Characterizing Variability in Shared Meaning through Millions of Sketches

Molly Lewis
mollyllewis@gmail.com
Dept. of Psychology
Carnegie Mellon University

Anjali Balamurugan
abalamur@andrew.cmu.edu
Dept. of Stats. & Data Science
Carnegie Mellon University

Bin Zheng
binzheng@andrew.cmu.edu
Dept. of Stats. & Data Science
Carnegie Mellon University

Gary Lupyan
lupyan@wisc.edu
Dept. of Psychology
University of Wisconsin-Madison

Abstract

The study of mental representations of concepts has historically focused on the representations of the "average" person. Here, we shift away from this aggregate view and examine the principles of variability across people in conceptual representations. Using a database of millions of sketches by people worldwide, we ask what predicts whether people converge or diverge in their representations of a specific concept, and which kinds of concepts tend to be more or less variable. We find that larger and more dense populations tend to have less variable representations, and concepts high in valence and arousal tend to be less variable across people. Further, two countries tend to have people with more similar conceptual representations when they are linguistically, geographically, and culturally similar. Our work provides the first characterization of the principles of variability in shared meaning across a large, diverse sample of participants.

Keywords: concepts and categories; drawings; cultural variability; large scale data

Introduction

Understanding how the human mind represents concepts is a fundamental question for psychologists, philosophers, and linguists (Margolis & Laurence, 1999). Researchers have developed a wide range of theories characterizing how people represent relatively simple concepts that are shared across people, like "chair" and "tree." Such theories make predictions about how an "average" person should perform in behavioral tasks, such as rating how typical members of the category are. Here we shift the focus from the representation of the "average" person, to *differences* in representations between people. This allows us to ask (1) what predicts whether people converge or diverge in their representations of a specific concept, and (2) which concepts are more vs. less variable. To answer these questions, we use a novel method: analyses of millions of sketches drawn by participants worldwide.

The question of variability in shared meaning across people has received remarkably little attention in part because it is difficult to study. One reason for this is that people tend to engage in "good enough" processing (Ferreira & Patson, 2007) when faced with behavioral tasks, thus often failing to reveal to outside observers underlying differences in their conceptual representations. Second, the psychological paradigms that are used to study individuals' concept representations, like typicality (e.g., Rosch, 1975) and word association tasks (e.g., De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019), are relatively course-grained. And, third, the differences in conceptual representations between people are likely small, requiring a large dataset to observe variability.

One way to get a window into people's shared meaning is through drawings. Like words, drawings are emergent cultural conventions that can be used to communicate about shared meaning. However, unlike words, drawings resemble the shared meaning they reference—a drawing of the concept "chair" *looks like* a chair. Drawings therefore provide an observable, quantifiable index of a person's conceptual representations (e.g., Fan, Yamins, & Turk-Browne, 2018; Long, Fan, & Frank, 2018).

Notably, drawings are not an unmediated window into a person's conceptual representations—a person's drawing of "chair" is not isomorphic to their cognitive representation of a chair (Cohn, 2019). This difference is due in part to cultural conventions about how to represent particular concepts in the drawing modality. For instance, a culture may converge on the convention of drawing the concept "house" with four windows and a chimney. The fact that drawings are conceptual representations mediated by drawing-specific conventions is unproblematic for the current purposes: drawings provide a test bed for understanding the dynamics of shared meaning in one particular modality.

Prior work has examined the emergence of drawing conventions in experimental paradigms (Garrod, Fay, Lee, Oberlander, & MacLeod, 2007). A key finding from this work is that drawings become both more consistent through repeated interactions and also more reliant on memory. For example, when two interlocutors are tasked with communicating the meaning "bunny" through sketch, they might draw a detailed picture with a nose, whiskers, and ears. But, with repeated interactions, this drawing will tend to become more schematic such that "bunny" is represented simply as two ears.

This prior work makes two general predictions about variability in shared meaning. The first is that the degree of interaction among people should be related to the degree of consensus in a concept: More social interaction should lead to higher degree of consensus (less variability between people). This prediction is consonant with findings in the language evolution literature demonstrating that languages with fewer people tend to have more complex (arguably, less variable) language systems (Lupyan & Dale, 2010). Second, psychological variables that influence memory should influence which concepts are more variable across people.

Prior work has shown that words that are more frequent (Hall, 1954), more concrete (Fliessbach, Weis, Klaver, Elger, & Weber, 2006), more positively valenced (Mather & Carstensen, 2005), and learned earlier (Barry, Hirsh, Johnston, & Williams, 2001) tend to be favored in learning and memory tasks. We predict that properties that make concepts more memorable should also make them less variable.

In what follows, we test these hypotheses using a database of millions of drawings by people worldwide. In Study 1, we develop a psychologically-valid method for quantifying the similarity of drawings at a large scale. In Studies 2 and 3, we then use this measure to examine predictors of variability in shared meaning across people and concepts.

Study 1: Estimating drawing similarity

Quantifying variability in drawings requires a psychologically-valid metric of drawing similarity. Because collecting human similarity judgments for millions of drawings is not feasible, in Study 1 we collect human similarity judgments for a sample of drawing pairs and use these to develop a computational measure of the similarity between two arbitrary drawings.

Methods

Participants We recruited 331 participants through Amazon Mechanical Turk and an undergraduate subject pool. After excluding participants who missed an attentional check, our final sample of included 267 participants.

Stimuli Drawings were taken from the Quick, Draw! dataset collected by Google (`https://github.com/googlecreativelab/quickdraw-dataset`). The drawings were collected through an online app in which participants were cued with an English word (e.g., "watermelon") and asked to sketch the corresponding object in under 20 seconds (Fig. 1). As participants sketched, a neural net trained on other participants' drawings made guesses about the cue word (this feature lead to some impartial drawings, but it is unlikely that this effect interacted with demographic features of the participant). Once the neural net guessed correctly, the app progressed to the next word cue. Each participant completed up to 6 drawings per session. Drawings are represented in the database as a series of x-y coordinates with individual strokes identified, and have been pre-processed to standardize position and scaling. Participant's country is also included as metadata.

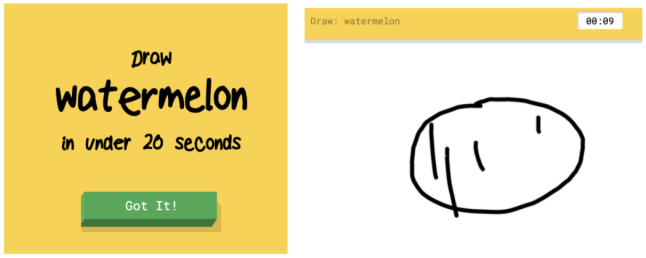

Figure 1: Screenshots of the Quick, Draw! App (https://quickdraw.withgoogle.com/).

For the current study, we sampled 1,000 drawing pairs for each of five word cues: "tree," "bread," "chair," "house" and "bird." In order to include a range of drawing similarities in our stimuli, we quantified the similarity between drawings in a pair using a computational measure of visual image similarity commonly used in machine vision, called Hausdorff distance (Huttenlocher, Klanderman, & Rucklidge, 1993). Informally, Hausdorff distance quantifies the similarity of two images by treating each image as a set of x-y coordinates, and calculating the Euclidean norm between each point in one image to the closest point in the other. The Hausdorff distance is the maximum of these pairwise distances (the distance between the most mismatched points). We calculated Hausdorff distance for each drawing pair and then sampled 20 drawing pairs from each distance decile (see Fig. 2). Our final stimuli list included 200 drawing pairs for each of the 5 target cues.

Procedure Participants were instructed to rate how similar pairs of drawings were to each other on a 7-pt Likert scale, ranging from "almost identical" to "completely different." Each participant rated a sample of 50 drawing pairs from a single cue word. As an attention check, we also included two additional trials where the two drawings were identical to each other. Participants were excluded from the final sample if they responded 3 or higher on the Likert scale for either of these two trials. Each drawing pair was rated by an average of 13.34 participants (SD = 7.04).

Results and Discussion

Log Hausdorff distance was moderately positively correlated with human judgments of visual dissimilarity ($r(998) = 0.4$, $p < .0001$; Fig. 3), accounting for 16% of the variance in human judgments.

We next tried to better predict human similarity judgment using additional computational measures of similarity. We examined five new measures in drawing similarity: Log average Hausdorff distance (Taha & Hanbury, 2015), Euclidean distance, Mahalanobis distance (Mahalanobis, 1936), log difference in number of strokes between, and the log difference in mean stroke length. Log average Hausdorff distance is similar to the Hausdorff distance metric described above, but is less sensitive to outliers. Euclidean distance is calculated as the average pairwise Euclidean distance between all points. Mahalanobis distance is similar to Euclidean distance, but takes into account the correlation of points in the drawings.

The five distance measures were moderately correlated with each other and with human judgments (see Table 1). We next fit an additive linear model predicting human judgments with each of these five predictors. This model accounted for 28% of the variance in human judgments (see Table 2 for model parameters). Figure 4 shows a two-dimensional scaling solution of the predicted human similarity ratings for a sample of one hundred "bird" drawings.

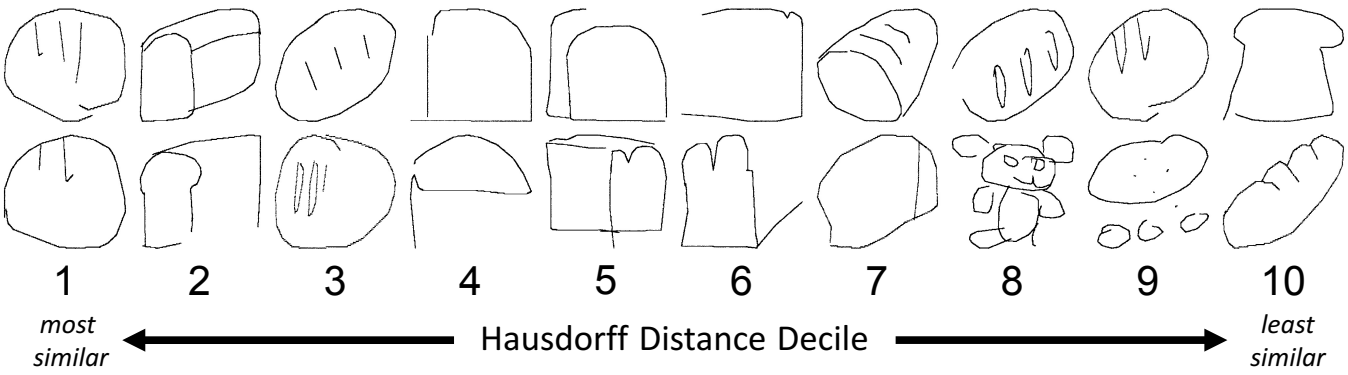

Figure 2: Example stimuli pairs for the cue "bread," sampled from each Hausdorff distance decile.

Table 1: Pairwise correlations (Pearson's r) between all five drawing distance measures. Human = average human disimilarity judgment; Haus. = log average Hausdorff distance; Euc. = Euclidean distance; Maha. = Mahalanobis distance; N strokes = log difference in number of strokes; Stroke Len. = log difference in stroke length (in pixels). * = $p < .01$.

	Human	Haus.	Euc.	Maha.	N Strokes
Haus.	0.34*				
Euc.	0.03	0.6*			
Maha.	0.4*	0.44*	0.24*		
N Strokes	0.27*	0.09*	-0.05	0.19*	
Stroke Len.	0.13*	-0.04	0.07	0.14*	0.14*

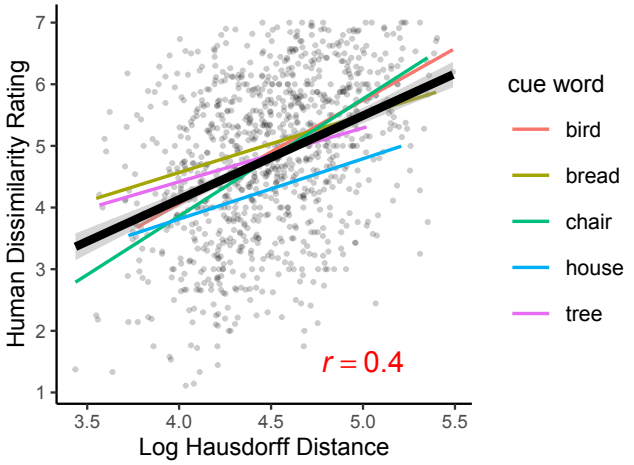

Figure 3: Relationship between human judgments of drawing similarity and drawing similarity estimated from a computational measure, log Hausdorff distance. Each point corresponds to a drawing pair (N = 1,000). The color lines show the best fit for each of the five individual cue words; black line shows the best fit for all drawing pairs and corresponding standard error. Pearson's r corresponds to the correlation estimate across all word cues.

Table 2: Fixed effect parameters for the additive linear model predicting human similarity judgment of 1,000 drawing pairs in Study 1 from five computational similarity measures. Log Avg. Haus. = log average Hausdorff distance.

	Estimate	SE	t-value	Pr($>\|t\|$)
Hausdorff	0.37	0.04	9.84	$<.01$
Mahalanobis	0.25	0.03	8.18	$<.01$
Euclidean	-0.26	0.03	-7.44	$<.01$
N Strokes	0.16	0.03	5.88	$<.01$
Stroke Length	0.10	0.03	3.71	$<.01$

In sum, Study 1 provides evidence that human judgments of sketches are partially predictable from simple computational measures of visual similarity. In the remaining studies, we use the parameters of the model we derived for predicting human similarity judgments to estimate similarity in a large scale analyses of drawings from the Quick, Draw! dataset.

Study 2: Variability in shared meaning

In Study 2, we ask what characteristics of populations predicts whether people converge or diverge in their representations of a specific concept, and which kinds of concepts tend to be more or less variable across people.

Study 2a: Population properties

Method The Quick, Draw! Dataset contains drawings for 345 cue words. We excluded cues that contained multiple

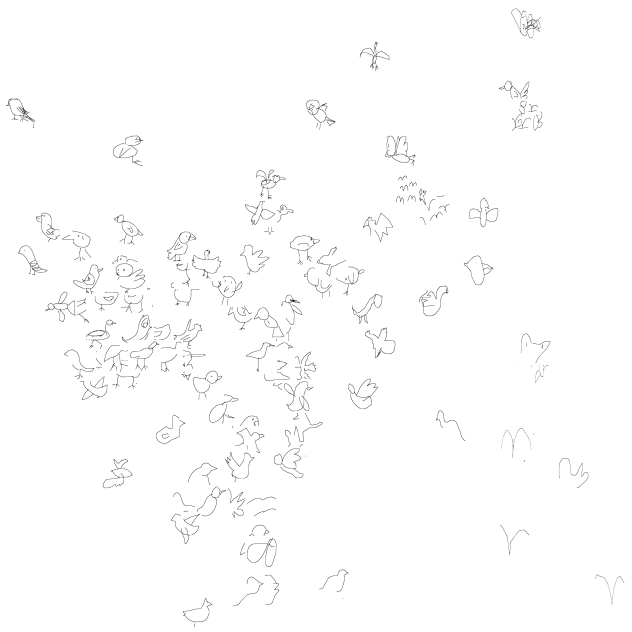

Figure 4: Multi-dimensional scaling solution of pairwise similarity of 100 bird drawings judged in Study 1. Similarity is estimated from as the predicted values from a model predicting human judgments with five computational similarity measures.

words (e.g., "sea turtle;" N = 54) or had synonymous meanings ("bat;" N = 1), leaving 288 words.

We analyzed drawings for each of these cue words for the 20 countries with the most drawings. For each country-cue combination (e.g., Thailand-bread), we sampled 1,000 drawings to create 500 drawing pairs (N = 5.76M total drawings). We then quantified the distance between drawings for each pair by calculating the five distance metrics described in Study 1 and then calculating the predicted human distance using the parameters of the additive linear model developed in Study 1. For each country-cue, we quantified the distribution of drawing similarity across people as the mean and standard deviation of predicted human distance drawings across the 500 pairs. Because mean and standard deviation were correlated with each other ($r(5758) = 0.3$, $p < .0001$), we quantified the variability in the distances using the coefficient of variation (the ratio of the standard deviation to the mean). The coefficient of variation quantifies the variability in a distribution, controlling for differences in the mean.

We examined the relationship between coefficient of variation for drawing similarity and two properties of populations that have been hypothesized to relate to linguistic variability: Population size and population density. For each country, we obtained estimates of population size (log thousands) and population density (N people/sq. km.) from Worldbank.

Results We next predicted the amount of variability for each cue within each country. Population density and population size were not significantly correlated ($r(18) = 0.03$, p = 0.91). We fit an additive mixed effect model predicting the coefficient of variation for each country-cue pair with population size and density, including random intercepts by cue and country. Each observation in our model was a country-cue pair. Both population size (β = -0.05, SE = 0.02, Z = -3.19) and density (β = -0.03, SE = 0.02, Z = -2.07) were reliable predictors of variability: countries with smaller and less dense populations tended to have more variable drawings (Fig. 5, left; Fig. 6a).

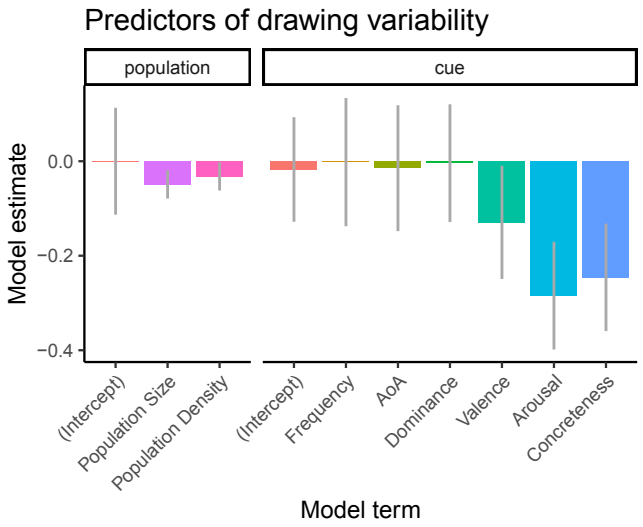

Figure 5: Standardized parameters from additive mixed effect linear model predicting the coefficient of variation in drawing distances with population (left) and cue (right) predictors. Ranges correspond to 95% confidence intervals.

Study 2b: Cue properties

Method We examined six psychologically-relevant properties of the cue words: (i) word frequency estimated from a spoken corpus, (log; Brysbaert & New, 2009), (ii) Concreteness (degree to which a word refers to a perceptible entity; Brysbaert, Warriner, & Kuperman, 2013), (iii) estimated age of acquisition (log AoA; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), (iv) arousal (intensity of emotion provoked by a stimulus), (v) valence (pleasantness of a stimulus), and (vi) dominance (degree of control exerted by a stimulus; Warriner, Kuperman, & Brysbaert, 2013). Complete data were available for 93.4% of cues.

Results We fit an additive mixed effect model predicting the coefficient of variation for each country-cue combination with each of the six word-level predictors with random intercepts by cue and country.

Drawings tended to have low variability across people when the cue words were associated with high arousal (e.g. "tornado"; β = -0.28, SE = 0.06, Z = -4.9; Fig. 5, right; Fig. 6b), positive valence (e.g., "angel"; β = -0.13, SE = 0.06, Z = -2.12), and high concreteness (e.g. "cup"; β = -0.25, SE

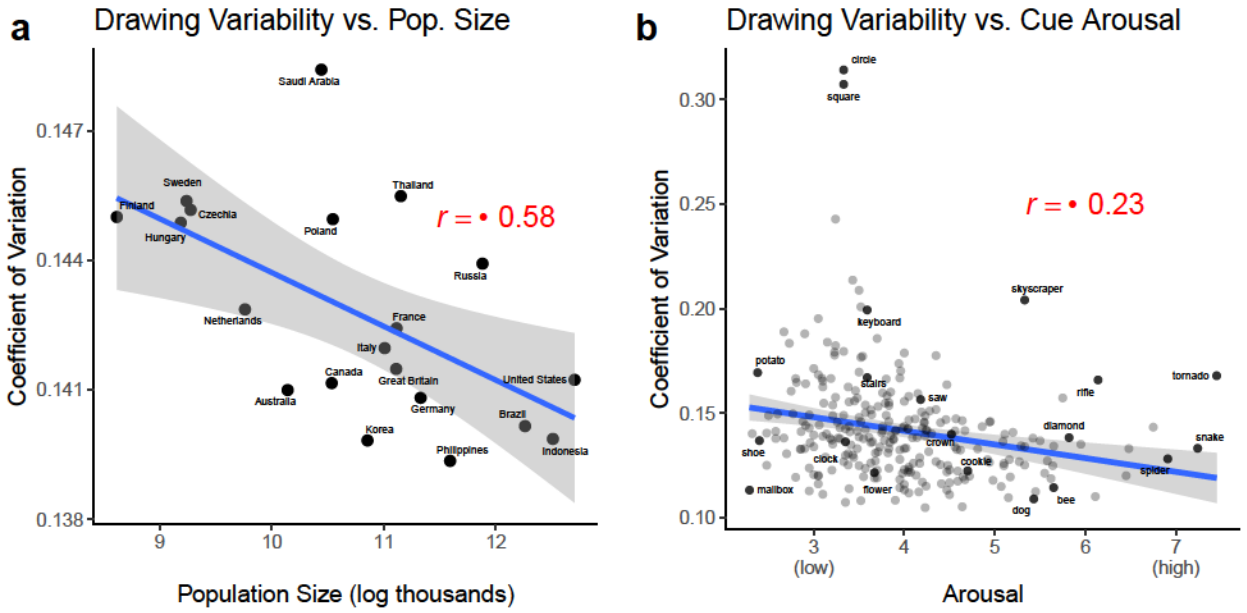

Figure 6: (a) Relationship between drawing variability and population size. (b) Relationship drawing variability and word cue arousal: Words that have higher arousal (e.g., "tornado") are associated with more similar drawings across people. The relationship is robust to the exclusion of two outliers ("circle" and "square;" $r = .24$, $p < .01$).

= 0.06, Z = -4.24). Word frequency, age of acquisition, and dominance were not reliable predictors of variability.

Follow-up analyses revealed that the concreteness effect was due primarily to the subset of cues related to shapes (e.g., "circle," "square"); with these items excluded, there was no longer a relationship between concreteness and variability in drawing similarity (β = -0.07, SE = 0.06, Z = -1.28). Notably, however, after excluding shape cues, the sample of items were highly concrete with relatively little variability (M = 4.87; SD = 0.15), making this effect difficult to detect in the current sample of items.

Study 2 provides evidence that shared meaning across people is detectable in drawings and varies in predictable ways: Smaller, less dense populations have more variable shared representations across a range of concepts, and concepts that are higher in arousal and valence tend to be more variable.

Study 3: Shared meaning across cultures

In Study 3, we examine a corollary of the prediction that greater social interaction should lead to greater conceptual similarity. In particular, we ask whether participants from more related countries produce drawings that are also more similar to each other.

Method

We used the same sample of drawings as in Study 2 to quantify the similarity between drawings among people from different countries (1,000 drawings for each country-cue). For each pair of countries in our dataset (N = 190), we created 1,000 drawing pairs, where each pair was composed of a drawing of the same item from a different country (e.g. a bird drawing from a Hungarian participant and one from a Brazilian participant). We quantified the similarity between drawings in a pair using the same method as in Study 2, and then averaged across drawing pairs to calculate the mean dissimilarity rating for each country pair and cue combination.

We examined three variables that relate to the amount of interaction between people in two countries: geographic distance, linguistic distance, and amount of migration. Geographic distance was calculated as the distance in log meters between the centroid of the two countries. Linguistic distance was quantified as the Levenshtein edit distance between a standard set of words in each country's most frequently spoken language (Dediu, 2017). Migration was quantified as the log number of people who migrated between two countries (average from country a to b and b to a; Worldbank, 2017).

Results and Discussion

Figure 7 visualizes the cross-country variability in drawings for a particular cue, "bread." To account for this variability, we fit an additive mixed effect model predicting mean dissimilarity rating for each country pair and cue combination (M = 4.78; SD = 0.36) with geographic distance, linguistic distance, and number of migrants. These three measures were weakly correlated with each other (geographic-linguistic: $r(188)$ = -0.01, p = 0.91; linguistic-migration: $r(188)$ = -0.34, $p < .0001$; migration-geographic: $r(188)$ = -0.21, $p < .01$). Both countries and cue were included as random intercepts.

Geographic distance (β = 0.008, SE = 0.001, Z = 5.23), linguistic distance (β = 0.006, SE = 0.001, Z = 4.59), and number of migrants (β = -0.003, SE = 0.001, Z = -2.1) each

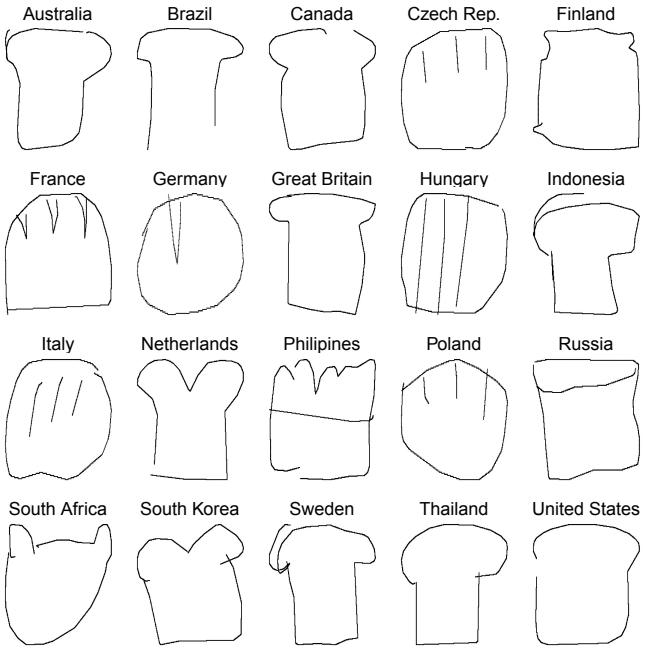

Figure 7: Prototypical bread drawings for each country in our analysis. The prototype was calculated as the drawing with the shortest average distance to all other drawings from the same country.

predicted independent variance in mean drawing similarity: Countries that spoke more similar languages, were closer geographically, and had more inter-country migration tended to have people who created more similar drawings. This suggests that countries that are more similar due to cultural contact tend to have a higher degree of shared meaning.

General Discussion

How does one person's representation of a simple concept, like "chair," differ from another's? Using millions of sketches collected from an international population, we provide a window into the principles of variability in shared meaning. We find that properties of populations—density and size—and properties of concepts—arousal and valence—influence the degree of convergence in meaning across people.

There are, however, a number of important limitations of this work. First, drawings are not a direct window into people's concepts both because they are conventionalized representations, and because they only represent a concept through shape features. Second, our computational metric of drawing similarity is only able to predict about a third of the variance in human judgments of drawing similarity. It is possible that with a better metric of drawing similarity, using, for instance, neural networks trained on drawings, our method would be more sensitive to smaller effects.

In sum, how the "average" person represents a concept is only part of a complete theory of conceptual representations—it is also critical to understand how people vary in these representations. Variability in shared meaning, for example, implies that some groups of people may be better able to communicate with each other than others. Our work provides the first large scale characterization of the principles of variability in shared meaning.

References

Bank, T. W. (2017). Migration and remittances data. Retrieved from `https://www.worldbank.org/en/topic/migrationremittancesdiasporaissues/brief/migration-remittances-data`

Barry, C., Hirsh, K. W., Johnston, R. A., & Williams, C. L. (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language*, *44*(3), 350–375.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 1–8.

Cohn, N. (2019). Visual narratives and the mind: Comprehension, cognition, and learning. In *Psychology of learning and motivation* (Vol. 70, pp. 97–127). Elsevier.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006.

Dediu, D. (2017). Language classifications as standardized newick phylogenetic trees with branch length. *GitHub repository*. `https://github.com/ddediu/lgfam-newick`; GitHub.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*(8), 2670–2698.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage*, *32*(3), 1413–1421.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987.

Hall, J. F. (1954). Learning as a function of word-frequency. *The American Journal of Psychology*, *67*(1), 138–140.

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850–863.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Long, B., Fan, J. E., & Frank, M. C. (2018). Drawings as a window into developmental changes in object representations. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS One*, *5*(1), e8559.

Mahalanobis, P. C. (1936). On the generalised distance in statistics, (6), 49–55.

Margolis, E., & Laurence, S. (1999). *Concepts: Core readings*.

Mather, M., & Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences*, *9*(10), 496–502.

Rosch, E. (1975). Cognitive representations of semantic categories. *JEP: G*, *104*(3), 192.

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*(1), 29.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.