Kinship terminologies reflect culture-specific communicative need: Evidence from Hindi and English

Gunjan Anand (gunj@berkeley.edu)
Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics, UC Berkeley
Berkeley, CA 94720 USA

Abstract

Systems of semantic categories vary across languages, and it has been proposed that this variation is constrained by a need for efficiency in communication. An important element of efficiency is communicative need, or how often a particular object needs to be referenced. Previous work has sometimes assumed for simplicity that the distribution of need over objects in a semantic domain does not vary across languages or cultures. Here, we explore culture-specific need as it relates to the kinship terminologies of Hindi and English. We assess the efficiency of each language's kin naming system under a variety of need distributions, including one based on that language's usage statistics, one based on the other language's usage statistics, and random permutations of each of those two distributions. Our results suggest that kinship terminologies reflect culture-specific communicative need.

Keywords: kinship; semantic categories; communicative need; efficient communication; language and culture

Introduction

Systems of semantic categories exhibit wide yet constrained variation across languages, and it has been argued that this pattern reflects a drive for efficient communication. On this view, languages generally, and semantic systems in particular, are under functional pressure to be both simple and informative, and the different semantic systems that are found in the world's languages represent different ways to efficiently trade these two desiderata off against each other (e.g. Zipf, 1949; Rosch, 1978; Haspelmath, 1999; Kemp et al., 2018). A maximally simple semantic system for a given domain (e.g. color, or number, or kinship) will have a single category covering the entire domain, and as a result the system as a whole will not be informative: knowing the category name will give no information about which specific object within the domain is intended. In a maximally informative system, in contrast, there will be a separate name for each object, so knowing an object's name will identify the referent precisely – but this will require many categories, and such a system will therefore be complex, not simple. An efficient way to navigate this tension is to have small, precise categories only in parts of semantic space with high *communicative need* – that is, parts of space that frequently need to be referenced – and fewer, broader categories elsewhere. That way, the system will support informative communication most of the time, at the price of only modest complexity. On this view, one should expect to find narrow, informative categories in high-need (frequently referenced) parts of semantic space.

Some previous studies (e.g. Kemp & Regier, 2012; Xu et al., 2016, 2020) have made the simplifying assumption that the distribution of communicative need over objects in a domain will be the same distribution for different languages. This assumption is grounded in the expectation that there are universal aspects of communicative need, reflecting universal tendencies of human thought and interest. However, it seems likely that there will also be culturally-specific aspects of communicative need. The reasoning laid out above predicts that if need does vary across cultures, that should produce corresponding variation in the semantic systems of languages spoken in those cultures. This is an idea with both a long history (Boas, 1911; Whorf, 1956) and some recent evidence supporting it (e.g. Regier et al., 2016; Gibson et al., 2017; Winter et al., 2018; Twomey et al., 2021).

Here, we consider this idea specifically in the domain of kinship. This domain suggests itself for two reasons. First, it has been claimed that kin terminologies correlate with characteristics of local social structure (e.g. Murdock, 1949; D'Andrade, 1971, but see also Guillon & Mace, 2016; Passmore & Jordan, 2020). When such correlations do exist, they could plausibly be mediated by culture-specific communicative need. Second, kinship is a semantic domain that has been analyzed in terms of efficiency (Kemp & Regier, 2012), but that earlier work provisionally assumed a universal need distribution. Here, we explore the efficiency of kinship terminologies under culture-specific need.

In what follows, we first briefly summarize the study of Kemp and Regier (2012), on which we build. We then present our own study, which compares kin naming in Hindi and English, in light of need distributions derived from corpora of those two languages, and variants of those two need distributions. To preview our results, we find that the Hindi and English kin terminologies are more efficient when assessed under their own native need distributions than under the need distribution of the other language, or under most random permutations of either language's need distribution. We conclude that each language's kin naming system appears to reflect culturally-specific patterns of communicative need.

The study of Kemp and Regier (2012)

Kemp and Regier (2012) proposed that kinship terminologies across languages achieve a near-optimal tradeoff between informativeness and simplicity. For a given kin type *i* (an object

in the domain, e.g. one's mother's older brother), they used the notation p_i to denote the probability of needing to refer to that kin type, and they called this measure of communicative need the *need probability* of that kin type. They also used the notation z_i to denote the category name used to refer to that kin type in a given kin terminology, e.g. *uncle* in English. They then defined the communicative cost of referring to a specific kin type i using kin term z_i as the surprisal associated with kin type i given kin term z_i:

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

Here, j ranges over those kin types that have the same kin term as i (e.g. over all *uncle*s), so the fraction inside the log is the probability that the speaker intended kin type i rather than some other kin type with the same kin term as i. They then took the communicative cost for the kin terminology (naming system) as a whole to be:

$$C = \sum_{i=1}^{n} p_i c_i \quad (2)$$

where n is the number of kin types in the domain – this is the expected cost over all kin types, weighted by need. They took a kin terminology to be informative to the extent that it exhibits low communicative cost C. They defined the complexity of a system to be the length of its description in a representation language based on primitives such as `PARENT( )` and `FEMALE( )`, and took a kin terminology to be simple to the extent that it exhibits low complexity. Finally, they took a kin terminology to be efficient to the extent that it is as informative as possible for its level of simplicity, and as simple as possible for its level of informativeness — i.e. to the extent that it lies along the Pareto frontier defined by these two dimensions. They calculated communicative cost and complexity for the kinship terminologies of 487 languages from a dataset compiled by Murdock (1970), and for a large set of hypothetical kinship systems intended to cover most of the space of possible systems. They found that the systems in Murdock's data lay near the Pareto frontier of possible systems, as shown in Figure 1, and thus tended to be efficient.

Kemp and Regier (2012) based their need distribution p_i on corpus frequencies for kin terms in English and German (see also Rácz et al., 2019 for a study of kin term frequency across a larger set of Indo-European languages). The need distributions for those two languages were found to be similar to each other, and were combined to yield a presumptively universal need distribution. As noted above, there are reasons to expect universal tendencies in need, and Kemp and Regier (2012)'s assumption of universality in need helped to make sense of the cross-language variation in attested kin terminologies, as seen in Figure 1. Their presumptively universal need distribution also helped to explain specific markedness constraints for kinship proposed by Greenberg (1990), in terms compatible with Greenberg's own reasoning. At the same time, Kemp and Regier (2012) noted that need distributions could in principle vary across cultures, reflecting the fact that "different cultures impose different communicative requirements" (p. 1051) — and that naming systems could vary accordingly.

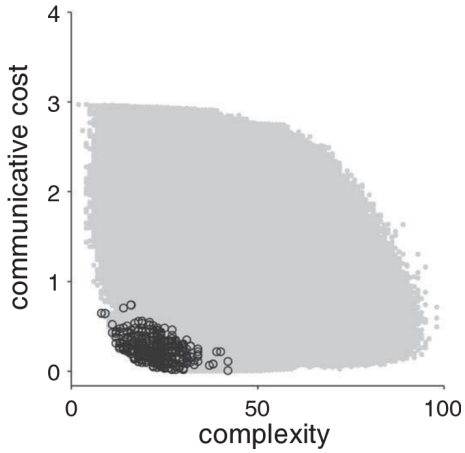

Figure 1: Communicative cost vs. complexity for a large set of hypothetical kinship systems (gray) and attested systems (black). Attested systems lie near the Pareto frontier. From Kemp and Regier (2012).

Methods

We wished to test this idea. To do so, we considered the kinship terminologies of Hindi and English, illustrated in Figure 2. Hindi was chosen because it is the native language of the first author, and English is convenient as a point of comparison. The Hindi naming system was contributed by the first author and double-checked against Shapiro (1989), and the English system was adopted from Kemp and Regier (2012). It can be seen that the two systems differ, and we draw attention to two points in particular. First, the Hindi system is finer-grained than that of English. For example, Hindi has distinct terms for the mother's mother vs. the father's mother, whereas English has a single term, *grandmother*. Similarly, Hindi has distinct terms for a younger sister vs. an older sister, whereas English has a single term, *sister*. Second, the Hindi system is also asymmetric in a way that the English system is not: Hindi has separate kin terms for the father's elder vs. younger brother, but has the same kin term for the mother's elder vs. younger brother. We wished to assess whether these differences in naming could be explained by differences in communicative need across the two languages.

We reasoned as follows. The efficiency view predicts finer-grained categories for regions of higher communicative need. This leads to two separate predictions, one at the level of the semantic domain of kinship as a whole, and the other at the level of specific objects (kin types) within this domain.

- At the level of the semantic domain of kinship taken as a whole, and considered as a subpart of the lexicon, the finer semantic grain of the Hindi kinship terminology predicts

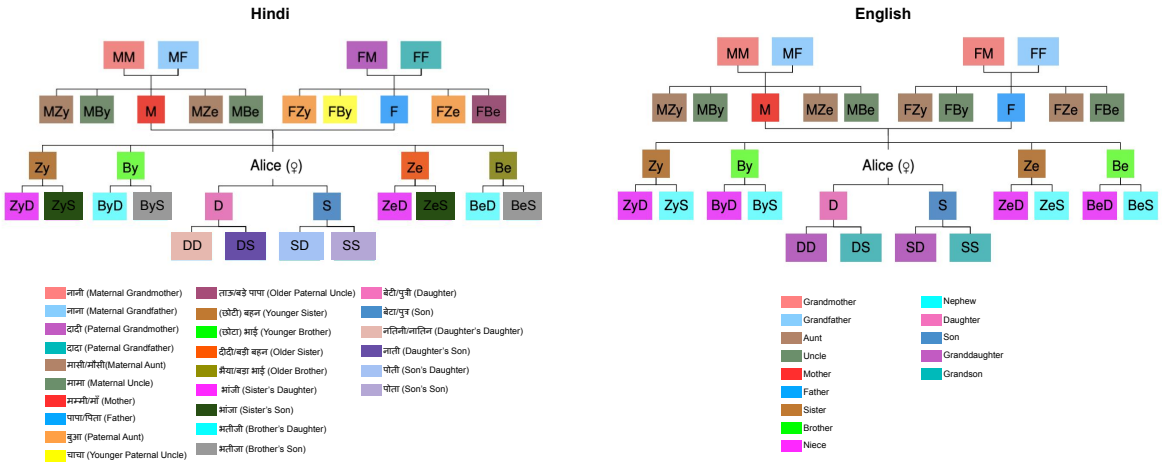

Figure 2: Kinship terminologies of Hindi (left) and English (right), in both cases relative to a woman named Alice; each language's kin terminology is unaffected by the gender of ego. Each node in the tree is a kin type; these are designated by combinations of M (mother), F (father), D (daughter), S (son), Z (sister), B (brother), e (elder), and y (younger), such that for example MBy designates the mother's younger brother. Colors denote the extension of kin terms, as shown in each legend.

that kin terms will constitute a greater proportion of Hindi language use than they will of English language use.

- At the level of specific kin types within the kinship domain, if naming reflects culture-specific need distributions over the family tree, we would expect each language's (Hindi, English) kinship terminology to "fit" a need distribution derived from the statistics of that language's usage. That is, a language's kinship terminology should be more informative when assessed under a need distribution derived from that language's usage, rather than the other language's usage, and rather than under a wide range of hypothetical need distributions.

Materials

We sought comparable corpora for Hindi and English. There are fewer corpora available for Hindi than for English, so Hindi placed stronger constraints on our search for corpora. We settled on a web-scraped corpus for each language. While this choice was driven in large part by availability, web-scraped corpora also capture a broad range of registers of speech, and therefore seem potentially more appropriate than corpora based solely on relatively formal registers such as news or Wikipedia. Specifically, for Hindi we used the 2015 1-million-sentence web-scraped Hindi corpus[1] from the Leipzig Corpora Collection (Goldhahn, Eckart, & Quasthoff, 2012), and for English we consulted the iWeb corpus, an English web-scraped corpus from 2017 containing 14 billion words.[2]

Procedure

We tokenized the Hindi corpus using code the first author wrote for this purpose. The English corpus was already tokenized. We then counted the number of occurrences of each kin term in each language.[3] On this basis we computed, for each language, the proportion of all word tokens in the corpus that were kin terms – this is a measure of the prominence of kin terms in that language's usage. We also computed, for each language, a need probability distribution p over kin types, as follows. We first calculated a need probability distribution over kin *terms*, based on relative frequency of kin terms, i.e. the frequency of each kin term in the language divided by the summed frequency of all kin terms in the language. Then, in those cases in which a kin term is used to name more than one kin type (e.g. English *grandmother* names both maternal and paternal grandmothers), we distributed the probability mass corresponding to that kin term uniformly over the kin types it names (Kemp & Regier, 2012, cf. Zaslavsky et al., 2019). We also produced 10,000 hypothetical variants of each language's need distribution p by randomly permuting the need probabilities p_i over kin types.

Finally, we calculated the communicative cost C (Equation 2) for each of the two languages (Hindi, English) under four conditions: (1) using that language's need distribution; (2) using the hypothetical variants of that language's need distribution; (3) using the other language's need distribution; (4) using the hypothetical variants of the other language's need distribution.

Results

Need probability

We found that the proportion of word tokens that are kin terms is greater in Hindi (2312 occurrences of kin terms per million

[1] https://wortschatz.uni-leipzig.de/en/download/Hindi

[2] https://www.english-corpora.org/iweb/

[3] Kemp and Regier (2012) searched for "my <kinterm>" or the equivalent. We deviated from their method because in Hindi the first person possessive determiner is often omitted in such contexts.

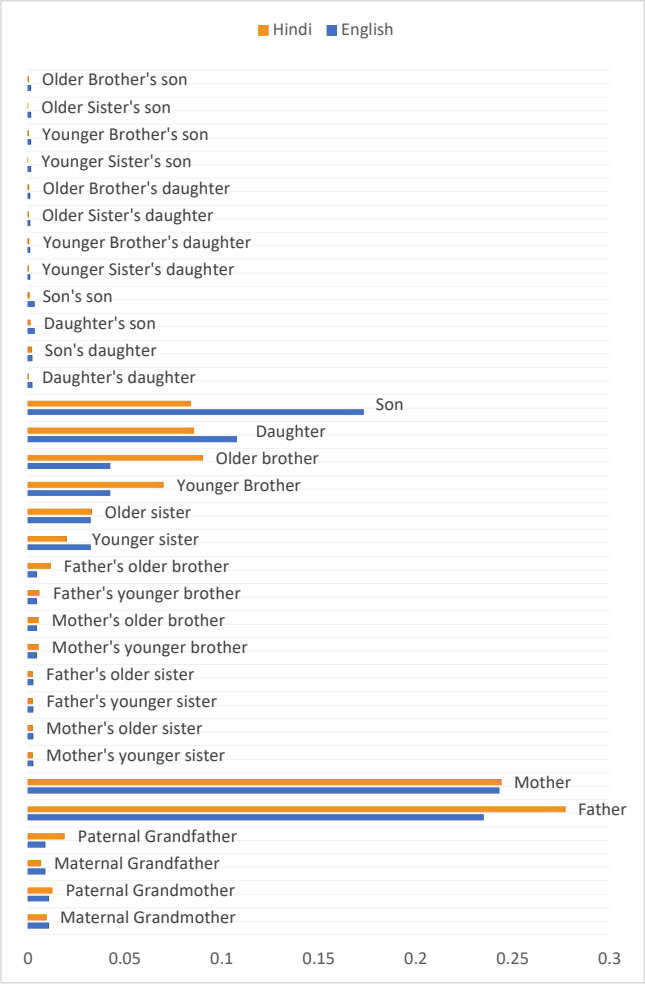

Figure 3: Need probabilities for kin types in Hindi and English, derived from Hindi and English web-scraped corpora.

words) than in English (654 occurrences of kin terms per million words). This confirms the prediction, based on the finer grain of the Hindi kin terminology, that Hindi usage will exhibit a greater tendency to communicate about kinship generally, compared to English.

Figure 3 shows the need probability distributions we obtained for Hindi and English. At a general level, it can be seen that the two distributions over kin types are similar ($R^2 = .92$, $F(1,31) = 352.4$, p .001), in line with the assumption of culturally shared aspects of need (see also Zaslavsky et al., 2019; Gao & Regier, 2022). Both need distributions also display documented patterns, such as the tendency for need to be higher for near relatives than for distant relatives, and higher for ascending (older than ego) generations than for descending (younger than ego) generations (Greenberg, 1966; Kemp & Regier, 2012). At the same time, there are also differences between the need probability distributions from the two languages — in line with the idea that there may also be culturally-specific aspects of communicative need. In particular, when compared with English, Hindi exhibits a general

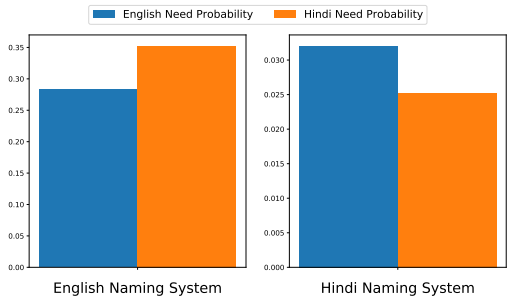

Figure 4: Communicative costs for English (left) and Hindi (right) kin naming systems, assessed under each language's need distribution. Communicative cost is lower, i.e. communication is more informative, under native-language need.

pattern of higher need probability for older male relatives on the father's side, specifically father, paternal grandfather, and father's older brother (Fisher's exact text comparing counts for these three kin types vs. all other kin types, in Hindi vs. English, yielded $p = 0.003$). This last example potentially aligns with the fact, noted above, that Hindi has separate kin terms for the father's older vs. younger brother. Because the father's older brother has relatively high need in Hindi, there would be substantial communicative cost contributed by confusing him with the father's younger brother in communication, as could happen if there were a single term covering the two kin types. This helps to explain the fact that Hindi has two separate terms for these kin types, yet no analogous distinction for the mother's brothers, where no such clear difference in need is seen.[4]

Communicative cost

This apparent partial alignment between cross-language differences in naming and cross-language differences in need suggests that each language's naming system may be especially well-suited to the statistics of usage of that language. Figure 4 confirms that this is the case. This figure shows the communicative cost C for each language when assessed relative to its own need distribution (i.e. when z and p in Equations 1 and 2 are from the same language), and when assessed relative to the other language's need distribution (i.e. when z is from one language and p from the other). It can be seen that each language exhibits lower communicative cost — that is, is more informative — under its own need distribution than under the other language's need distribution. Hindi is substantially more informative than English overall, consistent with the greater complexity (here, greater number of kin terms, yielding finer semantic grain, cf. Kemp & Regier, 2012) of the Hindi naming system. We also calculated the communicative cost for each language's naming system under the various permuted versions (see above, under Procedure) of

[4]The Hindi term for father's younger brother is sometimes used for non-kin, whereas the term for father's older brother is not. Thus, the greater frequency for father's older brother appears despite the use of father's younger brother for non-kin as well as kin.

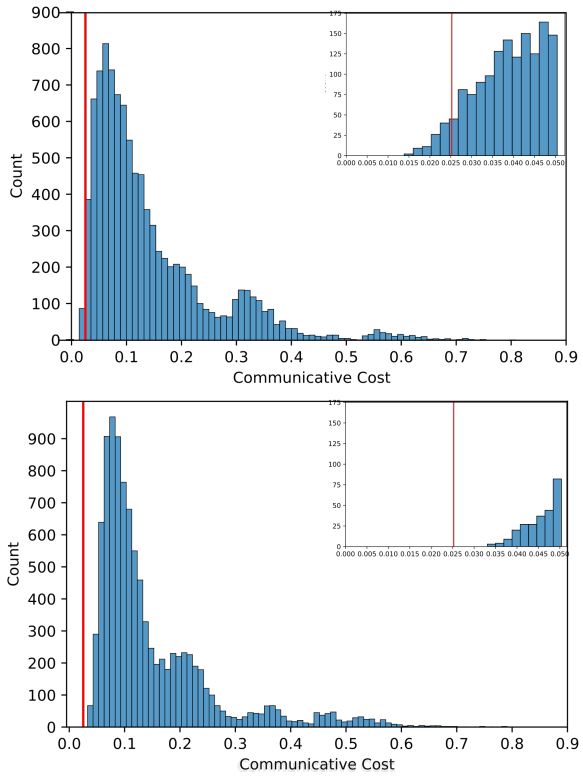

Figure 5: Communicative cost of the Hindi kin naming system, assessed under the native Hindi need probability distribution (red line), under permuted variants of that need distribution (blue histogram in top panel), and under permuted variants of the English need distribution (blue histogram in bottom panel). The insets zoom in to show the lower tail of each distribution. The Hindi naming system exhibits lower communicative cost when assessed under native Hindi need than it does when assessed under 99% of the permuted Hindi need distributions, and under 100% of the permuted English distributions.

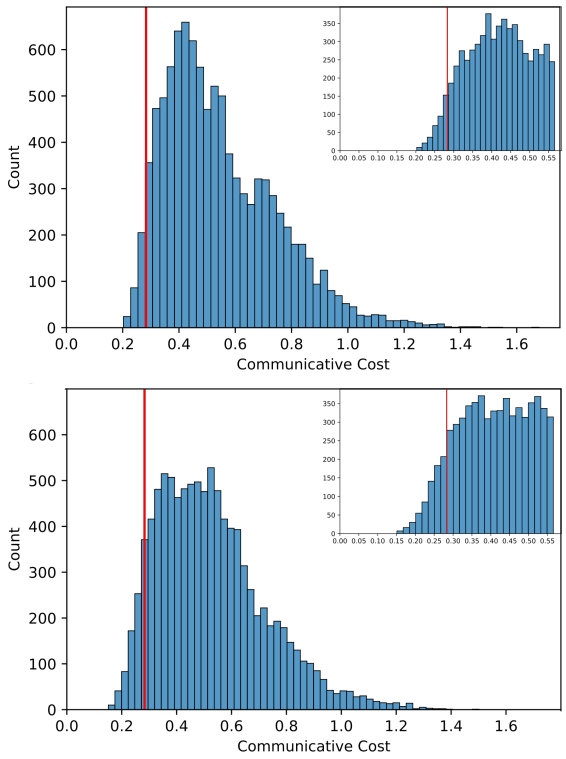

Figure 6: Communicative cost of the English kin naming system, assessed under the native English need probability distribution (red line), under permuted variants of that need distribution (blue histogram in top panel), and under permuted variants of the Hindi need distribution (blue histogram in bottom panel). The insets zoom in to show the lower tail of each distribution. The English naming system exhibits lower communicative cost when assessed under native English need than it does when assessed under 97% of the permuted English need distributions, and under 93% of the permuted Hindi distributions.

each language's need distribution. Figures 5 and 6 show that each language's naming system exhibits lower communicative cost, and therefore more informative communication, under native-language need than under most permutations of either language's need distribution.[5] These findings support the idea that the kin naming systems of these two languages are aligned well with the statistics of usage in the two languages.

Discussion

We have shown: (1) that the Hindi and English kin naming systems differ, (2) that usage statistics concerning kin terms in the two languages also differ, in a way that could in principle explain the naming differences, and (3) that each language's kin naming system is more informative under its own than under the other language's need probability distribution, and more informative than under many hypothetical need distributions derived from those two. We take these findings to support the proposal that systems of semantic categories may be shaped by culture-specific as well as universal aspects of communicative need. These findings and this conclusion raise a number of questions, and directions for future research.

First, there is the question whether these results, based on just two languages in part of a single semantic domain, will generalize to more of the kinship domain, to kinship systems in other languages, and to semantic domains other than kinship. The portion of the kinship domain that we have considered here, shown above in Figure 2, is the portion considered in the primary analyses of Kemp and Regier (2012). In secondary analyses they also considered cousins, which we have not considered here but which have been a major focus of cross-language work on kinship. A natural direction for ex-

[5] We also obtained qualitatively similar results when permuting need over kin *terms* rather than over kin types.

tension of this work would be to test it against cousins as well. We have considered only two languages, Hindi and English, both of them Indo-European. Despite the relatedness of these two languages, we have found evidence for cultural specificity of communicative need, but consideration of a larger and more diverse language sample would help to place our current results in a broader context. Finally, there is the question of other semantic domains. Recent work has made an argument analogous to ours in another semantic domain, that of names for household containers in English and Mandarin (Gao & Regier, 2022), using subjective need probabilities, and closely related ideas have also been recently explored (e.g. Gibson et al., 2017; Winter et al., 2018; Twomey et al., 2021) — but more work is needed to properly assess the generalizability of these ideas across languages and domains.

Efficiency is generally taken to mean an optimal tradeoff between simplicity and informativeness – and in this paper we have treated only informativeness. This allows a targeted analysis, in which simplicity is essentially controlled for: here we have manipulated only the need probability distributions, and these affect informativeness (via Equations 1 and 2), and not simplicity, so the simplicity of these systems is unaffected by our manipulations. A fuller analysis would be possible in future work, in which the Hindi and English naming systems are compared to a full space of hypothetical naming systems, all under varying need distributions.

That prospect in turn leads to an observation. Discussions of efficient communication naturally evoke an optimization process of some sort by which communicative systems become efficient. Given this, it is natural to imagine a space of possible systems like that shown in Figure 1, and to imagine a system's evolution as involving a trajectory within that space, heading toward greater efficiency or remaining near efficiency (e.g. Kemp et al., 2018, p. 113, Figure 3C; Zaslavsky et al., 2022). However, in some circumstances this notion may be incomplete. Specifically, if a system evolves under circumstances of a changing need distribution – e.g. to reflect ongoing social or cultural changes (e.g. Malt et al., 1999) – that change in need would affect the informativeness, and therefore the positions, of *all* systems in the space, meaning that the optimization would take place over a shifting rather than static landscape.

Another question raised by the prospect of a broader analysis concerns the nature of informativeness itself. Here we have assumed, following the work on which we build (Kemp & Regier, 2012), that the informational goal of communication about kin is to convey a kin type in a family tree. This is certainly an important sort of information conveyed by kin terms – but there are also other sorts, such as the social features and social roles commonly associated with specific kin types – features such as kindliness, trustworthiness, authority, and the like – i.e. the social and cultural content, and not just the identity, of that kin type. These are beyond the scope of the analysis we have presented here, but seem relevant to a more general analysis of communication about kin.

Finally, more work is needed to directly probe an idea briefly alluded to in our introduction. It has been claimed that aspects of kin terminologies correlate with characteristics of local social structure (e.g. Murdock, 1949, but see Guillon & Mace, 2016; Passmore & Jordan, 2020), and we have speculated that that when such correlations do occur, they may be mediated by culture-specific communicative need. The results we have presented here seem broadly consistent with that idea, but it is also possible that such correlations, when they exist, are undergirded by more than simple frequency of mention, e.g. by the perceived similarity of the social roles canonically associated with specific kin types in a given culture. Better understanding such mediating variable(s) is a natural goal for future research.

Acknowledgments

We thank Charles Kemp and 3 anonymous reviewers for helpful comments on an earlier version of this paper. Any errors are our own. Author contributions: GA and TR designed the research; GA performed the research; GA analyzed the data; and GA and TR wrote the paper.

References

Boas, F. (1911). Introduction. In *Handbook of American Indian Languages, Vol.1* (p. 1-83). Government Print Office (Smithsonian Institution, Bureau of American Ethnology, Bulletin 40).

D'Andrade, R. G. (1971). Procedures for predicting kinship terminologies from features of social organization. In P. Kay (Ed.), *Explorations in mathematical anthropology* (pp. 60–75). Cambridge, MA: MIT Press.

Gao, S., & Regier, T. (2022). Culture, communicative need, and the efficiency of semantic categories. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.*

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785-10790.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012).*

Greenberg, J. (1966). *Language universals.* The Hague: Mouton de Gruyter.

Greenberg, J. (1990). Universals of kinship terminology: Their nature and the problem of their explanation. In K. Denning & S. Demmer (Eds.), *On language: Selected writings of Joseph Greenberg.* Stanford, CA: Stanford University Press.

Guillon, M., & Mace, R. (2016). A phylogenetic comparative study of Bantu kinship terminology finds limited support for its co-evolution with social organisation. *PLoS ONE*, *11*(3), e0147920.

Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, *18*, 180-205.

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

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1).

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*, 230–262.

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

Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology*, *9*(2), 165–208.

Passmore, S., & Jordan, F. M. (2020). No universals in the cultural evolution of kinship terminology. *Evolutionary Human Sciences*, *2*(e42), 1–14.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS ONE*, *11*(4), e0151138.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (p. 27-48). New York: Lawrence Erlbaum Associates.

Rácz, P., Passmore, S., Sheard, C., & Jordan, F. M. (2019). Usage frequency and lexical class determine the evolution of kinship terms in Indo-european. *Royal Society Open Science*, *6*(10), 191385.

Shapiro, M. (1989). *A primer of modern standard Hindi*. Delhi: Motilal Banarsidass.

Twomey, C. R., Roberts, G., Brainard, D. H., & Plotkin, J. B. (2021). What we talk about when we talk about colors. *Proceedings of the National Academy of Sciences*, *118*(39).

Whorf, B. L. (1956). Science and linguistics. In J. B. Carroll (Ed.), *Language, thought, and reality: Selected writings of Benjamin Lee Whorf* (pp. 207–219). MIT Press.

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, *179*, 213-220.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*, 2081–2094.

Zaslavsky, N., Garvin, K., Kemp, C., Tishby, N., & Regier, T. (2022). The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*.

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2019). Communicative need in color naming. *Cognitive Neuropsychology*.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.