Linguistic Relativity and Word Acquisition: A Computational Approach

Eliana Colunga (ECOLUNGA@CS.INDIANA.EDU)
Michael Gasser (GASSER@CS.INDIANA.EDU)
Computer Science Department
Indiana University
Bloomington, IN 47405

Abstract

Language plays a pervasive role in our day-to-day experience and is likely to have an effect on other non-linguistic aspects of life. At the same time, language is itself constrained by the world. In this paper we study this interaction using Playpen, a connectionist model of the acquisition of word meaning. We argue that the interaction between linguistic and non-linguistic categories depends on the pattern of correlations in the world and on their relation to the correlations defined by words. We then discuss three kinds of possible interactions and present simulations of each using Playpen, a neural-network model of the acquisition of word meaning.

Introduction

Language plays a pervasive role in our day-to-day experience. Thus it is no surprise that people wonder to what extent language in general and the particular language one speaks affect the rest of our cognitive abilities. Does language affect thought; that is, do linguistic categories influence general cognitive categories? More generally, how do linguistic and non-linguistic categories interact?

The Linguistic Relativity Hypothesis, associated most closely with Benjamin Lee Whorf (Whorf, 1956), concerns the first question. The claim, in its strongest form, is that linguistic categories exert a direct influence on general cognitive categories. Since Whorf, many researchers have attempted to find evidence for this influence (see Lucy (1996) for a review), but there is as yet no agreement that the evidence has been found.

People have usually studied the effect of language on cognition by looking for differences in adult speakers of different languages. Negative results (Rosch, 1973; Kay & McDaniel, 1978) are met with the arguments that the experiments are biased towards Indo-European languages, are dealing with a part of perception not subject to linguistic effects, or involve irrelevant language distinctions which should not be expected to have an effect in the first place. Positive results (Carroll & Casagrande, 1958; Kay & Kempton, 1984; Bloom, 1981) are generally explained away as effects of culture, biased stimuli or the linguistic nature of the task being used.

More recently, linguistic relativity has been studied in the context of learning, and the picture there looks more promising for linguistic relativity. It has been shown that the order in which children learn certain words, as well as their patterns of overgeneralization, depend on the language being learned (Brown, 1994; Bowerman, 1996), evidence for an effect of language on the rest of cognition. This is further supported by converging evidence from studies showing how learning labels *in the laboratory* can affect children's performance in tasks like word generalization and analogical problem solving (Jones & Smith, 1993; Gentner, Rattermann, Markman, & Kotovsky, 1995). Other developmental studies also show parallels between linguistic and non-linguistic performance in various domains (Jones, Smith, Landau, & Gershkoff-Stowe, 1992; Smith & Sera, 1992).

We believe that learning is the right place to look for relativistic effects, but we also believe that the empirical work on development must be supplemented with a computational account, one which looks at how the demands of linguistic and non-linguistic tasks may lead to long-term effects. In the paper we present the beginnings of such an account. In the next section we discuss the role of correlations in the learning of linguistic and non-linguistic categories. Next we present a computational model and discuss the results of three simulations demonstrating possible kinds of relativistic effects. Finally we consider the implications of the model for future research on linguistic relativity.

Linguistic and Non-Linguistic Correlations

We propose that the way linguistic and non-linguistic categories interact depends on the nature of the correlations in the world and the way these relate to the correlations in the language. During their first year, babies experience the world without any of the biases that are built into language. It is by now clear that they learn a great deal about how the world works during this time (Baillargeon, 1994; Spelke, Breinlinger, Macomber, & Jacobson, 1992). One aspect of this learning is the discovery of correlations between features along different dimensions (Younger, 1990). These non-linguistic correlations define what we will call **non-linguistic categories**. In its second year, the child begins to learn language, which introduces its own categories, defined in terms of the correlations between words and non-linguistic dimensions. The linguistic categories may agree or disagree with the non-linguistic categories. Figure 1 shows some of the ways in which the two sorts of categories can be related to one another.

Given these sorts of correlations, linguistic and non-linguistic cognition could interact at several levels. First, some words could be rendered easier to learn than others. In fact, some categories of words are consistently learned before others, an effect which cannot be explained by frequency. Across languages nouns are learned before verbs (Nelson, Hampson, & Shaw, 1993) and instrument verbs are learned before other verbs (Huttenlocher, Smiley, & Char-

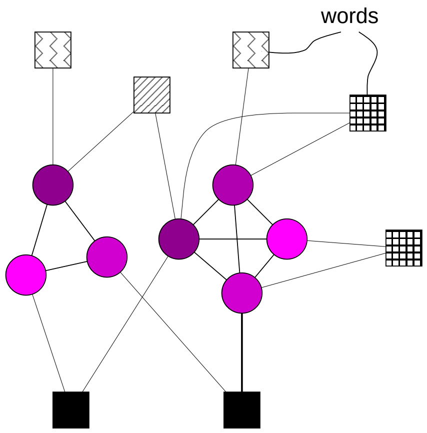

Figure 1: Possible correlations between linguistic and non-linguistic categories. Non-linguistic features, represented by circles, may be associated with words, represented by squares, in such a way that the words agree with non-linguistic categories (squiggly pattern), disagree with non-linguistic categories (solid pattern), subdivide non-linguistic categories (cross-hatched pattern), or combine non-linguistic categories patterns (diagonal hatched pattern).

ney, 1983; Behrend, 1990). These orderings have been explained in terms of the traditional view that categories are formed around strong correlational structure (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Kersten & Billman, 1997). This strength-of-correlations account should also hold for the correlations between words and other perceptual inputs. Words that agree with previously learned categories should be easier to learn than words that disagree with them.

Another place where an effect of language on cognition could be found is in the highlighting (or downplaying) of dimensions that are relevant (or irrelevant) to the language being learned. This effect on attention could be shown in both linguistic and non-linguistic tasks. Hunt and Agnoli (1991) suggest that learning a language that makes a distinction could make speakers of that language more sensitive to that distinction in non-linguistic tasks, a direct effect of language on perception.

An example of a linguistic task where effects of language on attention can be observed is the development of the shape bias. Children at around 18 months of age tend to generalize words to novel objects of the same shape as the exemplar, rather than to novel objects that share size, color or material with the exemplar. Because concrete nouns in most languages are organized mainly in terms of shape, this attentional bias helps them generalize correctly. The shape bias appears only after the child has learned roughly 50 nouns, most of them naming categories of things that are similar in shape (Jones et al., 1992). This suggests that it is the learning of words that drives the learning of the bias. The way in which the linguistic categories correlate with perceptual dimensions apparently causes the learner to attend to particular dimensions, at least in the context of linguistic tasks.

A more dramatic effect of language would be on the nature of the non-linguistic categories themselves. Non-linguistic categories are built up out of correlations between perceptual dimensions. Linguistic categories may agree with these non-linguistic correlations if words correlate with correlating perceptual dimensions. Alternatively, the non-linguistic correlations may be irrelevant for the linguistic categories. The child has both linguistic and non-linguistic tasks to perform. In performing the non-linguistic tasks, the child can rely on non-linguistic correlations, but if the language agrees with these correlations, she can rely on linguistic correlations as well. On the other hand, if the non-linguistic correlations have nothing to do with the language, non-linguistic tasks can only be performed using these correlations. This implies that the strength of the non-linguistic correlations might vary with the languages. While we know of no direct evidence for this possibility, the strong version of the Linguistic Relativity Hypothesis predicts this sort of effect may be found.

In what follows, we provide illustrations in the model of all three sorts of interactions, the influence of the match between linguistic and non-linguistic categories on the relative ease of words, the influence of linguistic categories on attention to perceptual dimensions, and the influence of linguistic categories on the way in which non-linguistic categories are represented.

The Model

Playpen (Gasser & Colunga, 1997) is a connectionist model of the acquisition of word meaning. For the purposes of this paper, the following features of the model are relevant:

1. The network is a generalization of a continuous Hopfield network. Units are updated randomly until the network settles.

2. Network units have relative phase angles in addition to activation, and feature binding is handled through the synchronization of unit phase angles. Units affect each other's phase angles via the weights on the connections joining them.

3. Units are of two types. Micro-object units (MOUs) represent object features. Micro-relation units (MRUs) represent relations between the features of separate objects. Relation words take the form of MRUs. Each MOU has a single phase angle; each MRU has two phase angles, one for each of the objects it relates.

4. Connection weights are adjusted via the contrastive Hebbian learning rule (Movellan, 1990).

5. Non-linguistic features and relation words interact through one or more intermediate layers of MRUs.

Three characteristics of Playpen make it especially well suited to the study of the interaction between language and perception. First, linguistic meaning and non-linguistic concepts are not rigidly distinguished. This is important because, if we are to enter the linguistic relativism debate without any biases, we should not assume from the start that linguistic and non-linguistic concepts are independent. The model allows correlations to develop in the layers of MRUs separating words and non-linguistic perception as learning takes place, and these correlations can have more or less of a linguistic character. Second, the model is designed to deal with relational knowledge. Languages vary more in the way in which they express relational information than in the way they express information about objects, so it is more likely that effects of language will be found in relational words (Gentner & Boroditsky, 1998) While we do not model these properties of relation words, they argue for focusing on relation words as a possible site of relativistic effects, and modeling these effects would require a system capable of handling relations. In particular, a model must be able to learn **relational correlations** (Gasser & Colunga, 1998). Third, the model can be "run" in both the "comprehension" and the "production" directions, allowing for the possibility of mutual effects of language and perception on one another.

Experiments

The three simulations we describe here were based on a set of pre-defined correlations among non-linguistic dimensions and correlations between the non-linguistic dimensions and words. There were two non-linguistic dimensions, and relations within each dimension correlated with relations within the other. That is, a pair of objects with particular values on one dimension tended to have particular values on the other. The relational correlations are shown in Figure 2a.

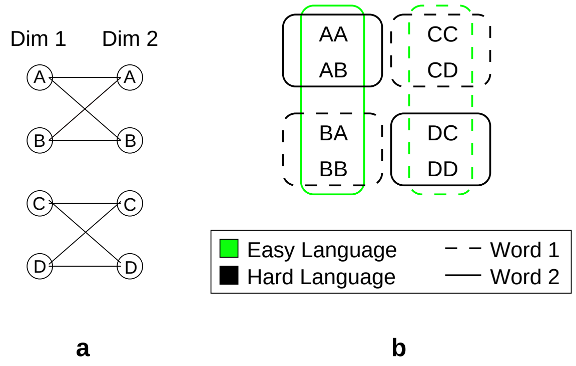

Figure 2: Correlations used in experiments. A, B, C and D represent possible micro-relations between features on Dimensions 1 and 2. (a) The micro-relations are associated with each other across the dimensions in the two clusters shown. (b) Possible pairings of relations on the two dimensions are associated with one or the other of two words. In the Easy language, the words agree with the non-linguistic correlations; in the Hard language, the words correlate only with micro-relations on Dimension 1.

We defined two "languages," an *Easy* language, which agrees with the non-linguistic correlations, and a *Hard* language, which disagrees with the non-linguistic correlations, as shown in Figure 2b. Each language consists of two relational words. For the *Easy* language, the categories in the world agree with the categories promoted by the words. That is, each of the two correlational clusters existing in the world is associated with one of the words in the language. For the *Hard* language, the words cut across the two prelinguistic correlational clusters in such a way that the word describing a pair of values along the two dimensions is determined by the value along Dimension 1 only. For example, according to the pattern of correlations between dimensions in Figure 2, the pairs of values labeled A-B and B-A should be in the same category but they are assigned to different words in the Hard language. That is, the value along Dimension 2 is not predictive of the linguistic category.

The architecture of the networks used in these simulations is shown in Figure 3.

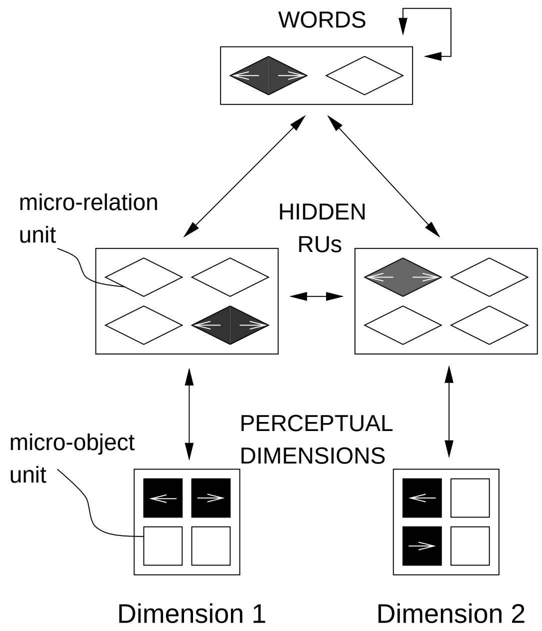

Figure 3: Network architecture. Micro-object units are represented by squares, micro-relation units by diamonds. Arrows indicate complete connectivity between layers. Each Hidden MRU is associated with a pair of Perceptual Dimension MOUs. A possible pattern across the network is shown. Darkness indicates activation, and arrow direction indicates relative phase angle.

The networks are trained and tested on two different tasks. For **Non-linguistic Pattern Completion**, they are presented with a pattern on one of the Perceptual Dimensions and expected to produce an appropriate pattern on the other. (Note that there are always two possibilities for the appropriate pattern.) The network can learn to solve this task using the connections joining the Perceptual Dimension and Hidden Relation layers or the connections between the two Hidden Relation layers. For **Production**, the networks are presented with a pattern on the Perceptual Dimensions and expected to output a word.

Experiment 1 - Difficulty of languages

The goal of this experiment is to see how the different correlational patterns both between dimensions and with the words affect the difficulty of learning the two languages. The networks were first trained in a Pre-linguistic Phase on Non-linguistic Pattern Completion alone for 30 repetitions of the relevant training patterns (epochs). Next, during a Linguistic Phase, Pattern Completion training was discontinued, and the networks were trained on Production for seven epochs. We predict that the Easy language will be learned faster than the Hard language during the Production phase because the Easy language categories agreed with the non-linguistic categories.

During the Pre-linguistic Phase, the networks mastered the Pattern Completion task by learning weights between the two Hidden layers representing the non-linguistic correlations. Results for the Linguistic Phase are shown in Figure 4.

The data were submitted to a 2(Language) $*$ 7(Epoch) anal-

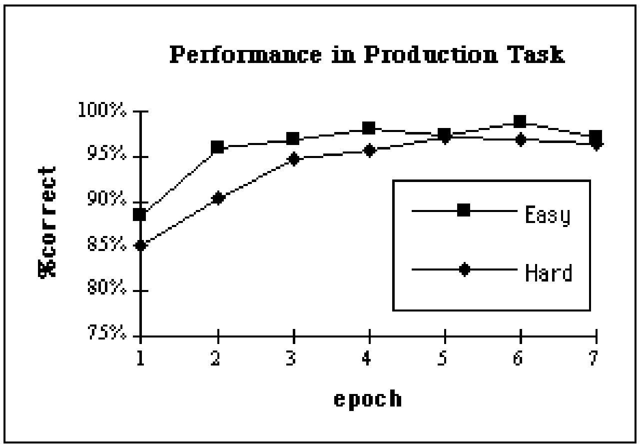

Figure 4: Results for Experiment 1. The Hard language is harder to learn than the Easy language.

ysis of variance for a mixed design. This analysis revealed a main effect of epoch, indicating that the networks get better as they receive more training. More importantly, as predicted, there is a main effect of language ($p < .001$). Thus, the Easy language is learned faster than the Hard language, although by the end of the training the two networks have comparable performance. No interactions between language and epoch were found.

The results make sense for two reasons. Given the correlations between the Perceptual Dimensions and the two language situations, a network learning the Easy language could choose to attend to Dimension 1, to Dimension 2, to both dimensions or even to different dimensions for different values along the dimensions. In contrast, to learn the Hard language the network needs to attend to Dimension 1 *and* ignore Dimension 2. Since the space of possible good solutions is larger for the Easy than for the Hard language, the words in the *Hard* language to be harder to learn than the words in the *Easy* language.

A second reason for the ease of the Easy language concerns the effect of the non-linguistic correlations on language learning. In the case of the Easy language, learning the right association between *one* perceptual input and its corresponding word should improve the chances of producing the right word for the other instances of that word. This is because of the previously existing correlations. At the beginning of Production training, any associations between a Hidden unit and the Word layer indirectly affect the other Hidden units involved in non-linguistic clusters with that Hidden unit. In the case of the Easy language, the correlations help since linguistic and non-linguistic categories agree; in the case of the Hard language, the correlations fail to help solve the Production task.

This experiment demonstrated how words can differ in ease of learning to the extent that they agree with non-linguistic categories. That is, given a particular set of perceptual dimensions, for example, the set of dimensions that is learned relatively early because of its salience or importance to the child, words will differ in the degree to which those dimensions define them. And this difference will lead to differences in ease of learning. The comparison holds within languages as well as across languages. If this is true, this would explain the facilitated learning of instrument verbs over other verbs. The instrument together with the action form a tight correlational cluster that is likely to be there prelinguistically. In contrast, for a more abstract verb, for example, *enter*, a child would have to concentrate on the one thing that matters (path) and ignore the other aspects of the situation to which the word applies. This is also consistent with findings that in Tzeltal, a Mayan language with an apparently complicated system for expressing spatial relations, context-specific spatial relation words are learned earlier than the more abstract spatial prepositions (Brown, 1994).

Experiment 2 - Highlighting dimensions

The goal of the second simulation is to verify that the networks trained on the Hard language do in fact pay more attention to the relevant than to the irrelevant dimension. To test this we presented the trained networks with novel perceptual input patterns. We predict that the networks trained on the Hard language will produce the word which is consistent with the relevant dimension (Dimension 1), while those trained on the Easy language should show no such preference. For example, in Figure 2, if the network is given the values for A in Dim1 and C in Dim1, the networks trained on the Hard language should tend to output Word 1, because only the pattern on Dim1 (A) counts. In the same situation networks trained on the Easy language could output either Word 1 (consistent with AA and AB) or Word 2 (consistent with CC or DC).

Networks were first trained in the Pre-linguistic Phase, then in the Linguistic Phase for 7 epochs of training on the Production patterns. We were only concerned with the performance following this training. To compare the performance of the networks, we subtracted the number of words agreeing with Dimension 2 from the number of words agreeing with Dimension 1. Thus a positive result indicates a preference for Dimension 1, a negative result a preference for Dimension 2.

The results are shown in Figure 5. A T-test revealed that, as expected, the networks trained on the Easy language had a different preference pattern from those trained on the Hard language. In fact, the networks trained on the Easy language showed no preference for either word while the networks trained on the Hard language showed a significant preference for the words consistent with Dimension 1.

Experiment 3 - Effect of language on non-linguistic categories

The goal of Experiment 3 is to determine whether the difference in the two languages can have an effect on the way in which the network learns the correlations between the Perceptual Dimensions. During pre-linguistic training in Experiments 1 and 2, the networks readily learned the weights on the connections joining the two Hidden layers representing these correlations. Since each hidden unit is associated with a pair of values along one of the Perceptual Dimensions, these weights are easily interpreted. Pre-linguistic training results in positive weights on each of the connections joining Hidden-layer MRUs representing pairs of perceptual features which correlate and negative weights on the other connections. As can be seen from Figure 6a, there are eight correlating pairs; hence eight of the weights joining the Hidden

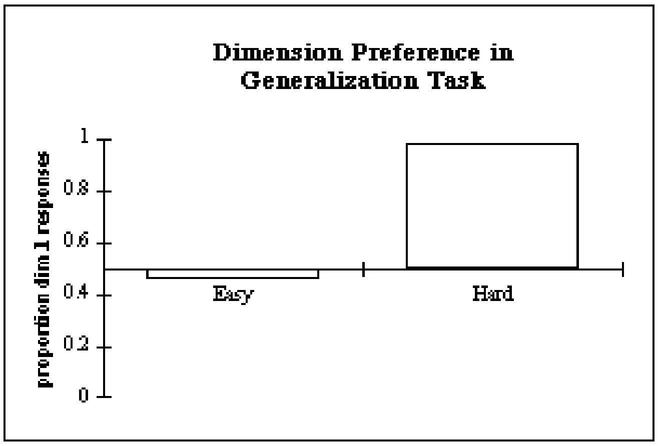

Figure 5: Results for Experiment 2. The networks trained on the Hard language responded with the word consistent with Dimension 1 97% of the time, while networks trained on the Easy language showed no preference.

layers are positive, while the other eight weights are negative. For example, the weight on the connection joining the hidden units representing A on Dimension 1 and B on Dimension 2 is positive, while the weight for A on Dimension 1 and C on Dimension 2 is negative.

In Experiment 3, rather than training the networks on the Non-Linguistic Pattern Completion task before training on the Production task, we trained them on the two tasks simultaneously by alternating between the two tasks.

For the Hard network, the two tasks must be solved using completely different weights. To learn to produce the correct word, the network must rely on the connections from the Dimension 1 Hidden layer to the Words layer. To learn to perform the Pattern Completion task, it needs to learn the inter-Hidden-layer correlation weights.

For the Easy network, on the other hand, because the linguistic and non-linguistic correlations agree, the two tasks can make use of the same weights. In particular, there are two ways in which the network could learn to solve the Pattern Completion task. It could make use of the inter-Hidden-layer correlation weights, as we expect in the Hard network. Alternately, it could rely on the Hidden-to-Words connections, using the word as a bridge between the two dimensions. These two paths are shown in Figure 6. Because the Easy network can perform the Pattern Completion task without the inter-Hidden-layer weights if it has the Hidden-to-Words weights, and because it needs the Hidden-to-Words weights anyway to solve the Production task, we predict the inter-Hidden-layer correlation weights will be smaller in the Easy than in the Hard network.

We trained 10 networks each on the Easy and Hard set of patterns, alternating Production and Pattern Completion tasks. For this experiment, we are interested only in the inter-Hidden-layer weights that resulted during training, not in the performance of the networks on the tasks. After four epochs of training, we compared the correlation weights for the A and B input patterns, that is, the A-A, A-B, B-A, and B-B inter-Hidden-layer weights, for the Easy and Hard networks. As we expected, the weights in the Hard network were significantly larger ($p < .02$) than the weights in the Easy network. This shows that the nature of the linguistic categories can directly influence the weights representing the non-linguistic correlations.

In this experiment we showed how the kind of language being learned can affect the way the same information is learned. More importantly, the same task was solved with or without linguistic knowledge depending on the correlation patterns between the words and the world. This points out a flaw in one of the most frequent complaints about relativism experiments, namely that whenever a cross-linguistic difference is found, the task is declared to be linguistic in nature. In our illustration, both networks solve the same task using different parts of the architecture. There was no behavioral difference between the two networks in either of the two tasks they were trained on and yet their representation of the knowledge necessary to solve the tasks was different. We think this is a direct effect of the structure of the languages being learned by the networks on cognition. This suggests that it is not the *task* that makes the process linguistic or non-linguistic, and to a certain extent, that it could be the structure of the language that does. The fact that we found no behavioral differences reflecting the weight differences in the networks should not be discouraging. Brain scan studies could be performed on people to look for effects analogous to the weight differences in the networks. Also, preliminary results show language effects during the course of learning suggesting that that is a good place to start looking for evidence for relativism.

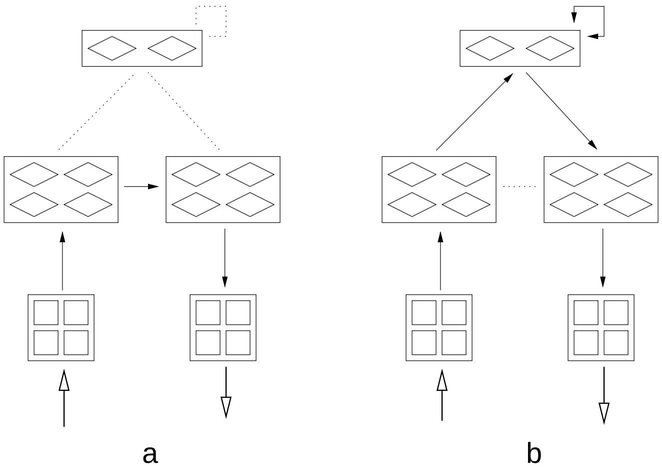

Figure 6: Two paths for performing Non-Linguistic Pattern Completion. (a) The network uses the between-Hidden-layer connections representing the correlations between the Perceptual Dimensions. This is possible with both the Easy and Hard networks. (b) The network uses the Hidden-to-Words connections. This is possible only with the Easy Network.

Conclusions

In this paper we have argued that linguistic relativity can be best studied in terms of the correlations between different perceptual dimensions, the correlations between linguistic categories and perceptual dimensions, and the way in which these correlations interact during the learning of language and of non-linguistic tasks. We focused on three specific relativistic effects and showed how each of these could be simulated with a simple neural-network model of word learning. We believe that such a model is crucial to the relativity debate. Without an explicit account of how the learning of linguistic and non-linguistic categories depends on different kinds of correlations, it will remain unclear precisely what linguistic relativity might mean for cognition.

References

Baillargeon, R. (1994). How do infants learn about the physical world?. *Current Directions in Psychological Science*, *3*, 133–140.

Behrend, D. (1990). The development of verb concepts: children's use of verbs to label familiar and novel events. *Child Development*, *61*, 681–696.

Bloom, A. (1981). *The linguistic shaping of thought: a study in the impact of language on thinking in China and the West*. Lawrence Erlbaum, Hillsdale, NJ.

Bowerman, M. (1996). Learning how to structure space for language: a crosslinguistic perspective. In Bloom, P., Peterson, M. A., Nadel, L., & Garrett, M. F. (Eds.), *Language and Space*, pp. 385–436. MIT Press, Cambridge, MA.

Brown, P. (1994). The ins and ons of Tzeltal locative expressions: the semantics of static descriptions of location. *Linguistics*, *32*, 743–790.

Carroll, J. & Casagrande, J. (1958). The function of language classifications in behavior. In Maccoby, E., Newcomb, T., & Hartley, E. (Eds.), *Readings in social psychology*, pp. 18–31. Henry Holt, New York.

Gasser, M. & Colunga, E. (1997). Playpen: toward an architecture for modeling the development of spatial cognition. Tech. rep. 195, Indiana University, Cognitive Science Program, Bloomington, IN.

Gasser, M. & Colunga, E. (1998). Where do relations come from?. Tech. rep. 221, Indiana University, Cognitive Science Program, Bloomington, IN.

Gentner, D. & Boroditsky, L. (1998). *Language acquisition and conceptual development*, chap. Individuation, Relativity and Early Word Learning. Cambridge University Press, England.

Gentner, D., Rattermann, M. J., Markman, A., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In Simon, T. J. & Halford, G. S. (Eds.), *Developing Cognitive Competence: New approaches to process modeling*, pp. 263–313. Lawrence Erlbaum Associate, Hillsdale, NJ.

Hunt, E. & Agnoli, F. (1991). The Whorfian hypothesis: a cognitive psychology perspective. *Psychological Review*, *98*, 377–389.

Huttenlocher, J., Smiley, P., & Charney, R. (1983). Emergence of action categories in the child: evidence from verb meanings. *Psychological Review*, *90*, 72–93.

Jones, S. S. & Smith, L. B. (1993). The place of perceptions in children's concepts. *Cognitive Development*, *8*, 113–140.

Jones, S., Smith, L., Landau, B., & Gershkoff-Stowe, L. (1992). The origins of the shape bias. Boston. Boston Child Language Conference.

Kay, P. & Kempton, W. (1984). What is the Sapir-Whorf hypothesis?. *American Anthropologist*, *86*, 65–79.

Kay, P. & McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*, 610–646.

Kersten, A. W. & Billman, D. (1997). Event category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*(3), 658–681.

Lucy, J. A. (1996). The scope of linguistic relativity: an analysis and review of empirical research. In Gumperz, J. J. & Levinson, S. C. (Eds.), *Rethinking Linguistic Relativity*, pp. 37–69. Cambridge University Press, Cambridge.

Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In Touretzky, D., Elman, J., Sejnowski, T., & Hinton, G. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*, pp. 10–17. Morgan Kaufmann, San Mateo, CA.

Nelson, K., Hampson, J., & Shaw, L. K. (1993). Nouns in early lexicons: evidence, explanations and implications. *Journal of Child Language*, *20*, 61–84.

Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *7*, 573–605.

Smith, L. B. & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*(1), 99–142.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*, 605–632.

Whorf, B. L. (1956). *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf.* MIT Press, Cambridge, MA.

Younger, B. (1990). Infants' detection of correlations among feature categories. *Child Development*, *61*, 614–620.