Title: **How much is a cow like a meow? A novel database of human judgements of audiovisual semantic relatedness**

Authors: Kira Wegner-Clemens[1], George L. Malcolm[2], Sarah Shomstein[1]

Author affiliation: [1] George Washington University, Psychological and Brain Sciences

[2] University of East Anglia, School of Psychology

Keywords: semantics, multisensory, audiovisual, naturalistic stimulus set

Corresponding author:

Kira Wegner-Clemens
Department of Psychological and Brain Sciences
George Washington University
E-mail: kira@gwu.edu

Abstract

Semantic information about objects, events, and scenes influences how humans perceive, interact with, and navigate the world. The semantic information about any object or event can be highly complex and frequently draws on multiple sensory modalities, which makes it difficult to quantify. Past studies have primarily relied on either a simplified binary classification of semantic relatedness based on category or on algorithmic values based on text corpora rather than human perceptual experience and judgement. With the aim to further accelerate research into multisensory semantics, we created a constrained audiovisual stimulus set and derived similarity ratings between items within three categories (animals, instruments, household items). A set of 140 participants provided similarity judgments between sounds and images. Participants either heard a sound (e.g., a meow) and judged which of two pictures of objects (e.g., a picture of a dog and a duck) it was more similar to, or saw a picture (e.g., a picture of a duck) and selected which of two sounds it was more similar to (e.g., a bark or a meow). Judgements were then used to calculate similarity values of any given cross-modal pair. The derived and reported similarity judgements reflect a range of semantic similarities across three categories and items, and highlight similarities and differences among similarity judgments between modalities. We make the derived similarity values available in a database format to the research community to be used as a measure of semantic relatedness in cognitive psychology experiments, enabling more robust studies of semantics in audiovisual environments.

Introduction

Semantic information is crucial to daily life. How we understand scenes, interact with objects, and navigate through environments is shaped by the meaning, or semantics, of these very scenes, objects, and environments. Despite the importance of semantics, its role on behavior has been less extensively studied than other features of sensory signals, such as loudness, brightness, or color. A major barrier to studying semantics has been the difficulty in quantifying how multiple objects are semantically related, especially across sensory systems. For a study investigating loudness, any two auditory stimuli can be directly compared by measuring the decibels of each, while for a study investigating semantic relatedness, any two signals could potentially be related in a number of different ways. Two signals might share a category (e.g., foods), be associated with the same event or object (e.g., a dog and its bark), or occur in the same location (e.g., kitchen items). Each of these possible relationships corresponds to a different aspect of semantic meaning that overlaps with and is available simultaneously with other aspects.

To compare stimuli in studies, researchers often select one aspect and define semantic relatedness in reference to that aspect. For example, a study might define semantic relatedness as whether two items belong to the same category. Under this definition (semantics-as-category), two items of clothing (a t-shirt, a pair of pants) would be defined as semantically related, while an item of clothing and a kitchen utensil would be defined as semantically unrelated (a t-shirt, a spoon). This category based definition has been widely used, in studies finding that same-category distractors disrupt visual search to a greater extent (Moores, Laiti, and Chelazzi 2003), same-category words are remembered better (Buchanan et al. 2006), and category guides attention between visual objects even when task-irrelevant (Malcolm, Rattinger, and Shomstein 2016). Categories themselves can be defined in various ways, with a major distinction between thematic relationships based on co-occurrence and taxonomic relationships based on feature similarity (Lin and Murphy 2001; Estes, Golonka, and Jones 2011; Wisniewski, E. J., & Bassok, M 1999)

However, category is not the only way semantics has been defined in studies of memory and attention. An alternative option is to define semantic relatedness by whether two signals have the same source. Under this definition (semantics-as-source), a visual image of a piano and an auditory sound of piano note would be considered semantically related, while a visual image of a piano and an auditory sound of a violin would not be considered semantically related. In an auditory context, two speech recordings might be considered semantically related if each was spoken by the same speaker. The source based definition has also been widely used, especially in multisensory contexts, with studies finding that sounds speed search for shared-source images (Iordanescu et al. 2008) and videos (Kvasova, Garcia-Vernet, and Soto-Faraco 2019) and improve memory for shared-source objects (Heikkilä et al. 2015), even when task irrelevant (Duarte, Ghetti, and Geng 2021; Mastroberardino, Santangelo, and Macaluso 2015), and images improve memory for shared-source sounds (Moran et al. 2013). Ostensibly, these studies and the studies described above using the semantics-as-category definition investigate the same aspect of sensory events, semantics, and depend on shared mechanisms of semantic processing. However, depending on what definition is used, the same pairing of stimuli could be considered either semantically related or not semantically related. Under a semantics-as-category definition, an image of violin and the sound of a piano would be considered related, but would not be considered related under the semantics-as-causality definition. These differences in definition have an impact on perception, with thematically related pairs being grouped together more quickly than taxonomic related pairs (Nah and Geng 2021). Each definition has provided key insights into how the corresponding aspect of semantics influences attention and memory, but taken together, leave a number of open questions about semantics.

A fundamental barrier to a more comprehensive understanding of semantic influence is that prior measures of semantic relatedness have most been relying on a binary classification (either semantically related or not semantically related), while human observers have more nuanced and continuous understandings of semantic relatedness. In the example of a shared-cause definition of semantic relatedness above, an image of a piano was defined as related to the sound of a piano note, but not related to the sound of a violin note. However, under a categorical definition of semantic relatedness, a piano and a violin would be defined as semantically related because both are musical instruments. A human observer would likely place these into a continuum of relatedness with the image of the piano more related to the sound piano note and less related to the violin note. Any differences in behavior that rely on this continuous understanding of semantic relatedness would be missed with either the categorical or causality-based definition of semantic relatedness.

Several studies have sought to tackle this issue by using machine learning algorithms to extract semantic relatedness values from massive text corpora. The algorithms produce models of semantic meaning, known as distributional semantics models, that use the context that a word appears in large language databases such as Wikipedia and news archives to define how that word relates to other words (Lenci 2018). In a distributional semantics model, any pair of words that appear in the database has a corresponding relatedness value, which provides a measure of relative strength of relatedness (a piano would be more related to violin than to a spoon). By using a continuous measure, studies based on distributional semantics models can more effectively represent the continuum of relatedness as human observers understand it and how that more complex representations of semantics influences human behavior. In one application of this definition, values from distributional semantic models have been shown to predict eye movements (Hwang, Wang, and Pomplun 2011; Hayes and Henderson 2021), suggesting that values derived from corpora do reflect human behavior.

However, despite the shown relationship between the corpora and behavior, the derived relationships extracted from how words describing that stimuli are used in writing might not be the most sensitive measure. The model is based on words representing sensory experiences, rather than human judgements about the sensory experience of the stimuli. Particularly in multisensory studies, it is possible that the judgement of semantic similarity for two items will depend on what sensory modality each item is being experienced through. Mixed results in direct comparisons of corpora-based semantic relatedness value and human ratings provide further evidence for the possibility sensory experience shapes semantic similarity. Algorithm judgments and human judgments are correlated (Richie, Zou, and Bhatia 2019), but distributional semantic models systematically fail to capture certain elements of how human raters understand semantics (Nematzadeh, Meylan, and Griffiths 2017; Bhatia, Richie, and Zou 2019). For example, human raters produce systematic asymmetric judgements, so object A will be judged as similar to object B, but object B will not be judged as similar to object A (Nematzadeh, Meylan, and Griffiths 2017). Distributional semantics models are incapable of providing different relatedness depending on the directionality; the relatedness values are always symmetrical. Additionally, distributional semantic models are also largely constrained to similarity relationships in nouns and struggle with position in a hierarchy (hypernyms), opposites (antonyms), and verbs. The models also cannot account for any differences between stimuli of different sensory modalities. Some models have incorporated visual information (Bruni, Tran, and Baroni 2014; Lazaridou, Nghia The Pham, and Baroni 2015) or auditory information (Lopopolo and van Miltenburg 2015), but even sensory-grounded models are limited to a single sensory modality rather than the multisensory world humans experience.

To better understand the role of semantics in multisensory contexts, we identified the need for constructing a database of visual pictures and sounds along with a set of corresponding semantic relatedness values that are recorded from human observers. Audiovisual stimulus sets already exist, such as the Multimodal Stimulus Set (Schneider, Engel, and Debener 2008), but do not include corresponding semantic relatedness values. Similarly, semantic ratings databases exist, but they rely exclusively on image pairs (as in Jiang, Sanders, and Cowell 2022) or word pairs (as in Landrigan and Mirman 2016). Here, we developed such a database for a naturalistic audiovisual stimulus set, providing a measure of semantic relatedness derived from human judgements for every possible item pairing within each of three categories. The values reflect the continuum of semantic relatedness human observers understand by providing a quantified value for each pairing, rather than a binary decision of related or not related. We share this database of pictures and images, along with corresponding semantic relatedness values, statistics, and larger versions of the figures in an Open Science Framework (available at osf.io/v9rgy/).

Methods

Participants: In Experiment 1 (audiovisual judgments), we analyzed judgments from 140 participants. An additional 19 were excluded due to low accuracy (<70% on catch matched trials). Forty-three were recruited from Amazon's Mechanical Turk service and 97 were recruited from the George Washington University participant pool. In Experiment 2 (word judgments), we analyzed judgments from a separate group of 140 participants. An additional 37 were excluded due to low accuracy (<70% on matched trials). Eleven were recruited from Amazon's Mechanical Turk service, and 129 were recruited from the George Washington University participant pool. The Amazon Mechanical Turk participants were US-based adults, expected to have similar demographics to previous studies of US mTurk workers (55% female; 50% under 33) (Difallah, Filatova, and Ipeirotis 2018). George Washington University participant pool is a typical sample of American undergraduate students, with similar demographics to the overall George Washington undergraduate population (62% female; 50% under 20). All participants were compensated financially or with course credit. All participants gave informed consent and the study was approved by the Institutional Review Board of George Washington University.

Power analysis: A traditional power analysis to determine sample size is not possible because the goal is to characterize the perceived relationship between stimuli, rather than test a hypothesis. In order to determine sample size, we calculated how many raters would be necessary in order to obtain the 43200 total ratings (20 ratings for each of 2160 stimuli trios) without fatiguing raters with an overly long experimental time.

Selection of stimuli: A total of 30 images and 30 corresponding sounds were selected for the stimulus set, split evenly between three stimulus categories (animals, instruments, and household items) with 10 images and 10 corresponding sounds in each category. The categories were selected to be fairly broad and allow for a wide range of semantic relatedness. The items were selected to be recognizable both as an image and a sound. Since audiovisual matching performance has been shown to depend on exemplars (Edmiston and Lupyan 2015), exemplars for each item were selected to correspond between the sound and image. Since a recording from an acoustic guitar was selected as the guitar sound, a picture of an acoustic guitar was selected as the guitar image. However, all images were shown in a "static" position to avoid showing hands for items operated by people (e.g., there was not a hand shown strumming the guitar). Items and exemplars were selected to be as familiar to as broad an audience as possible. For example, we avoided items like a seagull, that may be much more familiar to a participant that grew up on a coast, or an ambulance, where the sound of a siren differs from city to city.

Images were selected from the THINGS Database, a set of naturalistic images (Hebart et al. 2019). Among the exemplars for each item, images were selected to be clearly visible, recognizable, and did not have other objects in view or people interacting with the object. Sounds were collected from online databases of freely available sounds and were trimmed to 1 second and normalized for loudness in Audacity (Audacity Team, 2021). To ensure the sounds were readily recognizable, pilot testing was conducted. Sixteen participants listened to all exemplars of the sound items on the initial list, provided a description of it, and only sounds where the pilot participants provided the same description (e.g., "cat", "doorbell") were selected for the main experiment.

Task design: In Experiment 1 (audiovisual), participants completed a two-alternative forced choice task determining how similar visual images and auditory sounds were to one another (Fig. 1a). A forced choice task was selected over a direct rating task because of concerns participants would not use the entire rating scale and simply classify pairs as related or unrelated, as we had observed in pilots of other experiments in the lab. Before the trials started, participants completed a familiarization phase in which each image was presented with a simultaneously presented corresponding sound. The familiarization phase ensured that participants recognized each sound and each image. Participants were instructed to always select the matched pairs shown in the familiarization stage when they appeared as a stimulus and option (e.g., a dog and a bark). Matched trials served as catch trials, ensuring that participants were paying attention and making actual judgments about the stimuli they were hearing and seeing. Catch trials were included given that it was not possible to calculate a “correct” answer and evaluate accuracy for the unmatched semantic judgment trials. Participants with low accuracy (<70%) on match trials were excluded from further analyses.

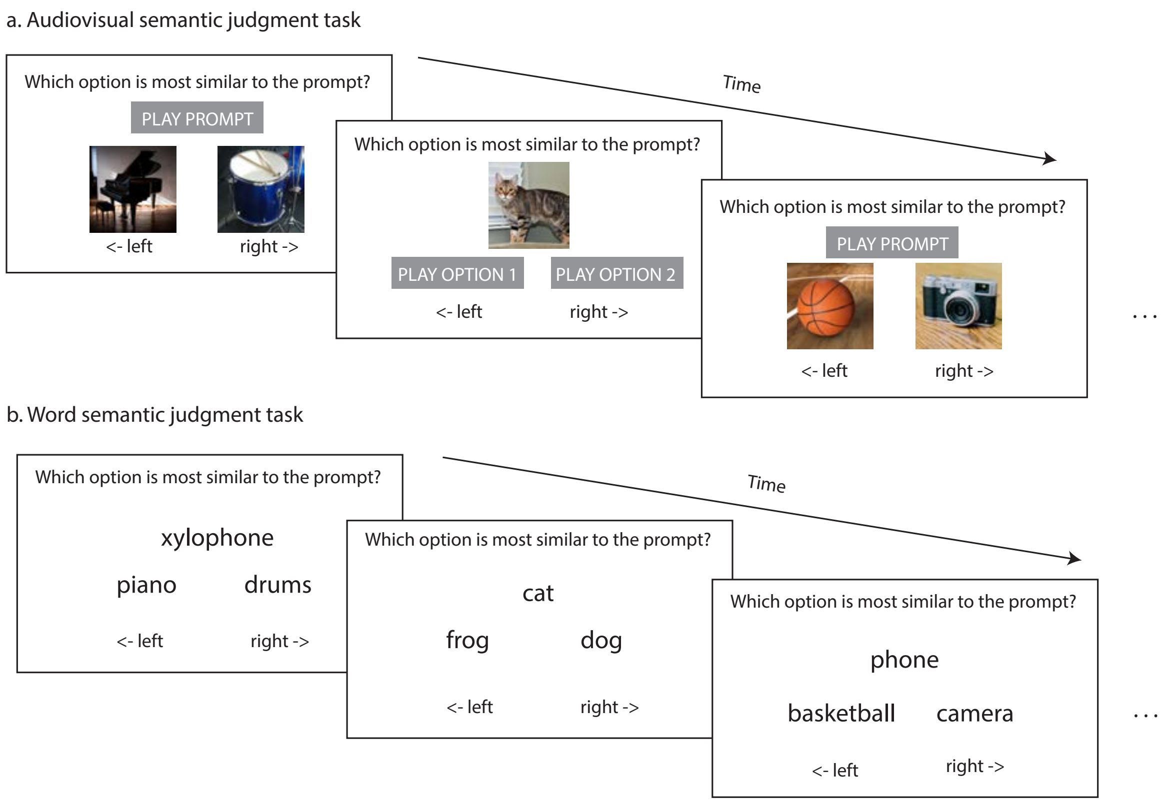

Fig 1. (a). Sample trials for audiovisual judgement task. Participants pressed gray play buttons to play auditory stimuli. After playing each sound option, responses were made by choosing either the left or right arrow associated with the corresponding sound. (b). Sample trial for word judgement task. Either left or right arrow associated with the corresponding word was chosen as a response.

On a “visual” trial, a prompt image was shown (e.g., an image of a cat) along with two placeholders for sounds. Participants clicked on each of the two sounds, and after listening to both, selected which of two sounds was most similar to the prompt image. On an “auditory” trial, a prompt sound was played and the participants selected which of two images was most similar to the prompt sound. Within a trial, the prompt and both options were selected from the same category (animals, instruments, household items). Categories were not presented in separate blocks of trials, but rather trials from different categories were presented randomly within the session. The trials were self-paced. Participants clicked a button to start the sound and could listen to the sounds multiple times if they chose to, but could not progress if they did not listen to each sound at least once. The next trial started once participants selected one of the options via a key press. In Experiment 2, a similar two-alternative forced choice task was used, with the difference that the images and sounds were replaced with written words (Fig. 2b). On each trial, a prompt word was presented and the participants selected which of two option words were most similar to the prompt word.

Randomization and counterbalancing: Due to the large number of comparisons, it was not possible for a single participant to provide a judgement for every possible trio combination of prompt and two options. There were 1080 trio combinations and every trio combination was judged 20 times with a visual prompt and 20 times for an auditory prompt, for a total of 43200 judgements on the audiovisual task. In the word task, each trio of words was judged 20 times for a total of 21600 trials. There are half as many trials in the word task because each trio was only presented in one modality (word) rather than two (auditory, visual). Every participant provided judgments for approximately $1/7^{th}$ of the trio combinations and saw every pair of prompt and option at least once. Including match trials, participants in the audiovisual task completed either 317 or 318 trials in audiovisual and participants in the word task completed 158 or 159 trials.

Data analysis: The likelihood of picking an option for a given prompt was calculated for each pairing for each participant. The likelihood is the percent of trials that option was picked given a specific prompt, independent of what the second option on that particular trial. To understand the variation between trials where the prompt was visual and trials where the prompt was auditory, we conducted a series of independent t-tests where individual participant likelihood values for visual prompt trials and auditory prompt trials was compared (bottom panels in Fig. 3-5). Semantic relatedness values that were averaged over modality, but not over prompt direction, were calculated in order to get to compute semantic relatedness for each possible prompt and option combination (Fig. 3a, 4a, 5a).

To understand whether a specific modality pairing (auditory prompt/visual option or visual prompt/auditory option) yielded more closely related relationships, we subtracted the raw values between the trials (visual – auditory) to identify the pairs where relatedness differed by modality as well as the directionality of that difference, (Fig. 3b, 4b, 5b). Positive values indicate that the pair was judged more similar when the prompt (on the y-axis) was visual and the option (on the x-axis) was auditory. To understand any variation based on whether the stimulus was a prompt or an option, we again conducted a series of independent t-tests where individual participant likelihood values for each prompt direction were compared. The values for each prompt direction were then subtracted to create the difference by prompt direction (Fig. 6b, 7b, 8b). The initial values for each option and pair were ultimately averaged over participant, modality, and prompt direction to get the final semantic relatedness values (Fig. 6a, 7a, 8a). A similar analysis pipeline was used to derive likelihood values for the word task (Fig. 9b, 9e, 9h), with the exception that there were no differences by modality since all words were presented in the same modality, as text.

Text corpora values: The text corpora values were extracted using the Gensim library for Python and a pre-trained model, “fasttext-wiki-news-subwords-300” (details of model available in (Mikolov et al. 2017). This model was trained on a total of 650 billion words including

Wikipedia from June 2017, two news corpuses (statmt.org news, UMBC news), and corpuses derived from a wide range of websites (Gigagword, Common Crawl). The words were identical to those used in the word task, with the exception of “cuckoo clock” which was substituted for clock because cuckoo clock as not available.

Results and Discussion

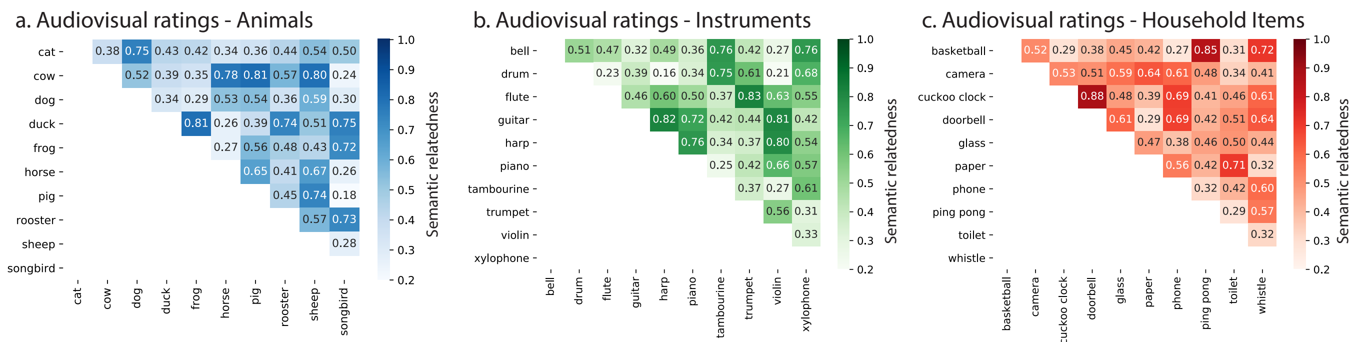

Fig 2 Measure of semantic relatedness based on human ratings of similarity between images and sounds for (a) animals, (b) instruments, and (c) audiovisual items. Values are derived from the likelihood a participant would judge that pair as more closely related. Higher values and darker colors indicate more relatedness (e.g., an exact match like a cat and a meow would have a value of 1).

We observed a wide range in semantic relatedness for both the audiovisual task (Experiment 1) and word task (Experiment 2), which reflects that some item pairs were judged to be more closely related to one another than other item pairs. Since this database is intended to be used for studies of differences in semantic relatedness, it is essential to have pairs with a low level of relatedness and pairs with a high level of relatedness. The wide range in semantic relatedness values also suggests that participants were making judgements based on a shared understanding of semantic relatedness. If each individual’s semantic judgements were highly idiosyncratic or participants were answering randomly, each pairing would have a value around 0.5 because neither option would be more likely to be selected than any other option. Instead, in the audiovisual task, semantic relatedness values ranged from 0.18 to 0.81 for animals (Fig. 3a), 0.16 to 0.83 for instruments (Fig. 4a), and 0.29 to 0.88 for household items (Fig. 5a). In the word task, semantic relatedness ranged from 0.18 to 0.94 for animals (Fig. 9b), 0.23 to 0.82 for instruments (Fig. 9e), and 0.21 to 0.89 for household items (Fig. 9h). The range of the values indicates that some items were considered more closely related to one another than other items and that there was at least some amount of consensus between participants about which those were. In an analysis of how many participants made the same choice for each stimulus trio, we found there was a high level of consensus for some trios and a lower level for others, as would be expected for stimuli that vary considerably in semantic relatedness. On average, 70% of participants made the same choice for a given trio, ranging between 97% agreement on some trio and 50% agreement on other trios (participants were as likely to pick one trio as another). Examining the most strongly and most weakly related items can also provide some insight into what factors participants used to make semantic judgements. Items likely to occur in the same location (e.g., cows and pigs both often are on farms; audiovisual relatedness = 0.81) seem to be more strongly related than items likely to occur in different locations (e.g., pigs are on farms while songbirds are in forests, audiovisual relatedness=0.18). Similarly, items with shared materials or components (guitars and harps both have strings; audiovisual relatedness=0.82) seem to be more strongly related than items without similar materials (basketballs and phones, audiovisual relatedness=0.27). However, since these observations are post-hoc interpretations, future studies would be necessary to determine the relative contribution of different components of semantics to the overall semantic understanding.

Differences due to modality and prompt direction: In Experiment 1, pairs were presented with the prompt as either a visual image or an auditory sound. We calculated differences between averages when item A was shown as a prompt compared to when item B was shown as a prompt

(Fig 3b, 4b, 5b). Our results show that while for most pairs the relatedness values did not differ as a function of prompt modality, for other pairs, the relatedness values were significantly different for visual prompt and auditory prompt. The modality differences provide a cautionary observation pointing to an important asymmetry that exists for some types of relatedness that is dependent on the modality of the primary source. For example, when hearing a guitar, participants might be more likely to think of other string instruments that create a similar sound, but when seeing a guitar, participants might think of other instruments made of wood. This interpretation, of course, is of a post hoc type but is an example of one possible explanation for the modality asymmetry.

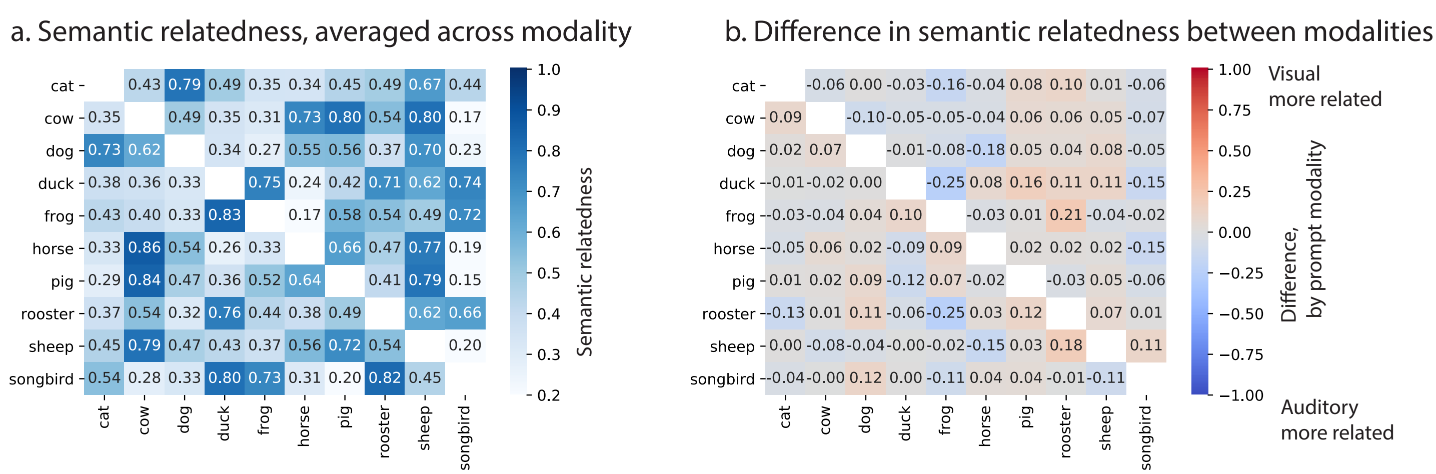

Fig 3. (a). Semantic relatedness value for animal items averaged across visual prompt and auditory prompt trials. Values are derived from the likelihood a participant would judge that pair as more closely related. Higher values and darker colors indicate more relatedness, such that an exact match would have a value of 1 if shown. Prompts are shown in the column and options are shown in the rows. (b). Difference in semantic relatedness (auditory prompt subtracted from visual prompt). Positive numbers and red shading indicate the pair was judged more related when the image was the prompt. Negative numbers and blue shading indicate that the pair was judged more related when the sound was the prompt. Prompts are shown in the column and options are shown in the rows

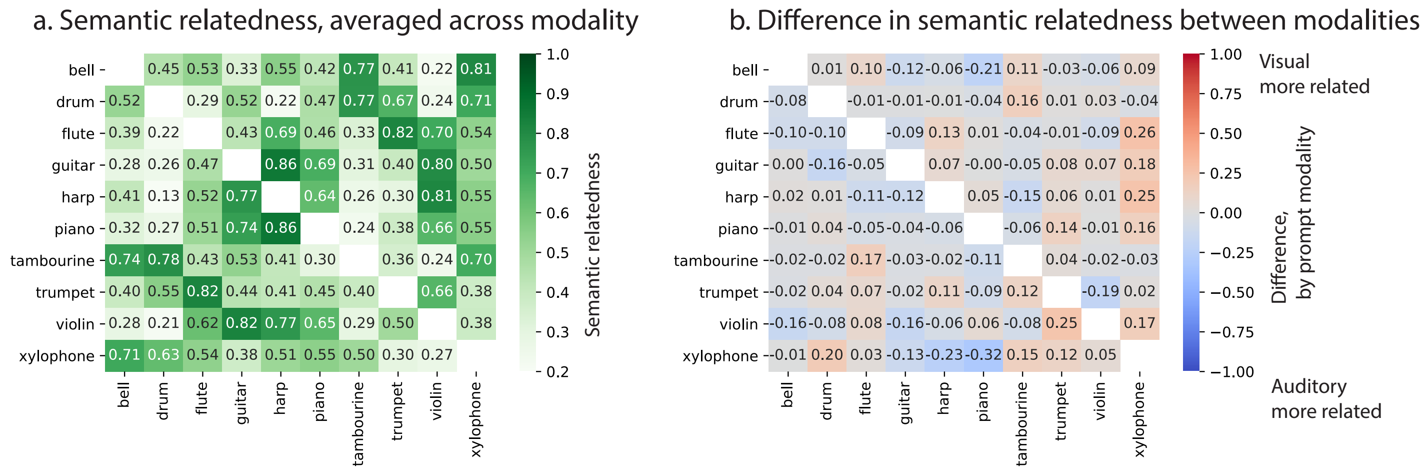

Fig 4. (a). Semantic relatedness value for instrument items averaged across visual prompt and auditory prompt trials. Values are derived from the likelihood a participant would judge that pair as more closely related. Higher values and darker colors indicate more relatedness, such that an exact match would have a value of 1 if shown. Prompts are shown in the column and options are shown in the rows. (b). Difference in semantic relatedness (auditory prompt subtracted from visual prompt). Positive numbers and red shading indicate the pair was judged more related when the image was the prompt. Negative numbers and blue shading indicate that the pair was judged more related when the sound was the prompt. Prompts are shown in the column and options are shown in the rows

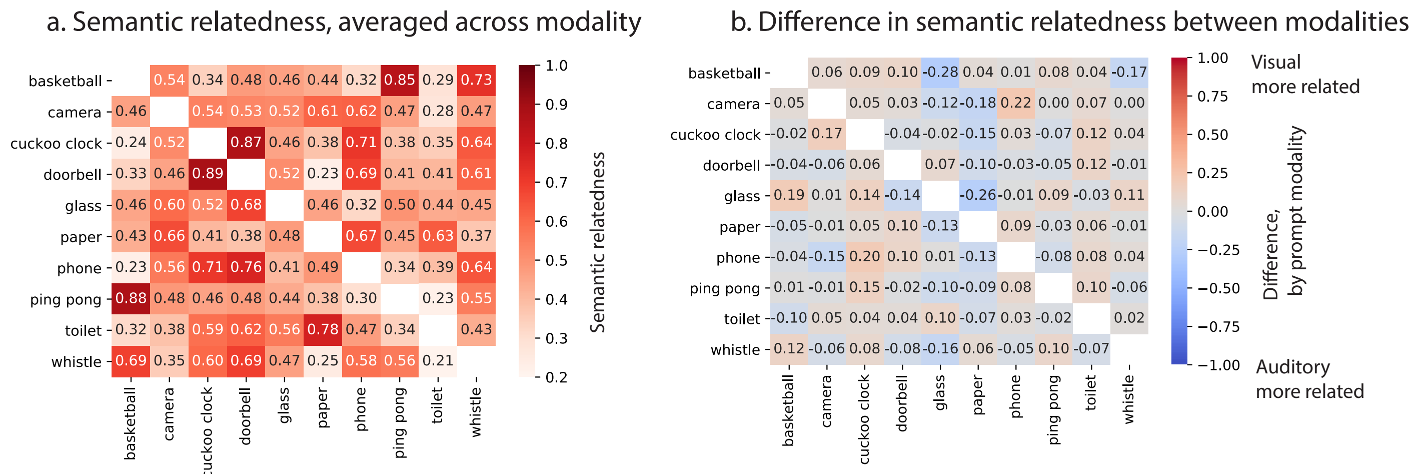

Fig 5 (a). Semantic relatedness value for household items averaged across visual prompt and auditory prompt trials. Values are derived from the likelihood a participant would judge that pair as more closely related. Higher values and darker colors indicate more relatedness, such that an exact match would have a value of 1 if shown. Prompts are shown in the column and options are shown in the rows. (b). Difference in semantic relatedness (auditory prompt subtracted from visual prompt). Positive numbers and red shading indicate the pair was judged more related when the image was the prompt. Negative numbers and blue shading indicate that the pair was judged more related when the sound was the prompt. Prompts are shown in the column and options are shown in the rows

Independent of modality, pairs could be presented with either item as the prompt (cat as a prompt with dog as an option vs. dog as a prompt with cat as an option). We calculated differences between averages when item A was shown as a prompt compared to when item B was shown as a prompt (Fig 6b, 7b, 8b). We again found that for certain pairs, there is a difference that depends on which item is the prompt and which is the option. For example, a flute and a harp are more related when a flute is the prompt (0.69) than when a harp is the prompt (0.52; Fig. 7b). These asymmetries depending on prompt directions could reflect differences in what features of the item is prioritized. For example, one possible interpretation is that when flute is the prompt, participants are more likely to focus on the feature "makes a high-pitched sound" which would make it more similar to a harp, while when harp is the prompt, participants are more likely to focus on the feature "has strings" which would make it less related to the flute.

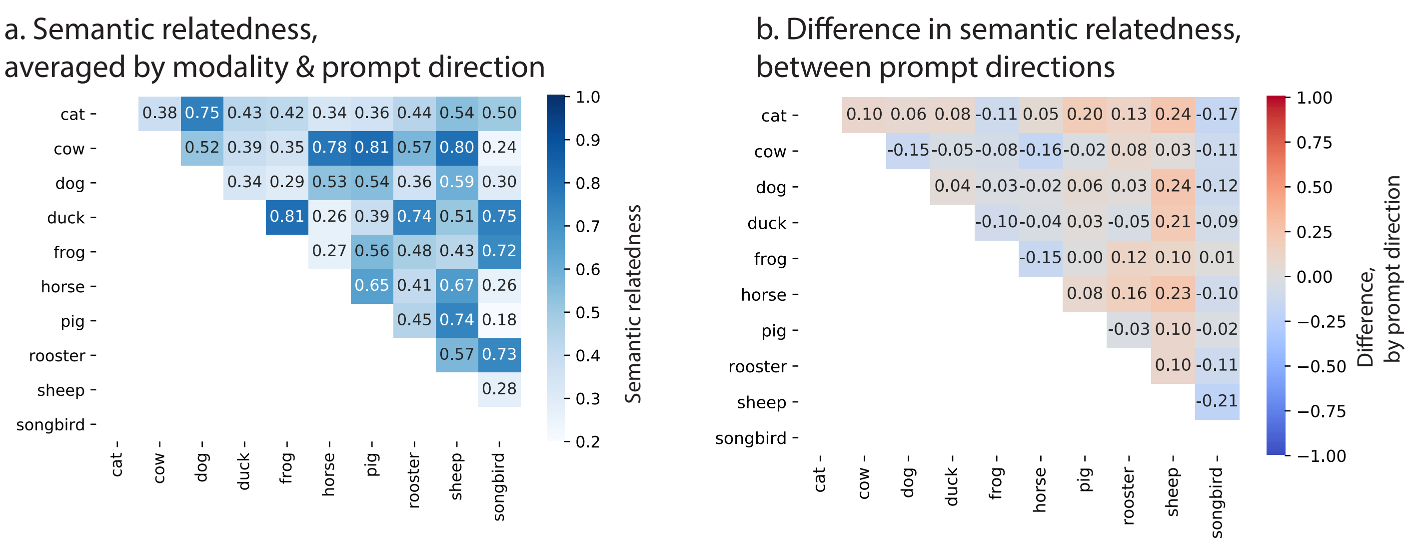

Figure 6. (a). Semantic relatedness values for animal items averaged across visual prompt and auditory prompt trials. Values are derived from the likelihood a participant would judge that pair as more closely related. Higher values and darker colors indicate more relatedness. (b). Difference in semantic relatedness by prompt direction. Positive numbers and red shading indicate the pair was judged more related when the item in the column was the prompt. Negative numbers and blue shading indicate that the pair was judged more related when the item in the row was the prompt.

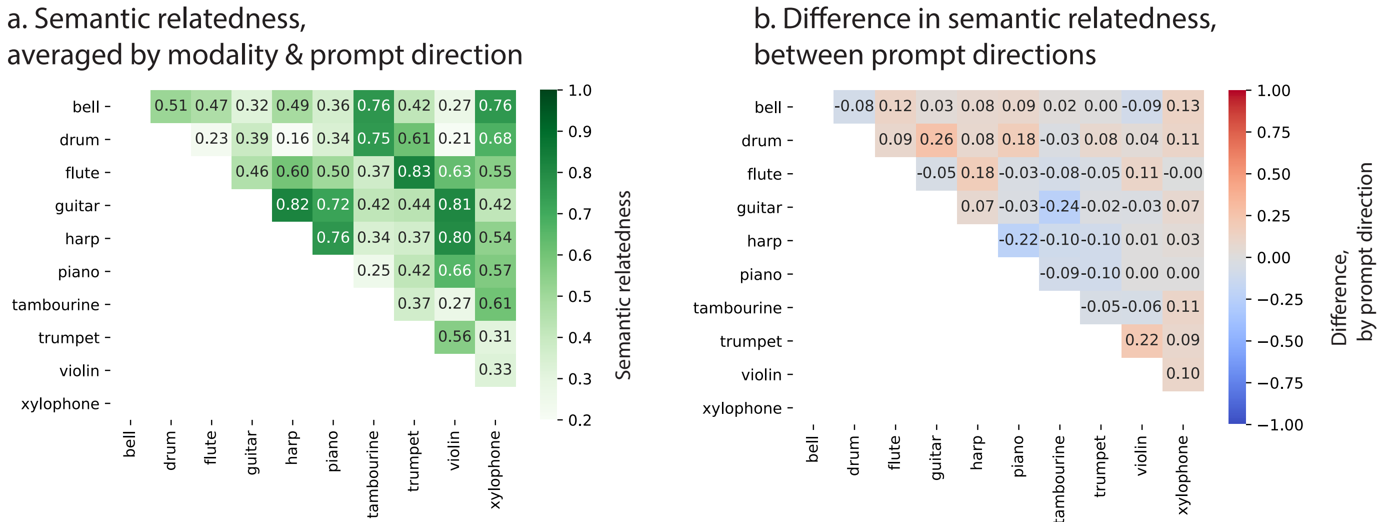

Figure 7. (a). Semantic relatedness values for instrument items averaged across visual prompt and auditory prompt trials. Values are derived from the likelihood a participant would judge that pair as more closely related. Higher values and darker colors indicate more relatedness. (b). Difference in semantic relatedness by prompt direction. Positive numbers and red shading indicate the pair was judged more related when the item in the column was the prompt. Negative numbers and blue shading indicate that the pair was judged more related when the item in the row was the prompt.

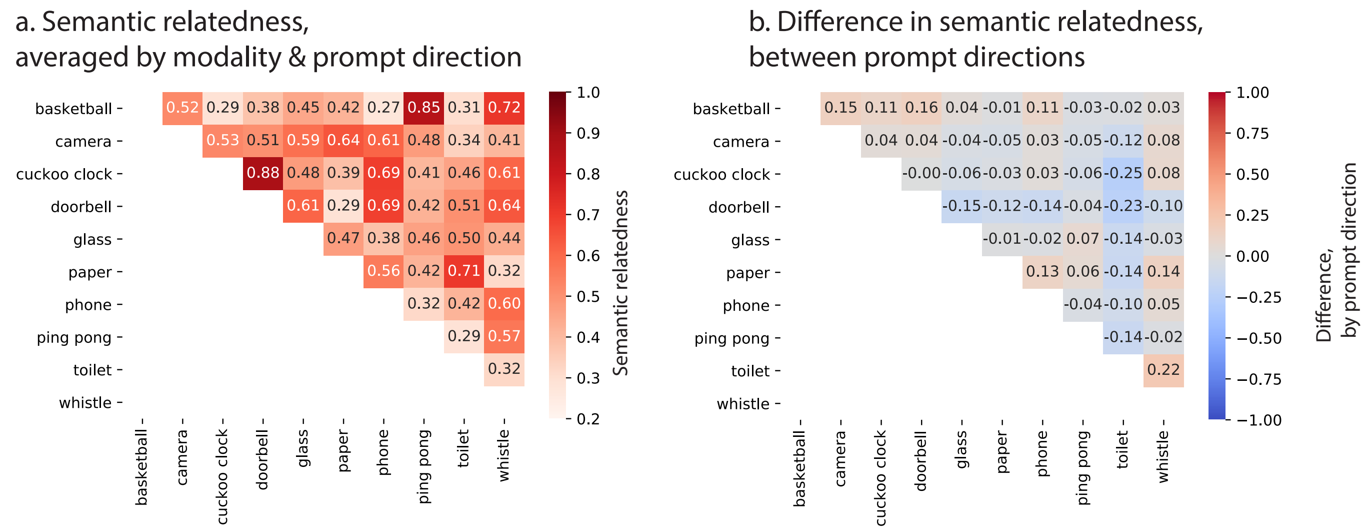

Figure 8. (a). Semantic relatedness values for household items averaged across visual prompt and auditory prompt trials. Values are derived from the likelihood a participant would judge that pair as more closely related. Higher values and darker colors indicate more relatedness. (b). Difference in semantic relatedness by prompt direction. Positive numbers and red shading indicate the pair was judged more related when the item in the column was the prompt. Negative numbers and blue shading indicate that the pair was judged more related when the item in the row was the prompt.

Regardless of the underlying reason for asymmetries in semantic judgement by prompt modality and direction, which cannot be conclusively interpreted without further studies, the differences by prompt modality and prompt direction suggests that researchers will need to carefully consider experimental design and determine whether their question of interest involves an explicit prompt and option where prompt modality and direction needs to be considered. If there is not a clear prompt directionality, the averaged value should be an effective estimate of semantic relatedness for items.

Comparison between audiovisual, word, and text corpora:

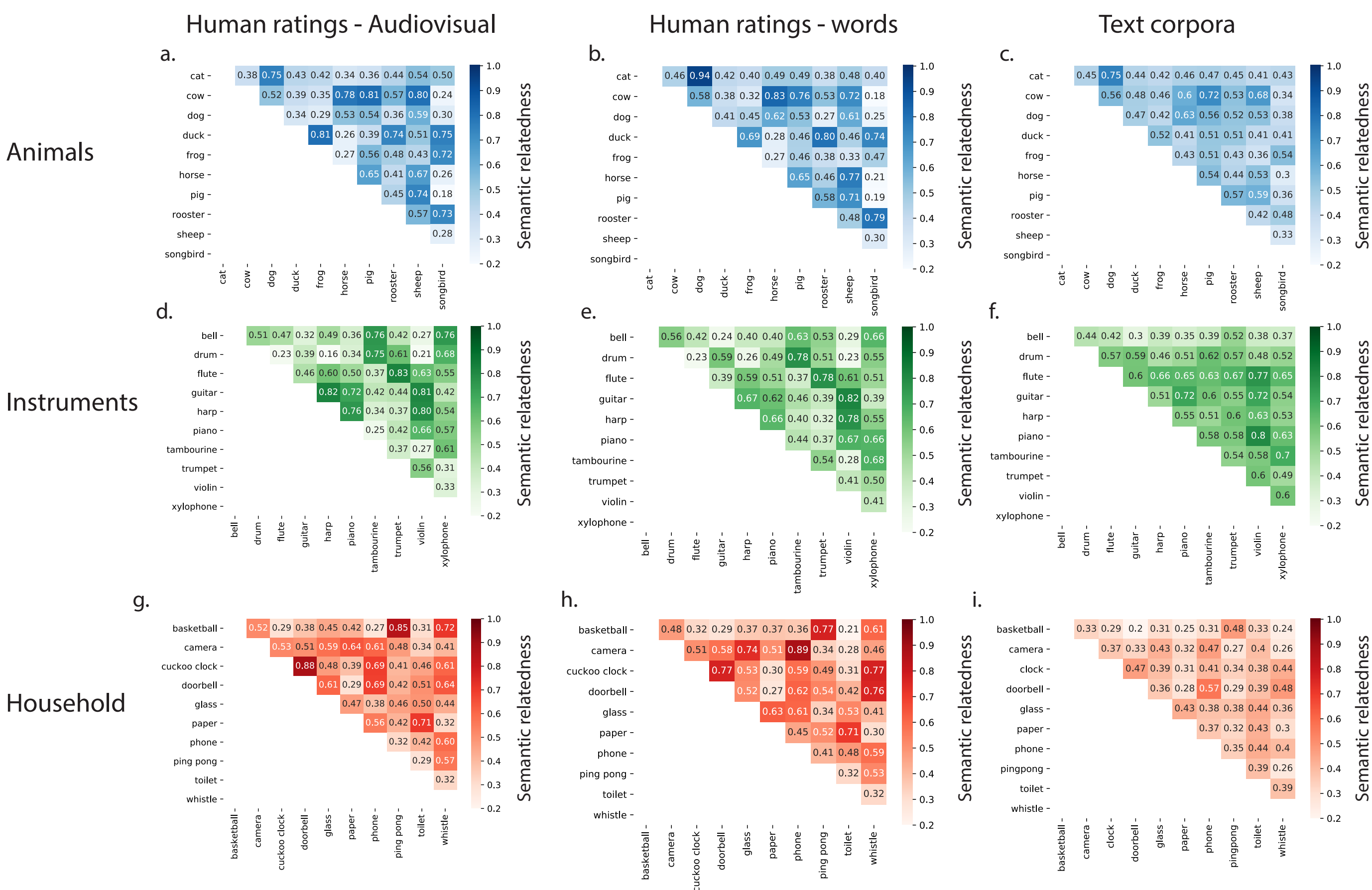

Figure 9. Semantic relatedness values averaged over prompt modality and direction for animal items on audiovisual task (a), animal items on words task (b), animal items in text corpora analysis (c); instrument items on audiovisual task (d), instrument items on words task I, instrument items on text corpora analysis (f); household items on audiovisual task (g), household items on words task (h), household items in text corpora analysis (i). Darker colors indicate a greater degree of relatedness.

The overall patterns for audiovisual, word, and text corpora were similar. Items that were related during the audiovisual task were also generally related for the word task and text corpora

(Fig. 9). The overall similarity between our tasks and the broader word corpus confirms that the similarity ratings derived from our tasks are broadly consistent with previous studies that have used text corpora. However, there was a much higher degree of variability in similarity ratings in the audiovisual and word tasks than in the text corpora. For the animals category, the values on the audiovisual task ranged 0.18-0.81, word task ranged 0.18-0.94, and the text corpora ranged 0.3-0.75. For the instruments category, the values on the audiovisual task ranged 0.16-0.83, word task ranged 0.23-0.82, and text corpora 0.3-0.8. For the household items category, the values on the audiovisual task ranged 0.27-0.88, word task ranged 0.21-0.89, and text corpora 0.2-0.57. The smaller amount of variance for the text corpora is notable because it differs from both of the human judgements tasks, suggesting that the text corpora may not effectively capture real human understanding of semantic relationships. Alternatively, the low variance in text corpora might be a result of the much larger semantic model that the pairings are embedded in. A pair might be the most similar to items in the stimulus set, but each item is likely more closely related to other items in the larger text corpora but not in the stimulus set, reducing the semantic relatedness value relative to the more constrain stimulus set. Since the purpose of this database is to characterize differences in responses to the stimulus set that depend on semantic relatedness, the higher amount of variance in the audiovisual and word tasks allows for a better characterization of the range *within* the actual stimulus set participants are viewing. Ultimately, the measure of semantic relatedness derived from the audiovisual task provides the most useful measure of semantic relatedness for studies based on this stimulus set.

Semantic information is important to understanding human behavior in real world environments, but studies of the influence of semantic information on behavior have been stymied by the difficulty of quantifying semantic relatedness. Past studies have used a binary classification, defining semantics as category (Moores, Laiti, and Chelazzi 2003; Buchanan et al. 2006; Malcolm, Rattinger, and Shomstein 2016) or semantics as source (Iordanescu et al. 2008; Kvasova, Garcia-Vernet, and Soto-Faraco 2019; Heikkilä et al. 2015; Duarte, Ghetti, and Geng 2021; Moran et al. 2013), or use algorithms to derive values based on text corpora rather than human judgments (Hayes and Henderson 2021). Human raters make more nuanced continuous judgments about semantic relatedness that have been shown to vary in key ways from both the categorical definitions and the continuous values produced by algorithms. Assuming that human behavior is based on the more subtle judgments human raters produce, the current methods present an issue for fine-grained questions of semantic relatedness and for multisensory studies in particular. A definition of semantic relatedness derived without actual judging sensory information may lose key information related to how that item is processed by a specific sensory system. Similarly, classifications of semantic related or not semantically related lose fine-grained information about human perception by simplifying the semantic relationship. The algorithmic methods fail to fully capture human judgments, as previously shown in the literature (Bhatia, Richie, and Zou 2019) and replicated here in our analyses comparing algorithm derived values to the values derived from the participant judgment data we collected (Fig. 9). Our semantic relatedness database, made available for research used, avoids these problems by providing semantic relatedness values based on human judgements for every possible pair in an audiovisual stimulus set. While it would be ideal to further validate these results by replicating an existing study showing a continuous relationship based on audiovisual semantics, it is not possible since the question of the role of continuous audiovisual understandings of semantics still needs to be explored in future studies.

Potential applications. This database is intended to be broadly useful for researchers in a number of fields interested in semantic information processing in audiovisual contexts. Psychologists can use the provided database to investigate more fine-grained differences in semantic relatedness across sensory modalities. Previously observed effects of semantics on attention can be studied in further detail to understand if they rely on category or causality specifically or a more generalized judgement of similarity that may be informed by multiple factors. It additionally could serve as a better baseline for researchers developing distributed semantics models and algorithms, particularly for those tied to perceptual experience. Comparing performance to real human judgments will better test how well they represent actual human experience of semantics.

Generalizability and future directions. While the database of related sounds and images provided here offers the needed quantification of semantic relationships between sounds and images, quantifications are derived on a finite set of images and sounds. The database that we provide here is based on relatively small number of stimuli. This stimulus set is large enough to allow for conclusions about the *relative* influence of semantic relatedness. Semantic information is highly dependent on context, with studies showing out-of-context items are less well remembered (Almadori et al. 2021; Santangelo et al. 2015).   Due to contextual influences, two objects within a category may seem closely related when compared to objects from another category, but more distantly related when compared within a category), meaning it is impossible to provide an *absolute* relationship of similarity between two given stimuli.

Similarly, different exemplars may differ slightly in semantic relatedness, with perhaps a small dog being seen as more similar to a cat than a large dog. It is important to carefully consider the relevant experimental paradigm when using this database. Certain questions and experimental designs may require a larger stimulus set with more categories or more exemplars for each item, but for many questions about the role of semantics in attention, memory, and perception, the relative relatedness between two pairs of objects will be sufficient. For example, it is possible to make conclusions about the role of semantics if a more semantically related distractor has a different behavior effect on the target than a less semantically related distractor, even if the exact semantic relatedness values are not meaningful beyond the stimulus set. In the future, the methods described here could be used to expand the database further by measuring semantic relatedness within modality (visual-visual and auditory-auditory) and between items in different categories. Certain household items may be semantically related to certain animals or instruments based on the purpose of the object or the scenes that object is likely to occur in. Cross-category values would allow researchers to tease out the role of semantics in general from the contribution of category or shared location.

The database could additionally be expanded in the future by examining differences in semantic relatedness judgements by demographic group. We sought to select items that would be familiar to many people, but the degree of familiarity or particular associations may differ if used in an older population or from outside of the United States. This generalizability is a problem universal to studies of semantics: since semantic understanding is shaped by culture, it is impossible to create a universal stimulus set and semantic relatedness values fully generalizable across all participant populations. Additionally, all of our participants were US based because we specifically sample from US-based mTurk workers and an US university, who could all share semantic understandings that the participants in other countries do not. However, since prior studies have relied on researchers' intuition about category or text corpora that have no explicit semantic judgements, even a database that is not fully generalizable like this can provide a more robust semantic measure than existing methods. In the future, the same methodology could easily be used to collect semantic judgements specific to a given demographic group or in cross-cultural comparison studies.

Ultimately, we hope that this database will allow for more robust studies and a better understanding of the role of semantics in human behavior.

References

Audacity Team (2021). Audacity(R): Free Audio Editor and Recorder. Version 3.0.0 retrieved from https://audacityteam.org/

Almadori, Erika, Serena Mastroberardino, Fabiano Botta, Riccardo Brunetti, Juan Lupiáñez, Charles Spence, and Valerio Santangelo. 2021. "Crossmodal Semantic Congruence Interacts with Object Contextual Consistency in Complex Visual Scenes to Enhance Short-Term Memory Performance." *Brain Sciences* 11 (9). https://doi.org/10.3390/brainsci11091206.

Bhatia, Sudeep, Russell Richie, and Wanling Zou. 2019. "Distributed Semantic Representations for Modeling Human Judgment." *Current Opinion in Behavioral Sciences* 29 (October): 31–36.

Bruni, E., N. K. Tran, and M. Baroni. 2014. "Multimodal Distributional Semantics." *The Journal of Artificial Intelligence Research* 49 (January): 1–47.

Buchanan, Tony W., Joset A. Etzel, Ralph Adolphs, and Daniel Tranel. 2006. "The Influence of Autonomic Arousal and Semantic Relatedness on Memory for Emotional Words." *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology* 61 (1): 26–33.

Difallah, Djellel, Elena Filatova, and Panos Ipeirotis. 2018. "Demographics and Dynamics of Mechanical Turk Workers." In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 135–43. WSDM '18. New York, NY, USA: Association for Computing Machinery.

Duarte, Shea, Simona Ghetti, and Joy Geng. 2021. "Object Memory Is Multisensory: Task-Irrelevant Sounds Improve Recollection-Based Recognition Memory." https://doi.org/10.31234/osf.io/pk4cf.

Edmiston, Pierce, and Gary Lupyan. 2015. "What Makes Words Special? Words as Unmotivated Cues." *Cognition* 143 (October): 93–100.

Estes, Zachary, Sabrina Golonka, and Lara L. Jones. 2011. "Thematic Thinking: The Apprehension and Consequences of Thematic Relations." In *Psychology of Learning and Motivation*, edited by Brian H. Ross, 54:249–94. Academic Press.

Hayes, Taylor R., and John M. Henderson. 2021. "Looking for Semantic Similarity: What a Vector-Space Model of Semantics Can Tell Us About Attention in Real-World Scenes." *Psychological Science* 32 (8): 1262–70.

Hebart, Martin N., Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 2019. "THINGS: A Database of 1,854 Object Concepts and More than 26,000 Naturalistic Object Images." *PloS One* 14 (10): e0223792.

Heikkilä, Jenni, Kimmo Alho, Heidi Hyvönen, and Kaisa Tiippana. 2015. "Audiovisual Semantic Congruency during Encoding Enhances Memory Performance." *Experimental Psychology* 62 (2): 123–30.

Hwang, Alex D., Hsueh-Cheng Wang, and Marc Pomplun. 2011. "Semantic Guidance of Eye Movements in Real-World Scenes." *Vision Research* 51 (10): 1192–1205.

Iordanescu, Lucica, Emmanuel Guzman-Martinez, Marcia Grabowecky, and Satoru Suzuki. 2008. "Characteristic Sounds Facilitate Visual Search." *Psychonomic Bulletin & Review* 15 (3): 548–54.

Jiang, Zhuohan, D. Merika W. Sanders, and Rosemary A. Cowell. 2022. "Visual and Semantic Similarity Norms for a Photographic Image Stimulus Set Containing Recognizable

Objects, Animals and Scenes." *Behavior Research Methods*, January. https://doi.org/10.3758/s13428-021-01732-0.

Kvasova, Daria, Laia Garcia-Vernet, and Salvador Soto-Faraco. 2019. "Characteristic Sounds Facilitate Object Search in Real-Life Scenes." *Frontiers in Psychology* 10 (November): 2511.

Landrigan, Jon-Frederick, and Daniel Mirman. 2016. "Taxonomic and Thematic Relatedness Ratings for 659 Word Pairs." *Journal of Open Psychology Data* 4 (1): e2.

Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. 2015. "Combining Language and Vision with a Multimodal Skip-Gram Model." *ArXiv [Cs.CL]*. arXiv. http://arxiv.org/abs/1501.02598.

Lenci, Alessandro. 2018. "Distributional Models of Word Meaning." *Annual Review of Linguistics* 4 (1): 151–71.

Lin, E. L., and G. L. Murphy. 2001. "Thematic Relations in Adults' Concepts." *Journal of Experimental Psychology. General* 130 (1): 3–28.

Lopopolo, Alessandro, and Emiel van Miltenburg. 2015. "Sound-Based Distributional Models." In *Proceedings of the 11th International Conference on Computational Semantics*, 70–75. London, UK: Association for Computational Linguistics.

Malcolm, George L., Michelle Rattinger, and Sarah Shomstein. 2016. "Intrusive Effects of Semantic Information on Visual Selective Attention." *Attention, Perception & Psychophysics* 78 (7): 2066–78.

Mastroberardino, Serena, Valerio Santangelo, and Emiliano Macaluso. 2015. "Crossmodal Semantic Congruence Can Affect Visuo-Spatial Processing and Activity of the Fronto-Parietal Attention Networks." *Frontiers in Integrative Neuroscience* 9 (July): 45.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. "Advances in Pre-Training Distributed Word Representations." *ArXiv [Cs.CL]*. arXiv. http://arxiv.org/abs/1712.09405.

Moores, Elisabeth, Liana Laiti, and Leonardo Chelazzi. 2003. "Associative Knowledge Controls Deployment of Visual Selective Attention." *Nature Neuroscience* 6 (2): 182–89.

Moran, Zachary D., Peter Bachman, Phillip Pham, Seong Hah Cho, Tyrone D. Cannon, and Ladan Shams. 2013. "Multisensory Encoding Improves Auditory Recognition." *Multisensory Research* 26 (6): 581–92.

Nah, Joseph, and Joy Geng. 2021. "Thematic Object Pairs Produce Stronger and Faster Perceptual Grouping than Taxonomic Pairs." https://doi.org/10.31234/osf.io/6u3sn.

Nematzadeh, Aida, S. Meylan, and T. Griffiths. 2017. "Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words."

Richie, Russell, Wanling Zou, and Sudeep Bhatia. 2019. "Predicting High-Level Human Judgment across Diverse Behavioral Domains." *Collabra. Psychology* 5 (1). https://doi.org/10.1525/collabra.282.

Santangelo, Valerio, Simona Arianna Di Francesco, Serena Mastroberardino, and Emiliano Macaluso. 2015. "Parietal Cortex Integrates Contextual and Saliency Signals during the Encoding of Natural Scenes in Working Memory." *Human Brain Mapping* 36 (12): 5003–17.

Schneider, Till R., Andreas K. Engel, and Stefan Debener. 2008. "Multisensory Identification of Natural Objects in a Two-Way Crossmodal Priming Paradigm." *Experimental Psychology* 55 (2): 121–32.

Wisniewski, E. J., & Bassok, M. 1999. “What Makes a Man Similar to a Tie? Stimulus Compatibility with Comparison and Integration.” *Cognitive Psychology*. https://doi.org/10.1006/cogp.1999.0723.