse language families and locations around the world. It contains a list of core concepts (e.g. "male ego's father", "female ego's father's sister's son") and for each language, if attested, the lexical form corresponding to each concept. Prior to the analysis, we removed duplicated records. In particular, some of the kinship terms in English and German were written in both Latin script and the International Phonetic Alphabet (IPA), possibly because the authors incorporated different sources when constructing the database. Since not all of the kinship terms in English and German were shown in both scripts, we removed the records written in IPA. Similarly, some kinship terms in Russian were available in both the Latin script and the Cyrillic script, and because not all of the kinship terms were presented in both scripts, we removed the records written in the Cyrillic script.

Computing the sociopolitical complexity index (SCI)

Following Chen et al. (2024), for each language we computed a sociopolitical complexity index (SCI), considering the following nine societal features. The first two are related to the language status in the society where it is spoken, as defined by the Expanded Graded Intergenerational Disruption Scale (henceforth EGIDS) in Ethnologue (Lewis, 2009): the first scale reflects the gradient nature of language status, ranging from 1 (an extinct language, corresponding to extreme esotericity) and 13 (a lingua franca, corresponding to extreme exotericity), whereas the second scale (called EGIDSnat) reflects whether a language is a national language or not, with 1 indicating it's not a national language and 2 indicating it is a national language. The third feature is the size of the language family a language belongs to, as measured by the number of languages belonging to the same language family, suggesting the degree of migration and expansion, according to the classification on Glottolog (Hammarström et al., 2022), ranging from 1 (language isolates, an extreme case of esotericity) to 1433 (languages in the Atlantic-Congo family, an extreme case of exotericity). The remaining 6 features are drawn from the D-Place database (Kirby et al., 2016): the number of jurisdictional levels above the local community (Feature EA033 in the database), the size of local communities (EA031), population size (EA202) and density (SCCS156), fixity of residence (SCCS150), and distance moved each year (B014). An exoteric society tends to have more jurisdictional levels, larger local communities, larger population size, and higher population density; moreover, people living in an exoteric society are also more likely to settle at a place (instead of being nomadic) and more likely to move around (Chen et al., 2024).

Following (Chen et al., 2024), based on the aforementioned 9 features, we calculated a sociopolitical complexity score using principal component analysis (PCA). We first imputed the missing values in the dataset with the missforest package (Stekhoven and Bühlmann, 2011) in R (R Core Team, 2021). Then, we ran a PCA on these 9 features after normalizing them and extracted dimensions that captured the most variance in the data, with the prcomp function in R. We considered the first principal component (PC1) as the SCI (called PC1 in Chen et al., 2024), as it explained 56.8% of the variance in the data (See Figure 1 in Chen et al., 2024, for a visualization of the SCI). The more negative the SCI is for a society, the higher complexity it has, and hence the more exotericity.

Computing kinship complexity for each language

To estimate the complexity of the kinship system in each language, we followed a descriptive complexity approach (e.g. Juola, 1998; Dahl, 2009; Kemp and Regier, 2012): given a number of primitive concepts and their compositions, a kinship system is more complex if the total description length to define each kinship term in the system is longer. Below is a rough outline of our procedures.

The KinBank database has a list of primitive concepts: FEMALE EGO (coded as f), MALE EGO (m), FATHER (F), MOTHER (M), BROTHER (B), SISTER (Z), SON (S), DAUGHTER (D), HUSBAND (H), WIFE (W), CHILD (C), BORN ON THE SAME DAY (BornSameDay), HUSBAND FROM THE SAME GROUP (HusbandSameGroup), and WIFE FROM THE SAME GROUP (WifeSameGroup). It also has a list of primitive modifiers aimed at qualifying the former: younger (y), elder (e),

exchange (exchange), agnatic (_a_), cognatic (_c_), co- (_co_), or (_or), and not (_not). These primitive concepts and modifiers compose and generate new concepts (e.g. female ego's father's younger sister is coded as fFyZ). One limitation of this coding is that some of the aforementioned primitive concepts are actually compositional. For example, the concept FATHER can be rewritten as a composition between a broader concept PARENT with a modifier male. Therefore, we modified the concepts by using a different set of primitive concepts: EGO(E), PARENT (P), SIBLING (S), CHILD (C), and SPOUSE (O), with the same set of modifiers but different abbreviations: male (m), female (f), younger (y), elder (e), exchange (X), agnatic (_a_), cognatic (_c_), co- (_co_), or ($\hat{}$), and not (). For example, now the concept FATHER will be coded as fP instead of F.

Then, for each kinship term in each language, we obtained a list of concepts that could be referred to by the term (called **extension** in Kemp and Regier, 2012). For instance, the extension of the word sister in English, represented in the revised code, is (fEefS, fEfS, fEyfS, mEefS, mEfS, mEyfS, representing {male, female} ego's {younger, elder, $\varnothing$ } female sibling). Next, we compiled a list of kinship terms in each language, and substituted each kinship term by its list of extensions, with different lists separated by a backslash ('/'). Using the gzip algorithm (Deutsch, 1996), we compressed the extension list of each language into a raw vector and then calculated the length of the vector. The length of this vector then serves as the complexity metric of the system. As with information compression in, for example, computer files, this process is mostly aimed at keeping the essential information while removing the redundant ones, mainly by looking at patterns in the data. The more essential information a kinship system has, the more complex we assume it is. For example, consider two strings 'abcabcabcabcabcabcabc' and 'f98&juc#87?[*fgcba'. The former string can be compressed to "repeat abc 7 times", whereas the latter string cannot be compressed into a shorter one, and therefore, even though both strings have the same original length, the latter string is considered to be more complex since it contains more essential information. Indeed, the compressed former string has a length of 13 under the gzip algorithm, whereas the latter has a length of 29.

Our method is an approximation of the method in Kemp and Regier (2012) on a computational level. In their method, Kemp and Regier (2012) defined a list of primitives and logical operations, based on which a list of concepts were generated, and for each concept, the extensions were also computed. In their work, the complexity was defined as the minimal number of concepts in a set such that 1) each kinship term has the same extension as one concept in the set and 2) each kinship term can be defined by at least one other concept in the set, plus one primitive and one logical operation. On the other hand, the gzip algorithm in our approach looks for repeated patterns in the extension list so that a list full of repeated patterns has a shorter vector than a list of the same length but without repeated patterns. Repeated patterns capture two critical components in Kemp and Regier (2012) in the following two senses. First, among the extensions of the same kinship term, repeated patterns reflect the existence of logical operations involved when defining the term. Second, among the extension lists of the same language, repeated patterns indicate that new kinship terms are defined based on old kinship terms.

As an illustration of the metrics, Table S1 shows a comparison between the sociopolitical features of the two specific societies speaking two specific languages: Eastern Panjabi and Dakota, and their corresponding SCI. Compared to Dakota, Eastern Panjabi belongs to a larger language family and is considered less endangered. Eastern Panjabi-speaking societies have more judicial

hierarchy, a larger population, community size, and population density. They are also more likely to be sedentary (in contrast to being nomadic) and are more likely to travel around. Based on these factors, Eastern Panjabi has an SCI of -2.42, whereas Dakota has an SCI of 4.19, and therefore Eastern Panjabi is on the more exoteric side of the spectrum, whereas Dakota is on the esoteric side.

To illustrate what aspect of kinship system increases the complexity score under this approach, Table S2 presents three pairs of toy languages, each describing a very limited subset of the kinship concepts. The first pair shows that languages with more kinship terms have a higher kinship complexity: Language A only has one word, referring to the concept fEfP (female ego's female parent), whereas Language B has two words, one referring to the concept fEfP, and another referring to the concept fEmP (female ego's male parent). Indeed, Language B has a complexity of 34, higher than Language A's complexity of 24. The second pair shows that languages discriminating more concepts have a more complex kinship system: both Language C and Language D have 4 kinship terms, but the terms in Language C only distinguish two concepts, whereas those in Language D distinguish four concepts. In particular, Language C contains two pairs of synonyms: the first two words both refer to fEfP, and the last two words both refer to mEmP (male ego's male parent). In contrast, each word in Language D refers to a different concept. Indeed, Language D has a complexity of 42, compared to 40 in Language C. The third pair shows that languages with more systematic extensions for each concept have a lower complexity: in this example, both Language E and Language F have only one word, each referring to four kinship concepts. The concepts that the word in Language E refer to are more systematic, since fEfSmC (female ego's female sibling's male child), fEmSmC (female ego's male sibling's male child), fEfSfC (female ego's female subling's female child), fEmSfC (female ego's male sibling's female child) can be compressed to female ego's {male, female} sibling's {female, male} child, which in turn can be compressed to female ego's sibling's child. In contrast, concepts that the word in Language F refer to are less systematic, as mEmSfC (male ego's male sibling's female child), mEfPfS (male ego's female parent's female sibling), fEfOmP (female ego's female spouse's male parent), and fEmPmS (female ego's male parent's male sibling) cannot be compressed in an efficient way. As a result, Language F has a higher complexity than Language E according to the algorithm (66 vs. 52).

Analysis

We then conducted a Bayesian mixed-effects linear regression between kinship complexity, quantified by the description length, and sociopolitical complexity, quantified by the principal component value, using the brms package (Bürkner, 2017) in R (R Core Team, 2021) To control for Galton's problem (Roberts et al., 2015), we specified two random effect structures in the form of covariance matrices, following previous works (Chen et al., 2024; Shcherbakova et al., 2024). The first covariance matrix specifies the language relatedness: if two languages are more closely related to each other, they will have a higher covariance (e.g. English and Dutch) compared to those that are not so related to each other (e.g. English and Finnish). The covariance between each language pair was calculated from their distance on a globally reconstructed phylogeny of languages (EDGE tree, Bouckaert et al., 2022), using the ape package (Paradis et al., 2019) in R. The second covariance matrix specifies how close languages are to each other, based on their

great circle distance calculated from their coordinates provided in the AUTOTYP database (Bickel et al., 2022). The great circle distances were first transformed to Matérn covariances using the geoR package (Ribeiro Jr and Diggle, 2006) and then normalized against the maximal covariance. SCI, representing sociopolitical complexity, is coded as a fixed effect. The regression can be written as the equation below, following the syntax in the brms package:

$$\begin{aligned}\text{kinship complexity} \sim \text{SCI} &+ (1 \mid \text{gr}(\text{Glottocode, A})) \\ &+ (1 \mid \text{gr}(\text{Glottocode, B}))\end{aligned}$$

The analysis was conducted under an uninformative prior over 4 chains, each with 10000 iterations and a warm-up period of 5000 iterations. We reported the posterior mean, the 2.5% quantile, and the 97.5% quantile. We say the result is significant if the 2.5% quantile and the 97.5% quantile are both positive or both negative. Specifically, because a more negative SCI value corresponds to a more complex society, and a more positive SCI value corresponds to a less complex society, a negative relation between SCI and kinship complexity implies a positive correlation between societal complexity and kinship complexity, and vice versa.

References

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., and List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, 4(1):21–53.

Atkinson, M., Mills, G. J., and Smith, K. (2018). Social group effects on the emergence of communicative conventions and language complexity. *Journal of Language Evolution*, 4(1):1–18.

Bentz, C., Gutierrez-Vasques, X., Sozinova, O., and Samardžić, T. (2022). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*, 9(s1):9–25.

Benítez-Burraco, A., Chen, S., and Gil, D. (2024). The absence of a trade-off between morphological and syntactic complexity. *Frontiers in Language Sciences*, 3.

Benítez-Burraco, A. and Kempe, V. (2018). The emergence of modern languages: Has human self-domestication optimized language transmission? *Frontiers in Psychology*, 9.

Berlin, B. and Kay, P. (1969). *Basic color terms: Their universality and evolution*. Univ of California Press.

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., Bierkandt, L., Zúñiga, F., and Lowe, J. B. (2022). The autotyp database.

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., and Bickel, B. (2019). Human sound systems are shaped by post-neolithic changes in bite configuration. *Science*, 363(6432).

Bolender, J. (2007). Prehistoric cognition by description: a russellian approach to the upper paleolithic. *Biology amp; Philosophy*, 22(3):383–399.

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., and Atkinson, Q. (2022). Global language diversification is linked to socio-ecology and threat status.

Bromham, L., Hua, X., Fitzpatrick, T. G., and Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7):2097–2102.

Brown, C. H. (1979). Folk zoological life-forms: Their universality and growth. *American Anthropologist*, 81(4):791–817.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1).

Chen, S., Gil, D., Gaponov, S., Reifegerste, J., Yuditha, T., Tatarinova, T., Progovac, L., and Benítez-Burraco, A. (2024). Linguistic correlates of societal variation: A quantitative analysis. *PLoS One*.

Dahl, O. (2009). *Testing the assumption of complexity invariance: the case of Elfdalian and Swedish*, page 50–63. Oxford University PressOxford.

Deutsch, P. (1996). Gzip file format specification version 4.3. Technical report.

Dixon, R. M. (1997). *The rise and fall of languages*. Cambridge University Press.

Dunbar, R. (2009). Mind the bonding gap: constraints on the evolution of hominin societies.

Dunbar, R. I. M. (2022). Managing the stresses of group-living in the transition to village life. *Evolutionary Human Sciences*, 4.

Ember, M. (1978). Size of color lexicon: Interaction of cultural and biological factors. *American Anthropologist*, 80(2):364–367.

Everett, C. and Chen, S. (2021). Speech adapts to differences in dentition within and across populations. *Scientific Reports*, 11(1).

Fenk-Oczlon, G. and Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznan Studies in Contemporary Linguistics*, 50(2).

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1).

Fox, R. (1983). *Kinship and marriage: An anthropological perspective*. Number 50. cambridge university press.

Fromkin, V., Rodman, R., and Hyams, V. (1998). *An Introduction to Language 6e*. Orlando, FL: Hartcourt Brace College Publishers.

Gil, D. (2021). Tense–aspect–mood marking, language-family size and the evolution of predication. *Philosophical Transactions of the Royal Society B*, 376(1824):20200194.

Goulden, R., Nation, P., and Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11(4):341–363.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2022). glottolog/glottolog: Glottolog database 4.6.

Hockett, C. F. (1958). A course in modern linguistics.

Hughes, A. L. (1988). *Evolution and human kinship*. Oxford University Press.

Jones, D. (2004). The universal psychology of kinship: evidence from language. *Trends in Cognitive Sciences*, 8(5):211–215.

Jones, D. (2010). Human kinship, from conceptual structure to grammar. *Behavioral and Brain Sciences*, 33(5):367–381.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Kemp, C. and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., Leehr, D., Low, B. S., McCarter, J., Divale, W., and Gavin, M. C. (2016). D-place: A global database of cultural, linguistic and environmental diversity. *PLOS ONE*, 11(7):e0158391.

Koplenig, A., Wolfer, S., and Meyer, P. (2022). Human languages trade off complexity against efficiency.

Lev-Ari, S. (2024). The influence of community structure on how communities categorize the world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Lévi-Strauss, C. (1971). *The elementary structures of kinship*. Number 340. Beacon Press.

Lewis, M. P., editor (2009). *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition.

Lupyan, G. and Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, 5(1):e8559.

Lupyan, G. and Dale, R. (2016). Why are there different languages? the role of adaptation in linguistic diversity. *Trends in cognitive sciences*, 20(9):649–660.

McWhorter, J. H. (2011). *Linguistic simplicity and complexity: Why do languages undress?*, volume 1. Walter de Gruyter.

Miestamo, M. (2009). *Implicational hierarchies and grammatical complexity*.

Miestamo, M. (2017). Linguistic diversity and complexity. *Lingue e linguaggio*, 16(2):227–254.

Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology*, 9(2):165.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1597):1829–1836.

Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H. S., Desper, R., and Didier, G. (2019). Package 'ape'. *Analyses of phylogenetics and evolution*, 2(4):47.

Passmore, S., Barth, W., Greenhill, S. J., Quinn, K., Sheard, C., Argyriou, P., Birchall, J., Bowern, C., Calladine, J., Deb, A., Diederen, A., Metsäranta, N. P., Araujo, L. H., Schembri, R., Hickey-Hall, J., Honkola, T., Mitchell, A., Poole, L., Rácz, P. M., Roberts, S. G., Ross, R. M., Thomas-Colquhoun, E., Evans, N., and Jordan, F. M. (2023). Kinbank: A global database of kinship terminology. *PLOS ONE*, 18(5):e0283218.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reali, F., Chater, N., and Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society B: Biological Sciences*, 285(1871):20172586.

Ribeiro Jr, P. J. and Diggle, P. J. (2006). Analysis of geostatistical data. *The geoR package, version*, pages 1–6.

Roberts, S. G., Winters, J., and Chen, K. (2015). Future tense and economic decisions: Controlling for cultural evolution. *PLOS ONE*, 10(7):e0132145.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., Chang, S., Lai, Y., Morozova, N., Arjava, H., Hübler, N., Koile, E., Pepper, S., Proos, M., Van Epps, B., Blanco, I., Hundt, C., Monakhov, S., Pianykh, K., Ramesh, S., Gray, R. D., Forkel, R., and List, J.-M. (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(1).

Rácz, P., Passmore, S., and Jordan, F. M. (2019). Social practice and shared history, not social scale, structure cross-cultural complexity in kinship systems. *Topics in Cognitive Science*, 12(2):744–765.

Sampson, G., Gil, D., and Trudgill, P. (2009). *Language complexity as an evolving variable*, volume 13. Oxford University Press.

Shcherbakova, O., Blasi, D. E., Gast, V., Skirgård, H., Gray, R. D., and Greenhill, S. J. (2024). The evolutionary dynamics of how languages signal who does what to whom. *Scientific Reports*, 14(1).

Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, 10(1):1–40.

Sinnemäki, K. (2008). Complexity trade-offs in core argument marking. pages 67–88. John Benjamins.

Sinnemäki, K. (2009). Complexity in core argument marking and population size. In *Language complexity as an evolving variable*, pages 126–140. Oxford University Press.

Sinnwell, J. P., Therneau, T. M., and Schaid, D. J. (2014). The kinship2 r package for pedigree data. *Human heredity*, 78(2):91–93.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press, USA.

Witkowski, S. R. and Brown, C. H. (1978). Lexical universals. *Annual Review of Anthropology*, 7(1):427–451.

Witkowski, S. R. and Burris, H. W. (1981). Societal complexity and lexical growth. *Behavior Science Research*, 16(1–2):143–159.

Wray, A. and Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, 117(3):543–578.

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.