Kinship categories across languages reflect general communicative principles

Charles Kemp[1] & Terry Regier[2]

[1]Department of Psychology, Carnegie Mellon University
[2]Department of Linguistics, Cognitive Science Program, University of California, Berkeley

Summary sentence: The competing principles of simplicity and informativeness help to explain which kin classification systems are found in the languages of the world.

Languages vary in their systems of kinship categories but the scope of possible variation appears to be constrained. Previous accounts of kin classification have often emphasized constraints that are specific to the domain of kinship and are not derived from general principles. Here we propose an account that is founded on two domain-general principles: Good systems of categories are simple, and they enable informative communication. We show computationally that kin classification systems in the world's languages achieve a near-optimal tradeoff between these two competing principles. We also show that our account explains several specific constraints on kin classification proposed previously. Because the principles of simplicity and informativeness are also relevant to other semantic domains, the tradeoff between them may provide a domain-general foundation for variation in category systems across languages.

Concepts and categories vary across cultures but may nevertheless be shaped by universal constraints (*1–4*). Cross-cultural studies have proposed universal constraints that help to explain how colors (*5,6*), plants, animals (*7,8*), and spatial relations (*9,10*) are organized into categories. Kinship has traditionally been a prominent domain for studies of this kind, and researchers have described many constraints that help to predict which of the many logically possible kin classification systems are encountered in practice (*11–15*). Typically these constraints are not derived from general principles, although it is often suggested that they are consistent with cognitive and functional considerations (*2, 11–13, 15*). Here we show that major aspects of kin classification follow directly from two general principles: Categories tend to be simple, which minimizes cognitive load, and to be informative, which maximizes communicative efficiency. Principles like these have been discussed in other contexts by previous researchers (*16–19*). For example, Zipf suggests that word-frequency distributions achieve a tradeoff between simplicity and communicative precision (*20, 21*), Hawkins (*22*) suggests that grammars are shaped by a tradeoff between simplicity and communicative efficiency, and Rosch (*23*) suggests that category systems "provide maximum information with the least cognitive effort" (p 190).

Figure 1A shows a simple communication game that helps to illustrate how kin classification systems are shaped by the principles of simplicity and informativeness. The speaker has a specific relative in mind and utters the category label for that relative. Upon hearing this category label, the hearer must guess which relative the speaker had in mind. The speaker and hearer communicate through a shared system of categories that specifies a category label for each relative. This system is simple to the extent that it can be concisely mentally represented, and therefore easily learned and remembered (*11*). The system is informative to the extent that it supports successful communication. The principles of simplicity and informativeness trade off against each other (*?, 20, 21, 23*). A system with a single category that includes all possible relatives would be simple but uninformative because this category does not help to pick out specific relatives. A system with a different name for each relative would be complex but highly informative because it picks out individual relatives perfectly.

Understanding how simplicity and informativeness trade off in a particular domain requires assumptions about the structure of that domain. Analyses based on generic assumptions can be productive (Figure 1B), but in-depth analyses of specific domains will need to formalize simplicity and informativeness in ways that are sensitive to the structural properties of those domains. For example, analyses of kin classification (Figure 1A) should reflect the fact that kinship categories are defined over relatives embedded within a genealogical system, and analyses of color classification (Figure 1C) should reflect the fact that colors are embedded within a continuous perceptual space. In order to explore whether kin classification systems are shaped by the tradeoff between simplicity and informativeness, we formulate versions of these general constraints that are appropriate for the domain of kinship.

The kin classification systems we consider include terms that refer to the kin types in Figure 2A, namely grandparents, parents, aunts, uncles, siblings, children, nieces, nephews, and grandchildren. This is the largest set of kin types for which we have kin naming data and for which our analyses are computationally tractable. Previous studies that chart the space of logically possible classification systems have focused on grandparents (*24*), siblings (*11*), or uncles and aunts (*13*) in isolation, and the classification systems that we consider are large by comparison. The systems in Figure 2A, however, do not include cousins, which have played a prominent role in previous taxonomies of kin classification systems (*2*). The supplementary material (*25*) describes how our approach extends to systems including cousins and other more distant relatives in the family tree. Here we show that our approach accounts for the substantial cross-language variation found in categorizing the kin types shown in Figure 2.

The first tree in Figure 2A includes 24 relatives of a woman labeled as Alice (here we assume one relative per kin type; the supporting material (*25*) reports analyses with different numbers of relatives per kin type) and the second tree includes 32 relatives of a man labeled as Bob. Only the second tree includes niblings (nieces and nephews), because our data include information about nibling categories for male speakers only. Each possible kin classification system that we consider partitions the 56 relatives in the two trees into a set of non-overlapping categories. That is, each possible system includes terms that allow Alice and Bob to refer to each of their relatives, and exactly one term is available for each relative. The colors in Figure 2A show a partition that corresponds to the English kin classification system. In the case of English, the partitions of the two trees are identical, but some other languages include different terms for speakers of different sexes. For example, Figure 2B illustrates the kin classification system of Northern Paiute, an indigenous language of the western United States, in which men and women use different terms to refer to their grandchildren (*26*). In Northern Paiute, unlike English, the kin terms for grandparents and grandchildren are self-reciprocal; for example, Alice and her maternal grandmother use the same term to refer to each other. We work with a cross-cultural data set compiled by Murdock (*27*) that includes kin classification systems for 566 languages. In compiling these data Murdock aimed to cover the set of kin classification systems that had been described in the literature, and took care not to include closely related languages with similar classification systems. Some of the kin classification systems in Murdock's data are incomplete and do not specify kin categories for all kin types in Figure 2A, but the data set specifies complete systems for 487 languages in total. These systems include 410 distinct types, of which the most frequent occurs 6 times, and these 410 types represent only a tiny fraction of the 10^{55} systems that are possible in theory.

We hypothesized that the tradeoff between simplicity and informativeness can help to explain which of the many possible systems are attested, or found to exist in actuality. Intuitively, a kin classification system is complex if it includes many terms that must be learned and remembered, and if each of those terms has a complex definition. We formalize this idea by assuming that kin classification systems are mentally encoded in a representation language, and that the complexity of a system corresponds to the length of its shortest description in this language.

The representation language is assumed to be universal, but kin classification systems for different cultures can be created by combining elements of this language in different ways (*28*). Figure 2C shows the shortest description of the English system in a representation language that we now describe. The representation language includes a small set of primitives, all drawn from previous accounts of kinship classification (*29–31*), including features like $\texttt{FEMALE}(\cdot)$ and relations like $\texttt{PARENT}(\cdot,\cdot)$. The full set of primitives is shown in Figure 3A. Rules for combining these primitives are again based on previous formal accounts (*30, 32–34*), and are shown in Figure 3B. The first six rules indicate that a new relation $\texttt{C}(\cdot,\cdot)$ can be defined as a conjunction or disjunction involving two relations or a relation and a feature. For example, Figure 2C uses the first rule in Figure 3B to define $\texttt{mother}(\cdot,\cdot)$ as the conjunction of $\texttt{PARENT}(\cdot,\cdot)$ and $\texttt{FEMALE}(\cdot)$. The seventh rule indicates that a new relation $\texttt{C}(\cdot,\cdot)$ can be defined as the relative product of two existing relations. For example, Figure 2C defines $\texttt{grandmother}(\cdot,\cdot)$ as the relative product of $\texttt{mother}(\cdot,\cdot)$ and $\texttt{PARENT}(\cdot,\cdot)$. The final three rules indicate that a new relation can be the inverse, symmetric closure ($\texttt{A}^{\leftrightarrow}(\cdot,\cdot)$) or transitive closure ($\texttt{A}^{+}(\cdot,\cdot)$) of an existing relation. Figure 2D uses the inverse rule to define the child of a man's sister ($\texttt{mansisterchild}(\cdot,\cdot)$) as the inverse of $\texttt{maternaluncle}(\cdot,\cdot)$. The symmetric closure rule is used in Figure 2D to define the four self-reciprocal terms that refer to grandparents and grandchildren. The transitive closure rule can capture systems where the same term is used to refer to a parent and a grandparent, or to a child and a grandchild. Given the conceptual resources in Figure 3, the complexity of a system is the smallest number of rules needed to define all terms in the system. For example, the complexity of the English system (Figure 2C) is 15, and the complexity of the Northern Paiute system (Figure 2D) is 24. The complexity of a system can exceed the number of terms in the system: The Northern Paiute system includes 18 terms, but Figure 2D requires 24 rules in order to define these 18 terms. Our algorithm for computing the complexity of a system is described in the supporting material (*25*).

Consider now the dimension of informativeness (*35, 36*). Suppose that the 24 individuals in Alice's family tree are numbered from left to right and top to bottom, and let $\boldsymbol{z}$ be a vector that represents a partition of the 24 individuals into categories, where z_i represents the kinship category used to label individual i. For example, if Alice is an English speaker, then z_1 will equal z_3 because individuals 1 and 3 are both grandmothers. Suppose now that Alice wants to refer to individual 1 (her maternal grandmother) and uses the phrase "my grandmother." Because Alice has two grandmothers, some additional information must be specified to pick out the individual she has in mind. Information theory holds that the additional cost c_i in bits when referring to individual i is

$$c_i = -\log_2 \left(\frac{p_i}{\displaystyle\sum_{z_j = z_i} p_j} \right) \tag{1}$$

where p_i is the probability that Alice will need to refer to individual i. For example, if Alice is equally likely to refer to her maternal and paternal grandmothers, then one extra bit must be communicated in addition to the phrase "my grandmother" to indicate which grandmother she has in mind. The communicative cost C for Alice is defined as the expected cost when Alice needs to refer to one of the individuals in her family tree:

$$C = \sum_{i=1}^{24} p_i c_i. \tag{2}$$

The communicative cost for Bob is defined similarly, and we define the communicative cost of an entire kin classification system as the average of the costs for Alice and Bob.

The need probabilities p_i play an important role in Equations 1 and 2, and can in principle capture the fact that different cultures impose different communicative requirements (*2, 37*). Because we lack data that would allow us to estimate need probabilities on a per-language basis, we provisionally assume a universal distribution of need probabilities. We estimated these probabilities by computing the relative frequencies of kin expressions of the form "my grandmother," "my mother," "my daughter," "my granddaughter" and the like across corpora for two languages: the Corpus of Contemporary American English (*38*) and the German Reference Corpus (*39*). Relative frequencies were similar for English and German, and Figure 3C is based on combined results across these two languages. Although the need probabilities in Figure 3C are based on English and German, some of the same qualitative patterns may apply to many cultures. For example, because every member of a society has parents but some members do not have children, it follows that references to parents should tend to be more common across the whole society (*40*). We will use the probabilities in Figure 3C for all analyses to follow, but future studies can explore whether our results can be improved by using different sets of need probabilities for different cultures.

Given our definitions of complexity and communicative cost, we will say that one kin classification system *dominates* another if it is superior along one dimension and no worse along the other. For example, Northern Paiute does not dominate English (English is simpler) and English does not dominate Northern Paiute (Northern Paiute enables more informative communication). A system is "near-optimal" (*6, 10*) if it is dominated by few alternatives, and we can now explore whether attested systems are near-optimal with respect to the space of possible systems.

We tested the near-optimality hypothesis using a series of analyses that range in scope from broad to focused. Because the complete space of possible systems is too large to enumerate, we began by exploring a large subset of this space that is likely to include all of the near-optimal candidates. We began by enumerating around $71,000$ distinct kinship categories that can be defined by starting with the primitives in Figure 3A and applying the rules in Figure 3B up to three times. Each of these categories corresponds to a subset of the 56 individuals in Figure 2A. The number of kin classification systems that can be built from these categories is extremely large, and we therefore sampled a representative subset of these systems. Figure 4A plots these systems along our two dimensions. The best systems according to our account are located along the *optimal frontier*, also known as the *Pareto frontier*, which corresponds to the bottom left boundary of the space. The majority of attested systems (black circles) are found near the optimal frontier. Whereas Figure 4A explores a sample from the space of all possible kin classification systems, Figure 4B shows the results of a more focused analysis that includes all and only the 8.3×10^8 systems that can be created by combining categories that appear in more than two attested systems. Again the attested systems tend to fall near the optimal frontier, indicating that they tend to dominate other systems built from the same collection of categories.

Figures 4A and 4B are based on partitions of the full family tree in Figure 2, but Figure 4C shows results for analyses that focus separately on grandparents, grandchildren, siblings, mother and aunts, father and uncles, and children and niblings (nieces and nephews). Attested systems are again shown in black, and the size of each black circle indicates its frequency. The results are consistent with the near-optimal pattern observed for the entire family tree. The results also support a related prediction of our account: that frequent systems should tend to lie closer to the optimal frontier than rare systems.

Our analyses so far have tested the near-optimality claim relative to large spaces of possible competitors. We now test this claim relative to smaller sets of competitors that might be expected to perform especially well: simple transformations of attested systems (*6*). If attested systems are near-optimal, then each attested system should tend to dominate transformations of that system. The transformations that we consider are based on permutations defined over five "chunks" of the family tree, each comprising four individuals: grandparents, grandchildren, siblings, maternal siblings, and paternal siblings. These chunks are shown in Figure 5A. For example, if we permute the system in Figure 2B by exchanging the category labels of the grandparents chunk with those of the grandchildren chunk, then Alice will use one term for her maternal grandparents, a second term for her paternal grandparents, and four distinct terms for her grandchildren. We considered all such permutations that respected category boundaries: that is, permutations that moved entire categories and did not move only parts of categories. Figure 5B summarizes the results when the full set of permutations is applied to the attested systems in the Murdock data set. In most cases attested systems dominate permutations of these systems, suggesting that attested systems tend to be near-optimal with respect to a focused set of related systems, not just the full space.

Previous researchers have described constraints that help to predict which kin classification systems are encountered in practice (*2, 11, 15, 35–37, 41–50*), and we now show that some of these constraints emerge as consequences of our account. Greenberg (*13, 24*) focuses on markedness constraints, including the constraint that near relatives (e.g. siblings) are more likely than distant relatives (e.g. parent's siblings) to be split into multiple categories, and the constraint that ascending generations (e.g. grandparents) are more likely to be split than descending generations (e.g. grandchildren) (*13*). As Greenberg (*24*) and others (*51, 52*) have argued, markedness constraints can often be usefully formulated in terms of probabilities, and some of Greenberg's specific constraints can be explained as a consequence of the non-uniform distribution of need probabilities over the tree in Figure 3C. In particular, need probabilities are higher for near relatives than distant relatives, and higher for ascending generations than descending generations. To demonstrate that our theory is sensitive to Greenberg's constraints, we show results for three specific permutations (Figure 5C). The first and second permutations exchange near relatives (siblings) with distant relatives (maternal or paternal siblings), and the third exchanges an ascending generation (grandparents) with a descending generation (grandchildren). In all three cases, attested systems tend to score better than permuted systems, illustrating that violations of Greenberg's constraints are penalized by our account.

Another commonly-invoked constraint is that kinship categories tend to have conjunctive definitions (e.g. parent AND female) rather than disjunctive definitions (e.g. parent OR female) (*11–13*). Because conjunctive definitions lead to smaller categories and disjunctive definitions lead to broader categories, our communicative cost measure predicts that conjunctive definitions will tend to be preferred. To test this prediction we developed a ranking measure that reflects the optimality of individual categories (*25*). By this measure, smaller ranks are better: Categories that belong to systems on the optimal frontier have ranks of zero, and categories that belong only to systems that lie far from the optimal frontier have large ranks. We computed the rank of every category that belongs to one of the hypothetical systems shown as light gray circles in the subtree analyses of Figure 4C. Figure 5D shows that for each subtree, conjunctive categories have smaller (i.e. better) ranks on average than disjunctive categories. Although our theory tends to penalize disjunctive systems, Figure S9 shows that the near-optimality results in Figure 4 are driven by more than just this fact.

We have argued that kin classification systems in the world's languages exhibit a near-optimal tradeoff between simplicity and informativeness. We have also argued that the tradeoff between these general principles accounts for some specific constraints on kin classification systems. Even so, kin classification is clearly shaped by factors that go beyond the account presented here. For example, kin classification systems tend to correlate with, and are presumably shaped by, local social patterns of marriage and residence (*2*). Our account places substantial constraints on kin classification systems, explaining several previously documented constraints, and we propose that social forces supply further constraints, such that kin classification systems are both near-optimal with respect to general communicative constraints and well-suited for local social purposes.

Although our theory predicts that kin classification systems achieve a near-optimal tradeoff between simplicity and informativeness, it does not capture the process of cultural evolution (*41, 53, 54*) that presumably led to this result. Previous researchers have developed models of cultural evolution that may help to explain how near-optimal kin classification systems emerge as a consequence of selective pressure for simplicity and informativeness (*55–57*). Studying the evolution of kin classification systems may reveal additional constraints on attested systems—for example, there may be systems that are near-optimal according to our analysis but unattested because they are not the outcomes of plausible evolutionary sequences (*41*).

We have relied here on kinship-specific realizations of the principles of simplicity and informativeness. Appropriate realizations of the same general principles may apply to semantic domains other than kinship, and some evidence suggests that they do. It has been proposed that color naming systems in the world's languages reflect partitions of perceptual color space that are near-optimally informative (*58*), and recent analyses support this view (*6*), including an analysis of lightness terms (*59*) that relies on a variant of the communication game in Figure 1. A similar analysis of color terms should be possible within our framework, where communication is considered successful to the extent that the color inferred by the hearer is close in perceptual space to that intended by the speaker. The domains of kinship and color are different in fundamental respects: Kin terms describe relations between discrete individuals, whereas color terms pick out regions of a continuous perceptual space. The fact that the same general principles help to explain semantic category systems in such dissimilar domains opens up the possibility of a domain-general foundation for categorization across cultures.

References

1. A. A. Goldenweiser, The principle of limited possibilities in the development of culture, *The Journal of American Folklore* **26**, 259 (1913).

2. G. P. Murdock, *Social structure* (Macmillan, New York, 1949).

3. D. E. Brown, *Human universals* (McGraw Hill, 1991).

4. M. Hauser, The possibility of impossible cultures, *Nature* **460**, 190 (2009).

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

6. T. Regier, P. Kay, N. Khetarpal, Color naming reflects optimal partitions of color space, *Proceedings of the National Academy of Sciences* **104**, 1436 (2007).

7. B. Berlin, *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies* (Princeton University Press, Princeton, 1992).

8. S. Atran, Classifying nature across cultures, *Thinking: An invitation to cognitive science*, E. E. Smith, D. N. Osherson, eds. (MIT Press, Cambridge, MA, 1995), vol. 3, pp. 131–174.

9. S. Levinson, S. Meira, The Language and Cognition Group, 'Natural concepts' in the spatial topological domain—adpositional meanings in crosslinguistic perspective: An exercise in semantic typology, *Language* pp. 485–516 (2003).

10. N. Khetarpal, A. Majid, T. Regier, Spatial terms reflect near-optimal spatial categories, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, N. Taatgen, H. van Rijn, eds. (Cognitive Science Society, Austin, TX, 2009), pp. 2396–2401.

11. S. B. Nerlove, A. K. Romney, Sibling terminology and cross-sex behavior, *American Anthropologist* **69**, 179 (1967).

12. D. B. Kronenfeld, Sibling typology: Beyond Nerlove and Romney, *American Ethnologist* **1**, 489 (1974).

13. J. Greenberg, Universals of kinship terminology: their nature and the problem of their explanation, *On language: Selected writings of Joseph Greenberg*, K. Denning, S. Demmer, eds. (Stanford University Press, Stanford, CA, 1990), pp. 310–327.

14. R. E. Valdés-Peréz, V. Pericliev, Computer enumeration of significant implicational universals of kinship terminology, *Cross-Cultural Research* **33**, 162 (1999).

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

16. A. Martinet, *A functional view of language* (Clarendon Press, 1962).

17. P. Grice, *Studies in the way of words* (Harvard University Press, Cambridge, MA, 1989).

18. S. Pinker, P. Bloom, Natural language and natural selection, *Behavioral and Brain Sciences* **13**, 707 (1990).

19. N. Evans, S. C. Levinson, The myth of language universals: Language diversity and its importance for cognitive science, *Behavioral and Brain Sciences* **32**, 429 (2009).

20. G. K. Zipf, *Human behavior and the principle of least effort* (Addison-Wesley Press, Cambridge, MA, 1949).

21. R. Ferrer i Cancho, R. V. Solé, Least effort and the origins of scaling in human language, *Proceedings of the National Academy of Sciences* **100**, 788 (2003).

22. J. Hawkins, ed., *Efficiency and complexity in grammars* (Oxford University Press, Oxford, 2004).

23. E. Rosch, Principles of categorization, *Cognition and categorization*, E. Rosch, B. B. Lloyd, eds. (Lawrence Erlbaum Associates, New York, 1978), pp. 27–48.

24. J. H. Greenberg, ed., *Language universals* (Mouton de Gruyter, The Hague, 1966).

25. Materials and methods are available as supporting material on Science Online.

26. A. L. Kroeber, *California kinship systems* (University of California Press, Berkeley, CA, 1917).

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

28. A. Wierzbicka, *Semantics, culture and cognition: Universal human concepts in culture-specific configurations* (Oxford University Press, New York, 1992).

29. A. L. Kroeber, Classificatory systems of relationship, *Journal of the Royal Anthropological Institute of Great Britain and Ireland* **39**, 77 (1909).

30. J. H. Greenberg, The logical analysis of kinship, *Philosophy of Science* **16**, 58 (1949).

31. S. Gould, ed., *A new system for the formal analysis of kinship* (University Press of America, 2000).

32. A. F. C. Wallace, J. Atkins, The meaning of kinship terms, *American Anthropologist* **62**, 58 (1960).

33. E. Woolford, Universals and rule options in kinship terminology: A synthesis of three formal approaches, *American Ethnologist* **11**, 771 (1984).

34. D. W. Read, C. A. Behrens, KAES: An expert system for the algebraic analysis of kinship terminologies, *Journal of Quantitative Anthropology* **2**, 353 (1990).

35. A. F. C. Wallace, On being just complicated enough, *Proceedings of the National Academy of Science* **47**, 458 (1961).

36. E. J. Hedican, Sibling terminology and information theory: An hypothesis concerning the growth of folk taxonomy, *Ethnology* **25**, 229 (1986).

37. R. G. D'Andrade, Procedures for predicting kinship terminology from features of social organization, *Explorations in Mathematical Anthropology*, P. Kay, ed. (MIT Press, Cambridge, MA, 1971), pp. 60–75.

38. M. Davies, The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. (2008). Available at http://www.americancorpus.org.

39. M. Kupietz, H. Keibel, The Mannheim German reference corpus (DeReKo) as a basis for empirical linguistic research, *Working papers in corpus-based linguistics and language education* (2009), pp. 53–59. Corpus available at http://www.ids-mannheim.de/cosmas2/.

40. H. H. Clark, E. V. Clark, *Psychology and language: An introduction to psycholinguistics* (Harcourt Brace Jovanovich, New York, 1977).

41. P. J. Epling, J. Kirk, J. P. Boyd, Genetic relations of Polynesian sibling terminologies, *American Anthropologist* **75**, 1596 (1973).

42. D. B. Kronenfeld, ed., *Plastic glasses and church fathers: Semantic extension from the ethnoscience tradition* (Oxford University Press, New York, NY, 1996).

43. P. Hage, Unthinkable categories and the fundamental laws of kinship, *American Ethnologist* **24**, 652 (1997).

44. M. Godelier, T. R. Trautmann, F. E. Tjon Sie Fat, Introduction, *Transformations of kinship*, M. Godelier, T. R. Trautmann, F. E. Tjon Sie Fat, eds. (Smithsonian, 1998), pp. 1–26.

45. F. K. Lehman, Aspects of a formalist theory of kinship: The functional basis of its genealogical roots and some extensions in generalized alliance theory, *Anthropological Theory* **1**, 212 (2001).

46. M. de l'Etang, P. J. Bancel, The global distribution of (P)APA and (T)ATA and their original meaning, *Mother Tongue* **IX**, 133 (2005).

47. D. B. Kronenfeld, Issues in the classification of kinship terminologies: Toward a new typology, *Anthropos* **101**, 203 (2006).

48. M. J. Leaf, Experimental-formal analysis of kinship, *Ethnology* **45**, 305 (2006).

49. D. W. Read, Kinship theory: A paradigm shift, *Ethnology* **46**, 329 (2007).

50. B. Milicic, Is there a kinship module? Evidence from children's acquisition of kinship terms in Pitumarca, Peru, *Kinship, language and prehistory: Per Hage and the renaissance in kinship studies*, D. Jones, B. Milicic, eds. (University of Utah Press, Salt Lake City, 2010).

51. M. Haspelmath, Against markedness (and what to replace it with), *Journal of Linguistics* **42**, 25 (2006).

52. E. Hume, Markedness and the language user, *Phonological studies* **11** (2008).

53. V. Pericliev, *Profiling language families by their kin term patterns: A computational approach* (LINCOM Etymological Studies 02, 2011).

54. F. Jordan, A phylogenetic analysis of the evolution of Austronesian sibling terminologies, *Human biology* **83**, 297 (2011).

55. L. Steels, T. Belpaeme, Coordinating perceptually grounded categories through language: A case study for colour, *Behavioral and Brain Sciences* **28**, 469 (2005).

56. T. L. Griffiths, M. L. Kalish, Language evolution by iterated learning with Bayesian agents, *Cognitive Science* **31**, 441 (2007).

57. T. C. Scott-Phillips, S. Kirby, Language evolution in the laboratory, *Trends in Cognitive Sciences* **14**, 411 (2010).

58. K. A. Jameson, R. G. D'Andrade, It's not really red, green, yellow, blue: An inquiry into perceptual color space, *Color categories in thought and language*, C. L. Hardin, L. Maffi, eds. (Cambridge University Press, 2010), pp. 295–319.

59. R. Baddeley, D. Attewell, The relationship between language and the environment: Information theory shows why we have only three lightness terms, *Psychological Science* **20**, 1100 (2009).

60. Code and data are available at `www.charleskemp.com/kinship`. We thank Susanne Gahl, Andrew Garrett, Marcel Just, Andrew Kemp, Brian MacWhinney, Yang Xu, Yuting Zhang, two anonymous reviewers, and especially Paul Kay for valuable suggestions. This work was supported by the National Science Foundation under awards CDI-0835797 and SBE-0541957 and by the Pittsburgh Life Sciences Greenhouse Opportunity Fund.

Supporting Online Material

www.sciencemag.org

Materials and Methods

References

Figs. S1,S2,S3,S4,S5,S6,S7,S8,S9

Tables S1,S2,S3,S4,S5,S6

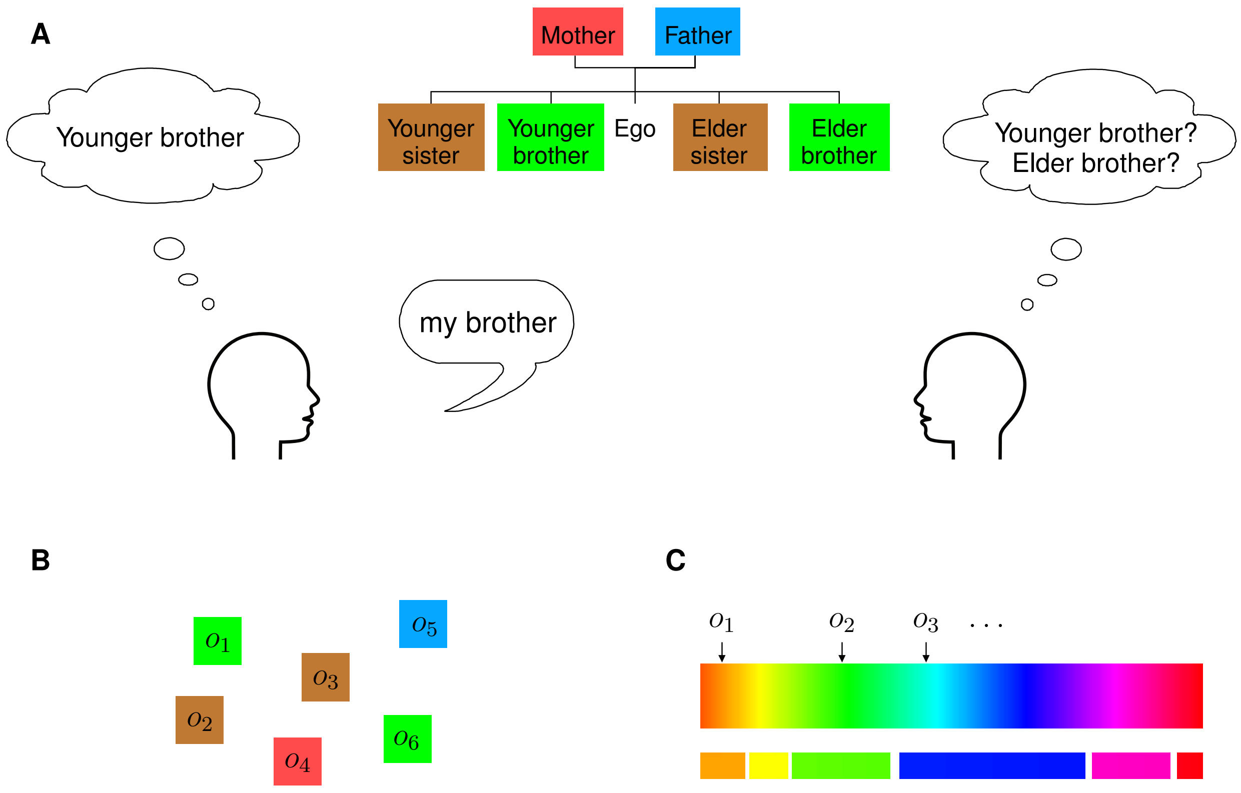
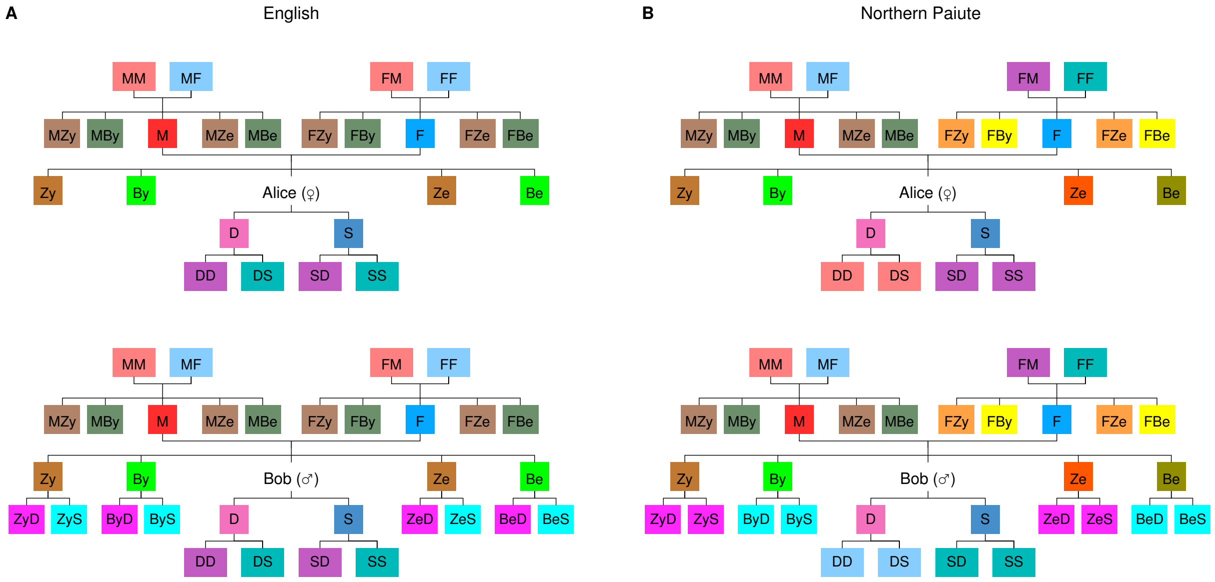


Figure 1: Communication games can be used to study the tradeoff between simplicity and informativeness. (**A**) A communication game for the kinship domain. The objects in the domain are six relatives of the speaker, who is labeled as "Ego" in the family tree. The speaker and hearer have both learned category labels for all of the relatives and the colors shown here represent English category labels. For example, a single term "brother" is used for both younger and elder brother. The speaker uses a category label to refer to one of her relatives and the hearer must infer which relative she has in mind. (**B**) A similar communication game can be formulated within a generic domain that has no structure except for the fact that objects are grouped into categories (*21*). The colors again represent category labels. (**C**) A similar communication game can be formulated within the domain of color, where the "objects" are now points in a continuous perceptual space, shown here for simplicity as a one-dimensional spectrum. The colored bars below the spectrum indicate ranges of color that belong to the same English color categories.



C

■	$\mathtt{\mathbf{mother}}(x, y)$	$\leftrightarrow \mathtt{PARENT}(x, y) \land \mathtt{FEMALE}(x)$
■	$\mathtt{\mathbf{father}}(x, y)$	$\leftrightarrow \mathtt{PARENT}(x, y) \land \mathtt{MALE}(x)$
■	$\mathtt{\mathbf{daughter}}(x, y)$	$\leftrightarrow \mathtt{CHILD}(x, y) \land \mathtt{FEMALE}(x)$
■	$\mathtt{\mathbf{son}}(x, y)$	$\leftrightarrow \mathtt{CHILD}(x, y) \land \mathtt{MALE}(x)$
■	$\mathtt{\mathbf{sister}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{daughter}(x, z) \land \mathtt{PARENT}(z, y)$
■	$\mathtt{\mathbf{brother}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{son}(x, z) \land \mathtt{PARENT}(z, y)$
	$\mathtt{sibling}(x, y)$	$\leftrightarrow \exists z\, \mathtt{CHILD}(x, z) \land \mathtt{PARENT}(z, y)$
■	$\mathtt{\mathbf{aunt}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{sister}(x, z) \land \mathtt{PARENT}(z, y)$
■	$\mathtt{\mathbf{uncle}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{brother}(x, z) \land \mathtt{PARENT}(z, y)$
■	$\mathtt{\mathbf{niece}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{daughter}(x, z) \land \mathtt{sibling}(z, y)$
■	$\mathtt{\mathbf{nephew}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{son}(x, z) \land \mathtt{sibling}(z, y)$
■	$\mathtt{\mathbf{grandmother}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{mother}(x, z) \land \mathtt{PARENT}(z, y)$
■	$\mathtt{\mathbf{grandfather}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{father}(x, z) \land \mathtt{PARENT}(z, y)$
■	$\mathtt{\mathbf{granddaughter}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{daughter}(x, z) \land \mathtt{CHILD}(z, y)$
■	$\mathtt{\mathbf{grandson}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{son}(x, z) \land \mathtt{CHILD}(z, y)$

D

■	$\mathtt{\mathbf{mother}}(x, y)$	$\leftrightarrow \mathtt{PARENT}(x, y) \land \mathtt{FEMALE}(x)$
■	$\mathtt{\mathbf{father}}(x, y)$	$\leftrightarrow \mathtt{PARENT}(x, y) \land \mathtt{MALE}(x)$
■	$\mathtt{\mathbf{daughter}}(x, y)$	$\leftrightarrow \mathtt{CHILD}(x, y) \land \mathtt{FEMALE}(x)$
■	$\mathtt{\mathbf{son}}(x, y)$	$\leftrightarrow \mathtt{CHILD}(x, y) \land \mathtt{MALE}(x)$
	$\mathtt{sister}(x, y)$	$\leftrightarrow \exists z\, \mathtt{daughter}(x, z) \land \mathtt{PARENT}(z, y)$
	$\mathtt{brother}(x, y)$	$\leftrightarrow \exists z\, \mathtt{son}(x, z) \land \mathtt{PARENT}(z, y)$
■	$\mathtt{\mathbf{youngersister}}(x, y)$	$\leftrightarrow \mathtt{sister}(x, y) \land \mathtt{YOUNGER}(x, y)$
■	$\mathtt{\mathbf{oldersister}}(x, y)$	$\leftrightarrow \mathtt{sister}(x, y) \land \mathtt{OLDER}(x, y)$
■	$\mathtt{\mathbf{youngerbrother}}(x, y)$	$\leftrightarrow \mathtt{brother}(x, y) \land \mathtt{YOUNGER}(x, y)$
■	$\mathtt{\mathbf{olderbrother}}(x, y)$	$\leftrightarrow \mathtt{brother}(x, y) \land \mathtt{OLDER}(x, y)$
■	$\mathtt{\mathbf{maternalaunt}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{sister}(x, z) \land \mathtt{mother}(z, y)$
■	$\mathtt{\mathbf{maternaluncle}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{brother}(x, z) \land \mathtt{mother}(z, y)$
■	$\mathtt{\mathbf{paternalaunt}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{sister}(x, z) \land \mathtt{father}(z, y)$
■	$\mathtt{\mathbf{paternaluncle}}(x, y)$	$\leftrightarrow \exists z\, \mathtt{brother}(x, z) \land \mathtt{father}(z, y)$
■	$\mathtt{\mathbf{mansisterchild}}(x, y)$	$\leftrightarrow \mathtt{maternaluncle}(y, x)$
■	$\mathtt{\mathbf{manbrotherchild}}(x, y)$	$\leftrightarrow \mathtt{paternaluncle}(y, x)$
	$\mathtt{maternalgrandmother}(x, y)$	$\leftrightarrow \exists z\, \mathtt{mother}(x, z) \land \mathtt{mother}(z, y)$
	$\mathtt{maternalgrandfather}(x, y)$	$\leftrightarrow \exists z\, \mathtt{father}(x, z) \land \mathtt{mother}(z, y)$
	$\mathtt{paternalgrandmother}(x, y)$	$\leftrightarrow \exists z\, \mathtt{mother}(x, z) \land \mathtt{father}(z, y)$
	$\mathtt{paternalgrandfather}(x, y)$	$\leftrightarrow \exists z\, \mathtt{father}(x, z) \land \mathtt{father}(z, y)$
■	$\mathtt{\mathbf{selfreciprocal1}}(x, y)$	$\leftrightarrow \mathtt{maternalgrandmother}^{\leftrightarrow}(x, y)$
■	$\mathtt{\mathbf{selfreciprocal2}}(x, y)$	$\leftrightarrow \mathtt{maternalgrandfather}^{\leftrightarrow}(x, y)$
■	$\mathtt{\mathbf{selfreciprocal3}}(x, y)$	$\leftrightarrow \mathtt{paternalgrandmother}^{\leftrightarrow}(x, y)$
■	$\mathtt{\mathbf{selfreciprocal4}}(x, y)$	$\leftrightarrow \mathtt{paternalgrandfather}^{\leftrightarrow}(x, y)$

Figure 2: Kin classification systems and their shortest descriptions in the representation language that we use. (**A**) The English kin classification system. Individuals with the same color (e.g. Alice's mother's mother (MM) and father's mother (FM)) are assigned to the same category (in this case "grandmother"). (**B**) The Northern Paiute system. Note that Alice and Bob use different kinship terms for their grandchildren. (**C**, **D**) The shortest descriptions of the English and Northern Paiute systems in the representation language that we use.

A

FEMALE(·)
MALE(·)
PARENT(·, ·)
CHILD(·, ·)
OLDER(·, ·)
YOUNGER(·, ·)
SAMESEX(·, ·)
DIFFSEX(·, ·)

B

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}(x,y) \wedge \mathtt{B}(x)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}(x,y) \wedge \mathtt{B}(y)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}(x,y) \wedge \mathtt{B}(x,y)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}(x,y) \vee \mathtt{B}(x)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}(x,y) \vee \mathtt{B}(y)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}(x,y) \vee \mathtt{B}(x,y)$

$\mathtt{C}(x,y) \leftrightarrow \exists z\, \mathtt{A}(x,z) \wedge \mathtt{B}(z,y)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}(y,x)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}^{\leftrightarrow}(x,y)$

$\mathtt{C}(x,y) \leftrightarrow \mathtt{A}^{+}(x,y)$

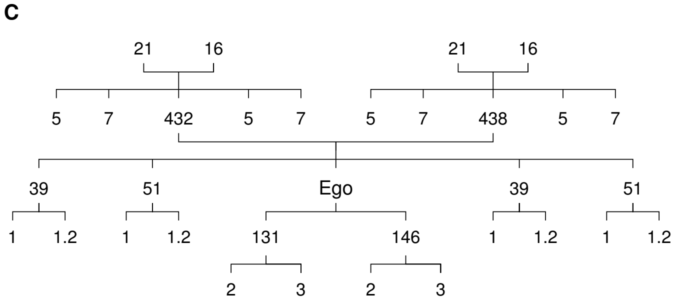


Figure 3: Components used to formalize the notions of cognitive complexity and communicative cost. (**A**) Primitive concepts. (**B**) Rules for combining these concepts. Each rule allows a new concept C() to be defined in terms of at most two concepts A() and B() which must be either primitive or previously defined. (**C**) Need probabilities for individuals in the family trees of Figure 2. The actual probabilities are derived from English and German corpus statistics and are proportional to the numbers shown here.

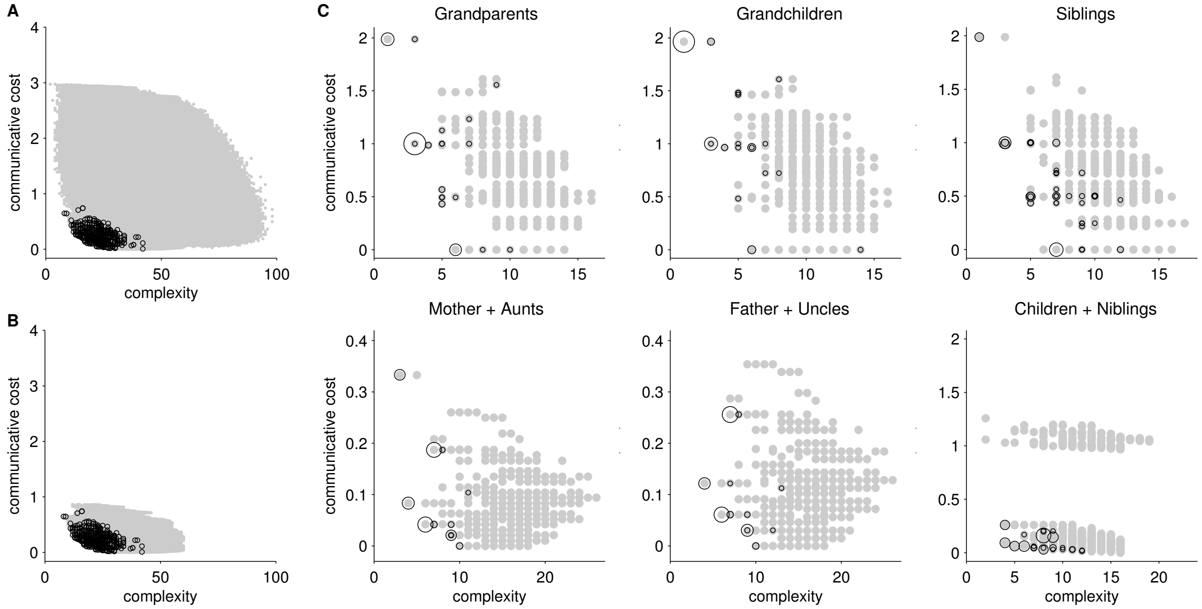


Figure 4: The optimal frontier. (**A**) Communicative cost versus complexity for a large space of possible kin classification systems. Attested systems are shown as black circles. (**B**) Communicative cost versus complexity for systems built from attested categories that appear more than twice in the Murdock data. (**C**) Optimality analyses for six subsets of the full family tree in Figure 2. In each plot the black circles represent real-world systems and the sizes of these circles represent frequencies within the Murdock data set.

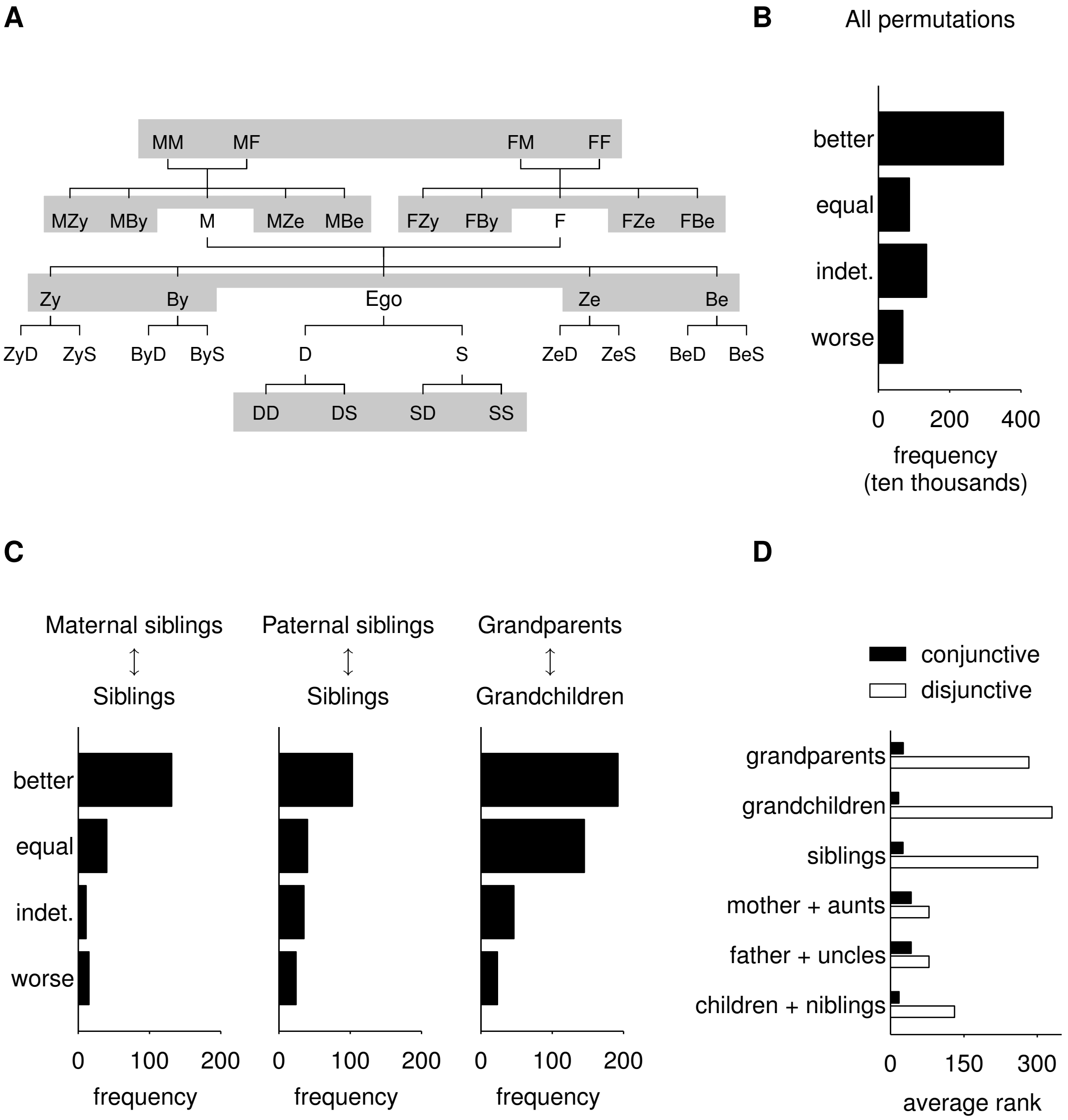


Figure 5: Fine-grained optimality analyses. (**A**) The gray bars indicate the five chunks used for the permutation analysis. (**B**) Results of the permutation analysis. Attested systems typically score better than permuted versions of these systems and rarely score worse. In some cases the attested and permuted systems are equal in both cost and complexity, and in others the attested system is superior along one dimension but inferior along the other and the comparison is indeterminate. (**C**) Results for three specific permutations that exchange near relatives (siblings) with more distant relatives (maternal and paternal siblings), and that exchange ascending relatives (grandparents) with descending relatives (grandchildren). Attested systems dominate those that permute near and distant relatives, or ascending and descending relatives, which explains specific markedness constraints proposed previously. (**D**) Comparison between conjunctive and disjunctive categories. Conjunctive categories contribute to systems of greater optimality, which may help to explain the cross-cultural predominance of conjunctive categories.