Season naming and the local environment

Charles Kemp (c.kemp@unimelb.edu.au)
School of Psychological Sciences
University of Melbourne, 3010, Australia

Alice Gaby (alice.gaby@monash.edu)
School of Languages, Cultures and Linguistics
Monash University, 3800, Australia

Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics and Cognitive Science Program
University of California, Berkeley, CA 94720 USA

Abstract

Seasonal patterns vary dramatically around the world, and we explore the extent to which systems of season categories support efficient communication about the local environment. Our analyses build on a domain-general information-theoretic model of categorization across languages, and we identify several qualitative predictions that emerge when this model is applied to season naming, including the prediction that systems with even numbers of categories should be more common than systems with odd sizes. We test the model quantitatively using a collection of season systems drawn from the linguistic and anthropological literature and data specifying temperature and precipitation in locations associated with these systems. Our results support the predicted even-odd asymmetry, and we also find that the model makes a number of successful predictions about the locations of boundaries between seasons.

Keywords: categorization; efficient communication; information theory

Imagine an alien geographer who has detailed knowledge about the natural environment in one part of our planet. The geographer knows how temperature, rainfall, humidity, wind speed, and wind direction vary over the course of the year. The geographer knows about clouds, fog, dew, storms, and lightning, and about the water levels in local streams, rivers and lakes. The geographer is intimately familiar with the flowering patterns of local plants and the breeding and migration patterns of local animals. In all of these cases the geographer knows about long-run averages as well as the variability that can be expected year to year. Before meeting any of the local people, what predictions could the geographer make about the categories named in their language? We focus on a special case of this question, and consider the extent to which named seasons reflect properties of the local environment. For example, we ask whether the geographer could predict how many seasons the local people might recognize, and where the boundaries between these seasons might lie.

Our approach builds on a growing body of work that explores ways in which languages support efficient communication (Rosch, 1978; Corter & Gluck, 1992; Gibson et al., 2019). Particularly relevant to our approach are information-theoretic accounts of variation in named categories across languages (Baddeley & Attewell, 2009; Kemp, Xu, & Regier, 2018). Regier, Kemp and colleagues have developed an information-theoretic formulation of the idea that named categories achieve a near-optimal tradeoff between complexity and communicative cost, and have applied it to domains including color (Zaslavsky, Kemp, Regier, & Tishby, 2018) and kinship (Kemp & Regier, 2012). Here we use the same formal framework to study season naming across languages.

Our work addresses an important question that is largely absent from previous formal treatments of categorization and efficient communication. The information theoretic framework that we adopt allows for different languages to reflect different communicative priorities. For example, the framework allows that systems of color categories may vary in part because speakers of different languages are embedded in environments (e.g. desert vs rainforest) with very different colour distributions. Previous authors acknowledge this point but typically implement models that assume that speakers all around the world encounter the same distributions over colors (Zaslavsky et al., 2018), kin types (Kemp & Regier, 2012), and other elements of their environments.

A notable exception is a project that explored words for ice and snow, and found that languages with a term that covers both of these concepts tend to be found in warm regions (Regier, Carstensen, & Kemp, 2016). That work focused specifically on environmental variation, but the naming behavior considered is extremely simple (one term versus two for frozen precipitation). Here we focus on environmental variation in a domain that offers the potential to make detailed predictions about not just the number of categories, but the locations of the boundaries between these categories.

Season naming has previously been studied by researchers from disciplines including linguistics, anthropology and geography. In a pioneering project Orlove (2003) compiled systems of season terms from twenty eight languages, and used them to document general tendencies in season naming. For example, Orlove suggests that seasons are usually characterized in terms of atmospheric phenomena such as rainfall, wind, and temperature. In some cases, however, seasons are based on changes related to plants (e.g. the flowering of a certain species), animals (e.g. the first appearance of a given species), or water levels in local rivers and lakes. Our approach builds on the work of Orlove and others by using computational methods to probe the relationship between season naming and the local environment.

A small amount of previous work has taken a computa-

tional approach to season naming. Hatfield-Dodds (2016) gives a detailed description of Yolngu seasons from the north east Arnhem land in Australia, and describes a computational model that uses climate data to detect when the seasons start and begin. Our work provides much less detail about any single language, but complements the approach of Hatfield-Dodds by using computational methods to explore season naming across a relatively large set of languages.

Previous authors have also discussed the notion of an optimal set of seasons for a given area. Entwisle, for example, proposes a set of five seasons for southeastern Australia that fits the local climate better than the four traditional European seasons (Entwisle, 2014). Proposals like these are often based in part on climate data, but are not typically derived from computational models. Our work builds on these approaches by connecting season naming with a domain-general account of categorization across languages.

Theoretical framework

This section introduces an information-theoretic approach that measures the extent to which a system of season terms supports informative communication about the environment. Consider a speaker who is talking about an event that falls within a standard year of 365 days. Let d indicate the day of the event. The prior distribution $p(d)$ captures the probability that the speaker will talk about an event that occurs on day d. For simplicity we assume that $p(d)$ is uniform.

Each day is associated with a distribution $p(\vec{s}|d)$ over a vector of season variables. We will consider three—precipitation (s^p), temperature (s^t), and temporal location within the year (s^y)—so that $\vec{s} = [s^p, s^t, s^y]$. Many other factors are relevant to season naming, and in principle we would like to include variables that capture information about the local climate, flora, fauna, and bodies of water. In future it may be possible to include some of these variables, but for now we work with two climate variables (precipitation and temperature) that are readily available for locations all around the world.

Each day is also associated with a distribution $p(w|d)$ over words for seasons. The speaker labels day d by sampling from the distribution $p(w|d)$. After hearing the label the listener uses Bayesian inference to compute a distribution over the season variables:

$$p(\vec{s}|w) = \sum_d p(\vec{s}|d)p(d|w) \propto \sum_d p(\vec{s}|d)p(w|d)p(d).$$

We assume that communication succeeds to the extent that the speaker distribution $s = p(\vec{s}|d)$ resembles the listener distribution $l = p(\vec{s}|w)$, and formalize this idea using the same information-theoretic measure of communication cost used by previous work on domains including color and kinship (Kemp & Regier, 2012; Zaslavsky et al., 2018). Communication cost is defined as the Kullback-Leibler divergence $KL[s||l]$ from the speaker distribution s to the listener distribution l, and is low when the distributions are similar to each other. This cost measure can be used to assess the overall communication cost associated with an entire system of season terms. This overall cost is defined as the expected cost when the speaker communicates about an event:

$$\text{system cost} = \sum_d P(d)KL[s||l] = \sum_d P(d)KL[p(\vec{s}|d)||p(\vec{s}|w)].$$

There is a tradeoff between the communication cost of a system of categories and its complexity. Complexity can be formalized in different ways (Kemp & Regier, 2012; Zaslavsky et al., 2018) and here we define the complexity of a system as the number of terms that it contains. A system with many terms (high complexity) can allow the listener to reconstruct the speaker distribution very precisely (low communication cost), but a system with few terms (low complexity) means that the listener is typically able to approximate the speaker distribution only roughly.

Previous work suggests that systems of kinship terms (Kemp & Regier, 2012) and color terms (Zaslavsky et al., 2018) are efficient in the sense that they achieve near-optimal tradeoffs between communicative cost and complexity. An optimal tradeoff is achieved if the communicative cost of a system cannot be reduced without increasing the system's complexity, and vice versa. We will explore the extent to which attested season systems support efficient communication by comparing them to hypothetical systems of equal complexity.

Synthetic climate data

To illustrate some qualitative predictions of the model we first apply it to a simple synthetic data set that specifies how a single climate variable s^c varies over a hypothetical 48 day year. Fig 1a shows a climate variable s^c that rises smoothly then falls over the course of the year, as temperature does in many parts of the world. We combined this climate variable with a temporal variable s^y so that $\vec{s} = [s^c, s^y]$. Fig 1a includes season systems that minimize communication cost for different levels of complexity. For example, when $n = 2$ the optimal categories divide the year into days when $s^c < 0.5$ and days when $s^c > 0.5$

Although the model allows categories to be disconnected the categories in these optimal systems are always connected regions of the year. This result emerges because connected categories ensure that category members have similar values of the two season variables s^c and s^y.

A second qualitative result is that the turning points of the climate variable (i.e. the peak and trough) always lie within a category rather than at a category boundary. Because the days on either side of a turning point have similar values of s^c and s^y, assigning them to the same category minimizes communication cost. A related but more subtle result is that for a fixed value of the system size n, categories containing turning points are longer than categories without turning points. For example, when $n = 4$ the categories that contain the peak and trough have 13 days each, and the remaining categories have 11 days each. In general, combining two intervals of length k that lie on either side of a turning point produces a category

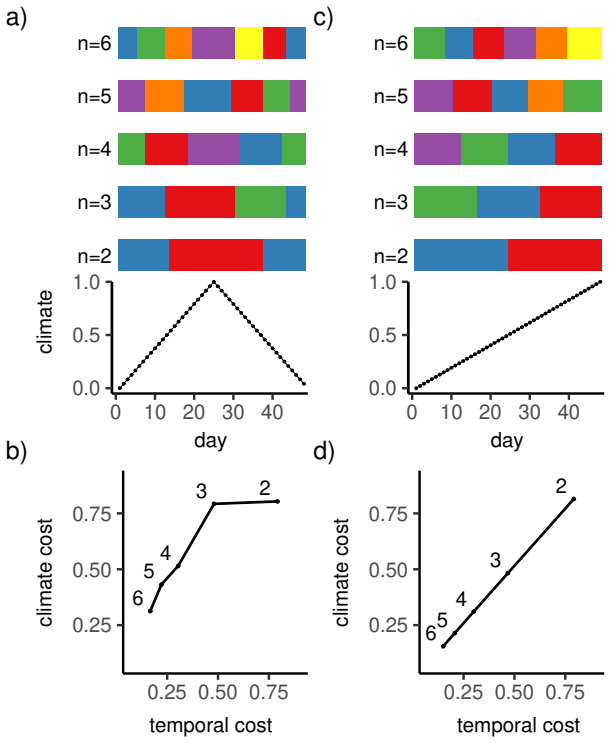

Figure 1: Analyses of synthetic climate data over a 48 day year. (a) Optimal systems of different sizes given a climate variables that varies smoothly. (b) Communication costs relative to the climate and temporal variables for the optimal systems in a. Labels indicate the sizes n of the 5 systems. (c), (d) Analogous results when the climate variable has a discontinuity at the end of the year.

of size $2k$ that has the same coherence (in s^c) as a category of size k in a region without a turning point.

A final qualitative result is that *even systems* (i.e. systems with even numbers of categories) are more effective than *odd systems* at capturing information about the climate variable. The $n = 2$ system in Fig 1a distinguishes naturally between low and high values of s^c, but distinguishing between low, medium and high values turns out to be less straightforward. If only three categories are used, then the medium category must have two disconnected components. If all categories are connected regions of the year than four categories (as for the $n = 4$ system in Fig 1a) are actually needed to distinguish between low, medium and high values of s^c. More generally, if categories are connected then at least $2k-2$ categories are needed to distinguish k levels of s^c. As a result, distinguishing between levels of s^c in a parsimonious way naturally leads to an even system.

Fig 1b compares the even and odd systems in Fig 1a by plotting communication cost with respect to variables s^c (climate cost) and s^y (temporal cost). Although communication cost was defined earlier with respect to the entire set of season variables $\vec{s}$, here we use the same approach to define communication cost with respect to one variable at a time. Fig 1b shows that moving from 2 to 3 categories produces a relatively small improvement in climate cost, but moving from 3 to 4 categories produces a relatively large improvement. A similar but less pronounced kink in the curve is visible when moving from 4 to 5 to 6 categories. Moving from 2 to 3 categories does allow a speaker to convey additional information about s^y, but Fig 1b shows that this increase in complexity provides little additional information about s^c.

Most of the qualitative results just discussed depend critically on the assumption that s^c varies smoothly over time. Figs 1c and 1d show analogous results if s^c increases smoothly over the year then drops very sharply to its original value before the year starts again. In this case optimal categories are still connected regions, but the turning point always lies at a category boundary, the categories within each system have equal sizes, and there is no even-odd asymmetry.

The simulated environment in Fig 1a is simple and highly stylized, and it is not clear whether qualitative results like the even-odd asymmetry still apply if the climate variable rises and falls at different speeds, or if additional climate variables are added. Even so, we propose that seasonal variation in real-world climates is more like Fig 1a than Fig 1c. Our analyses therefore identify several characteristics of real-world systems that might be expected purely on the basis that these systems support communication about periodic variables that vary smoothly through time.

Season naming data

We next evaluated the model using real-world naming and environmental data. Orlove's (2003) ethnoclimatology database was not available and we therefore consulted the primary literature to assemble our own data set.

The data set includes 53 languages in total. For 25 of these languages the set of season terms was described in enough detail to be roughly positioned relative to the Western calendar year, and the data set includes season boundaries for each season in each of these systems. Four examples of systems with boundaries are shown in Fig 2. For the remaining 28 languages the data set specifies only the number of season terms in each language. Our data have a strong Australian focus because our two biggest sources are collections of indigenous Australian seasonal calendars compiled by the Commonwealth Bureau of Meteorology and the CSIRO.[1]

The data set inevitably reflects a number of decisions that are somewhat arbitrary. There is no universally accepted definition of a season, and it is likely that our sources adopted slightly different notions of what qualifies as a season. Some of the systems are hierarchies with two levels: they include a number of major seasons which are in turn divided into minor seasons. Judgments about subjective seasons are likely to be especially subjective. For example, the Tiwi system includes three major seasons and thirteen minor overlapping seasons,

[1] Unless specified otherwise, all season systems discussed in this paper (including three of the four in Fig 2) are drawn from one of these resources (`http://www.bom.gov.au/iwk/index.shtml` and `https://www.csiro.au/en/Research/Environment/Land-management/Indigenous/Indigenous-calendars`).

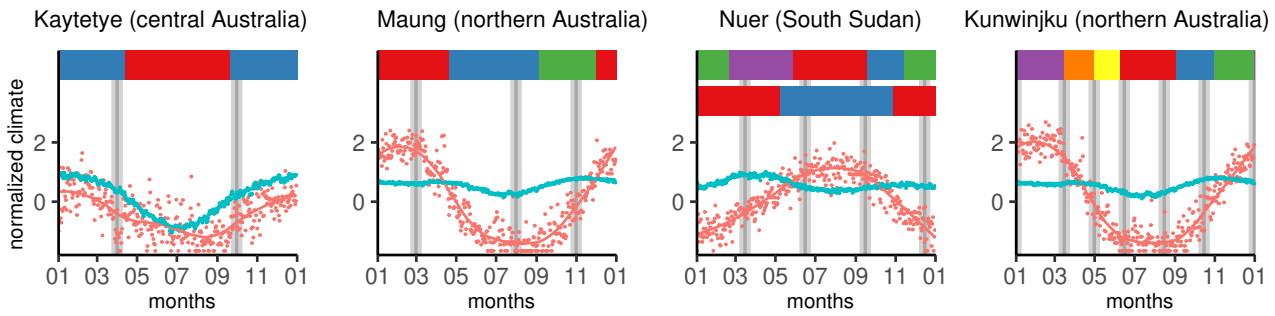

Figure 2: Climate data, seasons and optimal systems for four languages. Precipitation and temperature are shown using pink and cyan respectively. A cube-root transform has been applied to precipitation, and both variables are normalized to have zero mean and unit variance across the entire data set. Season boundaries are shown using vertical lines. The colored bars across the top show optimal systems according to the model with sizes matched to the linguistic systems. Two systems are shown for Nuer for comparison to the two major seasons and the four minor seasons recognized in this language.

including *Kurukurari* ("season of the mangrove worm", when these worms are easy to find) and *Tawutawungari* ("season of the clap sticks," when special yam ceremonies are held). It seems likely that some of the languages in our data set have minor seasons that are not documented in our sources.

Although 25 of the languages in the data set include season boundaries, our sources repeatedly stress that mappings of indigenous seasons onto the calendar year are approximate only. Seasons are often fuzzy categories with no sharp boundaries, and the boundaries between seasons often shift from year to year as a result of variability in the local climate and other factors.

Some of our sources describe overlapping seasons, and this overlap is preserved in our data set. When seasons overlapped the distribution $P(w|d)$ over season terms for a given day was taken to be uniform over all seasons including that day. None of our sources describes gaps (i.e. unnamed periods) between seasons, and as a result each system in our data assigns each day to at least one season.

Among our systems with season boundaries, seasons always correspond to connected regions of the year, but exceptions are known outside our data set. For example, Rukiga has two words for seasons, *orugazi* (rainy season) and *ekyanda* (dry season), but these seasons may alternate over the course of a calendar year so that there are two rainy seasons and two dry seasons (Orlove, 2003). For languages included in our data set, season terms may pick out disconnected regions of the year when actually applied by native speakers. For example, if an unusually cold spell occurred during the summer months, a Yolngu speaker might say that one season had "interrupted" another (Hatfield-Dodds, 2016). These interruptions mean that seasons can occur in different orders during the year, and that a particular season could occur multiple times. For all of these reasons the representations in our data set are best viewed as crude approximations of bodies of knowledge that are both rich and subtle.

Season variables

The precipitation (s^p) and temperature (s^t) variables are based on global gridded data available from the Climate Prediction Center (CPC) in the USA.[2] Our analyses used daily precipitation and daily temperature averaged over the period from 1979 to 2005 and excluding leap years. Following a common practice in climate modeling we applied a cube-root transform to the precipitation data. We then normalized both variables to have zero mean and unit variance, and normalized variables for four locations are shown in Fig 2.

We assigned Glottocodes manually to each language in the data set then retrieved the language family (e.g. Pama-Nyungan) and position (i.e. latitude and longitude) associated with each language in the Glottolog data base (Hammarström, Forkel, & Haspelmath, 2018). We then used these positions to extract precipitation and temperature data for each language from the CPC data.

The distribution $p(\vec{s}|d)$ for a given day and location was defined as a multivariate Gaussian distribution over a three-dimensional space. Two of the dimensions were the normalized precipitation and temperature dimensions already described, and the temporal dimension ran from 1 to 365 days and wrapped around so that day 366 was identical to day 1. The covariance was an axis-aligned distribution with standard deviation of 0.1 along the precipitation and temperature dimensions and standard deviation of 40 along the temporal dimension. The relative magnitudes of these standard deviations capture assumptions about the extent to which season categories should be informative about the three dimensions. For example, increasing the standard deviation along the temporal dimension would mean that there is less pressure for season categories to convey precise information about the location of an event within a year. As a result the temporal dimension would become less important and precipitation and temperature would effectively become more important. The numerical parameters used in our analyses (i.e. 0.1 and

[2]CPC data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at `https://www.esrl.noaa.gov/psd/`

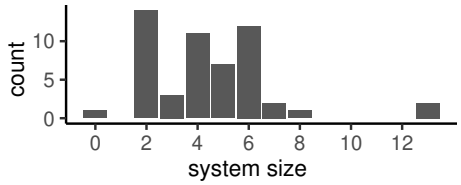

Figure 3: Distribution of system sizes.

40) were intended to give precipitation and temperature equal weight, and to capture the idea that season categories locate events within the calendar year only roughly.

Because climate data are noisy we smoothed the distributions $p(\vec{s}|d)$ using a linear kernel with a width of 9 days. This smoothing process meant that the distribution $p(\vec{s}|d)$ for a given day (e.g. Jan 15) was defined as a weighted average of distributions for Jan 11 through 19. As a final step we discretized these distributions over a regular grid for use in our information-theoretic analyses.

Analysis of system sizes

Fig 3a shows the distribution of system sizes across our data set. For languages with hierarchical systems, the system size is defined as the number of seasons at the finest level of resolution. The system of size zero corresponds to the Grand Valley Dani, who constitute "a significant exception to [the general statement] that all cultures recognize seasons" (Heider, p 212). The two systems of size 13 represent Tiwi (described earlier) and Ngan'gi, and both feature ecological seasons defined with respect to the local plants and animals. Other languages in the data set almost certainly have ecological seasons that were not documented in our sources, and our Fig 3 therefore exaggerates the difference between Tiwi and Ngan'gi and the other languages in our data set.

As suggested earlier the model predicts a preference for systems with even sizes, and Fig 3a reveals that 2, 4 and 6 are the most common sizes. Leaving aside the three systems with sizes of zero or 13, 38 out of 50 systems or 76% have even sizes. We evaluated the significance of this result using a Bayesian mixed effects binomial model based on the `rstanarm` package and its default priors (Goodrich, Gabry, Ali, & Brilleman, 2018). The binary outcome variable indicated the parity (even or odd) of a system, and we included both a fixed intercept and a random intercept for language family to acknowledge genetic relatedness between languages. [3] The median of the fixed intercept indicates a probability of 0.77 that a random system would have an even size, and the 95% posterior credible interval ($[0.59, 0.94]$) excludes the probability (0.5) that makes even and odd systems are equally likely. Our data therefore support the conclusion that even systems are more common than odd systems.

Orlove (2003) previously noted that systems with odd sizes are rare, and in his data 23 out of 28 systems, or 82% have an even size. He did not offer an explanation for this asymmetry, but we have argued that it emerges from a pressure for season systems to support informative communication about factors that vary smoothly over time.

[3] The model call was `stan_glmer(parity ~ 1 + (1|language_family), family='binomial')`

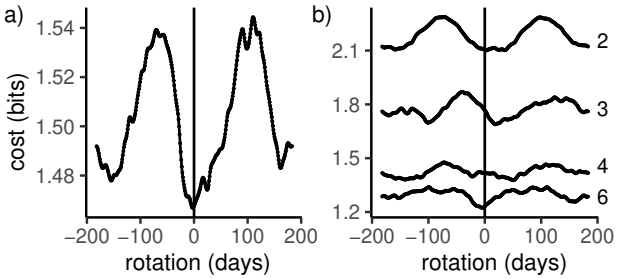

Figure 4: (a) Average rotation curve across all languages. (b) Averages across 2-term systems (4 languages), 3-term systems (3 languages), 4-term systems (8 systems) and 6-term systems (10 languages). The curve for 5-term systems (2 languages) is not shown because it overlaps the 6-term curve.

Analysis of season boundaries

Our remaining analyses focus on the 25 languages for which we have season boundaries. Four of these languages are shown in Fig 2 along with optimal systems according to our model.

Kaytetye has two seasons — *Watangka* (hot season) and *Yurluurp* (dry season) — and the boundaries between these categories roughly match the model predictions. Maung has three seasons: *Walmatpamalat* (heavy rain), *Wumulukuk* (cold weather) and *Kinyjapurr* (hot and humid). The model predicts three categories of roughly the right duration—in particular, the category that includes the steep increase in precipitation is shorter than the other two. The predicted season boundaries, however, are all shifted later in the year relative to the Maung system. Fig 2 also suggests that two of the Maung season boundaries lie close to simultaneous turning points in both temperature and rainfall. The Maung system therefore challenges the qualitative prediction that turning points in the climate data should lie within categories rather than at category boundaries.

Nuer has two major seasons: *tot* (mid-March to mid-September) and *mei* (mid-September to mid-March), each of which is divided into two minor seasons. The Nuer system provides additional evidence that season boundaries can be aligned with turning points in the climate data. Evans-Pritchard (1939, p 191) notes that "the *mei* season commences at the decline of the rains—not at their cessation." At the beginning of *mei* the Nuer start to anticipate the life they will lead when the dry weather arrives, and Evans-Pritchard (p 191) writes that their classification of seasons "aptly summarizes their way of looking at the movement of time, direction of attention in marginal months being as significant as actual climatic conditions."

Kunwinjku has six terms: *Kudjewk* (monsoon season), *Bangkerreng* (knock'em down storms), *Yekke* (start of dry time), *Wurrkeng* (cool weather time), *Kurrung* (hot, dry time), and *Kunumuleng* (humidity builds). The boundaries in the model system roughly match the linguistic data, and

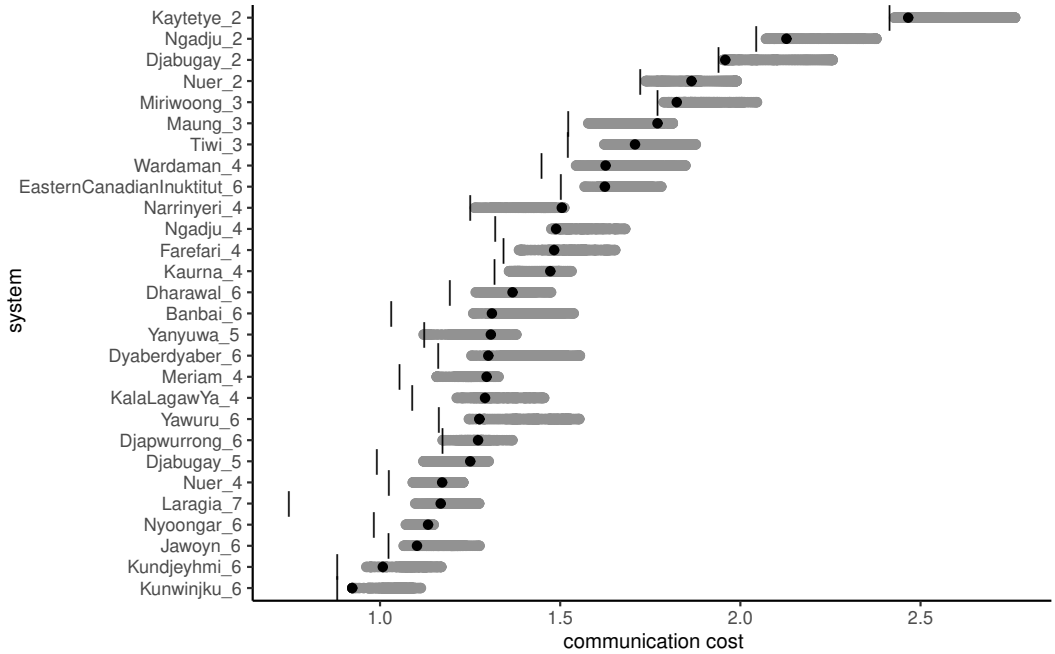

Figure 5: Attested systems (black dots) compared to rotations (grey dots) and size-matched optimal systems (black lines).

the model correctly predicts that there should be two short adjacent seasons during the part of the year when rainfall is declining sharply. Much more could be said about Kunwinjku and about each specific language in Fig 2, but we now turn to analyses that range more broadly across the entire data set.

A first general question is how closely season categories are aligned with variation in the two environmental variables (temperature and precipitation) included in our model. If a given system is closely aligned with the environmental variables, then rotating the system through the calendar year (i.e. incrementing all season boundaries by a constant while allowing for wrap around) should disrupt this alignment. Fig 4a plots communication cost against rotation size, and suggests that attested systems (i.e. systems rotated by zero days) tend to achieve lower communication cost than rotations of these systems. As shown in Fig 4, 0 day rotations score better than 99% of the 365 possible rotations. Fig 4b shows separate rotation curves for systems of size 2, 3, 4 and 6. The 2 term systems make an especially large contribution to the average result in Fig 4a, but a clear trough at zero days is visible also for the systems of size 6.

Fig 5 summarizes rotation results for individual languages. The three languages with hierarchies are included twice in the plot, once for each level of the hierarchy. Some systems (black dots) score better than most of their rotations (gray bar), including Kaytetye and Kunwinjku from Fig 2), but others (in particular Narrinyeri) do not. On average, each system scores better than 64% of its rotations.

Fig 5 also compares each system to the optimal system according to our model. Again, the pattern of results is mixed. Some systems (including Kaytetye and Kunwinjku) achieve scores close to the optimum, but others (including Laragia) do not. The most likely explanation is that our model and data set are limited in many respects. Perhaps the most glaring example is that we considered only two environmental variables even though we know that language groups around the world use many markers of seasonal transitions other than changes in precipitation and temperature. From this perspective, it seems striking that the model performs as well as it does given the limited information available to it.

Conclusion

We developed a computational model that assumes that systems of season terms are near-optimal at conveying information about the local environment. The model helps to explain why systems with odd numbers of terms are relatively rare, and makes a number of successful predictions about the locations of season boundaries.

Our results do not provide strong support for claims about optimality but nevertheless demonstrate the value of the efficient-communication approach to naming and categorization. Most interesting to us are the qualitative issues exposed by the model. We have touched on some of them already, including the even-odd asymmetry, and the relationship between season boundaries and turning points in environmental variables. Many others arise: for example, our approach could be used to test the hypothesis that systems with large numbers of terms are especially likely to be found in regions with variable climates, and the hypothesis that boundaries are more likely to be aligned with sharp transitions (e.g. the first major rainfall of the year) than gradual changes in variables such as temperature. Although our current model is extremely simple, we have found it to be a useful conceptual tool for thinking about season naming across languages.

Acknowledgments

TR's work on this study was supported in part by the Defense Threat Reduction Agency; the content of the study does not necessarily reflect the position or policy of the U.S. government, and no official endorsement should be inferred

References

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychological Science*, *20*(9), 1100–1107.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: feature predictability and information. *Psychological Bulletin*, *111*(2), 291–303.

Entwisle, T. (2014). *Sprinter and Sprummer: Australia's changing seasons*. CSIRO.

Evans-Pritchard, E. E. (1939). Nuer time-reckoning. *Africa*, *12*(2), 189–216.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy., R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.17.4)

Hammarström, H., Forkel, R., & Haspelmath, M. (2018). *Glottolog 3.3.* Max Planck Institute for the Science of Human History. Jena. Retrieved from `https://glottolog.org/`

Hatfield-Dodds, Z. (2016). *Integrating understandings of a Yolngu seasonal calendar* (Honours Thesis). Australian National University.

Heider, K. G. (1970). *The Dugum Dani: A Papuan culture in the highlands of West New Guinea*. Wenner-Gren Foundation.

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*, 109-128.

Orlove, B. (2003). How people name seasons. In S. Strauss & B. S. Orlove (Eds.), *Weather, climate, culture.*

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

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

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*.