UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Is Language a Window into the Human Mind? The Same Latent Features Guide Categorization in Language and Vision

Permalink
https://escholarship.org/uc/item/55m104sv

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Wu, Jingyi
Xu, Enjie
Bitterman, Ella Brooke
et al.

Publication Date
2026

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Is Language a Window into the Human Mind? The Same Latent Features Guide Categorization in Language and Vision

Jingyi Wu[1]
jingyiwu99@g.ucla.edu

Enjie Xu[1]
xuenjiek15@gmail.com

Ella B. Bitterman[2]
bitterman.el@northeastern.edu

Joshua B. Tenenbaum[3]
jbt@mit.edu

Joshua K. Hartshorne[4]
joshua.hartshorne@hey.com

Tao Gao[1,5]
taogao@ucla.edu

[1]Department of Communication, University of California, Los Angeles
[2]Northeastern University
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[4]Communication Sciences and Disorders, MGH Institute of Health Professions
[5]Department of Statistics, University of California, Los Angeles

Abstract

Do linguistic structures reflect the organization of human conceptual representations? A broad tradition in cognitive science argues that grammar is shaped by rich, compositional event representations that underlie cognition more broadly, motivating the use of linguistic patterns to probe the structure of thought. However, direct evidence linking linguistic and non-linguistic event cognition is sparse. Here, we test whether event categories are consistently represented across language and vision. Across two experiments, participants categorized event types using either linguistic descriptions or short video clips depicting the same classes of events. We find that humans reliably and consistently categorize events in both modalities, exhibiting parallel patterns of categorization across language and vision. Notably, this categorization is not inherent to every sophisticated cognitive system: State-of-the-art computational vision models failed at the same task. These results suggest that linguistic and visual event processing draw on a shared underlying system of event representation, providing evidence that language offers a window into human event cognition.

Keywords: event representation; verb semantics; conceptual structure; visual event perception

Introduction

Language imposes structure on the world, characterizing objects and events as agents, patients, paths, manners-of-motion, and causes, among other abstractions (Jackendoff, 2002; Talmy, 2000). A long-standing question in cognitive science is whether this is an idiosyncratic fact about language, or whether it reflects a more general structure of thought (Pinker, 2007). This might arise, for instance, if combinatorial conceptual representations (i.e., a language of thought; Fodor, 1975; Quilty-Dunn et al., 2023) preexisted language, which emerged as a mechanism for communicating those thoughts (Hartshorne & Snedeker, 2026). In that case, linguistic structure is constrained by the structure of the Language of Thought.

Promising results come from studies of objects and masses. For instance, people visually distinguish discrete, bounded entities that can be individuated and counted from substance-like material, such as water or sand, a contrast that parallels the count–mass distinction in language (Barner & Snedeker, 2005; Xu & Carey, 1996). Similar correspondences emerge for spatial relations: Linguistic prepositions encode abstract relational structure such as containment, support, and direction, and visual cognition likewise represents spatial relations in a structured and flexible way, supporting rapid extraction of relative configuration and spatial arrangement independently of object identity (Franconeri et al., 2012; Hafri et al., 2024; Majid et al., 2004).

However, where the Language-as-Window hypothesis (LWH) really shines is in event cognition. Decades of research has shown that a wide range of syntactic phenomena respect a relatively small number of semantic distinctions (Hartshorne et al., 2014; Jackendoff, 1992; Levin, 1993; Pinker, 2007; Pustejovsky, 1995). For instance, the syntactic behavior of verbs appears to be sharply constrained by meaning. A canonical example is the dative alternation, which concerns whether a verb may appear in both a prepositional object construction (*A verb B prep C*) and a double-object construction (*A verb C B*) (Levin, 1993). Speakers, including children, do not freely generalize the dative alternation across verbs (Ambridge et al., 2012; Oehrle, 1976; Pinker, 1989); instead, some verbs can only appear in the prepositional object construction, some only in the double object construction, and some in both. A major determinant is semantics: for instance, verbs involving punctate transfer of force can appear in the double object construction (*throw/kick him the ball*) while those involving continuous force cannot (**lift/push him the ball*). Such patterns indicate that subtle differences in argument structure are explained by event representations, and researchers have probed these phenomena to develop well-articulated theories of event representation (Jackendoff, 1992; Levin, 1993; Pinker, 1989; Talmy, 1988).

Other well-known examples involve how languages encode causation and force. Across languages, causal agents often appear as subjects of sentences and are more likely to appear earlier in the sentence than other event participants. This asymmetry reflects the directional structure of physical causation, in which force flows from an initiator to an affected entity (Croft, 2012; Dowty, 1991; Levin & Hovav, 2005). The transmission of physical force is also reflected in the constraint

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

that instruments can appear as subjects only when they can plausibly be construed as immediate transmitters of force, as in *the key opened the door* (Alexiadou & Schäfer, 2008). Similarly, some verbs readily license a direct object (*roll the ball*), while others systematically resist it (**cry him*), even when a causal interpretation is pragmatically imaginable. This contrast has been explained in terms of beliefs about what can be externally caused: physical changes, yes; mental states, no (Ambridge et al., 2008; Dixon & Aĭkhenvald, 2000; Pinker, 1989; Schäfer, 2008).

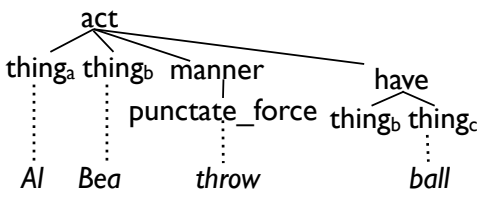

Figure 1: Example of the semantic representation for *Al threw Bea a ball*, based on Pinker (1989).

Particularly compelling are the existence of Levin verb classes, named for their discoverer (Levin, 1993). English admits of hundreds of syntactic constructions like the double object, prepositional object, and causative constructions described above. When verbs are grouped by which constructions they can appear in, the result is around 400 distinct classes, each of which appears to have a clear semantic basis (Hartshorne et al., 2014; Kipper et al., 2006; Levin, 1993). Hierarchical modeling suggests that these classes can be arranged in super-classes that share some (not all) of the syntactic constructions and which are also semantically related to one another — exactly as would be expected if a verb's ability to appear in a particular construction is determined by semantic distinctions that are shared across many verbs.

While not without controversy (see, for instance, Goldberg, 1995), a leading candidate for explaining the phenomena above involves positing that language is grounded in a structured, combinatorial event representation system (see Figure 1) made of primitives that themselves ground out in perceptual and motor systems (Hartshorne et al., 2016; Jackendoff, 2002; Levin & Hovav, 2005; Pinker, 1989) — that is, a language of thought.

Non-linguistic evidence for the Language-as-Window hypothesis (LWH)

While evidence from non-linguistic cognition is not inconsistent with the existence of the richly structured event representations described above, research to date provides only limited support around the margins. Much of the work on event cognition has focused on the largely-unrelated problem of event segmentation (dividing continuous videos into discrete "events"; Wang et al., 2024; Zacks & Swallow, 2007). Other work has focused on broad constructions like causation — certainly relevant to the LWH but consistent with many other possibilities.[1]

For instance, languages distinguish between telic and atelic predicates, reflecting whether an event is construed as having a natural endpoint. Observers can visually distinguish between actions that culminate in a clear endpoint and those that do not (Y. Ji & Papafragou, 2020), and they use this distinction to organize and categorize dynamic actions even when linguistic encoding is unavailable or disrupted (Y. Ji & Papafragou, 2020b). Similarly, as noted above, many linguistic phenomena revolve around whether an event is externally caused. Humans and even many animals are exquisitely tuned into the same distinction in non-linguistic tasks (Saxe et al., 2007; Vettori et al., 2024; Wilson et al., 2022; Wolff & Shepard, 2013). A third example involves motion events, with languages famously dissociated path (*enter, exit, circumnavigate*) from manner (*walk, run, crawl*) (Talmy, 2000), a distinction that also appears in non-linguistic cognition (Lakusta & Carey, 2015; Papafragou, 2010).

However, as described in the previous section, linguistic event representation theories involve dozens of more fine-grained distinctions, such as punctate vs. continuous force, creation vs. destruction, attachment vs. detachment, and beginning vs. ending (Levin & Hovav, 2005). Evidence that non-linguistic cognition is organized around these distinctions is far more sparse. Preliminary evidence comes from

[1]One complication is that much of the research comparing linguistic and conceptual representations has focused on Thematic Role theory (e.g., Frankland & Greene, 2015; Hafri et al., 2018; Papeo et al., 2024; Rissman & Majid, 2019; Segalowitz, 1982; Strickland et al., 2014; Vettori et al., 2024). Thematic role theory grew out of work in the 1960s attempting to explain syntactic phenomena not in terms of event representations but rather based on abstract categories of event participants, such as AGENT, PATIENT, and THEME (Fillmore, 1968; Gruber, 1965). Because thematic roles are descriptions of arguments and ignore the predicates themselves, they do not constitute a language of thought, any more than numbers by themselves comprise mathematics. It became clear by the 1990s that thematic roles by themselves cannot predict or explain the syntactic phenomena they were designed to explain (Jackendoff, 1992; Levin & Hovav, 2005). Many researchers abandoned thematic roles in favor of the complex event representations that are the focus of this paper (for a related approach, see Croft, 2012). Others — particularly in the generativist tradition — retained thematic roles but embedded them in complex abstract syntactic structures that do most of the work (e.g., Hale & Keyser, 2002). Others, mostly in the empiricist tradition, treat thematic roles as only one of many probabilistic predictors of syntactic phenomena (Goldberg, 1995, 2006). Thus, while thematic roles are still often used as convenient shorthand for otherwise-hard-to-name structural positions in more complex event representations (much as nuclear physicists long ago abandoned the notion that electrons orbit nuclei but still refer to "orbits"), there is broad (though not universal) agreement that these are approximations for something more complex, and the best one can hope for is a weak correlation with syntactic phenomena. Thus, the LWH would similarly predict that they are at best weakly correlated with non-linguistic cognitive behavior. Thus, observing such correlations is indicative of something, but not a strong test of the hypothesis. Conversely, finding that thematic roles do not perfectly predict behavior is expected and thus uninformative (*pace* Rissman & Majid, 2019).

Table 1: Overview of verb classes, tested verbs, and example sentence stimuli.

Verb Classes	Tested Verbs	Example Sentence Stimuli
Continuous force	lift, lower, pull, push, tote	David lowers a cup onto the table.
Instantaneous force	flick, kick, shove, slide, throw	Laura shoves the heavy gate open.
Attachment	fasten, pin, stick, strap, tape	Tina tapes a small box closed.
Destruction	chip, crush, rip, shatter, snap	Vanessa snaps a piece of bread in two.

the categorical perception of action types (e.g., twisting vs. rotating) (H. Ji & Scholl, 2024), and from the persistence of this categorical sensitivity even when linguistic encoding is suppressed (H. Ji & Scholl, 2025). However, these findings are limited to single actions, leaving open whether visual cognition — like language — supports generalization across higher-order action classes that abstract over surface motion or appearance.

In this study, we focus on four classes of verbs distinguished by their verb semantics and event structures proposed by Pinker (1989, 2007): (1) continuous force, (2) instantaneous force, (3) attachment via contact, and (4) destruction of an object. Specifically, we examine whether participants can infer a common event structure from either linguistic descriptions or videos and generalize this structure to novel verbs or actions. Methodologically, our design adapts category identification paradigms from visual event cognition (Y. Ji & Papafragou, 2020). We extend this approach by adopting a few-shot learning framework, in which participants are provided with only positive examples during learning and are then asked to generalize to novel verbs or actions that they have not previously observed, thereby increasing the demands on abstraction and category formation (Xu & Tenenbaum, 2007). Using the same learning-and-generalization logic across linguistic and visual experiments, we find that humans consistently categorize event types when presented with either visual or linguistic stimuli. This cross-modal consistency suggests that language and vision rely on a shared system for event representation, supporting the view that linguistic structure reflects underlying conceptual organization.

Experiment 1: Event Categorization based on Language

Experiment 1 aimed to validate the experimental paradigm by testing whether humans can reliably categorize verb classes using linguistic input alone. Guided by theories of verb meaning and argument structure (Pinker, 1989, 2007), we asked whether participants could distinguish among four verb classes defined by core event-structural properties. Because our focus is on broad semantic distinctions, each class subsumes multiple, more fine-grained Levin verb classes.

Methods

Participants The experiment employed a between-subjects design, requiring 40 participants per condition. This sample size was determined based on prior work on event perception and categorization (Y. Ji & Papafragou, 2020; Strickland & Scholl, 2015). In total, 160 participants (78 female, 81 male, 1 other; mean age = 43.21 years) from English-speaking countries were recruited via Prolific.

Materials We selected four verb classes that differ in their argument-structural properties. For each class, five representative verbs were chosen (see Table 1). All selected verbs denote concrete, physically observable actions, avoiding metaphorical or highly abstract meanings. For each verb, four distinct event scenarios were constructed, resulting in four unique sentences per verb. Each verb class comprised 20 sentences (5 verbs × 4 scenarios). To minimize the possibility that participants could rely on superficial linguistic patterns rather than the events described, all sentences were constructed using a uniform argument structure. Specifically, each sentence was constrained to include three core arguments. Table 1 shows the verbs and example sentences.

Design and Procedures The experiment used a few-shot learning paradigm to test whether participants could generalize verb-class knowledge to novel verbs from sentence descriptions. Each participant was assigned to one verb-class condition and completed three phases. In the learning phase, participants were told that all sentences belonged to the same verb class and read six sentences from three verbs in that class (two per verb). In the preview phase, participants read four sentences, each from a different verb class, to familiarize them with the range of categories. In the testing phase, participants read 16 sentences and judged whether each described event belonged to the learned class by responding "Yes" or "No." Figure 2 illustrates the stimuli selection and experimental flowchart.

Participants completed the experiment on their personal laptops. After providing informed consent, they read instructions and completed self-paced trials. No time limits were imposed. After all trials, participants completed a post-experiment questionnaire collecting demographic information and reporting the strategies they used for categorization.

Results

Performance was assessed using signal detection measures (Macmillan & Creelman, 2004). Collapsing across conditions, participants showed reliable sensitivity to verb classes ($d' = 1.28$), significantly above chance, $t(159) = 13.42$, $p < .001$. Sensitivity was also above chance in all four conditions, with the highest d' in destruction ($M = 2.06$)

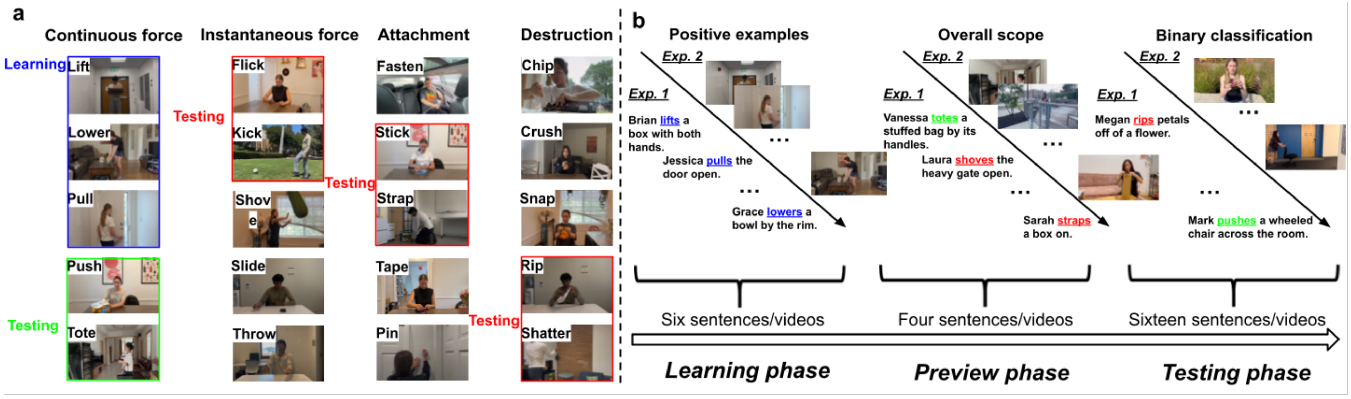

Figure 2: Stimuli selection and experimental flow. **(a)** Stimuli selection for each phase: participants learn a subset of verbs or actions and are tested on novel ones. Stimuli shown in green indicate positive examples drawn from the same verb class, whereas stimuli shown in red indicate negative examples drawn from a different class. **(b)** Flowchart of the experiment for both the language-based and video-based versions.

and attachment (M = 1.77), followed by instantaneous force (M = 0.80) and continuous force (M = 0.50; all ps<.001).

To assess differences in sensitivity across verb classes, we fit a linear mixed-effects model with d' as the dependent variable, verb class as a fixed effect, and random intercepts for participants. A likelihood-ratio test showed a significant effect of verb class, $\chi^2(3) = 55.05, p < .001$, indicating systematic variation in sensitivity. Using continuous force as the reference, sensitivity was higher for attachment and destruction (both $p < .001$), but did not differ reliably for instantaneous force ($p = .079$).

The confusion matrix (Figure 3) shows participants in the continuous force condition tended to overgeneralize to instantaneous force. Similarly, subjects in the instantaneous force condition tended to overgeneralize to continuous force. This explains the lower accuracy in these conditions. This confusion is not surprising, in that both classes involve caused motion. However, determining whether *in general* overgeneralization can be explained by similarity in event representations would require systematic study.

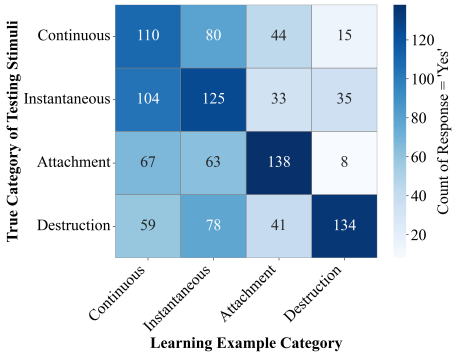

Figure 3: Confusion matrix of participants' affirmative ("Yes") responses for language stimuli. The x-axis shows the category learned during training, and the y-axis shows the true category of the test stimulus. Each cell indicates the number of trials in which participants judged a stimulus as belonging to the learned category (responded "Yes"), given its true category.

Experiment 2: Event Categorization based on Videos

Experiment 2 extends this investigation to the visual domain by asking whether the same event-class distinctions can be learned and generalized when actions are presented as visual events.

Methods

Experiment 2 followed the same design and procedure as Experiment 1, except as noted. A new set of 160 participants was recruited (59 female, 98 male, 1 other, and 2 who preferred not to disclose their gender; mean age = 41.54 years), evenly distributed across the four conditions.

Instead of reading sentences, participants viewed videos depicting the target verb actions. For each verb action, four

distinct action scenarios were recorded, varying actors and objects to ensure stimulus diversity. At the same time, experimental control was maintained: all videos were recorded under standardized conditions with a fixed camera position, a single fully visible actor, no occlusion, and no extraneous agents or contextual cues. Video duration ranged from approximately 2 to 5 seconds.

Results

Participants showed significantly above-chance sensitivity overall ($d' = 1.45$), $t(159) = 16.08$, $p < .001$, indicating robust discrimination between learned and unlearned action categories. Sensitivity was also above chance in all four conditions, highest for attachment via contact ($M = 1.95$), followed by destruction ($M = 1.62$), instantaneous force ($M = 1.55$), and lowest for continuous force ($M = 0.68$; all ps<.001).

We then fit a linear mixed-effects model with d' as the outcome, video class as a fixed effect, and random intercepts for participants. The model revealed a significant effect of video class, $\chi^2(3) = 30.01$, $p < .001$, with all non-reference conditions showing higher sensitivity than the continuous force condition (all ps<.001). This pattern indicates that categorization from visual input varied across video types.

Following Experiment 1, we examined overgeneralization patterns using a confusion matrix (Figure 4). Participants showed an asymmetric confusion pattern, overgeneralizing instantaneous-force videos as continuous-force videos but not vice versa.This may reflect broader conceptual coverage for continuous force, while the mechanisms underlying this asymmetry remain open.

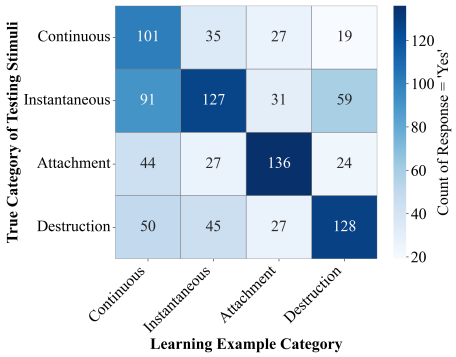

Figure 4: Confusion matrix of participants' affirmative ("Yes") responses for video stimuli. The x-axis shows the category learned during training, and the y-axis shows the true category of the test stimulus. Each cell indicates the number of trials in which participants judged a stimulus as belonging to the learned category (responded "Yes"), given its true category.

Experiment 3: Verb Class Categorization by Computer Vision Models

The correspondence between language and vision is less compelling if any complex cognitive system would categorize the events in the same way. Perhaps language and vision are simply dividing nature at its joints. Thus, in Experiment 3, we presented our video classification task to three recent video models as well as a multimodal language model (see below). These models are impressively successful at a variety of vision tasks, such as action recognition (Bertasius et al., 2021), and video-text retrieval (Radford et al., 2021); thus, if visual competency is all that is required to succeed at our video classification task, then these models should succeed as well.

Computational Categorization Framework

To test whether computational models have the capacity to categorize different actions as humans, we adopted two approaches, which are perceptual similarity-based decisions, and abstract, instruction-following inference.

Representation-Driven Categorization Categorization is performed by operating directly on learned embedding spaces derived from visual input. To standardize the visual information available to each model, we extracted five key frames from each video and provided these frames as input to the representation models. We consider two types of visual representation models. The first consists of video-only representations learned by transformer-based video models, such as VideoMAE (Tong et al., 2022) and TimeSformer (Bertasius et al., 2021), which encode spatiotemporal structure from visual input. The second consists of vision–language representations obtained from CLIP (Radford et al., 2021), in which visual inputs are aligned with linguistic semantics through contrastive pretraining.

Computational categorization was evaluated using a leave-one-action-out procedure, in which each action served as the held-out test instance once. For each fold, one action was held out from a target verb class, while one-class models were trained separately for all four classes using the remaining actions. The held-out action was subsequently evaluated against each class-specific model. This procedure yields four binary decisions for every held-out action, indicating whether the test action is classified as in-class or out-of-class under each method. We enumerated this process across all actions so that each action served as the held-out test case once. We implemented one-class learning using three decision methods: one-class support vector machines, Isolation Forests, and Local Outlier Factor.

Instruction-Driven Categorization To implement the instruction-driven categorization, we used GPT-5, which is one of the state-of-the-art multimodal language models capable of integrating visual and linguistic information to support vision-based reasoning and categorization. The task structure was designed to parallel the human learning paradigm as

closely as possible. Similar to representation models, GPT-5 process five key frames extracted from the video. Model responses were recorded as binary categorization decisions, indicating whether each test action was judged to belong to the target class.

Results

Table 2 reports sensitivity (d') for each computational model. Across models and approaches, sensitivity was at or below chance, indicating poor discrimination between target and non-target actions. This low sensitivity is driven by a systematic bias toward "no" responses, which suppresses true positives and reduces action categorization performance.

Table 2: Sensitivity (d') of computational models in action categorization

Models	**Representation-driven categorization**		
	SVM	**IF**	**LOF**
VideoMAE	-0.693	-0.825	-1.202
TimeSformer	-0.081	-0.584	-0.915
CLIP	-0.415	-0.304	-1.193
	Instruction-driven categorization		
GPT-5	-0.018		

Discussion

This study shows that, across both language and vision, human participants can infer the relevant verb category from minimal positive evidence and reliably generalize this category to novel verbs. In contrast, despite extensive training on large-scale video datasets, current vision models failed to show comparable generalization. These results suggests that human verb and action categorization relies on shared event-structural distinctions across modalities, supporting the Language-as-Window hypothesis that linguistic structure reflects an underlying conceptual system shared with vision.

By comparing performance in language- and vision-prompted tasks, we find that although humans are sensitive to event-type distinctions in both modalities, the degree of sensitivity differs across conditions (Figure 5). In particular, performance in the instantaneous condition was reliably higher in the vision task than in the language task ($t(78) = -3.592$, $p < .001$). This pattern is theoretically informative. Instantaneous force events are primarily defined by their manner of motion—that is, by how an action unfolds over time, characterized by brief, abrupt force application and a sharp temporal contour. Vision provides direct access to these motion dynamics, which are treated as primitive components of event representation. By contrast, manner of motion in language is typically lexicalized rather than systematically encoded in syntactic structure (Jackendoff, 1992; Pinker, 1989). As a result, accessing manner information in the language task requires verb-specific semantic retrieval and inferential mapping

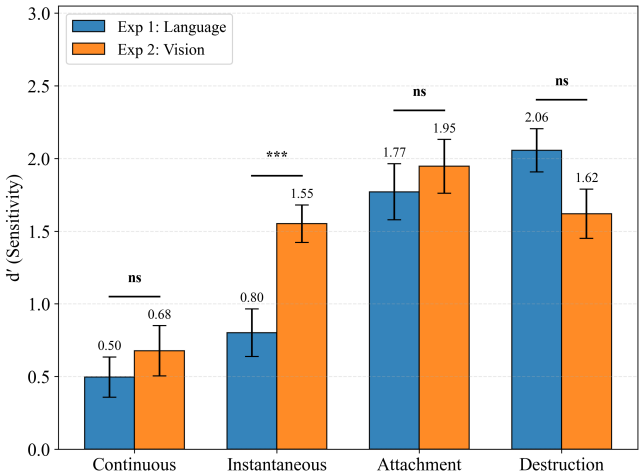

Figure 5: Comparison of sensitivity (d') across language and vision.

from lexical meaning to underlying event representations, introducing additional processing demands (Friederici, 2002, 2011).

We also found some interesting results when examining confusion patterns within each verb-class condition. It shows a clear directional asymmetry emerges between continuous and instantaneous force events: participants are more likely to overgeneralize instantaneous-force events as continuous-force events than to make the reverse error (Figure 3, Figure 4). This pattern suggests a tentative asymmetry in conceptual coverage, whereby continuous-force events may occupy a broader representational space that can accommodate instantaneous-force events as a limiting case, whereas instantaneous-force events are comparatively restricted and do not readily generalize to continuous force. This asymmetry may be informed by ideas from Event Segmentation Theory, which suggests that event categories can emerge from the range of dynamics supported by a single predictive event model (Zacks & Swallow, 2007). Continuous-force events are defined by sustained causal influence rather than by uniform force application; as a result, momentary changes in force—such as brief increases, decreases, or pauses—do not violate the event structure and can be treated as internal variation within an ongoing force process. This is analogous to perceiving someone as walking even if they slow down, speed up, or briefly stop: these fluctuations do not disrupt the interpretation of a continuous action. By contrast, an event defined as taking a step is inherently bounded, and extending it over time no longer fits the same event description.

Acknowledgments

This work was supported by ONR Vision Language Integration grant (PO 951147:1) and DARPA ECOLE grant.

An AI-based language model (ChatGPT, Claude) was used to assist with improving the experiment website code, and

refining the prompts used in the instruction-driven categorization task.

References

Alexiadou, A., & Schäfer, F. (2008). *Instrument subjects are agents or causers*. Universitätsbibliothek Johann Christian Senckenberg.

Ambridge, B., Pine, J. M., Rowland, C. F., & Chang, F. (2012). The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language*, *88*(1), 45–81.

Ambridge, B., Pine, J. M., Rowland, C. F., & Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*(1), 87–129. https://doi.org/10.1016/j.cognition.2006.12.015

Barner, D., & Snedeker, J. (2005). Quantity judgments and individuation: Evidence that mass nouns count. *Cognition*, *97*(1), 41–66. https://doi.org/10.1016/j.cognition.2004.06.009

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *Proceedings of the International Conference on Machine Learning (ICML)*, *139*, 1268–1278.

Croft, W. (2012). *Verbs: Aspect and causal structure*. OUP Oxford.

Dixon, R. M. W., & Aĭkhenvald, A. Y. (Eds.). (2000). *Changing valency: Case studies in transitivity*. Cambridge University Press.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619.

Fillmore, C. (1968). The case for case. *Universals in Linguistic Theory*.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227. https://doi.org/10.1016/j.cognition.2011.11.002

Frankland, S. M., & Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, *112*(37), 11732–11737.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84. https://doi.org/10.1016/S1364-6613(00)01839-8

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. OUP Oxford.

Gruber, J. S. (1965). *Studies in lexical relations.* [Doctoral dissertation, Massachusetts Institute of Technology].

Hafri, A., Sun, Z., & Firestone, C. (2024). The psychophysics of compositionality: Relational scene perception occurs in a canonical order. *Journal of Vision*, *24*(10), 681. https://doi.org/10.1167/jov.24.10.681

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, *175*, 36–52. https://doi.org/10.1016/j.cognition.2018.02.011

Hale, K., & Keyser, S. J. (2002). *Prolegomenon to a theory of argument structure*. MIT press.

Hartshorne, J. K., Bonial, C., & Palmer, M. (2014). The verb-corner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 397–402.

Hartshorne, J. K., O'Donnell, T. J., Sudo, Y., Uruwashi, M., Lee, M., & Snedeker, J. (2016). Psych verbs, the linking problem, and the acquisition of language. *Cognition*, *157*, 268–288. https://doi.org/10.1016/j.cognition.2016.08.008

Hartshorne, J. K., & Snedeker, J. (2026). Conceptual nativism: The origins of language are in the structure of thought. *PsyArXiv*. https://doi.org/10.31234/osf.io/ur3gj

Jackendoff, R. S. (1992). *Semantic structures* (Vol. 18). MIT Press.

Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Ji, H., & Scholl, B. J. (2024). "Visual verbs": Dynamic event types are extracted spontaneously during visual perception. *Journal of Experimental Psychology: General*, *153*(10), 2441–2453. https://doi.org/10.1037/xge0001636

Ji, H., & Scholl, B. J. (2025). Pouring, scooping, bouncing, rolling, twisting, and rotating: Does spontaneous categorical perception of dynamic event types reflect verbal encoding or visual processing? *Attention, Perception, & Psychophysics*, *87*, 2430–2441. https://doi.org/10.3758/s13414-025-03141-3

Ji, Y., & Papafragou, A. (2020). Is there an end in sight? Viewers' sensitivity to abstract event structure. *Cognition*, *197*. https://doi.org/10.1016/j.cognition.2020.104197

Ji, Y., & Papafragou, A. (2020b). Midpoints, endpoints and the cognitive structure of events. *Language, Cognition and Neuroscience*, *35*(10), 1465–1479. https://doi.org/10.1080/23273798.2020.1797839

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2006). Extending verbnet with novel verb classes. *LREC*, 1027–1032.

Lakusta, L., & Carey, S. (2015). Twelve-month-old infants' encoding of goal and source paths in agentive and non-agentive motion events. *Language Learning and Development*, *11*(2), 152–175.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levin, B., & Hovav, M. R. (2005). *Argument realization*. Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd). Psychology Press. https://doi.org/10.4324/9781410611147

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? the case for space. *Trends in cognitive sciences*, *8*(3), 108–114.

Oehrle, R. T. (1976). *The grammatical status of the english dative alternation* [Doctoral dissertation, Mass. Cambridge].

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive science*, *34*(6), 1064–1092.

Papeo, L., Vettori, S., Serraille, E., Odin, C., Rostami, F., & Hochmann, J.-R. (2024). Abstract thematic roles in infants' representation of social events. *Current Biology*, *34*(18), 4294–4300.

Pinker, S. (1989). *Learnability and cognition, new edition: The acquisition of argument structure*. MIT Press.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Pustejovsky, J. (1995). The generative lexicon.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, e261. https://doi.org/10.1017/S0140525X22002849

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic bulletin & review*, *26*(6), 1850–1869.

Saxe, R., Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental psychology*, *43*(1), 149.

Schäfer, F. (2008). *The syntax of (anti-)causatives*. John Benjamins. https://doi.org/10.1075/la.126

Segalowitz, N. S. (1982). The perception of semantic relations in pictures. *Memory & Cognition*, *10*(4), 381–388.

Strickland, B., & Scholl, B. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, *144*, 570–580. https://doi.org/10.1037/a0037750

Strickland, B., Fisher, M., Keil, F., & Knobe, J. (2014). Syntax and intentionality: An automatic link between language and theory-of-mind. *Cognition*, *133*(1), 249–261.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100. https://doi.org/10.1207/s15516709cog1201_2

Talmy, L. (2000). *Toward a cognitive semantics (vols. 1–2)*. MIT Press.

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, *35*, 10078–10093.

Vettori, S., Odin, C., Hochmann, J.-R., & Papeo, L. (2024). A perceptual cue-based mechanism for automatic assignment of thematic agent and patient roles. *Journal of Experimental Psychology: General*.

Wang, Y. C., Adcock, R. A., & Egner, T. (2024). Toward an integrative account of internal and external determinants of event segmentation. *Psychonomic Bulletin & Review*, *31*(2), 484–506.

Wilson, V. A., Zuberbühler, K., & Bickel, B. (2022). The evolutionary origins of syntax: Event cognition in nonhuman primates. *Science Advances*, *8*(25), eabn8464.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. In *Psychology of learning and motivation* (pp. 167–202, Vol. 58). Elsevier.

Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, *30*(2), 111–153. https://doi.org/10.1006/cogp.1996.0005

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84. https://doi.org/10.1111/j.1467-8721.2007.00480.x