Birds and Words: Exploring environmental influences on folk categorization

Anonymous CogSci submission

Abstract

How do we name things? What role does frequency of observation and physical size play in categorization of animals? Here we explore these questions using ideas from anthropology and ethnobiology, and utilizing large-scale citizen science datasets.

Keywords: ethnobiology; categorization; 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). 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.

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. 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 freqeuncy. 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.

Environmental factors in ethnobiological classification

Environmental factors have played a major role in scientific classification of species (Amadon, 1943; Hunn, 1999).

Here we talk about the data we use. We show how to utilize publicly available digitized information to explore these questions.

Language naming data

We focus on a single language for brevity. We use bird-naming data from (Hunn, 2008), also found online[1], a Zapotec language spoken in a small village in San Juan Gbëë, Oaxaca Mexico.

Describe the dataset and what we extracted from it (e.g., basic-level, terminal-level names, prototypes, etc.) in more detail here.

For example, the folk-generic name for hummingbirds is *dzǐng*, which covers 13 birds in Zapotec. Of these 13 birds, there are 3 other labels besides *dzǐng*, there are folk-specifics *dzǐng-dán-yǎ-guì*, *dzǐng-gué*, *and dzǐng-yǎ-guì*.

Frequency data

We utilize a citizen-based bird observation network, eBird (Sullivan et al., 2009). We sampled bird observational data

---

[1] http://faculty.washington.edu/hunn/zapotec/z5.html

from just the region containing the state of Oaxaca, Mexico[2]. This resulted in XXX entries.

Physical Size

Bird weights as an aid in taxonomy (Amadon, 1943).

We'll also look at bird size as a factor, following (Hunn, 1999) and using data from EltonTraits (Wilman et al., 2014). This data set provides information on key attributes for all 9993 and 5400 extant bird and mammal species, derived from key literature sources. Variables include relevance of select diet types and foraging strata, body size, and activity time.

We focus on the body mass data, separately sourced from (Dunning Jr, 2007), which is measured as the geometric mean of average values provided for both sexes.

What categories are given a name?

We first look at the distributions of birds with and without names in Zapotec, using the two environmental factors, in relation to all birds seen in OAX.

Frequency

Here we analyze the frequencies of birds named in Zapotec. We plot the densities of birds named and unamed birds, along with all birds observed in the state of Oaxaca in Figure 1.
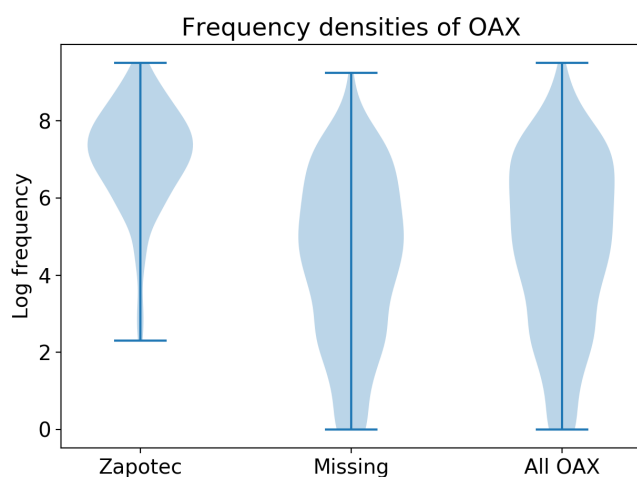

Figure 1: Frequency densities of birds named in Zapotec and those observed in the state of OAX.

Size

Here we analyze the masses of birds named in Zapotec. We plot the densities of birds named and unamed birds, along with all birds observed in the state of Oaxaca in Figure 2.

We see here that frequency is more informative than mass in predicting which birds in OAX are given a name in Za-
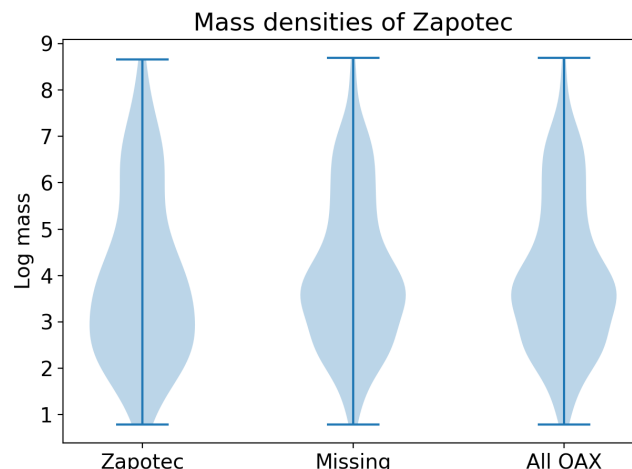

Figure 2: Mass densities of birds named in Zapotec and those observed in the state of OAX.

potec. In the next section, we will find a different set of results.

Hunn Analysis

Here we re-explore Hunn's analysis of how size affects the perceptual salience for folk biological classification, following the chapter (Hunn, 1999). In sum, he demonstrates a positive correlation between the groupings of named categories, and the average size of those organisms in those categories.

Hunn proposed that the size of an organism is influential in folk categorization and developed a measure which examines the degree to which the organisms are recognized taxonomically in the folk classification. Typical use of the measure considers how many scientific species are included within each folk-generic or folk-specific level taxon. Hunn called this the Scientific Species Recognition Ratio (SSRR).

Treating each scientific species as a "sampling unit" (the single species point method), calculating an average size for each species, then calculating the SSRRS of each species according to a slightly variant procedure.

The SSRR of a species this procedure is 1 if it corresponds 1:1 to a basic folk taxon, it is 0.5 if it is one of two species included within a single basic folk taxon; it is 0.33 if it is one of three such species; and so on.

We analyze this using both mass and frequency as predictors. Figure 3 presents SSRR plots of folk-generic names (top row) and folk-specific names (bottom row) for both mass (left column) and frequency (right column).

Here we find that mass rather than frequency is a better predictor, opposite of the results in the previous section.

---

[2]We use eBird observation of frequency from the Basic Dataset (EBD) on https://ebird.org/data/download, last accessed January 24, 2020.
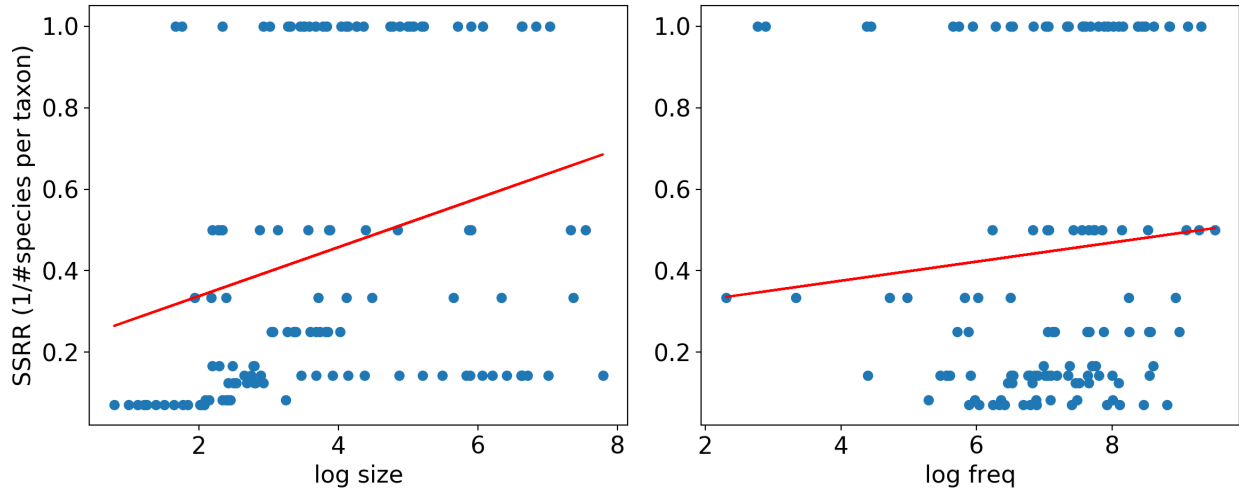
Figure 3: SSRR plots of single-species for both mass (left column) and frequency (right column).

Analysis of name-forms

Name-length

Here we analyze the frequencies of birds named in Zapotec in relation to the name length of the bird. See Figure 4
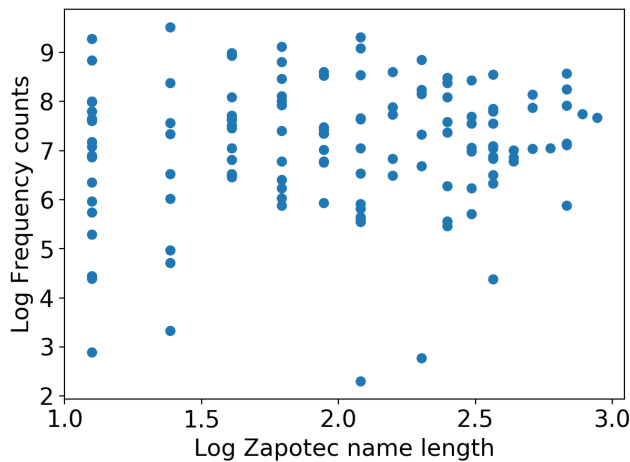

Figure 4: Frequency densities of birds named in Zapotec as a function of name length.

Compound names

Here we further examine names based on whether the Zapotec label is a single word (a monomial) or a compound of multiple words. First we examine frequencies, in Figure 5. Here we see that the monomials tend to be more frequently observed than compounds. The raw mean frequency counts are mono = 2465 and compound = 1715, for monomials and compound names, respectively.

We also explore how masses are distributed based on name form. See Figure 5. Here we see a similar trend as before, with raw mean masses of mono = 375g and compound = 152g.

Prototypes

Here we look at unmarked-prototypes in Hunn's data on Zapotec bird-naming (Hunn, 2008). These are bird names that Hunn noted were mono-syllabic and were recognized at the same level taxa (had the same folk generic and folk specific names).

We can address (Berlin, 1992): "Taxa of generic and subgeneric rank exhibit a specifiable internal structure where some members of a taxon, x, are thought of as being more prototypical of that taxon than others (i.e., are the best examples of the taxon). Taxa of intermediate and life-form rank may also show prototypicality effects. Prototypicality may be due to a number of factors, the most important of which appear to be taxonomic distinctiveness (as inferred from the scientific classification of the organisms in any local habitat), frequency of occurrence, and cultural importance (i.e., salience)."

Here we consider the question: in cases where nomenclature reveals prototype effects – is the prototype highest in frequency?

Following Hunn's data(Hunn, 2008), we analyzed six prototypes; instances in which it was clear there was a single bird labeled as an unmarked prototype. See Figure 6 to see the distribution for each of the prototype families. The top left bar chart shows the frequencies of vultures in Oaxaca, where we see the prototypical Turkey Vulture (in red) is clearly is more frequent. This trend holds across the other instances in the Hunn data, with an exception for the category of Owls (note this category has comparatively low frequency counts).

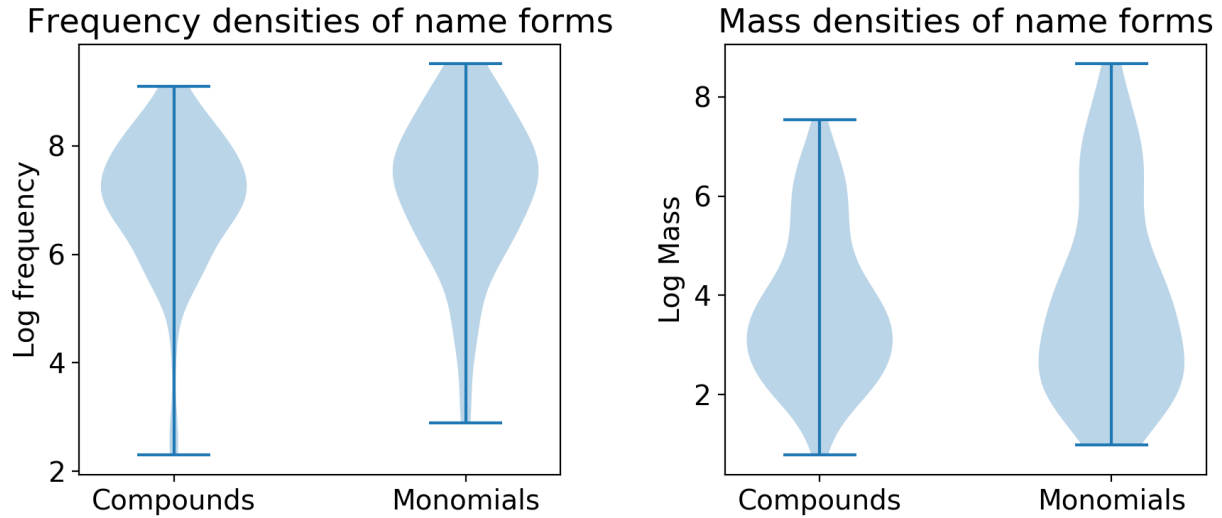We also examined the distribution of the frequencies of

Figure 5: Frequency and mass densities of birds named in Zapotec as a function of name form.
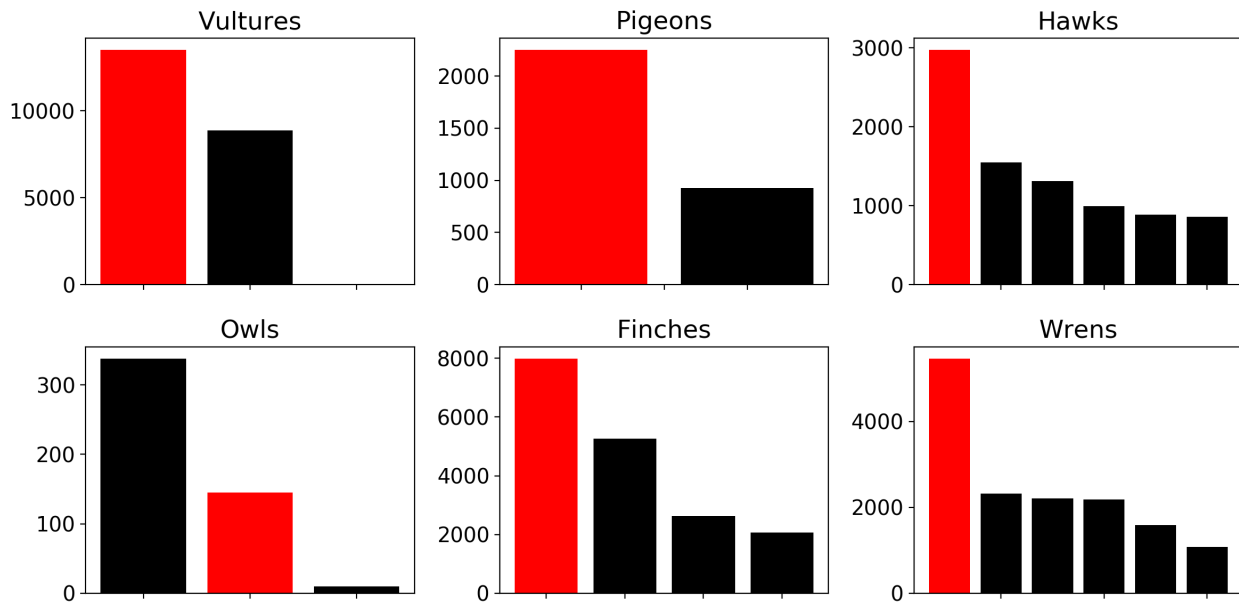

Figure 6: Frequency bar plots of birds named in Zapotec as unmarked prototypes and the frequencies of other birds with the same folk generic name.

birds in Zapotec split into groups based on whether the label is an unmarked prototype or not. We found those that are prototypes are highest in frequency (m=7.93 vs. m=7.02 for log frequency of prototypes and non-prototypes respectively).

Discussion

Summary of some of the questions we could explore using the methods detailed above.

Potential concerns

We address potential concerns that can arise in using eBird frequency of observation data here. Does frequency of observation in eBird accurately represent the statistic of interest? (SOME CITATIONS to back up this claim).

Also, note our observation about the owl prototypes above – this was an interesting insight which could be backed up by adding additional environmental feature data from Elton-Traits (Wilman et al., 2014), which would indicate that these

birds were all nocturnal (and thus potenially more difficult to reflect accurate numbers through eBird).

Also: These questions are interesting because we typically take for granted the categories of natural kinds. However, scientific taxonomies are just another human-constructed category system. When considering the set of birds in particular, it has been difficult for biologists to agree on a standardized taxonomy, which has been shown to severely impact decisions on conservation policy (Peterson, 2006; Garnett & Christidis, 2017).

Future Directions

The next step would be to expand these analyses to more languages. To do this one needs to find trustworthy ethnographies similar to the Zapotec naming data we used here from Hunn (2008), and one needs decent coverage in eBird over the geographic region in question. Clear next steps would be to analyze the Tzeltal language from Chiapas, Mexico, and the Tlingit language from the south-east Alaska, both published by Hunn as well (Hunn, 1977; Hunn & Thornton, 2012), which have decent coverage within their respective geographics regions in eBird observational data.

That said, it can be difficult to find languages with both expert ethnographries of the folk biological naming systems which also have good coverage in eBird. This has prohibited us from exploring bird naming data from known experts in regions with low coverage in eBird (e.g., naming data summarized in (Holman, 2002), including the Tobelo language from Indonesia (Taylor, 1990) and the Anindilyakwa language from Australia (Waddy et al., 1988), which do not have coverage in eBird currently).

Conclusion

References

Amadon, D. (1943). Bird weights as an aid in taxonomy. *Wilson Bull*, *55*(3), 164–177.

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

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

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

Garnett, S. T., & Christidis, L. (2017). Taxonomy anarchy hampers conservation. *Nature News*, *546*(7656), 25–27.

Holman, E. W. (2002). The relation between folk and scientific classification of plants and animals. *Journal of Classification*, *19*(1), 131–159.

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.

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.

Peterson, A. T. (2006). Taxonomy is important in conservation: a preliminary reassessment of philippine species-level bird taxonomy. *Bird Conservation International*, *16*(2), 155–173.

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.

Taylor, P. M. (1990). Folk biology of the tobelo people: A study in folk classification. *Smithsonian Contributions to Anthropology*.

Waddy, J. A., et al. (1988). *Classification of plants and animals from a groote eylandt aboriginal point of view*. The Australian National University.

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.