# Cognitive biases explain constrained variation in noun classification

Ponrawee Prasertsom*[1], Kenny Smith[1], and Jennifer Culbertson[1]

*Corresponding Author: ponrawee.prasertsom@ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, United Kingdom

## 1. Background

A central issue in language evolution is how to explain the apparently constrained variation across the world's languages. Here, we focus on explaining the constrained variation of noun classification (Seifart, 2010): grammatical genders (as in Romance languages), noun classes and classifiers. Cross-linguistically, only some domains, such as animacy, commonly form conceptual bases for classifying nouns (e.g., Swahili noun classes), whereas other potentially salient domains such as colour never do (e.g., no "warm-coloured" classes; cf. Talmy, 1985). Following works linking cognitive biases to typology (e.g., Maldonado & Culbertson, 2022), we hypothesise that this results from a cognitive bias for animacy and/or against colour in grouping nouns. In our pre-registered experiments (OSF project: `https://osf.io/b6yns`), we test if 1) such a bias exists in noun class learning (Exp 1a-1b) and if 2) the bias is specific to language (Exp 2a-2c), as sometimes suggested in the literature (Cinque, 2013).

## 2. Experiments 1a-1b: Artificial noun class learning tasks

In Exp 1a, participants were randomly assigned one of two conditions (Colour and Animacy, N=40 each). In both, they were trained on the artificial nouns through images (Fig 1, left) and audio to criteria (scoring 13 out of 16 at test). Each noun may be animate (a frog/a lizard) or inanimate (a box/a bag), warm- (red/yellow) or cool-coloured (blue/green). Participants then learned two noun classes through determiners that vary based on the noun colour/animacy depending on the condition. The Animacy participants scored higher ($\beta_{\text{Cond.}} = 1.86$, $p = 0.023$, deviation-coded, mixed-effects logistic model), suggesting an animacy bias. The effect is reliable but not large (Prop. correct $0.92 \pm 0.03$ in Animacy vs. $0.84 \pm 0.04$ in Colour), perhaps because the simplicity of the language masked the bias.

In Exp 1b, we use an extrapolation design to determine whether a stronger bias emerges when participants are trained on an ambiguous system and must decide at test whether classification is based on colour or animacy. Here, the procedures remained largely the same, but participants (N=80) were not assigned to conditions. Crucially, during noun class training, the stimuli were compatible with both

Figure 1. Visual stimuli for Exps 1a-1c, 2a, 2c (left) and for 2b (right). Only a subset is shown for 2b. stimuli to illustrate the colours and animacy types.

animacy- and colour-based systems (e.g., animates were always warm-coloured, inanimates always cool-coloured), but critical test trials asked the participants to select the determiner variant for unseen combinations (e.g., warm-coloured inanimates), forcing them to choose the criterion. The results show a strong bias to classify by animacy (Mean prop. animacy-based classification $= 0.78$, $p < 0.001$).

## 3. Experiments 2a-2c: Image sorting tasks

In Exp 2a, participants (N=30) were asked to sort the images used to represent nouns in Exps 1a-1b (Fig 1, left) into two groups. If participants have an animacy bias, they are predicted to prefer sorting by animacy to sorting by colour. This prediction was borne out, with most people (88%) sorting by animacy ($p < 0.001$, Wilcoxon signed-rank test on adjusted mutual information values).

Exps 2b-2c addressed the possibility that the stimuli were not representative of the world: Red and yellow may be unusually different for warm colours, and frogs and lizards too similar for animates. In Exp 2b, we increased the intra-category similarity in the stimuli for colour (e.g., using orange instead of yellow) and decreased it for animacy (e.g., using butterflies and fish instead of frogs and lizards). Fig 1 (right) gives a sample. In Exp 2c, we reduced the stimulus set to one category per domain (e.g., an animate is always a frog, a warm-coloured thing always red), eliminating all intra-category differences. We reproduced the animacy bias in Exp 2b (62% by animacy, 20% by colour, $p < 0.001$), but saw a different pattern in 2c (38.3% by animacy, 61.7% by colour, $p = 0.071$).

## 4. Discussion

Exps 1a-1b showed that an animacy bias exists in noun class learning, and could explain the prevalence of animacy and the absence of colour in noun classification. Exps 2a-2b showed that the bias is not domain-specific; it was also observed in non-linguistic categorisation. The contrast between Exp 2a-2b and Exp 2c results suggests that the animacy bias may result from the fact that an animacy-based classification offers more coherent, clear-cut clusters than a colour-based one *under variability* (Exp 2a-2b). With only one type of animacy/colour, animacy does not offer classificatory advantage (Exp 2c). When combined with the idea that cultural transmissions can amplify soft biases (Culbertson & Kirby, 2016), our results strengthen the explanatory power of cognitive biases in typology.

## References


Cinque, G. (2013). Cognition, universal grammar, and typological generalizations. *Lingua*, *130*, 50–65.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*.

Maldonado, M., & Culbertson, J. (2022). Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, *53*(2), 295–336.

Seifart, F. (2010). Nominal classification. *Language and Linguistics Compass*, *4*(8), 719–736.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (Vol. 3, pp. 36–149).