CHAPTER SEVEN

What Do Words Do? Toward a Theory of Language-Augmented Thought

Gary Lupyan

Contents

Abstract

Much of human communication involves language—a system of communication qualitatively different from those used by other animals. In this chapter, I focus on a fundamental property of language: referring to objects with labels (e.g., using the word "chair" to refer to a chair). What consequences does such labeling have on cognitive and perceptual processes? I review evidence indicating that verbal labels do not simply point or refer to nonlinguistic concepts, but rather actively modulate object representations that are brought on-line during "nonverbal" tasks. Using words to refer to concrete objects affects the learning of new categories, memory for and reasoning about familiar object categories, and even basic visual processing. Object representations activated

Psychology of Learning and Motivation, Volume 57
ISSN 0079-7421, DOI: 10.1016/B978-0-12-394293-7.00007-8

by verbal means appear to be different, and specifically, more categorical, than ostensibly the same object representations activated by nonverbal means. A connectionist model of "language augmented thought" provides a computational account of how labels may augment cognitive and perceptual processing.

1. Introduction

Much of human communication involves language. One of the fundamental ways in which language differs from non-human communication systems is in its use of words which, in spoken language, take the form of largely arbitrary sequences of sounds that denote external entities (Burling, 1993; Deacon, 1997; Hockett, 1966; Hurford, 2004). Attempts to understand how an essentially unlimited array of meanings can be communicated using finite ordered sequences of sounds has spawned disciplines from information theory to psycholinguistics, to pragmatics. Yet, a central question concerning this fundamental property of natural language has received relatively little attention: What are the cognitive *consequences* of naming? To what degree is normal human cognition actually language augmented cognition? (cf. Clark, 1998, 2006).

In this chapter I make three main claims:

(1) Verbal labels change (modulate) "nonlinguistic" representations.
(2) These effects run deep; language affects basic visual processing.
(3) Verbal labels appear to be "special." More precisely, concepts activated via labels appear to be different from ostensibly the same concepts activated by nonverbal means.

I will argue that this association of "nonverbal" representations with verbal labels results in conceptual representations that are under pervasive on-line influence by language. The same stimulus thus comes to be represented differently depending on the degree of linguistic influence.

These claims are also logically distinct from issues concerning the *format* of conceptual representations. It is also separate from the rather vague question of whether we "think in words" (see Boroditsky, 2010; Carruthers, 2002 for discussion).

The issue addressed here is not whether certain thoughts are "unthinkable" without language, but whether language augments our concepts in a systematic way. To illustrate the distinction, consider the following thought experiment:

In 1770 Captain James Cook (Lieutenant Cook at the time) landed in what is now Cooktown, Queensland, Australia. During his stay, he encounters a kangaroo, an animal he is unfamiliar with, and one for which he has no name. Certainly we can agree that Cook possessed the ability to

have "thoughts" about this new animal, as expressed by Devitt and Sterelny (1987, p. 219), "Captain Cook had thoughts about kangaroos without having any word for them simply on the strength of observing them". Curious about what this strange animal is called, Cook inquires about its name and is told by a Guuguu Yimidhirr-speaking local that the name is "ganguroo." Meanwhile, the scientist on board the *Endeavor*, (Sir) Joseph Banks, is otherwise occupied and remains ignorant of this name. Imagine further that Banks and Cook proceed to have an identical set of observations of kangaroos. Both individuals observe that kangaroos chew their cud and have ritualized fights, both get a sense for the typical length of their leaps, the color of their fur, and their odd gestational apparatus. Cook's observations comprise the perceptual data which is accompanied by a self-generated label. That is, Cook's thoughts become "indexed" by the category name while Banks's are products of observational experiences alone.[1] Does this produce a difference in the two men's cognitive and perhaps even perceptual processes?

The idea that words, and language more broadly, matter for our thoughts has been, of course, addressed by what has come to be known as the Sapir-Whorf hypothesis (Whorf, 1956; see Boroditsky, 2010; Wolff & Holmes, 2011 for some contemporary reviews, and Lee, 1996 for a deeper insight into Whorf's own writings). Much of the work in the domain of language and thought has drawn a sharp distinction between language and "thought" as done in most contemporary writing about the subject (e.g., Bloom & Keil, 2001; Gleitman & Papafragou, 2005) or conflating the two (e.g., Carruthers, 2002; Pinker, 1994; see Levinson, 1997 for discussion).

On the present position, our mental representations are to varying degrees under continuous influence of language and performance on nonverbal tasks such as categorization, visual memory, object recognition, and even simply detecting the presence of a visual stimulus is to varying degrees augmented by language. In the sections below I summarize a program of study that has attempted to understand the role language plays in cognition and perception by manipulating linguistic variables and observing the effects of these manipulations on putatively nonverbal tasks. The logic is that insofar as normal performance on these tasks is affected by aspects of language, manipulating linguistic variables should manipulate aspects of performance on the task. Thus, although there is nothing "verbal" in observing jumping kangaroos, insofar as language is

[1] It is conceivable that actual events may have mirrored this description to some degree. In his chapter "A Last Look at Cook's Guuguu Yimidhirr Word List," John Haviland (1974) lists "ganguroo" in Cook's but not Banks's word-list. Banks's list, on the other hand, contains the Guuguu Yimidhirr word for "nipple" which is missing in Cook's. On the other hand, Banks is generally credited with introducing "kangaroo" into English and his field notes on the animals do use this term (Cilento, 1971).

co-active during these observations, the representations produced by them may be systematically affected.

The idea that language and thought are intertwined and mutually reinforcing is certainly not new. The Hebrew scholar (and Noam Chomsky's father), William Chomsky wrote:

> Language is not merely a means of expression and communication; it is an instrument of experiencing, thinking, and feeling… Our ideas and experiences are not independent of language; they are all integral parts of the same pattern, the warp and woof of the same texture. (Chomsky, 1957, p. 3).

The German neurologist Kurt Goldstein speculated that the reason aphasic patients he was examining suffered from problems on ostensibly nonverbal tasks is that:

> Language is not only a means to communicate thinking; it is also a means to support it, to fixate it. Defect in language may thus damage thinking. (Goldstein, 1948, p. 115).

And Benjamin Lee Whorf himself in some of his writings took a strikingly connectionist position advocating for continuous interaction between language and thought:

> Any activations [of the] processes and linkages [which constitute] the structure of a particular language… once incorporated into the brain [are] all linguistic patterning operations, and all entitled to be called thinking (Whorf, 1937 p. 57—58 cited in Lee, 1996 p. 54).

Our conceptual content is derived from a multiplicity of sources: direct experience, observational learning, inference and deduction, formal instruction, innate biases, etc. and Banks could certainly make observations of kangaroos without having a simple way to refer to them linguistically. But Banks, like all of us, lives in a linguistic world in which our experiences are co-mingled with linguistic referents. Even in cases when we lack a single word for some entity, we can describe it verbally through circumlocutions.

Banks could presumably organize his field notes by referring to the "curious jumping beast". And so the central question addressed by the chapter is: what do words do? Apart from making linguistic communication possible (no small feat), do words augment our conceptual representations and perhaps even our perceptual processing, and if so, how?

2. From Labeling Our Concepts to Language Augmented Cognition

2.1. Labeling Our Concepts

The question of whether verbal labels (and language more broadly) affect cognition rests on a set of basic assumptions about the relationship between words and concepts and before proceeding it is useful to make these assumptions explicit. On one view, words basically label our concepts (e.g., Fodor, 1975; Pinker, 1994; e.g., Snedeker & Gleitman, 2004). This position is succinctly summarized by Li and Gleitman (2002):

> It is possible to suppose that these linguistic categories and structures are more-or-less straightforward mappings from a preexisting conceptual space, programmed into our biological nature: humans invent words that label their concepts (p. 266).[2]

This view does not preclude language from having an effect on cognition, via, for example, helping to bind different concepts (e.g., Hermer-Vazquez, Spelke, & Katsnelson, 1999) but, importantly, the strict separation and unidirectional relationship between verbal representations and conceptual "nonverbal" representations that characterizes this position means that the hypothesis that labeling affects concepts is actually ill-defined. If concepts are by definition non-verbal — not linguistic — what would it even mean that they are changed or even "affected" by language?

A schematic view of this position, sometimes referred to as the "cognitive-priority" hypothesis, is shown in Figure 1A. In the top panel we see an exemplar of a familiar and meaningful category, tree. Multiple perceptual exemplars map onto a common representation such that different trees are recognized as members of the same class. The concept is further mapped onto a lexical entry—the word "tree"—that enables a speaker to activate the same concept in a listener using the label (assuming that the listener's representation of "tree" is also mapped onto the tree concept). The bottom panel illustrates a parallel situation with a less familiar category (with which we have correspondingly less perceptual experience): a tram pantograph. This is a device used to connect the tram's motors to the overhead electrical cables. Although many people

[2] Li and Gleitman (2002) further state that "This perspective would begin to account for the fact that the grammars and lexicons of all languages are broadly similar, despite historical isolation and cultural disparities among their users." The idea that languages are broadly similar in their grammars and lexicons, popularized by generative linguistics, is hardly an assumption we should accept without question, and crumbles considerably when placed under scrutiny (e.g., Evans & Levinson, 2009).

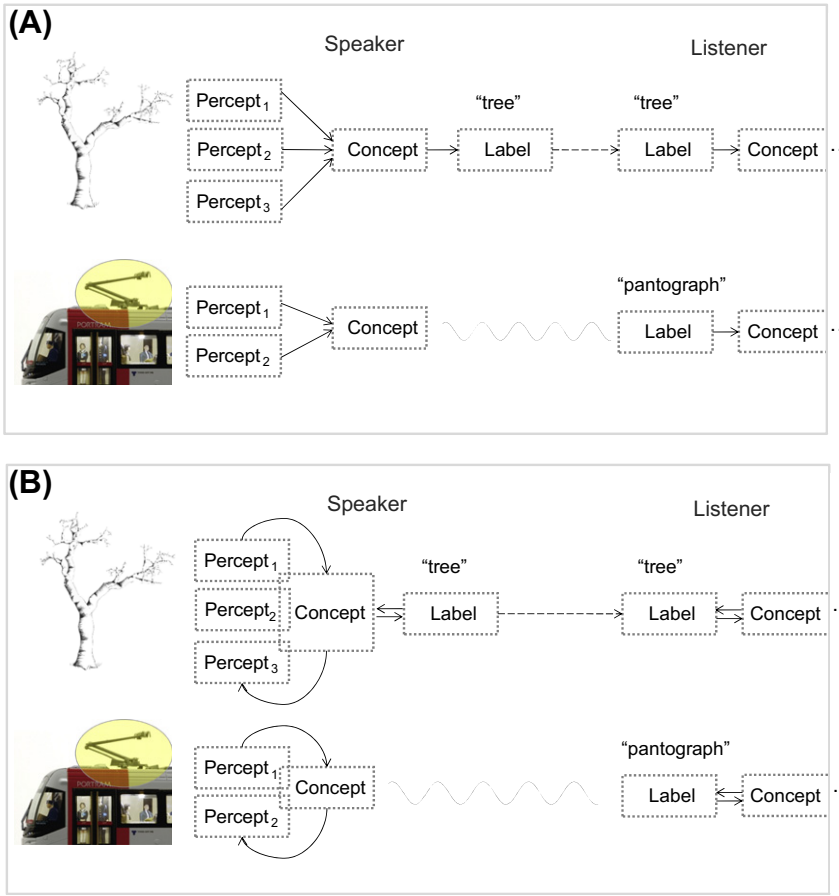

Figure 1 (A) A schematic view of the standard account in which words label our concepts (e.g., Gleitman & Papafragou, 2005). See text for details. (B): A schematic view of language augmented thought. All representational layers are recurrently connected. The overlap between perceptual and conceptual layers indicate the difficulty in drawing sharp distinctions between different types of representations in interactive frameworks. (For color version of this figure, the reader is referred to the web version of this book.)

have seen pantographs and have a rough idea of their function, few know what the device is called. On the standard account, the lack of a name makes communicating just what one means more difficult (the squiggly line indicates the likely need for circumlocutions and definitions of the kind I just used). Assuming the definition is sufficiently precise, it activates the pantograph concept in the listener. Critically, whether the listener knows the word "pantograph" does not affect their conceptual (or perceptual) representations. That is, the speaker and listener could have precisely the same concepts of pantographs except that, as shown in

Figure 1, the listener happens to have the label "pantograph" mapped onto her concept. On this view, the label is simply a reporting device and its role is limited to communication. To illustrate by analogy: verbal expressions are like a Caps-Lock indicator on a keyboard. Lacking such an indicator makes it more difficult to assess whether the computer is in Caps-Lock mode, but has nothing to do with the computer's ability to enter it. Similarly, whether or not we have a word for something has no effect on our ability to "have" a concept. In fact, on this view it is unclear how one can ever "acquire a concept that one could not antecedently entertain" (Gleitman & Papafragou, 2005, p. 634). This position, espoused notably by Fodor (e.g., 1975), and referred to by Gleitman and Papafragou as "the venerable view" (p. 634), is very difficult to reconcile with the extant empirical evidence on concept learning.[3]

2.2. Language Augmented Thought

An alternative position is schematized in Figure 1B. The information flow between all the layers is bidirectional. The label is not simply a means of accessing a concept. Rather, its activation affects the representation of the concept itself. The bidirectional information flow between the concept and the perceptual representations means that the label can indirectly affect even perception itself (see Section 5). The consequence of this bidirectional information flow is that the label (e.g., "pantograph") is not something the concept simply maps onto. Its activation can change the nature of the concept itself. Thus, the concept of a pantograph associated with a verbal label may be systematically different than the ostensibly same concept not associated with a label. Moreover, the representation of even highly familiar concepts like tree may be augmented, *on-line*, as the label "tree" is activated affecting the "nonverbal" representation of the tree concept. On this view words are not pointers to nonlinguistic concepts. Words are best described as operators on conceptual (and, via continued feedback, perceptual) representations. Words, and linguistic expressions more broadly, don't *have* meaning. Rather, they *provide clues* to meaning (Elman, 2004, 2009; Rumelhart, 1979).

This position, which I refer to here as *language-augmented-thought*, makes three broad predictions, stemming from the claims made in the introduction. (1) Insofar as verbal labels change "nonlinguistic" representations, associating a label with a concept should affect the acquisition of the concept. Namely, labeled categories should be easier to learn than

[3] The tension (and apparent incompatibility) between the philosophical thesis that learning conceptual primitives is impossible, and empirical work happening in the cognitive sciences over the past 30 years was on fascinating display at the 2005 Cognitive Science conference in Stresa, Italy, during a symposium entitled "Solutions to Fodor's Puzzle of Concept Acquisition." A transcript can be found at: http://www.wjh.harvard.edu/~lds/pdfs/Niyogi_Snedeker-2005.pdf.

unlabeled categories. (2) Given the bidirectional information flow between the different representational layers, the effects of labels should penetrate even perceptual processes. That is, language use can actually affect what we see. (3) Named concepts should be activated differently under the on-line influence of the label than when the labels are prevented from affecting the concept. These effects may be observed by increasing (up-regulating) or decreasing (down-regulating) the salience of verbal labels and observing the consequences on task performance.

In this chapter, tests of these predictions are restricted to concrete objects, omitting superordinate, relational, and abstract categories. There are two reasons for this: first, in order to study effects of language on cognition it will be necessary to set up experimental paradigms in which ostensibly the same information is communicated linguistically and nonlinguistically. It is much simpler to communicate the concept of a dog (e.g., by showing a picture of a dog) than to communicate the concept of mammals, predators, or the idea of evolution. Second, finding an effect of language on the cognizing of concrete categories is, arguably, a stronger test of the theory than finding an effect of language on abstract categories. While an argument can be made that our knowledge of many abstract and non-perceptual categories comes in large part *from* language in the form of formal education and reading, the same cannot be said of concrete and familiar categories which can be experienced directly (Sloutsky, 2010). Hence, if one finds effects of language on even the most concrete of categories, one might expect even larger (albeit harder to study) effects on more abstract categories (e.g., Gentner & Boroditsky, 2001).

3. Eskimo Snow, William James, and Grecious Aliens

No discussion of words and their potential effects on thought can be complete without Eskimos. In his highly entertaining chapter "Great Eskimo Vocabulary Hoax," George Pullum reviews the intellectual history of the idea, originally penned by the anthropologist Franz Boas (1966/1911), that Eskimos have some varyingly large number of words for snow. Pullum writes that "even if there *were* a large number of [word] roots for different snow types in some Arctic language… this would not, objectively, be intellectually interesting; it would be a most mundane and unremarkable fact. …Botanists have names for leaf shapes; interior decorators have names for shades of mauve … Utterly boring, if even true. Only the link to those legendary, promiscuous, blubber-gnawing hunters of the ice-packs could permit something this trite to be presented to us for contemplation" (Pullum, 1989, pp. 278–279).

Figure 2 Naming patterns of languages reflect the preoccupations of their speakers. An outstanding question is what are the cognitive consequences of naming. (© Dave Coverly www.speedbump.com. Used with permission.) (For color version of this figure, the reader is referred to the web version of this book.)

The fact that cultures specialize in different things is of, course, not in itself surprising (Figure 2), but it is far from trite. It is true that upon hearing that the Hanunóo of the Philippines have around ninety words for rice (Conklin, 1957, cited in Wierzbicka, 1997), we might reasonably conclude that rice cultivation is culturally important—a conclusion we would probably reach through simple observation, knowing nothing about the language. Within a single language as well, we expect people with specialized knowledge to have an enriched vocabulary in that domain. An oenophile not only has additional experience tasting wines, but a vocabulary for varietals, tastes and bouquets that is acquired concurrently with wine-tasting experience. The reason this is potentially important is that encountering a language community with specialized vocabulary in some domain shows that *at the very least* acquiring that kind of expertise is possible. For example, the ability to (accurately) name wine varietals denotes an ability to accurately categorize them. Similarly, a culture in which every individual reliably uses cardinal direction terms indoors and in unfamiliar environments speaks to human capacities (see Levinson, 1997; Majid, Bowerman, Kita, Haun, & Levinson, 2004 for discussion). Thus, observations of novel lexical patterning or elaboration can serve as the raw material for hypotheses and can inform theories of human cognition in the

same way as patterns of associations and dissociations observed through neurological case studies can inform theories of human cognition.

It was this connection between labeling and categorization that formed the crux of the original Eskimo snow example by Boas. Using a word involves a choice to select certain aspects of the experience. Being part of a community that uses a particular word requires the learner to perform the necessary acts of categorization to be able to use the word properly. In Boas's own (1966/1911) (somewhat dense) formulation:

> In our actual experience no two sense-impressions or emotional states are identical. Nevertheless we classify them, according to their similarities, in wider or narrower groups the limits of which may be determined from a variety of points of view. Notwithstanding their individual differences, we recognize in our experiences common elements, and consider them as related or even as the same, provided a sufficient number of characteristic traits belong to them in common. Thus the limitation of the number of phonetic groups expressing distinct ideas is an expression of the psychological fact that many different individual experiences appear to us as representatives of the same category of thought (pp. 20–21).

Thus, pointing out that English has sufficient vocabulary to accommodate the needs of skiers or meteorologists (Pinker, 1994) misses the point: the question is not whether a language can be used to articulate a particular proposition (see Sapir, 1924 on the formal completeness of natural languages), the question, is what are the *consequences* of learning a particular pattern of naming or *any* pattern of naming at all.

The connection between names and categorization was also discussed by William James in his *Principles of Psychology* (1890). James uses an example of learning to distinguish two wines: a Claret from a Burgundy. James writes that the wines have probably been drunk on different occasions and settings, and the next time we drink the wine, "a dim reminder of all those things" is recalled.

> After a while the tables and other parts of the setting, besides the name, grow so multifarious as not to come up distinctly into consciousness; but *pari passu* with this, the adhesion of each wine with its own *name* becomes more and more inveterate, and at last each flavor suggests instantly and certainly its own name and nothing else. The names differ far more than the flavors, and help to stretch these latter farther apart. Some such process as this must go on in all our experience (p. 511).

Speculating further on the importance of verbal labels in apprehending perceptual experiences, James comments that although it may seem that the

difference we feel between the two wines "we should feel, even though we were unable to name or otherwise identify the terms", this difference "is always concreted and made to seem *more substantial* by recognizing the terms." (p. 512). So, apart from any knowledge that may be communicate via labels, the labels themselves, even when communicating to no one in particular, may concrete or ground the experience. This idea is further illustrated by an example coincidentally involving recognition of a kind of snow:

> I went out for instance the other day and found that the snow just fallen had a very odd look, different from the common appearance of snow. I presently called it a 'micaceous' look; and it seemed to me as if, the moment I did so, the difference grew more distinct and fixed than it was before. The other connotations of the word 'micaceous' dragged the snow farther away from ordinary snow and seemed even to aggravate the peculiar look in question (p. 512).

James's description on how labels may alter category learning and recognition of novel exemplars is similar to how the process unfolds in a neural network presented in Section 6. As various exemplars become associated with a common label, the label begins to modulate the representations of these exemplars via feedback which acts to sharpen the category boundaries making "more distinct and fixed" the differences between them. Although "micaceous" is a descriptor rather than category name as such, the hypothesized mechanism is the same: "micaceous" highlights sparkliness (a distinctive feature of mica) just as "dog" highlights the combination of features that are most relevant for distinguishing dogs from non-dogs.

3.1. Grecious Aliens: Testing the James Hypothesis

Do category names actually facilitate the learning of novel categories as James speculated? One way to find out is to train two groups on the same category distinction, providing each with equal learning experiences, but providing only one of the groups with names for the categories. This was the precise approach used by Lupyan, Rakison, & McClelland (2007). The basic task required participants to learn to classify 16 "aliens" into those that ought to be approached and those to be avoided, responding with the appropriate direction of motion (approach/escape). The perceptual distinction between the two alien classes involved subtle differences in the configuration of the head and body of the creatures. On each training trial, one of the 16 aliens appeared in the center of the screen and had to be categorized by moving a character in a spacesuit (the "explorer") toward or away from the alien, with auditory feedback marking the response as correct or not. In the *label* conditions, a printed or auditory

label (the nonsense words, "leebish" and "grecious" depending on the category of the alien) was presented following the accuracy feedback. In the *no-label* condition, the alien remained on the screen by itself. All the participants received the same number of categorization trials and saw the aliens for exactly the same duration; the only difference between the groups was the presence of the category labels that followed each response. The labels, being perfectly predictive of the behavioral responses, constituted entirely redundant information.

The results are shown in Figure 3. Participants in the label conditions learned to classify the aliens about twice as fast as those in the no-label conditions (left panel). In a subsequent study (not shown), we introduced a control condition to determine if any redundant, but perfectly correlated, information would facilitate categorization. The labels were replaced with non-linguistic and non-referential cues in the form of the alien moving in one direction or another to signal where, on the planet, its kind lived. Although learned equally well as the referential labels, these cues failed to facilitate categorization.

After completing the category-training phase during which participants in both groups eventually reached ceiling performance, their knowledge of the categories was tested in a speeded categorization task that included a combination of previously categorized and novel aliens presented without any accuracy feedback and without labels (though the newly learned labels could modulate on-line performance via feedback: see Section 6). Results showed that those who learned the categories in the presence of labels retained their category knowledge throughout the testing phase. Those

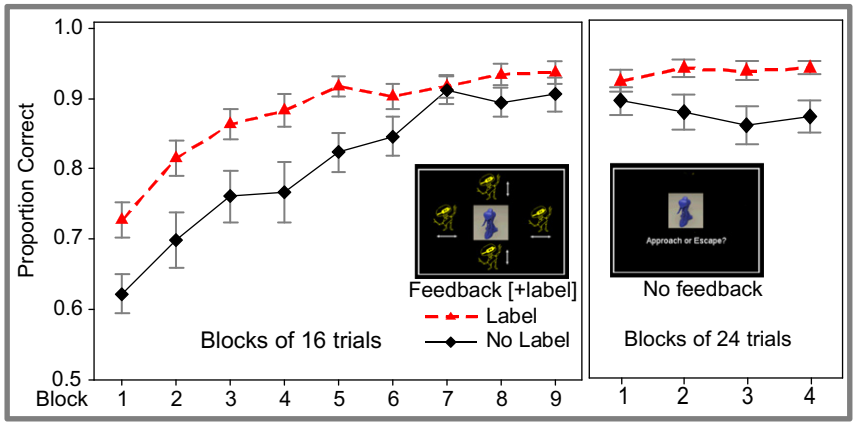

Figure 3 Mean classification accuracy in the initial training (left) and subsequent test phase (right) of Experiment 1 of Lupyan, Rakison, & McClelland (2007). Error bars indicate standard errors of the means. (For color version of this figure, the reader is referred to the web version of this book.)

who learned the categories without labels showed a slight performance drop probably due to the presence of previously unseen exemplars and lack of feedback (Figure 3-right). The difference between label and no-label conditions in this second session was observable even though the session immediately followed the supervised training session. It is likely that the difference would increase in size if a delay is introduced between the two sessions of the experiment.

Learning named categories appears to be easier than learning unnamed categories. More than just learning to map words onto pre-existing concepts (Li & Gleitman, 2002; Snedeker & Gleitman, 2004), words appear to facilitate the categorization process itself. The difference between referential verbal labels and nonreferential cues is further discussed in Section 7. To foreshadow the discussion: there are data showing that labels and evidently equally predictive nonverbal cues have different effects. The exact nature of this difference is still unclear (cf. Lupyan & Thompson-Schill, 2012; Waxman & Gelman, 2009).

4. Effects of Language on Visual Memory: The Categorization-Memory Tradeoff

> I realized that I had never acquired the habit of looking closely at things, and now that I was being asked to do it, the results were dreadfully inadequate. Until then, I had always had a penchant for generalizing, for seeing the similarities between things rather than their differences (Auster, 1990, p. 117).

Suppose we are tasked with designing an algorithm that detects airplanes. The algorithm should output "airplane" if and only if it is presented with an image of an airplane. Recognizing that an Airbus A380 and a Cessna 152 are both airplanes requires representing both as having certain properties in common while ignoring numerous differences. An algorithm whose sole purpose is to discriminate airplanes from non-airplanes may not care that the wingspan of the Airbus is 262 ft. and the wingspan of the Cessna is 33 ft. Of course, this is radically unlike human categorization.[4] Even as we classify (and name) an Airbus and a Cessna as airplanes, we remain cognizant of their differences. Yet, the act of

[4] The reason that within-category differences are never fully collapsed (e.g., see McMurray & Spivey, 2000; McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008 for the argument against invariance within phonemic categories) is that doing so would render the representations useful only for that single type of categorization. This is never the case. So, e.g., we need to know not only whether something is a car, but whether it is our car, whether it is moving, and whether it poses a present danger.

categorization may make the two objects ever more similar than they would be otherwise (e.g., Goldstone, 1994).

Importantly, these effects of categorization may occur on-line, that is, during the categorization process. In addition to whatever effect category learning has on e.g., gradual fine-tuning of feature detectors (Goldstone, 1998), the process of categorization may further augment how the exemplar is represented, *on-line*. Insofar as language requires us to engage in rapid categorization, an act of naming *is* an act of categorization. Thus, simply calling something by its name may shift the representation of the labeled object such that properties typical or diagnostic of the category are highlighted while properties irrelevant to the category are abstracted over. Because categorization is posited to minimize, however slightly, within-category differences, the involvement of category labels should result in enhanced categorization performance, but poorer ability to make within-category distinctions and to remember idiosyncratic details: a tradeoff between categorization and memory.

The prediction that labeling impairs within-category memory was tested in a series of visual recognition memory experiments (Lupyan, 2008a). Participants viewed pictures of common objects such as chairs and tables, and were prompted to label some of them with their basic-level name, e.g., "chair", and to provide a nonverbal response to others, e.g., indicating whether they liked that particular chair or not. Afterward, participants' recognition memory was tested by presenting the original items, one at a time, intermixed with visually similar foils (e.g., the same chair, but without armrests). As predicted, participants had substantially worse memory (*d*-prime) on the objects they had labeled. Item analysis showed that participants had no trouble discriminating a beanbag chair from its foil regardless of whether they overtly labeled it as a chair. But when they were tested on more typical exemplars, labeling resulted in a drastic impairment in memory. Notably, this decrease in performance came from decreased *hits* (from ~80% to ~60%) rather than increased false alarms: labeling a chair as a "chair" made participants less likely to recognize the same chair at test (Figure 4). The results were consistent with an account in which labeling resulted in activation of prototypical features: labeling a typical chair without armrests may have led participants to misremember it as having armrests which results in a higher likelihood of rejecting the original armrest-less chair when it is presented again at test.

The work described above (Lupyan, 2008a) was recently criticized by Richler, Gauthier, and Palmeri (2011) who argued that the observed detrimental effect of labeling is better understood as an enhancement in performance in the control (preference) condition. In currently ongoing studies aimed to addressed this critique, the categorization-memory tradeoff was examined more directly. Participants were shown a series of rectangles of varying aspect ratios. A few seconds after each one, an array of 12

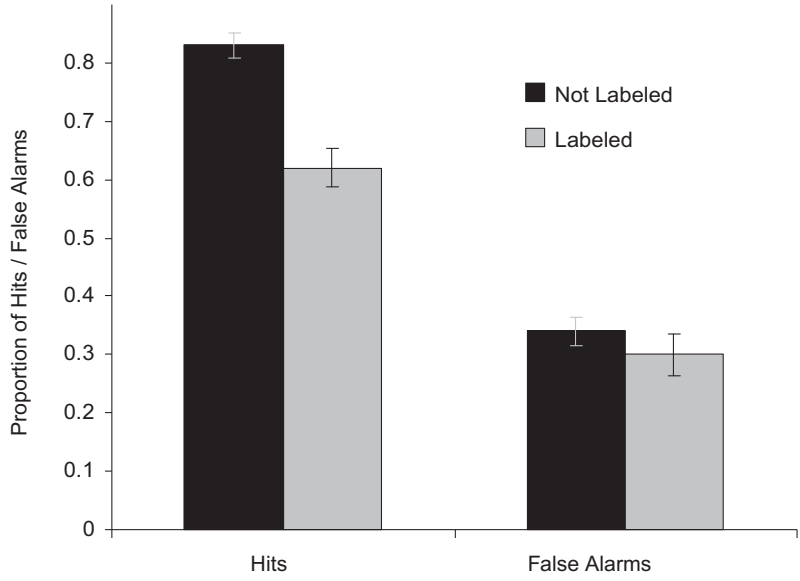

Figure 4 Recognition performance memory in Experiment 2 of Lupyan (2008a). Items that were labeled during the study session resulted in lower hit-rates than items for which participants gave a category irrelevant preference response. (For color version of this figure, the reader is referred to the web version of this book.)

alternatives appeared and participants had to select the rectangle with the exact aspect-ratio they just saw. Some of the rectangles were wider-than taller, and some taller-than-wider resulting in two implicit categories. It was reasoned that categorizing the rectangles into "tall" and "wide" categories (visual categories already well-known to the subjects) would produce poorer memory for precise shape, as would be predicted if labeling quickly produced more categorical representations of the labeled shape.

Four conditions were contrasted: in the *observation-only* condition, participants simply observed each rectangle and selected it from the array of choices as best they could. In the *forced-categorization* condition, participants were asked to categorize the rectangle as "tall" or "wide" while it was visible on the screen, and received accuracy feedback. In the *random-categorization* condition, after the rectangle disappeared participants were cued on 50% of the trials to respond with the category "tall" or "short," receiving no feedback this time, or, on the remaining trials, were instructed to withhold the response. Finally, in the *unrelated-categorization* condition was similar to the random-categorization condition except that instead of optionally categorizing the rectangles as tall or wide, participants were cued on a random 50% of trials to report the identity of a small letter embedded in the rectangle. The total viewing time of each "study" rectangle and the memory test were identical for all conditions.

Participants were told (and could easily observe) that the rectangles varied only in their aspect ratio leaving no ambiguity about the feature relevant to the task. Only a single feature had to be attended. Even so, categorization into "tall" and "wide" categories resulted in poorer memory: performance was poorer in the *forced-categorization* condition than in the *observation-only* condition. Critically, the results showed that explicit acts of categorization (i.e., of the kind involved in verbal naming) was dissociable from more implicit categorization that did not require a response. Thus, in the *random-categorization* condition, a virtually identical difference in memory performance was observed in a within-subject design: the random half of the trials that called for a categorization response produced poorer memory than trials that did not (even though participants did not know at the time of viewing the rectangle whether they would be asked to classify it). In the *unrelated-categorization* condition which served as a control to test whether performing any secondary task decreased visual memory, it was found that making categorization responses unrelated to the shape of the rectangle did not decrease performance; performance in this version of the task was comparable to the *observation-only* condition.

This work suggests that in addition to effects of introducing labels during training, the act of categorizing itself seems to augment the representation of the item being categorized. Specifically, labels appear to make stimuli more categorical (see also Lupyan, 2008a,b, 2009; Lupyan & Thompson-Schill, 2012 for further demonstrations). Section 5 describes effects of this augmentation on perceptual processing.

One reason why verbal labels may impair memory is that labeling enhances the categorization process inducing selective representing the feature of the stimulus that are most typical or diagnostic of the object category thus making individual items less distinctive; computational explorations of this idea are presented in Section 6.

4.1. Some Implications of the Categorization-Memory Tradeoff for Cross-Linguistic Differences

The finding that explicit categorization—the kind that occurs each time we name something—augments on-line the representation of the labeled item has clear implications for thinking about how using different languages can augment ongoing cognitive processing. As in James's example of micaceous snow, a word may drag apart certain aspects of the stimulus, while collapsing others.

Languages whose lexicons include words that refer to certain characteristics thus enable speakers (for better or worse) to selectively highlight those aspects. The point is not that language necessarily provides a unique way to accomplish this, but that simply speaking necessitates such categorization. For example, languages that possess systems of honorifics require their speakers to

decide what discrete level of formality/respect each addressee commands. On the present account, this act of categorization would act to augment on-line how the addressee is represented in the mind of the speaker.

Although the experiments above all concern concrete objects, the categorization-memory tradeoff and its link to language is much more general. For example, consider the following example of representing agency. Suppose that in representing an intentional act there exists a strong associative link between the representation of the action and of the actor. Given the act *John knocked over a glass of water during a political argument*, we ought to care very much that it was John who knocked over the glass because this can help guide future action: during the next argument we may want to place the glass farther away from John. In contrast, because accidental events correlate only weakly with their actors, actors of such events may be less centrally represented. John may thus be de-prioritized in a representation of him *accidentally* knocking over a glass of water. If a language uses a syntactic cue to indicate whether an act is accidental then that cue may quite automatically change the degree to which the agent is represented as being central to the action. The Spanish clitic *se* appears to play such a role, its one function being to signal the degree of intentionality inherent in the act. Speaking grammatically-correct Spanish may thus require speakers to rapidly categorize events as accidental or intentional which may affect their memory for agents. Indeed, as reported by Fausey and Boroditsky (2011), Spanish-speakers have poorer memory for agents of accidental events than English-speakers, whose language does not require in the same way signaling of the intentionality status of a given event.

5. Effects of Labels Run Deep: Penetrability of Visual Processing by Language

> Even comparatively simple acts of perception are very much more at the mercy of the social patterns called words than we might suppose (Sapir, 1929, p. 210).

As argued in a prescient paper by Churchland, Ramachandran, and Sejnowski (1994), the brain is only grossly hierarchical: sensory input signals are only a part of what drives "sensory" neurons, processing stages are not like assembly line productions, and later processing can influence earlier processing (p. 59).[5] The idea that neural processes are *intrinsically* interactive has since received overwhelming empirical support (e.g., Foxe

[5] The notion that perception is more than passive perception of the physical characteristics of a stimulus is quite old and was central, for example, to Locke's doctrine of primary and secondary qualities of objects. A view of perception as a constructive process is also seen in the early 20th century, as when Bergson writes, "Perception is never a mere contact of the mind with the object present; it is impregnated with memory-images which complete it as they interpret it" (Bergson, 1911, p. 133).

& Simpson, 2002; Freeman, 2007; Gilbert & Sigman, 2007; Koivisto, Railo, Revonsuo, Vanni, & Salminen-Vaparanta, 2011; Kveraga, Ghuman, & Bar, 2007; Lamme & Roelfsema, 2000; Mesulam, 1998; Mumford, 1992; Rao & Ballard, 1999; Reynolds & Chelazzi, 2004). To give two examples of gross violations of hierarchical processing in vision: (1) the "late" prefrontal areas of cortex can at times respond to the presence of a visual stimulus *before* early visual cortex (V2) (Lamme & Roelfsema, 2000 for review). (2) The well-known classical receptive fields of V1 neurons showing orientation tuning appear to be dynamically reshaped by horizontal and top–down processes. Within 100 ms. after stimulus onset, V1 neurons are re-tuned from reflecting simple orientation features, to representing figure/ground relationships over a much larger visual angle (Lamme, Rodriguez-Rodriguez, & Spekreijse, 1999; Olshausen, Anderson, & Van Essen, 1993).

An implication of pervasive top–down influences on even the lowest levels of visual processing (e.g., O'Connor, Fukui, Pinsk, & Kastner, 2002) is that even simple visual decisions such as whether some stimulus is present or whether two stimuli are identical depend on interaction between bottom–up and top–down processes. As stated by Foxe & Simpson:

> The rapid flow of activation through the visual system to parietal and prefrontal cortices (less than 30 ms) provides a context for appreciating the 100–400 ms commonly needed for information processing prior to response output in humans. It demonstrates that there is ample time for multiple cortical interactions at all levels of the system during this relatively long processing period (2002, p. 145).

Viewing perception as an interactive process means that non-perceptual influences such as semantic knowledge, goals, and expectations can affect vision (cf. Pylyshyn, 1999). Within the framework of language-augmented thought such top–down influences on perception are extended to linguistic influences. One way to examine the degree to which language augments visual processing is to test whether manifestly linguistic manipulations alter performance on standard visual tasks. This was the same approach used to investigate effects of language on categorization and memory detailed in Sections 3.1 and 4 and was applied here to perceptual processing.

In a series of experiments run by Lupyan and Spivey (2010a), participants viewed briefly presented displays of the numerals 2 and 5, with several from each category presented simultaneously. In Experiment 1 showing the basic effect, the participants' task was to attend to, for example, the 5s and to press a button as soon as a small dot appeared

next to one of the 5s—a category-based version of a Posner cuing task. The more selectively participants could attend to the 5s, and just the 5s, the better they should perform. The linguistic manipulation was implemented here by presenting the word "five" prior to the numeral display on a random 50% of the trials. On the remaining trials participants heard an auditory cue that omitted the category label (Figure 5). Because participants know what the task is—the task of attending to the 5s remained constant for the whole 45-min experiment—the word "five" (or "two") told them nothing they did not already know. Yet, on the randomly intermixed trials on which they actually heard the numeral label, participants responded more quickly

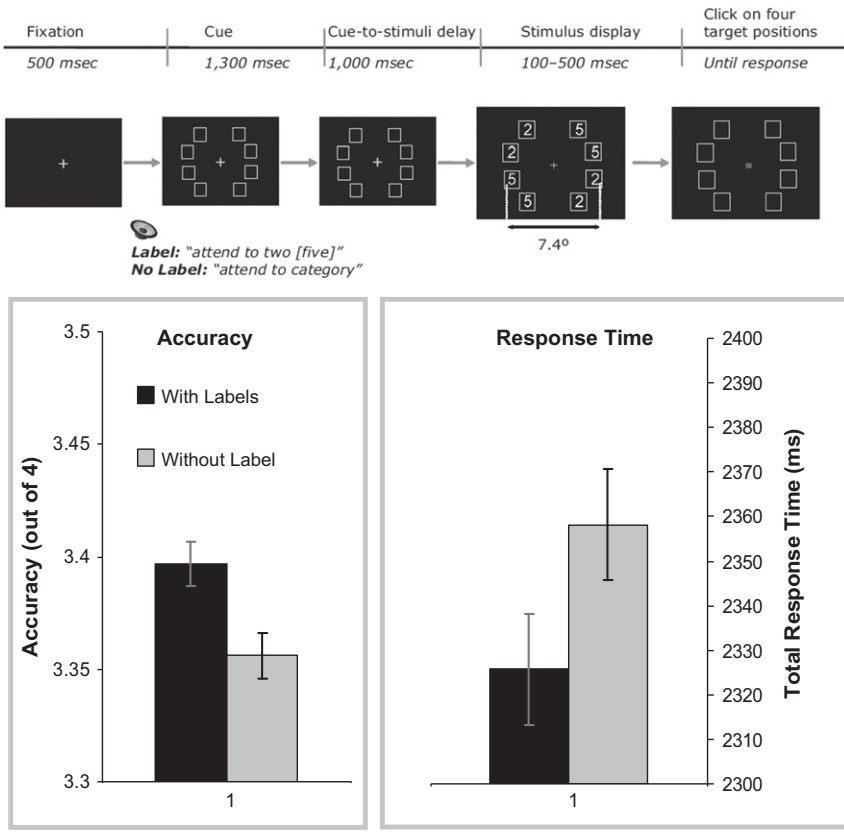

Figure 5 Top: Procedure of Results from Lupyan & Spivey's (2010a) Experiment 6 Bottom: Results showing improved performance in attending to all the items of a given category when it is cued explicitly (and redundantly) with its verbal label. (Figure adapted from Lupyan & Spivey (2010a). Used with permission.)

than on trials on which it was omitted. In another version of the task (Experiment 6) shown in Figure 5, participants had to attend to briefly flashed groups of numbers, being instructed to attend only to 2s (or, for a separate group, only to 5s). On some trials the actual label was heard right before the numbers appeared. The task was to click on all the (now-blank) locations that contained the target digits. We reasoned that if the label helps to activate (or keep active) a visual representation of the task-relevant category in a top−down manner, performance should be superior after the label (once again, even though the label is completely informationally redundant). This is indeed what we found (Figure 5-bottom). This facilitation occurred even when the items were seen for only 100 ms, a time too brief to permit eye movements, supporting the interpretation that the facilitation occurred in parallel throughout the visual display. Similar effects were obtained with more complex items such as pictures of chairs and tables (Lupyan & Spivey, 2010a Experiment 4). One possible confound concerned the finding that trial-by-trial cues have been shown to be more effective in spatial attention tasks than block-wide cues (Posner, Snyder, & Davidson, 1980). If participants did not make use of the block-wide instruction to attend to a particular category then the trial-by-trial cues were actually informative rather than redundant. This possibility was tested in several control studies (Lupyan & Spivey, 2010a Experiments 3A−3B). The results showed that subjects were in fact making use of the block-wide cues as indicated by faster responses of valid than invalid trials, thus ruling out this confound.

An important take-home message from the discussion above is that the observed patterns of finding are only possible if hearing labels induced *transient* effects, over and above whatever long-term effects there are of learning labels. If the facilitation due to hearing a word (i.e., a kind of linguistic up-regulation) carried through the entire experiment, the difference between the intermixed label and no-label trials would quickly vanish. Yet the difference persisted in most cases through the entire experiment lasting for hundreds of trials which was only possible if hearing a label affected perceptual processing in a transient, on-line manner. The finding that labels, which did not communicate any extra information, affected visual processing is entirely unexpected on accounts in which labels simply map onto concepts (Figure 1A).

The finding is accommodated by the language-augmented thought in the following way: The association between the word "five" and the visual form of the Arabic numeral means that hearing the word "five" is expected to activate visual features corresponding to 5s (a 5 prototype of sorts), transiently dragging the representations of subsequently appearing 5s and 2s further apart, while simultaneously making the perceptual representations of the various 5s on the screen more similar, and thereby easier to

simultaneously attend. Notice that this task did not *require* identification or naming. Verbal labels were certainly not needed to see that 2s and 5s are perceptually different. Yet, overt language use—a hypothesized "up-regulation" of what normally takes place during perception—had robust effects on perceptual processing. This verbal description is implemented in a computational model in Section 6.

How far "down" can effects of labels be observed? Consider a simple visual detection task in which the goal is to respond "yes" if a stimulus—any stimulus—is present, and "no" otherwise. Lupyan and Spivey (2010b) presented subjects with backward-masked letters with the contrast of the letter adjusted to each subject to produce about 60% detection rates. That is, on 40% of the trials subjects did not perceive a stimulus when there was one present. The linguistic manipulation involved presenting an auditory letter name prior to the detection phase (Figure 6-top). On these trials, subjects had increased visual sensitivity as measured by a greater *d-prime*. Simply hearing the name of the category enabled participants to detect the presence of briefly-presented masked objects that were otherwise invisible. Interestingly, showing participants a preview of the actual letter (i.e., a bottom−up cue) failed to facilitate simple detection (Figure 6-bottom). In an even stronger demonstration of the power of words to affect basic perception, Ward and Lupyan (2011) used a flash-suppression paradigm known to suppress visual representations at a low level (Tsuchiya & Koch, 2005). It was shown that simply hearing a word (e.g., "zebra") was sufficient to unsuppress otherwise suppressed images (e.g., of various zebras), again hearing a word enabled participants to see what was otherwise invisible.

These results showing that overt presentation of verbal labels affects visual processing are meant to speak to "normal" visual processing being augmented (or guided) to some degree by language. The interaction between vision, language, and categorization was further addressed in several studies that took advantage of a convenient dissociation between the visual and conceptual properties of the letters B, b, and p. The letters in the pairs B-b and B-p are have equal visual similarities, but B-b are more conceptually similar (in that both letters are members of the same class) than B-p. When tasked with performing speeded same-different judgments of physical identity (i.e., B-B = same, B-p and B-b = different), participants' judgments were equally fast for the within-category (B-b) and between-category (B-p) trials (Lupyan, 2008b Experiment 2; Lupyan et al., 2010). A category-effect, measured by the RT difference between B-p and B-b stimuli emerged, however, when a $\geq$ 150 ms delay was introduced between the presentation of the first and second letter in the pair (with the first letter always visible) (Lupyan, Thompson-Schill, & Swingley, 2010) thus showing a gradually unfolding effect of the conceptual category on perception. During the delay, the representation

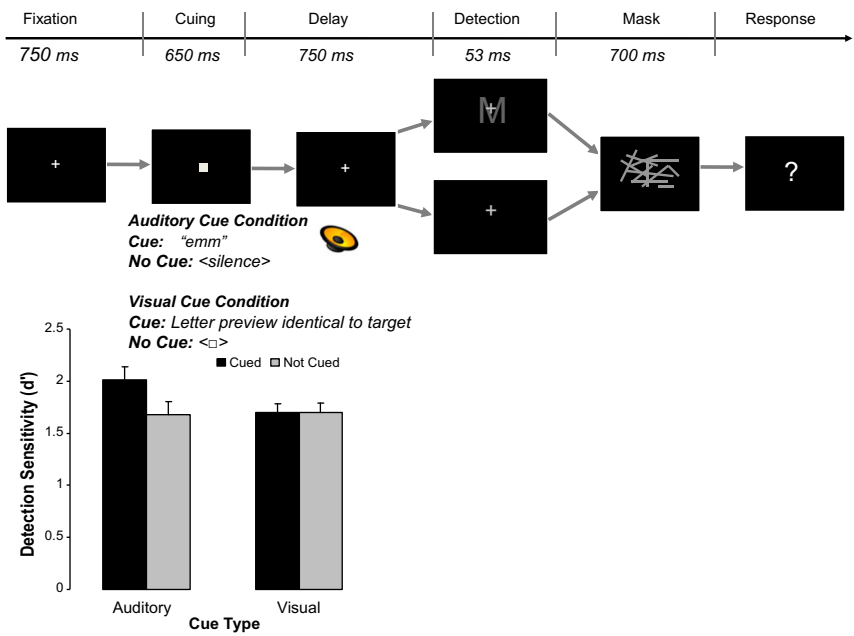

Figure 6 Top: Trial structure of the basic cued object detection paradigm (e.g., Experiment 1 of Lupyan & Spivey, 2010b). During the response part of the trial, participants respond 'present' or 'absent' depending on whether they detected a letter. Bottom. Effects of auditory and visual cues on the detection of cued visual objects. Bars indicate 1SE of the difference between the means. (Figure adapted from Lupyan & Spivey (2010b). Open access.) (For color version of this figure, the reader is referred to the web version of this book.)

of the first letter becomes augmented by its conceptual category, increasing the perceived similarity between B's and b's and decreasing the similarity between B's and p's.

These results show perception to be rapidly affected by the conceptual category of the stimulus, but do not directly implicate language *per se*.[6] Further evidence for the involvement of verbal labels in perception comes from a recent study in which we administered transcranial magnetic stimulation (TMS) to a classic verbal labels in perception region (Wernicke's area; pSTG: BA 22) while participants performed the same—different B/b/p task (Lupyan, Hamilton, & Thompson-Schill, in prep.). Insofar as slower responses to B-b relative to B-p are the result of feedback from labels, disrupting Wernicke's area may affect the category effect. The results showed that an inhibitory stimulation regime completely eliminated the RT difference between responding "different" to B-p and B-b letter pairs.

[6] Although see Lupyan (2008b Experiment 3) in which it is shown that overt presentation of the letter name affects performance on a pop-out visual search task.

Control stimulation to the vertex had no effect. There is, to my knowledge, no theory of visual processing on which Wernicke's area is involved in bottom-up visual processing. That disruption of activity in this region alters behavioral responses on a visual task supports the hypothesis that the effects of conceptual categories (here, letter categories) on visual processing are subserved in part by a classical language area, stimulation of which possibly disrupts its usual modulation of neighboring posterior regions of the ventral visual pathway.

6. Language Augmented Thought: A Model

In this section, I present a model of the language-augmented thought framework I have thus far described only verbally. The theory of language augmented thought laid out in this chapter derives naturally from connectionist principles. Mental representations are viewed as distributed patterns of activity arising from propagation of activations via weighted connections. In recurrent networks of the kind used here, a representation (pattern of activity) at a given point in time is a joint function of bottom—up activity, namely perceptual inputs, and top—down activity, namely constraints derived from prior experience, current task demands, etc. (Elman, 1990; McClelland & Rumelhart, 1981; Rumelhart, McClelland, & the PDP Research Group, 1986).

The model and simulations presented here are should be taken as an extended "intuition pump" (Dennett, 1984) demonstrating how phenomena responsible for the empirical results may emerge, rather than as fully explicated models of particular tasks. Consequently, I will be presenting only general methods and summaries of results; detailed methods and analyses of the network's performance will be reported elsewhere.

6.1. Methods

6.1.1. Network Architecture

All the simulations used the same network architecture shown in Figure 7: a 30-unit ("perceptual") input layer which can receive perceptual input from the "outside" world was connected bi-directionally to a 60-unit intermediate layer. This layer can be thought of as developing "conceptual" representations, but naming it as such serves as descriptive shorthand. As we shall see, the representations learned in this layer were more abstract than the learned perceptual representations. The conceptual layer was in turn connected bidirectionally to a two-unit label layer as well as back to itself. Each unit in the label layer

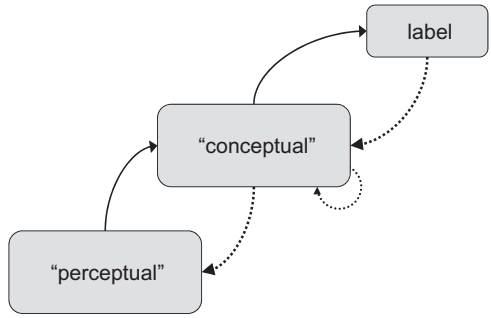

Figure 7 Architecture of the Solid and dashed lines denote feedforward and feedback connections respectively.

corresponded to a category label. The bi-directional architecture meant that the activity of the perceptual and label layers was a function of both external and internal inputs. The model was implemented and trained using the Lens v2.4 neural network simulator (Rohde, 1999). Details regarding additional parameter settings are available on request.

6.1.2. Materials

The networks were trained on exemplars of two categories. Let us call category 1 "goodies" and category 2 "baddies". The categories were generated from the two prototype patterns shown in Table 1. Each value denotes the probability of a particular feature being present for a given goody or baddy. For example, features 1–2 always had a 90% probability of occurring; features 11–12 had a 70% probability each of occurring for baddies, but only a 10% probability of occurring for goodies: Features 21–22 had the opposite pattern: 70% for goodies and 10% for baddies. The feature-set thus comprises two types of features: common features (1–10) and category-specific features (11–30) with some of the latter having a higher likelihood for goodies, and some for baddies.

Table 1 Prototype Patterns Used to Generate Training and Testing Exemplars. The Numbers Reflect Probabilities of Setting a Feature value to 1

	Common Features										Category Specific Features																					
Feature number:	1 … 10										11 … 20										21 … 30											
Categ. 1: (baddies)	.9	.9	.8	.8	.7	.7	.6	.6	.5	.5	.7	.7	.6	.6	.5	.5	.4	.4	.3	.3	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
Categ 2: (goodies)	.9	.9	.8	.8	.7	.7	.6	.6	.5	.5	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1	.7	.7	.6	.6	.5	.5	.4	.4	.3	.3		

6.1.3. Training Regime

The network's recurrent connectivity meant that it could be run in two directions. Given a perceptual stimulus, the network could be asked to name it. Given a name, the network could be asked to output the likely perceptual features of a stimulus with that name. Comprehension and production were trained simultaneously by including three types of intermixed training trials: on *naming trials*, the network was shown one of the category exemplars generated from one of the prototype patterns (that is, the visual units were "soft-clamped" to the incoming visual information). In response, the network had to produce the appropriate category label while also attempting to represent the original item as accurately as possible. The latter requirement corresponds to the fact that when we label a Toyota Camry as a "car" we are simultaneously representing its particular features, e.g., that it's silver and illegally parked. That is, item-specific details are never fully overriden by categorization. On *comprehension* trials, the input comprised only the label. The target pattern was one of the specific exemplars. In the absence of disambiguating information it is, of course, impossible to know what specific exemplar is being referred to (e.g., if I say the word "car" I can't expect the listener to know exactly the make, model, and color of the car I have in mind). Faced with this task, the network does the obvious thing: learn activate the most likely features that correspond to the labeled category. Finally, in *naming + comprehension* trials, the network was provided with both a name and a visual stimulus, and had to reproduce both. Weights were adjusted using a backpropagation through time algorithm.

6.1.4. Testing Procedure

All testing was done using a novel set of exemplars generated from the prototypes shown in Table 1. There were three testing conditions: on *disconnected-label* trials, the label units were prevented from having a top−down effect on the conceptual and visual representations. Thus, although the label layer would continue to output the category labels as before, the lower-level representations could not be affected by them. In the *self-generated-label* condition, the network produced the label itself (leaving open the possibility of mistaken classification). The activated label was then allowed to feed back affecting the conceptual and visual representations. Finally, on the *provided-labels* trials, the correct label was provided externally. This corresponds to the situation of hearing a verbal label applied to something we are currently experiencing.

During testing the weights were frozen. Therefore, the very same network (i.e., set of weights) could be run on different conditions to see how the on-line dynamics played out. It is worth pointing out that while the *network's* state can be frozen and its knowledge assessed, this is not true for humans. For *people*, the training, is never complete: each time we see, hear, or recall something is another learning opportunity. This makes hypotheses that allow for feedback

effects only during learning particularly baffling (e.g., Mitterer & De Ruiter, 2008; Norris, McQueen, & Cutler, 2000).

6.2. Results

6.2.1. Learning to Name

The first obvious way to assess the behavior of the network is to see whether it can successfully label items it has not seen before. As shown in the left panel of Figure 8, the network learns fairly quickly; after 1000 or so weight updates,[7] the network is unambiguously activating the correct label. Performance on the categories of "goodies" was similar. The right panel of Figure 8 shows the activation dynamics of the two label units over the course of a single trial following training. The profile shows typicality effects: certain items take longer to label (in fact, the time the network took to activate the appropriate category label was correlated highly with the Euclidean distance of the stimulus being presented from the category prototypes shown in Table 1).

6.2.2. Forming the Conceptual Representations

The output of the network's internal representation comprises a vector of activation outputs in the range 0−1 for each unit in a layer. So, for example, the visual representations of 100 examples produces a matrix of 120,000 values (100 examples × 30 layer units × 40 time intervals). There are numerous ways of analyzing such multivariate data. A simple method that produces easy-to-visualize results is principal components analysis (PCA) and this is the method used here. The full dataset (with each unit representing a separate dimension) was subjected to PCA. The first three components generate an x,y,z value for each item × time-point combination. For each individual item, a series of these points can then be strung together. Linearly interpolating the intermediate points produces what I call a "tassel plot".

The tassel plots in Figure 9 show training performance. Each string shows the conceptual (top row) and visual (bottom row) representation of a particular item at a given point during training. These representations represent the *end* of the network settling dynamics (i.e., the representation that the network on the last time-interval; cf. Figure 10). Not surprisingly, initially the network knows nothing about the structure of the visual or conceptual spaces and so the representations are entirely overlaid. As the network learns about the regularities, the two categories, shown in black and gray, diverge. Notice that although there are only two categories, the network's internal states continue to represent within-category differences. This is because in addition to learning to label, the network also learns to

[7] This number is a function of the learning rate and can be reduced considerably.

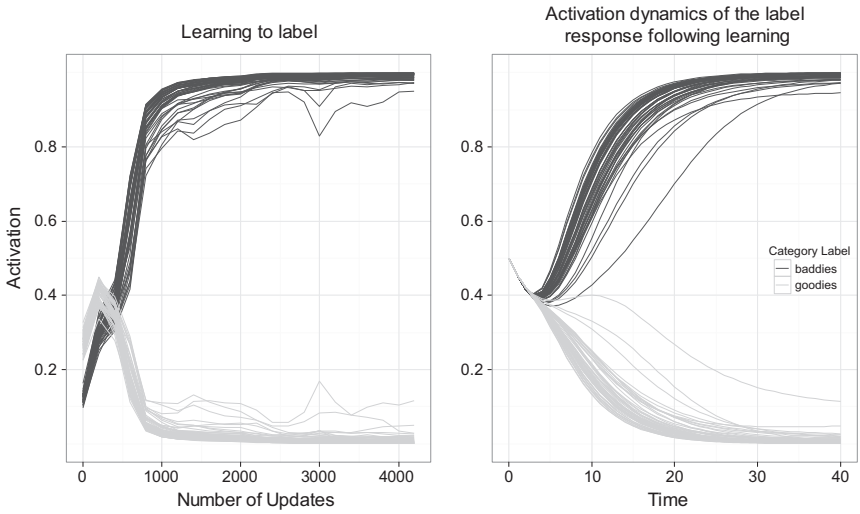

Figure 8 Left: Activation of the "goodies" and "baddies" label units in response to novel "baddies" following varying amounts of training. Right: the on-line activation of the two category labels to novel exemplars after 4200 training trials. Each line shows a different testing exemplar.

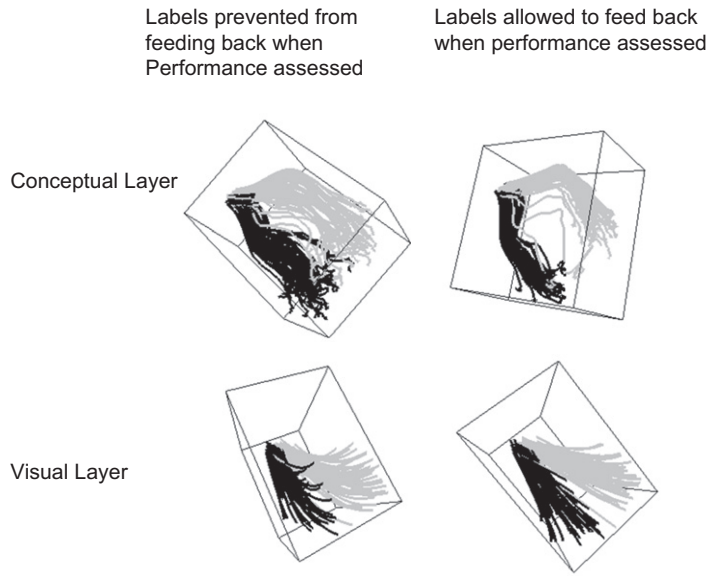

Figure 9 Tassel plots showing diverging representations in the conceptual layer (top row) and visual layer (bottom row) over the course of training. Both columns show performance of the same network in which labels are prevented from (right column) and allowed (right column) to affect the representations on-line during the test. Color represents the two categories of exemplars.

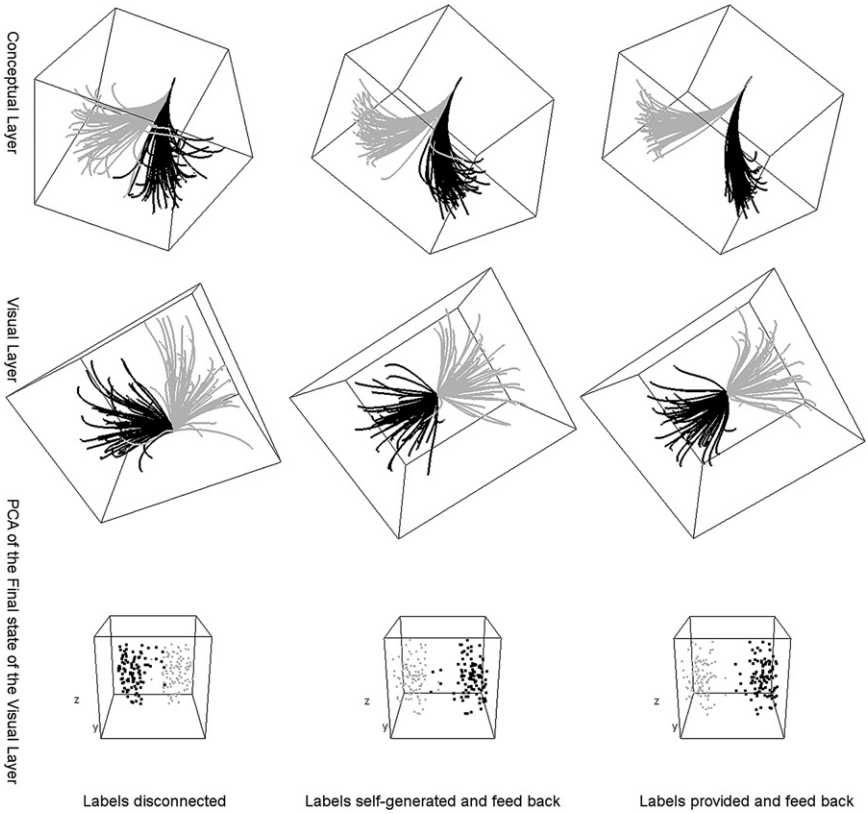

Figure 10 Tassel plots showing on-line activation dynamics following training in the conceptual layer (top row) and visual layer (middle row). The bottom row shows a principal component analysis of the representations at the last time-step. The left column shows representations when labels are disconnected, preventing top–down feedback. The middle column shows results when the labels are activated by the network. The right column shows representations under the influence of externally provided labels (an up-regulation of the automatic effects of the labels on network dynamics). Color represents the two categories of exemplars (see text for additional details).

represent the visual properties of individual exemplars—a burden that is shared unequally between the visual and conceptual layers, allowing the latter to represent the items in a more categorical way.

Although the networks were trained with labels using the procedure described 6.1.3, when the networks were tested at different points during training the labels could be selectively prevented from affecting the lower-level layers to determine if they were affecting these representations. The looser clustering of the tassels in the left column of Figure 9 indicates that disconnecting the labels resulted in less categorical representations

(see below and Figure 12 for a quantification of this difference). These results show that *even when* the training includes labels, allowing the labels to influence the representations in real-time via feedback can contribute to forming more categorical representations. Whether this is beneficial depends on the task. When the task involves distinguishing different classes as in the alien-learning experiments of Lupyan et al. (2007), this influence from labels is helpful. When the task requires representing a given item with high-fidelity as required by a within-category recognition task (Lupyan, 2008a), it is detrimental.

6.2.3. The Unfolding of Representations in Time

After training the network, I examined how conceptual and visual representations unfolded in time. This "unfolding" corresponds to the activation of the representation on a given trial. Figure 10 shows the multivariate analysis for the conceptual (top row) and visual representation (middle row). The three columns of Figure 10 show the three testing conditions, respectively: *disconnected-labels, self-generated-labels,* and *provided-labels*. It is apparent that, even when it is the network itself that generates the label, the resulting representations are different, and specifically, they are more categorical than when the output of the label is prevented from having and online influence. The bottom row shows a temporal slice of the visual representation (the last position of the tassels in the middle-row) for a slightly different perspective.

6.2.4. Quantifying Representational Change Due to Labels

One way to quantify the on-line effects of labels on the conceptual and visual representations is to measure the degree of clustering between exemplars within and between the two categories. One such analysis is shown in Figure 11. A K-means clustering algorithm was applied to the conceptual and visual representations outputted by the network at the end of the test (i.e., at time-interval 40)—these are the representations visualized in the bottom row of Figure 10. Degree of clustering was defined in terms of the average *within-cluster* and *between-cluster* distance between all pairs of exemplars. The y-axis in Figure 11 shows the within-to-between category ratio. A lower number indicates a more categorical representation. There are two main results: First, compared to the "normal" case of the network activating a label on its own that is then allowed to feedback (the self-generated labels condition), disconnecting the labels produces less categorical representations while presenting the label overtly results in slightly more categorical representations. Second, there is a difference in the degree of clustering between the conceptual and visual layers. When labels are allowed to feedback, the conceptual layer shows greater clustering of the items than the visual layer. When

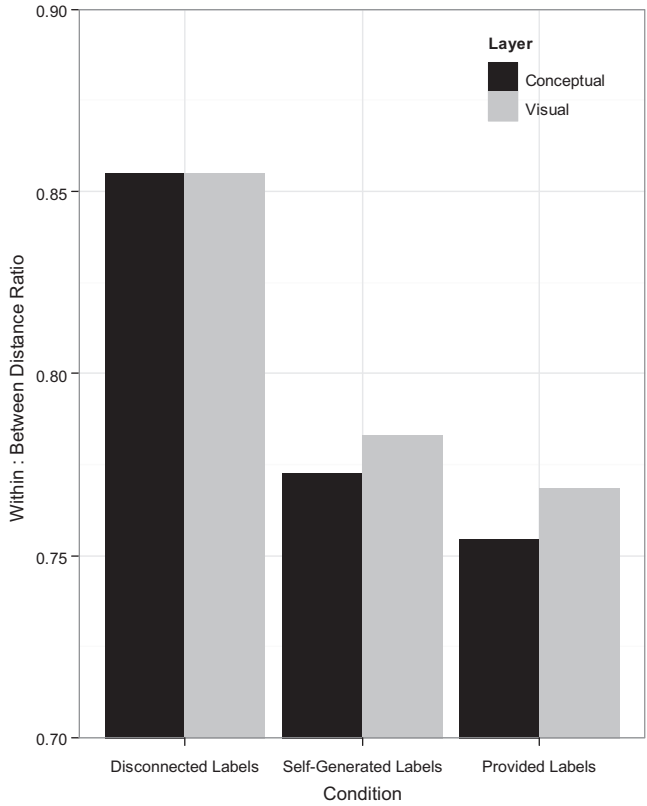

Figure 11 Results of a K-means clustering analysis of the conceptual and visual representations for the same labeling conditions shown in Figure 10. The y-axis shows the ratio of average within-category exemplar distances to the average between-category distance (see text for details).

labels are prevented from feeding back, both layers essentially reflect the visual properties of the items.[8]

Figure 12 shows a parallel analysis of clustering with one tweak. Here, a label is always externally presented, but what varies is the amount of time during which the network processes the label *before* the visual input is provided. Naturally, the best the network can do when presented with just the label is to activate its estimate of the prototype. However, one might expect that the longer that prototype is allowed to "linger" the more categorical the representation of the subsequently presented stimulus will be. This is exactly what was observed. The left-most part

[8] The fact that all the clustering ratios are less than 1 shows that between-category distances are always greater than within-category distances, a natural outcome of using categories having correlated visual features (Table 1). The labels increase clustering over and above that predicted by the visual features alone.

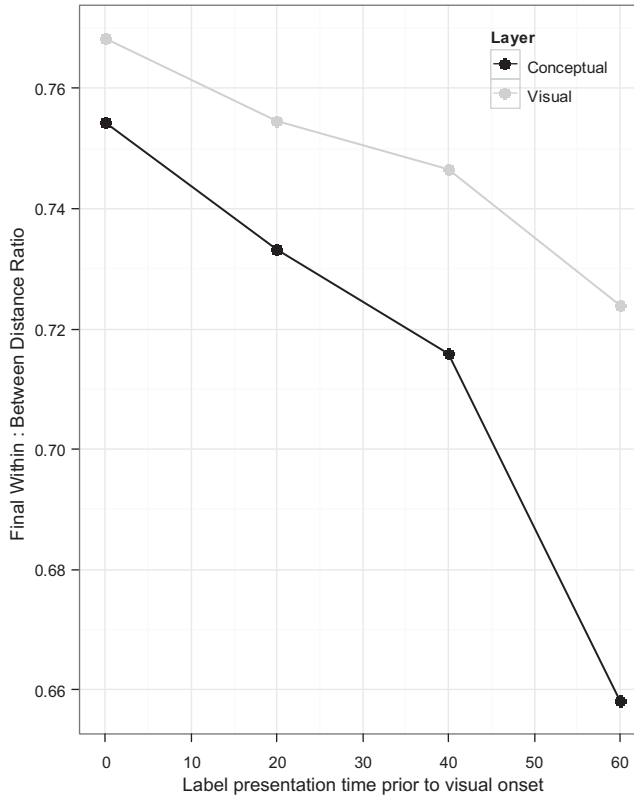

Figure 12 An analysis parallel to that shown in Figure 11 when the label is presented prior to the visual stimulus for varying amounts of time, shown on the x-axis. The visual stimulus was presented for a constant 40 ticks (identical to the network runs shown in Figures 10 and 11).

of Figure 12 just re-plots the provided-label clustering pattern shown in Figure 11. The subsequent time-points show the degree of clustering that results when the label is presented for an increasing amount of time prior to the presentation of the visual stimulus. The presentation of the visual stimulus is kept constant in all cases. Once again, there are two main results: first, the clustering increases as the label is allowed to have an increasing effect through feedback. Second, the difference in the degree of clustering between the two layers (a kind of division of labor) increases as well: When the label is active for a longer time the conceptual representations become progressively more clustered. That is, the network starts to "think" in prototypes, ignoring individual variability. This progressive increase in clustering is also evident in the visual layer, but to a lesser degree.

6.3. Summary of Results

In this section, I described a connectionist model of language-augmented thought. Allowing the labels to feed-back alters the dynamics of the model, resulting in greater clustering which increased as the influence of the label increased. Although the model was not intended to model performance in a particular experiment, it is not difficult to see the parallels between the model results and the human results presented above. For example, insofar as the ability to simultaneously attend to all the members of a given category is facilitated if they are represented in a more categorical way, hearing a label prior to performing this visual task ought to transiently improve performance, and it does (Lupyan & Spivey, 2010a). In analyses not reported here, the model can be extended to closely simulate the recognition memory data reported by Lupyan (2008a). Interestingly, the labels also enable the network to more quickly learn intra-category correlations, e.g., discovering that certain visual features are correlated with other visual features within a category. (e.g., Ross & Murphy, 2010).

All the results above are from networks that received identical past "experiences." The only difference was whether the labels were allowed to affect the representations in real-time using feedback. So, it is not only *learning* labels that can conceptual and visual representations (see Livingston, Andrews, & Harnad, 1998; Lupyan, 2005; Mirolli & Parisi, 2006; Plunkett, Sinha, Moller, & Strandsby, 1992 for additional discussion of effects of labels on bottom-up learning), but the activation of a verbal label in real-time appears to change how conceptual representations are brought on-line. These label-augmentations begin to occur as soon as a category label begins to be associated with exemplars.

Computational models such as these are sometimes criticized for parameter-tweaking: the modeler is thought to adjust the parameters to obtain the results they want. Although this critique is sometimes valid, the reality of the modeling enterprise is that there is always a class of behaviors that simply fall out of the system as soon as it is set up. Some of these are entirely unsurprising. For example, there is a strong relationship between stimulus typicality and naming latency. But among the behaviors that naturally fall out of the model are ones fundamentally incompatible with conceptions of words as simply labeling our concepts (Section 2.1). These model results provide a qualitative fit to the kind of tight interaction between language, categorization, memory, and perception that characterize the empirical results described in this chapter.

The claim that this is a model of "language-augmented thought" is, admittedly, too grand a phrase for what this model is doing: learning about two categories generated from partially overlapping prototypes. Yet, this model exemplifies, the basic claims made in the introduction: verbal labels

can change "nonlinguistic" representations, and these effects, although most readily observed in the higher-level "conceptual" layer, are also observed in the visual layer despite there being no direct connection between the labels and this perceptual layer. The model also addresses the last claim, showing that in the presence of a label, the same category (there are only two after all) is activated differently. The model, however, does not speak to the question of whether labels are "special." Would any cue reliably associated with a category have the same effect? The next section addresses this question empirically.

7. How Special are Labels?

Although I have referred to the top-most layer of the network (Figure 7) as a label layer, there is nothing inherently linguistic about it. In fact, one might expect that any reliable cues to categories—verbal or not—to have the same effects on conceptual and perceptual representations as linguistic labels both in the process of initial learning and in subsequent thinking about the category members. For example, learning that cats (and only cats) are called "cats" would be identical to learning that cats (and only cats) meow. In the first case, one learns to represent particular cats as instances to which the label "cat" is applied. In the second case, one learns to represent particular cats as instances that make a particular sound. As this association is learned, the categorical representation should exert an effect equally in both cases. In this Section, I review several empirical findings that speak to this issue. Computational explorations of this question will be presented elsewhere.

As described in the alien-categorization experiment (Lupyan et al., 2007 see Section 3), associating stimuli with labels facilitated category-learning while associating stimuli with equally correlated information, basically semantic facts about where the aliens lived, failed to facilitate categorization. One of the many unanswered questions is whether a similar dissociation can be observed not just when learning new categories but when activating knowledge about familiar concepts. If referential labels activate concepts in a particularly effective way, by e.g., selectively activating diagnostic information that is useful in recognizing an exemplar as a member of the given category (an immediate consequence of the kind of increase in clustering seen in the model), then individuals may recognize familiar items more readily when cued by verbal rather than nonverbal means.

In a series of studies conducted Lupyan and Thompson-Schill (2012), participants were cued by basic-level category names (e.g., "cat") or nonverbal sounds (e.g., a meowing sound) which extensive norming

showed to be unambiguously associated with the category. The first series of experiments comprised simple picture-verification: participants heard a verbal or nonverbal cue that was followed by a delay, after which a picture appeared that with 50% likelihood either matched or did not match the cue. The dependent variable was the time it took participants to make a "match" or "no-match" response to the picture. On the view that labels are just a convenient way to access a concept, responses should be equally fast regardless of how one accesses the concept provided the cue is unambiguous. The results showed a consistent advantage for verbal cues even for cue-offset to picture-onset delays as long as 1500 ms. (Figure 13). The finding that the label-advantage was not eliminated (and in fact, grew in size) with longer cue-to-picture delays rules out an interpretation of the finding purely in terms of speed of activation. For example, a differences in the speed of accessing a common concept may be predicted if people have more familiarity with the label "cat" than a meowing sound. Such a difference, however, is expected to diminish or disappear with longer delays. That it did not suggests that verbal labels do not simply activate conceptual representations faster, but that representations activated via verbal cues are *different* in some way than representations activated via nonverbal means.

The label advantage is entirely unexpected on the view that there is a single concept that is *accessed* by verbal cues, nonverbal cues, and the picture, and that the match/no-match response is generated based on the activation of this common concept (e.g., Gleitman & Papafragou, 2005; Jackendoff, 2002; Li, Dunham, & Carey, 2009; Snedeker & Gleitman, 2004; Snodgrass, 1984; Vanderwart, 1984). On the other hand, if labels more than other kinds of cues selectively activate the category-typical features (resulting in the kind of increased clustering seen in the model), then hearing a label would activate the "same" concept differently.

In subsequent experiments, Lupyan and Thompson-Schill (2012) showed that the verbal-cue advantage generalized to a visual discrimination task involving only minimal semantic knowledge. Instead of indicating whether the cue and picture matched—a task that requires full semantic processing of the target image—participants were instead flashed with two versions of the same picture: one upright, the other upside-down and had simply had to respond indicating which side of the screen contained the upright image. Matching verbal labels resulted in a greater validity effect (baseline RT − valid RT) compare to valid sound cues. Invalid labels produced a greater slowing down than sound cues. The last study in the paper extended these findings to novel categories. Participants were trained to associate alien musical instruments with either their names (nonsense words such as "whelph" and "shonk") or with their sounds (unfamiliar and meaningless sound effects). Participants showed equal facility with associating the visual exemplars

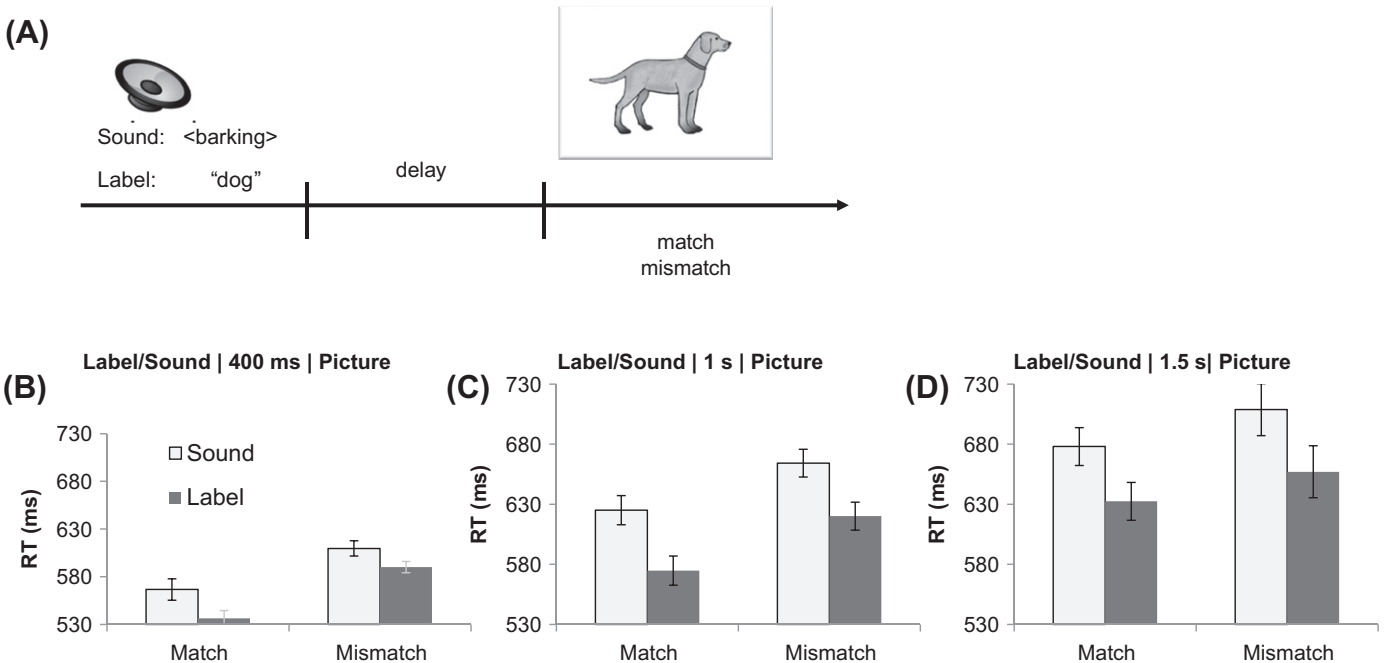

Figure 13 Verification times for the sound trials versus label trials in Lupyan and Thompson-Schill (2012) for Experiments 1A–1C which varied in the length of the delay between the offset of the auditory cue and the onset of the picture. The auditory cue and the picture matched on match trials and mismatched on mismatch trials. Error bars show 1 standard error of the mean difference between label and sound conditions. (For color version of this figure, the reader is referred to the web version of this book.)

with both types of cues. However, after only a 10-min training session (which was sufficient for participants to reach ceiling performance for this simple stimulus set), verbal cues were more effective at activating the category representation than nonverbal cues as determined by a performance pattern on the upright-picture location task strikingly similar to that seen with familiar objects.

7.1. Effects of Labels on Formally Defined Categories

Even seemingly simple categories like dogs exist in a vast feature space and have complex intra-category structure. In contrast, a category such as "triangle" has a formal definition: a three-sided polygon. In another set of experiments, Lupyan (2011, in prep) examined whether category labels activated categories like "triangle" differently from various circumlocutions that expressed the same formal definition (e.g., phrases such as "three-sided polygon" and "three-sided shape"). Consider the following set of results: When asked to draw a "figure with three sides", all drew triangles; 50% were isosceles/equilateral and 50% were parallel to the bottom of the page. When a separate group was asked to draw a "triangle", 91% drew isosceles or equilateral triangles; 82% drew triangles with bases parallel to the bottom of the page. A similar pattern was observed in a within-subject speeded recognition task. After hearing "triangle", participants were faster to verify isosceles than scalene triangles—a finding that is in line with typicality effects. However, this typicality gradient was only present on trials on which participants heard the word "triangle". On the randomly intermixed trials on which participants heard "three-sided", participants were equally fast and accurate regardless of the type of triangle shown to them (cf. Armstrong, Gleitman, & Gleitman, 1983).

In another study, participants were presented with pictures of triangles that were close to being equilateral. For each picture, one group was asked in written form "how many equal sides does this three-sided figure have?" Another group, presented with the same sequence of shapes was asked "how many equal-sides does this triangle have?" It would seem obvious that both questions are ostensibly the same. Yet, participants were much more likely to respond that all sides were equal (i.e., that the triangle was equilateral) when the question actually used the word "triangle". A second question asked participants to judge the angle of the figure's base relative to the bottom of the screen: zero for perfectly horizontal, and positive and negative for clockwise and counter-clockwise deviations, respectively. The results showed that the slope estimates were significantly more exaggerated, e.g., 20 degrees were judged as 25, and -20 as -25, when the term "triangle" was used than when it was omitted from the question. One explanation for why reading the word "triangle" would cause such

a distortion is that the category label activated a more typical or canonical instance of the category, which judging by recognition performance, has a horizontal base. This activated exemplar may have acted as an implicit contrast set, producing an increase in judged deviations from the true horizontal. In summary, these data show that even for categories that have known formal definitions, increasing the salience of the category label (and up-regulation of sorts) affected production, speeded recognition, and unspeeded visual reasoning. Despite the apparent equivalence in the *reference* of the two cue types, using the actual category name—"triangle"—appeared to reliably activate a more typical or canonical representation of the category which affected performance on a range of tasks.

Recent literature has seen vigorous debates regarding whether labels are special (see Waxman & Gelman, 2009 for review). On one view, labels are "merely" associations (e.g., Sloutsky & Fisher, 2004; Sloutsky, Lo, & Fisher, 2001): they are features of objects just as, to use the example of Waxman and Gelman (2009, p. 259) a black beret is a feature of the experience we associate with Jean Piaget. According to Waxman and Gelman, "this assertion runs aground because the words of human language are more than associations. Words refer".

This distinction dissolves somewhat on the theory of language-augmented thought. As argued by Lupyan and Thompson-Schill (2012), the fact that words refer is a property of language, not a mechanism for understanding the effect that words seem to have on human cognition. On the present position, words are indeed more than just simple features of the stimulus. Words appear to be special, but they *become* special by (1) the experience we have *associating* words with various category members and the high correlations that are formed between labels and diagnostic dimensions/features of the category, and (2) the ability of word activations to feed back and affect the unfolding of "lower-level" representations. The question of whether non-word cues that are also reliable category markers can come to function as words is an empirical one which ongoing work is actively attempting to address. Differences between e.g., words and nonverbal sounds may arise from stronger or more rapid feedback evinced by verbal labels.

8. So, What Do Words Do?

We live in a linguistic world. Human development is notably characterized by learning to refer to things, relations, ideas, etc. using language, and in the process, how we come to represent the external world is affected. Why shouldn't it be? Experience changes us and language is one very salient form of experience.

For example, learning to fly airplanes made me increasingly attuned to visual patterns that look like (and hopefully actually are) airports—visual patterns I've seen (and ignored) many times while flying as a passenger See if you can spot the airport in Figure 14. Did flight training rewire my visual system? I would guess, not very much. However, the learning process has allowed me to *guide* my vision and attention in particular ways. Learning words is similar. Learning to name colors does not in itself rewire our visual system. Given that the same visual system must be used for numerous tasks, it would be maladaptive for one task to dominate it (see footnote 4 for the case of categorical perception in speech). However, learning color name means that perceptual experiences from thereon become (potentially) perceptuo-linguistic experiences. Seeing a color now rapidly activates its name which can then feed back and modulate ongoing conceptual and perceptual processing. The degree of label activation is predictably greater if we are actively naming the color such as when we talk about it. But even "default" perception can be augmented by automatically co-activated color names (a parallel to the self-generated label condition of the model). Verbal interference on this view acts to interfere with this on-line effect of labels on ongoing processing.

Figure 14 Spot the airport. (For color version of this figure, the reader is referred to the web version of this book.)

In their overview of the language and thought literature, Gleitman and Papafragou note that,

> Inconveniently enough, it is often hard to study language development apart from conceptual and cultural learning or to devise experiments in which these factors can be prevented from interacting… [The] difficulty of even engineering such language—thought dissociations in the laboratory is one significant point in favor of a linguistic—relativistic view. Why should it be so hard to pry them apart if they are so separate? (2005, p. 653).

The answer is simple: it is hard to pry them apart because our brains are interactive systems in which different representational layers interact. An effect of language on thought and perception does not mean that perception is somehow verbal, or that our concepts are somehow linguistic in form. Nor does it mean that language must therefore inexorably constrain our thinking or perceiving (a truly strawman view of linguistic relativity). This interactive perspective *does* mean that we should take seriously the vast differences between languages (Evans & Levinson, 2009). Finding that a difference as subtle as using the word "triangle" when asking about an orientation of a triangular figure can affect orientation judgments, offers a hint at some of the more provocative differences that different languages may have on the unfolding of our mental states.

REFERENCES

Armstrong, S. L., & Gleitman, L. R. (1983). What some concepts might not be. *Cognition*, *13*(3), 263—308, doi:10.1016/0010-0277(83)90012-4.

Auster, P. (1990). *Moon palace*. New york. Ny: penguin.

Bergson, H. (1990). *Matter and memory*. Cambride, MA: The MIT Press.

Bloom, P., & Keil, F. C. (2001). Thinking through language. *Mind & Language*, *16*(4), 351—367.

Boas, F. (1966). *Introduction to handbook of American Indian languages*. U of Nebraska Press.

Boroditsky, L. (2010). How the languages we speak shape the ways we think: The FAQs. In M. J. Spivey, M. Joanisse, and K. McRae (Eds.), *The cambridge handbook of psycholinguistics (p. forthcoming)*. Cambridge: Cambridge University Press.

Burling, R. (1993). Primate calls, human language, and nonverbal communication. *Current Anthropology*, *34*(1), 53, 25.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, *25*(6), 657—674.

Chomsky, W. (1957). *Hebrew: The eternal language* (Stated 1st ed.) Philadelphia, PA: Jewish Publication Society.

Churchland, P. S., Ramachandran, V., & Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch, and J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 23—60). Cambridge, MA: The MIT Press.

Cilento, R. (1971). Sir Joseph Banks, F.R.S., and the Naming of the Kangaroo. *Notes and Records of the Royal Society of London*, *26*(2), 157—161.

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers, and J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162–183). Cambridge University Press.
Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, *10*(8), 370–374. doi: 10.1016/j.tics.2006.06.012.
Deacon, T. (1997). *The symbolic species: The co-evolution of language and the brain*. London: Allen Lane: The Penguin Press.
Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. The MIT Press.
Devitt, M., & Strelny, K. (1987). *Language and reality: An introduction to the philosophy of language*. Cambridge, MA: London: MIT Press.
Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.
Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306. doi: 10.1016/j.tics.2004.05.003.
Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, *33*(4), 547–582. doi: 10.1111/j.1551-6709.2009.01023.x.
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(05), 429. doi: 10.1017/S0140525X0999094X.
Fausey, C. M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review*, *18*(1), 150–157. doi: 10.3758/s13423-010-0021-5.
Fodor, J. A. (1975). *The language of thought* (1st ed.). Cambride, MA: Harvard University Press.
Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans – a framework for defining "early" visual processing. *Experimental Brain Research*, *142*(1), 139–150.
Freeman, W. J. (2007). The place of "codes" in nonlinear neurodynamics. *Progress in Brain Research*, *165*, 447–462. doi: 10.1016/S0079-6123(06)65028-0.
Gentner, D., & Boroditsky, L. (2001). *Individuation, relational relativity and early word learning. Language acquisition and conceptual development*. Cambridge, UK: Cambridge University Press.
Gilbert, C. D., & Sigman, M. (2007). Brain states: Top-down influences in sensory processing. *Neuron*, *54*(5), 677–696.
Gleitman, L., & Papafragou, A. (2005). Language and thought. In K. Holyoak, and B. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 633–661). Cambridge: Cambridge University Press.
Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology-General*, *123*(2), 178–200.
Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.
Haviland, J. B. (1974). A last look at cook's Guugu Yimidhirr word list. *Oceania*, *44*(3), 216–232.
Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, *39*(1), 3–36.
Hockett, C. F. (1966). The problem of universals in language. In (2nd ed.), Greenberg, J. H. (Ed.), *Universals of language*, *Vol. 2* (pp. 1–29) Cambride, MA: The MIT Press.
Hurford, J. R. (2004). Human uniqueness, learned symbols and recursive thought. *European Review*, *12*(04), 551–565. doi: 10.1017/S106279870400047X.
Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, and evolution*. Oxford, England: Oxford University Press.
James, W. (1890). *Principles of psychology*, *Vol. 1*. New York: Holt.

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(7), 2488–2492. doi: 10.1523/JNEUROSCI.3074-10.2011.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and Cognition*, *65*, 145–168.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Lamme, V. A. F., Rodriguez-Rodriguez, V., & Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral Cortex*, *9*(4), 406–413.

Lee, P. (1996). *The whorf theory complex: A critical reconstruction*. John Benjamins Pub Co.

Levinson, S. C. (1997). From outer to inner space: Linguistic categories and non-linguistic thinking. In J. Nuyts, and E. Pederson (Eds.), *Language and conceptualization* (pp. 13–45). Cambridge: Cambridge University Press.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, *83*(3), 265–294.

Li, P., Dunham, Y., & Carey, S. (2009). Of substance: The nature of language effects on entity construal. *Cognitive Psychology*, *58*(4), 487–524. doi: 10.1016/j.cogpsych.2008.12.001.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology-Learning Memory and Cognition*, *24*(3), 732–753.

Lupyan, G. (2005). Carving nature at its joints and carving joints into nature: How labels augment category representations. In *Modelling language, cognition and action: Proceedings of the 9th neural computation and psychology workshop* (pp. 87–96). Singapore: World Scientific.

Lupyan, G. (2008a). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348–369.

Lupyan, G. (2008b). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition*, *108*, 566–577.

Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review*, *16*(4), 711–718. doi: 10.3758/PBR.16.4.711.

Lupyan, G. (2011). Representations of basic geometric shapes are created Ad-Hoc. Concepts, Actions, and Objects Workshop. Presented at the Concepts, Actions, and Objects Workshop, Rovereto, Italy.

Lupyan, G., Hamilton, R., Thompson-Schill, S.L. Effects of TMS on conceptual influences on perceptual processing. Manuscript in preparation.

Lupyan, G., & Spivey, M. J. (2010a). Redundant spoken labels facilitate perception of multiple items. *Attention, Perception, & Psychophysics*, *72*(8), 2236–2253. doi: 10.3758/APP.72.8.2236.

Lupyan, G., & Spivey, M. J. (2010b). Making the invisible visible: Auditory cues facilitate visual object detection. *PLoS ONE*, *5*(7), e11452. doi: 10.1371/journal.pone.0011452.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology-General*, *141* (1), 170–186 doi:10.1037/a0024904.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1082.

Lupyan, G., Thompson-Schill, S. L., & Swingley, D. (2010). Conceptual penetration of visual processing. *Psychological Science*, *21*(5), 682–691.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3).

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive activation model of context effects in letter perception .1. An account of basic findings. *Psychological Review*, *88*(5), 375–407.

McMurray, B., & Spivey, M. (2000). The categorical perception of consonants: The interaction of learning and processing. *Proceedings of the Chicago Linguistics Society*, *34*(2), 205–220.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1609–1631. doi: 10.1037/a0011747.

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*, 1013–1052.

Mirolli, M., & Parisi, D. (2006). Talking to oneself as a selective pressure for the emergence of language. In A. Cangelosi, A. D. M. Smith, and D. Parisi (Eds.), *The evolution of language: Proceedings of the 6th international conference* (pp. 214–221). Singapore: World Scientific.

Mitterer, H., & De Ruiter, J. P. (2008). Recalibrating color categories using world knowledge. *Psychological Science: A Journal of the American Psychological Society/APS*, *19*(7), 629–634. doi: 10.1111/j.1467-9280.2008.02133.x.

Mumford, D. (1992). On the computational architecture of the neocortex II. The role of cortico-cortical loops. *Biological Cybernetics*, *66*(241), 251.

Murphy, G. L, & Ross, B. H. (2010). Category vs. object knowledge in category-based induction. *Journal of Memory and Language*, *63*(1), 1–17. doi: 10.1016/j.jml.2009.12.002.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *The Behavioral and Brain Sciences*, *23*(3), 299–325, discussion 325–370.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *5*(11), 1203–1209. doi: 10.1038/nn957.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*, 4700–4719.

Pinker, S. (1994). *The language instinct*. New York: Harper Collins.

Plunkett, K., Sinha, C., Moller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? *Connection Science*, *4*(3 & 4), 293–312.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology-General*, *109*(2), 160–174.

Pullum, G. K. (1989). The great Eskimo vocabulary Hoax. *Natural Language & Linguistic Theory*, *7*(2), 275–281. doi: 10.1007/BF00138079.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*(3), 341–365.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, *2*, 79–87.

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611–647, 15217345.

Richler, J. J., Gauthier, I., & Palmeri, T. J. (2011). Automaticity of basic-level categorization accounts for labeling effects in visual recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(6), 1579–1587. doi: 10.1037/a0024347.

Rohde, D. L. T. (1999). *Lens: The light, efficient network simulator*. School of Computer Science, Carnegie Mellon University.

Rumelhart, D. E. (1979). Some problems with the notion that words have literal meanings. In A. Ortony (Ed.), *Metaphor and thought* (pp. 71–82). Cambridge University Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, *Vols. 1 and 2*. Cambridge, MA: MIT Press.

Sapir, E. (1924). The grammarian and his language. *American Mercury*, *1*, 149–155.

Sapir, E. (1929). The status of linguistics as a science. *Language*, *5*, 207–214.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, *34*(7), 1244–1286. doi: 10.1111/j.1551-6709.2010.01129.x.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology-General*, *133*(2), 166–188.

Sloutsky, V. M., Lo, Y. F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*(6), 1695–1709.

Snedeker, J., & Gleitman, L. (2004). Why is it hard to label our concepts? In D. G. Hall, and S. R. Waxman (Eds.), *Weaving a Lexicon (illustrated edition* (pp. 257–294) Cambridge, MA.: The MIT Press.

Snodgrass, J. G. (1984). Concepts and their surface representations. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 3–22, doi: 10.1016/S0022-5371(84)90479-1.

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neurosci*, *8*(8), 1096–1101. doi: 10.1038/nn1500.

Vanderwart, M. (1984). Priming by pictures in lexical decision. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 67–83. doi: 10.1016/S0022-5371(84)90509-7.

Ward, E. J., & Lupyan, G. (2011). Linguistic penetration of suppressed visual representations. Presented at the Vision Sciences Society, Naples, FL. Retrieved from http://www.visionsciences.org/abstract_detail.php?id=36.328.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263. doi: 10.1016/j.tics.2009.03.006.

Whorf, B. L. (1956). *Language, thought, and reality*. Cambridge, MA: MIT Press.

Wierzbicka, A. (1997). *Understanding cultures through their key words: English, Russian, Polish, German, and Japanese*. Oxford University Press.

Wolff, P., & Holmes, K. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 253–265.