PHILOSOPHICAL TRANSACTIONS B

*Phil. Trans. R. Soc. B.*
doi:10.1098/not yet assigned

# Cross-domain semantic alignment: Concrete concepts are more abstract than you think

**Qiawen Liu[1] and Gary Lupyan[1]**

[1]*Department of Psychology, University of Wisconsin, Madison, WI 53706, USA*

**Keywords:** cross-domain mapping, concrete concepts, metaphor, abstract dimensions

## Abstract

We can easily evaluate similarities between concepts within semantic domains, e.g., doctor and nurse, or violin and piano. Here, we show that people are also able to evaluate similarities across domains, e.g., aligning doctors with pianos and nurses with violins. We argue that understanding how people do this is important for understanding conceptual organization and the ubiquity of metaphorical language. We asked people to answer questions of the form "If a nurse were an animal, they would be a(n)..." (Experiment 1 and 2), and asked them to explain the basis for their response (Experiment 1). People converged to a surprising degree (e.g., 20% answered "cat"). In Experiment 3, we presented people with cross-domain mappings of the form "If a nurse were an animal, they would be a cat" and asked them to indicate how good each mapping was. The results showed that the targets people chose and their goodness ratings of a given response were predicted by similarity along abstract semantic dimensions such as valence, speed, and genderedness. Reliance on such dimensions was also the most common explanation for their responses. Altogether, we show that people can evaluate similarity between very different domains in predictable ways, suggesting that either seemingly concrete concepts are represented along relatively abstract dimensions (e.g., weak-strong) or that they can be readily projected onto these dimensions.

## Introduction

It is easy—at least for adults—to compare a *dog* to a *cat*, or a *piano* to a *violin*. In making these comparisons, people draw on rich perceptual and functional knowledge: that dogs are more similar to cats than to elephants is, for example, well predicted by a greater overlap in semantic features that people list for dogs and cats, than those they list for elephants [1–3]. When reasoning about more relational concepts such as *doctor* and *nurse*, people likewise appear to recruit knowledge about individual items which can be aligned and contrasted. For example, people judge a *doctor* to be more similar to a *nurse* than to a *carpenter* because doctors and nurses overlap more in the functions they perform and their job knowledge, than do doctors and carpenters.

However, people's ability to appreciate semantic similarity is not limited to comparing items from the same semantic domain. When asked to align concepts from different semantic domains, people sometimes show uncanny convergence. For example, when asked "If science were a colour, what colour would it be", 40% respond with "green". When asked in the same manner, to map philosophy to a beverage, 20% map it to tea. When mapping professions to musical instruments, 32% map doctors to pianos while 26% map nurses to violins (these two responses are the modal responses and greatly exceed the probability of these instruments being mentioned by chance alone) [4]. A main goal of the three experiments we present here is to understand the bases for such cross-domain mappings. In what ways are animals and musical instruments, or professions and animals similar such that when asked to map between them people prefer some mappings over others?

*Author for correspondence (lupyan@wisc.edu).
†Present address: Department of Psychology, University of Wisconsin, Madison, WI 53706, USA

There are two main reasons why studying people's performance on these—admittedly odd—cross-domain mapping tasks is important. The first is that while some consistent cross-domain mappings can be explained through simple associations between words or through clear thematic relationships, others are not so obvious and require an explanation. For example, when asked to map a cow to a beverage, not surprisingly most people (86%) respond with "milk". When asked to map cities to colours, people's responses are often mediated by sports team colours (e.g., 35 % map Seattle to blue presumably because of the Seahawks. When asked to map Boston to a colour, the top two answers are "green" (27%) and "red" (20%) presumably because of the Boston Celtics and Red Sox, respectively). Occasionally people even rely on phonological similarity, e.g., when asked if a bear were a beverage, it would be a…, 20% responded with "beer"). But it is more puzzling why a plurality of respondents map "philosophy" to "tea", or "cello" to a "cloudy day". If some kind of alignment between the domains is involved [5–8] what exactly is being aligned and along what dimensions of meaning?[1]

The second reason for examining how people perform cross-domain mappings is that it can help us think more clearly about the differences—and similarities—between how people represent concrete and abstract concepts [9–15]. Because concrete concepts generally have clearly identifiable perceptual properties [indeed that is largely what makes them "concrete" in the first place, 16], it is tempting to assume that mental representations of concrete items are constituted by these properties [17,18] with similarity being a function of overlap [19–22]. On this account, it is easy to explain why, when asked to map a pilot to an animal, a plurality of respondents (22%) respond with "bird": both have "fly" as a common semantic feature. But as we shall see, many mappings people consistently make cannot be explained through shallow feature matching of this sort because the items from the two domains do not have any obvious features in common.

An alternative way to make sense of what people do when asked to map between semantic domains is to assume that even highly concrete items are constituted not just by their perceptual and motor information, but by their placement along more abstract dimensions that cross semantic domains [23,24] and it is similarity along these abstract dimensions that explains patterns of cross-domain mappings. One example of such a dimension is valence. It is reasonable to talk about positively and negatively valenced animals, jobs, beverages, colours [25], etc., and one can imagine that mapping between these domains is informed by how similarly valenced the items are. For example, if part of our representation of "rat" is its negative valence, then when we are asked "If a rat were a job, what job would it be", we tend to think of negatively valenced jobs. Further informing the alignment may be dimensions related to activity (e.g., passive/active, dull/exciting) potency (e.g., weak/strong, silent/loud), and dimensions such as size and weight, which although grounded in concrete perceptual qualities, also have readily accessible metaphorical extensions. For example, many people interpret a "heavy job" to be a job with many responsibilities and a "low job" as one lacking prestige.

If alignment along these relatively abstract semantic dimensions explains people's cross-domain mappings it suggests that representations of even very concrete items include a considerable amount of rather abstract information. If so, then what may distinguish abstract and concrete concepts is not that abstract concepts uniquely include certain abstract information that concrete concepts lack. Rather, all concepts—regardless of concreteness—can be positioned on these abstract dimensions, but more abstract/relational concepts may lack some of the more grounded information that partly constitutes concrete concepts.[2] If true, it would also

[1] When the second author was 2.5 years old, his mother recorded him as asking the following question: "What is the difference between Pinocchio and a sausage?" It is not clear what the second author was thinking at the time (the shape of Pinocchio's nose?). Our results show that despite the seeming ill-posed nature of this comparison (one can hear the exasperated parent say "What?! Nothing!"), people are capable of aligning items from different semantic domains in ways that are both predictable and sensible.

[2] An alternative is that abstract dimensions like valence and potency are not constitutive of concrete conceptual representations, but when tasked with mapping between two different semantic domains, people project the concepts into the more abstract space in which items from both domains can be compared. We address this possibility in the General Discussion.

help explain why people so readily arrange colours by, e.g., how exciting, new, or active they are (a task that involves placing concrete concepts along these more abstract dimensions) [26,27].

In addition to shared perceptual features and alignment along abstract dimensions, there are other sources of information people may be relying on when making cross-domain mappings. One is alignment on sensorimotor dimensions. For example, when asked "If thunder were a musical instrument, it would be a…", 53% people respond with "drum". This answer is clearly motivated by sensory similarity: thunder can literally sound like a drum. At a more general level, thunder and drums are both saliently experienced in the auditory modality, a relationship captured by, e.g., the Lancaster sensorimotor norms [28] (though this would not explain why the majority of participants responded with "drum" in particular).

Another source of information that can inform semantic relationships between concepts is language itself [10,29–33]. By keeping track of the contexts in which words occur (which can be done by simply trying to predict what words occur near another word), it is possible to construct semantic spaces (word embeddings) in which the distance between words approximates human judgments to a surprising degree [34–37]. Although it is people—embodied agents with rich sensory and perceptual experiences—who create language on which the word embedding models are trained, a substantial amount of an individual's semantic knowledge may be learned from the distributional patterns produced by the language community to which we are exposed, rather than from direct sensorimotor experiences [32,38]. We were interested in whether linguistic information as captured by word embeddings derived from large English corpora, and by word-associations produced by English speakers can account for cross-domain mappings. For example, is the tendency to map "philosophy" to "tea" caused by both words—for one reason or another—occurring in similar contexts and therefore becoming linked?

To further illustrate how the different sources of information just reviewed are brought to bear on cross-domain alignment, consider the mapping "If a rat were an occupation, it would be a thief". Figure 1 shows a schematic that contrasts different, but not mutually exclusive ways of explaining this mapping. In panel A, the similarity between "rat" and "thief" is computed simply by taking note of overlap in semantic features—the method that works reasonably well for computing similarity within domains, but we suspect would fail here owing to a lack of shared features for many mappings we ask people to make. Panel B draws on similarity based on the sensorimotor profiles of the two words taken from the Lancaster sensorimotor norms [28]. If these are to be useful for predicting people's cross-domain mappings, sensorimotor similarity between "rat" and "thief" should be greater than between "rat" and other (less often selected) occupations. Panels C and D show two methods of computing similarity based on people's word associations. Panel C uses first-order word-associations: given "rat" how likely are English speakers to produce "thief". The answer, based on the Small World of Words association norms [39], is that none of the tested participants did. We can go beyond first-order word associations by computing paths between the two words in a network where words are joined by attested association links [40]. One path between the two words is rat → snitch → steal → thief. Panel D shows a different type of linguistic similarity using a distributional semantic approach. Here, the similarity between the two words is computed based on the similarity of the linguistic contexts in which they occur [41,34]. Finally, panel E shows similarity as computed by projecting each word into a common space defined by a set of more abstract dimensions.

To understand which of these—if any—kinds of similarities are at play in cross-domain mappings, we conducted three experiments. Experiment 1 was an open-ended task in which participants were asked to complete prompts of the form "If an A were a B it would be a…" where A is a specific item (e.g., "doctor") and B is a category label from a different domain (e.g., "musical instrument"). Participants were asked to explain their responses after the open-ended task, and their explanations were coded by two raters into distinct types depending on the likely basis of the mapping. In Experiment 2, we predicted the likelihood of producing various responses from the various kinds of similarity depicted in Figure 1. Because this analysis was necessarily limited to attested responses, it restricted our ability to distinguish between some of the predictors. In Experiment 3 we therefore generated specific mappings (e.g., "If a bat were a job, it would be a plumber") and asked people to indicate how good each mapping was. This procedure allowed us to manipulate mappings along one type of similarity while holding another constant. Of particular interest was whether people's mappings can be predicted by alignment on relatively abstract dimensions (Fig 1E) even when

controlling for other sources of similarity. If so, it suggests that concrete concepts like “cat”, “violin”, and “rat” may be more abstract than is sometimes appreciated.

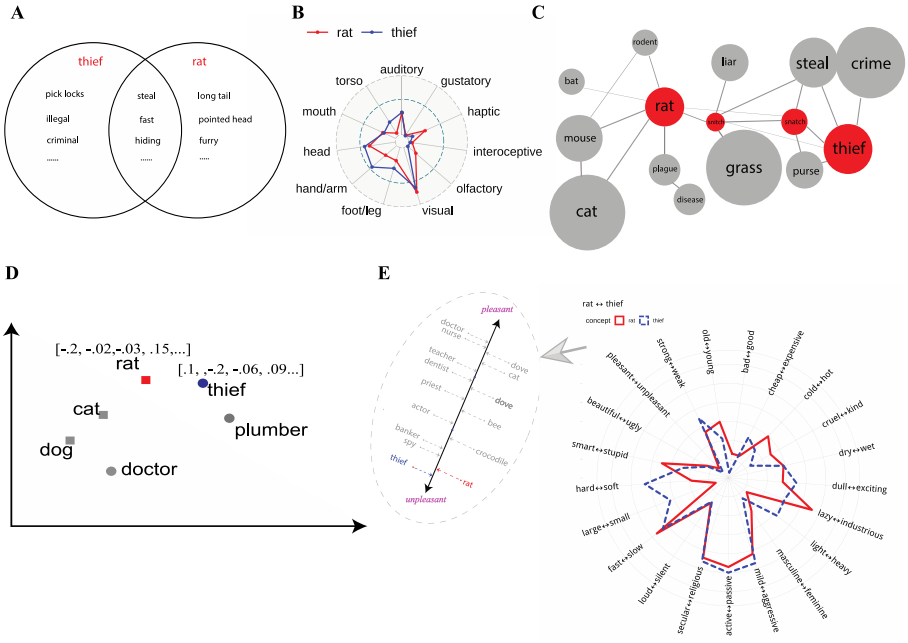


*Figure 1. Five ways of making sense of people’s tendency to respond with “thief” when asked “If a rat were an occupation, it would be a… A: The two concepts have shared semantic features. B: The two concepts match in their sensorimotor profiles (e.g., as quantified by the Lancaster norms [28]. C: The two concepts have common word associations. The image shows the shortest distance between rat and thief in the Small World of Words free association norms (rat -- snitch -- snatch – thief) [39,40].). D: The two concepts are close to one other in a word embedding space learned by tracking shared linguistic contexts. The actual space is 300-dimensional rather than 2-dimensonsal as schematized. E: The two concepts have similar profiles when aligned on relatively abstract dimensions that apply to many semantic domains.*

## Experiment 1: How do people answer questions like “If a nurse were an animal, they would be a(n)…?

The main goal of Experiment 1 was to determine the relative use of different kinds of similarity in understanding cross-domain mappings by evaluating people’s explanations of their mappings. We began by investigating people’s responses when they are asked to map from one semantic domain (e.g., occupations) to another (e.g., animals) by prompting them to complete sentences of the form “If an X were a Y, it would be a…” We refer to X as the “source domain” and Y as the “target domain”.
Participants were shown sentence of the form above, asked to provide an answer. They were then shown their original answers and asked to provide an explanation for why they answered in the way they did. These explanations were then coded according to different types of similarities that the participant seemed to rely on.

### *1) Initial Item Elicitation*

Our first step was to elicit words from which to use for the source items. We recruited 50 participants through Amazon’s Mechanical Turk (MTurk) crowdsourcing platform. Each participant was shown 12 superordinate terms denoting semantic domains: animals, jobs, sports, vehicles, cities, beverages, colours, branches of knowledge, words related to weather, musical instruments, fruit, and supernatural things. For each presented domain, participants were asked to provide 5 items. For example, prompted with “jobs”, a participant might list [“doctor”, “nurse”, “teacher”, “plumber”, “actor”. In the instructions participants were provided with the example: “If asked about insects, you might write down, fly, ant, bee, mosquito, and dragonfly.” These data serve two functions. First, they provide a principled list of items for use in Experiments 1-2. Second, they allow us to compute baseline probabilities of a person responding with, e.g., “cat”, when asked to list animals, which we use as a covariate in Experiment 2. Each participant was prompted with the 12 semantic domains in a random order.

### *2) Cross-domain mapping and explanations*

***Participants.***

We recruited an additional 80 participants from MTurk (40 Females, 40 Males, mean age = 40).

*Materials.*
We constructed 32 domain pairs, matching each domain with two or three other domains, e.g., animal → job/sports/beverage. We used a subset of all possible combinations of 12 semantic domains (12*11=132) because 1) Not all mapping are equally theoretically interesting (e.g., mapping animals to colours leads participants to just list colours characteristic of that animal) 2) Including all 132 unique pairs would require more data than we could realistically collect. For each unique domain pair, we randomly selected two to three statements (except for history → colour mapping, from which we selected six statements) of the form If an X [source item] were a Y [target domain] it would be a … [target response]. The source items were randomly chosen from the ten most frequent items of each domain during the phase of Initial Item Elicitation. For example, the job-animal trials included "If a doctor were an animal, they would be a…", "If an actor were an animal, they would be a …" etc. We ended up with 75 statements and we divided them into four trial lists such that each list contained about 19 trials. Each list was responded by a group of 20 participants.

*Procedure.*
Each participant was assigned to one of the four trial lists. Before beginning, they were given two example cross-domain mappings: a). If the sun was a job, what job would it be? (Example answer: King). B). If a cat were a branch of knowledge, what would it be? (Example answers: Philosophy/Psychology/Math).[3] After they were done with the mapping tasks, they were shown their previous answers, and for each, were asked to briefly explain why they responded in the way they did. For example, if to the prompt "If a dog were a sport, what sport would it be?" they answered "football", they would be later be prompted with: "When asked 'if a dog were a sport, what sport would it be?' you responded 'football'. Why? " Participants were instructed to be as precise as possible (e.g., avoid answers like "it's obvious", or "because it seemed correct). If they had no idea why they answered in the way they did, they could write "I was guessing".

*Response Coding.*
Two independent raters coded all responses into one or more of 7 categories using the following criteria:

1) *Phonological association*: e.g., if a bear were a beverage, it would be beer. Because they both starts with "b")
2) *Word associations*: e.g., If Beijing were a colour, it would be red. Because it reminds me of "red China"
3) *Perceptual similarity*: e.g., If rain were a musical instrument, it would be a drum. Because it sounds like a drum.
4) *Common mediators*: e.g., If a sunny day were a fruit, it would be an apple. Because apples are grown in the summer, and summer is sunny.
5) *Abstract alignment on certain dimensions*: e.g., if a cloudy day were a fruit it would be a banana. Because cloudy days are sad and mushy like a banana when it rots.
6) *Thematic association*: e.g., If a dog were a sport, it would be frisbee. Because they love games of fetch and frisbee.
7) *Guessing*: e.g., I was guessing/I don't know/It's the first thing that comes to mind.

If raters thought an explanation did not belong to any category listed above, they coded it as zero. (See also supplementary material for coding instruction)

Because multiple categories were possible, we used dummy coding for computing Cohen's kappa, common statistic for inter-rater agreement. [42]. After an initial coding pass, kappa was 0.62 indicating substantial agreement. The two raters then discussed disagreements until consensus were reached for all responses.

[3] To clarify what we meant by "branches of knowledge", we used it as part of the example in the instructions to cue people to what we mean by this phrase.

*Results and Discussion*

Figure 2 (left y-axis) shows the number of occurrences of each alignment type and the mean convergence rate for each type of explanation. Abstract Alignment was the most common type of for alignment (used for 718 out of 1520 different responses). It was the most common type of alignment for 26 out of 32 domain mapping.

We next computed convergence rates for all explanations. Convergence was operationalized as the proportion of participants who provided identical responses for each source-– target-domain pair. For example, if there were 20 total responses to the prompt "if a dog were a job, what job would it be?", and 5 people responded with "doctor", the convergence rate would be 5/20= 0.25.

Despite how common reliance on (seemingly) abstract dimensions was, these responses had relatively low convergence rates (11%), even lower than for explanations for which people said they were just guessing (12%). That is, although answers based on abstract alignment were very common, they were more variable than answers based on other types of similarity. As shown in Figure 2 (right y-axis), the highest convergence was achieved by responses coded as relying on word associations (21%), followed by perceptual similarity (18%), common mediator (15%), and thematic association (12%). The supplementary materials contains a version of Figure 2 broken down by each of the 32 source-target domain pairs (Figure S5). All the responses along with their coded similarity types can be found at https://osf.io/tkc84/).

These results provide initial evidence that when tasked with mapping between disparate semantic domains, people frequently rely on similarity along (relatively) abstract dimensions such as valence, size, and speed. The overall low convergence rates for these responses suggest that reliance on these dimensions is not as constraining compared to when people can use other strategies. For example, when asked to map "Demon" to a city, a plurality (6/20; 30%) responded with "Las Vegas" and of these, 4 mentioned that it was because Las Vegas is known as "Sin City" – a response coded as being based on word associations. When asked to map "Sunny day" to a colour, 50% (10/20) responded with "orange" and of these, 7 wrote that orange is the colour of the sun – a response coded as relying on perceptual similarity. So, when available, word associations and physical similarity can lead people to map between domains in relatively similar ways.

In the next experiment we sought to further understand the use of dimensional alignment in cross-domain mapping. Rather than relying on human coders, in Experiment 2 we predicted convergence rates from different, independently collected, similarity measures. This analysis allowed us to determine how semantic similarity along abstract dimensions contributes to cross-domain alignment as compared to other kinds of semantic similarity.

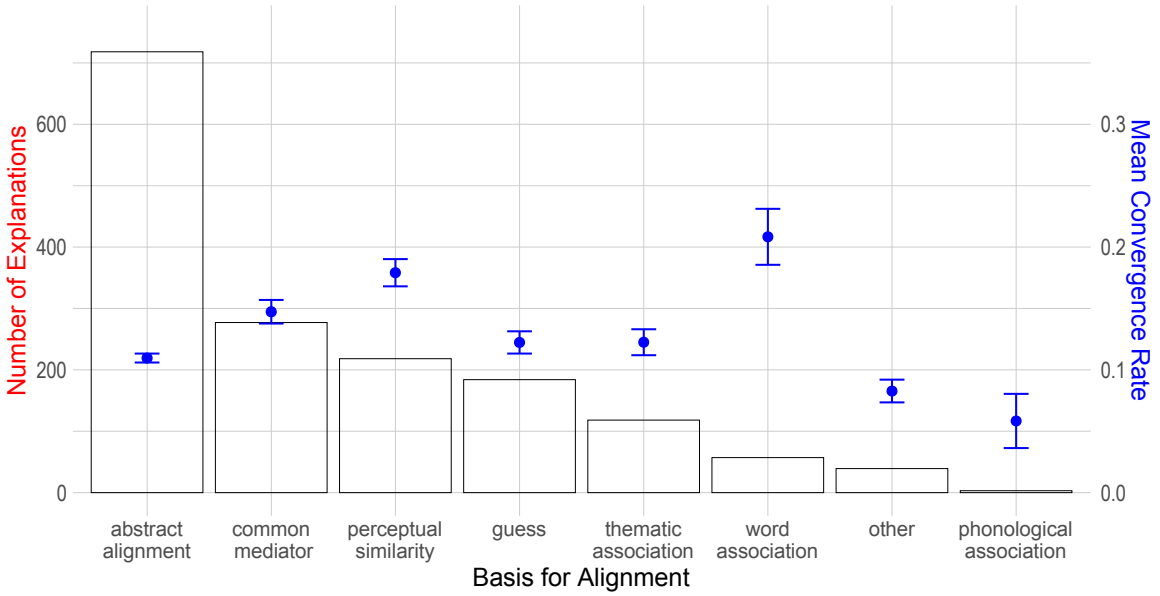


*Figure 2 Basis for alignment for cross-domain mapping task according to participants' explanation. Number of explanations for each basis are shown in red bars with error bars indicating 95% Binomial Confidence Interval; mean convergence rate are shown in blue circles with the standard error of the mean.*

## Experiment 2: How do different similarities predict convergence rates in questions like "If a nurse were an animal, they would be a(n) __"?

Experiment 1 provided qualitative evidence that when people are tasked with aligning different semantic domains, they frequently rely on—what looked to our raters to be—alignment based on abstract dimensions such as valence, size, and speed. However, people's explicit explanations do not always reflect the processes used to complete a task. In Experiment 2, we address this limitation by quantifying the relative contribution of similarity based on abstract dimensions in predicting convergence rates, and comparing this measure to more conventional forms of similarity such as those based on perceptual features and word associations.

Experiment 2 consisted of two parts: The first used the same type of cross-domain mapping task as in Experiment 1 (though without requiring participants to explain why they responded the way they did). 1). In the second, we computed a measure of dimensional similarity between the source items (the X in *If an X were a Y*) and a subset of the responses the participants produced in the first part.

This experiment was largely exploratory. We predicted though that participants would rely on dimensional similarity when generating their answers in which case higher dimensional similarity should predict greater convergence. However, we had no expectations about the predictive power of dimensional similarity compared to other forms of similarity, or even whether dimensional similarity would continue to predict convergence rates controlling for other, more familiar types of semantic similarity. Experiment 1 showed that abstract alignment was one of the most frequent strategies, yet was associated with lower convergence rates. It would therefore not be surprising to find that other forms of similarity are better predictors. Yet it may still be the case that people are consistently relying on this type of similarity, especially when other forms of similarity are not available.

### *1) Cross-domain mapping*

***Participants.***

We recruited 165 undergraduate students (104 Females, 59 Males, 2 Other, mean age = 18.6) from the University of Wisconsin-Madison psychology participant pool. We excluded four people for failing more than 2 out of four attention checks.

*Materials.*
For each of 32 domain pairs, we constructed 10 statements of the form "If an X [source item] were a Y [target domain] it would be a … [target response]. " Source items were the 10 most frequent items for each domain elicited in Experiment 1, Initial Item Elicitation. Each included pair also included its inverse, e.g., both animal → job and job → animal were included. We did not ask participants about both directions of mapping (e.g., answer both animal → job mappings and job → animal mappings) because thinking about one may bias the second, e.g., if they mapped dog → doctor, they might also map doctor → dog because of their earlier dog → doctor response. We divided the 320 stimuli into four lists such that each list contained 80 trials representing 16 domain pairs and 5 trials from each unique domain-pair.

*Procedure.*
The procedure was identical to the open-ended cross-domain mapping task in Experiment 1, except we didn't ask for explanation in Experiment 2. Each participant was assigned to one of the four 80-item lists. Participants were also explicitly told that they should try to provide single-word responses and avoid giving the same response multiple times for the same type of question. To further discourage repeated answers (e.g., responding with "guitar" whenever asked to map different items to musical instruments), we blocked the prompts by cross-domain mappings, i.e., all the trials under the same type of cross-domain mappings, e.g., animal → instrument were grouped together. A given participant might see the following trial sequence: "If a doctor were a musical instrument…", "If an actor were a musical instrument…", "If a nurse were a musical instrument…", etc.

## *2) Semantic Differential Ratings*
Semantic Differential ratings [23,43] ask people to rate concepts on scales that are anchored by two polar adjectives (e.g., "good-evil", "bright-dark"). The basic idea is that important aspects of a concept's meaning are its location in the space defined by these dimensions. The distance between concepts computed in this semantic differential space has been found to be a good measure of semantic similarity, especially in studies of metaphor and emotions [44–48].

We sought to obtain semantic differential ratings for the source words (the word in the X position in the "If the X were a Y it would be a…" prompt) and for people's responses. These semantic differential ratings were then aligned to obtain a semantic differential alignment score (Fig 1E). The total dataset (including all target words and answers that were produced by at least 1 person) comprised 2252 unique responses. Recruiting 10 raters per word and collecting 10 words per rater (on 21 dimensions per word) would involve recruiting over 2000 raters which was not feasible. Initial analysis of the data from the cross-domain mapping task (data and code are accessible at https://osf.io/tkc84/) revealed many trials to have high convergence for obvious (and uninteresting) reasons (cow→beverage = milk, ghost→colour = white, goblin→colour = green). It was of more interest to collect semantic differential ratings for words comprising less obvious cross-domain mappings. With this in mind, we focused on subset of trials that included 130 concepts from three semantic domains: animals (56 words), jobs (54 words), and musical instruments (28 words).

We elicited semantic differentials by prompting participants with a given word and asking them to position it on 21 dimensions anchored by two polar adjectives (e.g., position "dog" on a dimension of "bad (1)-good (7)". Original work on semantic differentials [43] revealed that many semantic dimensions group into three dominant factors: An Evaluative factor (represented by scales such as bad-good, unpleasant-pleasant, positive-negative), a Potency factor (represented by scales such as strong-weak, heavy-light, hard-soft), and an Activity factor (represented by scales such as fast-slow, active-passive, and excitable-calm). We selected 15 dimensions based on the Evaluative-Potency-Activity (EPA) framework and consulted other studies using semantic differentials [49,50] to add 6 dimensions that did not obviously load on only one of these factors, (see Table 1).

*Table 1 The 21 dimensions used for computing dimensional similarity between source and target in the cross-domain mapping task.*

| Evaluation | Potency | Activity | Other |
|---|---|---|---|
| Bad/good | Weak/strong | Passive/active | Small/large |
| Cheap/expensive | Soft/hard | Lazy/industrious | Dry/wet |
| Stupid/smart | Light/heavy | Dull/exciting | Masculine/feminine |

| | | | |
|---|---|---|---|
| Ugly/beautiful | Silent/loud | Slow/fast | Religious/secular |
| Cruel/kind | Mild/aggressive | | Cold/hot |
| Unpleasant/pleasant | | | Old/young |

We collected semantic differentials from 170 participants recruited through Amazon's MTurk crowdsourcing platform. They were directed to a Qualtrics survey, and each participant rated 10 concepts on 21 dimensions on a 1-7 scale. Each scale was anchored by two polar adjectives (e.g., 1=Bad, 7=Good). Each concept-dimension pair (e.g., dog: bad↔good) was rated by at least 10 raters.

After each response, participants were also asked to indicate their confidence level in rating from just guessing (1) to very confident (5). These ratings were used to help construct stimuli for Experiment 3.

## Outcomes and Predictors

We sought to predict the likelihood that people answer in a certain way using various measures of similarity (e.g., sensorimotor similarity, co-occurrence in context, abstract dimensional alignment, feature similarity) between X and the provided answer as a way to understand what kinds of similarities people rely on when performing cross-domain mappings.

*Outcome Measure.*

As in Experiment 1, we computed convergence rates for each prompt (a source-item – target-domain pair) as the proportion of people who provided identical responses.

*Predictors.*

*Baseline probability:* To the extent that some items are more typical (or salient) of their respective category, people are more likely to produce them as a response regardless of the source. For example, to the extent that "dog" is a more typical animal than "platypus", we expect more "dog" than "platypus" responses for "If X were an animal, it would be a…". We therefore need to take into account the baseline likelihood of a person producing a given item. We did this by using baseline frequency as one of the predictors, defined as the proportion an item was mentioned by participants in the item elicitation task. For example, "dog" comprised 38 out of 250 animal responses we recorded making its baseline probability 15.2%.

*Sensorimotor similarity:* Cross-domain concepts could have similar perceptual/sensorimotor profiles (e.g., drum and thunder belong to different domains but are strikingly similar in auditory representation). We obtained sensorimotor profiles for 713 words in our dataset from the Lancaster Sensorimotor norms [28]. These norms contain sensorimotor strength ratings across six perceptual modalities (touch, hearing, smell, taste, vision, and interoception) and five action effectors (mouth/throat, hand/arm, foot/leg, head excluding mouth/throat, and torso). We computed a similarity score between the source word and each answer by computing the cosine-based similarity between the 11-dimensional vectors corresponding to the source words and answers. Zero cosine similarity corresponds to words whose sensorimotor profiles are orthogonal while larger similarities indicate that the words have related sensorimotor experience.

*Extended word associations using random walks*: We obtained word associations from the "small world of words" (SWOW) mega-study [39]. Participants in that study were cued with one word and asked to provide up to three words the word made them think of. Instead of using the first-order word association based on *direct* associations (i.e., the probability of responding with the target word when given the source word as a cue), we examined *indirect* associations using a decaying random walk process [40]. Compared to direct (first-order) word association, extended word associations using random walks can identify both direct associates and more distance associations between words that are not directly linked [51]. Consistent with a spreading activation mechanism, the random walk considers both immediate neighbours of a word and indirect paths via chains of associates. As random walk distance between two nodes increases, the probability that a random walk starting at $word_1$ and arriving at $word_2$ decreases. For example, no one produced "hunter" when prompted with "tiger" and no one produced "lawyer" when prompted with "violin". Yet, tiger→hunter has a much shorter random walk distance thus higher probability of arrival (0.23) than violin→lawyer (0.004).

*Word embedding similarities*: Word embeddings allow for computing similarity between any pairs of words based on the similarity of the linguistic contexts in which they occur [52,53]. We obtained the semantic similarity between source concept (e.g., doctor) and participants' answers (e.g., "piano" when mapping "doctor" to a musical instrument) by computing the cosine distance between embeddings that represent source concepts and target concepts derived from applying the *fastText* algorithm to the Wikipedia corpus [54].

*Dimensional similarities*: We obtained concepts' abstract meaning using the semantic differential technique described earlier. We computed the average rating of concepts on each dimension, which was used as entries for a 21-dimensional vector representing each word's dimensional profile (e.g., Fig 1E). This allowed us to measure dimensional difference between two words by comparing their profiles which we did using cosine similarity.

*Feature-based similarities:* The vast majority of cross-mapping trials included in this analysis had no shared semantic features as listed in existing semantic-feature norms [2,3], which meant feature-based similarity was mostly zero. We therefore did not use semantic feature-based similarity as a predictor.

## Results

To understand what kinds of similarity people rely on when performing cross-domain mappings, we predicted convergence rates (the frequency of a provided response) from baseline convergence, word embedding similarity (cosine similarity), extended word association using random walks, sensorimotor-based similarity (cosine similarity), and semantic-differential based dimensional similarity. These similarities were entered as predictors in mixed-effects linear regression using lme4 package in R [55]. The model included by-participant and by-item random intercepts where items were defined as unique source-concept → pairs (e.g., dog → job). All predictors were zero-centered. Below, we present results only for trials for which we obtained semantic differential ratings. Analyses of all trials using the other five predictors—baseline probability, two-word association measures, word-embedding similarity, and sensorimotor similarity—are presented in the Supplementary materials (see Figure S1).

The results are shown in Figure 3A. Baseline probability was, unsurprisingly, a strong predictor ($\beta = 0.40, se = 0.05, \mathrm{t} = 7.50, p < .001$). For example, dog was the most frequently named animal in the item elicitation task (baseline probability = 15.2%), and people frequently mapped things to "dog" (on average, X → dog had an 8% convergence rate). Sensorimotor similarity was not a significant predictor of convergence ($\beta = -0.05, se = 0.05, \mathrm{t} = -1.10, p = 0.27$). This was surprising given that perceptual similarity was associated with the second highest convergence rate in Experiment 1. It's probably due to the nature of such similarity relies on capturing perceptual details that are not very likely to be reflected in sensorimotor norms. Convergence was also predicted by a word-association-based measure—the length of a random walk from source-to-target: shorter distances corresponded to greater convergence ($\beta = 0.22, se = 0.07, \mathrm{t} = 3.11, p < .01$). Word embedding similarity also did not significantly predict convergence ($\beta = 0.07, se = 0.05, \mathrm{t} = 1.3, p = 0.13$). It was, however, moderately correlated (r=.43) with random-walk based word associations. Excluding random-walk word-association as a predictor led word-embedding similarity to become a significant predictor ($\beta = 0.17, se = 0.07, \mathrm{t} = 2.43, p = .01$). Excluding word-embedding similarity as a predictor led random-walk word-association to become a stronger predictor ($\beta = 0.23, se = 0.07, \mathrm{t} = 3.45, p < .001$). Lastly, dimensional similarity was a significant predictor of convergence in free response. This was true both when it was the only predictor other than baseline probability ($\beta = 0.13, se = 0.04, \mathrm{t} = 2.39, p = .01$) , and when it was entered alongside all the other predictors ($\beta = 0.09, se = 0.04, \mathrm{t} = 2.12, p = .03$) suggesting that people might do dimensional-based alignment such that concepts with similar projections on semantic differential dimensions were easier to be mapped onto each other.

To summarize, in Experiment 2 we found that cross-domain alignment in a free-response task could be reliably predicted by 1) baseline frequencies, concepts with higher convergence without mapping also tend to have higher convergence with mapping, e.g., dog has the highest frequency when participants were asked to simply name an animal, and it's also a highly convergent answer for various mappings towards animal. 2) extended word association using random walks: words with higher contextual associations also tended to have better alignment, e.g., cow → farmer/lion → king in animal → job mapping. 3) dimensional similarity: people tended to align words from different semantic domains based on their similarity along (relatively) abstract dimensions. Consistent with Experiment 1, dimensional similarity leads to convergence to some extent by constraining the

answers which are aligned on certain abstract dimensions. Despite its popularity as a major basis for alignment, it does not lead to the most converged answers --- and it probably shouldn't be. We'll further address this seeming paradox in the discussion.

## Experiment 3: How do different similarities predict goodness of statements "If a nurse were an animal, they would be a cat."?

In Experiment 2, we found that when asked to freely map a concept (e.g., dog) to a different semantic domain (e.g., job), people readily obliged, producing varied, but partially predictable answers. The design of the task limited us to studying cross-domain mappings that were produced by our participants. Another limitation is that participants sometimes produced mappings based on their stored knowledge (e.g., cow → milk), which weakens the predictive power of abstract alignment. In Experiment 3 we constructed our own statements (e.g., "If a dog were a musical instrument, it would be a guitar") allowing us to parametrically vary the contribution of different semantic similarity measures. We then asked participants to indicate for each sentence how good they thought the mapping was. The task was similar to the aptness rating task used in [44]. This method allowed us to better isolate the contribution of different types of similarity and also allowed us to focus on cross-domain mappings that cannot be answered by stored knowledge, and to examine which specific dimensions (e.g., bad↔good, small↔large, masculine↔feminine) best predicted people's endorsements of the goodness of each mapping. For example, do dimensions that are more clearly applicable to a certain domain (e.g., animals have literal size) predict alignment more than dimensions whose appropriateness is more metaphorical (the size of a job is a more abstract, metaphorical construct).

*Participants.* We recruited 132 participants from Amazon's Mechanical Turk (61 Females, 67 Males, 4 Other gender, mean age = 36). We planned to exclude those who failed more than 1 (out of 3) catch trials, but none met this criterion.

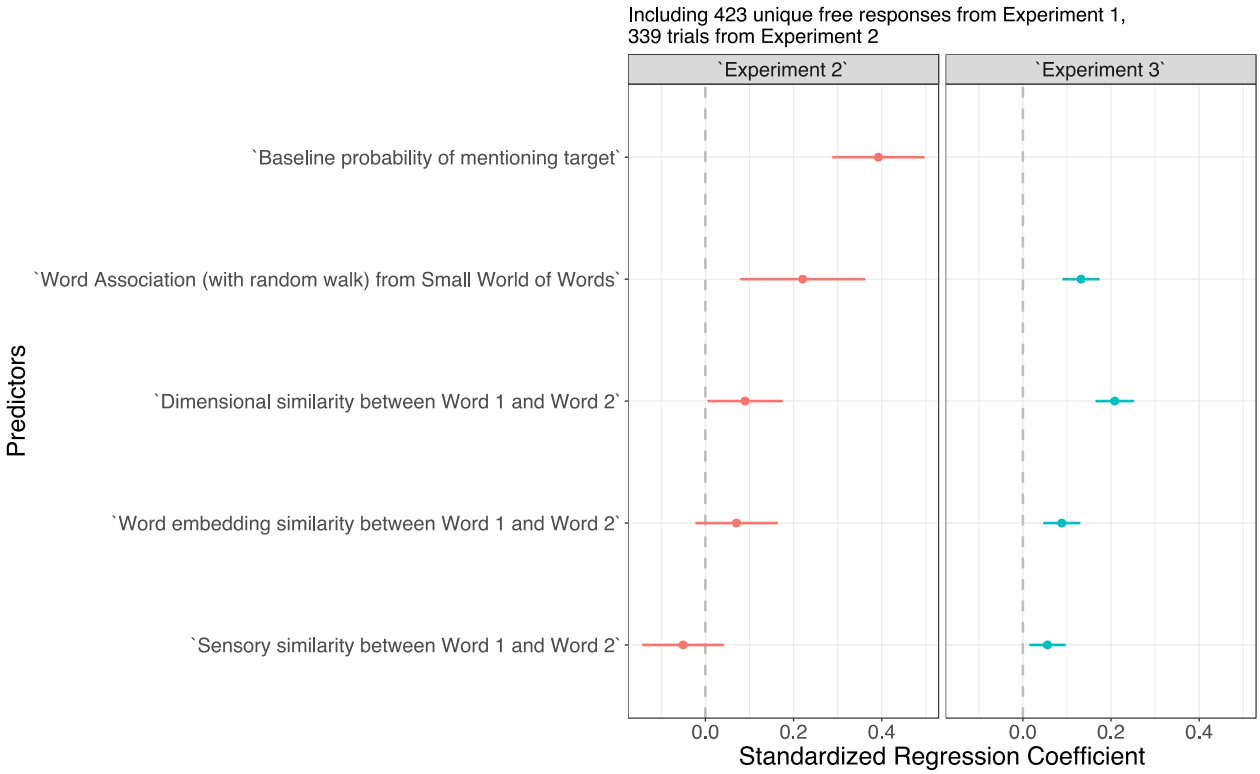


*Figure 3. Standardized multiple regression coefficients and 95% Confidence Interval for all predictors used in Experiment 2 (free-response task) and Experiment 3 (goodness rating task). Positive values indicate that greater similarity predicts that the answer was provided by a larger proportion of participants (Experiment 2) or had a high goodness rating (Experiment 3).*

***Materials.***

*Choice of Source words*

We selected 15 words from each of the three domains used in Experiment 2 (animals, jobs, musical instruments). For each domain, we chose the 10 words with highest overall confidence ratings in the semantic

differential task and 5 words with the lowest confidence ratings. For example, when rating various jobs along the 21 dimensions listed in Table 1, people were most confident in their ratings of "teacher" and least confident in their rating of "banker".

*Choice of Target words*

For each source word, we selected a subset of the target words from those produced by participants in Experiment 2, parametrically varying the semantic differential distance between the source and target. For example, for statements mapping "dove" to a musical instrument, we selected four targets: "wind chimes" (which had the smallest distance, i.e., greatest alignment along the 21 tested dimensions shown in Table 1), "bell" (second quartile of semantic distance), "guitar" (third quartile), and "organ" (largest semantic distance).

One shortcoming of this approach is that a short distance between two words can be obtained when participants provide intermediate ratings across many dimensions which they tend to do when they are not sure of how to respond, e.g., when rating an animal on a religious↔secular dimension. We therefore also computed the distance using only the 7 dimensions with the highest confidence for each domain (e.g., for animals, these would include small↔large and slow↔fast, but exclude cold↔hot and religious↔secular) and added the target word closest to the source word on these high-confidence dimensions if it was not already selected. This procedure generated 417 unique statements.

***Procedure**:* Each participant was shown approximately 70 statements of the form "If a(n) X were a Y, it would be a y" and asked to indicate "how good is this comparison" on a scale of 1 (terrible) to 7 (excellent). The comparisons included six source-to-target mapping pairs: animal ↔ job, animal ↔ instrument, job ↔ instrument). Each participant was shown only one type of mapping (e.g., only animal→job).

The instructions included two examples: (1) If a cat were a job, it would be a manager, and (2) If an apple were an animal, it would be a giraffe." The instructions acknowledged the oddness of the task, suggesting that despite the oddness, some of the comparisons are better than others: "Although these are both odd comparisons, you can probably more easily imagine a cat as a manager than an apple as a giraffe." In addition to the experimental stimuli, three catch trials were interspersed to ensure participants were paying attention, e.g., when asked "Please choose the right-most option", participants should choose the right-most option.

***Measures**:* We used the same predictors as in Experiment 2, with two exceptions. First, unlike Experiment 1 and 2 which used a free-response measure, Experiment 3 uses fixed materials for all participants obviating the need to control for baseline frequency; there's no theoretical reason to believe a mapping should be rated as better just because concepts involved are more frequent or come from small (e.g., colours) vs. larger (e.g., animal) domains.

***Results***: To understand what makes a cross-domain mapping a good one, we predicted people's goodness ratings from semantic differential similarity, word embedding similarity, extended word association using random walks, and sensorimotor similarity using an identical analytic framework as in Experiment 2.

The regression coefficients are shown in Figure 3. Extended word associations (random walks in word-association space) were positively associated with goodness ($\beta = .13, se = .02, p < .001$) as were word embedding-based similarities between the source and target words ($\beta = .09, se = .02, p < .001$). Unlike in Experiment 2, sensorimotor-based similarity was a significant, albeit weak, predictor ($\beta = .05, se = .02, p < .01$): words with more similar sensorimotor norms were judged to be better matches. Controlling for these three types of similarity, dimensional similarity was again a significant predictor of goodness rating ($\beta = .21, se = .02, p < .001$), with a substantially greater effect size than the other predictors. The effect of source and target dimensional similarity on goodness is also shown in Figure 4 which shows mean goodness averages for the four trial types that have the same source and target domains but vary the dimensional alignment between the source and target concepts.

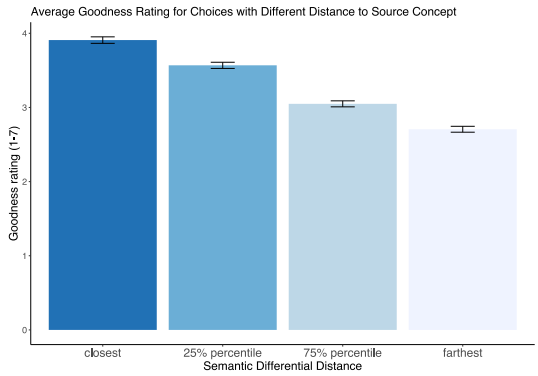


*Figure 4. Average goodness rating for cross-domain mappings as a function of semantic differential distance between source and target words. As dimensional similarity decreased (from closest to farthest), so did people's goodness ratings. Error bars correspond to standard errors of goodness rating for each quartile.*

The results so far show that pairs of words with more similar profiles on the 21 dimensions shown in Table 1 are seen as better matches in a cross-domain mapping task. But which dimensions contribute most to this mapping? Perhaps it is determined primarily by similar valence. Or perhaps valence is irrelevant and it is the relatively more concrete dimensions such as size that are most important. Alternatively, it may be that which dimensions are most important depend on the semantic domain such that size is only relevant if mapping between domains that have literal size (e.g., animals and musical instruments, but not jobs).

Recall that in addition to including target words that are closest along all 21 dimensions, we also included words that were closest only on the 7 dimensions that led to the highest confidence responses. Interestingly, statements containing source and targets words closest on all 21 dimensions (e.g., if a priest were a musical instrument, he/she would be a harp) were rated as better (M=3.91) than statements containing source and target words matched on dimensions with the highest confidence ratings (e.g., if a priest were a musical instrument, he/she would be a bass) (M=3.61), ($b = .3, se = .07, \mathrm{t} = -4.33, p < .001$), suggesting that the "low-confidence" dimensions may be playing a non-negligible part in cross-domain alignment.

We next examined *which* dimensions contributed to the various cross-domain mappings. Intuitively, dimensions that make more literal sense for a specific domain may play a larger role. For example, it makes more sense to people to arrange various animals or instruments on a small ↔ large dimension, compared to a cheap ↔ expensive dimension. Is size more important when people map between items that have literal size (i.e., animals and musical instruments), or is it also implicated in the musical instrument↔jobs and animals↔jobs mappings, cases in which the source or target items do not (it would seem) include size as part of their representation.

Rather than relying on confidence ratings from the same participants who generated the semantic differential ratings (which introduces some circularity to the analysis), we recruited an additional group of 40 participants from Amazon Mechanical Turk and asked them to indicate how meaningful is it to arrange "animals","jobs","musical instruments" on each dimension. The "meaningfulness" ranged from "makes no sense at all" (0) to makes a lot of sense (5). The meaningfulness of each dimension as applied to animals, jobs, and musical instruments is shown in Figure S 2.

We next predicted participant's goodness ratings from (1) the alignment on each individual dimension and correlated regression coefficients with the meaningfulness of that dimension as it applied to the source domain. As shown in Figure 5 the majority of dimensions are predicting cross-domain mappings in the

expected direction[4]: the closer participants' rating of two concepts on a particular dimension are, the better two concepts are aligned. Overall, dimensions that are most predictive are also more meaningful ($\beta = .32, se = .09, \mathrm{t} = 3.773,\ p < .001$). For example, slow/fast was one of the most important dimensions for predicting animal→job mappings, while also being one of the most meaningful dimensions for animals. However, there were many exceptions. For example, cruel↔kind was rated as one of the least meaningful dimensions for both jobs and musical instruments, but it was among the most predictive dimensions when mapping musical instruments to jobs. We also checked if a dimension was more predictive if it had more similar meaningfulness in the source and target domains. For example, is size more predictive when mapping between musical instruments and animals given that it is meaningful to both, as compared to when the source or target domain includes jobs for which the size dimension is less meaningful. Difference in meaningfulness was unrelated to the importance of the dimension ($\beta = -.12, se = .09, \mathrm{t} = -1.33,\ p = 0.2$). (see Figure S3).

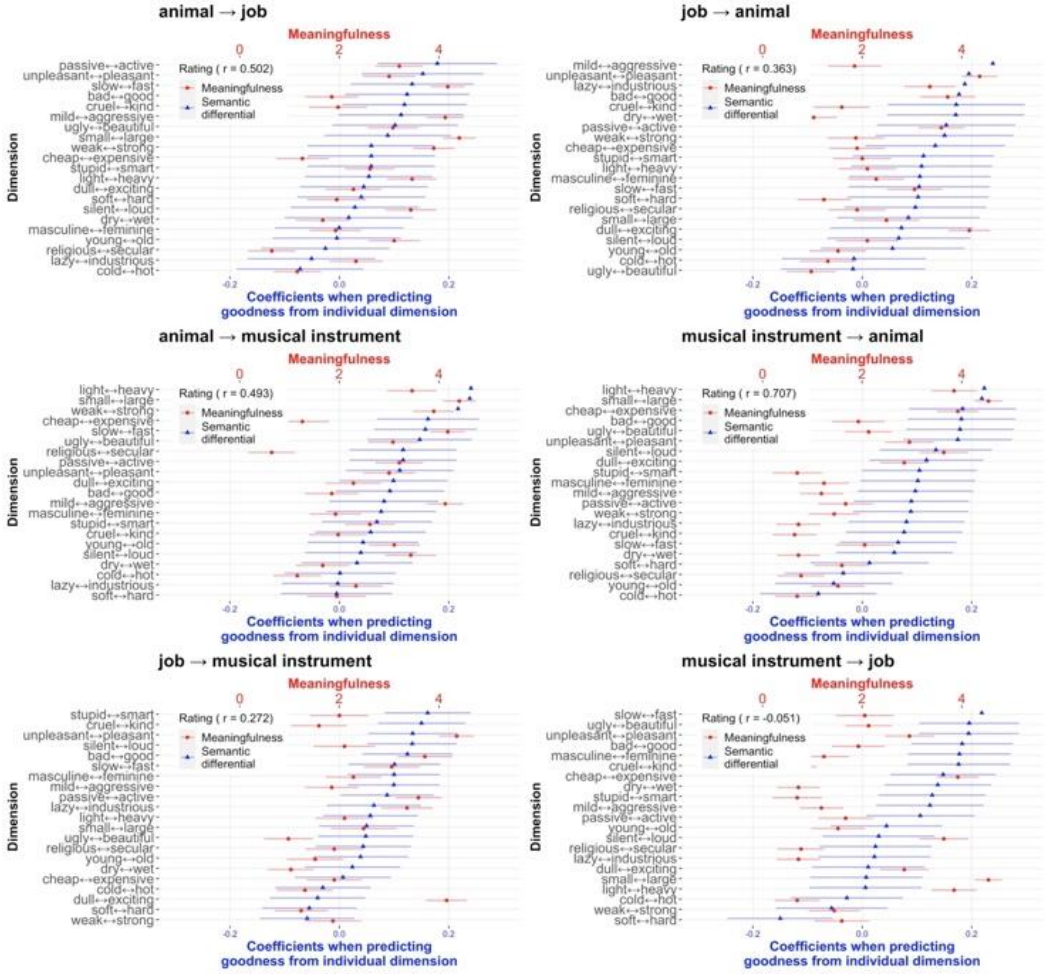


*Figure 5. Coefficient plot for 21 dimensions when each dimension is used as an independent predictor (blue triangles) and meaningfulness of the source dimension (red circles) as rated by independent raters. Error bars show 95% CI of the coefficient estimate.*

[4] We observed several cases such that aligning *less* well on certain dimensions predicted higher goodness ratings. We think this is because those dimensions are in "conflict" with other dimensions. For example, weak ↔ strong dimension in jobs →musical instrument negatively predicts goodness rating. This is potentially because two concepts that align well on weak ↔ strong tend to align worse on many other dimensions. E.g., thief and piano are both rated as moderately "strong", but thief is rated as bad, silent, aggressive, and stupid, while piano is rated as good, loud, mild, and smart.

To further examine the relative importance of various dimensions to cross-domain mappings, we predicted the goodness rating for each trial from alignment on each dimension simultaneously, allowing us to determine their unique contributions. We removed one predictor at a time and calculated the reduced chi-squared statistic. To avoid multicollinearity from including highly correlated dimensions, we computed composite scores for dimensions with correlations greater than $r$=0.7. Bad↔good, unpleasant↔pleasant, cruel↔kind, and ugly↔beautiful were collapsed into a single valence dimension. Small↔large, and light↔heavy were combined into a single magnitude dimension. Within each domain mapping, we ranked dimensions from the least meaningful (first quartile) to the most meaningful (fourth quartile). As shown in Figure 6 removing less meaningful dimensions reduced model fit to a similar extent as removing more meaningful dimensions. The two analyses just described appear to contradict one another. Figure 5 suggests that there's a moderately positive relationship between meaningfulness and "importance" (as judged by standardized coefficient size). Figure 6 suggests that both meaningful and meaningless dimensions are roughly equally important. The difference between the two analyses is that in the first analysis (shown in Figure 5) each alignment on each dimension is entered into the model as a sole predictor. In the second analysis (shown in Figure 6) all dimensions are entered into the model simultaneously and so the contribution excludes shared variance. In the latter case, removing highly meaningful dimensions reduces model fit to a similar extent as removing less meaningful dimensions because the remaining predictors make up for much of the removed variance, suggesting that the predictive power of dimensional alignment is quite distributed rather than being concentrated in the most meaningful dimensions.

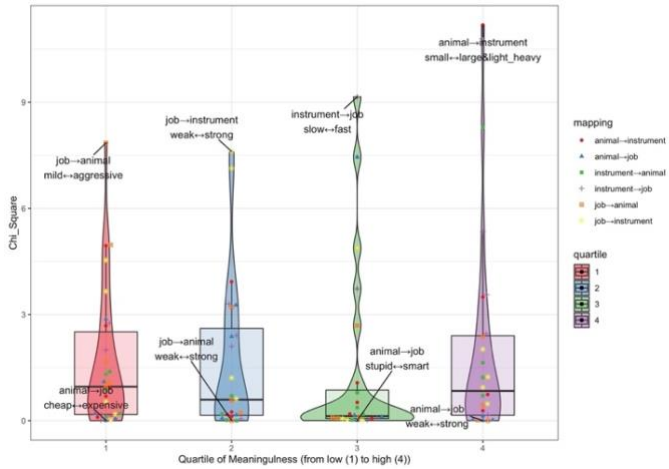


*Figure 6. The relationship between dimensional meaningfulness (with respect to source domain) and importance of that dimension to model fit. X-axis shows quartile of meaningfulness from least meaningful (1) to most meaningful (4). Y-axis shows the extent to which removing individual dimensions reduced model fit (chi-square statistic). Each dot represents a unique domain mapping-by-dimension combination. Dimensions most important (high Chi-square values) and least important within each quartile are labelled.*

In Experiment 3 we found that when people are asked to evaluate the "goodness" of cross-domain mappings rather than generate them *de novo*, people rate mappings as better if two concepts have 1) higher contextual associations (as reflected by larger extended word association using random walk and word embedding cosine similarities generated from large corpus) 2) higher sensorimotor similarities (albeit a weak effect) and 3) most critically, higher similarities along relatively abstract dimensions. Although more meaningful dimensions (e.g., animal size vs. religiosity) accounted for more variance on their own, removing less meaningful dimensions from the full model was as detrimental as removing more meaningful dimensions suggesting that these seemingly meaningless dimensions contribute to cross-domain alignment.

## General Discussion

We asked people to perform an odd task: map from one domain (e.g. musical instruments) to another (e.g., jobs). Despite producing many idiosyncratic answers (e.g., *saxophone → web developer*, *trumpet → works in dog training places*), people often provided similar, in many cases identical answers. For example, when asked "If a flute were a job" 20% of people said "teacher", a much higher rate than the baseline probability (7%) of listing teacher as a job when asked to list jobs. Mappings between some domains are easy to explain. For example, when asked to map animals to colours, people predictably list the colour characteristic of that animal, treating the task as a type of constrained feature-elicitation task. However, this strategy is ineffective when

mapping between, e.g., jobs and musical instruments and yet people are able to converge surprisingly often. How do they do this and what might it mean for how we represent concepts/word meanings?

In Experiment 1 we found abstract alignment to be the most popular basis for alignment compared to other alignment types such as word association, perceptual similarity, common mediator, etc. We then showed across two experiments that a consistent predictor of people's responses when asked to generate (Experiment 2) or evaluate (Experiment 3) cross-domain mappings, is alignment between source and target concepts in an abstract semantic space. This semantic space is not only defined by abstract semantic dimensions such as valence (evaluation), activity, and potency—the three general semantic factors originally identified by Osgood [43], but also by dimensions such as gender, religiosity, size, and age – even when these do not apply in a literal way, e.g., jobs do not have a literal size. Removing these seemingly meaningless dimensions from the model reduced model fit to a similar extent as removing the more meaningful dimensions.[5] Put another way: when asked to indicate their confidence in rating musical instruments along stupid↔smart, people indicate that they have low confidence in their answers and that stupid↔smart is not a meaningful dimension for musical instruments. However, people's placements along these ostensibly meaningless dimensions are actually somewhat systematic (see Figure S4A). Their answers may be informed by stereotypes about what kind of people are likely to play different instruments, or because musical genres that use certain instruments are accorded high prestige (e.g., violins, harps, and flutes being used in classical music). Either way, when mapping from, e.g., musical instruments to jobs, people were more likely to list jobs that were rated (by other participants) as having similar values along the stupid↔smart dimension (Exp. 2) and rated mappings with similar values along this dimension as being better (Exp. 3). It is interesting to compare how people rate musical instruments along the stupid↔smart dimension to soft↔hard, another dimension they indicate (see Figure S3 in supplementary material) to not be very meaningful for musical instruments. People also produce systematic ratings, rating bugles and trombones as hard, but moving downward (from hard to soft) one can see interesting cases of bimodality (see Figure S4B): some people think flutes are soft; others think they are hard; the same goes for cellos and pianos. As confirmed by subsequent investigations, these differences are due to some people interpreting the dimension in a more literal sense—the hardness of the material from which the instrument is made—while others interpret it in a more abstract, metaphorical sense—the hardness of the sound the instrument makes.[6] In short: the reason there is signal in seemingly meaningless dimension is that they are not, in fact, meaningless.

Dimensional similarity was found to be a type of similarity that predict convergence rates in Experiment 2, and the best predictor of goodness ratings in Experiment 3. At the same time, explanations relying on dimensional similarity offered in Experiment 1, despite being very frequent, had relatively low convergence, i.e., there was a larger variety of such responses. Is there a paradox here? We think not. What is likely happening is that some types of similarities are more constraining than others. For example, constraining answers to "If a cat were a musical instrument, it would be a___" to musical instruments having the largest word association to "cat", might generate high convergence (i.e., similar responses). Constraining answers to this prompt to musical instruments having the largest dimensional similarity to "cat" might lead to lower mean convergence insofar as

---

[5] Explaining cross-domain mapping using semantic differentials is reminiscent of an theory of metaphor known as domain-interaction theory which proposed that the dimension-structure of the source domain is mapped onto the dimensional structure of the target. [44–47]. For example, in *a wolf is a shark among fish, wolf* is close to *shark* in a two-dimensional factor space (a prestige factor, and a power-aggression factor), implying that both are high in aggression and strength. Our findings can be thought of as a larger-scale test of this idea, establishing the specific role of dimensional alignment (while controlling for other measures of similarity), and investigating the contribution of dimensions that are seemingly irrelevant for the mapping.

[6] We observed similar bimodality when asking people to rate jobs on a low↔high dimension: some people rated it according to a metaphorical interpretation (low vs. high prestige) while others rated it according to whether the job is literally high (e.g., pilot) vs. low (e.g., miner).

there are different musical instruments may have equally high dimensional similarity, albeit on different dimensions.

Beyond demonstrating that people are able to map between different semantic dimensions with apparent ease, and that people's mappings are partially predictable, our results suggest that concrete concepts/word meanings like "piano" and "giraffe" are readily placed along relatively abstract dimensions even when they seem quite meaningless, such as placing musical instruments along stupid↔smart or animals along religious↔secular. One interpretation is that the representational space within which these concepts reside contains task-independent information corresponding to these dimensions. Alternatively, the smartness of an instrument or the religiosity of an animal may only be defined when people are asked to make an explicit dimension rating as they were when indicating where the words fall along various semantic dimensions. It may also emerge as a consequence of people projecting conceptual representations into a common semantic space for the purposes of mapping between disparate domains. We cannot distinguish between these possibilities with our current data.

A final question is where does information allowing people to project, e.g., musical instruments onto the stupid↔smart dimension come from? We speculated earlier that it may reflect social stereotypes, but this just pushes the question up a level: how do people learn those stereotypes? We were curious whether some of the information may be encoded directly in the statistics of language such that, e.g., "elephant" is more likely to occur in shared contexts with words like "large" than words like "small" or, as shown previously, different jobs are differentially associated with masculine and feminine contexts [56].

To find out if people's projections of words onto the 21 dimensions listed in Table 1 were learnable purely from language, we used the method described in [57] [see also 56, and 58 for earlier, but more limited uses of this technique] to project words onto the dimension formed by the anchor labels. For example, to project "elephant" onto the small↔large dimension, we compute the cosine similarity between the vector representing "elephant" and the *size* vector obtained by subtracting "large" from "small". The correlations between each word's place along the 21 dimensions as rated by people and estimated from language statistics are shown in Table 2. The largely positive correlations suggest that distributional patterns in language contain information that people could, in principle, use to learn where various words fall on these dimensions (Table S1 shows the correlations broken down by jobs, animals, and musical instruments).

The existence of these positive correlations does not mean that people learn this information from language. For information readily accessible from vision (e.g., one's knowledge that elephants are large animals), resorting to learning from language seems unnecessary. For other features, however, such as the gender of jobs (or, for that matter, the stereotypical gender of animals; people rate dinosaurs as predominantly male and hummingbirds as predominantly female), exposure to linguistic contexts may be playing an important role. Even when the information is, in principle, readily available through perception, exposure to language statistics may be playing an important role in aligning people's conceptual representations in the face of varied perceptual experiences, helping to explain why blind people's knowledge of visual information is surprisingly similar to sighted people's [59–61].

*Table 2. Correlations between embedding-simulated dimensions and human-rated dimensions.*

| Dimension | Correlation between word-embedding-based semantic differentials and human ratings | p value |
|---|---|---|
| Masculine/feminine | 0.52 | <.001 |
| Silent/loud | 0.50 | <.001 |
| Mild/aggressive | 0.46 | <.001 |
| Lazy/industrious | 0.46 | <.001 |
| Ugly/beautiful | 0.39 | <.001 |
| Unpleasant/pleasant | 0.39 | <.001 |
| Small/large | 0.37 | <.001 |
| Bad/good | 0.37 | <.001 |
| Dull/exciting | 0.36 | <.001 |
| Stupid/smart | 0.33 | <.001 |

| | | |
|---|---|---|
| Young/old | 0.29 | <.001 |
| Slow/fast | 0.28 | <.001 |
| Cheap/expensive | 0.28 | <.001 |
| Dry/wet | 0.26 | <.001 |
| Weak/strong | 0.25 | .01 |
| Light/heavy | 0.25 | .01 |
| Cold/hot | 0.15 | .11 |
| Passive/active | 0.15 | .12 |
| Cruel/kind | 0.14 | .42 |
| Soft/hard | 0.07 | .48 |
| Religious/secular | -0.25 | .01 |

## Conclusion

How do people project concrete concepts between semantic domains? Using two free-response tasks and a goodness-rating task, we show that people do cross-domain mappings between relatively concrete concepts by aligning on abstract dimensions. Cross-domain mappings that have been studied in the context of metaphors have emphasized the importance of mapping between a more concrete source domain onto a more abstract target domain so that the abstract concept can be understood as an entailment of a more concrete concept [62]. Our results show that people tend to converge in how they map between concrete domains and that they do so by either making use of abstract information encoded as part of the concrete concepts that are being mapped, or by actively projecting these concepts into a more abstract semantic space so that they can be aligned.

## Acknowledgments

We would like to acknowledge Hannah Bombeck, Annie Gense, Henry Barford for help with response coding and data collection.

# References


1. Vigliocco G, Vinson DP, Lewis W, Garrett MF. 2004 Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognit. Psychol.* **48**, 422–488. (doi:10.1016/j.cogpsych.2003.09.001)

2. McRae K, Cree GS, Seidenberg MS, Mcnorgan C. 2005 Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* **37**, 547–559. (doi:10.3758/BF03192726)

3. Buchanan EM, Valentine KD, Maxwell NP. 2019 English semantic feature production norms: An extended database of 4436 concepts. *Behav. Res. Methods* **51**, 1849–1863. (doi:10.3758/s13428-019-01243-z)

4. Liu Q, Lupyan G. 2020 The Limits of Cross-Domain Mappings: Why Is Philosophy Purple? Poster presented at the 61st Annual Metting of the Psychonomic Society.

5. Gentner D. 1983 Structure-mapping: A theoretical framework for analogy. *Cogn. Sci.* **7**, 155–170.

6. Wolff P, Gentner D. 2011 Structure-Mapping in Metaphor Comprehension. *Cogn. Sci.* **35**, 1456–1488. (doi:10.1111/j.1551-6709.2011.01194.x)

7. Bowdle BF, Gentner D. 2005 The career of metaphor. *Psychol. Rev.* **112**, 193.

8. Gentner D, Clement C. 1988 Evidence for Relational Selectivity in the Interpretation of Analogy and Metaphor. In *Psychology of Learning and Motivation* (ed GH Bower), pp. 307–358. Academic Press. (doi:10.1016/S0079-7421(08)60044-4)

9. Barsalou LW, Dutriaux L, Scheepers C. 2018 Moving beyond the distinction between concrete and abstract concepts. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170144. (doi:10.1098/rstb.2017.0144)

10. Borghi AM, Binkofski F, Castelfranchi C, Cimatti F, Scorolli C, Tummolini L. 2017 The challenge of abstract concepts. *Psychol. Bull.* **143**, 263–292. (doi:10.1037/bul0000089)

11. Borghi AM, Barca L, Binkofski F, Tummolini L. 2018 Varieties of abstract concepts: development, use and representation in the brain. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170121. (doi:10.1098/rstb.2017.0121)

12. Dove G. 2016 Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychon. Bull. Rev.* **23**, 1109–1121. (doi:10.3758/s13423-015-0825-4)

13. Pecher D, Zeelenberg R. 2018 Boundaries to grounding abstract concepts. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170132. (doi:10.1098/rstb.2017.0132)

14. Fernandino L, Humphries CJ, Seidenberg MS, Gross WL, Conant LL, Binder JR. 2015 Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia* **76**, 17–26.

15. Binder JR, Westbury CF, McKiernan KA, Possing ET, Medler DA. 2005 Distinct Brain Systems for Processing Concrete and Abstract Concepts. *J. Cogn. Neurosci.* **17**, 905–917. (doi:10.1162/0898929054021102)

16. Brysbaert M, Warriner AB, Kuperman V. 2014 Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911. (doi:10.3758/s13428-013-0403-5)

17. Crutch SJ. 2005 Abstract and concrete concepts have structurally different representational frameworks. *Brain* **128**, 615–627. (doi:10.1093/brain/awh349)

18. Montefinese M. 2019 Semantic representation of abstract and concrete words: a minireview of neural evidence. *J. Neurophysiol.* **121**, 1585–1587. (doi:10.1152/jn.00065.2019)

19. Collins AM, Quillian MR. 1969 Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* **8**, 240–247.

20. Linton M, Norman DA, Rumelhart DE. 1975 Explorations in cognition.

21. Tversky A. 1977 Features of similarity. *Psychol. Rev.* **84**, 327.

22. Mcrae K, Seidenberg MS, Sa VRD. 1997 On the nature and scope of featural representations of word meaning. *J. Exp. Psychol. Gen.* , 99–130.

23. Osgood CE. 1952 The nature and measurement of meaning. *Psychol. Bull.* **49**, 197–237. (doi:10.1037/h0055737)

24. Osgood CE. 1960 The cross-cultural generality of visual-verbal synesthetic tendencies. *Behav. Sci.* **5**, 146–169.

25. Palmer SE, Schloss KB. 2010 An ecological valence theory of human color preference. *Proc. Natl. Acad. Sci.* **107**, 8877–8882. (doi:10.1073/pnas.0906172107)

26. Saysani A, Corballis MC, Corballis PM. 2021 Seeing colour through language: Colour knowledge in the blind and sighted. *Vis. Cogn.* **29**, 63–71. (doi:10.1080/13506285.2020.1866726)

27. van Paridon J, Liu Q, Lupyan G. 2021 How do blind people know that blue is cold? Distributional semantics encode color-adjective associations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp. 2671–2677.

28. Lynott D, Connell L, Brysbaert M, Brand J, Carney J. 2020 The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behav. Res. Methods* **52**, 1271–1291. (doi:10.3758/s13428-019-01316-z)

29. Andrews M, Frank S, Vigliocco G. 2014 Reconciling embodied and distributional accounts of meaning in language. *Top. Cogn. Sci.* **6**, 359–370.

30. Landauer TK, Dumais ST. 1997 A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211.

31. Lund K, Burgess C. 1996 Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* **28**, 203–208.

32. Lupyan G, Lewis M. 2017 From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Lang. Cogn. Neurosci.* **34**, 1319–1337. (doi:10.1080/23273798.2017.1404114)

33. Yee E, Jones MN, McRae K. 2018 Semantic Memory. In *The Stevens' Handbook of Experimental Psychology and Neuroscience, Fourth Edidition* (ed JW& S Thompson-Schill), pp. 319–356. UK: Wiley.

34. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013 Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc.

35. Baroni M, Dinu G, Kruszewski G. 2014 Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238–247. Baltimore, Maryland: Association for Computational Linguistics. (doi:10.3115/v1/P14-1023)

36. Kiela D, Hill F, Clark S. 2015 Specializing Word Embeddings for Similarity or Relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2044–2048. Lisbon, Portugal: Association for Computational Linguistics. (doi:10.18653/v1/D15-1242)

37. Boleda G. 2020 Distributional Semantics and Linguistic Theory. *Annu. Rev. Linguist.* **6**, 213–234. (doi:10.1146/annurev-linguistics-011619-030303)

38. Thompson B, Roberts SG, Lupyan G. 2020 Cultural influences on word meanings revealed through large-scale semantic alignment. *Nat. Hum. Behav.* **4**, 1–10. (doi:10.1038/s41562-020-0924-8)

39. De Deyne S, Navarro DJ, Perfors A, Brysbaert M, Storms G. 2019 The "Small World of Words" English word association norms for over 12,000 cue words. *Behav. Res. Methods* **51**, 987–1006. (doi:10.3758/s13428-018-1115-7)

40. De Deyne S, Perfors A, Navarro DJ. 2016 Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1861–1870. Osaka, Japan: The COLING 2016 Organizing Committee.

41. Firth JR. 1957 A synopsis of linguistic theory, 1930-1955. *Stud. Linguist. Anal.*

42. McHugh ML. 2012 Interrater reliability: the kappa statistic. *Biochem. Medica* **22**, 276–282.

43. Osgood CE. 1964 Semantic differential technique in the comparative study of cultures. *Am. Anthropol.* **66**, 171–200.

44. Tourangeau R, Sternberg RJ. 1981 Aptness in metaphor. *Cognit. Psychol.* **13**, 27–55.

45. Tourangeau R, Sternberg RJ. 1982 Understanding and appreciating metaphors. *Cognition* **11**, 203–244.

46. Trick L, Katz AN. 1986 The Domain Interaction Approach to Metaphor Processing: Relating Individual Differences and Metaphor Characteristics. *Metaphor Symb. Act.* **1**, 185–213. (doi:10.1207/s15327868ms0103_3)

47. Katz AN. 1989 On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *J. Mem. Lang.* **28**, 486–499. (doi:10.1016/0749-596X(89)90023-5)

48. Kajić I, Schröder T, Stewart TC, Thagard P. 2019 The semantic pointer theory of emotion: Integrating physiology, appraisal, and construction. *Cogn. Syst. Res.* **58**, 35–53.

49. Martin RA. 2007 CHAPTER 4 - The Cognitive Psychology of Humor. In *The Psychology of Humor* (ed RA Martin), pp. 83–111. Burlington: Academic Press. (doi:10.1016/B978-012372564-6/50023-X)

50. Hahn U, Heit E. 2015 Semantic Similarity, Cognitive Psychology of. In *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (ed JD Wright), pp. 579–584. Oxford: Elsevier. (doi:10.1016/B978-0-08-097086-8.53026-8)

51. De Deyne S, Navarro DJ, Perfors A, Storms G. 2016 Structure at every scale: A semantic network account of the similarities between unrelated concepts. *J. Exp. Psychol. Gen.* **145**, 1228–1254. (doi:10.1037/xge0000192)

52. Harris ZS. 1954 Distributional Structure. WORD **10**, 146–162. (doi:10.1080/00437956.1954.11659520)

53. Rieger BB. 1991 On Distributed Representation in Word Semantics.

54. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. 2017 Advances in Pre-Training Distributed Word Representations. *ArXiv171209405 Cs*

55. Bates D *et al.* 2020 *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. See https://CRAN.R-project.org/package=lme4.

56. Caliskan A, Bryson JJ, Narayanan A. 2017 Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186. (doi:10.1126/science.aal4230)

57. Grand G, Blank IA, Pereira F, Fedorenko E. 2018 Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. *ArXiv Prepr. ArXiv180201241*

58. Lewis M, Lupyan G. 2020 Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat. Hum. Behav.* , 1–8. (doi:10.1038/s41562-020-0918-6)

59. Kim JS, Elli GV, Bedny M. 2019 Knowledge of animal appearance among sighted and blind adults. *Proc. Natl. Acad. Sci.* **116**, 11213–11222. (doi:10.1073/pnas.1900952116)

60. Bedny M, Koster-Hale J, Elli G, Yazzolino L, Saxe R. 2019 There's more to 'sparkle' than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition* **189**, 105–115. (doi:10.1016/j.cognition.2019.03.017)

61. Lewis M, Zettersten M, Lupyan G. 2019 Distributional semantics as a source of visual knowledge. *Proc. Natl. Acad. Sci.* **116**, 19237–19238. (doi:10.1073/pnas.1910148116)

62. Lakoff G, Johnson M. 1980 Conceptual Metaphor in Everyday Language. *J. Philos.* **77**, 453–486. (doi:10.2307/2025464)