Kinship Categories Across Languages Support General Communicative Principles

Elena Wang (elenazy.wang@mail.utoronto.ca)
University of Toronto, 27 King's College Cir
Toronto, Ontario M5S 1A1 Canada

Nanyi Wang (vannie.wang@mail.utoronto.ca)
University of Toronto, 27 King's College Cir
Toronto, Ontario M5S 1A1 Canada

Abstract

Despite linguistic and cultural differences, the scope of possible variation in kinship categories appears to be constrained. Previous studies mainly focused on investigating some of the domain-specific constraints that shape kin classification. Here, we use computational analyses to explore the kinship systems in 651 languages and to examine the domain-general principles that underlie kinship: simplicity and informativeness. We find that the kin classification systems in languages worldwide achieve a near-optimal trade-off between complexity and informativeness, aligning with existing arguments in other semantic domains, including color and numeral systems. Our findings suggest that the trade-off between simplicity and informativeness, which are also applicable to other semantic domains, may serve as a domain-general foundation for the observed variation in category systems across languages.

Keywords: Kinship; Communicative efficiency; Classification systems; Informativeness; Simplicity; Information Theory; Linguistic relativity

Introduction

Although concepts and categories differ among cultures and languages, they seem to be influenced by some universal constraints (Regier et al., 2005; Kemp & Regier, 2012). Numerous cross-cultural studies have investigated potential constraints in various domains such as color categories (Berlin & Kay, 1969; Regier et al., 2007), animal classification (Atran, 1995), and spatial relations (Majid, 2015). Among these domains, kinship classification systems have traditionally been significant for such studies, and previous research has revealed many constraints to exploring the fundamental logic behind them.

Unlike previous research that typically focused on domain-specific constraints for kin classification systems, Kemp and Regier hypothesized in "Kinship Categories Across Languages Reflect General Communicative Principles" that the major aspects of kinship classification follow directly from two general principles: categories tend to be simple, which reduces cognitive load, and informative, which maximizes communicative efficiency (2012). For a long time, scholars such as Zipf (1949) have contended that languages are shaped by functional demands for effective communication, which entails the need to communicate accurately while minimizing cognitive effort. Zipf showed that word-frequency distributions achieve a trade-off between simplicity and communicative precision, and in recent years, this notion has gained increased attention from researchers across various domains. For instance, Xu et al. (2020) studied numeral systems and found a near-optimal tradeoff between communicative cost and complexity across attested numeral systems, while Regier et al. (2007) studied color categories and suggested that color naming reflects optimal partitions of color space.

Inspired by the paper "Kinship Categories Across Languages Reflect General Communicative Principles" by Kemp and Regier, our current research aims to explore universal constraints by investigating the domain-general principles of simplicity and informativeness using computational methods. In Kemp and Regier's study, they analyzed kinship terms across over 600 languages and found strong cross-linguistic patterns in how kinship terms were structured. Their results of the optimality analyses support the hypothesis that kinship categories are shaped by the domain-general principles of simplicity and informativeness. However, a limitation of their study is that the data they relied upon is outdated and contains some gaps in information. To address this, we sourced a new dataset that includes 1229 languages and has been constantly updated since 2022. In the current study, we employ similar analytical approaches on 651 languages to test our hypothesis, which posits that the world's kin classification systems across different languages achieve an optimal trade-off between informativeness and simplicity. Our primary objective is to investigate the near-optimality of complexity and informativeness in the 651 languages, as well as to explore any cross-linguistic patterns. To accomplish this, we will provide definitions of complexity and informativeness within the kin classification system, followed by the computation of these measures. The results will be subsequently compared with permuted data to determine their statistical significance. Finally, the findings will be explained in detail, and any atypical observations will be discussed.

Materials and Methods

Kemp and Regier employed a sophisticated method using information theory and need probability to calculate complexity and communicative cost. They used these two metrics to graph real kin classification systems with self-generated kin classification systems to measure optimality. We achieved a similar result with our study, using modified methods that fit the scope of this project. The code can be found here: `https://github.com/elenawzy/COG403-final-project`

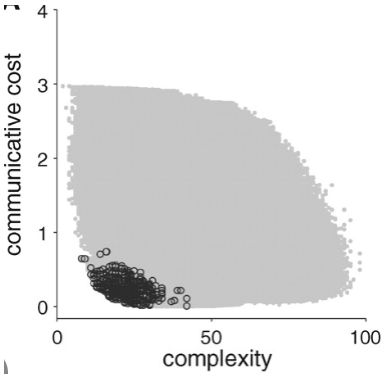

Figure 1: Kemp and Regier's visualization of real languages vs. created languages, where real systems seem optimized

1. Parse the dataset by language. The dataset has a website to visualize the data in a more readable format: `http://www.kinbank.net/`. Kin parameters are the codes that are used to identify specific kin types (ex. "fFyB": female speaker's father's younger brother). Kin terms/words are the names used in each language to refer to a kin parameter (ex. parameter: "fFyB", word (english): "uncle"). Each language in the dataset has a csv file with a column for kin parameter and a column for kin terms/words. A master list of kin parameters was included in the dataset, so we used this to create lists of kin parameters of interest for data parsing.

6	fBW	female ego's brother's wife
7	fD	female ego's daughter
8	fDD	female ego's daugher's daughter
9	fDH	female ego's daughter's husband
10	fDS	female ego's daughter's son
11	fF	female ego's father

	parameter	word
2	meB	ainé
3	mP	parent
4	mC	enfant
5	mA	ancêtre
6	mH	conjoint

Figure 2: A snippet of the master list of parameters (on the left) and the csv for the language French (on the right)

For this study, we focused on a set of 24 kin parameters, which include parents, parents' siblings, grandparents, siblings, children, and grandchildren. We created a list of female ego and male ego parameters to parse through each language's csv. We used pandas to read each language's csv and created dictionaries of each language with keys as kin parameters, and values as the corresponding kin term/word. Some languages did not have all the kin parameters that we were looking for. We filtered these languages out, leaving us with 651 real languages to measure complexity and informativeness.

```
male_ego_parameters = ["mMM", "mMF", "mFM", "mFF",
"mMyZ", "mMyB", "mM", "mMeZ", "mMeB", "mFyZ", "mFyB", "mF", "mFeZ", "mFeB",
"myZ", "myB", "meZ", "meB",
 "mD", "mS",
 "mDD", "mDS", "mSD", "mSS"]
female_ego_parameters = ["fMM", "fMF", "fFM", "fFF",
"fMyZ", "fMyB", "fM", "fMeZ", "fMeB", "fFyZ", "fFyB", "fF", "fFeZ", "fFeB",
"fyZ", "fyB", "feZ", "feB",
 "fD", "fS",
 "fDD", "fDS", "fSD", "fSS"]
```

Figure 3: Set of 24 parameters for the female and male ego

2. Complexity. Intuitively, a kin classification system is considered complex if it requires individuals to learn and remember a large number of terms, each with a complicated definition. For instance, Northern Paiute has separate terms for "older sister", and "younger sister", whereas English uses a single term "sister" to refer to both sisters; English has separate terms for "father", "mother", "brother", and "sister", while some languages use a single term for "parent" or "sibling" to refer to both male and female relatives. There are several factors that can affect the count of kinship terms in a given language, some languages may use the same term for a wider range of relatives, or may have more or fewer gender distinctions. To account for these factors and give a precise definition of complexity, we measured the complexity of the system by summing up the number of distinct terms in each language. By using this method, we would be able to construct a scale of complexity and order the kinship system from less complex (fewer distinct terms) to more complex (more distinct terms).

3. Informativeness To measure the informativeness of a kin classification system, we applied the Shannon entropy theorem.

$$H(X) = \sum_x p(x)s(x) = \sum_x p(x)\log_2\frac{1}{p(x)} = -\sum_x p(x)\log_2 p(x). \tag{1}$$

Specifically, we calculate the summation of the communicative cost C, which represents the extra bits of information required to refer to a specific individual within a kin group. For instance, let z be a vector that represents a partition of the 24 individuals in the kin group, if Bob is an English speaker, then z1 will equal z3 since both different aged sisters are referred to as "sister". Suppose Bob wants to refer to the younger sister (z1), and uses the phrase "younger sister", he would need to use additional information to differentiate her from his older sister. Thus, by information theory, the additional information needed when referring to an individual is ci:

$$c_i = -\log_2\left(\frac{p_i}{\sum_{z_j=z_i} p_j}\right) \tag{2}$$

where pi represents the likelihood that Bob will need to refer to individual i. While need probabilities can vary across different cultures and communicative requirements, for the current study, we assume a universal distribution of need probabilities by using the relative frequencies of kin expression across English and German Corpus. However, future studies can explore different sets of need probabilities for different cultures.

To determine the communicative cost C for Bob, we then calculate the anticipated cost for him when referring to any of the 24 individuals in his family tree:

$$C = \sum_{i=1}^{24} p_i c_i \tag{3}$$

The same communicative cost calculation process is used for both female and male egos, and the final communicative cost of a kin classification system is defined as the average of the costs for both.

4. Create new kin systems for comparison. To make new kin systems, we needed to randomly assign kin terms to kin parameters. To do this, we took a random number of kin terms and enumerated 1 to the number (e.g. 1, 2, . . . , 16, 17). Then, we randomly assigned these kin terms to 24 kin parameters (or 32 parameters for our extension). For example, one permutation of 17 unique kin words would be randomly distributed onto 24 parameters. This would be one permutation, and we repeated this 10000 times to create 10000 possible new kin systems. Doing this allows for repeats to happen (like how "sister" is used for both younger sister and older sister) in a random fashion.

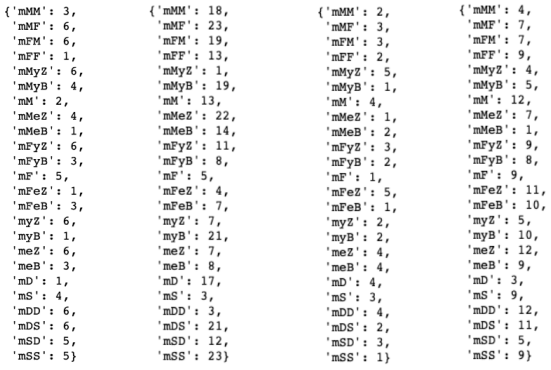

Figure 4: Example permutations on kin terms assigned to kin parameters

5. Graphing. To graph existing and newly generated languages together, we used steps 2 and 3 to find the complexity and informativeness of each of the languages. We used the dictionaries we created from each language to do this. Then, we scatter-plotted the real language systems and new language systems on a graph with the x-value being complexity, and the y-value being communicative cost. Graphing the kin systems by complexity vs. informativeness follows from the paper's method, and we believe it is appropriate to show how the real languages compare to other possible ways that languages could have formed their kinship categories.

6. Focus on subsets of the family tree. In addition to measuring the complexity and informativeness of a whole language, we also focused on subsets of the family tree. These 4 subsets were: mother and aunts, father and uncles, grandparents, and siblings. To do this, we calculated the complexity and informativeness of a subset of parameters that corresponded to the family tree subset. For example, grandparents would focus on the kin parameters: "fMM", "fMF", "fFM", and "fFF" for the female ego. We graphed real systems against permuted systems for each of the subsets in terms of their complexity and communicative cost.

Results

We test the near optimality hypothesis by using the partitions of the full family tree presented in Figure 5 and Figure 6. These partitions are compared to the permuted data to determine optimality. Figure 5 focuses on partitions of the full family tree that excludes siblings' children (e.g. niece, nephew), whereas Figure 6 examines an extended family tree that includes 32 kin parameters and takes these kin members into account. The optimal frontier, which is located near the bottom left boundary of the space, is used to identify the best systems. Since our study does not account for cousin relationships, the complexity scores would be higher and the optimal frontier, in this case, would shift to the right side of the graph. The majority of the attested systems (shown in red circles) are found near the optimal frontier for both the basic family tree and the extended family tree.

In Figures 7 to 10, we present results for analyses conducted on subsets of the family tree. The attested systems shown in red are compared with the permuted systems shown in grey, and the size of the circle indicates its frequency. These results are consistent with the near-optimal pattern observed for the complete family tree and support our hypothesis that the frequent kin classification systems tend to be more optimal than random systems in terms of the trade-off between complexity and informativeness.

While our study has successfully predicted the optimality of kin classification systems, it is important to note that our results do not account for all factors that may have contributed to this optimal pattern. For instance, local social patterns of marriage and residence, as well as cultural evolution, likely played a role in the development of these systems and are not explicitly accounted for in our study. Therefore, further

research is needed to fully understand the complex interplay factors that have shaped kin classification systems across different cultures and societies.

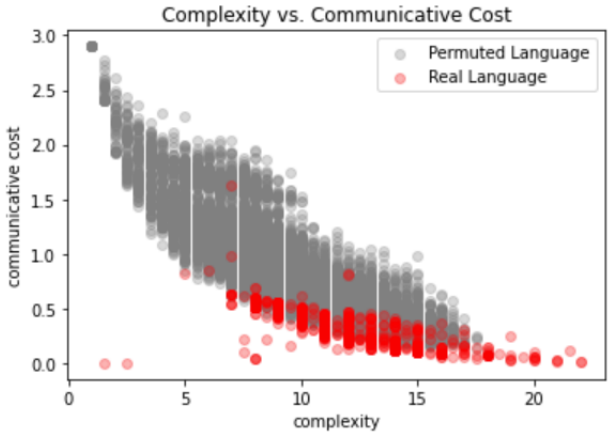

Figure 5: Complexity vs. Communicative cost of real languages vs. new languages measured with 24 parameters

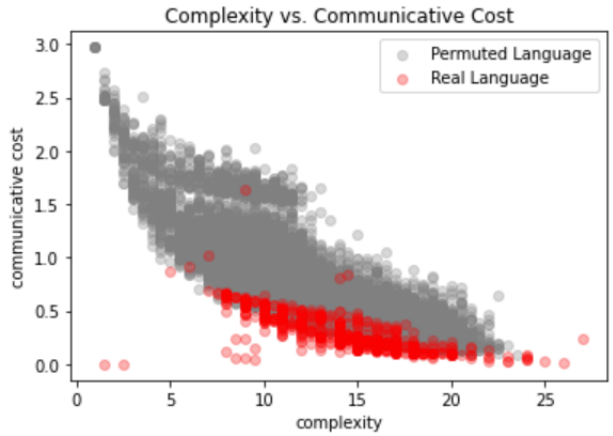

Figure 6: Complexity vs. Communicative cost of real languages vs. new languages measured with 32 parameters

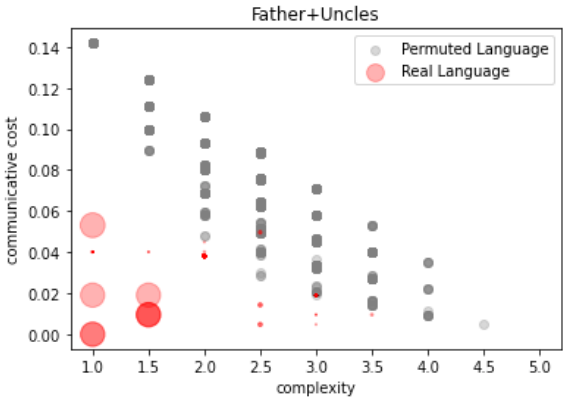

Figure 7: Complexity vs. Communicative cost of real languages vs. new languages measured with fathers and uncles

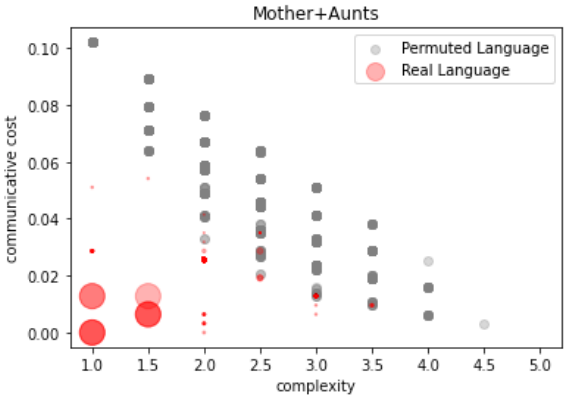

Figure 8: Complexity vs. Communicative cost of real languages vs. new languages measured with mothers and aunts

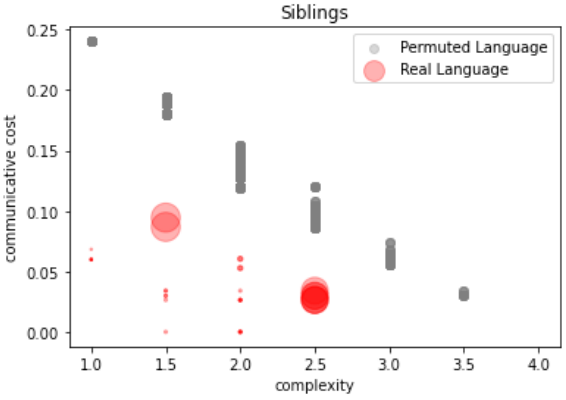

Figure 9: Complexity vs. Communicative cost of real languages vs. new languages measured with siblings

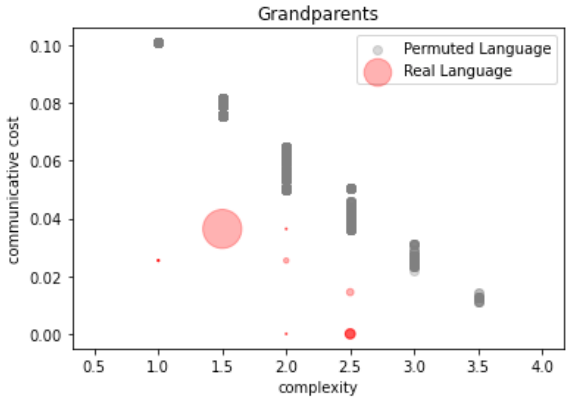

Figure 10: Complexity vs. Communicative cost of real languages vs. new languages measured with grandparents

Conclusion

We achieved very similar results to Kemp and Regier in our study. In all of our graphs, we saw those real language systems are near-optimal in terms of complexity and communicative cost. Kinship systems vary across languages, but they follow general communicative principles. These principles suggest that languages tend to show a near-optimal trade-off between simplicity and informativeness.

From this study, we were able to improve on Kemp and Regier's study. We used a more up-to-date dataset and investigated more languages/cultures. However, our study was also limited in a few ways. Firstly, we used only the need probabilities from the English and German Corpus like Kemp and Regier. These need probabilities were used in our calculations for informativeness and we used them to generalize to all languages. In reality, it may not be the case that need probabilities generalize to all languages. For example, some cultures may need to refer to the female side of the family more than the male side. The need probabilities should reflect this. Since we do not know if the need probabilities generalize, an extension would be to explore need probabilities from different culture families and use them for the informativeness calculation for each language. Another limitation was that we did not consider cousin relationships in our study. Because of this, the complexities are seen in the graphs cap at 24 or 32, which makes it seem that languages are very complex since they cluster towards the end of the graph. If we were to consider cousin relationships, then we would have much more parameters to query and the complexity cap would be higher. In the future, this would be an extension that is worth exploring.

Although the optimality trade-off trend was observed in our graphs and results, there are also many other factors that impact kinship systems, such as cultural evolution and different communicative requirements. For example, the Mosuo have a unique kinship system where all family relationships are traced through the mother's side, and there is no concept of a nuclear family (Thomas et al., 2018). Therefore, the role of the father is less significant in Mosuo society, and there is no specific term for "father" as a distinct kinship relationship. However, there are terms for male relatives on the mother's side, such as "maternal uncle" and "maternal grandfather". "Special" languages such as these which are different from the typical Western language systems that we see in our society can be further studied to understand the impact that culture has on language systems.

References

Atran, S. (1995). Causal constraints on categories and categorical constraints on biological reasoning across cultures. *D. Sperber, D. Premack, A. J. Premack (Eds.), Causal cognition: A multidisciplinary debate (pp. 205–232). Oxford University Press.*, 205–232.

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

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

Majid, A. (2015). Space under construction: Language-specific spatial categorization in first language acquisition. *Language Learning and Development*, *11*, 221–238.

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, *23*, 8386–8391.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436–1441.

Thomas, M. G., Ting, J., Jiajia, W., QioaQiao, H., Yi, T., & Ruth, M. (2018). Kinship underlies costly cooperation in mosuo villages. *R.Soc. open sci.*.

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

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