Categorization in early language acquisition - accounts from a connectionist model

Georg Dorffner
Dept. of Medical Cybernetics and Artificial Intelligence
University of Vienna
and
Austrian Research Institute for Artificial Intelligence

Abstract

In this paper we introduce a connectionist model for early word learning as an important part of language acquisition. In the spirit of previous connectionist models (e.g. [Plunkett et al. 1993]) we evaluate the performance of the model with respect to important phenomena such as over- and underextensions, the comprehension over production asymmetry, and the naming insight leading to vocabulary spurts. The main motivation behind the model – based on extensions of the well-known competitive learning paradigm – is the focus on categorization as a major foundation of language and language learning. We give a detailed motivation of this approach and demonstrate how this model can help in identifying the roots of some of the phenomena in word learning. We also discuss the relationship between this modeling approach and more "conventional" connectionist models based on multilayer perceptrons and backpropagation, and how our model can overcome some of the apparent shortcomings of those models.

1 Introduction

In this paper we introduce a connectionist model of the acquisition of early word meanings, most prominently nouns. This model is centered around the important aspect of *categorization* and its role in language learning. Inspired by cognitive linguistics, it puts large emphasis on the structure of human category systems – such as prototype effects, basic level effects, or taxonomical relations – and the interplay between input-based categorization ("unsupervised learning") and the role of language in forming categories. Contrary to many previous models on the same topic, we use variants of competitive learning as the basis for the model. We will give a detailed motivation behind this modeling approach and discuss its contributions to the understanding of language learning, vis a vis the previously published results using MLPs (e.g. [Plunkett et al. 1993]).

Similar to work like that by [Plunkett et al. 1993], the model will be evaluated with respect to its replication of well-known phenomena in early word learning in children, such as

- the asymmetry between comprehension and production in learning
- over- and underextensions of word meaning
- the "naming insight" and vocabulary spurts

Since one of the goals of any model of linguistic aspects is to provide the basis (or, in some sense, a "vocabulary") for understanding and explaining phenomena like these, we will discuss in detail what the main underlying reasons for them might be, according to literature, and how the model can help put a finger on some of their aspects. In this critical evaluation, the limits of such a model will also become clear.

Most recent connectionist models of language acquisition have focused on distributed representations in multilayer perceptrons (MLP, "backpropagation networks"). Their hypotheses focus on the power of such representations in expressing similarities between concepts and the interactions between representations, as well as the famous "one process" argument by [Rumelhart & McClelland 1986, Seidenberg & McClelland 1989] that regular *and* irregular linguistic forms are processed by the same mechanisms, namely the associative mapping between patterns. These approaches also neglect the existence of a separate lexicon, explaining entire phenomena involving words by the mechanisms of activation state spaces.

As compared to MLP models, we want to argue that the strengths of our approach can be found, among others, in the following aspects:

- the ability to model complex categorical structures explicitely, and to study their role in language learning
- the ability to learn in a much more stable manner without unwanted inferences between representations, which can often be found in MLPs (e.g. [McCloskey & Cohen 1989])
- the capability for fast learning, and learning of new words without interference with previously learned mappings
- the ability to combine unsupervised and supervised learning, as proposed by cognitive linguistics

The simple model introduced in this paper exhibitis interesting properties compared to psycholinguistic data, at the same time overcoming some of the shortcomings of distributed MLP models, discussed below. We will argue for a reformulation of connectionist models of language learning involving representations which are less distributed in the classical sense but nevertheless can exhibit the necessary associative cross-dependencies between patterns. The ultimate goal is to unite the two approaches in order to make models of language acquisition more powerful.

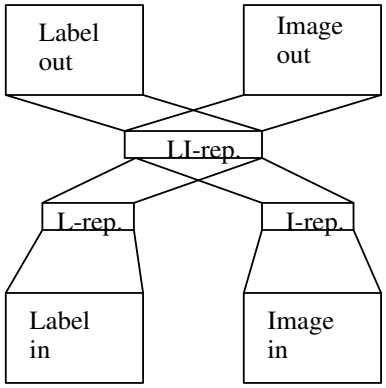

Figure 1: An outline of the model by Chauvin and Plunkett et al. consisting of a multilayer perceptrons with two separate and one common hidden layer

2 Previous models of word learning

Several connectionist models for aspects of word learning and language acquisition have been proposed in literature. [Cottrell et al. 1990], for instance, describe a model consisting of two separate multilayer perceptrons, each auto-associating a visual (image) and a locally encoded name pattern with each other, respectively, so as to form internal distributed representations of those patterns in the hidden layers. These hidden layers are then viewed as input and output to a third multilayer perceptron, learning to associate them with each other. After learning, a word pattern can be input and the corresponding image activated, and vice versa. Although no extensive psycholinguistic analysis is made, it is hinted that this kind of model can be seen as a first step toward a connectionist theory of meaning ("grounding meaning to perception").

[Chauvin 1989] and, in an extension, [Plunkett et al. 1993] also use two separate input in an autoassociative mode, but unite the two components into one multilayer perceptron, trained with backpropagation [Rumelhart et al. 1986]. They do this by introducing another common hidden layer, via which the autoassociations of both inputs onto themselves are learned (fig. 1). By this they suggest that learning of both patterns not only achieves a cross-association between them (permitting, similarly to [Cottrell et al. 1990], the activation of one pattern given the other), but also influence each other in an intricate, distributed way. 32 classes of binary visual patterns, with 5 random distortions each, were used as input on the visual side (preprocessed by projecting them onto a simple "retina"). Like in [Cottrell et al. 1990] orthogonal patterns (local representations) are used as label input. Their model exhibits many interesting effects, such as a vocabulary spurt (the rapid increase of learnt association between labels and visual patterns after a relativly slow progress), the temporal precedence of comprehension over production, and a prototype effect in the sense that several instances of a pattern category can lead to the generalization onto their prototype, although it had never been presented during training.

[Sales 1995] presents further analysis of this type of model.

Another interesting model is that by [Schyns 1991], who uses selforganizing feature maps to perform explicit categorization of sensory patterns. He suggests a separation of categorization and subsequent naming, reflected in separate components to achieve the two mechanisms.

[Grossberg & Stone 1986] propose a model based on ART modules which encompasses word learning from perception and production to meaning. However, they do not present simulations of their model and thus its value with respect to psycholinguistic data cannot be evaluated. Nevertheless, it presents one example corresponding closely to what we are suggesting.

3 Problems with MLP models

Multilayer perceptrons (MLPs), consisting of an input layer, an output layer, and one or several hidden layers, cannot only be found in the domain of lexical learning, but are the basis of a great many other connectionist models for aspects of language acquisition. Examples are the models by [Rumelhart & McClelland 1986] for the acquisition of the past tense, by [Seidenberg & McClelland 1989] for learning the mapping between graphemes and sounds, or by [Gasser & Lee 1990] for the acquisition of rudimentary phonology and syllable structures. The appeal of these models comes from several interesting properties:

- such models develop representations in the hidden layer(s), usually termed *distributed*, which reflect similarities between them based on both common input features and common output characteristics. [Hinton et al. 1993], for instance, have argued that such representations give rise for a plausible explanation for phenomena like acquired deep dyslexia, in that the model exhibits semantic errors in reading when applying unspecific lesions in deep layers of the model. For instance, people with deep dyslexia who see the word 'uncle' would often say 'nephew' when asked to read the word aloud. Such a semantic "mix-up" can be explained when resorting to representations where semantically related concepts such as *uncle* and *nephew* are closer to each other (in activation state space) than to other, unrelated, concepts.

- such models can perform linguistic mappings involving both regular and irregular cases without resorting to two separate mechanisms. [Rumelhart & McClelland 1986], for instance, have presented such a model for the acquisition of the mapping between the present tense form of a verb and its past tense. Regular mappings (e.g. 'talk' $\rightarrow$ 'talked') can be learned by the network side by side with irregular ones (e.g. 'ride' $\rightarrow$ 'rode'), wheras traditional approaches postulated two separate processes for the two (i.e. rules for regulars, lexical look-up for irregulars). Similarly, [Seidenberg & McClelland 1989] have presented such a "one-process" model for the mapping between the orthographic representation of a word and its pronunciation. They have also explained certain forms of dyslexia which affects irregulars more than

regulars by the effect of limited resources (hidden units) to form the underlying internal representations.

- such models are *similarity-sensitive*, meaning that they tend to exhibit similar responses to similar inputs (unless explicitly trained not to), naturally giving rise to generalization to new inputs. In a later version of their model, Seidenberg and McClelland, for instance, have demonstrated how an extension of their previous model could correctly process words the network had never seen during learning.

However, such models appear to show equally many problems with respect to plausbile learning:

- they tend to show *"catastrophic interference"* [McCloskey & Cohen 1989], meaning that learning new patterns can destroy or even erase previously learned mappings to an extent which is not psychologically plausible. This is related to the *stability–plasticity dilemma* raised by [Grossberg 1987]

- they are generally not able to learn fast, i.e. based on two or three presentations of a pattern – although fast learning is a phenomenon widely observed in human learning

- they show problems with respect to aspects of human category learning (see, for instance, [Kruschke 1993]).

- they are basically designed to implement one-to-one mappings and cannot be easily extended to many-to-one mappings (e.g., in language, the mappings of homonyms or synonyms to their meanings)

- they are very difficult to scale up in size

- their representations often appear to be too dependent on the particular learning context the network has been trained in (see, e.g., [Clark 92])

- their representations are highly implicit and often rather inaccessible, making them hard to control (e.g. to tune a model to the characteristics of psycholinguistic data) and difficult to sort out the reasons for a particular learning behavior.

Our model, described in the next section, attempts to address some of these problems and to overcome them by relying on quite different representations. The main motivation for this model is the important process of categorization, but as a "side-effect" it turns out to be not as prone to the above difficulties as MLPs trained with backpropagation.

4 Our model for the acquisition of early word meanings

The basic component for our approach is a model for categorization of feature vectors based on a variant of competitive learning. We will give a detailed motivation of this approach and argue that rather than replace, this model can complement more "traditional" neural network approaches based on multilayer perceptrons. We will introduce the categorization module and the full-blown model based on two such components and a module forming links between them (in the spirit of models like [Cottrell et al. 1990, Chauvin 1989, Plunkett et al. 1993]) separately, starting with a discussion on categorization.

4.1 Categorization in language

The model for word learning discussed here – an extension and more mature version of previously published work [Dorffner 1989, Dorffner 1992a, Dorffner 1992b] – is centered around the notion of *categorization*, or in a more sophisticated form, *concept formation*. The assumption is that words categorize the world, and categories determine what words can mean. This is a two-way process, viewed from the cognitive linguistic standpoint (see, e.g.[Taylor 1989]). According to this relatively recent, but more and more widely accepted, view language meaning is determined by at least three factors:

- inherent structures in perceived stimuli which permit carving out categories even without reinforcement or feedback through a teacher,
- the internal structure of the cognitive apparatus, set in its environment, that favors or permits the recognition of certain invariances in stimuli,
- language itself, which through its expressions selects situations or stimuli to belong to common categories.

In contrast, the structuralist linguistic tradition has largely put forward the idea that the way linguistic expressions carve up the world is more or less arbitrary, meaning that any categorization would be possible, such that it must be learned completely through the acquisition of language.

A prototypical example to illustrate this point is the perception and naming of colors (see, for instance, [Taylor 1989] for a discussion and literature overview). Given the fact that different languages have different color terms, and even different numbers of color terms, many structuralists have proposed that the continuous frequency spectrum is cut up into classes completely arbitrarily by a given language. Cognitive linguistics, on the other hand, based on recent findings on perception and the visual system, as well as the natural occurences of colors, proposes a view in which both the way the eye can perceive colors and the predominant roles of colors (e.g. blue and green) in our environment restricts categorization to a small number of possible schemes, from which language must select one.

Hardly anyone would contest the view that categorization is a central process in both cognition in general and in language. Cognitive linguistics (for an introduction, see [Langacker 1987]) now offers a theory of human categories that largely goes against more classical views, also widely held in structuralist linguistics. The classical view of a category is a set of objects or situations which have one or several semantic features in common. Dating back to [Wittgenstein 53], via the seminal work by [Rosch 1978] to more recent work by [Lakoff 1987] or [Taylor 1989], many researchers have proposed, and given strong evidence for, a different and intricate framework for how categories are structured, including the following aspects:

- membership in a category is *graded*, ranging from rather central members – members close to so-called *prototypes* – to border-line cases, which, depending on the context, can more or less easily become members of a different, adjacent, category.

- categories can consist of members that do not all share a single common feature, i.e. the intersection of all their feature sets can be empty. Instead there is mutual overlap with the prototype or among each other. Such a structure is frequently termed "family resemblances" [Wittgenstein 53].

- categories can be extended through such mututal overlaps in a chain effect (so-called *radial categories* [Lakoff 1987]), including members which do not share any features with the more central members.

- certain features can come to represent the whole category, leading to further extensions and chain effects. This is called *metonymy*.

- categories can be hierarchically structured in *taxonomies* of sub- and super-categories (e.g. *dog*[1], *poodle* and *animal*), with a distinct level on which categorization comes most natural based on perceptual or other similarities. This level is called *basic level* [Rosch 1978]. While structures in the environment and the cognitive apparatus are dominant in deciding upon class membership on the basic level, reinforcement through language (or other mechanisms) seems to be required for other levels (compare [Nelson 1988, Benelli 1988]).

While some of these phenomena are usually used to explain the diachronic development of language meaning (e.g. why 'pupil' today can refer both to student and to the opening in one's eye [Taylor 1989], p. 103), most of them are also active aspects of categorization during language learning, giving rise to observation such as overextensions in word usage in young children. [Clark 1993], for instance, reports about an example where one child "extended 'moon' to a half-grapefruit seen from below, to a lemon slice, to a dial on a dishwasher, to a shiny leaf, crescent-shaped paper, the inside of a lampshade, pictures of vegetables on the wall, circles on a wall-hanging, and so on" ([Bowerman 1978, Dromi 1987],

[1]Throughout this paper we write concepts or features in italics (e.g. *dog* or *roundness*), and words under single quotes (e.g. 'dog').

as cited in [Clark 1993], p.35.). Obviously, different features like crescent or round shape, and luminance, were used to extend the category, leading to members which do not all share the same features.

These observations have led us to build a connectionist model which primarily focusses on categorization with these properties, both in an "unsupervised" (i.e. given perceptual similarities and the architecture of the model alone) and "supervised" manner (primarily using language input as a "teacher"). The basics of this model are the ideas raised in several models of *competitive learning* [Rumelhart & Zipser 1985, Grossberg 1987], which can cluster feature vectors in an unsupervised or supervised [Carpenter et al. 1991] mode. The major extension to these models concerns the replacement of the usual "winner-take-all" mechanism which selects the most active unit in competition and sets it to maximum activation, by a graded mechanism of training both the weights attending to the features and the inhibition between the competitive units. This way, graded membership can be expressed by a value ("goodness-of-categorization" or "goodness-of-fit") describing the sucess of competition in leading to a winning unit supressing all the others.

Before describing this categorization model in more detail, we would like to discuss the major modeling assumptions behind using such an approach. Categorization is usually said to be a process of *abstraction* by "treating all members of a category the same way" [Neisser 1987]. That is, although the decision which stimulus should be classified in which category is based on similarities, the result should be a representation that abstracts these similarities away, including – as we want to argue – the ones between categories. In other words, categorization is a process making representations *dissimilar* to each other, even though the stimuli of different categories might have shown some similarity. For instance, when classifying an object as *cat* one abstracts away the similarities to, say, *dog*. There are several motivations for why this should be so:

- **Distinct responses:** Learned or required responses to categories can be quite distinct (e.g. *pet* or *run away*). Even in low levels of intelligence, the learning of such distinct responses is supported or eased when the representations of categories do not exhibit large similarities.

- **Fast learning:** Learning of responses often is or should be rather fast, i.e. cannot rely on multiple training cycles. If category representations are dissimilar, such fast learning is greatly enhanced (see below).

- **Arbitrary categories:** When introducing categories through language, similarities might not be given in terms of features, but solely by the fact that several members are named by the same expression. If similarities among category representations would be assumed, such arbitrary categories would not fit in easily with the others.

- **Categories mistaken as symbols:** Another evidence for the hypothesis that category representations should be dissimilar to each other is the common mistake to view categories (concepts) as symbols. The latter are arbitrary signs in the sense that similarities in their form does not reflect similarities in their meaning (see section 8).

We have argued elsewhere [Dorffner et al. 1993] that concepts and symbols are distinct entities, but the frequent mix-up points to some underlying representations that are quite distinct from each other.

Behind these arguments already lies a distinct connectionist view of things. When speaking of 'similarities', we mean vector similarities a connectionist network can handle and react sensitive to. 'Dissimilar', on the other hand, refers to two vectors which a connectionist learning algorithm can treat independently without interference between them. This is how and why this modeling approach differs from most connectionist models based on associative networks like perceptrons or multilayer perceptrons (MLPs). Consider a simple two-layer structure with feedforward connections (perceptron), where one layer contains a conceptual representation and the other layer contains a pattern corresponding to a response that is to be learnt. If the representations of different categories (e.g. *cat* and *dog*) in the first layer are distributed to express similarities (as in [Hinton et al. 1993] – it could also be the hidden layer of an MLP), distinct responses in the second layer cannot be learnt in a single step. Instead, they require repeated training cycles to gradually erase the interferences due to the similarities. If, however, the representations of two categories are *orthogonal* to each other (in vector terms), single-step learning is possible, when necessary. We will see below that this property, along with others, is rather important for language learning.

Based on these observations, we propose a representation scheme for categories in which *orthogonality* between categories is the important property. The activation states resulting from competitive learning ("unitized" patterns with a single unit highly active) are the most simple example of patterns which fulfill this requirement (since all such vectors are unity vectors, by default orthogonal to each other).

Although, at first sight, this proposal seems to argue against the achievements of connectionist models using distributed representations, it should rather be seen as an important extension to that view. One of the most interesting properties of connectionist networks is that they are *similarity-sensitive*, i.e. they can learn and generalize based on overlaps between representations. This property is exploited in models making use of explicit (hand-coded) or implicit (self-organized) distributed representations. We would never deny the usefulness of this aspect. After all, a large part of human and animal cognition indeed seems to be based on similarity-sensitivity – the ability to recognize similarities between preceived situations in order to act intelligently in a world where never the exact same perceptual pattern appears more than once. Nevertheless, on top of that, we propose that categorization is a process that, at the moment of activating a category, erases similarities in the corresponding representations. For reasons above, and more given below in the special context of language, such a process offers decisive advantages for a cognitive being.

As will be exemplified in the detailed description below, the model includes (at least) two levels (see fig. 2):

- a level of feature vectors, as a result of perception and action

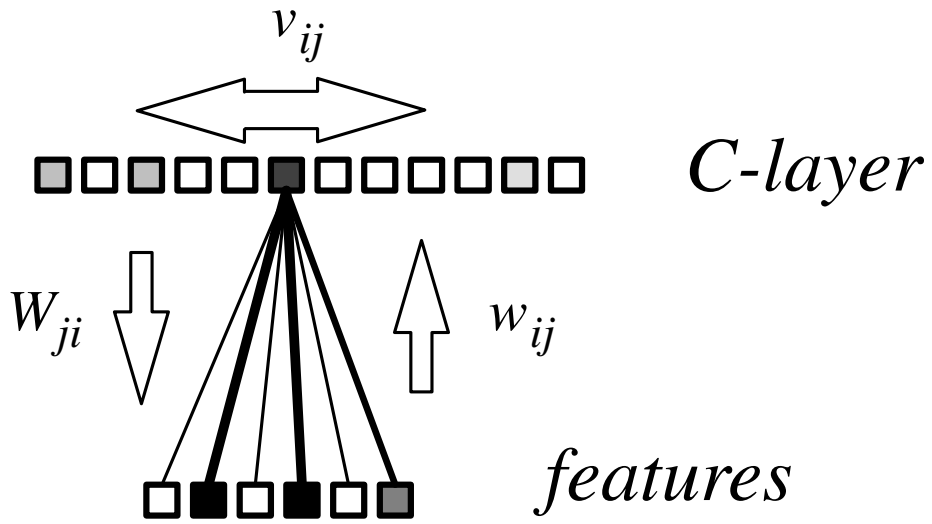

Figure 2: The basic architecture of the categorization model. A input layer containing feature vectors is connected to a categorization lazer (C-layer) in both direction. Through an interactive activation rule, the units compete in the C-layer. Learning is done through instar, creating a prototzpe of the category in the corresponding weights. Inhibition between units in the C-layer is also increased

- a level of category representations (in a layer called *C-layer*), which – among others – is the interface to language, or, in a sense, the substratum for the meaning of expressions

It is the level of feature vectors which is distributed and permits similarity-based overlap between representations. On top of that, the C-layer leads to orthogonal vectors for different category representations, the reasons for which are listed above.

Finally it should be mentioned that MLPs, or the representations in their hidden layers, can exhibit categorization through orthogonalization, as well. Especially when trained in the autoassociative mode (as in [Plunkett et al. 1993]) it can be shown that linear variants of MLPs, and in some sense also their usual non-linear instantiations perform a so-called *principal component analysis* (see [Baldi & Hornik 1989]) leading to a transformation of the coordinate system onto one which is aligned along the largest variances in the data, and the axes of which are naturally orthogonal to each other. This property, among others, is discussed by [Plunkett et al. 1993] when interpreting their model. If data is clustered along the principal components the resulting representations will tend to be orthogonal (but not necessarily unitized). However, this orthogonalization process is *implicit*, i.e. not explicitely induced by designing the architecture, and thus not under complete control by the modeler – and often not intended, since distributed representations in hidden layers are often aimed at, preserving similarities between representations. Nevertheless, most of

what has been said about orthogonal representations (and some more arguments below) would apply to orthogonal patterns in hidden layers which are not unit vectors, as well.

4.2 The connectionist categorization model

As opposed to other competitive learning models, categorization here is seen as a process that is developed gradually during learning. Orthogonality, as discussed above, should only be the limit state of learning, being only approximated in most cases. Instead of applying a strict winner-take-all (WTA) process (which sets the most active unit to maximum activation, and all others to zero), competition is achieved through inhibitive connections and a "rich-get-richer" effect. Through learning the ability of the winner to suppress the other units' activations is enhanced. This is a mechanism very similar to the ones implemented in the ART 3 model [Carpenter & Grossberg 1990]. Compared to ART3, in our model the mechanisms are greatly simplified, mainly due to the reason that we do not pay much attention to biological plausibility. The update rule used is known as *shunting equation* [Grossberg 1982] or *interactive activation* [McClelland & Rumelhart 1981].

$$x_j(t+1) = \begin{cases} x_j(t) + (1 - x_j(t))y_j(t) & \text{if } y_j > 0 \\ x_j(t) + x_j(t)y_j(t) & \text{otherwise} \end{cases} \tag{1}$$

where x is the activation of a unit; y is the so-called net input (the input to the unit before applying this rule) and is computed as

$$y_j = \sum_{i=1}^{N} w_{ij}x_i + \sum_{i=1}^{K} v_{ij}s_i x_i \tag{2}$$

The first term is the weighted sum over input activations. w_{ij} are the weights between the input units (index i) and the units in the C-layer (index j), and N is the number of input units. The second term represents the inhibition to unit j. v_{ij} are the negative intra-layer weights between units in the C-layer. In the simulations reported in this paper, $v_{ij} = 0.5$ for all i and j. s_i is an additional long-term activation value of each unit in the C-layer, called the *winner status*. It represents a long-term memory for categorization, implicitly indicating how many times a unit has been the winner in the competition before. The winner is defined as the unit j, which, after a fixed number of times (e.g. 3) applying eq. (1), has the largest activation x_j. The winner status of this unit is increased by

$$s_j = \eta_1(1 - s_j) \tag{3}$$

thus converging toward a maximum value of 1. From eq. (2) it can be seen that the more often a unit wins the competition, the larger will be the inhibition excerting from that unit to others. In other words, the unit will better be able to suppress the non-winning units. The weights between the input layer and the C-layer are adapted as in regular competitive learning, by an instar rule:

$$\Delta w_{ij} = \eta_2 x_j(x_i - w_{ij}) \tag{4}$$

The feedback weights W_{ij} between the C-layer and the input layer are adapted similarly by an outstar rule:

$$\Delta W_{ij} = \eta_2 x_i (x_j - w_{ij}) \tag{5}$$

All input vectors are normalized to unit length, and so are all weight vectors, after applying one of the learning rules (4) or (5).

As [Grossberg 1987] has pointed out, this type of competitive learning is rather unstable, in that new patterns can influence categories that have already been learned. Therefore, like in adaptive resonance theory (ART, [Grossberg 1987]), mechanisms for *reset* and *recruitment* are introduced, triggered whenever an input pattern is not sufficiently close to the prototype built up through learning rule (4). Unlike in ART, here this situation (*mismatch*) is detected rather simply through the value of the net input y_j of the winning unit. If y_j is smaller than a parameter β (called *vigilance* like in ART), then the current winner j is inhibited and competition is run again to select another winner. This process is repeated until the winner's prototype is sufficiently close to the input pattern, or a unit which has never been winner before (detectable through the winner status s_i) is found, i.e. a new winner is *recruited*. In the latter case, all the weights leading to that winner are set identical to the corresponding input values, rather than applying rule (4). Initially, all weights start with small random values.

Another effect of introducing this ART-like mechanism is that the vigilance parameter controls the "widths" of the categories in terms of the deviation from the prototype that is permitted within a category. A small vigilance leads to rather wide categories including many different patterns, while a large vigilance leads to narrow categories.

Through this learning algorithm, clear categories are developed only after repeated presentation of stimuli. The degree of "clearness" is measured by a "goodness-of-fit" value g which is defined as a value expressing the extent by which the winner can suppress the other units. This value is basically the activation of the winner minus the average activation of all the other units, plus an additional difference of the activation of the winner and that of the second-most activated unit (to prevent two highly active units to lead to large values of g):

$$g_j = \frac{1}{2}((x_j - \frac{1}{K-1}\sum_{i=1}^{K} x_i) + x_j - x_{j'}) \tag{6}$$

if unit j is the winner and unit j' is the second-most activated unit. This goodness-of-fit value is used in building the link to label categories, as described in section 4.3. It is important to note that g contains more information about category membership than the net input y in that it is not simply a distance measure to the prototype but takes frequency of category members and different sub-spaces within the category into account.

This learning scheme is basically a model for unsupervised learning, i.e. categorization without an explicit target or reinforcement. However, it is easly turned into a more supervised categorization model by

- permitting other model components to trigger a reset and/or recruitment of winning units (thus making a decision with respect to the appropriate category, or selecting a new one).

- or by letting other model components to control the vigilance parameter

In the simulations presented below, it will be mainly the second aspect that will be exploited. However, as will be briefly discussed in section 10, extensions to the model with respect to hierarchical categorizations induced by language [Nelson 1988, Benelli 1988] or other complex aspects of language and categorization are rather straightforward (see also [Dorffner 1992a]). Figure 3 illustrates how (a) unsupervised learning is based on feature vectors alone, (b) how supervised learning can be done by selecting a winner from outside, and then attending to the features that define this category, and (c) how activating the feature layer through the feedback connections can lead to similarity-based processes (e.g. priming of a category that is similar on the level of features).

It is worth taking a quick look at how this categorization model can be made to roughly fit the above-mentioned aspects of human categorization. For this, it should be mentioned that we are careful not speak about sensory stimuli when talking about the inputs to the categorization model. Even though this model is to reflect the whole process from sensory input up to the formation of categories, it would be premature to assume that, for instance, visual categorization is as simple as the competitive learning scheme described above. Thus we assume that inputs are actually the result of some more or less complex "pre-processing" mechanisms (e.g. in higher visual layers) resulting in the extraction of primitive features. This is why we frequently speak of *feature vectors*. Note, however, that we do not assign a priori meaning to those features. Instead we base the properties of these features on probabilistic distributions (similar, for instance, to what is done in [Plunkett et al. 1993], however without the reference to some sensory-like input layer such as "retina").

- **Prototype effects:**
  As already hinted upon, competition and instar learning lead to the development of weight vectors which can be considered pattern prototypes of the category. This leads to the effect that patterns close to this prototype will activate the winner most strongly, and will subsequently lead to high values of goodness-of-fit.

- **Graded membership:**
  Through the mechanisms of leaving the units activated as they are after competition, and the defined goodness-of-fit value, graded membership to a category is accounted for by the model. Borderline case, for instance, could be described as inputs that lead to a small goodness-of-fit (because the winner is activated more weakly, and some of the other units more strongly, leading to a less successful competition). This information can be exploited by other model components.

- **Basic level effects:**
  In a completely unsupervised mode, there will be a level of "natural" similarity that

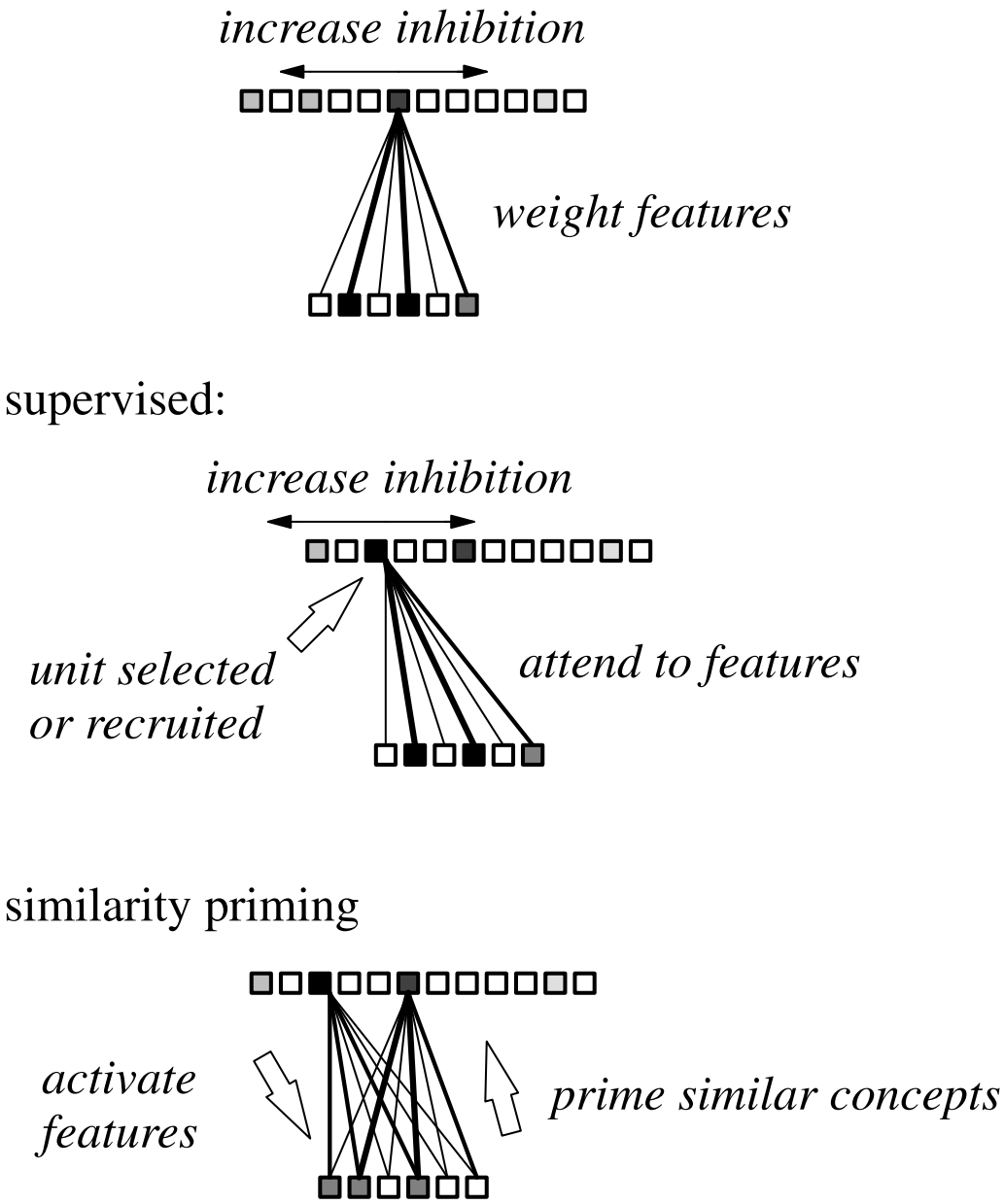

Figure 3: An illustration of unsupervised learning – based on feature vectors alone – supervised learning – where the winner is selected externally (e.g. induced through a word) –, and similarity-based priming of categories via activation of the feature layer. The latter process demonstrates that, although category representations tend to be orthogonal to each other (dissimilar), similarity between categories based on features can still play a role in interdependencies between categories.

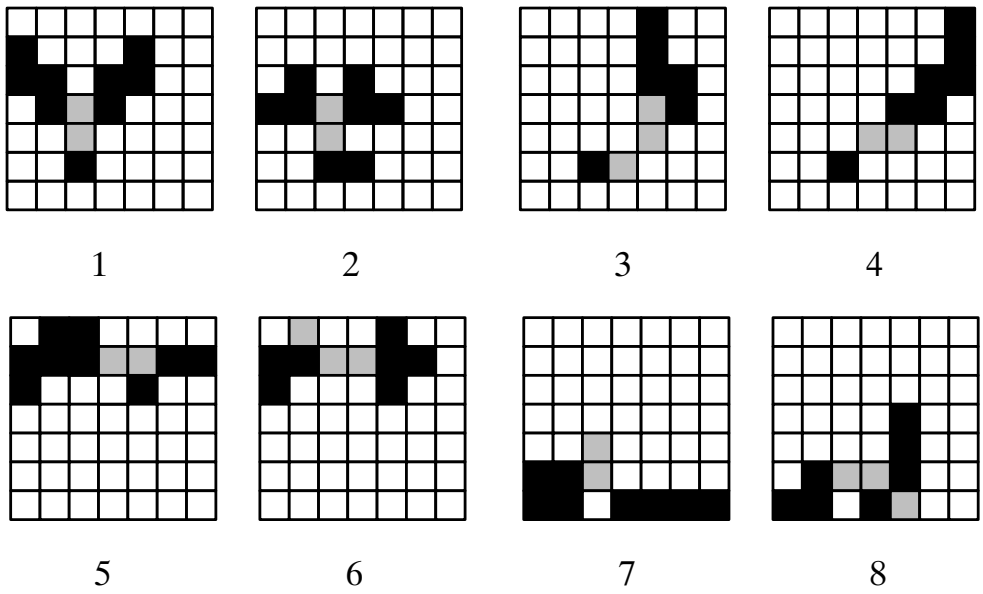

Figure 4: Eight patterns are most naturally categorized by grouping pattern 1 and 2, pattern 3 and 4, pattern 5 and 6, and pattern 7 and 8, respectively, although other groupings (e.g. patterns 1 through 4) would also show common features. This roughly corresponds to the basic level effect observed in human categorization

will lead most likely to a categorization (dependent on the value of vigilance). Categories which would require either higher or lesser similarity would need a "teacher", such as some reinforcement signal. Consider, for instance, fig. 4. Given a medium value of vigilance, these eight patterns will be most likely categorized such as to group pattern 1 and 2, 3 and 4, 5 and 6, 7 and 8, respectively, even though there could be other classes of patterns based on overlap of (binary) features (e.g. patterns 1 through 4). This property of "natural" level of similarity roughly coincides with the observations by [Rosch 1978], who defines basic level categories in terms of "cue validity" (the sum of the predicitiveness of all features) as "information-rich bundles of perceptual and functional attributes [...] that form natural discontinuities."[2]

- **Non-classical categories:**
  Even though we are talking about "feature vectors" refering to the inputs to categorization, this should not be mixed up with the assumption that categories could be defined as classes of members which all have a number of defining features in common. To the contrary – a winner is activated solely on the basis of the closeness of a pattern to the prototype. It is easy to see that one category can be activated

[2]It is interesting to note that Rosch even suggests a rather formal definition of cue validity and category membership in terms of weighted sums of features, which is very reminiscent of the type of connectionist model discussed here.

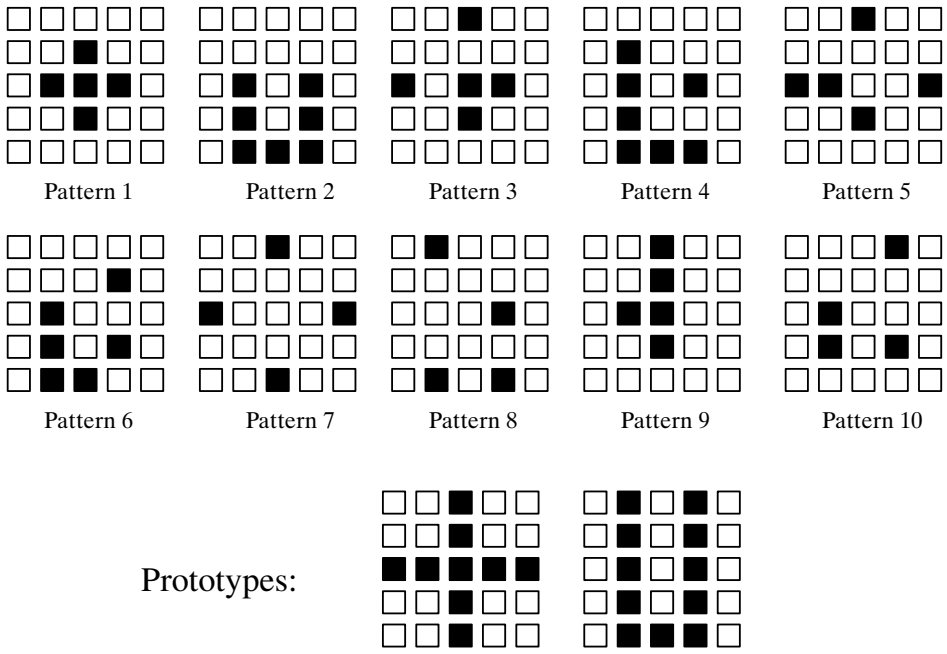

Figure 5: Eight patterns can be categorized in two classes (patterns 1, 3, 5, and 7 on one hand, patterns 2, 4, 6, and 8, on the other) corresponding to the two prototypes below, even though some patterns are disjunct with respect to their features (e.g. patterns 8 and 10)

by a number of patterns, among which some might be mutually disjunct in terms of features (see fig. 5). Furthermore it should be noted that the winner need not be activated by the input alone, but can receive activation of a great many other layers. In fact, ultimately we view C-layers as embedded in large structures of several C-layers, receiving inputs not only via sensory input, but also from other components, such as motor output, described in [Dorffner, in press]). Therefore, categories need not in all cases be inextricably be tied to feature vectors, but can be dynamically assigned to a much grater variety of situations (thus, in previous papers, we have always talked about 'situation' instead of 'sensory input' – e.g. [Dorffner 1992a]).

- **Metonymical extensions:**
  If, for some external reason, one or a few of the features in the feature vector are given more weight (salience), temporarily category membership can be extended to cases that only share these features. This property has not been used in the simulations yet, but will be discussed below as an interesting property with respect to some cases of word meaning overextension.

Before we describe the combination of at least two categorization modules into a model of word meaning acquisition, we briefly discuss the relationship of our model to other

connectionist models of human category learning.

One of the most well-known categorization models aimed at replicating aspects of human category learning is ALCOVE [Kruschke 1991, Kruschke 1993]. Like our model, ALCOVE is decisively distinct from MLP models with backpropagation, and implements categorization as a local process of attending to distinctive features in a network very similar to radial basis function networks [Broomhead & Lowe 1988]. ALCOVE is an *exemplar-based* model in that many units can specialize to different areas in vector space, which together form the representation of a category. This seems to be different from the prototype-based learning described here. However, our model could also be extended to a more exemplar-based version by choosing a rather large vigilance and adding additional C-layers to form more complex categories, based on other C-layers rather than the feature input. This possibility is not explored in this paper, which mainly focuses on prototype and basic level effects – issues not directly addressed by ALCOVE. Nevertheless, ALCOVE and our model can be viewed as being in the same "spirit", especially with respect to apparent problems with backpropagation learning (see [Kruschke 1993]).

Another categorization model is CALM by [Murre et al. 1992]. It is based on an array of modules which learn to specialize for categories using a novel WTA mechanism. Using winner-take-all, this model distinguishes itself from ours. Unfortunately, no evaluation with respect to human categorization performance was done – the focus of CALM is more on desribing categorization in a biologically plausible manner. It shares with out model the exploitatiuon of both unsupervised and supervised learning, and the organization in modules, which we are ultimately aiming at, as well (see section 10).

4.3 The naming model

A category formed based on sensory stimuli (or, more appropriately, based on feature vectors) is viewed as the potential internal substrate for the meaning associated to a word. In addition – contrary to most previous models – the identification of words is also seen as a categorization process, where words are physical entities causing patterns on a sensory input, which must be categorized. This more closely corresponds to the real-world situation than simply assuming orthogonal representations of words to be given (as is done in [Plunkett et al. 1993]). It can lead to a model that can account for effects rooted in the physical nature of words, and can avoid artefacts like assymetries caused by the fact that the categorization process of words would otherwise be side-stepped. The hypothesis raised herwith is that identifcation and categorization of words is similar to categorization in other compents. Given results from the nature of linguistic categories (e.g the prototype nature of parts-of-speech [Taylor 1989]) – although they are beyond the scope of this model – the hypothesis seems plausible.

The entire model therefore consists of two categorization modules and an additional layer (or array of layers), called a "(referential) link layer", to connect two categorization layers, one from each side (see fig. 6). For the sake of describing the model, we call one side (the one that categorizes words) the "acoustic" side (assuming words are presented acoustically) and the other one the "visual" side.

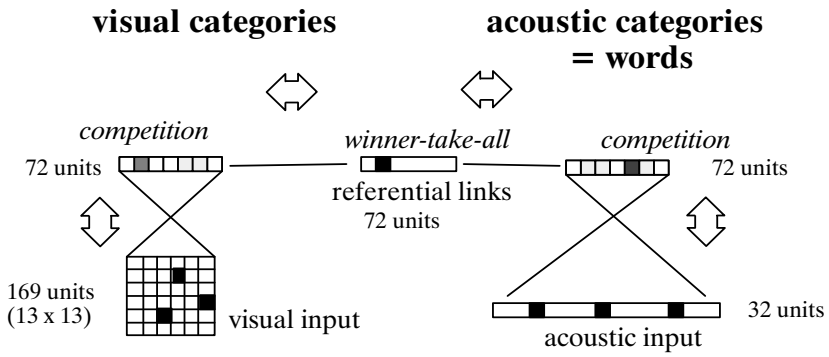

Figure 6: The model proposed in this paper, consisting of two inputs, two categorization layers and a layer linking two category representations

The purpose of the link layer is to develop links between word categories and visual categories that obey the properties of word meanings, most importantly the words' *arbitrariness*. In other words, similarities among words as physical entities must not generally be allowed to influence the corresponding visual categories (see the discussion in 8 below). To achieve this the following mechanisms were introduced:

- Each time the link layer is updated to form a link between category layers, a winner-take-all process is initiated, meaning that only one unit in the link layer is kept active (at maximum level), and all others are made inactive.

- A learning rule applied to connections between the link layer and the category layers is introduced that results in strong weights only between the winners in the link layer and the corresponding category layers. In other words, without explicit knowledge about the categories and their winners, the link layer learns to activate exactly one category unit. Thus, in spite of the permission of co-active units in the category layer, the weights are strengthened such that after learning only the winner (the unit representing the corresponding category) will be activated, erasing the influences of similarities to other category responses.

  This learning is defined as follows. Each unit in the link layer is basically treated like the units in a C-layer. Thus, they also possess a winner status s_j, which is raised like in (3) every time this unit fires in a winner-take-all process. Weights V_{ij} between a unit i in a C-layer and this link unit j are adapted according to

$$\Delta V_{ij} = \eta_3 x_i - \eta_4(\theta_s - V_{ij})s_j \tag{7}$$

  where η_3 and η_4 are learning rates (typically 0.1) and θ_s is a threshold value (typically 0.8). If V_{ij} is below θ_s, and $s_j > 0$, V_{ij} will rise only for large values of x_i, the activation of a unit in the C-layer. The more often this rule is applied (i.e. the

more often the link unit fires), the larger s_j will get. At the same time, V_{ij} will increase, thus decresing the influence of the second term in (7). Thus, large weights will stabilize. As a result, only weights from highly active units in the C-layer (i.e. only the winners, if the goddness-of-fit is large enough) will grow.

Weights in the other direction (from the category to the link layer) are adapted by simple outstar rules, analogous to (5).

When making a unit in the link layer fire, the goodness-of-fit value of the C-layers is taken into account. In particular, winner-take-all is only triggered if the g-values of both connected C-layers is above a threshold θ_g (in the simulations below, 0.5). A second mode, not discussed in this paper, permits firing of a link unit when at least one of the g-value is above threshold (for cases where language, i.e. naming, must induce a new category). This mechanism puts forward another hypothesis: A word can only be associated to some meaning – at least in early language acquisition – after at least some fuzzy category has been learned for this context. For instance, if a child hears the word 'apple' when faced with a certain object, they can map the word onto some meaning only if they have formed at least a rudimentary category (e.g. round objects). Unfortunately, we still lack conclusive data for either verifying or falsifying this hypothesis, although work in this direction is in progess[3].

Once links between words and their meaning have been built, a "mismatch" can be detected when a category is named with a different word. This can be done by comparing the contributions of the two different C-layers to a link unit with winner status above threshold. In case a visual C-layer would activate a different link unit than the acoustic C-layer, there must be an inconsistency in naming. In the current model version, such a mismatch triggers the increase of the vigilance parameter permitting a narrowing of the categories in order to correct this inconsistency (see section 6). More complex schemes, including the shift to categories on different taxonomic levels can also be initiated by mismatch (see [Dorffner 1992a].

The full algorithm of learning in the entire model is the following:

```
update both C-layers

if goodness-of-fit>threshold on both sides

          Pick winner in link layer and apply WTA

          if winner status is above threshold, but there is mismatch
                    raise vigilance

          else
              Apply learning rule (7)
```

[3]This is joint work with Catherine Harris at Boston University, to be reported in a forthcoming paper

```
        to weights from link layer to C-layers
Apply instar learning rule
        to weights from C-layers to link layer
```

Like above, we can briefly discuss how the main observations with respect to learning the words for different category types, are reflected in the model (compare, e.g., [Rosch 1978]):

- **Prototype effects:**
  Through the inclusion of the goodnes-of-fit in the decision whether a link unit can fire, patterns close to the prototype will (a) lead to better learning (as opposed to learning the mapping only based on border-line cases), and (b) will be learned earlier.

- **Basic level effects:**
  Since basic level categories (as described above) are learned first without the influence of language (unsupervised), they will be learned to be named earlier than categories which need language or other input. This phenomenon was raised as a hypothesis by [Rosch 1978], and was widely observed (e.g. [Nelson et al. 1993]), although it is not completely uncontested (e.g. [Callanan et al. 1994]). Learning words for categories on different levels of a taxonomic hierarchy is beyond the focus of the model version described here, but section 10 will briefly discuss extensions in this direction (see also [Dorffner 1992a]).

- **The influence of language:**
  Even though currently implemented in a rather simple way, the control of the vigilance parameter through mismatches in the link layer constitutes an important mechanism for accounting for some of the influences language plays on categorization. In this case, the width of the categories is narrowed through the consistent naming of the category. More sophisticated schemes, to be explored in the future, include the indivudal and temporary rising of the vigilance parameter in each case, such that different widths of categories are permitted (see also the ARTMAP model by [Carpenter et al. 1991]).

- **Learning of new words:**
  Since category representations tend to be orthogonal to each other, learning links between labels and visual categories shows only minimum interference. As will be discussed in section 8 this property is critical for language learning. It also means that there is no catastrophic interference [McCloskey & Cohen 1989] when learning new word-meaning mappings after having learnt a set of such mappings already. Whenever a new word or a new category is encountered, it can be learned without distorting other links, unless it constitutes a true extension to the previously learned mapping (e.g. a new pattern at the visual input is named by a word already used for a category, in which case that new pattern will have to be included in that category, if possible, or a new category will have to be recruited, which is mapped onto the same word).

- **Fast learning:**
  It has been reported at several places (e.g. [Clark 1993], p.49) that a few presentations of a word can be enough for stable learning. Contrary to an MLP model, our model permits the short-term increase of learning rates which can lead to rather fast learning, again without deleterous influence on previously learned mappings.

5 Data and Simulations

To make the results from this model comparable to previous results by [Plunkett et al. 1993], for all simulations reported in this paper the following data was used:

- 32 pattern prototypes, permitting some small random overlap between them, at the visual input. The patterns are binary and contain 30 active bits on an overall number of 169 units. From each prototype, the original, two patterns with a random distortion (flipping of bits) of 5%, and two patterns with a random distortion of 10% were presented, leading to an overall number of 160 patterns.

- 32 different binary patterns (4 active bits of 32 units each), permitting about 10 % overlap between patterns, for the acoustic input.

Note that the acoustic patterns indeed show similarities with each other, the influence of which must be diminished during learning. We did not introduce noise like for visual patterns, reflecting the fact that usually variations in visually presented objects are much larger than variations in acoustically presented words (see section 6). Each class of visual patterns was arbitrarily assigned one of the 32 labels as the "correct" one. Each learning step consisted of randomly picking one visual pattern and the corresponding acoustic word pattern. Both patterns are presented, categorization is initiated, as well as the learning algorithm for the connections between category and link layers. Two performance measures are distinguished [Plunkett et al. 1993]:

- Comprehension: For this, an acoustic pattern is input, the corresponding category layer, the link layer, the visual category layer and the visual input layer are updated (in that order) – provided that the goodness-of-fit in the acoustic C-layer was above threshold – applying winner-take-all to the link layer. A response is counted as being correct, if the input pattern (among the 5 times 32 patterns) which is closest to the resulting pattern in the visual input (according to its Euclidean distance) is in the "correct" category (the one that was associated with the word input).

- Production: For this, a visual pattern (one out of the 5 times 32 patterns) is input, the corresponding category layer, the link layer, the acoustic category layer, and the acoustic input are updated (in that order), applying winner-take-all to the link layer – again taking goodness-of-fit into account. A response is counted as correct, if the input pattern (among the 32 possible acoustic patterns) which is closest to the resulting pattern in the acoustic input layer is the "correct" word.

Figure 7: The results of a typical training run. The x-axis shows the number of pattern presentations, and the y-axis shows percentages of correct performance, both comprehension (upper curve) and production (lower curve).

Comprehension and production scores in fig.7, a typical run of the model, are percentages correct over all (5 times 32) visual and (32) acoustic patterns, respectively. The run depicted started with a vigilance of $\beta = 0.3$ and used a threshold $\theta_s = 0.7$, preceded by a run where only categorization on both sides was performed (thus building pre-linguistic categories). Initially, with rather low vigilance, less than 32 categories are learned on the visual side, leading to a number of incorrect namings (see section 6). Vigilance was then increased by a constant, whenever mismatch was detected. This leads to narrowing of the categories, the discovery of new categories (until, finally, 32 different classes are detected).

We will now discuss the results and several important phenomena in some detail.

6 Over- and underextensions

A well-known aspect of early language acquisition is that children tend to over- or underextend the meaning of words. Overextension means that the word is made to refer to more objects than the adult usage would infer (e.g. using 'ball' for all round objects, or 'apple' for all fruits). Underextension, on the other hand, means that the word is made to refer to only a subset of objects, as compared to the adult use of the word (e.g. the

word 'dog' refering only to poodles and dachshunds, or only to dogs passing by the yard). [Plunkett et al. 1993] discuss over- and underextensions in their model by showing that both phenomena can be replicated (with the former being more dominant).

It is worth looking at these two phenomena in more detail to demonstrate how our model permits putting the finger on the underlying reasons for them. According to [Clark 1993], the main mechanisms behind over-extensions are

- either, 'over-inclusions' making a category more widely defined than in adult use (e.g. 'ball' for all round objects). This is akin to extending a category by introducing a super-concept.

- or 'analogical extensions' due to the dominance of one or a few features (e.g. calling the inside of a lampshade 'moon'). This is akin to metonymical extensions of categories.

Both cases can be nicely reflected in the model. Over-inclusion could be equated with categories that are still represented more widely than by adults. By starting categorization in the C-layer with a small vigilance parameter, the model generally tends to form a small number of categories, each containg a large number of members. This modeling assumption postulates that children tend to form rather coarse pre-linguistic categories, which are then refined later, largely through the influence by language. Each time a category term has been learned, and a mismatch is detected (i.e. a word other than the previously learned one is uttered), the vigilance parameter is increased slightly. In the simulation, this is done globally, thus effecting other categorizations as well. However, it could also be done individually, by rising it for a short term (to affect the current category), and then decrasing again (compare the learning mechanism in the supervised ARTMAP network [Carpenter et al. 1991]).

Analogical overextension must obviously be attributed to a different mechanism. It can come about in the model if we assume that, especially at early learning phases, certain features are more salient (i.e. can be recognized more easily and with higher activations). Category learning will tend to form categories mainly according to those features (e.g. *crescent shape*). If such salient features are introduced into the model, it will begin forming prototypes centered around them. Over-extensions then come naturally. If salience is slowly decreased relative to the others (meaning that the other features become more easily recognized as well), the prototypes will be shifted toward the new features, gradually reducing memberships to the members of the adult categories (who, presumably, recognize those features rather than only the initially salient ones).

Underextensions are yet a different story. First of all, they appear to be much less frequent and dominant than overextensions, and some forms appear to occur only at very early stages. According to [Clark 1993] and [Elliot 1981] again two types can be distinguished (although there is much less detailed discussion on them):

- thematic dependencies, e.g. using 'ball' only when a ball is handed to the infant by its mother.

- restrictions to central members of a category [Clark 1993], e.g. using 'dog' only for prototypical dogs and not for more borderline members.

The first kind is clearly mainly out of the focus of the model. [Markman 1989] distinguishes between two types of relations that can be applied to group objects: *taxonomic* groupings according to common features (e.g. different dogs) and *thematic* relationships (e.g. a dog and a leash). Even though she postulates the *taxonomic assumption* which children appear to apply when deciding what a word might refer to (i.e. children most often make a word refer to taxonomic groupings than thematic ones, e.g. they would not learn 'dog' as refering to *leash*), thematic relationships are rather dominant in early infants (e.g. when asked to group things according to what belongs together, they would more likely put dog and leash together than two dogs - [Markman 1989], p. 23ff). The taxonomic assumption gives a good motivation for the model being based on categorizations of features. Nevertheless, thematic relationships apparently sometimes play a role in underextensions, when more than one object in the relationship is required to define the meaning of the word (e.g. the ball and the mother). To include that in the model we would have to assume a more complex perceptual component analysing scenes instead of just single objects, which is beyond the current focus.

The second type of underextension can, however, be replicated. Through the prototype effects in categorization and the gradual development of the goodness-of-fit, prototype members will be named earlier, while still excluding borderline cases (this is also observed by [Rosch 1978]).

Fig. 8 shows the development of three sample categories and their corresponding labels, dependent on the number of training cycles (pattern presentations), from a run that started with vigilance $\beta = 0.3$ and using a threshold $\theta_g = 0.5$. White bars show the number of patterns in the category correctly named, while black bars show the number of patterns from other categories named by the label belonging to this category. Since each category consists of five patterns, values on the white bars smaller than 5 correspond to underextensions (not all patterns are named correctly). All other namings, depicted by the black bars, represent over-extensions (patterns from other categories named by this label). The three chosen categories behave quite differently: The first shows large overextension, and some underextension in the beginning, then reducing to zero (no naming), until finally correct performance is reached. The second shows underextensions and overextensions persisiting for a while. Finally, the third shows no underextensions, but overextensions the number of which even increase initially. All three behaviors appear plausible when comparing to data from child language acquisition. Sudden jumps in overextensions coincide with relatively sudden "discoveries" of new categories, based on the increase in vigilance (akin to a split of categories, such as the split into *cookie* and *ball*, after initially the word 'ball' was used for both.

Fig. 9 shows the average number of over- and underextensions over all words that have already been learned at each time step. The upper (dotted) curve depicts the average number of patterns from a category that are correctly named (value 5 again corresponds to optimal performance). The lower curve depicts the average number of patterns from a

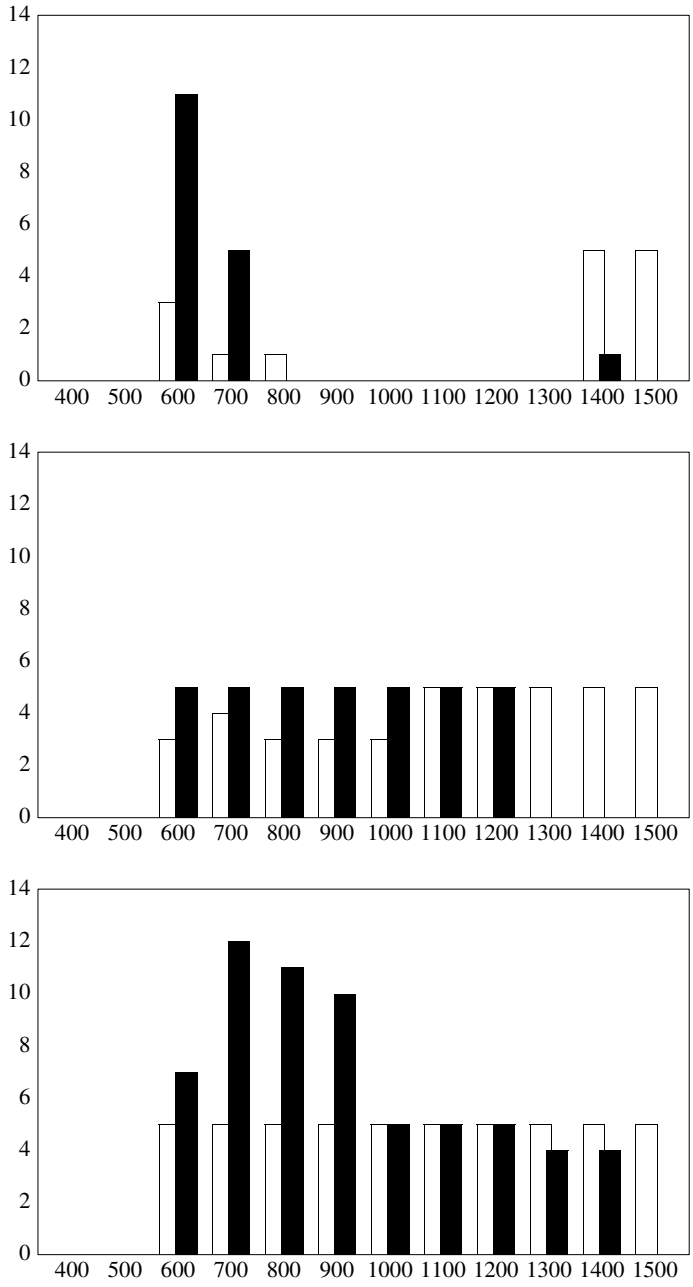

Figure 8: The number of correctly labeled patterns (white bars) of three selected categories, and the number of overextended patterns (patterns from other categories named by the same label – black bars) in dependence on the number of training cycles. A value of 5 on the white bars, and a value of 0 on the black bars corresponds to perfect performance.

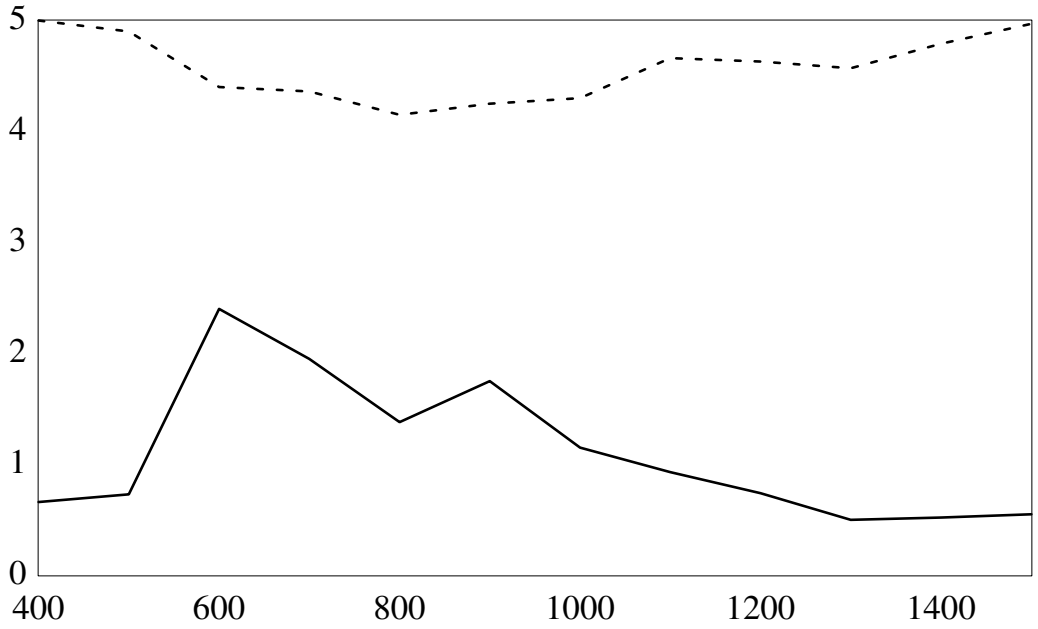

Figure 9: The average number of correctly labeled patterns per category (upper, dotted curve) and overextensions (lower curve), in dependence on the number of training cycles. Again values of 5 and 0, respectively, correspond to perfect performances.

different category named by the same label (overextensions). Both over- and underextensions increase after the first mappings are learned (around cycle 400), before they decrease again. This is mainly due to the fact that in the beginning only few labels are used (only 3 at cycle 400), which happen to be ones that correspond to a previously learned basic level categories. This initial behavior is not statistically significant, while the later parts of the curves (decrease in over- and underextension) appears to be more reliable, and also to be plausible when compared to language learning in children.

7 Comprehension and production

It is generally agreed that language in both children and adults shows an asymmetry between comprehension and production: The former always appears to be more developed than the latter – people comprehend more words than they can or do produce. With respect to child language acquisition this means that comprehension appears to be learned earlier and more quickly than production (see [Clark 1993] and [Elliot 1981] for discussions). [Chauvin 1989] and [Plunkett et al. 1993] demonstrate that their models can replicate this asymmetry: Learning curves for correct comprehension of words rise earlier and more quickly than those for correct production (see, for instance, [Plunkett et al. 1993], p. 303). They explain this in terms of an asymmetry of the identification of label vs. visual categories:

> Label vectors are far better predictors (and transparent representations) of category membership than retinal vectors which can be characterized as fuzzy and, for the high level distortions, sometimes equivocal in terms of their category membership. Until the network has established an accurate representation of the clustering of the images, labels provide more reliable cues to category membership ([Plunkett et al. 1993], p. 306).

One might argue that this asymmetry is an artefact of using local representations for labels, while visual categories are presented through several distributed prototypes. However, our model, even though working on distributed label input representations, can replicate the same comprehension-before-production advantage, mainly due to the fact that categorization on the visual side means classifying many distortions, while on the label side patterns are more distinct (and a higher level of vigilance is used). It is probably a safe assumption to propose that this aspect reflects reality. Consider the category *chair*: There are many more possible visual representations of chairs than there are varieties of the spoken word 'chair'. Thus, categorization on the visual side is a much lengthier process to achieve a state where the category reflects the adult's use, whereas on the label side the categorization process is simpler and more clear-cut. As a result, even though links between categories develop symmetrically in our model, an asymmetry comes from the fact that it takes longer to include all members into a category on the visual side than on the label side.

Fig. 7 above illustrates this. Even though at the beginning both comprehension and production scores rise in parallel, after about 750 pattern presentations a decisive advantage of comprehension over production emerges. Also, the comprehension score rises more smoothly, while production continually shows some oscillations (compare the results in [Plunkett et al. 1993]). The split of the two scores coincides with the discovery of several new categories (through the increase in vigilance) and thus an increase in overextension (see also fig. 9).

This phenomenon thus coincides with the phenomena of over- and underextensions (see section 6): Overextension, for instance, affects production much more than it affects comprehension. When hearing a word like 'ball', most likely the prototype of that category is activated, even though the category might still be based on limited features like *roundness*. The current prototype is more likely to fulfill the adult criteria of the category than borderline or "incorrectly" included cases (like a round cookie). Thus comprehension will often not reflect the fact that the category is actually overextended. Production, on the other hand, will reveal overextensions, especially when borderline or "incorrectly" included members are to be named (i.e. when the child calls a cookie a 'ball').

A yet more detailed analysis is in place, though. [Clark 1993] discusses the asymmetry between comprehension and production in terms of representations including the actual production of a word. Both the model in [Plunkett et al. 1993] and our model have been tested on production by checking whether the original label pattern could be reproduced. This approach entirely leaves out the process of articulating a word, and the fact that perceiving and producing a word are quite distinct. [Clark 1993], p. 246, proposes the existence of *C-representations* (containing information of the auditory properties of a word) and *P-presentations* (containing articulatory information). In this view, a visual category might well be able to activate the C-representation of a word (which, in the models discussed, would be the representations built up in the label modules of the models), but unless appropriate P-representations have been learned, the word cannot actually be produced (articulated). P-representations are not contained in either [Plunkett et al. 1993] or our model (although some preliminary work has been devised to go a step in this direction [Dorffner & Schoenauer 1993]). Thus the asymmetry in the models' learning curves cannot directly be compared to asymmetries in psycholinguistic studies (e.g. [Clark & Hecht 1983], cited in [Clark 1993]). In this context the fact that in the model the comprehension–production asymmetry in fig. 7 could not be observed with all parameter settings must be seen.

Furthermore, [Elliot 1981] cautions against overestimating a child's understanding of a word: "... children may appear to understand much more language than they actually do" ([Elliot 1981], p. 81). For instance, apparent understanding might come to a large part by contextual effects (e.g. deixis or miming). Since we can only observe children's behavior, but we can test our models in a quite detailed way, generalization of the models' predictions must be handled with care.

In sum, although many aspects of the asymmetry between comprehension and production appear to be outside the focus of the discussed models, they appear to reflect one important part of it – the asymmetry between varieties in visual and label categories, and

the relation of this asymmetry to over- and underextensions.

8 Symbol emergence and the naming insight

Both [Chauvin 1989] and [Plunkett et al. 1993] stress the issue of *symbol emergence* in their respective work. In other words, learning the meaning of words is seen as the learning of symbol use. We believe that this issue has not been treated sufficiently in the context of language acquisition and connectionist models. Thus we discuss it to some detail in this section.

What is a symbol? Besides the many definitions and treatments in computer science and artificial intelligence, the most relevant view on symbols in language seems to come from semiotics (e.g. [Sebeok 1994]). There, a symbol is defined as a type of sign with special properties, most prominently *arbitrariness*. This refers to the property of symbols that their shape (form) does not reflect anything about their referent (meaning). According to [Sebeok 1994], p. 33, a symbol is

> [a] sign without either similarity or contiguity, but only with a conventional link between its signifier and its denotata, ...

This type of sign is usually distinguished from other types like *index*, *symptom*, or *icon*:

> A sign is said to be iconic when there is a topological similarity between a signifier and its denotata ([Sebeok 1994], p. 28).

Of course, as [Sebeok 1994] emphasizes, there is hardly any sign that is purely a symbol, or an icon, etc. Instead signs are usually identified with their most prominent property. In this sense, words (or better: morphemes) are dominantly symbols (for iconic aspects of language, see for instance [Haiman 1985]), i.e. a word form does not reflect any properties of its meaning. Learning words as refering to a category thus requires the acquisition of the conventionalized link, as given by the speakers of the community (e.g. an adult speaking to a child). This induction process [Markman 1989] cannot be based on stimulus similarities but must consist of building (almost) completely arbitrary links between word form and meaning.

Complete neglection of stimulus similarities does not only apparently run against the more "natural" way by intelligent beings of dealing with stimuli (see above), but also runs against the "natural" mechanisms of connectionist networks. A standard neural network architecture, such as a perceptron or an MLP, always tends to be similarity-sensitive, generalizing to new inputs based on vector similarities. Even though an MLP can be proven to be able to implement arbitrary functions between input and output [Hornik et al. 1989], it will always tend to interpolate across pattern sub-spaces that have not been used for training. For the learning of the mapping between word forms and their meanings, this means that they tend to show "blending" effects (see also [Prem 1994]). For instance, if a word form, which is intermediate between two learned word forms, is presented, the MLP

tends to produce an output which is somehow intermediate between the learned outputs, as well. In other words, those networks consider words to be much more iconic than they actually are. With respect to words as true symbols, this blending effect must be avoided to prevent implausible phenomena.

Our model attempts to deal with this problem explicitly. The first step toward arbitrariness is the categorization process which leads to orthogonal near-unitized responses. If representations of different categories are perfectly orthogonal to each other, an arbitrary link without blending can be realized. However, as discussed above, category membership is graded and develops gradually, thus still giving rise to intermediate, non-orthogonal states. Therefore, the referential link layers must employ an additional mechanism to prevent blending on one side (label categories or visual categories) to be transfered to the other side – exactly the mechanisms (WTA, special learning rulre) described above. We view such a mechanism as essential when attempting to understand the basics of language reference.

The MLP models by [Chauvin 1989] or [Plunkett et al. 1993] side-step one half of the blending problem by assuming label representations to be discrete and orthogonal to begin with. However, in a language learner such discrete states cannot be assumed to be given. It is the learner's task to identify both the category a label refers to and the label itself, perceived as physical stimulus like everything else. This is why we introduced a catgeorization component for labels, as well, and have assumed the label inputs also to be distributed feature patterns. "True" symbol emergence, according from the semiotic definition, is thus the learning process consisting of

- the identification of (object or other) categories
- the identification of labels as distinct entitites in terms of categories
- the realization that labels (words) can refer to (object) categories independent of their form
- the implementation of the corresponding arbitrary mapping

Except for the third aspect, our model can give connectionist accounts for these processes. The third aspect could be equated with something often termed the "naming insight" ([McShane 1979], cited in [Plunkett et al. 1993], sometimes seen as the reason behind *vocabulary spurts*, i.e. a period of sudden rapid increases of vocabulary size in infants after a period of relatively slow learning. In this view, a child is seen as "discovering" that words can productively stand for things, knowledge which is widely exploited thereafter. Before such an insight, iconic properties of words appear to be dominant, reflected in acts of imitating sounds, or rather limited, situation-dependent use of words. An interesting account of this observation in a completely different setting is given by [Schaller 91], who describes her process of teaching a deaf adult without language some basic signs of ASL (American sign language). In their first meetings, the man would continually imitate her

signs, and reference could be made only through rudimentary, iconic properties of miming. Only after several days would he, relatively suddenly, discover that a sign can have an arbitrary reference to a category of things. After this insight, language learning could proceed (albeit rather slowly).

We would like to argue that most mechanisms behind this kind of naming insight are still beyond the focus of connectionist models, which cannot include such complex aspects as motivation, intention, goals of behavior, and the like. However, some basis of this appears to be possible to be included. [Plunkett et al. 1993], and before them, [Chauvin 1989] discuss the apparent vocabulary spurts in their models in terms of the discovery of clusters or principal components in the representations the build up. Speaking metaphorically, representations in the hidden layer, starting at random, have to "unfold" before major progress in learning the word–meaning mappings can be made. Once this unfolding has proceeded to a certain extent, the acquisition of mappings comes rather quickly and easily.

In our model this would correspond to the "unfolding" in terms of clear categories which take time to develop, and which are considered as a precondition for first links to be built. Admittedly, the resulting spurt in vocabulary size is not as distinct as in [Plunkett et al. 1993]. This points to some weakness of our model as compared to MLPs on one hand, and on the necessity for more careful evaluations (e.g. when are which categories learned?) and taking into account other possible mechanisms behind the naming insight, on the other.

9 Discussion

The results discussed in the previous section show that our model is not only able to reproduce many of the most important aspects of early word learning, but also permits a pinpointed discussion of the underlying mechanisms in terms of model components. The model is able to reproduce phenomena like

- Over- and underextensions children apply to word meanings
- The asymmetry between comprehension and production in language learning
- The emergence of "true" symbols in the semiotic sense.

The discussions above show that all these phenomena apparently involve several underlying roots, only some of which can be included in a connectionist model like the one presented. By this, such a model helps identify those roots and provides a step toward a better understanding of them.

While being motivated by the importance of categorization in language learning, the model can also overcome several of the shortcomings of previous models based on multilayer perceptrons and backpropagation (compare section 3):

- Through the orthogonalization process in categorization it does not show the property of catastrophic interference MLPs are plagued with.

- It is able to perform fast learning and the learning of new words without distortion of previously learned words
- It reflects many properties of human categorization before and during language learning
- Through competition and selection of winners it is basically able to implement many-to-one mappings such as the ones needed for synonyms or homonyms
- Through the lack of interference in representation it can be scaled up much more easily than MLPs.
- Its representations are less context-dependent
- Its representations provide a more accessible handle for the modeler to control the model's functions and to obtain an understanding of the processes underlying language learning.

Not all of these properties have really been demonstrated in this paper, and will have to be evaluated more carefully in the future.

As hinted upon in the introduction, we do not view this kind of model as completely replacing multilayer perceptrons and backpropagation. As the discussion has shown, our model does show some weaknesses with respect to properties that MLP can nicely exhibit. One such property is the effect of "unfolding" leading to a natural naming insight. Although arguments have been given that the roots of this phenomenon might lie elsewhere, the MLP appears to have some strength in this respect, not paralleled by our model. Therefore, we propose a combination of MLP and explicit categorization in the future. One possible such combination is depicted in fig. 10. There, MLPs are suggested to implement similarity-sensitive mappings, such as the mappings between orthography and sound (like in [Seidenberg & McClelland 1989]) or the building up of distributed representations based on perception *and* action. Categorization then works with those representations as input, providing an interface to language.

Such a combination of distributed models with a categorization module suggests the reintroduction of a more explicit lexicon than authors like [Seidenberg & McClelland 1989] have suggested. It could also provide the basis for the reintroduction of a moderate "two-process" model for aspects like the grapheme-to-phoneme mapping (e.g. rare exceptions could be modeled by association to the word category rather than through the similarity-based process of the MLP). Future research will be concerned with the evaluation of this proposal.

Like every connectionist model, our model has a large number of parameters (e.g. vigilance, thresholds, aspects of the algorithm) which will have to be fine-tuned to match psycholinguistic data more adequately. However, the model appears restricted enough in order to form the basis of a theory in its own, by raising certain important hypotheses about language learning, such as

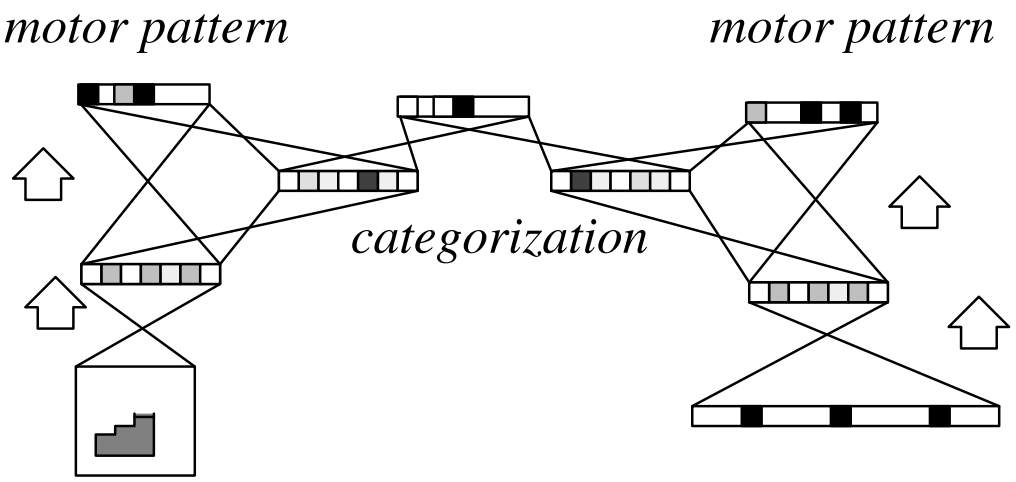

Figure 10: A proposal for combining MLP components with categorization modules. The MLPs are used to map perception to action, or orthographical input to phonemes, while categories form the interface to language, like above.

- Prelinguistic categories play a large role in language learning.
- For a link between a word and its meaning to develop strongly, a category must have been built up at least to some degree
- Representations of categories are dissimilar to each other
- etc.

Again, more research and also more data will be needed to validate these hypotheses.

10 Future research

As hinted upon several times, the model presented in this paper is only a simple version of a much larger setting, envisioned to be implemented and tested in the future. Here are some aspects of how this model can still be extended and further tested:

- Replacing single C-layers by pools of several C-layers, interconnected with each other, more complex category structures can be modeled. For instance, in the acquisition of taxonomic hierarchies, categories in different layers can evolve, partially based on other catgeories rather than feature inputs.

- Several constraints children appear to apply when learning word meanings, such as the ones suggested by [Markman 1989], can be implemented by further mismatch-and-reset mechanisms and adaptive foci of attention with respect to the feature vectors (see [Dorffner 1992a] for a rough discussion of such an extension).
- We hinted above that words are not pure symbols but also do show some iconic properties (e.g. sound symbolism). By side-stepping the winner-take-all process in the link layer under certain conditions such similarity-based associations between word form and meaning can be reintroduced.
- By embedding the model in the framework of an autonomous agent, aspects such as the grounded and situated nature of representations and the role of language in action can be studied (see, for instance, [Dorffner & Prem 1993]).

We are currently conducting research in most of these directions.

11 Conclusion

In this paper we have presented a model for early language learning, focusing on categorization as a central process in both cognition and language. We have shown results from several simulations and have demonstrated how the model can replicate many important phenomena in language acquisition. We have also discussed the relationship of this type of model to previous connectionist models based on multilayer preceptrons and backpropagation, and have shown how the two might be able to complement each other in the future. As such, we consider this model as a contribution to increased understanding of language and language acquisition via connectionist networks.

Acknowledgements

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Sciene, Research, and the Arts. I thank Georg Thurner and Michael Hentze, currently working on some further aspects of the model, for helpful input and comments. I further thank Prof. Robert Trappl for his support of this research.

References

[Baldi & Hornik 1989] Baldi P., Hornik K.: Neural Networks and Principal Component Analysis, Neural Networks, 2 (1) p.53-58, 1989.

[Benelli 1988] Benelli B.: On the Linguistic Origin of Superordinate Categorization, Hum.Dev., 31:20-27, 1988.

[Bowerman 1978] Bowerman M.: The acquisition of word meaning: an investigation of some current conflicts, in Waterson N., Snow C.E. (eds.): The development of communication, Wiley, New York, pp. 263-287, 1978.

[Broomhead & Lowe 1988] Broomhead D.S., Lowe D.: Multivariable Functional Interpolation and Adaptive Networks, Complex Systems, 2,321-355, 1988.

[Callanan et al. 1994] Callanan M.A., Repp A.M., McCarthy M.G., Latzke M.A.: Children's hypotheses about word meanings: is there a basic level constraint?, J. of Experimental Child Psychology, 57, 108-138, 1994.

[Carpenter & Grossberg 1990] Carpenter G.A., Grossberg S.: ART 3: Hierarchical Search using Chemical Transmitters in Self-Organizing Pattern Recognition Architectures, Neural Networks, 3, pp.129-152, 1990.

[Carpenter et al. 1991] Carpenter G.A., Grossberg S., Reynolds J.H.: ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network, Neural Networks, 4, 1991.

[Chauvin 1989] Chauvin Y.: Toward a Connectionist Model of Symbolic Emergence, *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum, pp. 580-587, 1989.

[Clark 92] Clark A.: The Presence of a Symbol, Connection Science, 4(3-4), pp.193-206, 1992.

[Clark 1993] Clark E.V.: The Lexicon in Acquisition, Cambridge University Press, 1993.

[Clark & Hecht 1983] Clark E.V., Hecht B.F.: Comprehension, production and language acquisition, Annual Review of Psychology 34, 325-349, 1983.

[Cottrell et al. 1990] Cottrell G.W, Bartell B., Haupt C.: Grounding Meaning in Perception; *Proc. of the German Workshop for AI (GWAI)*, Heidelberg: Springer, 1990.

[Dorffner 1989] Dorffner G.: A Sub-Symbolic Connectionist Model of Basic Language Functions, Indiana University, Computer Science Dept., Dissertation, 1989.

[Dorffner 1992a] Dorffner G.: Taxonomies and Part-Whole Hierarchies in the Acquisition of Word Meaning - A Connectionist Model, *Proceedings of the Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum, 1992.

[Dorffner 1992b] Dorffner G.: A Step Toward Sub-Symbolic Language Models without Linguistic Representations, in Reilly R., Sharkey N.(eds.): *Connec tionist Approaches to Language Processing (Vol. I)* , Hove: Lawrence Erlbaum, 1992.

[Dorffner, in press] Dorffner G.: Radical Connectionism - A Neural Bottom-Up Approach to AI, to appear in Dorffner G.(ed.), Neural Networks and a New AI, Thompson, London, (in press).

[Dorffner et al. 1993] Dorffner G., Prem E., Trost H.: Words, Symbols, and Symbol Grounding, Oesterreichisches Forschungsinstitut fuer Artificial Intelligence, Wien, TR- 93-30, 1993.

[Dorffner & Prem 1993] Dorffner G., Prem E.: Connectionism, Symbol Grounding, and Autonomous Agents, Proceedings of the 15th Annual Meeting of the Cognitive Science Society, Boulder, CO, pp. 144-148, 1993.

[Dorffner & Schoenauer 1993] Dorffner G., Schoenauer T.: Unsupervised Learning of Simple Speech Production Based of Soft Competitive Learning, in Eeckman F.H. & Bower J.M.(eds.), Computation and Neural Systems, Kluwer Academic Publishers, Boston/Dordrecht/London, pp.363-368, 1993.

[Dromi 1987] Dromi E.: Early Lexical Development, Cambridge University Press, 1987.

[Elliot 1981] Elliot A.J.: Child Language, Cambridge University Press, 1981.

[Gasser & Lee 1990] Gasser M., Lee C.-D.: Networks that Learn about Phonological Feature Persistence, Connection Science, 2(4), pp.265-278, 1990.

[Grossberg 1987] Grossberg S.: Competitive Learning: From Interactive Activation to Adaptive Resonance, *Cognitive Science* **11(1)**, pp.23-64, 1987.

[Grossberg 1982] Grossberg S.: Studies of mind and brain, Reidel Press, Boston, 1982.

[Grossberg & Stone 1986] Grossberg S., Stone G.: Neural dynamics of word recognition and recall: attentional priming, learning, and resonance, Psychological Review, 93(1)46- 74, 1986.

[Haiman 1985] Haiman J.(ed.): Iconicity in Syntax, John Benjamins, Amsterdam, 1985.

[Hinton et al. 1993] Hinton, G.E., Plaut, D.C., Shallice, T.: Simulating Brain Damage, *Scientific American*, October, 1993.

[Hornik et al. 1989] Hornik K., Stinchcombe M., White H.: Multi-layer Feedforward Networks are Universal Approximators, Neural Networks, Vol. 2, pp.359-366, 1989.

[Kohonen 1984] Kohonen T.: Self-Organization and Associative Memory, Springer, Berlin, 1984.

[Kruschke 1991] Kruschke J.K.: ALCOVE: A Connectionist Model of Human Category Learning, in Lippmann R.P., et al.(eds.), Advances in Neural Information Processing 3, Morgan Kaufmann, San Mateo, CA, pp.649-655, 1991.

[Kruschke 1993] Kruschke J.K.: Human Category Learning: Implications for Backpropagation Models, Connection Science, 5(1), pp.3-36, 1993.

[Lakoff 1987] Lakoff G.: Women, Fire and Dangerous Things; What Categories Reveal about the Mind, University of Chicago Press, Chicago, 1987.

[Langacker 1987] Langacker R.W.: Foundations of Cognitive Grammar, i, Theoretical Prerequisites, Stanford University Press, 1987.

[Markman 1989] Markman E.M.: Categorization and Naming in Children, MIT Press/Bradford Books, Cambridge, 1989.

[McCloskey & Cohen 1989] McCloskey M., Cohen N.J.: Catastrophic interference in connectionist networks: the sequential learning problem, in Bower G. (ed.): The Psychology of Learning and Motivation, Vol. 24, Academic Press, New York, 1989.

[Murre et al. 1992] Murre J.M.J., Phaf R.H., Wolters G.: CALM: Categorizing and Learning Module, Neural Networks, 5(1), pp.55-82, 1992.

[Neisser 1987] Neisser U.(ed.): Concepts and Conceptual Development, Cambridge University Press, Cambridge, UK, 1987.

[Nelson 1988] Nelson K.: Where Do Taxonomic Categories Come from?, Hum.Dev., 31, p.3-10, 1988.

[Nelson et al. 1993] Nelson K., Hampson J., Kessler-Shaw L.: Nouns in early lexicons: evidence, explanations and implications, J. of Child Language, 20, 61-84, 1993.

[Plunkett et al. 1993] Plunkett K., Sinha C., Moller M., Strandsby O.: Symbol Grounding or the Emergence of Symbols? Vocabulary Growth in Children and a Connectionist Net, *Connection Science* **4(3&4)**, 1993.

[Rosch 1978] Rosch E.: Principles of Categorization, in E.Rosch, B.B.Lloyd (eds.), Cognition and Categorization, Erlbaum, Hillsdale, NJ, 1978.

[McClelland & Rumelhart 1981] McClelland J.L., Rumelhart D.E.: An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings, Psychological Review, Vol.88, 375-407, 1981.

[McShane 1979] McShane J.: The development of naming, Linguistics 17, 879-905, 1979.

[Prem 1994] Prem E.: Symbol Grounding: Die Bedeutung der Verankerung von Symbolen in reichhaltiger sensorischer Erfahrung mittels neuronaler Netzwerke, Institut fuer Med.Kybernetik u. AI, Universitaet Wien, Dissertation, 1994.

[Rumelhart & Zipser 1985] Rumelhart D.E., Zipser D.: Feature Discovery by Competitive Learning, *Cognitive Science* **9(1)**, pp.75-112, 1985.

[Rumelhart & McClelland 1986] Rumelhart D.E., McClelland J.L.: On Learning the Past Tenses of English Verbs, in McClelland J.L. & Rumelhart D.E., Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol II: Psychological and Biological Models, MIT Press, Cambridge, MA, 1986.

[Rumelhart et al. 1986] Rumelhart D.E., Hinton G.E., Williams R.J.: Learning Internal Representations by Error Propagation, in Rumelhart D.E. & McClelland J.L., Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol 1: Foundations, MIT Press, Cambridge, MA, 1986.

[Sales 1995] Sales N.: Dissertation, Imperial College, Univ. of London, 1995.

[Schaller 91] Schaller S.: A Man Without Words, Summit Books, New York, 1991.

[Schyns 1991] Schyns P.G.: A Modular Neural Network Model of Concept Acquisition, Cognitive Science, 15(4), 1991.

[Sebeok 1994] Sebeok T.A.: Signs, University of Toronto Press, 1994.

[Seidenberg & McClelland 1989] Seidenberg M.S., McClelland J.L.: A distributed, developmental model of word recognition and naming, Psychological Review, 96(4)523-568, 1989.

[Taylor 1989] Taylor J.R.: Linguistic Categorization, Clarendon Press, Oxford, 1989.

[Wittgenstein 53] Wittgenstein L.: Philosophical Investigations, MacMillan, New York, 1953.