Divergence in Word Meanings and its Consequence for Communication

Yuguang Duan (yduan38@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

How similar are people's meanings of common words? Do differences in word meanings lead to miscommunication? We examined divergence between word meanings using similarity ratings such as "Is a penguin more similar to a chicken or a dolphin". We found that given the identical instructions some people prioritized taxonomic relationships more than other people. Moreover, this taxonomic bias generalized from animals to artifacts suggesting a more general difference in semantic organization. When people with different biases are paired in a matcher-director type task, they were less likely to achieve communicative success.

Keywords: individual differences, semantics, word meaning, concepts, features, taxonomic bias, director-matcher

Introduction

At the core of every language is a vocabulary—-a set of building blocks from which language users can construct arbitrarily complex meanings. Each member of a speech community must learn both the form and meaning of each word. It is generally assumed that for a language to function as an effective communication system, both the forms and meanings must be closely aligned among speakers (e.g., Hutchins & Hazlehurst, 2006). But how aligned are people's word meanings? Whereas speakers can directly observe the forms of words, adjusting their pronunciations to match those of others from the speech community, speakers can only infer a word's meaning from the way it is used in context and its effects on other people. Here, we measure differences in meanings of common animals and artifacts among adult English speakers and investigate whether the observed differences lead to communicative consequences. Do people who diverge more in their word meanings have a harder time communicating?

There are several kinds of divergence in word meanings. The first and perhaps most obvious is that words vary by speech community. In a classic paper, Clark (1998) discusses how depending on a speaker's membership in various partially overlapping speech communities (regional, occupational, religious, etc.), the same word-form may have rather different meanings. For example, the word-forms *raise*, *cut*, and *drift*, have commonly known meanings, but in addition have specialized meanings in mining (e.g., a *raise* as a vertical shaft connecting two levels of a mine).

A second, more surprising type of divergence concerns differences in extensions of common (non-specialized) words even (seemingly) within the same speech community. Consider the question "Have you smoked at least 100 cigarettes in your entire life?" taken from the Tobaccoo Use Supplement survey administered by the US Census Bureau and used as a branching question, determining whether respondent's should be asked more detailed questions about their smoking. We are tempted to assume that differences in responses have a clear mapping onto differences in the responder's smoking experience. And yet, it turns out that people's meanings of both "smoking" and "cigarettes" diverge to a surprising extent (Suessbrick, Schober, & Conrad, 2000). In followup questions, about half indicated that when answering the question they were thinking of smoking as referring to puffs they inhaled. The other half included all puffs whether inhaled or not. Twenty-three percent indicated that they counted only cigarettes they finished. Fifty-four percent indicated that they counted a cigarette that they took a single puff off as having been "smoked". Similar divergence is observed for what counts as a cigarette: tobacco, yes, but what about hand-rolled, clove, marijuana? For the 37 concepts the authors studied in this way, only 51% of respondents endorsed the majority interpretation. Our own work confirms these kinds of differences for more abstract words (Lupyan et al., 2022). For example, the word "pressure" has physical and social meanings that are part of everyday non-specialized language (compare tire pressure, peer pressure). When asked to endorse a variety of meanings of "pressure", 35% favored glosses like "force that pushes toward a single point on an object" and "force inside or outside pushing on something". The majority (65%) also endorsed these physical senses (though to a lesser extent), but gave much stronger endorsement to social senses like "manipulate someone's actions or thoughts". In both Suessbrick et al. (2000) and our own work, we see that divergence is both present and principled. It's not that some people think cats are dogs. Rather, people are associating words with somewhat different parts of meaning space.

The two types of divergence just discussed focus on specific word-forms (or lemmas). For words like "smoking", differences in meaning may derive in a transparent way from differences in personal experience. Someone who only smokes tobacco might only think of tobacco cigarettes when asked about smoking. And so there would be that differences in the meaning of "smoking" have anything to do with differences in the meaning of "pressure". But another uncovered

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

type of word meaning divergence raises the possibility that some differences may stem from differences in more general cognitive processes. In a recent study, Martí, Wu, Piantadosi, and Kidd (2023) assessed word meaning differences through similarity judgments. People were asked to indicate whether, e.g., a penguin was more similar to a chicken or a dolphin, a robin or a salmon, etc. People's responses varied. For example, 55% of people indicated that a seal was more similar to a penguin than to a whale. A non-parametric, Bayesian clustering model ("Chinese restaurant process", Pitman, 1995) was then applied to the judgments, showing that people's responses were best accounted for by positing different underlying concepts of (in this case) "penguin". Once again, there is divergence, but it is principled rather than haphazard. A possible basis for this divergence may stem from some people prioritizing more abstract aspects of meanings (e.g., the taxonomic class of an animal) while other people prioritize more situational aspects (e.g., an animal's typical habitat or other thematic relationships). If true, we should find that meaning divergence may be quite principled: someone with a stronger taxonomic bias may have systematically different word meanings [1].

The studies below sought to answer two main questions. First, how generalizable are word meaning differences of the sort revealed by similarity-rating tasks in Martí et al. (2023) (also see Hill, Reichart, & Korhonen, 2015; Phillips & Boroditsky, 2003; Waxman & Senghas, 1992)? For example, do people who differ in their judgements of whether a seal is more similar to a whale or a penguin differ in their knowledge of seals (the semantic representation of *seal*)? Perhaps for some people, the most salient thing about a seal is that it lives in cold water, making it semantically similar to a penguin which they think of as likewise living in cold water. If so, how one responds to a seal should not predict how one responds to a question about chickens or leopards. Alternatively, divergence in similarity ratings may derive from more general differences in semantic organization. For example, for some people, the organization of animal concepts (at least as instantiated by English words) may be primarily taxonomic. For others, taxonomic relatedness may be less important than other types of similarity such as habitats and behavior. This question is addressed in Experiments 1 and 2.

Our second question is, do such differences in word meanings have consequences for communication? It is entirely possible, for example, that even if people have stable and generalizable differences in whether they rate this or that words as being more similar, these differences only reveal themselves in explicit rating tasks of the sort we use here, but are of no consequences during actual language use. This question is addressed in Experiment 3.

[1] and by extension, concepts, though we will stick with discussing word meanings here. For related discussion, see Casasanto and Lupyan (2014).

Experiment 1: Individual differences in similarity ratings

Participants rated words on similarity using a triad task. On each trial, participants indicated whether a target word was more similar to one choice or another. We then clustered people's responses and assessed whether the clusters differed in reliance on different types of similarity. For example, did some people consistently rely on taxonomic similarity and other people on thematic similarity? It is well-known that participants' reliance on taxonomic and thematic similarity is flexible, varying both by word-type and subtle differences in the provided instructions (Lin & Murphy, 2001). Here, however, the instructions were identical for all.

Methods

Participants We recruited fifty-four participants were recruited from Amazon Mechanical Turk. Six participants were excluded for failing more than one catch trials.

Materials We used McRae, De Sa, and Seidenberg (1997) (also see McRae, Cree, Seidenberg, & McNorgan, 2005) feature norms, which contain 2526 unique features (e.g, flies, made of wood) collected from 725 participants for 541 living (e.g., *sparrow*) and nonliving (e.g., *chair*) concepts. The original participants (n=725) were asked to list different types of features for the concepts the words referred to, such as physical (perceptual) properties (how it looks, sounds, smells, feels, and tastes), functional properties (what it is used for and where and when it is used), and other facts about it, such as the category it belongs in (e.g, is a bird) or other thematic facts (such as where it is from). Cree and McRae (2003) then classified features into ten knowledge types: three corresponding to visual information (visual–color, visual–form and surface properties, and visual–motion), four corresponding to other primary sensory-processing channels (smell, sound, tactile, and taste), one corresponding to functional/motor information regarding the ways in which people interact with objects (function), one corresponding to taxonomic categories (taxonomic), and one corresponding to all other knowledge types consisting of habitats and other thematic features, e.g., holiday-relevant features (thematic). Some example feature norms are listed in Table 1.

Table 1: Examples of feature norms.

Concept	Feature	Type
chicken	A bird	taxonomic
chicken	Lays eggs	thematic
chicken	Lives on farms	thematic
chicken	Has wings	Visual-form and surface
chicken	Has a beak	Visual-form and surface
chicken	Is edible	function
chicken	Eaten by frying	function

Following Martí et al. (2023)'s work, we began by examin-

ing divergence in similarity for words denoting animals. We first conducted a preliminary experiment to examine which types of feature contribute to explaining variance in people's similarity ratings. We computed cosine similarities for each unique pair of animal words in McRae et al. (2005)'s dataset using each type of feature vectors (e.g., each animal word has a thematic feature vector consisting of Boolean values for all thematic features). The cosine similarities from different types of features were used to define the "gold standards" of feature-type similarity for each word pair [2].

There are different types of features predicting different similarity for the same word pairs, and some feature types confounded with each other (taxonomic and visual features) so they may not explain the variance independently. Our preliminary experiment showed that only taxonomic features and thematic features independently contributed to explaining variance in individual's similarity ratings, which suggests if there are systematic differences in people's feature reliance, they are most possibly about these two types of features. We therefore designed 150 triads that pit taxonomic similarity against thematic similarity, as shown in Table 2.

Table 2: Examples of animal word triads.

Center Word	Taxonomic Pair	Thematic Pair
beaver	chimp	turtle
cow	caribou	rooster
gorilla	sheep	rattlesnake
lion	horse	alligator
lamb	rat	rooster

Procedure Following (Martí et al., 2023), we generated similarity rating tasks on triads of animal words. A sample trial is shown in Figure 1. We used a continuous response scale allowing participants to indicate how much more similar they thought the target word was to choice 1 vs. choice 2.

Participants saw the following instructions:

In this part, you will be asked to indicate how similar a target word is to two comparison words. For example, you may be asked to indicate whether the word dog is more similar to cat or horse. In this case, you would probably think it's more similar to cat and so you would move dog toward cat and away from horse. Please try to use the whole scale, so if you think cat is only slightly more similar to dog than to horse, you can put it just a bit toward dog rather than all the way.

Figure 1: An example of similarity rating task for animals.

Five catch trials were inserted in random order, e.g., "Tiger is more similar to Tiger or Turtle?". Participants needed to place the the slider towards the correct answer in catch trials within a distance of 15 on the scale. Those who failed more than one catch trial were excluded.

Analysis To see if people made use of different kinds of similarity, we developed a method called *iterative linear regression*: for each participant's responses, we selected a feature type whose "gold standard" similarity explained most of the variance in the participant's ratings and added it to the linear regression model as the first predictor. We then regressed the other feature type on it and computed the residual, which is added as the second predictor to the model. Both predictors were scaled at the beginning and after residualization. This method can scale to more features. While this method is very similar to the stepwise linear regression, we added a residualization process to make sure predictors were orthogonal to each other. We took the order that each feature type being added to the model as indicators of their importance because it reflects how much variance in individual's data can be explained by each type of feature.

To evaluate response consistency we performed 10-fold cross-validation. For evaluation, we used two metrics: (1) we first computed the root mean squared error (RMSE) between the model prediction and the true rating values, and then (2) computed the categorical accuracy by checking how many model predictions suggested the same direction of choices as the real human ratings (either the left animal word or the right animal word). We used t-test to compare each person's evaluation results with a baseline where the cross validation was done on 150 random ratings that are inconsistent on feature reliance.

We then re-trained an iterative linear regression model on all the data from each individual to check how much variance in the data can be explained by taxonomic and thematic features respectively. Then we clustered participants according to their variance explained by the two feature types using K-means clustering method, which partitions multi-dimensional vectors into k groups according to their Euclidean distance in the multi-dimensional space (Hartigan & Wong, 1979). And the best value of k can be computed through Silhouette method (Rousseeuw, 1987) which in our case was two.

Results

Thirty-five participants had significantly lower RMSE than the random baseline, and 36 had significantly higher categor-

[2] A complication with using feature norms to derive taxonomic similarity is that participants often do not list taxonomic features of animals and when they do, they use them idiosyncratically. For example, taxonomic-labeled features for "pony" included: <an animal>, <a horse>, and <a small horse>, but omitted the feature <a mammal>). In contrast, taxonomic features for "zebra" were <an animal>, <a mammal>, and <a horse>. To address these idiosyncrasies and to simplify the taxonomic-based similarity prediction, we coded each animal into one of six taxa: <mammal>, <bird>, <reptile>, <fish>, and <amphibian> and <a shellfish>.

ical accuracy than the random baseline. That is, 67% people consistently relied on either taxonomic or thematic features to judge word similarity. Some participants were inconsistent in their responses possibly because they were relying on unmodeled dimensions of similarity or relied on different features for different animals.

As shown in Figure 2, whereas most participants relied strongly on taxonomic similarity, others relied far less on taxonomic similarity and slightly more on thematic features. This finding partially answers to our first question – there are principled differences in patterns of people's similarity ratings suggesting somewhat different underlying word meanings. In particular, the differences are partially explained by different taxonomic bias in how people respond and these differences in taxonomic bias are consistent for individuals rather than the divergence in responses being driven by differences in specific words. However, until now our finding is limited to animals. The next experiment examines whether a preference for a certain type of similarity in judging animals extends to judgments of artifacts.

Figure 2: Clusters for trials pitting thematic and taxonomic dimensions. Error bars are 95% confidence intervals.

Experiment 2: Generalizing taxonomic bias to artifact words

Do people who prioritize taxonomic information only do so for concepts that have clear taxonomic organization such as animals? Or does this bias also predict their organization of other semantic domains? We investigate this by reproducing Experiment 1 with both animals and artifacts. We predicted that participants who prioritize taxonomic relations for animals would also do so for artifacts. Here, following Landrigan and Mirman (2016), we define taxonomic similarity for artifacts as"looking alike or belonging to the same category", and thematic similarity as "occurring in the same time or place even if they are perceptually dissimilar". While the taxonomic similarity for artifacts may not be as clear as that for animals, this definition of taxonomic similarity can apply to both artifacts and animals.

Methods

Participants We recruited 86 participants from Amazon Mechanical Turk. Six participants were excluded for failing more than one catch trials. We analyzed the data from the rest 80 participants.

Procedure Since there isn't enough taxonomic data for artifacts in McRae et al. (2005)'s dataset, we obtained similarity ratings between artifacts and their taxonomically and thematically related words from Landrigan and Mirman (2016). The ratings are based on a 7 likert scale. The instructions tell participants to rate two words as *taxonomically* similar if they look alike or belong to the same category and to rate them as thematically similar if they are *thematically* connected or related or if they occur in the same time or place even if they are perceptually dissimilar. For example, *dots* and *stripes* are *taxonomically* similar because both are types of patterns or designs, while *shirt* and *stripes* are *thematically* related because stripes are often found on shirts.

We selected 24 animal triads from Experiment 1 that were especially diagnostic of participants' taxonomic bias for animals. We then created 20 new artifact triads from Landrigan and Mirman (2016)'s dataset, pairing each target word with a taxonomic and a thematic choice. The triads were created to ensure that the taxonomic-based similarity of the taxonomic choice was as strong as the thematic-based similarity of the thematic choice. For example, in the triad asking participants to indicate whether *train* was more similar to *platform* (thematic choice) or *motorcycle* (taxonomic choice), the taxonomic similarity between *train* and *motorcycle* (M=5.38, SD=1.56) was similar to the thematic similarity between *train* and *platform* (M=5.78, SD=1.17). Conversely, the thematic similarity between *train* and *motorcycle* (M=2.72, SD=1.9) was similarly low to the taxonomic similarity between *train* and *platform* (M=2.14, SD=1.46). Animal and artifact trials were blocked and the block order was counterbalanced between participants. Example triads are shown in Table 3.

Table 3: Examples of artifact word triads.

Center Word	Taxonomic Pair	Thematic Pair
train	motorcycle	platform
vase	bucket	bouquet
int	paint	printer
spaghetti	rice	fork
football	basketball	kick
menu	map	recipe

Procedure Same as Experiment 1.

Analysis As the trials for both animal words and artifact words were designed to pit taxonomic similarity and non-taxonomic-based similarity (thematic dimension for animal words and thematic dimension for artifact words), we can get a person's taxonomic bias for each trial by looking at how

much an individual's rating is directed toward the taxonomic choice (the distance between center and the slider). The larger this score, the higher the taxonomic bias. Because for animal trials, the taxonomic similarity and thematic similarity were not balanced, we adjusted the the center before computing taxonomic bias, so that the center is equally distant from the gold standard slider position predicted by taxonomic features and that predicted by thematic features. We performed the same computation on artifact trials, and examined whether each individual's average taxonomic bias on animal trials correlates with that on artifact trials.

Results

People's taxonomic biases on animal trials were moderately correlated with their taxonomic biases on artifact trials (Pearson $r = 0.33$, $p < 0.01$), figure 3.

Since both the animal word experiment and the artifact word experiment were designed to contrast taxonomic and thematic choices, this correlation also means that the thematic biases are correlated. This suggests that the differences we observed in both experiments actually come from prioritizing either taxonomic or thematic relations in meaning.

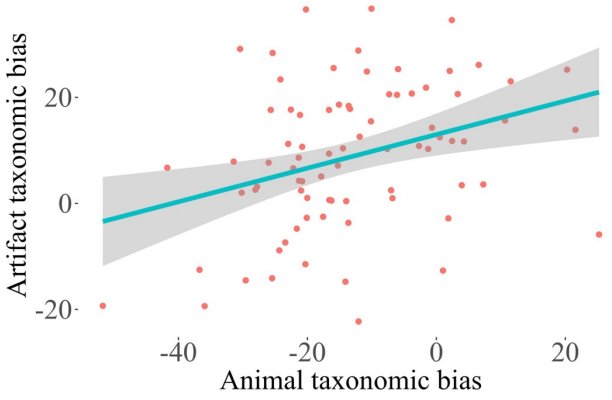

Figure 3: The relationship between the taxonomic bias for animal and artifact trials. Each point represents a participant.

Experiment 3: Do differences in similarity biases impact communication?

Although finding stable differences in participants' patterns of similarity ratings is interesting, these results are more interesting if the differences have behavioral consequences beyond similarity ratings. If different patterns of similarity ratings are caused by participants having (somewhat) different conceptual representations of the target words (and/or the choices), we may observe consequences of these differences in a communication task. As one test of this basic prediction, we had pairs of participants play a cue-target guessing game (Sulik & Lupyan, 2018). One participant (the cuer) was shown a word and asked to produce three cues that would allow another participant to guess the target word. The second participant (the guesser) was shown the generated cues and had to guess the word that generated them. We then predicted how successful each cuer-guesser pair was from the strength of the taxonomic bias of the cuer and guesser.

Methods

Participants We recruited 56 participants from Amazon Mechanical Turk to play the role of a cuer. Seven were excluded because of failing more than one catch trial, and another 1 was excluded because they responded with whole phrases instead of words. We then recruited an additional 163 participants to serve as guessers. Their job was to guess the target words using the cues generated by the cuers. Nine guessers were excluded because of failing more than one catch trial.

Materials and Procedure We selected 10 animal and 10 artifact words from the previous experiments to serve as targets. We first recruited a group of participants to act as cuers. Each participant was shown 20 words and asked to generate three cues for each word such that another person could guess the word by looking at the cues.

After collecting these data, we corrected spelling errors, and lowercased and lemmatized the responses. We then used these responses to generate lists of cues shown to the guessers who attempted to guess the word the cuer saw.

Examples are shown in Table 4.

Table 4: Examples of cues provided in response to select target words

Target Word	Cue 1	Cue 2	Cue 3
bat (the animal)	dark	wings	nocturnal
crocodile	Nile	reptile	river
vulture	bird	beak	scavenger
leash	dog	rope	walk
truck	loaded	wheels	heavy
vase	glass	water	flowers

Following the cuing or guessing task, we measured each participant's taxonomic bias by averaging their responses to 17 animal similarity rating trials and 10 artifact similarity rating trials from Experiment 2 that had no overlapping words with the words used in the cuing task. The cuing/guessing trials and similarity ratings trials were blocked. Block order was counterbalanced between participants.

Each guesser was paired with a high taxonomic-biased cuer and a low taxonomic-biased cuer for both animal trials and artifact trials, allowing within-subject comparisons. Each set of cues from each cuer was guessed by at least three guessers to guarantee enough data for the following analysis.

Analysis We used a mixed-effects logistic regression model to regress guess accuracy on cuers' taxonomic bias score, guessers' taxonomic bias score, and the absolute difference between their taxonomic biases, i.e., `accuracy cuers taxonomic bias + guessers taxonomic bias`

`+ absolute difference`. In particular, we examined 2 types of taxonomic bias: 1) animal taxonomic bias, 2) artifact taxonomic bias. We separated the animal and artifact taxonomic biases because we suspected that they may be differentially predictive of communication success. We also included by-cuer, by-guesser, and by-target word random intercepts, as well as by-cuer and by-guesser random slopes for trial types (animal or artifact), i.e., `(1 + trial type | cuer) + (1 + trial type | guesser) + (1 | target word)` because those random effects could also be a source of variance in the data.

By looking at the fixed effects, we could tell whether any types of cuer's, guesser's, or absolute differences in taxonomic bias influence people's coordination in communicative tasks.

Results

As shown in Figure 4, a higher taxonomic bias was associated with higher guess accuracy. This relationship held for both cuers (b = .35, $p < .05$, 95%CI [.08,.63]) and guessers (b = .15, $p < .05$, 95%CI [.001,.3]). Only the artifact taxonomic bias was significantly predictive of accuracy, but it predicted accuracy for both (other, held out) artifact trials and for animal trials. Holding constant the mean taxonomic bias between cuer and guesser, the more similar cuer and guesser's biases were, the higher the accuracy (b = -.13, $p < .05$, 95%CI [-.26,0]). The effects of the difference between cuer and guesser's biases are further unpacked in Figure 5. We see that cues produced by cuers with a stronger taxonomic bias lead to greater accuracy, but only if the guesser also had a high taxonomic bias (i.e., the absolute difference between guesser's and cuer's biases was small).

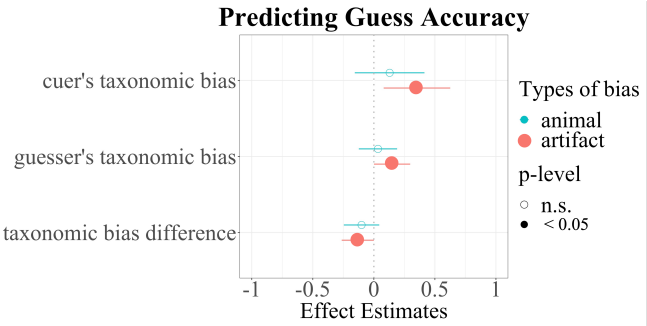

Figure 4: Effects of cuer's and guesser's taxonomic biases and cuer's and guesser's bias difference on guess accuracy

Why was accuracy only predicted by the artifact taxonomic bias? One possible reason is that for animal words the thematic similarity between the center word and the thematic paired word is always lower that the taxonomic similarity between the center word and the taxonomic paired word. Therefore, compared to artifact taxonomic bias, animal taxonomic bias is a less discriminative indicator of differences in semantic organization.

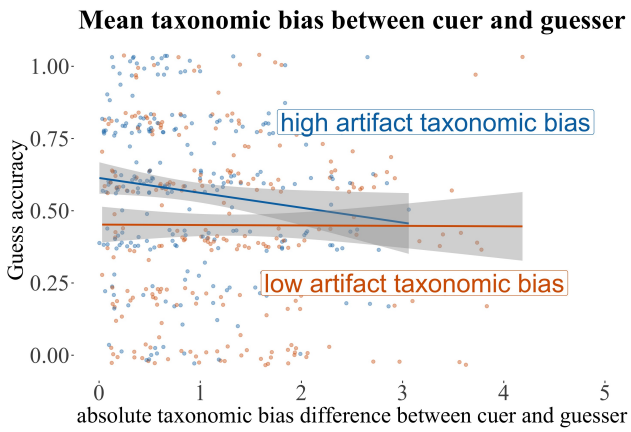

Figure 5: Guessing accuracy as a function of difference between cuer and guesser's taxonomic biases (x-axis) and mean taxonomic bias (median split). Dark regions are 95% confidence bands.

General Discussion

The main motivation for our study was to examine whether divergence in word meanings (e.g., Suessbrick et al., 2000; Martí et al., 2023; Wang & Bi, 2021) has a more principled basis than idiosyncratic differences in semantic knowledge of specific words/concepts. By using similarity judgments to measure aspects of people's word meanings, interpreting differences in similarity judgments as informative of underlying differences in semantic knowledge (Martí et al., 2023; Wang & Bi, 2021) and examining the underlying basis for people's ratings, we found that differences in ratings for animals (Experiment 1) and artifacts (Experiment 2) were consistent within domains, an to a lesser degree, across semantic domains such that some people relied primarily on taxonomic similarity while others less so.

In Experiment 3, we further test whether these differences in word meanings (captured here by a difference in people's taxonomic bias) have consequences for communication. People with more similar taxonomic biases had greater communicative success (as long as the mean taxonomic bias was relatively high).

Our work has several notable limitations. It has little to say about what gives rise to differences in word meanings (but see Mirman, Landrigan, & Britt, 2017; Mirman & Graziano, 2012, for work relevant to individual differences in taxonomic biases). An intriguing possibility is that a stronger taxonomic bias may come in part from exposure to written material (Nation & Snowling, 1999). Another limitation is that the communicative task in Experiment 3 is a rather contrived task involving only asynchronous communication. Typical face-to-face language use allows for conversational repair. It's up to future work to see whether the types of divergence in word meaning we observe here also affects interactive language use as evidenced by more other-initiated repairs requests and/or lower success in resolving these requests.

References

Casasanto, D., & Lupyan, G. (2014). All Concepts are Ad Hoc Concepts. In E. Margolis & S. Laurence (Eds.), *Concepts: New Directions* (pp. 543–566). Cambridge: MIT Press.

Clark, H. H. (1998). Communal Lexicons. In K. Malmkjær & J. Williams (Eds.), *Context in Language Learning and Language Understanding* (p. 63). Cambridge University Press.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of experimental psychology: general*, *132*(2), 163.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hutchins, E., & Hazlehurst, B. (2006). How to invent a lexicon: the development of shared symbols in interaction. In *Artificial societies* (pp. 144–171). Routledge.

Landrigan, J.-F., & Mirman, D. (2016). Taxonomic and thematic relatedness ratings for 659 word pairs. *Journal of Open Psychology Data*, *4*(1).

Lin, L., & Murphy, G. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology-General*, *130*(1), 3–28.

Lupyan, G., Piantadosi, S., Martí, L., Kidd, C., Liu, E., van Paridon, J., & Bi, Y. (2022). How much do we agree on what words mean? In *Joint conference on language evolution.*

Martí, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent diversity in human concepts. *Open Mind*, *7*, 79–92.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99.

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General*, *141*, 601–609. doi: 10.1037/a0026451

Mirman, D., Landrigan, J.-F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological bulletin*, *143*(5), 499.

Nation, K., & Snowling, M. J. (1999). Developmental differences in sensitivity to semantic relations among good and poor comprehenders: Evidence from semantic priming. *Cognition*, *70*(1), B1–B13.

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? grammatical gender and object concepts. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 25).

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, *102*(2), 145–158.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Suessbrick, A. L., Schober, M. F., & Conrad, F. G. (2000). Different respondents interpret ordinary questions quite differently. In *Proceedings of the american statistical association.*

Sulik, J., & Lupyan, G. (2018). Perspective taking in a novel signaling task: Effects of world knowledge and contextual constraint. *Journal of Experimental Psychology: General*, *147*(11), 1619.

Wang, X., & Bi, Y. (2021, September). Idiosyncratic Tower of Babel: Individual Differences in Word-Meaning Representation Increase as Word Abstractness Increases. *Psychological Science*, 09567976211003877. Retrieved 2021-09-29, from https://doi.org/10.1177/09567976211003877 doi: 10.1177/09567976211003877

Waxman, S. R., & Senghas, A. (1992). Relations among word meanings in early lexical development. *Developmental Psychology*, *28*(5), 862.