Constraint on covariation: It's not meaning

Arthur M. Glenberg and Sarita Mehta

Covariation among words is certainly related to meaning, meaning similarity, and psychological processing. We argue, however, that the causal arrow is from meaning (and meaning similarity) to covariation, not *vice versa*. Consequently, covariation is not meaning, it is unlikely to provide an accurate metric for similarity of meanings, and embodied learning mechanisms, rather than computation of statistics, underlie effects of covariation on psychological processing. We report the results from two experiments that provide the first empirical test of the strong covariation claim that meaning can be derived from covariation structure. In the experiments, people studied the covariation among unnamed features taken from a familiar domain. In the first experiment, after learning the covariation structure of the features, participants were unable to choose the correct domain on a forced choice test, and they were unable to use the learned structure to grossly classify unnamed features even after the domain and majority of features were named. In the second experiment, the majority of the features was named during the study of the covariance structure. Nonetheless, participants were unable to use the learned structure to classify the few remaining unnamed features. Thus, contrary to the strong covariance claim, covariance structure alone is not particularly useful for deriving meaning.

Keywords: meaning, distribution hypothesis, embodiment, covariation, connectionism

Constraint on Covariation: It's not meaning

The distributional hypothesis is that words with similar distributional properties have similar meanings (e.g., Baroni & Lenci, this volume; Sahlgren, this volume). The hypothesis has great appeal and makes intuitive sense, but to what degree is it correct? Some (e.g., Burgess & Lund 1997; Landauer & Dumais 1997; Sahlgren, this volume) have also proposed that "distributional representations *do* constitute full-blown accounts of linguistic meaning" (Sahlgren, emphasis in the original). In contrast, we argue that a) distributional analyses cannot provide a sufficient basis for similarity of meaning, b) the claim that distributional information is a full-blown account of meaning is, except under a solipsistic interpretation of meaning, unlikely

 (ricevuto nell'ottobre 2008)

to be correct, and c) using the data from two experiments, that real people have a difficulty using distributional information alone to determine even an approximate meaning for new symbols.

The papers in this special edition of the *Italian Journal of Linguistics* demonstrate the appeal of distributional information, particularly for machine-based natural language processing. For example, Fazly and Stevenson (this volume) propose that distributional information can be used to automatically classify multiword expressions (MWEs) into one of four classes ranging from the idiomatic (e.g., "shoot the breeze") to the mostly compositional (e.g., "give a present"). They describe ten distribution-based measures, such as institutionalization (joint co-occurrence probability), and fixedness (e.g., lack of syntactic variation) that might predict the MWE class. Combining the ten distribution-based measures leads to modest success, in that the proportion of correct classifications of the MWEs ranged from .39 to .67.

Rumshisky (this volume) sets for herself a similarly daunting task: using distributional analyses to distinguish among the multiple senses of polysemous verbs. The computational technique is based on the concept of selectors, words with which the target verb forms syntactic dependencies. Her five-step algorithm attempts to identify different sets of selectors for the different senses of the target verb so that the occurrence of the target with members of a selector set should identify the relevant meaning of the target.

Baroni & Lenci (this volume), Sahlgren (this volume), and Schulte im Walde & Melinger (this volume) echo Fazly & Stevenson in suggesting that different distributional measures may pick up on different aspects of meaning. For example, Baroni & Lenci analyze two distributional approaches to finding the properties of concepts. One approach, Singular Value Decomposition, examines the co-occurrence of words within five words of a target. The dimensionality of the resulting co-occurrence matrix is reduced using singular value decomposition, and the nearest (in SVD space) neighbors of a target word are taken as properties of that target word. The second distributional approach, StruDEL (for Structured Dimensional Extraction and Labeling), finds words with particular relations to a target (not just co-occurrence) before attempting to generalize patterns for that target. For both distributional analyses, the question of interest is how the properties of the target word identified through distributional analyses compare to the properties determined from human-generated norms. The results are revealing: Properties generated by SVD are dominated by category coordinates (e.g, cat-dog) and situationally

associated entities (e.g., spoon-bowl). In contrast, the properties generated by StruDEL are more evenly distributed among six different property types.

Although each of these papers has documented some success in using distributional analyses as a way to assess similarity of meaning, there are at least two good reasons to suspect that no matter how large the corpus and no matter how creative and complex the analyses, distributional analyses of similarity will never be completely successful. The first reason is that language use is creative: the meanings of words are shaped by non-linguistic situations and human goals, and those goals and situations are endless. Hence, distributional analyses based on past usage is unlikely to reflect new, creative usage of words. French (1997) provides several interesting examples of the problem. He suggests that one can choose any word or concept X and any word or concept Y and find ways in which they are similar. For example, a credit card is like a) a hotel door key in shape, size, and rigidity; b) a Braille book in that each has raised letters; c) a ruler in that it can be used to draw straight lines; d) an autumn leaf in regard to wind resistance; e) a breeze because it can be used to cool one off when used as a fan; f) fingernails in that it can create annoying sounds when scrapped on a blackboard; and g) a bad friend in that both can get you in trouble. With a bit of thought, one can find similarities between a credit card and a rose, a horse race, or the Spanish Inquisition! Because each of these similarities is a new, creative response to a new goal (in this case, demonstrate how X can be similar to Y), the results from distributional analyses of previous uses of *credit card* will be close to irrelevant.

French's example also illustrates the second reason why distributional analysis are unlikely to reveal much about meaning. Namely, meaning is not inherent in the words, but in the qualities and uses of the objects and events that the words refer to. Thus, the meaning of *credit card* appears to be flexible because of the many qualities and usages of the object, not (only) because the words themselves may be used flexibly.

A linguist interested only in the putative structure of language and not language use might dismiss these examples as fanciful. But the examples should give pause to any one interested in the psychology of meaning, similarity of meaning, and natural language processing. These examples suggest that human meaning can only be ascertained by a system with a human-like body, human-like experiences (e.g., that the sound of fingernails on a blackboard is annoying), and human-like goals.

The aforementioned papers share the goal of using distributional analyses to further automatic, machine-based, natural language processing. Onnis *et al.* (this volume) have a different goal, namely, to identify how distributional statistics might be used in human language comprehension. They note that some words seem to have semantic valence tendencies (SVTs). For example, the verb *cause* is often used in descriptions of negative events (e.g., cause trouble), whereas the verb *provide* is typically positive (e.g., provide work). Their research demonstrates that a) when people are asked to complete sentences using verbs such as these, the completions often reflect the SVTs, b) human reading is slowed by violation of an SVT, and c) a semi-automatic algorithm can successfully extract SVTs from a corpus. Onnis et al. suggest that word co-occurrence statistics "are likely computed by the human brain during the processing of language". Later, we will suggest that the data are more likely to reflect the brain's learning about situations rather than the computation of statistics.

Among the authors included in this special volume, only Sahlgren suggests that distributional representations are equivalent to meaning, as noted in the quote in the first paragraph of this article. His claim, however, is couched within the particular paradigm of structuralist linguistics. As Sahlgren notes, within this framework "linguistic meaning is inherently differential, and not referential (since that would require an extra-linguistic component); it is *differences* of meaning that are mediated by *differences* of distribution" (Sahlgren, emphasis in original). Later he writes, "…the only meanings that exist within a structuralist account of language are the types of relations distributional methods acquire". Sahlgren's analysis appears to be correct by virtue of the following tautology: Structuralist linguistic meaning is only concerned with differences in meaning; words with different distributional properties will have different meanings; hence distributional properties captures all (structuralist) meaning.

Of course, once we become concerned with the psychology of meaning and language use, then other forms of meaning, such as referential meaning, become central. Nonetheless, several theories of psychological (including referential) meaning are built on a foundation of distributional analyses. Consider, for example, that Landauer & Dumais's (1997) Latent Semantic Analysis (LSA) is meant to be a theory of acquisition, induction and representation of knowledge. The LSA mechanism begins by noting the frequency of occurrence of about 60,000 words across some 30,000 texts, forming a matrix with words as rows, texts as columns, and frequencies as the cell

entries. After some pre-processing, the matrix is submitted to a singular value decomposition to reduce the dimensionality and thereby enforce consistency. The result is a matrix in which the words are coded with values on about 300 dimensions (in contrast to the 30,000 texts). Landauer & Dumais (1997:215) state, "… we suppose that word meanings are represented as points (or vectors; later we use angles rather than vectors) in *k* dimensional space…". Later in the paper they write: "Given the strong inductive possibilities inherent in the system of words itself, as the LSA results have shown, the vast majority of referential meaning may well be inferred from experience with words alone" (p. 227). The obvious implication of this claim is that experience with the world is not important for the "vast majority of referential meaning…" Note that all perceptual information is stripped away before coding in LSA in that (a) words are descriptions not the objects themselves, and (b) it is frequencies of words that matter. That is, in LSA frequency (or covariation in frequency after dimensional reduction) is the currency of meaning, not say, redness, or loudness, or spatial extent or the neural coding of redness, loudness, or spatial extent.

Landauer & Dumais (1997:227) do recognize the need for some sort of symbol grounding: "But still, to be more than an abstract system like mathematics words must touch reality at least occasionally". To ground a word such as *rabbit* they suggest, "judiciously add[ing] numerous pictures of scenes with and without rabbits to the context columns in the encyclopedia corpus matrix, and fill[ing] in a handful of appropriate cells in the *rabbit* and *hare* word rows." Nonetheless, the currency remains frequency because all processing in the LSA theory is based solely on the frequencies (or the dimensional values after singular value decomposition), not directly on anything to do with the pictures or perceptual or action systems. For example, using LSA to judge that a rabbit and a hare are similar, even with the addition of picture contents, would not require accessing visual information in the pictures. Instead, the judgment would be based on a mathematical comparison of the similarity of the vectors representing rabbit and hare. Similar claims are made by Burgess and Lund (1997) about Hyperspace Analogue to Language (HAL), another system that tracks covariation amongst words.

There are many demonstrations that covariation, as implemented in LSA and HAL, could serve as a basis of meaning. For example, Landauer & Dumais report that similarity between LSA vectors can be used to pick out synonyms about as well as non-Native English speakers applying for admission to U.S. colleges. Nonetheless, all of

the demonstrations are correlational, and consequently, the causal relations cannot be determined. Consider, for example, the finding that the angle between two LSA vectors predicts the extent to which the corresponding words will prime one another in a lexical decision task. One possibility is that covariation amongst words determines meaning, and hence covariation correlates with lexical decision, a putative measure of relational meaning. Another possibility, however, is that words that share meaning (e.g., *cow* and *bovine*) tend to appear in similar texts because of the overlap in meaning. Thus, the covariation does not determine meaning, it is the shared meaning that determines the covariation.

In contrast to models that depend on covariation for meaning, there are data (e.g., Glenberg & Robertson 2000) and arguments (e.g., Harnad 1990; Searle 1980) suggesting that word meaning requires more than covariation. Consider Harnad's (1990) symbol merry-go-round argument. Imagine someone traveling to another country who does not speak the language, but who has a dictionary written in that language. The traveler sees a sign and wishes to translate it. The traveler looks up the first word in the dictionary, but of course the definition is only in terms of other words in the unknown language. Undaunted, the traveler looks up the first word in the definition, but it too is defined solely in terms of other unknown words. No matter how many words the traveler looks up, that is, no matter how much covariation the person tracks, the traveler will never be able to induce the meaning of even the first word in the sign.

Furthermore, there is strong empirical evidence for the grounding of word meaning in perception and action, not (just) in the word's covariation with other words. For example, Kaschak et al. (2005) demonstrated that watching displays of visual motion affects the comprehension of sentences describing visual motion. Glenberg & Kaschak (2002) demonstrated that understanding a sentence about directional action differentially affects literal action in compatible and incompatible directions. Hauk *et al.* (2004) found that listening to verbs produced enhanced activation in areas of motor cortex corresponding to the effectors used in the actions the verbs named.

In summary, there are reasons to suppose that covariation plays a critical role in determining the meaning of words, but there is equally good evidence to suggest that that role may be limited. Importantly, there do not seem to be any direct tests of the claim that 'vast majority of referential meaning' can arise from covariation alone. That is, the evidence supporting covariation is mainly correlational, and the evidence demonstrating that some words are grounded

in perception and action does not rule out the possibility that some (or even the vast majority of) meaning may be based on covariation. The experiments reported next are designed to fill the evidential gap.

The experiments contrast performance of two conditions, a Learning condition and a Control condition. In the Learning condition, participants are exposed to the coherent covariation of stimuli (radio buttons on a computer interface). We know that the covariation is coherent, because they are taken directly from a random sample of real-life stimuli: 102 examples of two-wheeled vehicles found around campus. Each of the examples was coded on 29 features such as *Is a road bike*, *has a chain*, and *is noisy* (see Table 1). For the participants in the Learning condition, however, these verbal descriptions of the features were replaced during learning with the on or off occurrence of radio buttons (see Figure 1, although the verbal labels were suppressed). Thus, a particular example of a road bike would be presented with particular radio buttons highlighted (but not named); other examples of road bikes would be presented with a similar patter of highlighted radio buttons as determined by their natural co-variation in the sampled population; and other two wheeled vehicles would occur with an overlapping (but less similar) pattern of radio buttons.

The participants were explicitly directed to learn patterns of covariation, and that learning was demonstrated through a series of benchmark learning tests. These benchmark tests demonstrated that participants have derived at least some of the covariation structure. Following the benchmark tests, the participants were given a series of Final meaning tests designed to measure the extent to which meaning can be derived from the covariation structure. For example, participants were asked to choose the domain of the examples from amongst choices such as *two-wheeled vehicles* and *celestial bodies*.

In the Control condition, participants proceeded immediately to the Final meaning tests, and they were told to make their best, educated guesses on the tests. Thus, participants in the control condition had no opportunity to learn the covariance structure of the 102 exemplars before the Final meaning tests.

If the strong covariation claim is correct, then participants in the Learning condition, who have demonstrably learned part of the covariation structure, ought to be more successful on the Final meaning tests than the participants in the Control condition.

Experiment 2 was similar except that the majority of the radio buttons were accompanied by the verbal labels (much as in Figure 1). Given that the majority of the examples and properties were readily identified, this experiment is closer to the cases of both reading and

Table 1. Cross categorization of feature types and relations used in the experiments

		Relations			
		ISA (category)	Parts	Properties	How changes
Feature Types	Abstract (Categories)	Two-wheeled* Motorized Road bike* Mountain bike* Recumbent bike Scooter Motorcycle* Moped			
	Abstract (Features)			Inexpensive*	Can go a short distance Can go a medium distance* Can go a long distance Gets flat tires easily Disassembled easily* Can carry another person
	Visible		Chain visible Mirror* Red light in back Keys needed	Not achromatic* Used on sidewalk*	
	Auditory			Noisy*	
	Haptic (touch)		Smooth tires	Hot Not heavy*	
	Proprioceptive (body position)			Legs still* Sit upright Lean forward* Recumbent	

* Features that were labeled during learning in Experiment 2

perceptual observation in which the majority of the input is readily identifiable. Then, the theoretical question becomes, can people use the learned covariation to enhance their domain knowledge and thereby induce the meaning of the properties (radio buttons) that are not identified.

Experiment 1

Experiment 1 has three goals. The first is to produce clear evidence that people can learn about covariation of unnamed symbols.

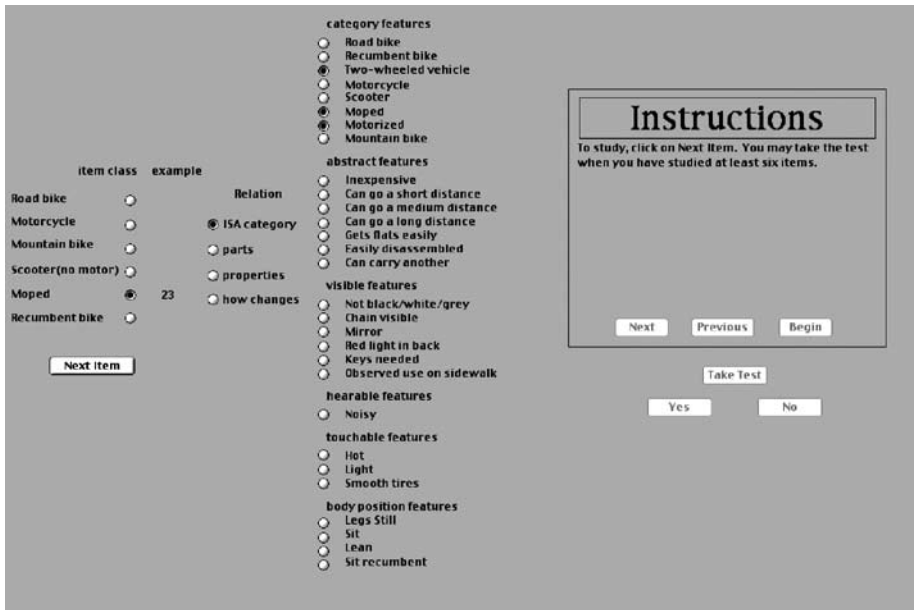

Figure 1. Screen shot of the user interface showing the names of the items and features. The computer program selected the item class of moped and the particular example (23). The participant has selected the ISA category relation and the categories corresponding to a moped are highlighted. Note: In Experiment 1, none of the item class names or feature names was visible during learning; in Experiment 2, four of the item names and 14 of the feature names were visible.

The second goal is to determine if meaning can be inferred from the learned covariation structure. The third goal is to determine whether naming some of the symbols allows people to use the covariation structure to infer the meaning of the other symbols (cf. Goldstone *et al.* 2005).

We began by selecting a semantic domain familiar to the student participants in our experiments, namely, the domain of two-wheeled vehicles. Then, over the course of several weeks we made quasi-random observations at various locations on campus resulting in 102 instances of the domain, including 20 road bicycles (e.g., those with downward curving handlebars), 30 mountain bicycles, 7 recumbent bicycles, 7 leg-powered scooters, 11 motorcycles, and 26 mopeds. We will refer to these five types of vehicles as five item classes (the terminology used for the participants). We coded each example of an item class on 29 binary features that could be cross-classified by mode of observation (e.g., auditory or visual) and type of relation between the

vehicle and the feature. These features and the cross-classification are given in Table 1. The four relations we used are those used by Rumelhart (1990) and Rogers and McClelland (2004) in their simulations of semantic memory.

We constructed a user interface illustrated in Figure 1. Note, however, that in the main part of the experiment, most of the specific verbal labels next to the radio buttons (e.g., *road bike*, *inexpensive*, *noisy*) were not present. However, the interface did label the groups of radio buttons (as in Figure 1) using the labels *Item class*, *Relations*, *Category features*, *Abstract features*, *Visible features*, *Hearable features*, *Touchable features*, and *Body position features*. In addition, the names of the relations (*ISA category*, *parts*, *properties*, and *how changes*) were displayed.

Participants in the Learning condition were specifically instructed to learn which features, that is, which radio buttons, tended to occur together. Learning preceded by having the participant click on the *Next Item* button. The computer would randomly select one of the 102 exemplars, display an example number (e.g., the number corresponding to a particular road bike), and display the features corresponding to that example for one of the relations (e.g, the ISA relation). For example, if the computer chose example 23 from class 5 (moped), and the ISA relation was selected, then the computer would highlight Category radio buttons corresponding to two-wheeled (activated for all items), motorized (activated for motorcycles and mopeds), and moped (activated for all examples of mopeds). If the participant then choose the relation Parts, the computer would highlight visible feature radio buttons 3, 4, and 5 corresponding to *mirror*, *red-light in back*, and *keys needed*, assuming that these features did occur for the specific moped that contributed to example 23. Participants were free to select different relations for the particular example (e.g., example 23 of Moped) to study all of the properties for that example, or they could select the next item (by clicking on the *Next Item* button) at any time. Participants had unlimited time to study.

The learning of at least part of the covariance structure was ensured by having participants study until two types of tests were passed. On each trial of the ISA test, both an item class radio button was highlighted and the ISA category radio button was highlighted. The computer program also highlighted a selection of category radio buttons. Either all of the category radio buttons were correct (that is, the hidden labels described the item class and these radio buttons had been presented with each example from this item class) or one was changed. The participant made a decision as to whether all the

selected categories were correct or not. There were 12 trials on the ISA test, two for each item class, with one trial containing correct categories and the other trial an incorrect category. To pass the ISA test, the participant had to make at least nine correct decisions. If the participant had fewer than nine correct, then he or she was forced to study again before retaking the test.

After meeting the ISA test criterion, the participant was given the Feature test. On each trial of this test, an item class was highlighted as well as all of the correct category radio buttons. The participant would then cycle through the other relations (parts, properties, and changes). For each relation, either the correct features (radio buttons) were highlighted, or one was changed. The participant was to determine if all of the highlighted radio buttons were correct. Again, the criterion was nine correct out of twelve trials. If the criterion was not met, then the participant was forced to study again.

Meeting both the ISA test and Feature test criteria ensures that the participant has learned a good deal about the covariance structure. At this point, a series of five Final tests (Domain, Item 4, Item 5, Change, and Part) were given to determine whether or not the participant could infer meaning from the covariance relations. For the Domain test, the participant was told that he or she a) had been studying a coherent semantic domain, b) was to indicate which of six choices was the domain studied and c) was to give a confidence rating on a four-point scale ranging from guessing to positive. The domain choices consisted of two natural, non-biological categories (celestial objects and geological features), two natural, biological categories (mammals and single-celled organisms), and two artifact categories (furniture and two-wheeled vehicles). The dependent variables were the choice made, confidence in the choice, and the time taken to make the choice. Although we cannot be sure, it is unlikely that the distractor categories have the same covariance structure as two-wheeled vehicles. Thus, if covariance structure allows the induction of meaning on the basis of structure, the choice of two-wheeled vehicles should be obvious.

Following the Domain test, the participants were shown the names of four of the item classes (all but *moped* and *scooter*), as well as the names of 24 of the 29 buttons (not displayed were: *motorized*, *moped*, *scooter*, *easily gets flat tires*, and *keys needed*). At this point, the meaning of the majority of radio buttons that contributed to the covariance structure becomes available. Thus, the question of interest is whether the meanings of those buttons along with the learned covariance structure can be used to determine the meaning of the

remaining, unnamed, buttons. For the Item 4 Final test, participants were asked to pick the name of the fourth item class (*scooter*) from six choices that included three human-powered vehicles (*scooter*, *skateboard*, *tandem bicycle*) and three motorized vehicles (*moped*, *Segway*, *bicycle with motor*). The dependent variables were the probability of selecting from the correct category (human-powered), the confidence, and the time to make the selection. For the Item 5 Final test, participants were asked to pick the name of the fifth item class (*moped*) from the same choices.

For the Change Final test, participants were given the names of all of the item classes, and all of the categories so that only two buttons remained unnamed. The participant was asked to select the name of the missing change button (*easily gets flat tires*) from among three features typical of human powered vehicles (*easily gets flat tires*, *easily falls on ice*, and *easily raise or lower seat*) and three features typical of motorized vehicles (*new registration required yearly*, *oil is added*, and *can run out of gas*). The dependent variables were the probability of selecting a name from the correct category (human-powered), confidence, and time to make the decision.

On the Part Final test, the participants were asked to select the name of the part button (*keys needed*) from among three parts typical of human-powered vehicles (*curved handlebars*, *basket*, *thin spokes*) and three parts typical of motorized vehicles (*keys needed*, *speedometer*, *exhaust pipe*). The dependent variables were the probability of selecting a name from the correct category (motorized), confidence, and time to make the decision.

The performance on these five Final tests for participants in the Learning condition was compared to performance of participants in a Control condition. Initially control participants were treated identically to the participants in the Learning condition (see procedure). However, instead of studying the covariance structure, these participants were given special instructions and then proceeded immediately to the Final tests. The special instructions were:

> You have been chosen to participate in a special version of this experiment. Other students study the item classes and features. They then take the same Final test that you will take shortly. As a special participant, your job is to help us to estimate the probability of educated guessing on the Final test. Because, you have not studied anything, your answers will have to be educated guesses. In what sense might your answers be 'educated?' There are several clues that you might use in making your guesses. For example: You are a college student. You are in an experiment. You know that the

> item classes are not trees or fish [These two classes were examples used in the instructions that all participants received, and all participants knew that these were not the semantic domain studied.] From these hints and others, you may be able to deduce the correct answers, or at least eliminate some of the obviously wrong answers. Why should you bother? Why not just respond randomly? There are five questions on the final test. For each one you answer correctly, you will earn a bonus of $1. If you get all five correct, you will earn an additional bonus of $5. We hope that this money will motivate you to make educated guesses rather than random guesses.

If learning the covariance structure of a set of stimuli a) can be used to infer the semantic domain or b) can be used in conjunction with named features (e.g., the 27 of 29 features names provided after the Domain test), then the participants in the Learning condition should do substantially better than those in the Control condition.

Method

The 91 participants were students enrolled in introductory psychology classes at the University of Wisconsin – Madison. Of these, 33 participants were randomly assigned to the Control condition. A total of 58 participants were assigned to the Learning condition, and of this total, 43 completed the experiment. The remaining 13 elected to discontinue the experiment after repeated failures on the ISA or Feature tests. Only data from the 76 participants who completed the experiment are reported.

Procedure

Participants in both the Learning and Control conditions were treated identically at first. The computer-controlled instructions introduced the participants to the interface using three, named item classes: pine tree, maple tree, and trout. The 29 features were selected to fit these items (e.g., has leaves, grows, can be eaten). The instructions explained how features could be cross-classified by "how a person comes to know a feature" such as through vision and touch, and "how they (the features) relate to the item" such as categories and parts. The participants were encouraged to observe how some features tend to occur together, such as "living thing" (a category feature) and "grows" (a change feature). They were also encouraged to attend to patterns of covariation rather than specific examples "It would be very difficult to learn all of the features for all of the examples in this experiment, so your goal should be to develop some general ideas, such as all living pine trees have needles, all living maple trees have leaves, but not needles." At this point, the names of the example item

classes and features were erased from the computer screen, and the participants were instructed:

> Of course, you already know a lot about pine trees, maple trees, and fish. So, in this experiment you are not going to know the names of the items or the names of their features! (If you want to see the example features again, use the Previous button.) Instead, your goal is to learn things such as: For Item class 1, ISA categories are 1 & 2, and features 14 & 20 go together. For Item class 2, ISA categories are 1 & 3, and features 17 & 18 go together.

The participants were encouraged to explore the interface by clicking on the *Next Item* button and the relations buttons and observing the corresponding features. The participants were instructed further that the item classes they were studying were not fish or trees and that all of the names of the buttons differed from those observed previously. To begin the formal study, the participant clicked on a button labeled *Begin*.

For participants in the Control condition, on clicking the Begin button, they received further instructions (those quoted above) and then moved immediately to the Final tests.

For participants in the Learning condition, on clicking the Begin button, they could begin studying the items and relations. After studying at least six examples, a participant was given the option of further study or of proceeding to the ISA test. There were 12 trials on the ISA test, two for each item class. One trial for the item class was correct (all of the categories were correct) and one trial for the item class was incorrect (one of the category features was incorrect). If the participant did not get at least 75% correct on the ISA test, he or she was directed to restudy the items. He or she had to study at least six examples before being able to re-take the ISA test. This procedure continued until the ISA test criterion was met, or the participant asked to be excused from the experiment.

Once the ISA test criterion was met, the participant was given the Feature test. There were 12 trials on the Feature test, two for each item class, and one trial was correct and the other incorrect. On incorrect trials, the computer randomly picked a relation to modify (e.g., part relation) and randomly chose a radio button of that relation type (e.g., another part) to substitute for the correct radio button. If the participant did not get at least 75% correct on the Feature test, he or she was directed to restudy the items. He or she had to study at least six examples before being able to re-take the test. This procedure continued until the Feature test criterion was met, or the participant asked to be excused from the experiment.

After meeting the Feature test criterion, the participant was given the series of five Final tests as previously described. After the Domain test, the domain was revealed along with the names of four of the six item classes and 24 of the 29 features. The participant was then asked to select the names of the remaining two item classes (Item 4 and Item 5 Final tests) and the remaining two features (Change Feature Final test and Part Feature Final test). Although the items tests always preceded the feature tests, within each type of test the order of testing (e.g., the order of Items 4 and 5) was randomly determined for each participant.

Results

The extent of study, as well as performance on the ISA and Feature tests, are indices of the degree to which participants in the Learning condition mastered the covariance structure of the stimuli. On average, participants clicked 139 (SD = 113) times on the Next Item button and 143 (127) times on different relations. On average, participants spent 32.4 minutes (11.5 min) studying (excluding time spent on instructions and the Final tests). Participants needed an average of 3.00 (2.5) attempts to meet the ISA test criterion, and the average number correct on the last ISA test was 9.93 (.91) out of 12. Participants needed an average of 8.05 (7.62) attempts to meet the Feature test criterion, and the average number correct on the last test was 9.49 (.80). These data testify to an impressive amount of study and accomplishment.

Clearly, people can learn something like the covariance structure of a set of ungrounded stimuli, the radio buttons. Nonetheless, the question remains as to whether that learning can be a source of meaning. To answer that question, we turn to the data from the Final tests. Two caveats are required, however. First, the data from the Learning condition can only be interpreted in relation to the data from the Control condition. The reason is that it is unlikely that the six choices offered on the Final tests are all equally attractive. Thus, the baseline or guessing rate can only be determined from the Control condition. The second caveat is that it seems unreasonable to expect the participants in the Learning Condition to be able to infer specific item classes and features from the covariance structure. Consequently, except for the Domain test, we scored whether or not the participant selected a choice within the appropriate class of motorized or non-motorized vehicles.

The results from the various Final tests are given in Table 2. For the Domain test, can the participants in the Learning condition select

Table 2. Means (and SDs) from Experiment 1 Final tests

	Domain test		Item tests		Feature tests	
	L	C	L	C	L	C
Correct	.05 (.21)	.12 (.33)	.50 (.38)	.47 (.39)	.59 (.35)	.53 (.30)
Confidence	2.00 (.90)	2.03 (1.13)	1.92 (.74)	2.04 (.71)	2.12 (.76)	2.24 (.75)
Time (sec)	35.47	44.85	33.96	43.70	32.39	51.37

L = Learning condition; C = Control condition

the correct domain (two-wheeled vehicles) more accurately than participants in the Control condition? In fact, just the opposite occurred, although the difference was not statistically significant, $\chi^2(1) = 1.43$. There was not a significant differences in the confidence of the choices, $F(1, 74) = .02$.

Because of the positive skew in the selection times, the data were transformed into logarithms before the statistical analyses. The data reported in Table 2 are the means of the logarithms transformed back into seconds. The difference between the mean selection times was not significant, $F(1,74) = 2.74$. Nonetheless, the times are interesting for several reasons. First, in both conditions, the participants took a considerable amount of time to make the choice; apparently, they treated the task seriously. Second, the times provide little or no evidence that the participants in the Learning condition were able to solve the task by simply referring to a semantic memory. Instead, judging from the length of time required, it would appear that participants in both conditions were applying some sort of reasoning strategy.

For each participant, we determined if he or she chose an item from the correct category (motorized or not) for Items 4 and 5, and then we obtained the proportion correct out of 2. The difference between the conditions (see Table 2) was not significant, $F(1, 74) = .12$, nor was the difference between the confidence ratings significant, $F(1, 74) = .57$. However, the participants in the Learning condition took less time to make their choices than did the participants in the Control condition, $F(1, 74) = 4.06$. Once again, the times are quite informative. The long durations are more consistent with some sort of reasoning strategy rather than direct consultation of semantic memory.

The last two tests involved identifying the Change relation radio button *easily gets a flat tire* and the Parts relation radio button *keys needed*. For each participant, we determined if he or she chose an item from the correct category (motorized or not) for the two tests, and then we obtained the proportion correct out of 2. There were no significant differences in the proportions correct or confidences, Fs

< 1. As with the Item tests, participants in the Learning condition tended to make their choices faster, $F(1, 74) = 10.24$.

Discussion

The data are clear in three respects. First, people can learn at least part of the covariance structure of ungrounded symbols. This is demonstrated by performance on the ISA and Features tests. Second, that covariance structure cannot be used to determine the domain of study; participants in the Learning condition were no more accurate on the Domain test than participants in the Control condition. Third, even after most of the radio buttons were named, people cannot easily use the covariance structure to determine even a coarse categorization (motorized or not) for the unnamed buttons.

Experiment 2

In Experiment 1, the names were revealed only after the covariance structure was learned. Perhaps the covariance structure is more useful when some of the items in the structure are named while learning takes place. In that way, an already known meaning structure can be updated on the basis of new learning. Thus, in Experiment 2, three of the item classes (road bike, mountain bike and motorcycle) were named throughout the study period, as were 14 of the 29 features (see Table 1). Thus, if learning the covariance structure is useful when the names of some of the radio buttons are known, participants in the Learning condition should demonstrate a clear superiority on the Final tests compared to participants in the Control condition. In addition, we might expect the selection times to be much shorter for the Learning condition.

Experiment 2 addresses many of the criticisms that could be leveled at the methodology of Experiment 1. One such concern is that radio buttons are an odd representational medium. Perhaps one needs to develop some skill in the representational medium in order to attend to and use the covariance structure. Naming half of the radio buttons makes the task closely analogous to a reading task in which some of the words are unknown. In the following example, each sentence begins with the (always available) name of a relation or name of a group of radio buttons. The sentence continues with names of the labeled radio buttons and ends with the names of the unlabled buttons. For a typical moped (not named in Experiment 2), reading the display results in:

Category Features: is a two-wheeled vehicle; is not a Road bike; is not a motorcycle; is A (A corresponds to "motorized," an unnamed

radio button, in Figure 1); is not B; is not C; is not D; is E. Abstract Features: is expensive; can travel a medium distance; cannot be easily disassemble; can F; can G; cannot H; can I. Visible features: has a mirror; is colored; is not used on a sidewalk, does not have J; does have K; does have L. Audible features: is noisy. Touch features: is heavy; does not have M; is N. Body Part features: legs are still; do not lean forward; has O; not P.

Thus, the input is substantially similar to what is given the LSA program. By tracking the covariation and computing dimensional reduction, the LSA program can determine similarities (that is similar co-occurrence structures) between unnamed features such as A, B, and C and named features such as *road bike*. The question is: Can people do the same?

Finally, one of the supposed benefits of a theory such as LSA is that it gives rise to covert concepts based on previously encountered covariation structure (and dimensional reduction). For example, according to a theory such as LSA, it is possible to infer that mopeds do not easily get flat tires (unnamed abstract Change feature H in the above listing), based on a moped's similarity to motorcycles and cars which do not get flat tires easily. Given that half of the radio buttons are named during learning in Experiment 2, if the strong covariation claim is correct, it should be much easier for the participants in the Learning condition to map the covariation of H (it occurs with mopeds and motorcycles but not with road bikes and mountain bikes) with *easily gets a flat tire* compared to participants in the Control condition.

Method

A total of 93 participants were randomly assigned to the Learning (n = 61) and Control (n = 32) conditions. A total of 21 participants in the Learning condition failed to finish the experiment. Thus, the data are based on 40 participants in the Learning condition and 32 participants in the Control condition. The method was identical to Experiment 1 except as noted above and except that the Domain Final test was eliminated. Because the names of three of the item classes were given, the choice of domain (two-wheeled vehicles) would be obvious.

Results

On average, Learning condition participants clicked 81.8 (SD = 78.7) times on the Next Item button, 105.2 (70.2) times on different relations, and studied for 28.32 minutes (10.28). Participants needed

an average of 2.00 (1.55) attempts to meet the ISA test criterion, and the average number correct on the last test was 10.10 (1.01) out of 12. Participants needed an average of 4.60 (3.06) attempts to meet the Feature test criterion, and the average number correct on the last test was 9.32 (.57). As in Experiment 1, there is clear evidence that the participants in this condition have learned at least some of the covariance structure.

The data from the Final tests are presented in Table 3, and in many respects, they mirror the data from Experiment 1. For the combined Final tests on Items 4 and 5, there were no significant differences in proportion correct or confidence, Fs (1, 70) < 1. Although the participants in the Control condition did take longer to make a decision, $F(1, 70) = 5.30$, it should be noted that the participants in the Learning condition took a considerable amount of time. Similarly, for the combined Final tests on features and parts, there was not a significant difference for proportion correct, $F<1$, or for confidence, $F(1, 70) = 1.47$, but the participants in the Control condition took longer to make a selection, $F(1, 70) = 16.95$. Apparently, even when about half of the radio buttons are named throughout the learning of the covariance structure, that structure is of little use in classifying whether the other radio buttons correspond to human-powered or motorized vehicles.

Table 3. Means (and SDs) from Experiment 2 Final tests

	Item tests		Feature tests	
	L	C	L	C
Correct	.59 (.39)	.62 (.31)	.64 (.32)	.59 (.37)
Confidence	2.12 (.74)	1.95 (.78)	2.01 (.76)	2.23 (.79)
Time (sec)	34.90	46.80	31.81	51.64

L = Learning condition; C = Control condition

General Discussion

The experiments were designed to answer three questions: Can people learn the covariance structure of ungrounded symbols? Can people derive meaning from that covariance structure? When part of the covariance structure is named, can people use the covariance structure to infer meaning (in this case, classification as human-powered or motorized) for the unnamed radio buttons? We will consider the answer to these questions, and then discuss two more: If covariation plays a limited role in acquiring new meaning, how are those

meanings acquired? And, what accounts for the impressive relations between covariation and meaning that have been documented?

If we can trust the ISA and Feature tests, then the data from the experiments provide a clear answer to the first question: People can learn at least a portion of the covariance structure of ungrounded symbols. Should the ISA and Feature tests be trusted? Not if the tests could have been passed by chance. Each test had 12 two-choice questions. With the probability of answering correctly by chance equaling.5, the probability of getting at least 9 correct is.073. On average, participants took about two tries to pass the ISA test. The probability of getting at least 9 correct (by guessing) on either the first or second test is only.139. Participants needed about five attempts to pass the Feature test. If they were guessing, the probability of passing the test on any of five tries is still only.312. Thus, given that people succeeded in passing these tests, it seems clear that they have learned something about the relations among the radio buttons.

The answer to the second question also seems clear: People were unable to map the covariance structure of the unnamed symbols (the radio buttons) to the correct general domain (two-wheeled vehicles). Nonetheless, there are several constraints on this conclusion. First, we used only one domain. Perhaps it has a covariance structure that is difficult to map onto semantic memory structures, or we picked features that are not features normally attended.

The experiments do seem to meet the conditions needed to test Landauer & Dumais (1997). Namely, words in texts are to a computer program no different from a collection of radio buttons (abstract symbols that are presented visually), and the operation of the theory is to compute the covariances (and dimensional reduction) among those symbols. The theory would work exactly the same if the order of the words in each text used as input to LSA were randomized, or if each word was replaced by a random number (as long as the same random number was used each time that word was presented in any text). In fact, the theory would work exactly the same if it were given a coding representing the spatial location of each radio button in the user interface of the two experiments. That is, the LSA model would note the frequency that each radio button occurs with each example (treating an example as equivalent to a text), and the singular value decomposition could be applied to the resulting matrix.

Experiment 2 provides an even stronger test of the LSA-type of mechanism. In that experiment, much of the information was presented verbally (the named radio buttons) that could be read much like a paragraph of the sort that goes into establishing the matrixes

used in LSA (see the sample in the Introduction to Experiment 2). Nonetheless, on the Final tests participants in the Learning condition were unable to assign the correct category (e.g., human-powered) to the radio buttons.

Two arguments against accepting these results as tests of the strong covariance claim remain. One could argue that the covariance structure formed from the 102 examples used in the experiments is unlikely to match the covariance structure for a particular participant. At the very least, that participant may own a two-wheeled vehicle and the vast amount of experience with that vehicle will perturb the structure. However, if this argument were to be made, theories such as LSA could never work: The probability that the covariance structure derived from words in texts would match the particulars of an individual's experience would be infinitesimally small.

The second argument is that much of the learning accomplished by LSA and people is implicit. In these experiments, participants were explicitly told to learn covariance relations. Indeed, it is probably worthwhile devising an implicit form of the experiments. However, even if the results were quite different and in line with the strong covariance claim, the results of the current experiments would be telling. That is, they would demonstrate that separate acquisition mechanisms and data structures are required for implicit and explicit learning via covariation.

The answer to the third question (can covariance be used when some of the symbols are named) is also clear. In Experiment 1, the majority of the symbols were named after the Final Domain test. Nonetheless, participants in the Learning condition could not use that knowledge in conjunction with covariance knowledge to identify (or even grossly classify) the remaining radio buttons. In Experiment 2, about half of the symbols were named during learning. Nonetheless, participants were unable to use the learned covariance structure to classify the remaining symbols.

If covariance structure is not the source of concept acquisition and demonstrable fast mapping of linguistic structures to concepts (e.g., Casenhiser & Goldberg 2005), what else might be needed? To expand on an example from Landauer & Dumais, if we already know a lot about rabbits, how is it that we can then learn a lot about hares from simply being told that hares are like large rabbits with elongated hind legs or even by encountering the word "hare" in the same linguistic contexts as the word "rabbit?" Our proposal is that it is not the covariation that gives the meaning, but that the covariation provides the opportunity for creating embodied representations of the objects

and events that do carry meaning. For example, when told that a hare is like a rabbit with long back legs, we use embodied knowledge about rabbits, such as their size, shape, and haptic qualities to simulate or create a likely representation of hares. [cf., Barsalou's (1999) idea of creating a simulator.] From this simulation, it is possible to derive affordances or inferences, such as the fact that hares run fast given the long hind legs. Consistent with this sketch, Smith (2005) has demonstrated that toddlers extend the name of an object to other objects that the toddler can manipulate in similar ways. In other words, it is not simply the covariation of hares and long hind legs that matters. Instead, what matters is how those hind legs affect perception systems and how those hind legs interact with action systems. This type of information is not inherent in covariation structure.

On the other hand, it is worth reiterating that embodied accounts of meaning value covariation because it is the covariation that provides the opportunity for comparison and construction of embodied simulations. Also, it is covariation that allows for the sort of learning identified by Onnis et al. (this volume). When the verb *to cause* is consistently paired with negative situations, then the stage is set for learning an association between the verb and the emotional reaction generated by directly experiencing the negative situation, observing it, or simulating the experience during language comprehension. Thus, findings by Onnis et al, may reflect previous associative learning between the verb *to cause* and an experienced emotion rather than the on-line computation and use of statistics. (See Havas *et al.* 2007, for a demonstration of how the manipulation of emotional experience can affect language processing.) Thus from an embodied perspective, covariation is very useful, even if covariation alone has little to do with meaning.

If covariation by itself does not play a major role in determining meaning, what accounts for the impressive relations between aspects of covariation and meaning? We think that there are two likely answers. First, covariation does provide opportunities for learning, such as the opportunities for creating embodied mental models of hares from models of rabbits. Second, as we suggested earlier, the relation between covariation and meaning may arise from a reversal of the causal arrow (Zwaan & Madden 2005). For example, Landauer & Dumais, propose that the meaning of Word A is determined by its covariation with Words B, C and so on. In contrast, we think that Word A happens to occur with Words B, C and so on because those words are useful in describing the actions and events to which Word A refers. That is, meaning causes word covariance structure, not the other way around.

Address of the Author:
Arthur Glenberg, Department of Psychology, Arizona State University, 950 S. McAllister, Tempe, AZ 85287
<glenberg@asu.edu>

Author Note

This research was supported by NSF grants BCS-0315434 and INT-0233175 to Arthur Glenberg. The opinions expressed in this paper are those of the authors and do not necessarily reflect those of the funding agency. We thank Lawrence Angrave, Megan Brown, Jerry Federspiel, David Havas, Emily Mouilso, Michal Riscall, and Bryan Webster for contributing to data collection and discussion of the issues raised in this article. Correspondence should be directed to: glenberg@asu.edu.

Bibliographical References

Barsalou Lawrence W. 1999. Perceptual symbols systems. *Behavioral and Brain Sciences* 22. 577-609.

Baroni Marco & Alessandro Lenci (this volume). Concepts and properties in word spaces.

Burgess Curt & Kevin Lund 1997. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes* 12. 177-210.

Casenhiser Devin & Adele E. Goldberg 2005. Fast mapping between a phrasal form an meaning. *Developmental Science* 8. 500-508.

Fazly Afsaneh & Suzanne Stevenson (this volume). A distributional account of the semantics of multiword expressions.

French Robert M. 1997. When coffee cups are like old elephants or why representation modules don't make sense. In Riegler Alexander & Markus F. Peschl (eds.). *Proceedings of the 1997 International Conference on New Trends in Cognitive Science*. Austrian Society for Cognitive Science. 158-163.

Glenberg Arthur M. & Michael P. Kaschak 2002. Grounding language in action. *Psychonomic Bulletin and Review* 9. 558-565.

Glenberg Arthur M. & David A. Robertson 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language* 43. 379-401.

Goldstone Robert L., Ying Feng & Brian J. Rogosky 2005. Connecting to the world and each other. In Pecher & Zwaan 2005. 292-314.

Harnad Stevan 1990. The symbol grounding problem. *Physica D* 42. 335-346.

Havas David A., Arthur M. Glenberg & Mike Rinck 2007. Emotion simulation during language comprehension. *Psychonomic Bulletin & Review* 14. 436-441.

HAUK Olaf, Ingrid JOHNSRUDE & Friedemann PULVERMULLER 2004. Somatopic representation of action words in human motor and premotor cortex. *Neuron* 41. 301-307.

KASCHAK Michael P., Carol J. MADDEN, David J. THERRIAULT, Richard H. YAXLEY, Mark AVEYARD, Adrienne BLANCHARD & Rolf A. ZWAAN 2005. Perception of Motion Affects Language Processing. *Cognition* 94 (3). B79-B89.

LANDAUER Thomas K. & Susan T. DUMAIS 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104. 211- 240.

ONNIS Luca, Thomas A. FARMER, Marco BARONI, Morten H. CHRISTIANSEN & Michael J. SPIVEY (this volume). Generalizable distributional regularities aid fluent language processing: The case of semantic valence tendencies.

PECHER Diane & Rolf A. ZWAAN (eds.) 2005. *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge, UK: Cambridge University Press.

ROGERS Timothy T. & Jay L. MCCLELLAND 2004. *Semantic Cognition*. Cambridge, MA: The MIT Press.

RUMELHART David E. 1990. Brain style computation: Learning and generalization. In ZORNETZER Steven F., Joel L. DAVIS, Clifford LAU & Thomas M. MCKENNA (eds.) 1990. *An introduction to neural and electronic networks*. San Diego, CA: Academic Press. 405-420.

RUMSHISKY Anna (this volume). Resolving polysemy in verbs: Contextualized distributional approach to argument semantics.

SAHLGREN Magnus (this volume). The distributional hypothesis.

SCHULTE IM WALDE Sabine & Alissa MELINGER (this volume). An in-depth look into the co-occurrence distribution of semantic associates.

SEARLE John R. 1980. Minds, brains and programs. *Behavioral & Brain Sciences* 3. 417-424.

SMITH Linda B. 2005. Action alters shape categories. *Cognitive Science* 29. 665-679.

ZWAAN Rolf A. & Carol J. MADDEN 2005. Embodied sentence comprehension. In Pecher & Zwaan 2005. 224-245.