ELSEVIER

COGNITION

Cognition 82 (2002) B75–B88

www.elsevier.com/locate/cognit

Brief article

Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency

Thad A. Polk*, Charles Behensky, Richard Gonzalez, Edward E. Smith

Department of Psychology, University of Michigan, 525 E. University, Ann Arbor, MI 48109-1109, USA

Received 20 March 2001; accepted 27 June 2001

Abstract

When judging the similarity of two stimuli, people's ratings often differ depending on the order in which the comparison is presented (A vs. B or B vs. A). Such directional asymmetries have typically been demonstrated using complex concepts that have a large number of semantic features and a standard explanation is that different sets of features are emphasized depending on the direction of the comparison. In this study, we show that directional asymmetries in the similarity of simple perceptual stimuli can be predictably manipulated merely by presenting each member of a pair with different frequency. Participants rated the similarity of color patches before and after performing an irrelevant training task in which a subset of colors was presented ten times more frequently than others. The similarity ratings after training were significantly more asymmetric than the ratings before training. We discuss the implications of these findings for models of similarity judgment and propose a computationally explicit explanation based on asymmetries in representational stability.

Keywords: Similarity judgment; Neural network; Color perception; Prototypical; Reference points; Frequency; Attractor; Simulation; Contrast model

* Corresponding author. Fax: +1-734-763-7480.
E-mail address: tpolk@umich.edu (T.A. Polk).

0010-0277/02/$ - see front matter
PII: S0010-0277(01)00151-2

1. Introduction

Similarity plays a central role in a variety of cognitive processes. For example, object recognition is often assumed to require judging the similarity of a perceptual representation with representations in memory (Biederman, 1987; Ullman, 1989). Many theories of categorization are based on a similar judgment involving semantic, rather than perceptual, similarity (Hintzman, 1986; Medin & Schaffer, 1978; Smith, 1995). Even among higher cognitive processes like analogy, similarity plays a central role (Gentner, 1983; Gentner & Markman, 1997; Ross, 1989; Ross & Kilbane, 1997). The question of how people judge similarity is clearly of critical importance in cognitive psychology.

Early models of similarity judgment assumed that similarity could be conceptualized as a metric distance between concepts: similar concepts are nearby in semantic space, while dissimilar concepts are far apart. A straightforward version of such a geometric model would predict that similarity judgments would satisfy metric axioms such as symmetry (i.e. $\mathrm{Sim}(a,b) = \mathrm{Sim}(b,a)$ for all a and b). Tversky (1977) demonstrated, however, that people's similarity judgments can differ depending on the direction of the comparison. For example, many people judge North Korea to be more similar to Red China than Red China to North Korea.

One common account of such directional asymmetries assumes that the surrounding context exerts a top-down influence on which features are emphasized, and that the emphasized features differ depending on the direction of the comparison (Glucksberg & Keysar, 1990; Medin, Goldstone, & Gentner, 1993; Ortony, 1979; Ortony, Vondruska, Foss, & Jones, 1985). Accordingly, most of the empirical work demonstrating similarity asymmetries has used complex, cognitive concepts (e.g. countries, famous people, animals) that have a large number of semantic features. On the other hand, Rosch (1975) suggested that similarity asymmetries could arise from a more fundamental bottom-up asymmetry in the representations themselves. The idea is that some representations are more prototypical than others, independent of context, and that they serve as cognitive reference points, that is, as representations that other stimuli are seen "in relation to". Rosch hypothesized that non-prototypical stimuli would be more easily assimilated to (and therefore judged more similar to) prototypical reference stimuli than vice versa, and that similarity judgments would therefore exhibit predictable directional asymmetries (see Tversky, 1977, for a variant of this hypothesis based on salience). According to this view, similarity asymmetries could arise even with simple perceptual stimuli that do not have a large number of associated semantic features.

There are a few experiments that suggest that similarity asymmetries can indeed arise with simple perceptual stimuli. Tversky (1977) found evidence of directional asymmetries when participants rated the similarity of geometrical forms that varied in their goodness of form (e.g. how symmetric they were) or that varied in their complexity. Similarly, Rosch (1975) found directional asymmetries using straight lines that varied in orientation and color patches that varied in hue. In these studies, participants rated the salient/prototypical stimuli (symmetric forms, complex forms, vertical/horizontal lines, focal colors) to be less similar to non-prototypical stimuli than vice versa.

One interpretation of such effects is that they reflect an asymmetry in the representations themselves. An alternative interpretation is that even similarity asymmetries involving simple perceptual stimuli reflect a contextual effect on the weighting of different features. The idea is that perceptual stimuli might have a number of semantic associations that differ between prototypical and non-prototypical stimuli. For example, focal red might be associated with anger, with stop signs, and with fire whereas other shades of red might have a different set of associations (or none at all). If so, and if people do emphasize the features of one stimulus over another depending on the direction of the comparison (Medin et al., 1993), then one might naturally expect to observe similarity asymmetries even with simple perceptual stimuli. One appealing aspect of this interpretation is its parsimony; it accounts for similarity asymmetries involving both perceptual and semantic stimuli in terms of a single mechanism.

On the other hand, the Rosch (1975) hypothesis that there are asymmetries in the prototypicality of representations themselves fits very naturally with work in neural computation that has been developed independently. Neural representations are typically assumed to correspond to a distributed pattern of activity across a network of interconnected neurons. Furthermore, communication in such networks is not unidirectional; rather such networks are *recurrent*. It is well known that in recurrent networks, unlike in simpler feed-forward networks, some distributed patterns of activation are more stable than others (they have lower energy states; Hopfield, 1982, 1984). One natural computational instantiation of the Rosch (1975) notion of prototypicality (or the Tversky (1977) notion of salience) is that more prototypical/salient stimuli correspond to more stable distributed representations.

Furthermore, the Rosch (1975) hypothesis that non-prototypical stimuli are more easily assimilated to prototypical stimuli than vice versa maps very naturally onto the dynamics of recurrent networks. When new input is presented to a recurrent network, the activation does not immediately assume a new, fixed pattern, but evolves over time until it settles into a final stable pattern (a so-called *attractor* pattern). Furthermore, the amount of time it takes the network to settle depends on how similar the initial and final activation patterns are as well as on their relative stability or strength. If two patterns are identical then no changes need to be made and the settling time (and switch cost) is zero. Conversely, if the two patterns are quite dissimilar, then it is harder for the network to switch between them. In keeping with the Rosch (1975) assumption that it is easier to assimilate a non-prototypical stimulus to a prototypical stimulus than vice versa, it is also easier to switch from a less stable (higher energy) activation pattern to a more stable (lower energy) activation pattern than vice versa.

In short, Rosch's hypothesis regarding directional asymmetries in similarity ratings can be naturally instantiated in a recurrent network in which more prototypical stimuli are represented using more stable distributed patterns and in which the difficulty of assimilating one stimulus to another is modeled in terms of settling time. We implemented these ideas in a simple simulation to confirm their feasibility (see Appendix A for details). We repeatedly presented a recurrent network with five activation patterns and used a correlation-based Hebbian learning rule to modify the connection strengths. To simulate differences in prototypicality, we presented some of the patterns more frequently than others, which led the network to develop more stable

representations for those patterns. The model therefore explicitly predicts that prototypes are not fixed but can change, which seems to differ from the Rosch (1975) idea of cognitive reference points and may be more consistent with the Tversky (1977) notion of salience. (Nosofsky (1988) demonstrated that frequent presentation of color patches does indeed increase their rated typicality.) Finally, we simulated similarity judgments by measuring the number of processing cycles the network required to switch from an initial activation pattern to another target pattern. As expected, the simulation produced asymmetric similarity judgments: it required fewer processing cycles to switch from a less stable pattern to a more stable pattern than vice versa.

This computational investigation suggested a way to test whether similarity asymmetries can arise from representational asymmetries, independent of context effects on feature weighting. As previously discussed, similarity asymmetries involving perceptual stimuli could still reflect directional effects on feature weightings, because perceptual stimuli may have different semantic associations. We therefore tested whether similarity asymmetries could be influenced by manipulating the prototypicality/salience of perceptual stimuli, without changing the stimuli themselves. If similarity asymmetries are influenced by changes in salience and/or prototypicality when the comparisons themselves are identical (same stimuli, same direction/context), it would suggest that asymmetries in the prototypicality of the representations themselves can play a role, independent of context.

Based on our simulations, we hypothesized that presenting some stimuli more frequently than others would influence similarity asymmetries, even if the comparisons themselves were identical.[1] We asked participants to rate the similarity of color patches that differed in hue (pre-test). We then manipulated the frequency of exposure to different hues while participants performed an irrelevant training task (training). Finally, we again collected similarity ratings on the same hues (post-test). We predicted that after training, participants would rate the infrequent hues to be more similar to the frequent hues than vice versa.

2. Methods

2.1. Participants

Forty-five University of Michigan undergraduates from the Introductory Psychology Subject Pool participated.

[1] Exposure frequency is not the only factor that influences stability in the model. In particular, the stability of a target pattern also depends on the other patterns on which the network is trained and on their similarity to the target pattern. For example, presenting a set of patterns that are all variants of the same prototype can actually lead the prototype itself (which was never presented) to develop the most stable activity pattern. Thus, higher exposure frequency does not always translate into greater stability/typicality. That said, the model does predict that, all else being equal, increasing exposure frequency will increase stability and typicality.

2.2. Materials and stimuli

Color patches were generated using Adobe Photoshop 5.0. We constructed five shades of blue (labeled Blue1, Blue2, …, Blue5) and five shades of green (Green1, …, Green5) by varying Photoshop RGB values (Table 1). The experiment was conducted on Apple G3 Power Macintosh computers with 17-inch monitors, 1024×768 screen resolution, color depth millions. Participants were seated about 24 inches from the monitor.

2.3. Procedure

The experiment consisted of four parts: a pre-test, a training phase, a post-test, and a post-experiment questionnaire. During the pre-test, participants rated the similarity of pairs of color patches on a scale from 0 to 9 (0 = highly dissimilar, 9 = highly similar). These pairs were presented in a text question to emphasize the direction of the comparison: "How similar is (color patch 1) to (color patch 2)?" The text was presented in 48 point Chicago font. Each color patch was 140×140 pixels. The sentence was centered both horizontally and vertically. The text "Blue1" or "Green1" (depending on the color) appeared under the first color patch and "Blue2" or "Green2" appeared under the second in 32 point Chicago font. The sentence and color patches remained on screen until a response was made. After each response, the screen was cleared and the next comparison appeared after 500 ms. Comparisons were always between different hues of the same color. Each pair of hues was presented four times, twice in each direction, for a total of 160 trials. Trials were randomized except that the same hue was not presented in consecutive trials.

During the training phase, two squares of different sizes but exactly the same hue were presented, and participants judged which was larger (color was irrelevant). Participants were instructed to press '1' on the number keypad if the left square was larger and '2' if the right square was larger. Four different sizes were used: 125×125, 131×131, 138×138, and 144×144 pixels. All four sizes appeared

Table 1
Photoshop RGB values for the ten color patches in the experiment

Hue	Photoshop RGB values		
	Red	Green	Blue
Blue1	0	170	255
Blue2	0	122	255
Blue3	0	0	255
Blue4	79	36	255
Blue5	106	36	255
Green1	147	189	34
Green2	99	189	34
Green3	0	189	65
Green4	0	189	100
Green5	0	189	129

with equal probability. The left square appeared at 30% of the screen width, the right square at 70%, and both were centered vertically.

Two of the blue hues and two of the green hues were presented 110 times each while the other hues were only presented 11 times each (a 10:1 frequency ratio). Half of the participants (the 4–5 group, selected at random) were presented with Blue4, Blue5, Green4, and Green5 frequently and the other hues infrequently. The other participants (the 1–2 group) were presented with Blue1, Blue2, Green1, and Green2 frequently and the other hues infrequently.[2]

The post-test was identical to the pre-test in all respects. In a post-experiment questionnaire, participants were asked to describe any strategies they had adopted, to indicate whether they had tried to remember previous responses in arriving at ratings, to indicate if they had any vision problems, to make a guess about the purpose of the experiment, and to indicate whether they maintained concentration throughout the experiment.

3. Results

Ten participants who failed to achieve 90% accuracy on size judgments or who reported on the post-experiment questionnaire a failure to maintain their effort/concentration throughout the experiment were excluded from the analysis.

Mean similarity ratings were obtained for all possible hue comparisons. Hues seen frequently during the training portion were labeled trained; those seen infrequently during training were labeled untrained. A repeated measures ANOVA was performed that included training group (1–2 group or 4–5 group) as a between-subjects factor and the following five within-subjects factors: direction (forward when untrained hue appeared first and trained hue appeared second, backward when trained hue appeared first), color (blue or green), untrained hue (three values, one for each of the three untrained hues), trained hue (two values), and test (pre-test or post-test).

As predicted, presenting some hues more frequently than others during training increased similarity asymmetries in the post-test relative to the pre-test: the size of the direction effect (the difference between untrained–trained (forward) comparisons and trained–untrained (backward) comparisons) was significantly greater in the post-test than in the pre-test (left panel of Fig. 1, $F(1, 33) = 5.17$, $P < 0.05$). After training, participants tended to judge less frequent hues to be more similar to more frequent hues than vice versa ($F(1, 33) = 10.92$, $P < 0.005$), but before training the

[2] Another approach would be to train arbitrary pairs in different subjects. We chose not to adopt this approach because manipulating the frequency of a color could potentially influence the perceived salience/prototypicality of neighboring/intervening colors. For example, if we made colors 1 and 3 frequent, and if the representations of these colors overlap with the representation of color 2, then color 2's representation might also be influenced. Because we do not know the extent to which such interactions occur, we would not be able to make unambiguous predictions about asymmetries involving color 2 (should it be treated as trained or untrained?). By training the 1–2 pair in one group and the 4–5 pair in the other, we avoided this problem while also demonstrating that the effect was not restricted to one specific pair of colors that were presented frequently.

comparison's direction did not significantly affect the rated similarity ($F(1,33) = 1.32, P > 0.05$). The directional asymmetry also exhibited the predicted increase when the participant groups (1–2, 4–5) and colors (Blue, Green) were analyzed separately (four graphs on the right of Fig. 1). These analyses lacked the power of the full analysis (they excluded half the data), but in all four cases the effect was significant at the 0.10 level and in two of the cases (4–5 group, Blue pairs) it was significant at the 0.05 level. As expected, training did not lead to significant changes in similarity asymmetries for trained–trained and untrained–untrained pairs ($F(1,33) = 0.75,\ P > 0.05$; in such pairs, both members were presented with equal frequency during training). There were no significant main effects for color (blue vs. green, $F(1,33) = 0.36,\ P > 0.05$) or for training group (1–2 vs. 4–5, $F(1,33) = 0.25,\ P > 0.05$), but there was a main effect for test with similarity ratings being significantly higher in the post-test than in the pre-test ($F(1,33) = 5.51, P < 0.05$).[3]

4. Discussion

These results demonstrate that manipulating the frequency with which stimuli are presented can influence directional asymmetries in similarity judgment. As previously discussed, a standard explanation of similarity asymmetries assumes that different features are emphasized depending on the direction of the comparison. Can this hypothesis account for the present results? The major perceptual dimension along which the color patches differed was their hue.[4] Although it is possible that participants emphasized hue more when making comparisons in one direction rather than the other, it seems unlikely. Hue was by far the most salient feature, regardless of the direction of the comparison. It is therefore unlikely that hue was systematically de-emphasized when the comparison was made in one direction relative to the opposite direction.

More importantly, any hypothesis must explain why the frequency manipulation would lead to changes in the observed asymmetry. Identical color patches were used in the pre-test and post-test so the changes in the asymmetry cannot be attributed to

[3] We did not predict this effect a priori, but the neural network model does suggest one possible interpretation. Most people consider the trained hues (1 & 2 or 4 & 5) to be non-focal versions of blue and green (most people consider Blue3 and Green3 (which no one saw frequently) to be the most focal). Assuming that presenting these non-prototypical hues more frequently does indeed increase their representational stability, then the differences in representational stability between these non-prototypes and more prototypical hues might be expected to be reduced leading to increased similarity ratings overall. Another possibility is that some subjects realized that they had never been asked to compare identical hues in the pre-test and therefore re-calibrated their ratings in the post-test to include higher ratings.

[4] The color patches probably differed slightly along other perceptual dimensions as well (e.g. saturation, brightness), because we did not adopt elaborate safeguards to ensure that they did not. (Although the Photoshop brightness and saturation settings were the same, there is no guarantee that the stimuli on the screen were perfectly matched.) Perfect matching of the stimuli along these other dimensions was unnecessary, because the same stimuli were used in the pre-test and post-test and therefore served as their own control. The observed changes in asymmetry therefore cannot be attributed to perceptual features of the color patches themselves.

perceptual features of the color patches themselves. One possibility is that participants explicitly encoded the frequency with which different hues were presented during the training phase, that this feature was used in the similarity comparisons, and that it was differentially weighted depending on the direction of the comparison. This hypothesis could potentially explain both how asymmetries could arise and why they would depend on the frequency manipulation.

One problem with this hypothesis is that there is no evidence that participants noticed the frequency with which different hues were presented. During the training phase (when frequency was manipulated), the participants were performing a size comparison task that required them to focus on the size of the color patches and to ignore their hue. To the extent that they succeeded in ignoring hue, they would not be expected to attend to the difference in frequencies with which the hues were presented. Consistent with this interpretation, none of the participants mentioned the frequency manipulation when asked (in the post-experiment questionnaire) what they thought the experiment had been about or what strategies they had employed.

Alternatively, perhaps asymmetries were indeed due to differential weighting of a frequency feature, but this feature was encoded implicitly rather than explicitly. Even under this assumption, it is difficult to explain the direction of the effect that training had on the observed asymmetry. The assumption by Medin et al. (1993) is that people emphasize features of the second base/referent concept in a directional comparison. Under this assumption, frequency would be emphasized in the forward comparisons (in which the frequent hues appear in the base/referent position) more than in the backward comparisons. Because the two patches *differ* in frequency, one

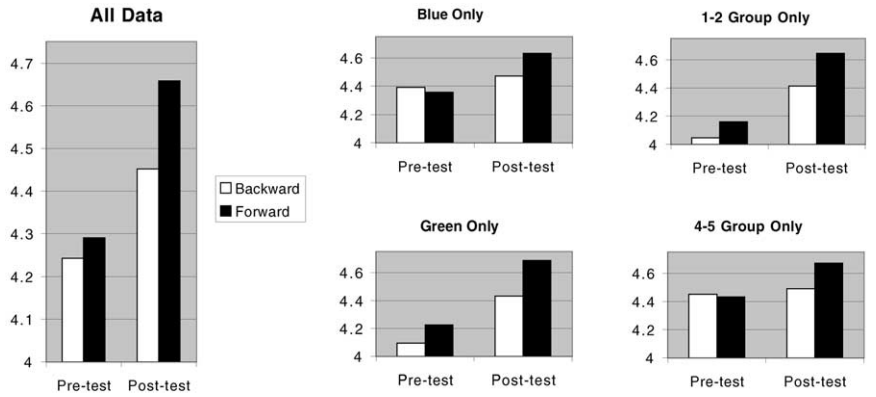

Fig. 1. Average similarity rating as a function of the direction of the comparison both before and after exposure frequency was manipulated in an irrelevant size comparison task. The graph on the left shows all the data collapsed together and the four graphs on the right show the data broken down by color (blue trials only, green trials only) and by participant group (1–2 group only, 4–5 group only). The directional asymmetry in similarity ratings (the difference between forward and backward comparisons) was consistently larger after training than before. Forward comparisons refer to trials in which a low frequency color patch was compared to a high frequency color patch. Backward comparisons refer to trials in which a high frequency patch was compared to a low frequency patch. Similarity was rated on a 0–9 scale (0 = highly dissimilar, 9 = highly similar).

might expect that emphasizing frequency would lead to lower similarity ratings and that the infrequent–frequent pairs (forward comparisons) would therefore be judged less similar (more distinctive) than the frequent–infrequent pairs (backward comparisons). The opposite pattern was observed.

We are not claiming that context plays no role in similarity judgment. Indeed, in many situations people do appear to emphasize different features depending on the direction of the comparison (Medin et al., 1993; Ortony et al., 1985). Our point is rather that asymmetries in the prototypicality and/or salience of representations themselves may also play a role in similarity asymmetries, independent of context. A natural explanation of the present results is that the frequency manipulation led to an asymmetry in the prototypicality or salience of the color representations (Nosofsky, 1988), and that this representational asymmetry gave rise to the observed similarity asymmetries. We operationalized this idea in terms of the stability of different patterns of activation in a recurrent network. This architecture is independently motivated by a simple fact about neural computation (namely, that it is recurrent), it is known to produce some representational states that are more stable than others (Hopfield, 1982, 1984), and it has been found to be useful in explaining a variety of other cognitive and neuropsychological phenomena (e.g. Becker, Moscovitch, Behrmann, & Joordens, 1997; Cree, McRae, & McNorgan, 1999; Farah, O'Reilly, & Vecera, 1993; Hinton & Shallice, 1991; McClelland & Rumelhart, 1981; Mozer & Behrmann, 1990; Mozer, Halligan, & Marshall, 1997; Plaut, McClelland, Seidenberg, & Patterson, 1996; Plaut & Shallice, 1993; Tanaka, Giles, Kremen, & Simon, 1998).

Acknowledgements

Thanks to Melissa Carmody, James Christensen, and Jamie Loundy for their help in administering the experiment and to Todd Stincic and LeeAnn Mallorie for their help in preparing this manuscript. We also gratefully acknowledge helpful conversations with David Meyer and William Gehring about this manuscript as well as the constructive comments of three anonymous reviewers. This research was supported by a grant from the University of Michigan Rackham Graduate School and the Office of the Vice President for Research.

Appendix A. Recurrent network simulation

A.1. Motivation

How could a recurrent network compute the similarity of two activation patterns over the same set of units, even in principle? Intuitively, one would like to measure the degree of overlap between the patterns (are the same units ON in both patterns) and, although this is trivial to do as an outside observer, it is not at all clear how a network could do this itself. Even if the network could represent the superposition of the two patterns at the same time (which is difficult in a recurrent network because

different patterns compete), there would then be no way to compare them. One obvious approach is for the network to represent one pattern and then represent the other. If the patterns overlap substantially then only a few aspects of the pattern will need to be changed, whereas if the patterns do not overlap much then a lot more changes will need to be made. Assuming that more changes require more work, then switching between similar patterns will be easier than switching between dissimilar patterns. Consequently, the ease of switching between patterns is a natural measure of similarity. The ease of switching between patterns is also influenced by their relative stability and it was this realization that led us to explore the possibility that asymmetries in stability could both predict and explain asymmetries in similarity ratings. Another independent motivation for the model's measure of similarity is that it maps very naturally onto the Rosch (1975) hypothesis that it is easier to assimilate a non-prototypical concept to a prototypical concept than vice versa.

A.2. Architecture

The neural network architecture consisted of 100 units that were fully interconnected except that units did not have connections to themselves. All the connection weights between the units were initialized to 0. In addition to receiving input from the other units in the network, each unit also received an external input signal making it possible to control the external input to the network.

Network activity was updated synchronously at each time point t according to the following sigmoid function:

$$A(t) = 1/(1 + e^{-(\text{NetInput}(t) - \text{Bias})})$$

$A(t)$, NetInput(t), and Bias are all column vectors with 100 rows (one row per unit). $A(t)$ represents the activation level of each unit at time point t and NetInput(t) represents the net input to each unit at time point t. This function produces activations that vary between 0 (when the NetInput to a unit is very negative) and 1 (when the NetInput is much larger than Bias, which was set to 5 for all units). When NetInput equals Bias, the resulting activation is 0.5. The Bias term is therefore a kind of threshold: when NetInput exceeds Bias, the activation is between 0.5 and 1, but when it does not the activation is between 0 and 0.5.

The NetInput to each unit at each time point t is the sum of the external input plus the activity of all the other units weighted by the strength of the connection linking them to the target unit:

$$\text{NetInput}(t) = \text{ExternalInput} + \text{Weights}^{*}A(t-1)$$

Here ExternalInput is a 100-row column vector representing the external input to each unit, Weights is a 100×100 matrix in which the entry at position (i,j) represents the connection strength from unit j to unit i, $A(t-1)$ is the activation vector at the previous time point $t-1$, and the * represents matrix multiplication.

A.3. Learning

After each processing cycle, the connections between the units were modified according to a correlation-based Hebbian learning rule. First, the activation of each unit was converted to a (−1,1) scale reflecting percentage above/below a baseline activation level of 0.5 (by subtracting 0.5 from the activity and multiplying by 2). Next, for each pair of units, the product of these scaled activations was computed. This product ranged between −1 and 1 depending on the relationship between the scaled activations: if both were very active or very inactive then the product would be close to 1; if one were very active and the other were very inactive, then the product would be close to −1. The product was then compared to the current connection strength between the units. If the product of the scaled activations was larger than the current connection strength, then the connection strength was increased by a proportion of the difference (controlled by a PositiveLearningRate parameter). Conversely, if the current connection strength was larger than the product, then the connection strength was weakened (controlled by NegativeLearningRate, always significantly smaller than PositiveLearningRate so that associations would get learned faster than they would be forgotten). These modifications in connection strength were scaled by the amount of positive activation in the presynaptic unit so that the presynaptic unit's activation had to be above baseline for a connection's strength to be modified.

A.4. Training

We trained the network by repeatedly presenting five patterns on the external input lines. These patterns were all variants of a single prototype pattern that had ten arbitrary units active (1.0) and the other 90 units inactive (0.0). Each of the five training patterns had one of the active units in the prototype pattern turned off and one of the inactive units turned on. We adopted this approach for two reasons. First, we assumed that the patterns corresponded to specific examples of the same category (e.g. different shades of blue) and would therefore be quite similar to each other. That is why we used patterns that share a number of units. Second, we wanted to be sure that all the patterns were equally stable before the frequency manipulation was instituted (so that differences in stability after the frequency manipulation could be unambiguously attributed to that manipulation). By making all of the patterns equidistant from a prototype pattern, we ensured that none would be more stable than any other to begin with.

These five patterns were presented to the network one at a time via the external input lines to each unit and this input was weighted very strongly (connection weights fixed at 20) so that the input pattern would be guaranteed to be instantiated after a single processing cycle. After the presentation of each input pattern, the connections between units were modified according to the correlation-based learning algorithm described previously and the activation was then reset to zero in all the units. These five patterns were each presented once, in the same order, in each

training epoch and the total number of training epochs varied from 50 to 200 in different simulations.

After the initial training in which the five patterns were presented equally frequently, we then adopted a skewed training regimen in which two of the five patterns were presented more frequently than the other three. In each training epoch, the three lower frequency patterns were presented once, and the two higher frequency patterns were presented five, ten, or 20 times in different simulations. The total number of these extra training epochs varied between one and five in different simulations. Our goal was to investigate whether manipulating frequency would lead to asymmetries

A.5. Testing

After training, we measured the number of processing cycles that the simulation required to switch from a higher frequency pattern to a lower frequency pattern and vice versa. We did this by first initializing the network in one of the higher frequency patterns, setting the external input to be one of the lower frequency patterns, and then counting the number of processing cycles the simulation required to converge (the convergence criterion was based on the square root of the sum of squared differences between previous and current activations being less than 0.001). Next, we initialized the network in one of the lower frequency patterns, set the external input to be one of the higher frequency patterns, and again counted the number of processing cycles the simulation required to converge.

We ran a variety of simulations in which we varied the PositiveLearningRate (0.1, 0.2, and 0.4), the ratio between the PositiveLearningRate and NegativeLearningRate (six, eight, and ten), the number of initial training epochs in which all five patterns were presented equally frequently (50, 100, and 200), the number of extra training epochs in which two of the patterns were presented more frequently than the other three (one, two, and five), and the ratio of the higher frequency patterns to the lower frequency patterns in these extra training epochs (five, ten, and 20).

A.6. Results

The simulation consistently produced a directional asymmetry in switch cost: it required more processing cycles to switch from the more frequently trained pattern to the less frequently trained pattern than vice versa. Indeed, in the parameter space we searched, there were only two parameter settings that failed to produce this asymmetry (in both settings: PositiveLearningRate = 0.1, NegativeLearningRate = 0.01, and initial training epochs = 50; in one of the settings, the number of extra training epochs = 1 and the ratio of higher to lower frequency patterns = 10; in the other setting, the number of extra training epochs = 2 and the ratio of higher to lower frequency patterns = 5). For both of these parameter settings, the simulation required the same number of cycles (six) to switch between patterns in both directions. It appears that the lack of an asymmetry in these cases was an artifact of the specific convergence threshold that we adopted: when switching from the higher frequency pattern to the lower frequency pattern, the simulation barely managed to pass the

convergence threshold after six cycles whereas in the other direction it had barely failed to pass the threshold after five cycles. When we relaxed the convergence threshold (square root of sum of squared differences less than 0.01 instead of 0.001), both of these parameter settings also produced the asymmetry.

A.7. Discussion

One way of thinking about stability in recurrent networks is to consider a network of just two units and the connections between them. If the connections are excitatory, then the pattern in which both these units are ON will tend to be more stable than a pattern in which one unit is ON and the other is OFF. Indeed, if either unit is ON then it will excite the other unit and turn it ON as well. Consequently, with these connection strengths, the ON–ON pattern is stable whereas the ON–OFF pattern cannot survive without strong external input (it is extremely unstable). Furthermore, assuming a Hebbian learning rule like the one we used, each time the ON–ON pattern is presented it will make the connections between the units even more excitatory, making this pattern even more stable. Exposure frequency therefore increases stability. In the case of multiple interconnected units, one can think of the connection strengths as soft constraints (excitatory connections want units to have the same value, inhibitory connections do not want both units to be ON). A pattern's stability then corresponds to how many constraints it satisfies (and with many units/connections, not all of the constraints have to be satisfied for a pattern to be stable). With each presentation of a pattern, Hebbian learning makes small weight changes that tend to lead more constraints to be satisfied, and so exposure frequency tends to increase stability.

This analysis can also provide insight into how changes in stability influence asymmetries in the time to switch between patterns. Suppose the network is currently representing a pattern that is relatively stable and then receives input that votes for a new pattern that is significantly less stable. Because the initial pattern is stable, it requires more work (and more time) to overcome all the activation that is reverberating within the network itself and move to the new pattern. Conversely, switching to a stable pattern is easy (and fast) because once it gets started, the network itself helps out. For example, in the two-unit network described previously, the network will spontaneously move to the ON–ON pattern (it is easy and fast) whereas it would take significant external input to overcome the internal weights and make it stay in an ON–OFF pattern.

References

Becker, S., Moscovitch, M., Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: a computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1059–1082.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*, 115–147.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: simulating semantic priming. *Cognitive Science*, *23*, 371–414.

Farah, M. J., O'Reilly, R. C., & Vecera, S. P. (1993). Dissociated overt and covert recognition as an emergent property of a lesioned neural network. *Psychological Review*, *100*, 571–588.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: beyond similarity. *Psychological Review*, *97*, 3–18.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America – Biological Sciences*, *79*, 2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America – Biological Sciences*, *81*, 3088–3092.

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

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Mozer, M. C., & Behrmann, M. (1990). On the interaction of spatial attention and lexical knowledge: a connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience*, *2*, 96–123.

Mozer, M. C., Halligan, P. W., & Marshall, J. C. (1997). The end of the line for a brain-damaged model of unilateral neglect. *Journal of Cognitive Neuroscience*, *9*, 171–190.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 54–65.

Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, *86*, 161–180.

Ortony, A., Vondruska, R. J., Foss, M. A., & Jones, L. E. (1985). Salience, similes, and the asymmetry of similarity. *Journal of Memory and Language*, *24*, 569–594.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: a case-study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*, 532–547.

Ross, B. H. (1989). Distinguishing types of superficial similarities: different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 456–468.

Ross, B. H., & Kilbane, M. C. (1997). Effects of principle explanation and superficial similarity on analogical mapping in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 427–440.

Smith, E. E. (1995). Categorization. In D. N. Osherson & E. E. Smith (Eds.), *Thinking: an invitation to cognitive science* (2nd ed., Vol. 3, pp. 33–53). Cambridge, MA: MIT Press.

Tanaka, J., Giles, M., Kremen, S., & Simon, V. (1998). Mapping attractor fields in face space: the atypicality bias in face recognition. *Cognition*, *68*, 199–220.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Ullman, S. (1989). Aligning pictorial descriptions: an approach to object recognition. *Cognition*, *32*, 193–254.