Birds and Words: Exploring environmental influences on folk categorization

Joshua T. Abbott (joshua.abbott@unimelb.edu.au)
Charles Kemp (c.kemp@unimelb.edu.au)
School of Psychological Sciences
University of Melbourne, Parkville, Victoria 3010, Australia

Abstract

Anthropologists and psychologists have long studied how living kinds are organized into categories, and a recurring theme concerns the relationship between folk categories and the structure of the environment. We ask whether the frequency and physical size of a species affect how it is classified, and address this question by linking frequency data from eBird (an online database of bird observations) with an existing taxonomy of Zapotec bird names. A first set of analyses explores whether frequency and size predict whether a bird is named and how many other birds it is grouped with. A second set explores whether frequency and size predict the word forms used as category labels. We find some evidence that frequency affects both category extensions and naming, but the results hint that frequency may be dominated by other factors such as perceptual similarity.

Keywords: folk biology; ethno-ornithology; categorization; cognitive anthropology; bird naming

Introduction

Languages around the world include rich systems of names for plants and animals, and each system can be viewed as the outcome of a natural experiment in which generations of speakers have organized their local environment into categories. A classic line of work in cognitive anthropology addresses the question of how named categories reflect the structure of the local environment (Berlin, 1992; Malt, 1995). One prominent theme is that folk taxonomies often align well with Western scientific taxonomies, suggesting that folk taxonomies are shaped more by environmental structure than by the idiosyncratic needs and concerns of a particular culture (Berlin, 1992)

Much of the cognitively-oriented work on folk biology took place last century, and in recent years new data sets have made it possible to characterize the structure of the environment in ways that were previously difficult or impossible (Sullivan et al., 2009; Wilman et al., 2014). Here we draw on these resources to revisit the classic question of the relationship between named categories and the environment. We focus on birds in particular, and begin by compiling properties of the bird species in a given area (e.g. how big is each species, and how often is it observed?) We then study how these properties relate to named bird categories in the local language. In particular, we ask whether the frequency of a species influences whether the species is named, and if so whether frequency influences the form of the name for that species and how many other species it is grouped with.

The effects of frequency on categorization have been previously studied in the psychological literature (Parducci, 1983; Nosofsky, 1988; Barsalou, Huttenlocher, & Lamberts, 1998). One relevant finding is that categories tend to be relatively broad in low-frequency regions of stimulus space, but relatively narrow in regions including frequently encountered stimuli (Parducci, 1983). We might therefore predict that bird species encountered frequently are more likely to be assigned to their own distinctive categories.

Our focus on frequency also connects with a prominent debate between *intellectualist* (Berlin, 1992) and *utilitarian* (Hunn, 1982) accounts of folk classification. The intellectualist view holds that named categories reflect "fundamental biological discontinuities" that are perceptually salient (Berlin, 1992, p 53), and assigns a minimal role to frequency. The utilitarian view emphasizes ways in which categories are useful for a given culture, and naturally accommodates frequency effects because assigning a label to a category is especially worthwhile if there are many occasions to use it.

The next section introduces the data sets that we use, and we then address two broad questions. First, we focus on category extensions, and ask whether environmental factors predict whether a species is named, and how the set of named species is organized into groups. Second, we focus on category labels, and ask whether environmental factors predict the relative lengths of category labels, and which labels have the structure of unmarked prototypes.

Data sets

The literature contains detailed folk classifications of birds from several languages around the world, and we focus here on named bird categories from Zapotec (Hunn, 2008), a language spoken in Oaxaca, Mexico. We used two data sets that characterize the frequency and size of bird species found in Oaxaca, and a third that specifies how these species are organized into named categories.

Frequency data

Our frequency data are drawn from eBird, a citizen-science based bird observation network managed by the Cornell Lab of Ornithology (Sullivan et al., 2009). eBird data are contributed by bird lovers (both professional and amateur) who use the site to record the time and place of bird sightings. We used data from just the region containing the state of Oaxaca, Mexico.[1] An observer who sees a group of 5 vultures may record both the species (e.g. *Cathartes aura*) and the number of birds in the group (5), but we treated each case like this as a single observation of the species in question. Our data for Oaxaca include 660,223 unique observations of 922 distinct species.

We will take eBird counts as a very rough proxy for the frequency with which a species is encountered in the course of everyday life. The fact that nocturnal species will tend to have lower counts than equally common diurnal species is therefore a strength of the data rather than a limitation. eBird, however, does not provide an unbiased measure of frequency in everyday life. As a group, eBird contributors are more interested in some species than others, and counts for rare but iconic species (e.g. the bald eagle in the USA) will overestimate the frequency with which they are encountered relative to other species. Even so, eBird is a valuable resource following strict data quality standards (Kosmala, Wiggins, Swanson, & Simmons, 2016) that allows rough estimates of a variable (frequency) that would otherwise be extremely difficult to measure.

Size data

Beyond frequency it is plausible that physical and behavioral characteristics of birds both influence folk categorization (Alcántara-Salinas, Ellen, & Rivera-Hernández, 2016). Hunn (1999) has documented that smaller species are more likely to be lumped together into large categories, and that larger species are more likely to be given distinct names. Following his lead we evaluate bird size as an influence on categorization, and use size data from EltonTraits (Wilman et al., 2014) which includes information on key attributes for all 9993 extant bird species, including those from Oaxaca. We use the body mass variable, separately sourced from (Dunning Jr, 2007), which is defined as the geometric mean of average values provided for both sexes. Beyond body mass EltonTraits includes variables related to diet types, foraging strata, and activity patterns, and future studies can explore whether and how these variables influence naming.

Naming data

Our naming data are based on a detailed folk taxonomy of the Zapotec language provided by Hunn (2008) based on his his fieldwork in San Juan Gbëë, a small village in Oaxaca, Mexico.[2] Folk classifications include names at different taxonomic ranks, and our data set includes a scientific name, a folk-specific name and a folk-generic name for each species listed. For example, *Colibri thalassinus* (Mexican violetear) is named *dzǐng̱-yǎ-guì* (mountain hummingbird) at the folk-specific level and *dzǐng̱* (hummingbird) at the folk-generic level. According to Hunn's taxonomy the folk-generic category *dzǐng̱* (hummingbird) includes 14 different species. These 14 species are partitioned into 4 categories at the folk-specific level: *dzǐng̱*, *dzǐng̱-dán-yǎ-guì*, *dzǐng̱-gué*, and *dzǐng̱-yǎ-guì*. As this example suggests, in some cases the folk-specific and folk-generic names for a species are identical: for example, *Amazilia Beryllina* is known simply as *dzǐng̱* (hummingbird) at the folk-specific level.

In total Hunn's taxonomy includes 153 species that are organized into 94 distinct folk-specific categories, which in turn are organized into 68 folk-generic categories. The scientific species labels given by Hunn did not always match those used by our other sources of data (eBird and EltonTraits). We took the Clements checklist (used by eBird) as our gold standard (Clements, 2007), and some manual preprocessing was required to align the labels used by all three resources.[3]

Analysis of category extensions

Given the three data sets just described we ask whether the frequency and body mass data predict aspects of Hunn's naming data. We focus first on the extensions of folk categories, and subsequently consider the labels or names given to these categories.

Our first analysis considers whether frequency and mass predict whether a species is likely to be named. Intuitively, one might expect that common species are more likely to be named, and that larger species are especially salient perceptually and therefore more likely to be named. Most descriptions of folk classification systems in the literature do not systematically describe species found in the local area that are *not* named by the local residents. Our eBird data, however, include species that were documented in Oaxaca but not included in Hunn's taxonomy. Some of these species are probably rarely if ever seen in the village (San Juan Gbëë) where Hunn carried out his fieldwork. We expect, however, that some species missing from the taxonomy would be occasionally encountered in San Juan Gbëë.

Figure 1 shows distributions of frequency and size for birds with and without Zapotec names. As expected, on average birds that are named tend to be more frequent than birds that are not named. The mass distributions for named and unnamed birds, however, are very similar. To confirm these impressions we ran a logistic regression including log frequency and log mass as predictors of a binary variable that indicates whether a species was named. The estimated coefficients were $\beta = 0.76 \pm 0.09$ (log frequency) and $\beta = 0.005 \pm 0.07$ (log mass). We compared the full logistic regression model to alternatives that removed either log frequency or log mass as a predictor, and found that removing log frequency produced a significant impairment ($\chi^2(1) = -123.52, p < 1e-10$), but removing log mass did not ($\chi^2(1) = 0.006, p = 0.94$). Akaike information criterion (AIC) scores supported the conclusion that the model with log frequency but without log mass is the best among the three.

[1] We used all eBird observation of frequency from the Basic Dataset (EBD) on `https://ebird.org/data/download`, last accessed January 24, 2020.

[2] Also available online at `http://faculty.washington.edu/hunn/zapotec/z5.html`

[3] All data and analysis code is available at `https://github.com/joshabbott/birdnaming`.

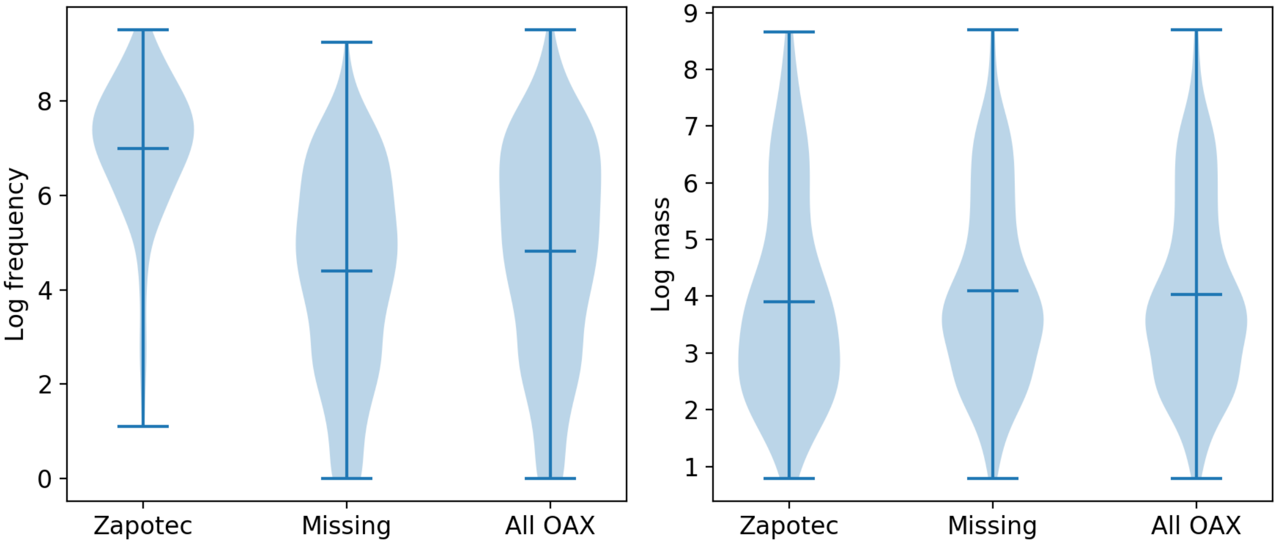

Figure 1: Frequency densities (left) and Mass densities (right) for species named in Zapotec, species found in Oaxaca (OAX) but missing from the Zapotec taxonomy, and all species in OAX.

Category size

Next we explore how named species are organized into categories. In particular, we explore whether the frequency and mass of a species partially predict the number of other species it is grouped with at the folk-generic level. Following prior work on frequency effects in the experimental literature (Parducci, 1983), we hypothesized that frequent species would tend to be grouped with fewer others, because investing in a distinctive name for a species makes most sense if there are many occasions to use it. We also expected to replicate the work of Hunn (1999), who reported that larger species tend to be grouped with fewer others, which makes sense given that larger species are especially salient perceptually.

The *category size* of each species is defined as the total size of the folk-generic category to which it belongs. For example, *Colibri thalassinus* (Mexican violetear) is grouped with 13 other species called *dz̦ǐng̱* (hummingbird) at the folk-generic level, and therefore receives a category size of 14. Figure 2 shows plots of category size (at the folk-generic level) against both frequency and mass, and shows that mass is a stronger predictor ($r^2 = 0.10$) than frequency ($r^2 = 0.005$). The coefficients of a linear regression also suggest that mass ($\beta = -1.18 \pm 0.20$) is a stronger predictor than frequency ($\beta = -0.55 \pm 0.23$). Comparing the full model with both predictors to models that remove one predictor, however, suggests that removing mass significantly impairs the fit of the model ($\chi^2(1) = 504.82, p < 1e-5$), as does removing frequency ($\chi^2(1) = -78.89, p < 0.05$). AIC values support this same conclusion.

Overall these findings support Hunn's finding that larger species tend to be assigned to smaller categories, but suggest that frequency predicts category size only weakly. Our analyses, however, did not include a third environmental factor which probably interacts with body mass and frequency. The category size of a species almost certainly depends on how many other similar species are found in the environment. For example, one reason why *dz̦ǐng̱* (hummingbird) is the largest folk-generic category in our data is that Oaxaca has many species of hummingbirds that look relatively similar to each other and relatively distinct from other species in the environment. In future work we plan to further explore the influence of perceptual similarity on folk categorization systems.

Analysis of category labels

The previous section focused on category extensions, but we now ask whether frequency and mass influence the form of the names for each species. We focus on names at the folk-specific level and consider three lexical properties of these names: name length, whether the name is a compound or monomial, and whether the name is an unmarked prototype.

Name length

Zipf's law of brevity (1936, see also Ferrer-i-Cancho et al., 2013) is the well-established regularity that word lengths are inversely related to word frequency. Intuition suggests that the frequency with which a species is named should roughly track the frequency with which it is observed, and we therefore hypothesized that more frequent species would tend to have shorter names. In contrast, we expected that there would be no relationship between body mass and name length.

Because we do not have phonemic representations of the Zapotec names, we used a crude measure of length based on the number of characters in the written form of each

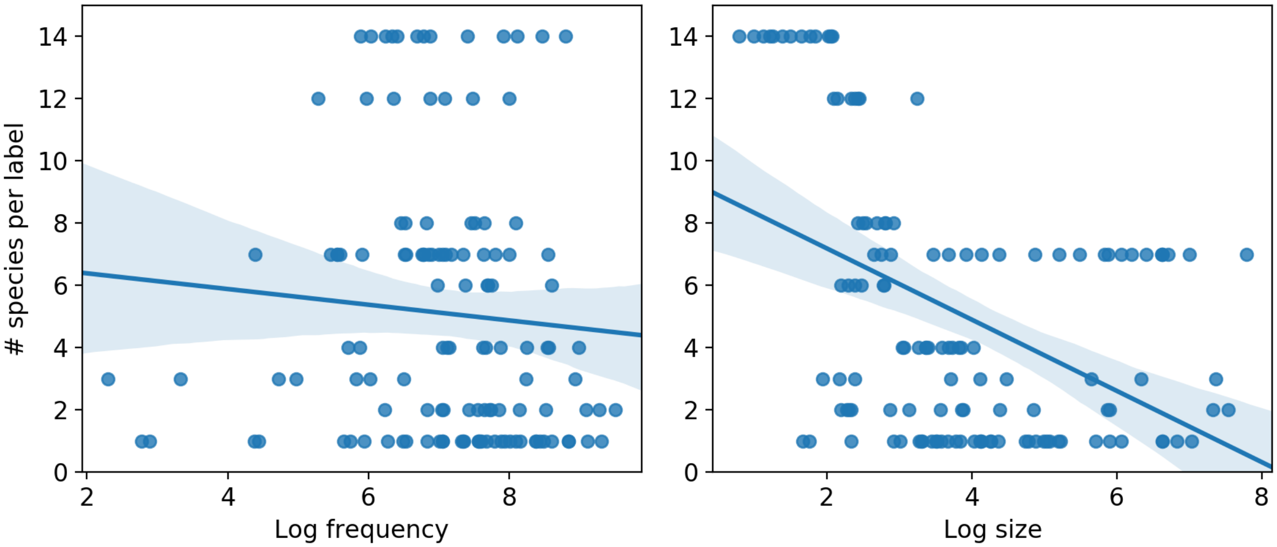

Figure 2: Category size plots for both frequency (left column) and mass (right column). Category size is defined as the number of other species a bird is grouped with under the same folk-generic name. Each point represents a bird species named in Zapotec.

name. We analyzed both the frequencies and masses of birds named in Zapotec in relation to the name length of the bird. In a direction opposite to our predictions, birds with longer names had a slight tendency to be more frequent, while having a slight tendency to be of smaller size. However, the coefficients of a linear regression suggest that log frequency ($\beta = 0.04 \pm 0.03$) and log mass ($\beta = -0.06 \pm 0.03$) are both weak predictors of log name length. Comparing the full linear regression model with both predictors to alternatives that dropped mass as a predictor significantly impaired model performance ($\chi^2(1) = 1.15, p < 0.03$) but dropping frequency did not ($\chi^2(1) = -0.48, p < 0.16$). AIC scores support the same conclusion.

Compound names

As suggested earlier, some hummingbirds have compound names at the folk-specific level (e.g. *dzǐng-yǎ-guì* (mountain hummingbird)) but others do not (e.g. *Amazilia Beryllina* is known simply as *dzǐng* (hummingbird)). Compound names are notated in Hunn's taxonomy with a dash ('-'), and we explored whether frequency and mass could predict whether the folk-specific name for a species is compound or monomial.

Compounds tend to be longer than monomials, and consistent with our analysis of name lengths we found that neither log frequency nor log mass predicts whether a species has a monomial name. The coefficients of a logistic regression suggest that log frequency ($\beta = -0.04 \pm 0.12$) is a stronger predictor than log mass ($\beta = -0.04 \pm 0.11$), and that as the frequency or mass of a species increases it becomes less likely to have a compound name. However, comparisons of the full model to alternatives that remove either log frequency or log mass suggest that neither predictor is significant ($\chi^2(1) = 0.10, p = 0.75$ for log frequency and $\chi^2(1) = 0.13, p = 0.72$ log mass). AIC scores support the same conclusion.

Prototypes

The literature on folk categorization proposes a link between monomial labels and category prototypes (Berlin, 1972, 1992). If some vultures (e.g. Turkey vulture, *Cathartes aura*) are simply called *pěch* at the folk-specific level but others have a compound name (e.g. *pěch-rúx*, or Black vulture), then vultures with the monomial name might be expected to be more typical than those given a distinctive folk-specific name. Several factors could contribute to typicality: for example, typical vultures could be those encountered most frequently, or those that are perceptually most representative of the folk-generic category *pěch* (Berlin, 1992). Here we test the hypothesis that frequency predicts typicality.

For us, any folk-specific category (e.g. the one that includes *Cathartes aura*) with the same label as the folk-generic category to which it belongs will be called an *unmarked prototype*. Although this definition of a prototype is based purely on linguistic form, we expect that it lines up with the psychological notion of a prototype (Rosch, 1973). Hunn's data include 11 unmarked prototypes, and for simplicity we focus on the 6 prototypes that include a single species each, which means that the species in question can be treated as a category prototype.

Figure 3 shows ranked raw frequency distributions that compare the frequency of each category prototype to the frequencies of other members of the same folk-generic category. The top left chart indicates that the prototypical Turkey Vulture (in black) is more frequently observed in Oaxaca than

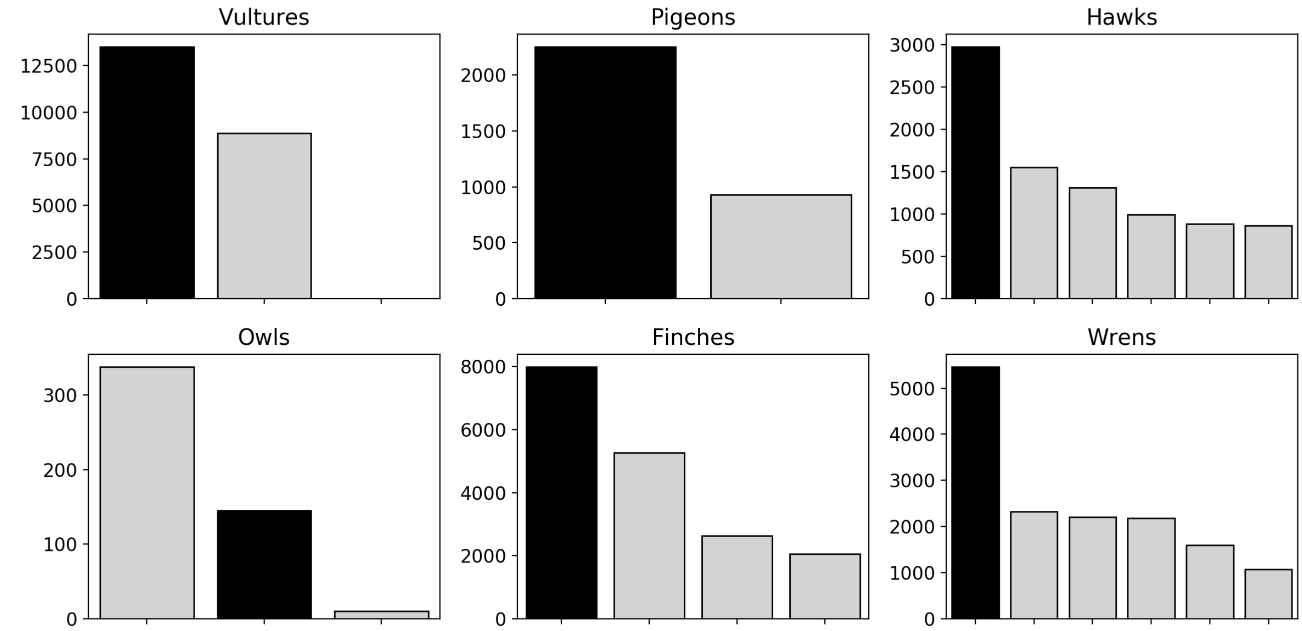

Figure 3: Frequency bar plots of birds named in Zapotec as unmarked prototypes along with frequencies of other birds with the same folk-generic name. The bar highlighted in black is the unmarked prototype.

other vultures. This trend holds across 5 of the 6 cases, with a notable exception for the folk-generic category including owls. Although the counts for this category are comparatively low, the unmarked prototype (Great Horned Owl) is only the second most frequent owl in the eBird data. Hunn notes that the Great Horned Owl is considered an ill-omen by many, and the cultural salience of this species may therefore explain why it receives an unmarked name despite being infrequently observed. Overall, these results suggest that Zapotec unmarked prototypes can be predicted by frequency of observation.

Discussion

We drew on large-scale digital data sets to ask whether the physical size of a bird species and its frequency of occurrence predict how it is classified in a Zapotec folk taxonomy. Our first set of analyses found that frequency (but not mass) predicts whether or not a species is named, and that both frequency and mass predict category size (i.e. how many other species a given species is grouped with at the folk-generic level). Hunn previously reported that mass predicts category size, and we found mass to be a stronger predictor than frequency. The relatively weak effect of frequency may seem at odds with previous experimental studies that report strong effects of frequency on categorization (Parducci, 1983). Experimental work, however, is often able to tightly control stimulus similarity, but we analyzed biological species that belong to a similarity space with rich naturalistic structure. Although perceptual similarity and frequency both affect categorization, the weak effect of frequency in our category size analysis suggests that perceptual similarity may be the stronger of the two factors.

Our second set of analyses focused on category names. Although our results suggest that frequency influences some aspects of naming (e.g. whether the folk-specific name for a species is an unmarked prototype), to our surprise, we found that frequency did not predict the length of a species' name, or whether the name is compound or monomial. Two possible explanations seem plausible. Consistent with Zipf's law of brevity, it is possible that frequently used names do tend to be short, but that these names do not pick out the species that are observed most frequently. Commonly discussed species may include cases (e.g. the bald eagle in American culture) that have cultural significance even though they are observed relatively rarely. The second possible explanation is that the effect of frequency is again dominated by perceptual similarity. For example, compound names may be most useful in "crowded" regions of perceptual space where they can serve to distinguish one species from its neighbors. In extreme cases in which a species (e.g. the Australian emu) is the only member of a folk-generic category, there is no reason to give it a compound name at the folk-specific level regardless of how frequently the species is observed.

Frequency plays a central role in theories of communicative efficiency. For example, efficiency-based theories of categorization predict narrow categories in frequently-observed regions of stimulus space (Regier, Carstensen, & Kemp, 2016), and Zipf's law of brevity can be explained in terms of communicative efficiency (Piantadosi, Tily, & Gibson, 2011) In turn, the notion of communicative efficiency is linked with the utilitarian approach to folk classification (Hunn, 1982), which suggests that folk categories are best understood by explaining the purpose they serve in a given culture. Our results, however, suggest that frequency-related effects may be dominated by perceptual similarity. If supported in future work this conclusion would be broadly compatible with the intellectualist account of folk classification (Berlin, 1992), which focuses on perceptual salience rather than communicative utility. A pressing goal for future work is to combine our frequency data with a perceptual similarity space and to explore whether and how the two interact in shaping folk classification. The similarity space for this analysis can potentially be derived from the work of Pigot et al. (2020), who generated a 9-dimensional space that includes most of the world's bird species and is based on body mass in addition to 8 variables related to beak shape and body shape.

A second important goal for future work is to expand our approach to languages other than Zapotec. The best candidates are languages for which a reliable folk taxonomy is available and for which eBird data is relatively plentiful for the geographic region in question. Clear next steps are to analyze folk taxonomies for Tzeltal (Chiapas, Mexico; Hunn, 1977) and Tlingit (south-east Alaska; Hunn & Thornton, 2012).

Conclusion

Psychologists, linguists and anthropologists have all studied how naming and categorization are affected by the structure of the environment. Working in this tradition we explored how Zapotec folk categories for birds are influenced by the physical size of bird species and the frequency with which they are observed. Our frequency data were drawn from an online database of bird observations, and our work therefore illustrates how large-scale digital data can be used to characterize environmental structure in new and useful ways. Our analyses so far have been extremely simple, but we see them as initial steps in a research program that combines large-scale environmental data and folk taxonomies to yield new insight into categorization and naming across cultures.

References

Alcántara-Salinas, G., Ellen, R. F., & Rivera-Hernández, J. E. (2016). Ecological and behavioral characteristics in grouping Zapotec bird categories in San Miguel Tiltepec, Oaxaca, Mexico. *Journal of Ethnobiology*, *36*(3), 658–682.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, *36*, 203–272.

Berlin, B. (1972). Speculations on the growth of ethnobotanical nomenclature. *Language in society*, *1*(1), 51–86.

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

Clements, J. F. (2007). *Clements checklist of birds of the world*. Comstock Pub. Associates/Cornell University Press.

Dunning Jr, J. B. (2007). *CRC handbook of avian body masses*. CRC press.

Ferrer-i Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., & Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, *37*(8), 1565–1578.

Hunn, E. S. (1977). *Tzeltal folk zoology: The classification of discontinuities in nature*. New York: Academic Press.

Hunn, E. S. (1982). The utilitarian factor in folk biological classification. *American Anthropologist*, *84*(4), 830–847.

Hunn, E. S. (1999). Size as limiting the recognition of biodiversity in folkbiological classifications: One of four factors governing the cultural recognition of biological taxa. *Folkbiology*, *47*, 47–69.

Hunn, E. S. (2008). *A Zapotec natural history: Trees, herbs, and flowers, birds, beasts, and bugs in the life of San Juan Gbëë*. University of Arizona Press.

Hunn, E. S., & Thornton, T. F. (2012). Tlingit birds: An annotated list with a statistical comparative analysis. In *Ethno-ornithology* (pp. 211–240). Routledge.

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560.

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*, *29*(2), 85–148.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: learning, memory, and cognition*, *14*(1), 54.

Parducci, A. (1983). Category ratings and the relational character of judgment. In *Advances in psychology* (Vol. 11, pp. 262–282). Elsevier.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pigot, A. L., Sheard, C., Miller, E. T., Bregman, T. P., Freeman, B. G., Roll, U., ... Tobias, J. A. (2020). Macroevolutionary convergence connects morphological form to ecological function in birds. *Nature Ecology & Evolution*, *4*, 230–239.

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

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292.

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M. M., & Jetz, W. (2014). Eltontraits 1.0: Species-level foraging attributes of the world's birds and mammals: Ecological archives e095-178. *Ecology*, *95*(7), 2027–2027.

Zipf, G. K. (1936). *The psycho-biology of language: An introduction to dynamic philology*. Routledge.