Semantic similarity and event categorization

Aron Marvel and Jean-Pierre Koenig

What we have looked at

- Similarity of word senses in lexical retrieval
- Similarity of meaning of words in syntactic frame selection and as cues to syntactic memory retrieval
- Similarity of participant role fillers as the underpinning of plausibility
- Similarity of the set of possible participant role fillers and its effect on reading

Similarity of participant role fillers and categorization

- How does similarity of verbs and objects predict event categories?

    1. John raised his arm/The forklift raised the pallet

- Lexicographic research assumes discrete, recognizable number of senses per verb

    - More for splitters (WordNet)

    - Less for lumpers (Cobuild)

- Are we undercounting the number of “senses”?

The reality of verb uses

- Verb senses and sentences denote event categories

- Event categories named by a verb sense can also be named by the combination of a verb and a dependent

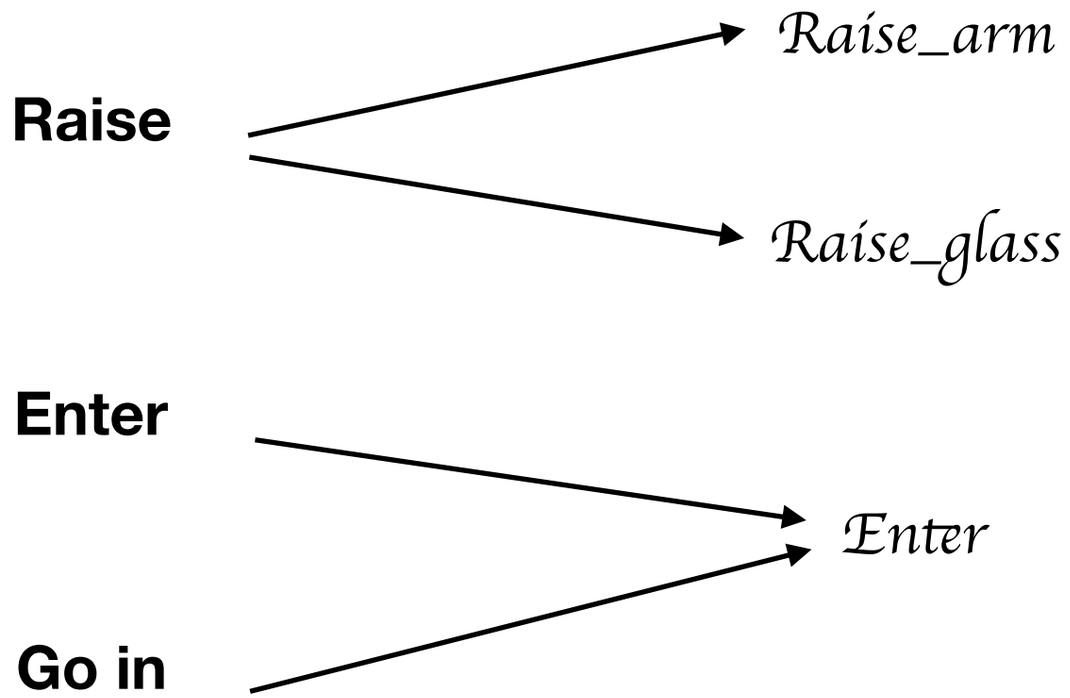
1. The officer entered the building/The officer went into the building

- Are event categories named by a single sense of *cut* or *raise* the same?

2. Marc cut the lawn/Marc cut the cake/cut his hair.

3. The senator raised a glass in celebration/The crane raised the car out of the water.

The relation between verb senses and categories



What is wrong with a lexical approach to event categorization

- Combination of words can contribute to an event category (c.f. *enter* vs. *go in*)
    - Languages vary in how many event categories map onto word senses vs. sequences of verb senses
    - Wagiman (Wilson): 500 verbal expressions most with one sense compared to English 4,000+ verbs with 3 to 4 senses
- A single word sense may contribute several categories (c.f. *cut*, *raise*)

Determining dimensions of event categorization

- If event categories might not be identifiable with just verb senses, we need to start with sentences (the *true* Frege principle)

- How do people group together events described by sentences that include the same verb (i.e. form event categories)?

- We started with a few dimensions we could find evidence for in the literature to establish “our” gold standard

Event complexity

- Does the event have recognized event subparts

1. He refused to sell any of his antiques

2. The support staff sells their expertise to the community beyond the school

Time scale

- How long does the event typically last?

1. Royal Bank of Scotland bought Bank Worcester at the end of 1990.

2. I stopped at a bar just long enough to buy two cheese rolls

Agent type

- Is the agent animate or inanimate? Is is a group or an individual?

1. A Genoese fleet rescued the city.

2. Archeologists rescue information about the past before it is destroyed

Socio-cultural salience

- Some event categories are distinguished by the fact that they are part of a socio-culturally important activity

1. The room is for pupils to borrow books.

2. Can you borrow an iron for me?

Inferences

- Some event descriptions lead to lots of inferences and those are likely to be event categories

1. She adjusted the scarf to cover the bruises forming on her neck.

2. The children covered their eyes and turned away as the needle went in

Specific motion sequence

- Some event categories are characterized by a sequence of motions

1. Charlery pulled the ball behind Halsall.

2. The General shouted at his men to pull the barricade down.

The role of event complexity in discourse

- Often assumed that discourse processing is subject to general principles
    - Aktionsart: events move forward reference time, states do not
    - Strong iconicity: successive sentences describe successive, *contiguous* eventualities (Dowty, Zwaan)
- Strong iconicity meshes well with simulation semantics

Expectation-based processing

- Hypothesis: Temporal update is sensitive to particulars of described situations, not just general principles
- How complex an event is matters for temporal update
    - If you are a paleoontologist giving a lecture the “next” event can be quite distant from the preceding event

TABLE 2

Examples of Materials Used for Sentence Continuation Tasks in Experiments 1–3 and the Self-Paced Reading Task in Experiment 4

	Example Stimuli for Experiment 1
Temporary states	The baby required a vaccination.
	Alex craved spicy chicken wings.
Permanent states	Andy loved science fiction movies.
	Claire trusted her bank.
	Example Stimuli for Experiments 2–3
Simple events	William picked up his tennis racket. (Then, ...)
	Joshua buttoned his coat. (Then, ...)
Complex events	Washington DC reduced the rate of violent crime. (Then, ...)
	The research hospital tested the new vaccine. (Then, ...)
	Example Stimuli for Experiment 4
Simple events	Mary poured water in a glass. After a few $seconds_{short}$/$weeks_{long}$, she drank it.
Complex events	The hospital collected DNA sample from AIDS patients. Several $minutes_{short}$/$months_{long}$ later, they tested them and analyzed the data.

Prompts described temporary and permanent states in Experiment 1 and simple and complex events in Experiment 2. Experiment 3 used identical materials with the addition of the forward-moving temporal connective *then*. Materials for Experiment 4 used the simple and complex events used in Experiments 2–3, followed by a second sentence describing an event that is plausible to happen next. Temporal connectives describing both short and long temporal intervals connect both sentences.

TABLE 1

Predictions Made by *Aktionsart*-, Iconicity-, and Expectation-Based Models of Temporal Update, When the First Sentence Describes Temporary and Permanent States (as in Experiment 1) or Simple and Complex Events (as in Experiment 2)

Predictions for Experiment 1

	Basis of Predictions	
	Aktionsart	*Narrative Expectations*
No movement	Permanent = temporary	Permanent > temporary
Forward movement	Permanent = temporary	Temporary > permanent
Backward movement	Permanent = temporary	Temporary > permanent

Predictions for Experiment 2

	Basis of Predictions	
	Iconicity	*Narrative Expectations*
No movement	Complex = simple	Complex > simple
Forward movement	Complex = simple	Simple > complex
Backward movement	No prediction	No prediction

Temporal update patterns are predicted to occur more on the state/event type on the left of the inequality symbol more often than the ones on the right.

TABLE 3

Raw Counts of Elicited Responses Grouped According to Temporal Movement and State Type for Experiment 1 and Temporal Movement and Complexity Type for Experiment 2

	Forward	*Backward*	*Static*
	Experiment 1: temporary and permanent states		
Permanent states	63	101	287
Temporary states	155	166	130
	Experiment 2: simple and complex events		
Simple events	414	203	156
Complex events	304	177	314

Not all states or events are created equal

- Some states evoke boundaries and readers expect movement of narrative time

- Complex events are more likely to keep narrative time static (leading to elaborations) than simple events

- In both cases, readers pay attention to properties of particular event category being described

- Size of narrative time movement is a function of event complexity (*what happens next? Time between two events*)

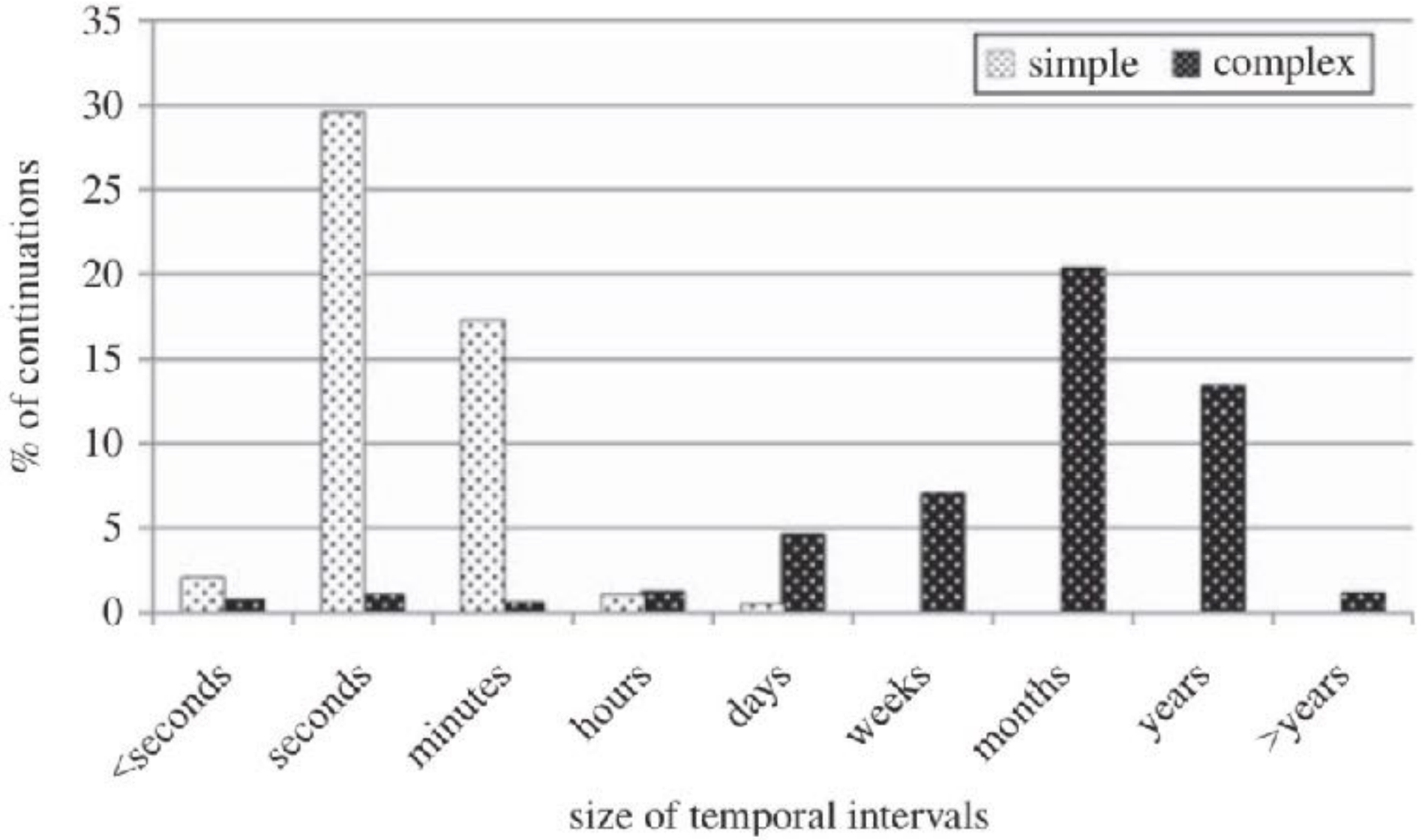


FIGURE 1 Distribution of size of temporal intervals between eventualities described in the continuations and the previous discourse segment in Experiment 3.

Cost of temporal update is a function of expectations

- If narrative time mirrors structure of real time, it should take longer to process phrases that describe longer temporal intervals between events

- If narrative time does not mirror real time, processing of the temporal phrases should be a function of readers' expectations (given the first event)

TABLE 4
Residualized Reading Times and the Fixed-Effect Structure of the Regression Models for the Temporal Connective and the Subject of the Second Clause in Experiment 4

Reading Times (ms)

	Event Type			
	Simple		*Complex*	
Region	*Short*	*Long*	*Short*	*Long*
Temporal connective	352	350	396	390
Subject of second clause	371	383	390	362

Model Summaries

	Estimate	*SE*	*t*	*p*
	Temporal connective			
(intercept)	371.85	16.06	23.16	
Complexity	43.22	30.32	1.43	.154
Temporal size	−3.47	14.72	−.24	.814
Probability of elaboration	7.53	15.2	.49	.621
Complexity × temporal size	−5.09	29.44	−.17	.863
	Subject of second clause			
(intercept)	378.8	18.12	20.91	
Complexity	2.91	26.4	.11	.912
Temporal size	−9.75	10.59	−.92	.357
Probability of elaboration	.75	13.28	.06	.955
Complexity × temporal size	−45.99	21.18	−2.17	.0301*

Subjects, items, and lists are random effects.

*Predictors that reached significance.

Hand-annotation of event categories

- 10 verbs: *bake*, *borrow*, *buy*, *cover*, *deliver*, *frighten*, *immerse*, *pull*, *rescue*, *sell*

- Sample sentences with unique non-pronominal, non-proper name subjects and direct objects (20 lists) x 100 sentences each (if that many in BNC) = 1,602 sentences

    - More pronouns and proper names in subject position (49.64% vs. 19.51% and 12.23% vs. 3.18% respectively)

- Categorization using the 6 *a priori* dimensions for each verb; reconciliation among two raters

More categories than verb senses

- Can native speakers categories match verb senses?
    - American Heritage Dictionary: 3.8 senses per verb for our sample sentences; our rating: 16.5 event categories
- What contributes more to distinguishing categories, properties of subjects or objects?
    - 62% of distinct categories came from object list, so objects seem to contribute more to distinguishing event categories

Verb	AHD senses	Categories
bake	2	10
borrow	2	18
buy	3	18
cover	8	30
deliver	7	17
frighten	2	14
immerse	3	8
pull	6	24
rescue	1	13
sell	4	13
Average	**3.8**	**16.5**

Table 1. Comparison of AHD senses to event categories discovered by application of the parameters discussed in Section 3.

There were differences in impact of 6 dimensions

- Agent-type (plurality, animacy, abstractness) mostly for subjects
- Time scale and specific motion sequence did not play much of a role
- Reversal of relative importance of subjects and object for *frighten*: 64% of *frighten* categories distinguished by combinations of subject and verb
- The more semantically similar the subjects or objects of a verb were, the more likely the verb+subjects or verb+objects were to be put in the same category
    - Cover their feet/their hands/their city

Testing the effect of semantic similarity of subjects and objects on categorization

- LSA can be used to test how much semantic similarity of subjects or objects predicts raters' categorization

- 400-dimension semantic space created from BNC

- Pair-wise relatedness values (5,050 values for subject and object lists)

- Clustering of subjects and objects based on their pairwise semantic similarity using average-linkage dendrogram

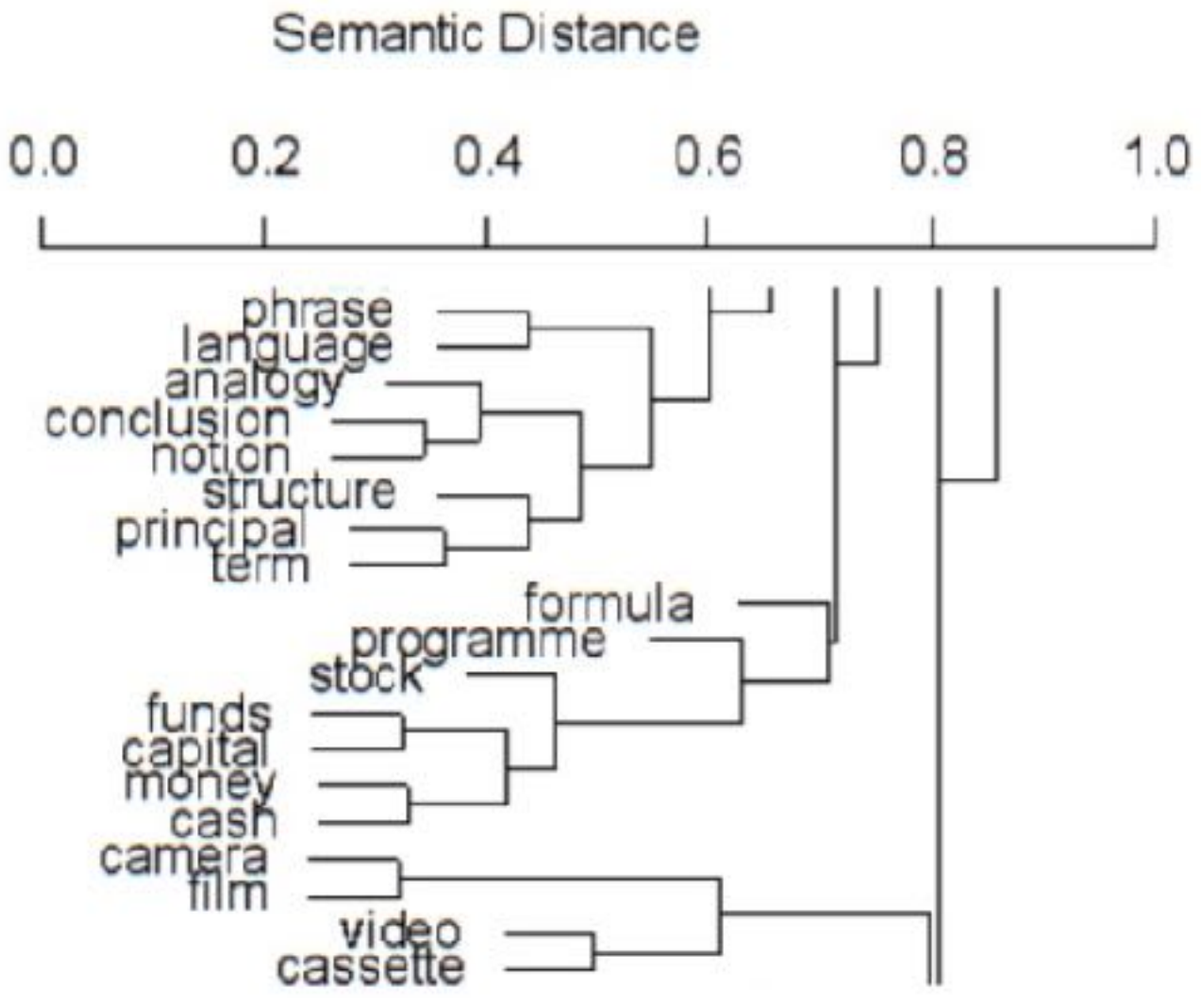


Figure 1. A section of the dendrogram created by using LSA semantic distance values to group direct objects of the verb *borrow*.

Semantic similarity predicts event categories

List	P LSA	R LSA	F LSA	F Rand	Ratio
Subj	40 %	80 %	0,53	0,39	1,38
DO	35 %	66 %	0,46	0,32	1,46
Overall	38 %	73 %	0,50	0,35	1,42

Do similarity judgments match raters' categorization

- Two sentences put into the same event category should be judged more similar semantically than two sentences put in different event categories

- 96 sentence pairs balanced across 3 groups for 8 verbs (*immerse* and *bake* excluded due to data sparsity)

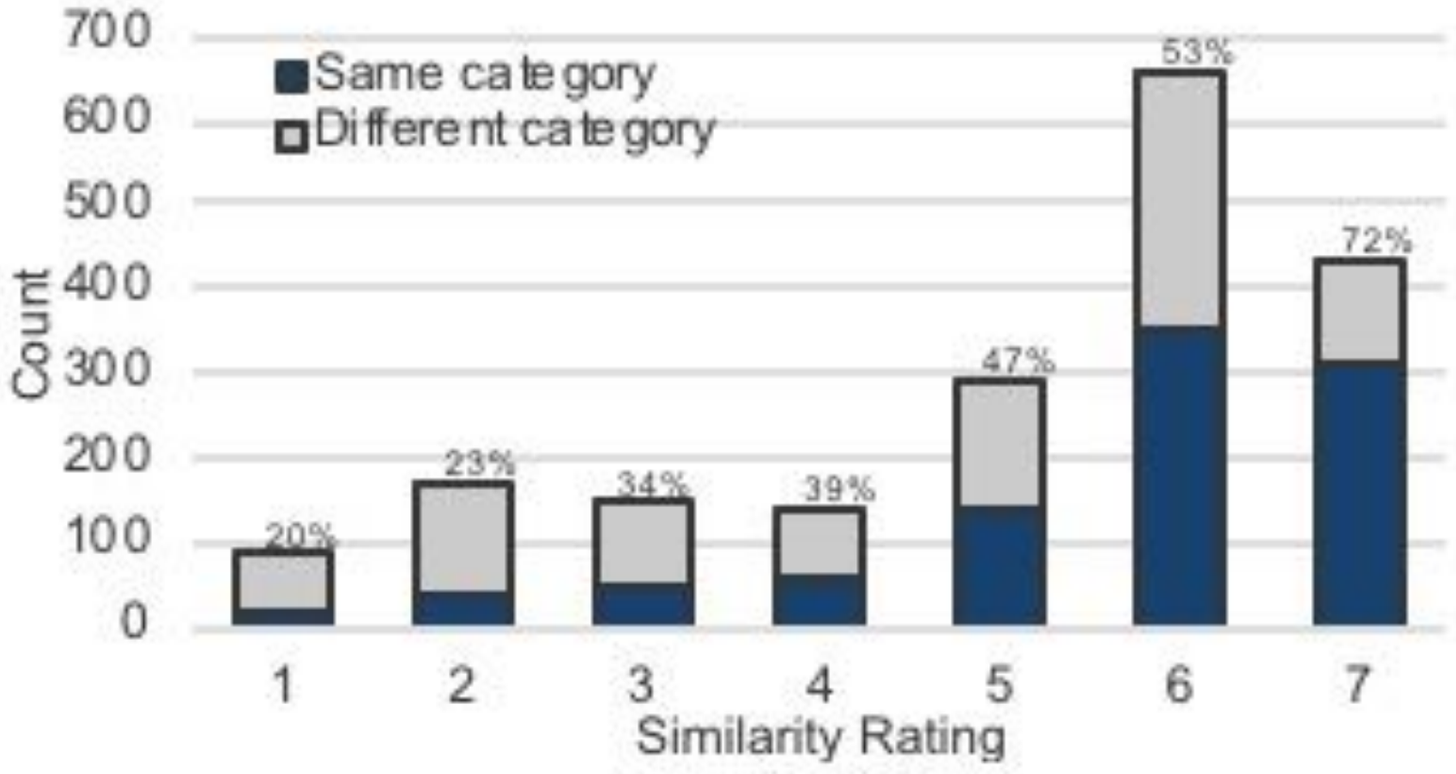
- Participants asked to judge the similarity of situations described by pairs of sentences on a 1-7 scale

Item group	Verb sense	Rater
1	same	same
2	different	different
3	same	different

- Judgments binarized into “above participant’s median score” and “below participant’s median score”

- Scores agreed with categorization if above (below) median and in the same (different) rater category

More similar = more likely to be in the same category



- **78%** of judgments were below the participant's median when events in different rater categories
- There was a significant relationship between similarity judgments and category assignment with a medium to large effect size (X^2=218.64, N=1129, p<.001, V=.44)

“Deriving” the dimensions of categorization from participants

- Our 6 dimensions of categorization are motivated, but what if others are also important?

- We conducted a sorting + justification task to see what dimensions are relevant for “ordinary” participants

- 24 verbs in 6 distinct semantic domains (to cover as much as possible of semantic space with limited set of verbs)

- For each verb we extracted 20 sentences from ANC, which were simplified a bit (e.g., avoiding very complex relative clauses)

The sorting task

- Each participant saw 6 verbs, 120 participants in all
    - 6 semantic domains: feeling, physical action, perception/mental attitude, possession, change of state
    - 24 verbs from list of 1,000 most frequent English verbs (COCA)
    - 20 sentences ps-randomly extracted from ANC
- Task:
    - Sort 20 sentences (with the same verb) into as any number of groups based on similar described events are
    - List any features you used to identify each group

Feature extraction

- 1,948 unique features
- Justification was a free form task, so we need to “standardize” responses
- We used LSA semantic similarity and clustering to group responses into “features”
    - R’s NbClust package was used to find an optimal number of feature clusters
    - Using k-means clustering, two indices had high optimality:
        - Hartigan, which optimizes based on maximum distance between hierarchy levels resulted in 11 standardized features
        - SDbw, which optimizes based on compactness and separation between clusters and resulted in 31 standardized features
- One cluster is “junk” (the *Other* category)

How we used the standardized features

- Standardized features were ranked according to both frequency of use and distinctiveness (as per cue validity, *p(c|f)-p(c)*)

- Mixed effect models to determine whether particular verbs or domains predict increase use of specific standardized features during categorization

- Agreement between participants w.r.t. categorization measured for each verb (for which verbs do participants agree the most on pairs of sentences being in the same or different categories?)

Hartigan-based features

Hartigan index (11 standardized features)		
Rank	**Most frequent**	**Most distinctive**
1	(GRAB BAG)	gender men/women
2	business, environment, community interactions, prepositional phrases (in/of), causes/reasons, sensory perception	animals
3	things, human/nonhuman, individuals	government, military, politics
4	people	inanimate objects
5	groups	items, money, metalinguistic
6	inanimate objects	one person
7	individuals/one person	groups
8	government, military, politics	people
9	items, money, metalinguistic	things, human/nonhuman, individuals
10	animals	business, environment, ++
11	gender men/women	*

SDbw index-based features

- See file
- The verbs with highest agreement tend to describe more concrete events:
  - participants agree more on which physical action events belong to the same category than on which perception/mental attitude and feeling events belong to the same category

Does similarity of subjects and objects help predict categorization?

- We grouped together sentences based on similarity of subjects and, separately, similarity of objects

- We compared that categorization to participant categories

- Used two measures to compare automatically generated categories and participant categories: Adjusted Rand index (ARI) and the harmonic mean of precision and recall

Yes, similarity of Ss and Os help

- ARI scores (0 = change category overlap; 1 = identical category):
  - GloVe: 0.14; LSA: 0.05; Word2vec: 0.12
- F values (improvement over randomly generated categories):
  - GloVE: 35% improved overlap; LSA: 17%; Word2vec: 37%