Lions, tigers and bears: Conveying a superordinate category without a superordinate label

Lilia Rissman (lrissman@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Gary Lupyan (lupyan@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

We asked whether categories expressed through lists of salient exemplars (e.g., *car, truck, boat, etc.*) convey the same meaning as categories expressed through conventional superordinate nouns (e.g., *vehicles*). We asked English speakers to list category members, with one group given superordinate labels like *vehicles* and the other group given only a list of salient exemplars. We found that the responses of the group given labels were more related, more typical, and less diverse than the responses of the group given exemplars. This result suggests that when people do not see a superordinate label, the categories that they infer are less well aligned across participants. In addition, categories inferred based on exemplars may be broader in general than categories given by superordinate labels.

Keywords: categories; concepts; semantics; superordinate nouns; exemplar theory

Introduction

Suppose you want someone to think about a dog. You have a variety of communicative tools at your disposal to achieve this goal — you could start barking, you could draw a picture of a dog, or you could say the word *dog*, for example. While each of these strategies would likely elicit thoughts about dogs, these strategies would likely not elicit *the same* thoughts about dogs — in other words, they would not activate identical conceptual representations. For example, English speakers have been shown to be faster and more accurate at identifying pictures of dogs when they hear labels such as *dog* rather than characteristic sounds (e.g., barking) (Lupyan & Thompson-Schill, 2012). In this paper, we asked whether different cues to a concept result in identical conceptual representations focusing on different linguistic means for communicating superordinate categories. Specifically, we investigated exemplar lists (e.g., *horses, monkeys, dogs, and so on*) as a cue to conceptual knowledge, asking whether such exemplar lists convey the same type of categorical information that is conveyed through a conventional superordinate term (like *mammal*).

Across the world's languages, communicating superordinate categories through lists of salient exemplars is a widely attested strategy (Mauri, 2017; Mauri & Sansò, 2018). Kannada speakers, for example, can use a reduplicative strategy: for example, if *pustaka* ('book') has a reduplicative marker — *pustaka-gistaka* — it conveys 'books and related stuff' (Mauri & Sansò, 2018, ex. 15). For ad-hoc categories, exemplar lists may provide an efficient means of communicating the category (e.g., as in *the job of the royal family is to be symbols of our better selves, to turn up at schools and hospitals and military events and so on*).[1]

To what extent do exemplar lists cue the same conceptual content as superordinate labels? There are at least two reasons why this is an interesting question to pose and answer. First, many superordinates in one language lack translational equivalents in another language (Goddard & Wierzbicka, 2014; Kemmerer, 2019; Mihatsch, 2007; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). For example, while English *seafood* includes fish, its Spanish translation equivalent— *mariscos* — refers to shellfish in particular. Such differences abound. The Chinese superordinate label *tiáowèipǐn* translates to something like "common ingredients used to flavor food" and includes herbs, spices, vinegar, salt, and sugar. Even seemingly basic concepts such as animal and color are not universally expressed as superordinates across the world's languages (Goddard & Wierzbicka, 2014; Kemmerer, 2019; Mihatsch, 2007). ǁGana divides living things not into categories of *plant* and *animal* but into categories such as *kx'ooxo* ('living things which are edible') and *paaxo* ('living things which are harmful to humans') (Harrison, 2007). If English speakers wanted to express *kx'ooxo* or if ǁGana speakers wanted to express *animal*, they could not rely on a single easy-to-translate term. Given such diversity, it is useful to understand whether alternate strategies to communicate superordinate categories, such as exemplar lists, cue identical conceptual content.

Second, making this comparison helps distinguish between different theories of word meaning. In a variety of theories, word meanings are encoded as collections of individual exemplars (Habibi, Kemp, & Xu, 2020; Voorspoels, Vanpaemel, & Storms, 2008). Exemplar lists (e.g., *roses, daisies, petunias, and so on*) and superordinates

[1] https://www.foxnews.com/transcript/lara-trump-we-would-love-to-beat-hillary-clinton-for-a-second-time

(e.g., *flowers*) may have similar semantic representation because a few prominent examples may be sufficient to construct the larger exemplar space. Alternatively, the meanings of exemplar lists and superordinates may differ in systematic ways. Access to verbal labels has been shown to affect memory, category learning and induction (Loewenstein & Gentner, 2005; Lupyan & Thompson-Schill, 2012; Özçalışkan, Goldin-Meadow, Gentner, & Mylander, 2009; Zettersten & Lupyan, 2020). While these and related studies have compared categorization of non-linguistic stimuli with and without verbal labels, we compare different linguistic strategies, testing the hypothesis that exemplar lists and superordinate labels are equivalent modes of conveying categorical information. If they are, it would support the idea that a small set of category exemplars is a good proxy for activating a general meaning, a reasonable inference if both map onto the same conceptual category.

Approach

One way to test the hypothesis that a superordinate meaning can be effectively conveyed through a set of exemplars is to compare speakers of a language that has a superordinate term against speakers of a language that lacks a translational equivalent. This is complicated by the nonlinguistic cultural differences that accompany such linguistic differences. For example, if we found that Chinese speakers had difficulty conveying *tiáowèipǐn* to English speakers, this might be the result of differences in cooking traditions rather than linguistic differences.

In the current study, we avoided this problem by testing a monolingual group of participants whom we expect to be highly familiar with the superordinate categories we are testing. We cue some participants (*label* group) with the superordinate cue (e.g., *vehicles*) and ask them to list 6 typical members. We then use the first three responses as cues to another *exemplar* group of participants and ask them to generate three more responses (see Figure 1). By comparing the qualities of responses 4-6 generated by the two groups, we are able to measure whether a set of salient category exemplars and a superordinate term activate the same conceptual knowledge or whether it is systematically different in the two cases.

We analyzed the groups' responses using three measures: the first is ***relatedness*** – the semantic similarity between the label and the response as measured by distance in word embedding space learned by a model trained on a large English corpus (see below). High relatedness signifies that the label and the response occur in more similar contexts and provides initial evidence that the responses are more central to the superordinate label category. The advantage of this measure is that it can be easily computed for the more than 2000 unique responses we collected. The disadvantage is that differences in relatedness can be obtained for a large number of reasons. We therefore also sought to collect human ***typicality*** ratings for a subset of responses. The typicality ratings allow us to determine whether the responses generated from superordinate labels elicit more or less typical responses compared to those generated from three exemplars. If superordinate labels elicit responses that have higher typicality with respect to the superordinate, this also constitutes evidence that the responses are more central to the superordinate category. Lastly, we quantified the ***diversity*** of people's responses given a superordinate label as opposed to a yoked set of exemplars. One function of superordinate terms may be to align people's mental representations so that they are more similar to each other. This would lead to greater similarity across the responses produced by the label group than across the responses produced by the exemplar group.

Method

Participants

We recruited 174 English-speaking adults on Amazon Mechanical Turk (N_{female} = 89, N_{male} = 82; age range = 20 – 70, median age = 36). An additional 37 participants were tested and excluded for providing repetitive or non-word responses (N = 6) or because they viewed a duplicate set of trials as another participant (N = 31; see Design). Participants received $1.00 for completing the study.

Materials

We tested 20 superordinate labels: *animals, appetizers, chores, clothing, desserts, diseases, flowers, food, furniture, games, hobbies, mammals, pets, plants, tools, toys, vegetables, vehicles*, and *weapons*. We used the following criteria in choosing these labels: inclusion of both natural kinds and artifacts, inclusion of labels that have been studied in previous research on superordinates, and inclusion of labels that are familiar and have many members (cf. a label like *precipitation*).

Design and procedure

Participants were assigned either to the Label condition or the Exemplar condition (see Figure 1). Participants in the Label condition were told that they would be given a word naming a category and would need to list six members of this category. As examples, they were given the labels *colors* and *beverages* along with some typical category members (e.g., *red, blue, green, yellow, orange, pink*). Each participant in the Label condition listed six category members for each of 10 superordinate labels, selected randomly from the total set of 20 and presented in a random order. Each participant in the Exemplar condition was yoked to a participant from the Label condition, viewing the first three category members listed by the participant from the Label condition. Participants in the Exemplar condition were told that they would be given three words that are members of a category and they would need to list three more members of the category. They were given the same examples of *colors* and *beverages* as participants in the Label condition, for example "if you were given the words *red, blue*, and *green*, you might list *yellow, orange*, and *pink*." Each Exemplar participant was yoked to a specific Label

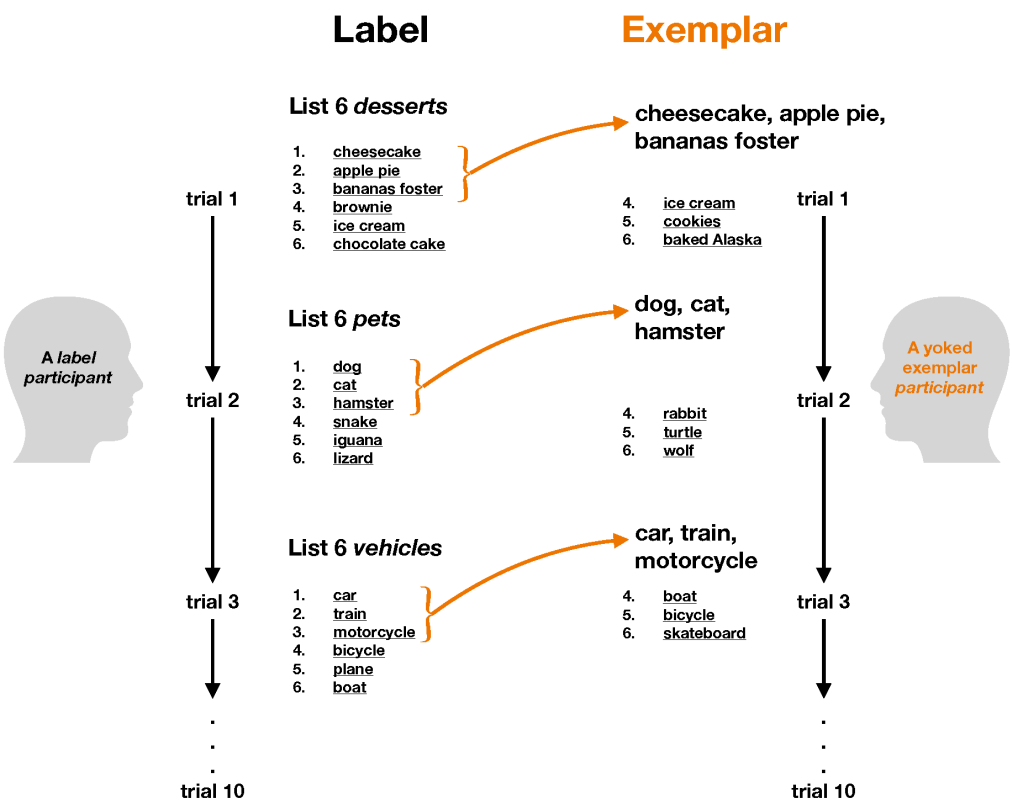

Figure 1. Schematic representation of design with sample participant responses.

participant. Participants in both conditions were instructed to list the first category members that came to mind.

Data preprocessing

We standardized spelling and inflectional variants to reduce minor variability. For example, *action-figure* and *action figure* were replaced by with *action figures*; *cleaning the room* was replaced with *cleaning room*. We retained the variant that was most common across all responses.

Label-to-response relatedness

We computed relatedness between the label and response using word embeddings trained on English Wikipedia+Statmt news corpus using the fast-text algorithm (300 dimensions with subwords) (Bojanowski, Grave, Joulin, & Mikolov, 2017).[2] Using these embeddings, we computed the cosine distance between the cue and each response. One shortcoming of this method is that pre-trained embeddings are available only for single words whereas 14% of our responses were multi-word (e.g., "heart disease"). For these, we treated the response as the vector sum of the content words. This simple procedure produces surprisingly good representations of compound words such as those we are dealing with here (Boleda, 2020). As additional verification, we replicated all our analyses using the subs2vec embeddings derived from movie and TV show subtitles (van Paridon & Thompson, 2020) and contain entries for many compound words. The results, omitted here for brevity, were nearly identical to those obtained using our original method.

Typicality ratings

We collected typicality ratings from 134 English speakers on Amazon Mechanical Turk. Given the large number of response types produced in the main experiment, we collected typicality ratings for a subset of label/response pairs. We sampled the label/response pairs by first dividing the responses for each label into three categories: responses that were produced in the Label condition only, responses produced in the Exemplar condition only, and responses produced in both conditions. We then used the semantic distance metric from the previous analysis to divide each

[2] wiki-news-300d-1M-subword.vec.zip available from https://fasttext.cc/docs/en/english-vectors.html

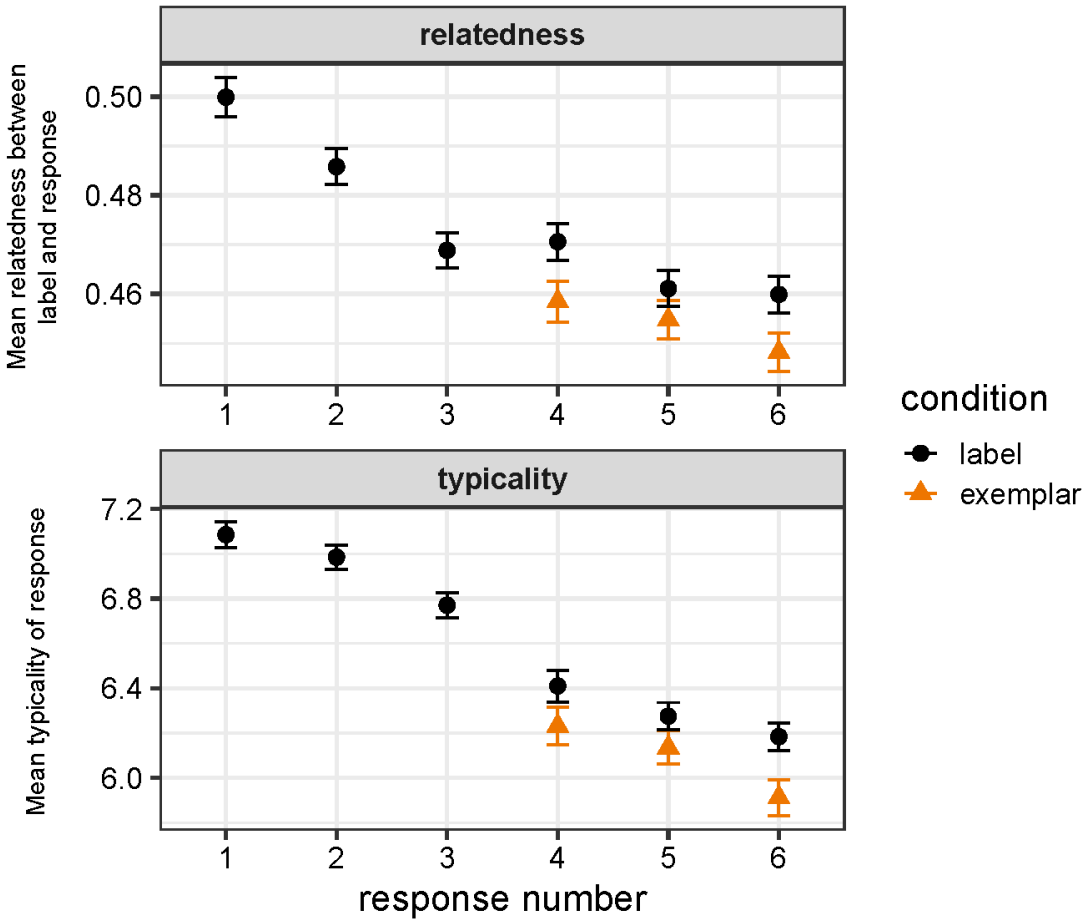

Figure 2. Mean relatedness and typicality scores by condition and response number. Error bars show within-subject standard error of the mean.

group of responses into three tertiles: responses with high, medium, and low similarity to the label. We randomly selected four responses from each tertile from each group, resulting in 12 responses per group and 36 responses per label. If there were four or fewer responses in each tertile, all responses were tested. Participants were asked "How typical is example [response] of the category [label]?" and rated each label/response pair on a scale from 1 to 8, where 8 corresponded to Very Typical, 2 to Not typical at all, and 1 to Not an example. Participants were given the example: "if the category is 'sports', you would probably rate football as more typical than lacrosse." Participants viewed two labels chosen at random from the total list of 20, rating all of the responses for each label in two blocks (e.g., a *plants* block followed by an *appetizers* block).

Results

Figure 1 shows example responses from the Label and Exemplar conditions. We compared the fourth, fifth, and six responses produced by participants in the Label condition against the responses produced by participants in the Exemplar condition. Across conditions and participants, an average of 89 responses per label were produced (range across labels = 42 – 156).

In the analyses below, we used mixed-effects regression models modeled using R (R Core Team, 2017) and the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014). We used the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) and Satterthwaite approximation to compute *p*-values for fixed effects (see Luke, 2017). For the relatedness and typicality analyses, we fit baseline models with by-subject and by-label random intercepts and response number-by-subject random slopes. For the diversity analyses, we fit baseline models with by-label random intercepts.

Relatedness analysis

As shown in Figure 2 (top), relatedness between the superordinate label and the responses decreased as participants produced more responses (b = -.008, SE = .0006, $t(4016)$ = -13.2, $p < .001$). Restricting our analysis to 4-6 — the responses we can directly compare between the Label and Exemplar conditions — reveals that relatedness progressively decreased for these responses as well (b = -.005, SE = .001, $t(3701)$ = -3.45, $p < .001$). Crucially,

responses were more related to the superordinate in the Label condition than in the Exemplar condition (b = .01, SE = .003, $t(174) = 3.01$, $p < .01$). That is, responses 4-6 produced by people who were cued with the superordinate label were more related to the label than responses 4-6 produced by people who were cued with responses 1-3. There was no significant interaction between response number and condition (b = -.0003, SE = .003, $t < 1$).

Typicality analysis

We next assessed whether the responses produced in each condition differed in how typical they were of the originally cued label (e.g., how typical is the response *action figures* of the category *toys*?). Figure 2 shows mean typicality of responses with respect to the superordinate across condition (Label vs. Exemplar) and response number (1 through 6). As with the relatedness analysis, we found that initial responses were more typical than later responses (b = -.20, SE = .015, $t(261) = -13.5$, $p < .001$). This decrease remains highly reliable when including only responses 4-6 (b = -.13, SE = .03, $t(1660) = -3.94$, $p < .001$). As with relatedness, we found that cueing people with the superordinate label yields more typical responses (4-6) than the analogous responses cued by an exemplar list (b = .24, SE = .08, $t(178) = 2.96$, $p < .01$). In addition, raters were more likely to judge that a response was 'Not an example' of the superordinate label when the response was produced in the Exemplar condition (b = .026, SE = .0068, $t(182) = 3.84$, $p < .001$). We found no significant interaction between response number and condition (b = .06, SE = .06, $t(2045) < 1$).

As expected, human typicality ratings were correlated with our word-embedding based relatedness measure, though only moderately (rho = .32, p < .001) suggesting that (again, as expected) the two measures capture meaningfully different types of relations. Interestingly, typicality of exemplar-produced responses was lower even when controlling for relatedness (b = -.20, SE = .08, $t(177) = -2.6$, $p < .01$) and relatedness was lower for exemplar-produced responses even when controlling for typicality (b = -.009, SE = .003, $t(162) = -2.5$, $p < .05$).

Response diversity analysis

Cueing people with lists of exemplars led to less related and less typical responses than cueing people with a superordinate cue even though everyone in our study is familiar with the superordinate categories we test here. This shows that explicit use of superordinates is more effective at evoking the category than what can be evoked through a small set of exemplars. In this last analysis, we examined whether superordinates help people *converge* as measured by their producing more similar responses than when ostensibly the same category is cued through exemplars.

We compared response diversity using measures: 1) the proportion of responses that are unique, 2) modal response agreement, i.e., the proportion of responses that match the modal response(s), and 3) Simpson's diversity index D (see Majid et al., 2018; Zettersten & Lupyan, 2020 on previous use of this index). D-values range from 0 to 1, with 0 corresponding to no overlap in responses and 1 to complete homogeneity in responses. Each of these measures was computed for each of the 20 labels in each condition.

Mean values of each measure were comparable across the Label and Exemplar conditions (proportion unique responses: .43 vs. .47; modal response agreement: .085 vs. .084; Simpson's diversity D: .027 vs. .029, respectively). The raw diversity measures revealed no difference between conditions. However, we unexpectedly found that response length (i.e., number of words) differed across conditions (1.20 vs. 1.14 words per response in the Label and Exemplar conditions, respectively), where response length correlates with our diversity measures. All else being equal, it is harder to converge on a longer response. We therefore included response length as a covariate. Controlling for response length, the proportion of unique responses was lower when participants had seen the superordinate label (b = -.065, SE = .015, $t(26) = -4.28$, $p < .001$). D-values were higher (signaling greater convergence) when participants had seen the superordinate (b = .0049, SE = .002, $t(26) = 2.4$, $p < .05$). Modal response agreement did not differ significantly by condition (b = .0050, SE = .0064, $t(23) < 1$).

Discussion & conclusion

We began by posing the question of whether different cues to a concept activate identical conceptual knowledge. We focused on the contrast between superordinate labels such as *vehicle* and exemplar lists (e.g., *cars, trucks, boats and so on*), the latter being a well-attested strategy for communicating categories in the absence of a conventional superordinate label. Do such exemplars activate the same conceptual knowledge as the superordinate term? Our results suggest that they do not. When people were not given a superordinate label, they listed category members that were less semantically related, less typical, and (controlling for length) more diverse than the category members provided by people who had seen a label. This result is particularly striking because all the participants in this study, as English speakers, had access to the same superordinate nouns. Given an exemplar list like *cars, trucks, boats,* participants could — and likely often did — recode this list as *vehicles*. Nonetheless, the absence of labels appeared to lead participants to infer categories that were less aligned than the categories used by participants who did see labels.

One interpretation of our data is that participants in the Exemplar condition inferred categories that had a broader extension that the categories provided by the superordinate, leading the exemplar-based responses to be less semantically related to the label, less typical of the label, and more diverse. Indeed, of the 37 label/response pairs that received typicality ratings of less than 2.0, 78% of these were produced by participants in the Exemplar condition. Such highly atypical responses suggest that a category broader than the label is being inferred. An alternative possibility is that participants in the Exemplar condition provided more fine-grained,

subordinate-level responses (e.g., *electric car, Chevy Malibu*) than participants in the Label condition, which may lead to the same overall pattern: responses that are less semantically related to the label, less typical of the label, and more diverse. One way to find out whether the responses exhibit over- or under-generalization is to have the responses rated on their generality (Lewis et al., in prep). Such an analysis may reveal that exemplar lists lead to over-generalization for some categories (e.g., more socially constructed categories like *hobbies*) and under-generalization for others (biological kinds).

Our results suggest that participants in the Exemplar condition did not always guess the correct label that generated the responses (otherwise their responses would have been indistinguishable from the responses in the Label condition). This may be because participants always generated guesses but sometimes their guesses were wrong. For example, if a Label participant was given *mammal* and listed *lion, wolf, tiger*, an Exemplar participant may have incorrectly guessed that the category was *animal* and listed *tiger, squirrel, shark*. It may also be that the conditions are different because Exemplar participants did not always generate guesses, i.e., they did not always use a superordinate label to characterize the category they were being given. Evidence for or against these contrasting explanations could be obtained by asking English speakers to guess the category that generated the three exemplars. For example, if a participant were given *lion, wolf, tiger* and guessed "animal," this would indicate that competition between existing superordinate terms is one factor leading to the different patterns of responses we observed between the Label and Exemplar conditions. If, by contrast, participants guessed ad-hoc descriptive categories such as "mammals that like to hunt," this would suggest that participants in the Exemplar condition are inferring categories that do not map onto conventional superordinate labels.

At the outset of this paper, we argued that if exemplar lists activate the same conceptual categories as superordinate labels, this would provide strong evidence for exemplar theories of word meanings. We did not find this supportive evidence, as responses in the Label and Exemplar conditions differed. This result does not on its own invalidate exemplar theories, as it could be that three exemplars is simply too few to infer the correct category. However, if participants in the Exemplar condition were not consistently activating superordinate labels for the categories (as we suggest above), this would present a challenge for exemplar theories. That is, if word meanings are represented as collectives of exemplars, it raises the question of why words would not automatically be activated upon viewing exemplar lists.

How do these results speak to issues around translating superordinate word meanings across languages? Our results suggest that without the ability to rely on a shared superordinate term, the ability to convey the kind of broad category that these terms typically denote is compromised, and to a likely much greater extent than what we see here with speakers of the same language. Providing lists of salient exemplars, no matter how good, may be insufficient. The efficacy with which dedicated superordinate terms convey broader categories is presumably a key reason why they exist (i.e., culturally evolve) in the first place. Our experiments take an initial step toward a more precise understanding of their communicative function and the possibility that superordinates help in aligning and perhaps learning the conceptual categories they denote.

Acknowledgments

This research was supported by NSF-PAC #1734260. Thank you to CogSci reviewers for helpful suggestions on improving this paper. Thank you to all study participants.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using S4 classes. R package version 1.1-7. *URL http://CRAN. R-project. org/package= lme4*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics, 6*(1), 213-234. doi:10.1146/annurev-linguistics-011619-030303

Goddard, C., & Wierzbicka, A. (2014). *Words and Meanings : Lexical Semantics Across Domains, Languages, and Cultures*. Oxford: Oxford University Press.

Habibi, A. A., Kemp, C., & Xu, Y. (2020). Chaining and the growth of linguistic categories. *Cognition, 202*, 104323. doi:https://doi.org/10.1016/j.cognition.2020.104323

Harrison, K. D. (2007). *When languages die : the extinction of the world's languages and the erosion of human knowledge*. Oxford; New York: Oxford University Press.

Kemmerer, D. (2019). *Concepts in the Brain: The View From Cross-linguistic Diversity*. Oxford: Oxford University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive psychology, 50*(4), 315-353. doi:http://dx.doi.org/10.1016/j.cogpsych.2004.09.004

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior research methods, 49*(4), 1494-1502. doi:10.3758/s13428-016-0809-y

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General, 141*(1), 170.

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., . . . Levinson, S. C. (2018). Differential coding of perception in the world's languages.

Proceedings of the National Academy of Sciences, 115(45), 11369-11376. doi:10.1073/pnas.1720419115

Mauri, C. (2017). Building and Interpreting Ad Hoc Categories: A Linguistic Analysis. In J. Blochowiak, C. Grisot, S. Durrleman, & C. Laenzlinger (Eds.), *Formal Models in the Study of Language: Applications in Interdisciplinary Contexts* (pp. 297-326). Cham: Springer International Publishing.

Mauri, C., & Sansò, A. (2018). Linguistic strategies for ad hoc categorization: theoretical assessment and cross-linguistic variation. In *Folia Linguistica* (Vol. 52, pp. 1).

Mihatsch, W. (2007). Taxonomic and meronomic superordinates with nominal coding. In A. C. Schalley & D. Zaefferer (Eds.), *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts* (pp. 359-377). Berlin: Walter de Gruyter.

Özçalışkan, Ş., Goldin-Meadow, S., Gentner, D., & Mylander, C. (2009). Does language about similarity play a role in fostering similarity comparison in children? *Cognition, 112*(2), 217-228.

R Core Team. (2017). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/. from R Foundation for Statistical Computing https://www.R-project.org/.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology, 8*(3), 382-439. doi:https://doi.org/10.1016/0010-0285(76)90013-X

van Paridon, J., & Thompson, B. (2020). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior research methods*. doi:10.3758/s13428-020-01406-3

Voorspoels, W., Vanpaemel, W., & Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic bulletin & review, 15*(3), 630-637. doi:10.3758/PBR.15.3.630

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition, 196*, 104135. doi:https://doi.org/10.1016/j.cognition.2019.104135