A Study of Perceptually Grounded Polysemy in a Spatial Microdomain

Jordan Zlatev*

TR-92-048

August 1992

Abstract

This paper attempts to exemplify the advantages of *perceptually grounded semantics* with respect to traditional formalist approaches in elucidating the nature of the controversial notion of linguistic *polysemy*, or multiplicity of meaning. It is also suggested how some aspects of language typically associated with *compositionality* could be modeled, without there being a strictly "compositional semantics".

This is done through a series of experiments, using modifications of Terry Regier's connectionist system for learning spatial relations [Regier, 1992] which constitutes a part of the L_0 project concerned with associating descriptions in an arbitrary language with an analog environment, (sequences of) pictures of simple 2-dimensional scenes.

The emphasis is above all on the English preposition 'over', famous for its polysemy, and analyzed in detail by [Brugman, 1981] and [Lakoff, 1987], but some modeling has been also done of the meaning of 'under', as well as some rudimentary semantics for simple verbs such as 'be', 'go' and 'fly' that combine with the two prepositions.

Three kinds of connectionist architectures have been used in trying to capture what might be called a "polysemous *over*". It is suggested that the first seems to treat polysemy like what has traditionally been regarded as *generality*, where distinctions

*The author has been supported by a scholarship from *The Swedish Institute* and may be reached by e-mail as **zlatev@icsi.Berkeley.EDU** or **jordan@ling.su.se**.

are neutralized and "senses" are not distinct, while the second reduces polysemy to *homonomy* where they are distinct but not related. It is the third type of (structured) connectionist architecture that managed best in both learning different "senses" and reflecting the "polysemous structure" of the lexical item in analyses of the relevant hidden layers. In this architecture polysemy emerges as an effect of the combinatorics of words and their pairing with the environment.

The main theoretical claim is that polysemy is best regarded as a *contextual* rather than a purely lexical phenomenon. This on its part suggests support for the claim made in [Geeraerts, 1992] that the distinction between polysemy and generality is unstable, and for a semantics that is radically anti-reificational. The results from this study suggest that such a semantics *can* account for the generativity and systematicity of language, despite claims to the contrary made by formalists.

1 Introduction

1.1 Perceptually Grounded Semantics

Recently in cognitive science there is an increasing interest in a new type of approach to semantics, which I in accordance with most common usage will refer to as **perceptually grounded semantics**. To appreciate its novelty, we might briefly contrast it with more traditional approaches to explaining "meaning", which without much coercion can be grouped by the term **formalist semantics**. A few well-known formalist projects are for example Kamp's Discourse Representation Theory [Kamp, 1984] (in logic), Jackendoff's Conceptual Semantics [Jackendoff, 1990] (in linguistics), and Schank's Conceptual Dependency Theory [Schank and Abelson, 1977] (in artificial intelligence). Despite all the considerable differences among such theories on what words, sentences and discourse "really mean", the account of this meaning is to be provided by something like a two-step procedure:

First, the fragment of "natural language" that is to be "given a semantics" is translated by some well-defined set of formalization rules into some sort of symbolic description or other ("semantic/conceptual representation", "logical form" etc.) in a pre-defined "formal language".

Then, depending on the variety of the theory, the meaning of the symbols and sentences in this formal language is defined precisely by either some set of "semantic primitives" (in more "mentalistic" approaches) or by an "interpretation function" to a "model" (as in model-theoretic approaches), thus giving the language a "denotational semantics". In any case, both "mentalist" and "realist" accounts of natural language semantics ignore the actual experienced world of human action and perception, in which language takes place, either substituting it with a set-theoretic model, or by staying, in a solipsist manner, within the realm of "mental representations", leaving it to "pragmatics" or "performance" to give an account of how these "representations" are learned, communicated and used in everyday life.

In reaction to formalist semantics, one can say that the representations the formalist uses lack any actual meaning, since they are not "grounded" in actual action and perception. This is exactly the problem that perceptually grounded semantics is to address. And unlike some 10 years ago, when there was essentially "no other game in (cognitive science) town" than formalist semantics, currently there are a number of cognitive science projects such as those of Hofstadter et al. (cf. [Chalmers *et al.*, 1991; French and Hofstadter, 1991]), Nenov (cf. [Nenov and Dyer, 1988; Nenov, 1991]), and Feldman et al. (cf. [Feldman *et al.*, 1990; Weber and Stolcke, 1990]) that try to tackle the issue of "Whence Perceptual Meaning?" [1].

Furthermore such projects use undeniably "scientific" tools such as computational modeling, and can not be accused of being "merely speculative", the stock reply by formalists to such serious criticisms of their presuppositions as those expressed in [Dreyfus, 1991] and [Winograd and Flores, 1987]. To allow modeling, however, all current perceptually ground-

[1] This is the title of an intriguing polemical paper by Varela [Varela, 1992], that proposes the notion of *enaction* as a new paradigm in cognitive science where cognition is seen as "Effective action: History of structural coupling which enacts (brings forth) a world." and this is achieved "[t]hrough a network of interconnected elements capable of structural changes undergoing an uninterrupted history." (p.256) The current work is (to an extent) in affinity with that proposal.

ing projects need to *restrict the domain* which is to constitute the "world" for their systems, bringing back the currently unfashionable in AI notion of "micro-worlds". The hope, as expressed in [Chalmers *et al.*, 1991] quite clearly, is that:

> "While microdomains may superficially seem less impressive, the fact that they are explicitly idealized worlds allows the issues under study to be thrown into clear relief — something that generally speaking is not possible in a full-scale real-world problem. Once we have some understanding of the way cognitive processes work in a restricted domain, we will have made genuine progress towards understanding the same phenomenon in the unrestricted real world." (pp. 20-21)

Whether this hope holds out or not, we should note that the microdomains of perceptually grounded systems are qualitatively different from the "micro-worlds" of AI in the 70's. The important difference is that they are *analog* rather than propositional, which forces the system to "extract" those aspects of the "situation", which are considered relevant for its current purposes, much in the way living organisms are thought to function in an environment. This makes it impossible to have a "mapping" between the "objects" and "relations" in the microdomain and a set of pre-fixed representations, as in the average "blocks world" AI system. Meaning has to "brought forth" through a history of interaction.

This leads to the possibly most important characteristic of perceptually grounding systems (that deserve their name), namely that they are *learning systems* [2] which means that they are allowed to "evolve" the structures that make it possible for them to carry out a given task. No matter what stance one takes on the status of these structures, i.e whether one regards them as "mental representations", or one adopts a non-representational stance and prefers to talk of a "structural coupling" of the system/organism with the environment, it is clear that these structures are not pre-defined, discrete, symbolic expressions with a "well-defined semantics" but rather emerge as a consequence of the interaction of the system with the environment. Virtually all perceptually grounded learning systems are *connectionist* in one form or another, due to the exteremly useful properties of (artificial) neural networks as a means for "the study of change" (cf. [Bates and Elman, 1992]).

1.2 The task of this paper

The present work is not committed to any of the theoretical stances on the status of the emergent structures that arise in a perceptually grounding learning system, but makes the following *methodological assumption*:

> Through examining the emergent structures of a perceptually grounded connectionist system in a relatively theory-independent way — in the cases when the system succeeds and when it fails in the learning task — it is possible to gain some insight into two usually opposed semantic phenomena: *polysemy* and *compositionality*.

[2] Chapman's work [Chapman, 1990], is a counterexample, but his "concrete-situated" approach to perception, action and language use, characteristically comes very close to both learning and connectionism, while operating in a much more realistic domain than, for example, the work discussed in this paper.

1.2.1 Polysemy

The first and major aim of this work is to provide some empirical evidence about the nature of the controversial notion of *linguistic polysemy* or "multiplicity of meaning". A typical example of polysemy would involve a word like 'foot', which has undoubtedly different meanings in the contexts of sentences like:

> The house lies at the foot of a mountain.
>
> My foot hurts.

Still, the two uses of 'foot' seem to be not unrelated, and a possible *motivation* for the relatedness of the polysemy of 'foot' would be a (fairly degenerate) 'mountain as body' metaphor (cf. [Lakoff and Johnson, 1980]).

Polysemy is most often contrasted with *homonymy*, on the one side, and *generality* or *vagueness*, on the other. The first of these is usually characterized as lacking the property of "relatedness", so that the senses corresponding to the phonological string would be totally different and unconnected with each other, the standard example in English being /bank/ (river bank vs. financial institution). *Generality* is said to be the property of words to be *neutral* with respect to certain distinctions, leaving these to be "filled in" by the background knowledge of speakers and hearers in a given context. For example in the sentence:

> John cut the grass.

— the action of cutting could be performed in a great variety of ways: with scissors, with a razor blade, with a kitchen knife, horizontally, vertically, diagonally etc. No one would claim that these are "distinct senses" of *cutting* though, and in a culture like ours only one or two interpretations e.g. "by using a lawn-mower" would be appropriate. It is sometimes said that polysemy and homonymy, despite possible differences, involve *intercategorial ambiguity*, i.e. between "related" or "unrelated" but separate categories/concepts while generality is *intracategorial ambiguity*, that is, involving variations within a concept such as e.g. *cutting*.

Polysemy has been a phenomenon about which the formalist approach has had practically nothing to say, with its insistence on well-defined, and disambiguated meanings. However, a recent school within linguistics with unmistakably anti-formalist leanings, namely *cognitive linguistics* (cf. [Langacker, 1987; Lakoff, 1987]) has taken polysemy — and the phenomenon of *metaphor* with which it is closely related — as an important indicator of the nature of language and meaning. After all, polysemy is usually characterized as *the property of certain lexical items to have a number of distinct, but related meanings* [3]. So to explain polysemy, one would have to give answers to questions such as:

> What *is* a "sense"/"meaning" of a lexical item (word)?
>
> In which ways can "distinct" senses be "related"?
>
> Why should the words of human languages display polysemy at all?

[3] [Regier, 1992], for example, introduces his section on polysemy with the sentence "Polysemy is a linguistic phenomenon whereby a single word has a cluster of distinct but related senses." (p.21)

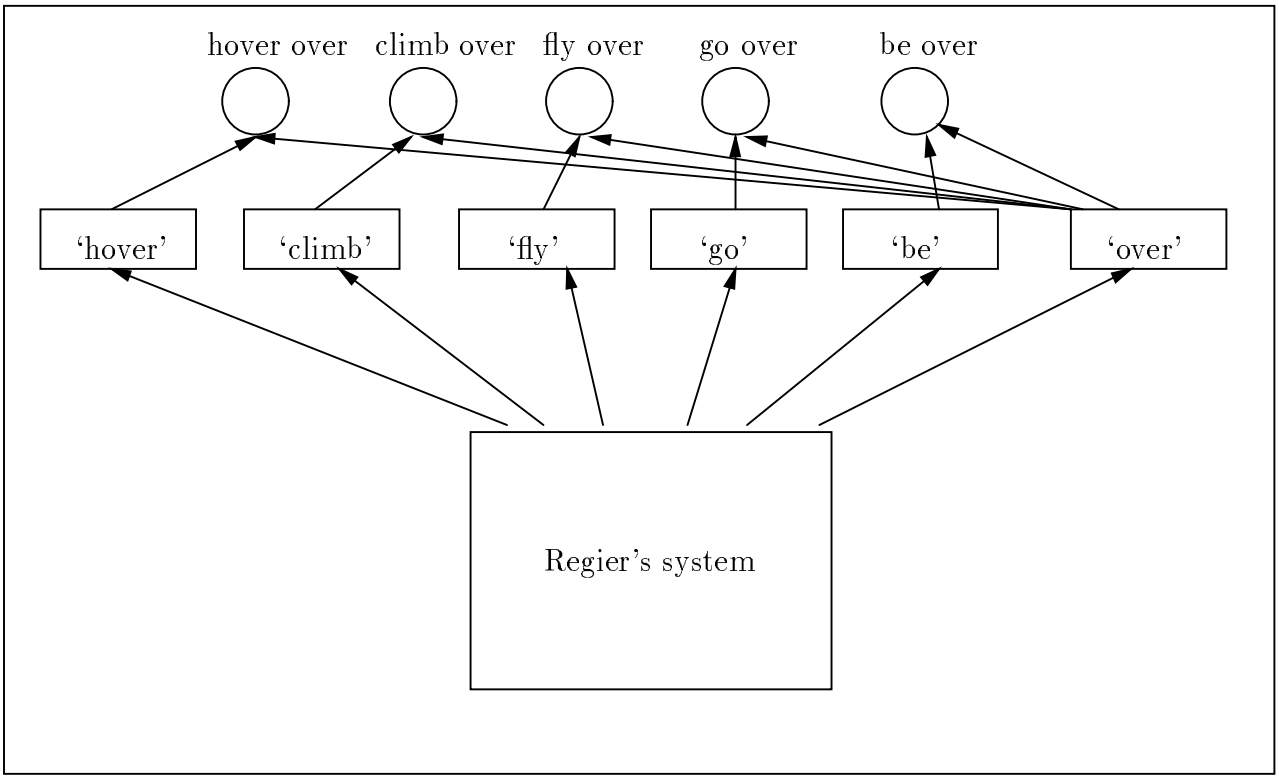

Figure 1: A very schematic overview of the perceptually grounding system that emphasizes the contextual character of polysemy: the activation patterns that form in e.g. the 'over'-layer depend on those that it combines with.

Two very broad types of answers to questions like the above are those that regard meaning as *reified* — "meanings" as (discrete) objects standing in certain relations to each other and subject to manipulations — and those that see meaning in *context-dependent, process* terms, where the interaction between the semantic contributions of the various elements and their coupling with the situation is the sourse of meaning of linguistic expressions. The perceptually grounded approach adopted in the present study and some of the more concrete results from it, support and aim to explicate the latter view.

More on this issue will be said in section 5, but for an idea of what the argument will be based on, the reader can glance at Figure 1. What is referred to as **Regier's system**, is the connectionist system of Terry Regier [Regier, 1992] that is described in section 2.2. In experimenting with different extensions of it, I found that *the most successful of these* had the structure displayed in Figure 1: there are separate layers for every "word" the system is to learn, and the activation vectors in one such layer will be crucially dependent on the activation vectors of the layers it combines with. With such an architecture, something

resembling a "polysemous structure" can be observed in the activation patterns of the units in the different layers, using techniques such as *hierarchical cluster analysis* (cf. section 3.1).

1.2.2 Compositionality

The other phenomenon discussed in this paper is, on the contrary, possibly the most often addressed one in formalist semantics (at least of the logical variety), namely *compositionality*. The claim often made is that unless there is a clearly defined function for "deriving the meaning of the whole, from the meaning of the parts", language comprehension — especially of novel sentences — would be impossible. And this claim often continues in saying that non-formalist approaches to language are doomed from the start since they have no way of accounting for these "structure-sensitive" operations of language (cf. [Fodor and Pylyshyn, 1988]).

While my emphasis is on polysemy, I will try to show that some experiments with the connectionist perceptually grounding system sketched above, used for example for learning the meaning of 'over' and 'under' in different contexts suggest a way of how clearly non-discrete, distributed activation patterns can combine, producing what can be interpreted as two-word sequences (as suggested in e.g. Figure 1). The system can also be trained on e.g. examples of 5 word-pairs, and interpret appropriately a sixth "novel" scene, by allowing the weights leading to this node to share with appropriate weights to the other output nodes.

1.3 Plan of the paper

In **Section 2** I will present in brief the L_0 project and more specifically the work of Terry Regier. This is done for the purpose of providing a concrete example of the perceptually grounded approach and an understanding of the context in which the present work arose. The question of how polysemy enters the particular microdomain is also addressed.

Section 3 describes the empirical studies of polysemy done using modifications of Regier's system and **section 4** will briefly discuss some of the preliminary results regarding compositionality.

Section 5 will sum up the results from the studies and suggest some theoretical conclusions about polysemy and perceptually grounded semantics in general.

2 The L_0 project, Regier's system, and the problem of polysemy

2.1 The L_0 project

The task of the L_0 project has been described in [Feldman *et al.*, 1990] as that of "Miniature Language Acquisition": the goal is a system that would learn to associate 2-dimensional pictures of simple geometric objects with descriptions of these pictures in a variety of different languages, with the hope of making the system flexible enough to be able to learn any of the semantically relevant distinctions that *any* human language can make within this

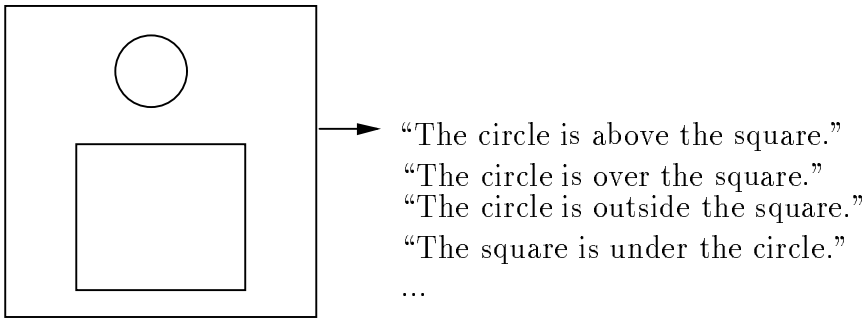

Figure 2: The L_0 task: learning to associate scenes with descriptions in an arbitrary language (here English).

---

limited domain. So the 0-subscript can be seen as suggesting both that in many ways the languages learned would be "micro-languages" with respect to the actual human languages they are meant to reflect (because of the "micro-world" they are to learn to describe), but also the desire to achieve a language-neutral system, which would shed insights on the human capacity of language learning in general.

An illustration of the kind of picture, and kind of sentences (in English) the L_0 system it to learn to associate is given in Figure 2.

2.2 Regier's model of perceptual grounding

The work of Terry Regier [Regier, 1990; Regier, 1991a; Regier, 1991b; Regier, 1992] is the most perceptually oriented within the project. Regier constructed a perceptually grounding connectionist system, which — in the case where the input consists of single pictures, rather than sequences of pictures — uses networks with a topology such as the one presented in Figure 3. [4]

After some preprocessing his system takes as input bitmaps of the two objects together with information on their orientation to one another (and to some reference orientations like *upright vertical*) expressed in angles. The system consists of 3 "modules". Both **D** and **ND** are heavily structured modules, having the job of extracting from the scenes presented as input *directional* and *non-directional* features, respectively. Examples of the directional features include *proximal orientation*: the orientation of the line connecting the two objects where they are closest, and the *major-axis orientation* of the larger object. Such orientations are compared with the reference orientations by the nodes in the *theta-node layer* and the resulting angles influence the output of module **D**. The output of module **ND**

[4] For learning dynamic spatial concepts like *through*, which require sequences of pictures ("movies"), the system has an extra module that takes input from **D** and **ND** and gives output to **U**. Even though the experiments with polysemy reported here used the extended system, this *motion module* was not the focus of any attention, and so for the sake of simplicity will not figure in the presentations in this paper.

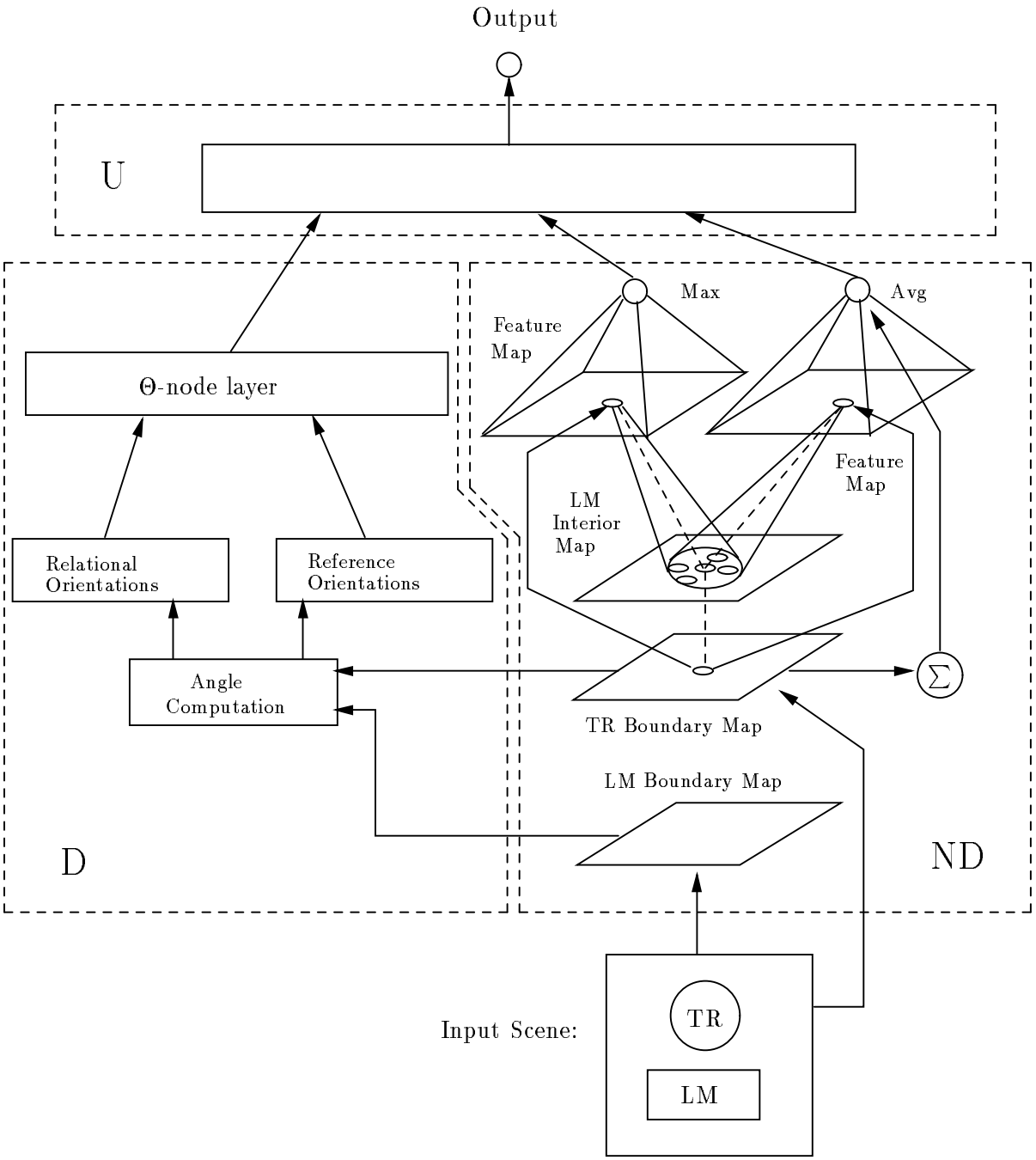

Figure 3: A graphical presentation of a typical architecture in Regier's system. (borrowed from [Regier, 1992] p.62). Not shown is the *motion module* which preceeds **U** in the extended system.

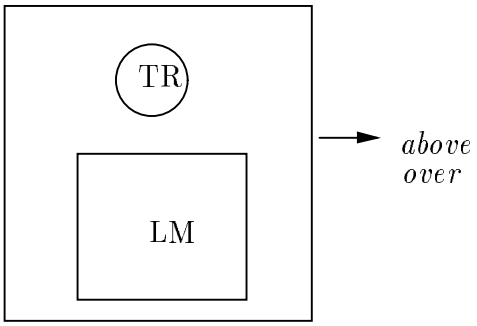

Figure 4: Regier's task: Learning to associate scenes with spatial relations between the object marked TR (trajector) and LM (landmark) that are linguistically expressed in an arbitrary language (here English).

---

is determined, above all, by the degree of overlap between the two objects, allowing for the system to learn features such as *contact* and *inclusion*. As to module **U**, Regier conceived it as an unstructured upper layer, leading to the output node. On the whole, the system is a representative of what Regier calls "partially structured connectionism":

> "This partially structured design is an attempt to capture some of the best features of both structured and unstructured network design, namely:
>
> > — the *tractability* in learning and *enhanced generalization* that results from structuring, as the dimensionality of the search space is typically dramatically reduced, and
> > — the flexibility that results from an unstructured, fully-connected network design. Flexibility is clearly a critical feature of a system which must be able to adapt itself to the spatial system of any natural language." (p. 63)

Using a version of the backpropagation algorithm [Rumelhart *et al.*, 1986], Regier's system is to learn to output the appropriate *spatial relation* between the two objects. The subtask performed by Regier's system is illustrated in Figure 4 [5].

So, Regier essentially simplifies the L_0 task in two ways:

First, he avoids the question of object perception and categorization, by having only two objects and prespecifying which one is to be regarded as the *trajector* i.e. the more focal object whose orientation is to be described *with respect to* the other object, the *landmark* [6].

[5] It is not the purpose of this paper to give a presentation of Regier's work, which includes learning static relations like *above*, dynamic like *through*, learning without explicit negative evidence, which are all features which his system performs. A fully comprehensive exposition of this and more is [Regier, 1992], warmly recommended for both linguists and connectionists. Some more details will be given along the way, as they become relevant.

[6] The terms "trajector" and "landmark" where originally proposed in [Langacker, 1987].

Second, and importantly for the work reported in this paper, Regier's system can be said to learn spatial relations or "concepts", that are expressed in different languages, but (in my opinion) can not be said to "perceptually ground" the *spatial terms* of the language, unless the sentences of the given language consist only of the triple *trajector, spatial term, landmark* (in any given order). My main criticism of Regier's work is that he uses "spatial concept" and "spatial term" interchangeably [7]. The point is that since Regier associates an event with a given spatial relation such as *over*, this relation is not to be mistaken with the meaning (or "meanings") of the word/grammatical element in the language that expresses it most often: in English — the spatial preposition, e.g. 'over'. Rather, it should be correlated linguistically with the whole description of the event, a sentence. And given that the trajector and landmark are pre-defined, the output of the system would be better interpreted as corresponding to, in English, the *verb-preposition* pair.

One of the main claims made in this paper hinges on the idea that the mapping between "concepts" and terms is not a direct and simple one. To avoid ambiguity I will use italics when talking about *spatial concepts/relations* and single quotes for 'spatial terms'.

2.3 Polysemy in the L_0 domain

Regier's system has proved successful in learning a variety of spatial concepts, expressed in a number of different languages, for example such that roughly correspond to English 'in', 'through', 'above'; German 'an', 'auf'; Bengali 'bhethoray', 'moddhay', Russian 'izpod' etc.

However, it has not been clear how to treat relations such as *over* described by English 'over' in the contexts of descriptions such as:

(1) The circle is over the rectangle.

(2) The circle is going over the rectangle.

(3) The circle is flying over the rectangle.

(4) The circle is hovering over the rectangle.

(5) The circle is over the rectangle from the square.

(6) The rectangle fell over.

(7) The square is lying over the rectangle.

(8) The circle is moving (all) over the square.

If we were not constrained by the limits of the L_0 domain, the list could go on much longer: As [Brugman, 1981] and [Lakoff, 1987] have demonstrated in a descriptive framework and [Harris, 1989] in a connectionist, but non-perceptual system, English 'over' is a perfect example of *polysemy*. Even if we restrict ourselves to (1-8), however, it is obvious that

[7] cf. [Regier, 1992]: "each movie has been correctly labelled as a positive instance of some *spatial concept* from a natural language... This system concerns the design and construction of a system which takes a set of such movies, each labelled as a positive example of some natural language *spatial term* from some language and learns the association between the words and the events or relations which they describe." (p.1, *my italics*)

there would be considerable variation in the pictures they would describe — perhaps great enough to make *over* unlearnable by the kind of system envisioned by the L_0?

Before this question can be empirically investigated using Regier's model, however, we must limit the set of "senses" of *over* further. First of all, it is unrealistic to expect an adequate notion of *covering*, which seems to be necessary for (7-8), to be learned from only 2-dimensional pictures.

Sentences like (6) involve according to [Brugman, 1981] a *reflexive landmark*: a single object is viewed as both trajector and landmark, thus is oriented with respect to itself (at a different location). Such sentences, though more clearly within the domain, are beyond the scope of Regier's system because of its requirement that trajector and landmark are two distinct, and pre-labelled objects.

Likewise with (5), which would, again according to [Brugman, 1981], be an instance of *end-point focus*, which is the term used in cognitive linguistics when the end-point of a *path* — in the case of (5) the imaginary path from the square to the circle — is especially prominent. The problem with such uses of *over* for Regier's system is that not only do they involve (at least) 3 objects, but one of them, in this case the square, has to function as *deictic center*, the reference point with respect to which a scene is construed. Since this point is most often the position from which a scene is viewed when the sentence is being uttered[8] this will conflict with the usual way scenes are regarded as being viewed, i.e. frontally. While [Regier, 1992] (7.1.2) discusses an extension to his system where there is a separate entity as a deictic center and with its help relations like *behind* are learned, the learning sessions assumed a "bird's eye" point of view. It is far from clear how to handle different perspectives in current perceptually grounding systems.

Apart from these shortcomings, there are a few that are less obvious, but all the more pervasive. I wish to mention two in passing:

(a) The abstractness of the objects involved (coupled with lack of a 3-rd dimension), leads to a loss of intuition on what is to be considered as an appropriate description of a scene — e.g should only the "feature" *contact* separate 'fly over' from 'go over'? This is related to the generally acknowledged assumption that spatial relations in human languages are not "purely spatial", i.e. disembodied from the physical and functional properties of the given trajectors, landmarks and motions.

(b) The lack of any role in the microdomain of gravity and other "force-dynamic" aspects in the relation between the objects (cf. [Talmy, 1987]), while such factors still obviously influence judgements of adequacy in the formation of *training sets* (the sets of pictures presented to the system as input during training) as well as the evaluation of the performance on *test sets* ("novel" pictures presented to the system, after it has learned the training set, in order to test how well it has generalized from the training sets). For example [Regier, 1992] acknowledges that the understanding that the trajector would fall on the landmark if released, should play a role in learning the perceptual semantics of relations such as *over* and *above*, but has no easy way of incorporating it in his system.

[8] cf. the example in the section on deixis in [Regier, 1992] (p.20) "San Francisco is just across the Bay Bridge." which is adequate if the speaker is in Berkeley or Oakland but not in Los Angeles (without prior "anchoring").

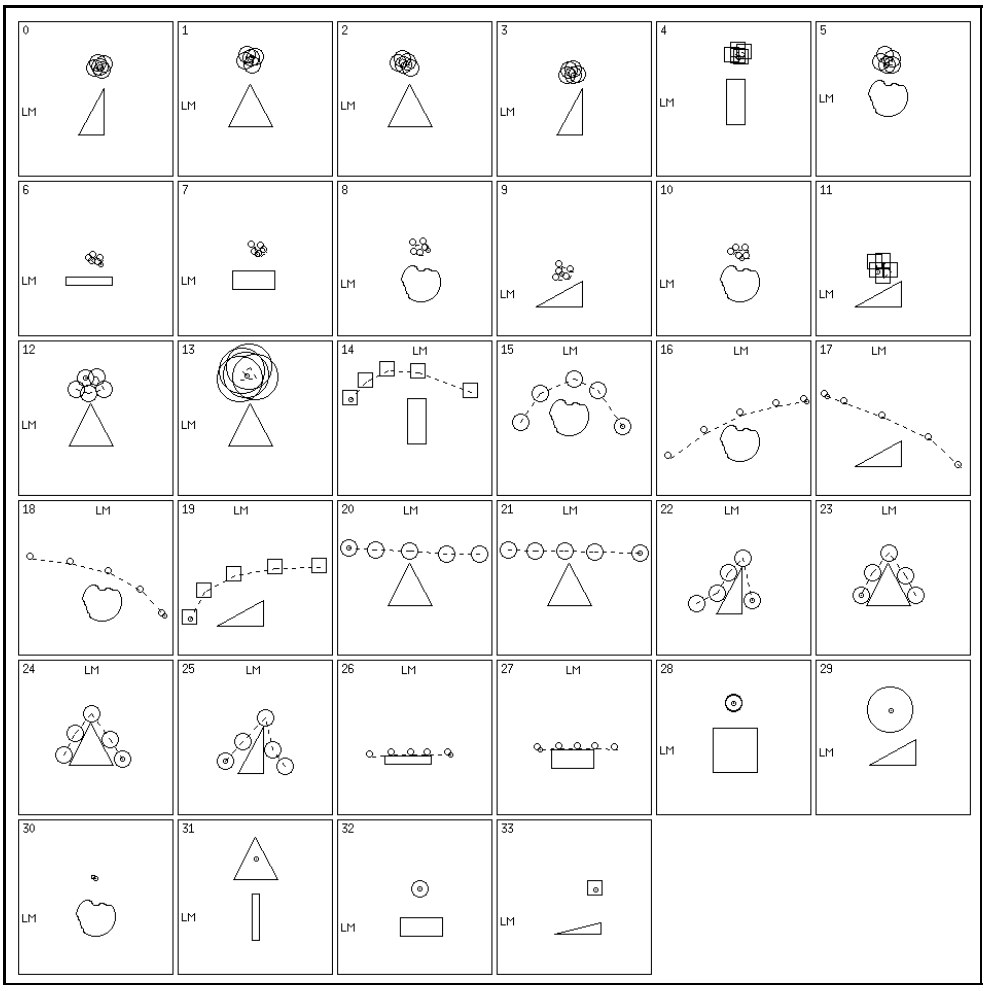

Figure 5: A set of scenes for "polysemous" *over*.

These reservations are serious and one possible, negative conclusion would be that it is pointless to try to deal with polysemy using Regier's system because (a) it misses some of the "most interesting" uses of e.g. 'over', and (b) those that it can (possibly) capture, it would do so "unnaturally". I, however, chose to adopt a less negative attitude about the undertaking. After all, there still seems to be enough perceptual information in the scenes that Regier's system "sees", to allow it to "understand" at least uses of *over* such as those that figure in sentences (1-4). Consider for example the pictures in Figure 5.

Examples 0-13 have a number of differences among them, but could all appropriately be described as cases of the trajector 'hovering over' the landmark. Examples 14-21, given the limited options to chose from would rather be 'flying over', while 22-27 is preferably described as 'going over'. 28-33 could be said to be instances of 'is over'. So in this set of examples, we can say that we have at least 4 different "senses" of *over*. The question of how to learn them, and once learned, how to "disambiguate" them, i.e. how does the appropriate

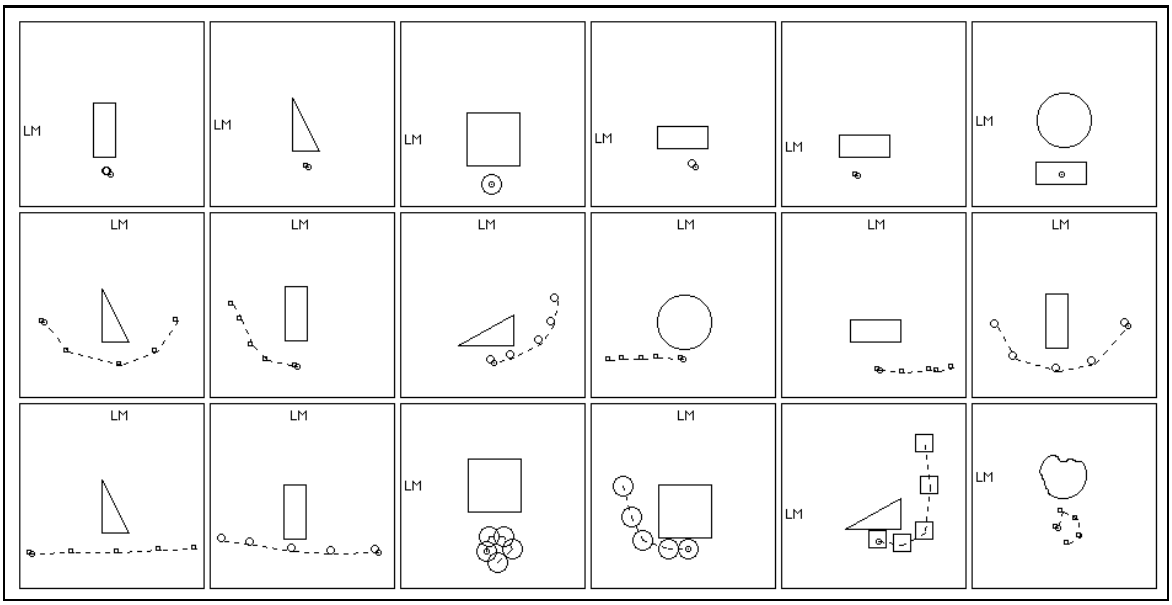

Figure 6: A set of scenes for "polysemous" *under*.

"sense" get "picked" in the right context, still remains open. This question, together with the more theoretical ones mentioned in the introduction, is the one that interested me the most, and I believe that it can be pursued even in the limited domain of Regier's system. (It is, however, possible that the answers obtained would not extend to a richer domain, so that some amount of caution is warranted.)

And unless it might be thought that the only spatial concept in English that offers an opportunity for studying polysemy in the limited domain is *over*, a similar set to the one above, but for *under* is shown in Figure 6.

The problem of the *context-dependency* of the meaning of words, even such ordinary words like 'over' and 'under' is of course far from being a problem for only perceptually grounded semantics. On the contrary, it poses what seem to be practically unresolvable problems to the formalist approach of denotational semantics and (strong) compositionality. The "subscript problem" described in 3.2 is, I believe, one such problem.

My claim is that a perceptually grounded, connectionist approach to semantics, by acknowledging the problems of *learning* different "senses" of words and that of *using the appropriate one* in the right context, and furthermore by showing that the two sub-problems are intimately related, can address "the problem of polysemy" in a much more successful fashion.

3 Dealing with polysemy in Regier's system

Figure 3 in the previous section presented a schematic overview of the kind of networks, or architectures, Regier's system operates with. With one exception, the experiments de-

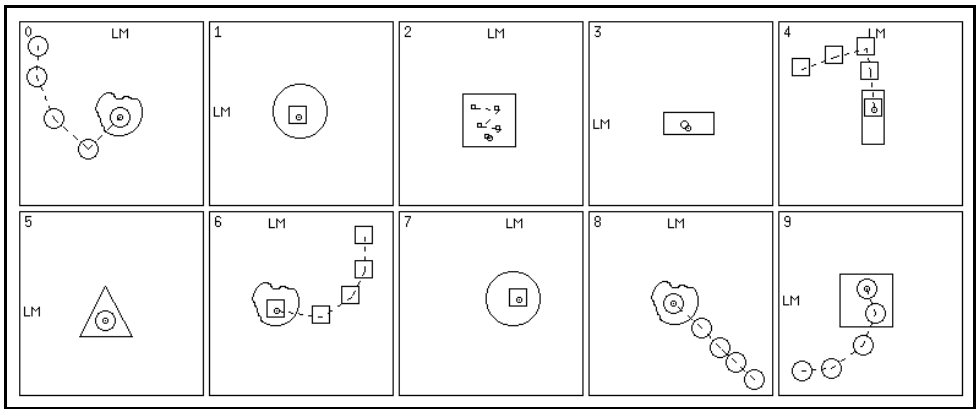

Figure 7: The positive examples of *in* in a typical training set.

scribed in this section did not involve changes in the structured part of the system, i.e. modules **D** and **ND**. However, the module Regier refers to as **U** has undergone a number of modifications so that the architectures that I have experimented with (altogether about 50) can be said to fall roughly into three groups, depending on the nature of this module:

— *uniform hidden layer* (UHL) architectures: those with a uniform hidden layer and one or more output nodes,

— *no hidden layer* (NHL) architectures: those without any hidden layer at all and with one or more output nodes, and

— *structured hidden layer* (SHL) architectures: those with a structured hidden layer and a number of output nodes.

The nature of and reason for these divisions will become clearer below.

3.1 Uniform hidden layer (UHL) architectures

Architectures with a single, uniform hidden layer in module **U** are the type Regier considered most suitable for his system. This is motivated by the "partially structured connectionism" approach mentioned above, in which a uniform, totally connected layer is to contribute for the flexibility of the system.

So far such architectures had indeed proved to be highly adequate. Relations, corresponding to e.g. English 'in', 'above", German 'an', 'auf', Bengali 'bhethoray', 'moddhay' (where the notion of "partial inclusion" is significant) were learned relatively unproblematically (cf. [Regier, 1991c], section 7).

For our purposes English 'in' is interesting, because at first glance it seems to involve something resembling polysemy. Figure 7 shows the positive examples of such a training set actually used by Regier. This characteristic training set includes examples where:

(a) the trajector is immobile within the landmark (scenes 1, 3, 5, 7),

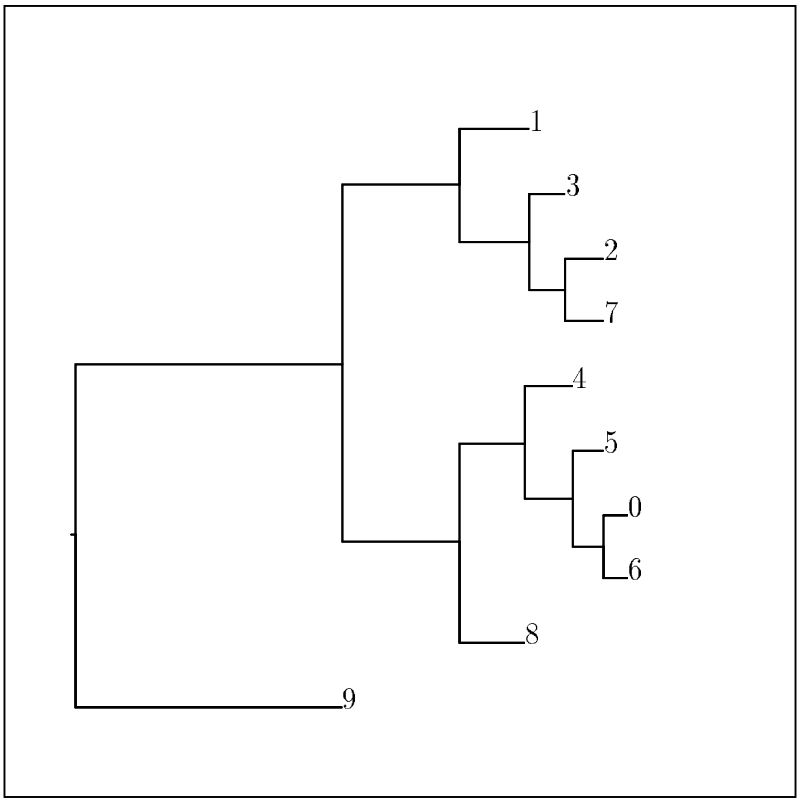

Figure 8: Hierarchical cluster diagram for the examples of *in* in Figure 7.

(b) the trajector enters the landmark (scenes 0, 4, 6, 8, 9), and

(c) the trajector is moving within the landmark (scene 2).

In presenting a training set consisting of these positive examples (together with an appropriate set of negatives) to the system, using a UHL architecture with 10 nodes in the hidden layer, the error fell below 0.01 in less than 30 epochs, i.e. extremely quickly. Once this was done, a natural step to take was to examine the activation patterns of the hidden layer, using some of the techniques developed by connectionist modelers, and applied for similar purposes by e.g. [Elman, 1988], [Harris, 1989] , and [Bartell and Cottrell, 1991]. These techniques include *hierarchical cluster analysis* (HCA), *principal component analysis* (PCA) and a simple method for display of the activation level of units, developed by Terry Regier; each will be explained when used in a particular case bellow.

Equipped with such tools, it was easy to see that the architecture had successfully learned the training set in Figure 7, but had not in any (intuitive) way learned to "cluster" the examples is sub-groups, and certainly not along the (a), (b), (c) division made above. Figure 8 , for example, shows that — if the activation patterns that develop in the hidden

layer once the training set has been learned are taken as indicative of categorization [9] — scenes like 2 (type c) and 7 (type a) are judged to be very similar, while comparatively different from 5 (also type a), which is most similar to scenes 0 and 6 (type b). (The interpretation of HCA diagrams is very simple: the vectors corresponding to the activation levels of a given set of nodes are compared and the "closer" these vectors, the closer will they be placed in the binary tree of the diagram.)

The reason for this odd "polysemous structure" seems to be that, encouraged by the fact that it learns only in *the last time step* of motion sequences, [10] the system "decides" that the trajector being within the landmark interior at the end of a sequence (if the trajector is static, all steps are identical), is all that is essential for the relation between the two objects to count as *in*, and all other details in the "movie": e.g. whether the trajector is moving or not, whether it entered, or not, whether it entered from the right or from the left, are considered more or less equally (un)important. This notion of *in* is neutral with respect to these features.

This is similar to the notion of *cutting* which was given as an example of *generality* in section 1.2.1, essentially neutral with respect to manner, direction etc. of cutting, but once placed in the context of a particular utterance would be "specified" as to these attributes in accordance with the conventional situations in which such utterances are used. The main problem with saying that Regier's system in learning the set of Figure 7, has learned the "meaning" of the English word 'in', is that no hint is given as to how this contextual "filtering" is to be done.

Another way to describe how the network learns *in* is to focus on the common aspect in all scenes, *inclusion during the last time step*, and say that the system has "found an abstraction" — it has picked out a characteristic feature present in all possible examples and absent in all negative examples which suffices to correctly discriminate between them. The fact that the activation vectors for the 10 positive examples in the training set (despite small differences which yield the clustering in Figure 7) are practically identical is shown using the activation display over the hidden layer for the last time frame: cf. Figure 9. In this figure the ten activation vectors, each corresponding to an example of *in* in Figure 7, are shown to be almost the same. (The size of the circles shows how strongly each unit of the 10-node hidden layer is activated.)

The natural question to pose is: How would the system manage when there is no such abstraction (or at least it is not obvious what it would be)? For an answer to this question, *over* was an obvious candidate to experiment with.

For the purpose of analysis, I distinguished between 4 "kinds" of *over* that were all within the domain of the system (as discussed in section 2.3):

[9] It should be mentioned that this and the other analyses of the activations of the hidden layers presumes that categorization is done on the basis of *similarity*. [Lakoff, 1987], for example, has argued that there are other factors such as mataphorical extensions and "image-schema transformations" that play an important role in categorization.

[10] This is an important structural feature of Regier's system, captured by the nature of the *motion module*. While activation is propagated forward for all steps in a "movie", the error is estimated and "back-propagated" and so weights are adjusted only for the last, resultant, step. This leaves Regier, in the good linguistic tradition, with a more *constrained* system, allowing him to make some predictions concerning the learnability and occurrence of certain spatial relations in human languages.

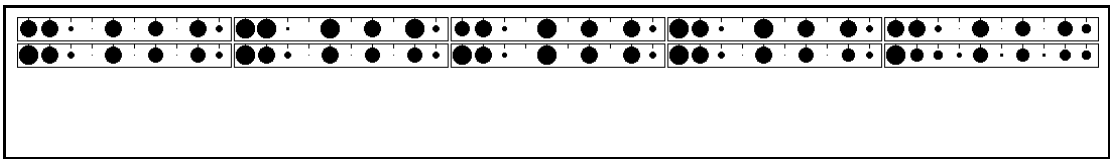

Figure 9: A display of the activations of the 10 units in the hidden layer at the last time frame, when tested on the positive examples of *in* (cf. Figure 7) after this spatial relation has been learned alone in a UHL architecture: the vectors are practically identical.

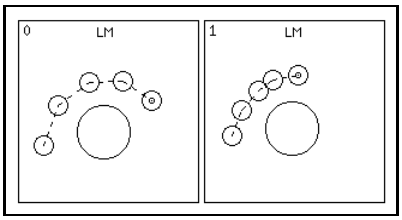

Figure 10: A positive and a negative example of *over*, that could not be separated with the original features.

(a) "static *over*" (described e.g. as "The circle is over the square.")

(b) "across-*over*" where the trajector moves from one side of the landmark to the other, touching the landmark (e.g. "The circle is moving over the triangle.")

(c) "flying-*over*": like (b), but without any contact (e.g. "The circle is flying over the triangle.")

(d) "hovering-*over* where the trajector is moving about above the landmark, without going very far in any direction: (e.g. "The circle is hovering over the square.")

Then, I constructed a number of training sets which consisted of one or more of these types of *over*, each type being represented by at least 6 examples. The maximum number of such positive examples was 34, the minimum 6. All sets included a uniform sub-set of negative examples that included scenes describable using English words such as 'on', 'to the left of', 'around', 'through', 'above' (in its "dynamic sense", since its "static" is partially synonymous with *over*, cf. Figure 4): altogether 80 scenes.

After multiple attempts at training, only sets that included positive examples of type (a) and (d) converged at all. Not even the other two "non-polysemous" kinds managed well. This pointed towards an anticipated difficulty not directly related to polysemy but detrimental for anything like an intuitive treatment of terms like 'over' and 'through' — the system lacked a notion of *sideness*.

For instance the system — even though in principle it should be able to distinguish between examples like those in Figure 10, which were given as a positive and a negative example of *over* in the training sets, using the available features (in this case, probably

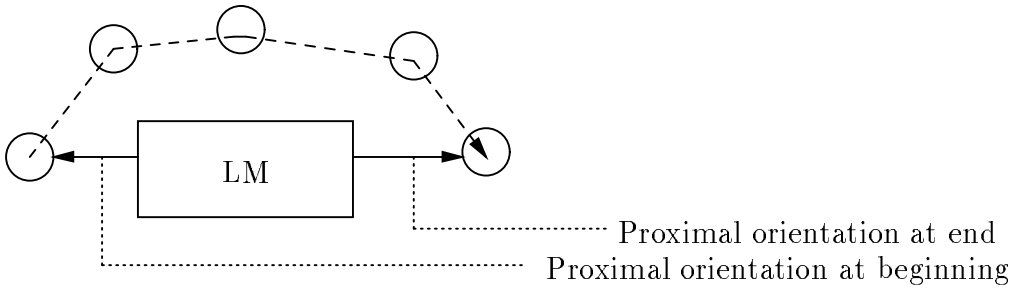

Figure 11: A situation that would yield a maximal response for the theta-node comparing the proximal orientation during the first time step with the current one.

---

alignment with upward vertical) — had considerable difficulty in doing so and time after time failed.

I did not come up with a good solution to this problem, but in order to be able to deal with (b) and (c) examples of *over* and see how they interact with "static" and "hovering" ones, Regier and I augmented the system (module **D**) with *two more theta-nodes*. One of these compares the current proximal orientation with the proximal orientation at the first time frame of the sequence (which is thus regarded as a reference orientation) and returns a maximal response when the two are diametrically opposed. So if there is a strong response of this "feature" during the last frame of the sequence (which is most important for learning) this would signify that the trajector is on the other side of the landmark from its starting position. Figure 11 shows a situation in which this node would be maximally active. The other theta-node performs the same kind of comparison between tbe *center-of-mass orientations* (the line connecting the centers of mass of the two objects) of the current and first time frame.

With the revised system and architectures that had uniform hidden layers with 10, 20 and 30 units, the system did better than before. It learned all the non-polysemous training sets much easier and much better than before and seemed like it had some success with polysemy: in less than 100 epochs it converged on a training set that included "flying" and "hovering" examples of over, type (b) and (d) according to the description above. Figure 12 shows such a training set.

This time, unlike with *in*, both hierarchical cluster analysis (HCA) and principal component analysis (PCA) [11] showed that the network had indeed learned to distinguish between (at least) two types of *over* examples, and these corresponded approximately to the "hovering" and "flying" kinds of *over* (cf. Figure 13 and Figure 14).

[11] This is how e.g. [Bartell and Cottrell, 1991] describe this latter method: PCA is used "...for the decomposing a very large sample of data vectors (in our case, vectors of activation). The spanning vectors are chosen by principle components such that the first component is along the axis of maximum variance in the data, the second along the second largest variance, etc. This decomposition can potentially localize the highly distributed representations in a set of hidden units (...), and particular components can often be correlated with the environment features (empirically)."

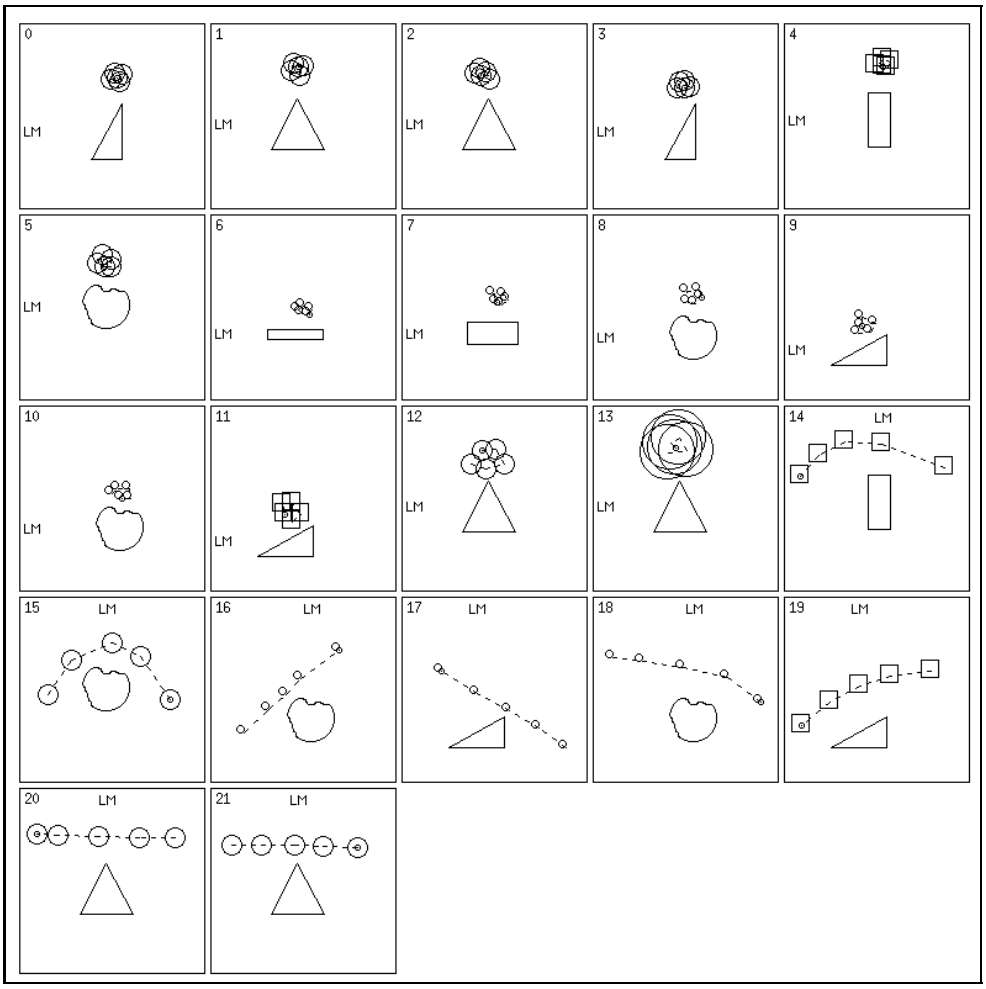

Figure 12: The positive examples of a set with "hovering" and "flying" types of *over*.

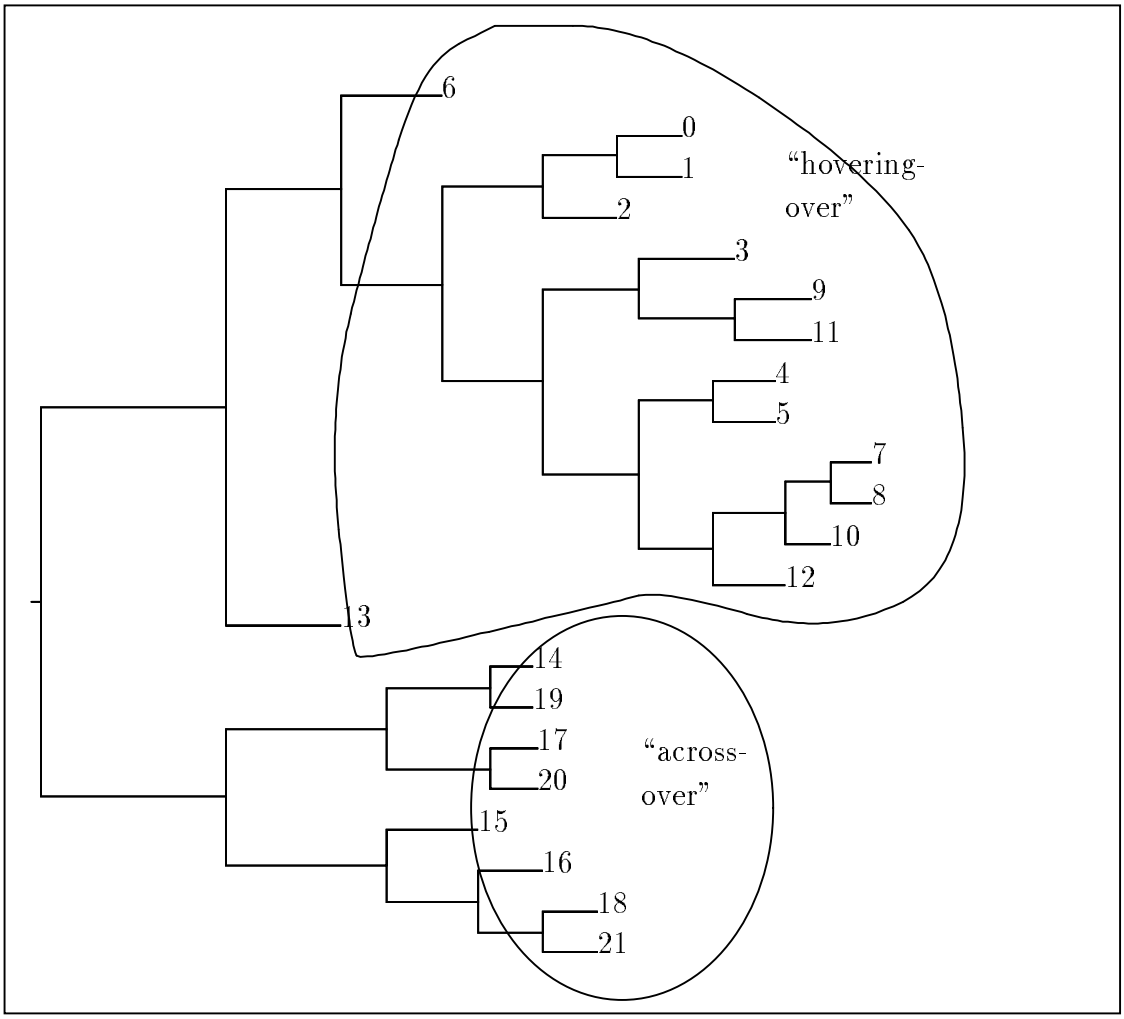

Figure 13: Hierarchical Cluster Analysis (HCA) diagram for the examples in Figure 12.

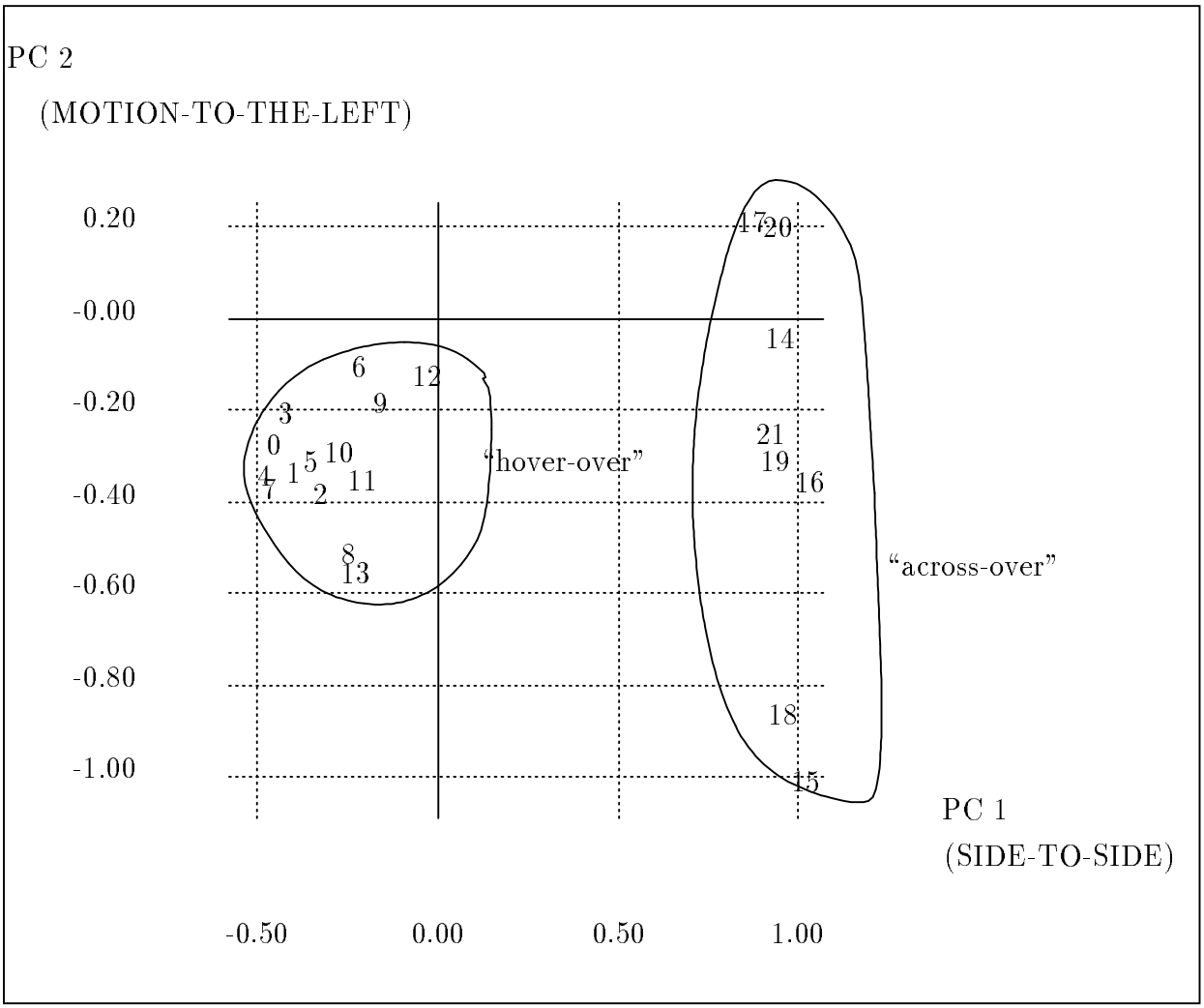

Figure 14: Graph displaying Principal Component Analysis (PCA) for the first (PC1) and second (PC2) principal components of the examples in Figure 12. These appear to be respectively: *motion from side to side* and *motion to the left.*

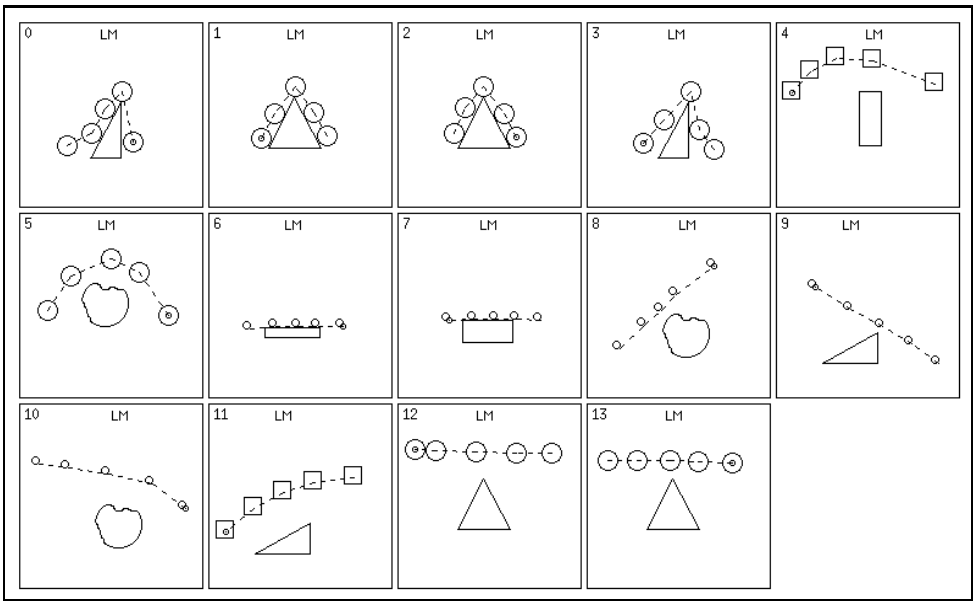

Figure 15: The positive examples of a set with "flying" and "going" examples of *over*.

Not every combination of two "types" of *over* yielded such clusterings. For example in learning a set of "hovering" and "static" examples, the situation was parallel to that of *in*: the network seemed again to establish a "general concept", an "abstraction" and there were no clear differences between the "senses".

However, the learning behavior described in the preceding two paragraphs is hardly one of "disjunction" vs. "abstraction". An interesting case, somewhat in between, was provided by a set that included "across-over" and "flying-over", i.e. with and without contact (cf. Figure 15).

The system did learn this set with a 30-node UHL architecture in about 150 epochs and I was expecting it to cluster the examples, based on the difference contact/no-contact. However, as the HCA diagram shows (cf. Figure 16), the most salient distinction is rather that of direction of motion.

There were a few more partial learning successes for architectures of this type, but they failed consistently when there where more then two of the four kinds of *over* in a single training set.

So the conclusion that I draw concerning UHL architectures with one output unit, both from cases when they learn and when they don't, is that with a single, uniform hidden layer, there is always strong pressure for homogeneity, i.e. of leveling the differences between examples. The result is that either the differences between types are entirely ignored or a loose gradient structure emerges, where the examples are distinguished along some feature, perhaps unintuitively. If any of the three terms from the introduction applies to this phenomenon, it is not polysemy in the classical sense, but *generality*. However — if Regier's system provides an adequate model of human learning — this makes a claim

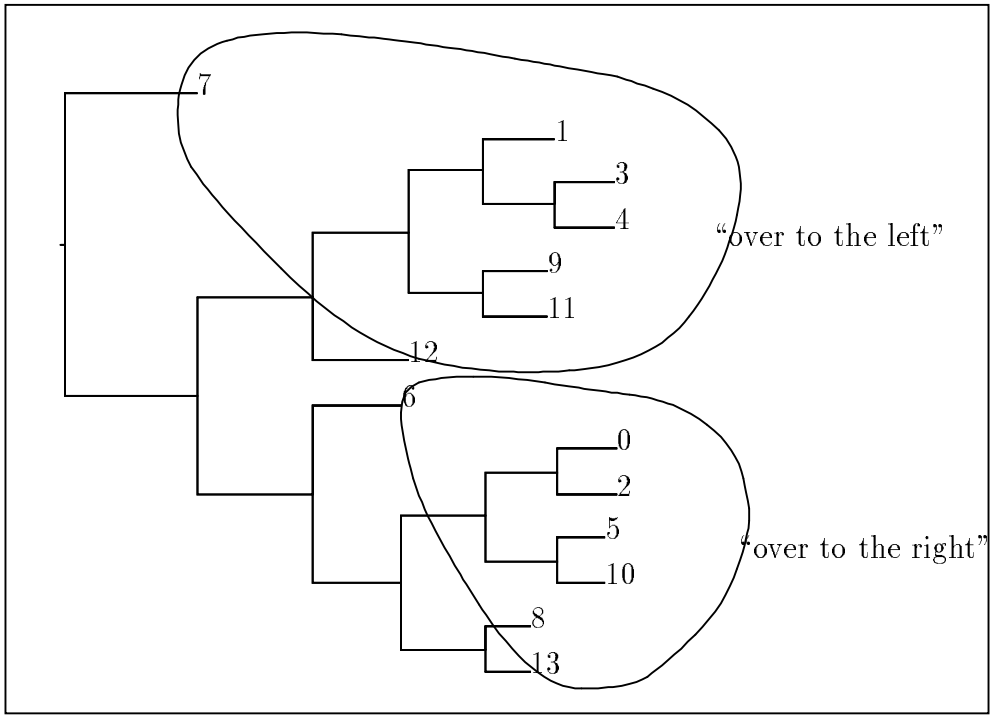

Figure 16: Graph displaying HCA of the examples in Figure 15 showing that the examples have clustered based of *common direction of motion*.

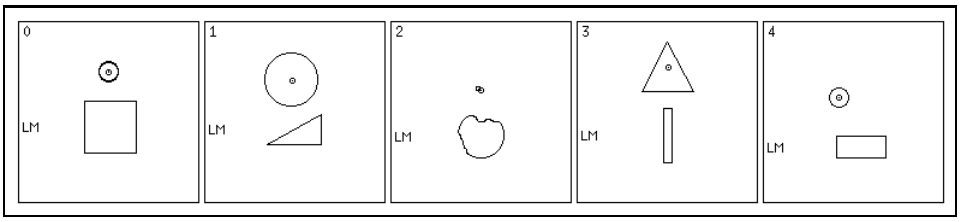

Figure 17: Positive examples from the "be-over" set.

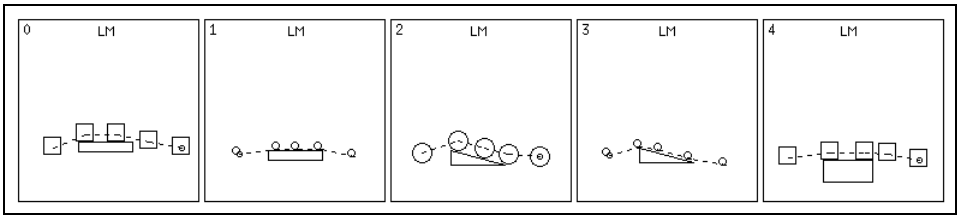

Figure 18: Positive examples from the "go-over" set.

about the scope of linguistic generality: namely, that without contextual information, only very limited differences can be tolerated within a single category. I will return to this in the section on "structured hidden-layer architectures".

The picture was somewhat different, however, when I presented as input to the system, not one single training set that included all the types of *over*, but separate sets and for every set — a corresponding output node. The system could learn sets with examples, describable as "static-over", "across-over" and "flying-over" and even these together with a set that could be described as "climbing-over", (thus distinguishing these examples from "across-over", depending on the major-axis orientation of the landmark) with a UHL architecture with a 30-node hidden layer and 4 output units.

Figure 17, Figure 18, Figure 19, and Figure 20 show the kind of positive examples in these sets. When analysed, for example with HCA, the hidden layer shows very reasonable clustering of the scenes of the "polysemous" set from Figure 5, here used as a test set. (cf. Figure 21).

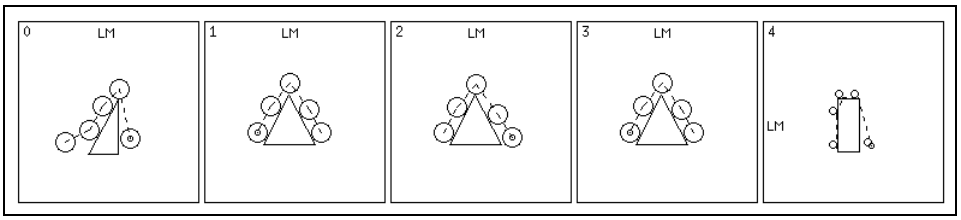

Figure 19: Positive examples from the "climb-over" set.

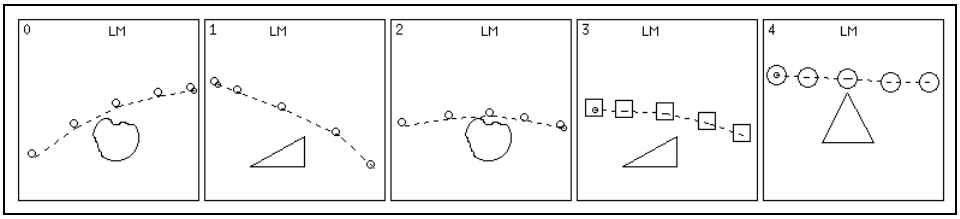

Figure 20: Positive examples from the "fly-over" set.

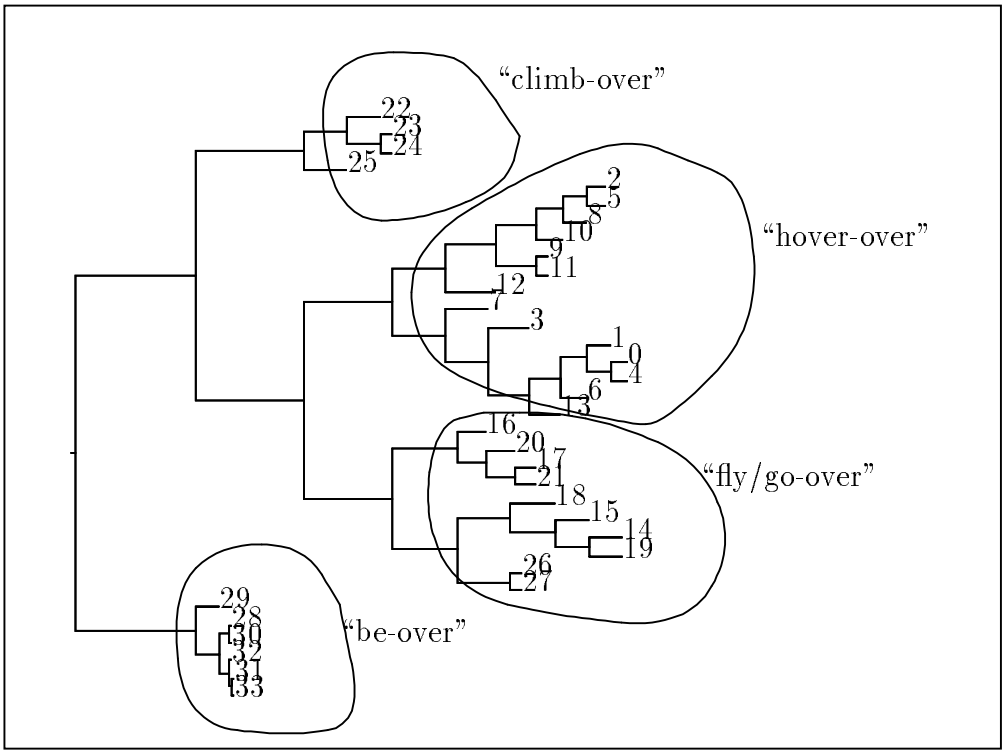

Figure 21: Graph showing HCA of the examples from Figure 5 used as a test set, after the network was trained on the sets exemplified in Figures 17-20, using a UHL architecture with 4 output nodes and a 30-node hidden layer.

The drawbacks with this type of architecture were that it still could not learn all the different kinds of *over* with a single set of weights — it would not converge whenever a "hovering-over" set ("*over* type d") was introduced together with the other three. The other more important problem, is that if the different output nodes are interpreted as different "senses", it becomes unclear how they are separated in learning and in comprehension. If, on the other hand, they are rather seen as verb-preposition pairs, as I briefly suggested in section 2.2 and as I will be arguing for below, then something must be said about the individual contributions of the verb and the preposition. These issues will come up again later, but before that I will discuss a "minimalist" hypothesis: doing without a hidden layer entirely.

3.2 No hidden layer (NHL) architectures

The idea of using NHL architectures (i.e. such with no **U** module at all) came about as a coincidence. In experimenting with hidden layers of various sizes, I made a mistake in the declaration of the number of units: I specified one less. The result was that the last unit in the "hidden layer" became in fact the output unit. The surprise came when I noticed that this by chance created architecture was learning much faster and better than any previous one! It learned in less than 100 epochs training sets that had up till then been unlearnable, like such that included all 4 kinds of *over*.

The fact that a NHL architecture seemed so good at "polysemy" led me to consider such issues as the relation between polysemy and *linear separability*. It is well known that hidden layers are necessary for learning problems that are at least of the complexity of *exclusive logical or*, where the presence of any of two "features" should be correlated with a positive response but when both are present or absent, the response should be negative. Perhaps the "senses" of polysemous concepts do not require anything of that complexity. A counterexample would be if there were e.g. examples of *over* that involved *motion*, and examples that involved *contact*, but there were non which involved both motion and contact — which is of course not the case. So maybe the hidden layers in architectures like those described in the previous subsection are totally unnecessary and maybe even harmful in arbitrarily increasing the "search space"?

I refrained from drawing the above conclusion, because given the features that Regier's system operated with, it was trivial to come up with counterexamples to the "linear separability hypothesis" regarding polysemy: *on the side of*, for example, can involve alignment with left horizontal or right horizontal, but obviously not both. Now, one might argue that this could show that the features used are inappropriate, rather than anything else, but I declined from pursuing this line of thought since NHL architectures seemed indeed to lack the kind of flexibility that module **U** was intended to yield:

The positive results with single polysemous training sets and single output nodes were quite stable as long as I used a particular set of landmarks, where triangles and "blobs" predominated (see e.g. Figure 5, again). As soon as I changed the "blobs" with more regular shapes, such as squares, rectangles and circles the performance of the system sharply deteriorated. [12]

[12] One possible reason for this was a fault in the representation of the major-axis-orientation (MAO)

So NHL architectures with one output unit did not have anything conclusive to suggest about polysemy. Despite occasional differences, UHL architectures and NHL architectures had performed roughly equally well with polysemous *over*, that is, not very well. There was, however, one noteworthy difference: when multiple training sets with non-polysemous examples of *over* were presented to a NHL architecture (with a corresponding number of output nodes) it converged within 50-60 epochs! I increased the training sets to 6 and still the network seemed to converge with ease.

Actually, the explanation for this is, I believe, straightforward: the different "types" of *over* that were represented in the training sets were not related in the network in anyway at all, and so did not interfere with each other. All the sets of weights from the theta-node and feature-map layer to the different output nodes were disjoint. So, one may say, this "treatment" of polysemy, actually reduces it to *homonymy*. This view is enhanced if one interprets the output nodes as $over_1$, $over_2$, $over_3$ ... i.e. phonological strings that "by chance" have been associated with events that are — at least according to the system — totally unrelated. This reduction of polysemy to homonymy, or "mere ambiguity", is reminiscent of formalist semantics: if we only disambiguated the meanings of words, say by a subscript, then perhaps we could define them with necessary and sufficient conditions, and combine these unambiguous meanings compositionally... Both formalist treatments and this type of connectionist architectures are, however, faced with, what may be called "the subscript problem":

Far from offering an advantage in learning, treatments of polysemy that separate the "senses" arbitrarily only push backward, or "inward", the question of how a given event, in being described with a given word, e.g 'over' is associated with "over-with-the-right-subscript". This makes it very hard to explain both learning — after all, not even analytic philosophers talk to their children in something like the manner of: "Don't walk over-subscript-one that puddle, jump over-subscript-two it!" — and comprehension, which thus seems to require an enormous amount of "ambiguity resolution" and "inferencing".

A beginning to an answer to the subscript problem can be found if something else than a subscript can be shown to distinguish the "senses" of polysemous words on the linguistic level. The humorous example above, and the way the "types" of *over* have been described and exemplified so far provide a straightforward suggestion: look at the rest of the sentence and — in the domain of Regier's system — above all at the verb, rather than just at the preposition, in the analysis of the polysemy of a lexical item.

In a number of tries (all successful) I, for example, regarded the output nodes not

for objects such as squares and circles. To test this, MAO was removed altogether from a number of architectures with the consideration that they ought to succeed in learning *over* anyway, since MAO is probably not essential for learning the restricted uses of *over* under consideration. The result was, however, quite ambiguous: only once so far has such an architecture come close to learning a training set that included all 4 kinds of *over* when the error fell to 0.05, only to go up afterwards. Otherwise, "no-MAO-architectures" have been just about as successful as the hidden-layer architectures described in the previous section, and in some cases less. On the other hand, with a consistent (though arbitrarily aligned with the horizontal for squares and circles) MAO and the original polysemous training set, the network did converge once (from more than 20 tries). This might signify that MAO is important, especially, for the "go-over" examples, but without in detail analyzing the individual weights in the successful and unsuccessful cases, (which I haven't done) I can hardly say anything more detailed.

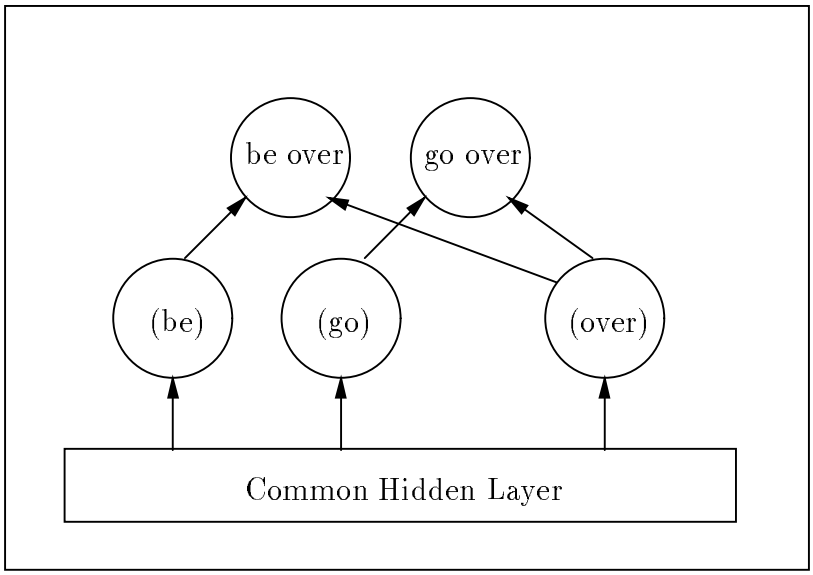

Figure 22: The **U** module of a simple SHL architecture with one common hidden layer, used for the learning of two word-pairs, here 'be over' and 'go over'.

---

as corresponding to the different "senses" of *over*, but as different *verb*-'over' pairs, and respectively constructed the training sets so that the examples in each would be naturally described, using the given pair. Thus, I had sets for 'be over', 'go over', 'climb over', 'jump over', 'fly over' and 'hover over'.

However, I still lacked some kind of architecture that, unlike the "disjoint" interpretation outlined above, would:

(a) reflect some of the commonality that the events described by a *verb*-'over' pair would have and which can be correlated with the "semantic contribution" of 'over', and

(b) begin to capture some of the linguistic *composability* of language, i.e. the ability of words to combine into larger, possibly novel, phrases (which need not imply a "compositional semantics" in the sense that the meaning of words has to be invariant with respect to context).

The structured hidden layer (SHL) architectures, described next, where meant to reflect both considerations (a) and (b).

3.3 Structured hidden layer (SHL) architectures

The basic idea behind SHL architectures is rather trivial: to "reserve" nodes (or sets of nodes) in the pre-output layer for separate "words" and then connect them to the output nodes in a way preserving linguistic structure.

Figure 22 shows the simplest such architecture: one that, for example, was used to learn

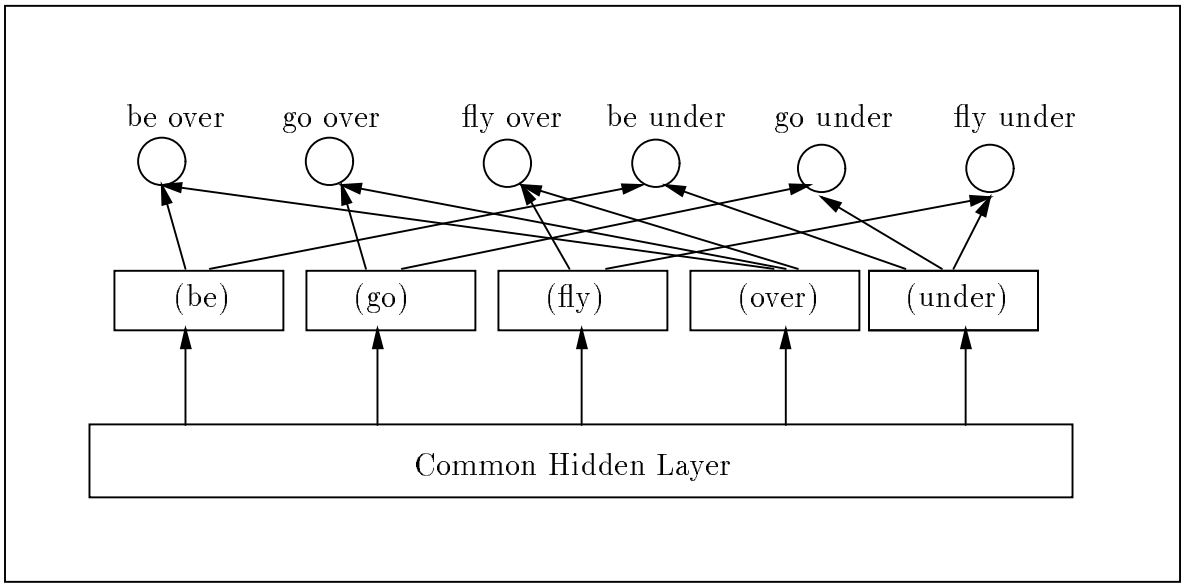

Figure 23: The **U** module of a SHL architecture for learning 3 x 2 word-pairs, e.g 'be over', 'go over', 'fly over', 'be under', 'go under' and 'fly under'.

'be over' and 'go over' examples simultaneously.

This architecture converged in less than 50 epochs, demonstrating again that the problem for the learning of polysemy in Regier's system is not the common hidden layer itself, [13] but *the single output unit.*

However, all architectures with single units for "words" could not learn more than 3 training sets simultaneously. The breakthrough came when I exchanged "word-nodes" for "word-layers", i.e. sets of nodes, that otherwise functioned like the individual nodes. Also, apart from the *over* training sets, in order to make more plausible the interpretations of particular "word-layers", I presented to a number of SHL architectures training sets for another spatial concept: *under*. These were combined with the *over* training sets in such a way that the "verb-layers" too would have more than one set of links leading to the output nodes. For example, the architecture in Figure 23, was used for attempting to learn training sets for 'be over', 'go over', 'fly over', 'be under', 'go under', and 'fly under'. Note that each output node receives input from exactly two "word-layers": one for the verb and one for the preposition.

Both with and without a common hidden layer, with single nodes in the pre-output

[13] Actually, when the structured layer was "broken up" into single units like the one above, architectures with a common hidden layer had an advantage to such that lacked one. An architecture with a four-node SHL, and 3 output units (a straightforward extension to the one in Figure 22) learned successfully training sets for 'be over' 'go over' and, 'fly over', while even the architecture like that in Figure 22, but without the common hidden layer failed to converge on a 'be over'-'go-over' training set pair.

layer, this architecture could not converge with such diverse training sets. But with e.g 5-node "word layers", the architecture performed very well, even without a common hidden layer. It did not converge during every, but rather, during roughly every second learning session. When it did learn successfully, it converged in less than 100 epochs and generalized very well, i.e. could classify — within the six categories — reliably test sets with "novel" scenes. It did over-generalize though, since there was no *mutual negative evidence* in the training sets: no positive examples of one category, e.g. 'be over', were presented as negative ones of another, e.g. 'be under'.

With *some* mutual exclusivity (two examples of another "verb-over" set for each *over* set, and two examples of another "verb-under" set for each *under* set) the learning became considerably more difficult (convergence between 100 and 250 epochs and on roughly every third try), but then, there was much less over-generalization. The 'go under' node, for example, responded positively to only "go-under" scenes, not to "be-under" and "hover-under" examples, which is demonstrated in Figure 24.

Furthermore, now that there was a special layer reserved for each of the 5 words, including 'over', each of these could be probed using the techniques for analyzing hidden layers. When, for example, the net was tested on the set of examples in Figure 25 and the 'over'-layer was analyzed, an interesting clustering of "senses" could be observed. Figure 26 shows the result from HCA, where the different kinds of examples do tend to cluster together. The are, however, not clearly separated from one another, but rather tend to form a kind of "family resemblence" structure. If we look at the activations, using the graphical display, the picture is quite different from the uniformity in Figure 9 (cf. Figure 27). It is clearly visible that there are at least two major types of "signatures" reflected in the activation patterns, with more subtle distinctions among them. Each of these signatures can be thought as a "distributed representation" of 'over' and each would combine with the signature in the respective "verb-layer" to determine how good an example of 'be over', 'fly over' and 'go over' a given scene is.

The activations of the units in the verb layers too fall into sub-groups, depending on whether the scene is characterized as a verb-'over' or a verb-'under' scene. Figure 28 shows a small test set with examples of 'go over' and 'go under' (with and without contact) and Figure 29 shows the corresponding activation vectors of the 5-node layer reserved for 'go' when tested on these examples with the architecture trained on 'over' and 'under' in combination with 'be', 'go' and 'fly'. The differences among sub-categories is quite obvious, and when probed with HCA, (cf. Figure 30) an even more interesting picture emerges. As it turns out, the network clusters the examples in 2 major types of "going": -over and -under. The "going-under" is itself divided quite clearly in 2 types, where the motion ends beneath the landmark, and where the trajector goes from side to side (the latter also subdivided according to the direction of motion). It is less clear what the 2 sub-types of "going-over" reflect, but it seems to be whether the motion is roughly parallel to the landmark, or not. Anyhow, the network has shown capabilities of not only dividing into sub-types, but also sub-dividing the sub-types themselves, i.e. of "hierarchical representations".

Regarding polysemy, this analysis suggests an interesting conclusion: that it is just as justified to speak of "different senses of *over*", e.g. a "static", an "across", a "flying" etc. as of "different senses of *going*", e.g. a "going-over", a "going-under-from-side-to-side", a

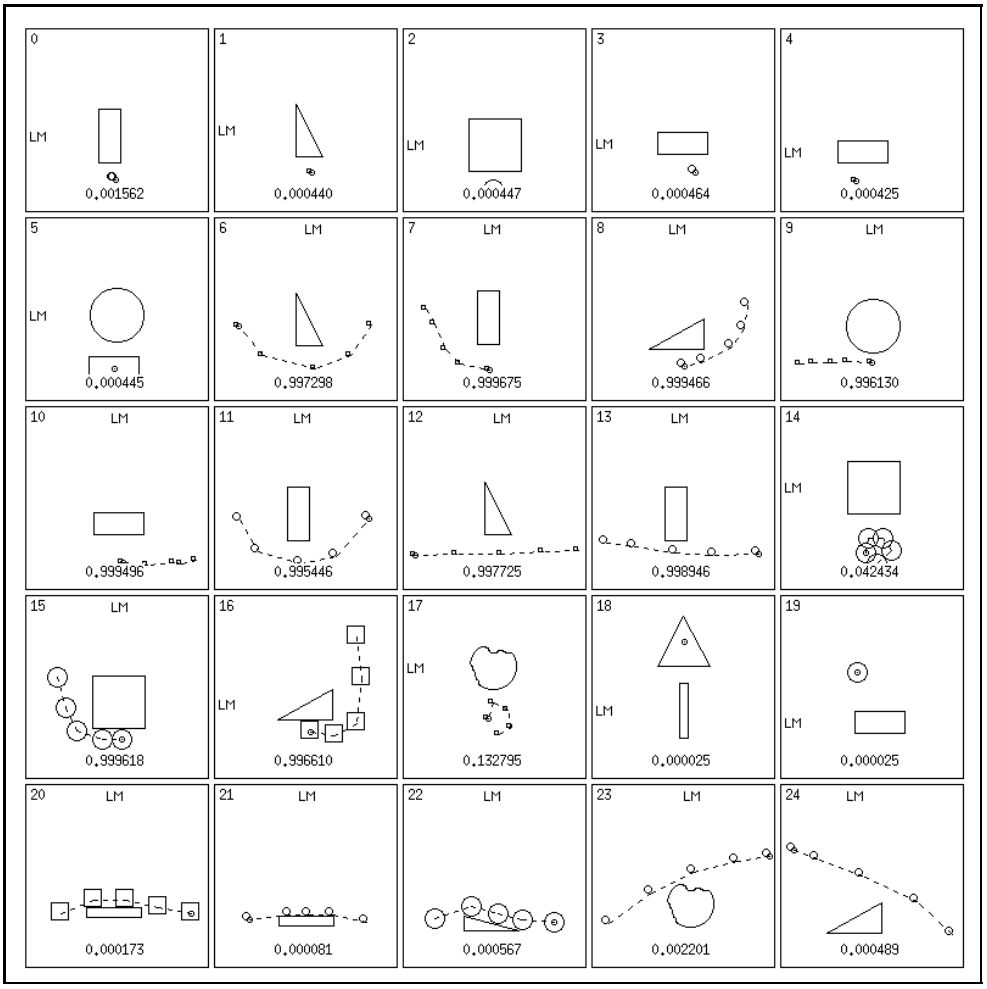

Figure 24: Probing the activation of the 'go under' output node, showing that trained with some mutual exclusive evidence, the system successfully separates "go-under" examples (6, 7, 8, 9, 10, 11, 12, and 13) from both examples of *over* and "be-under" and "hover-under".

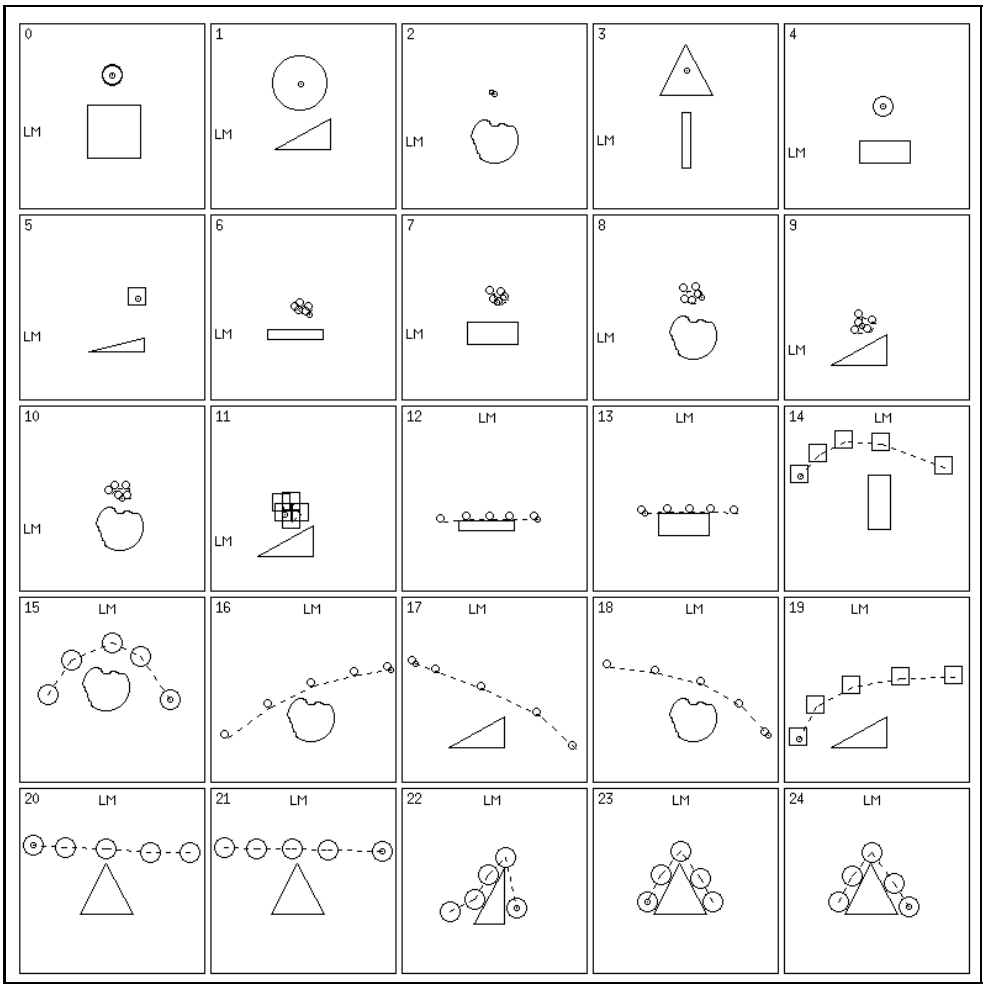

Figure 25: A test set of different examples of *over*.

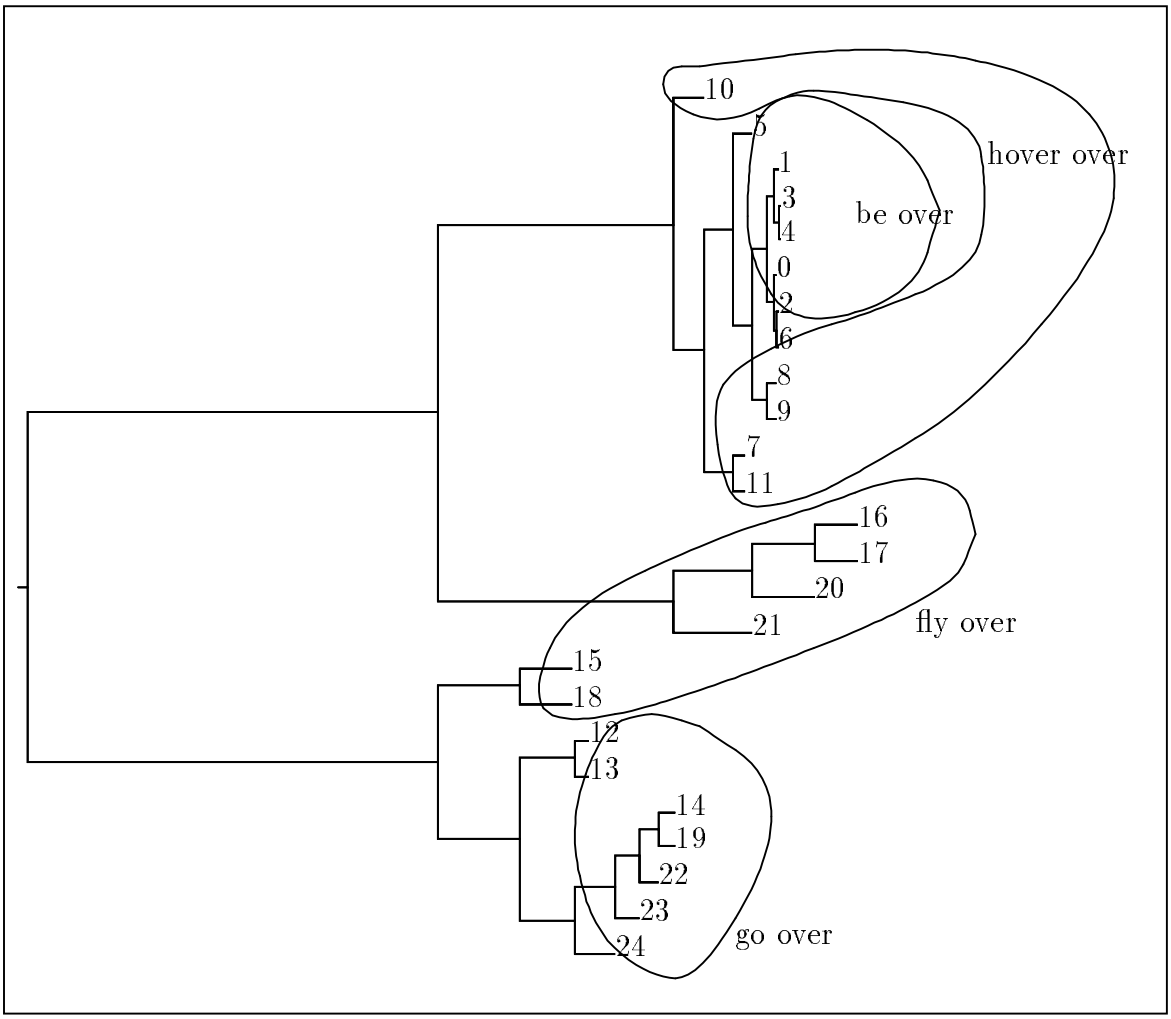

Figure 26: Graph showing HCA of the activations in the 5-node layer reserved for 'over', tested on the examples in Figure 25.

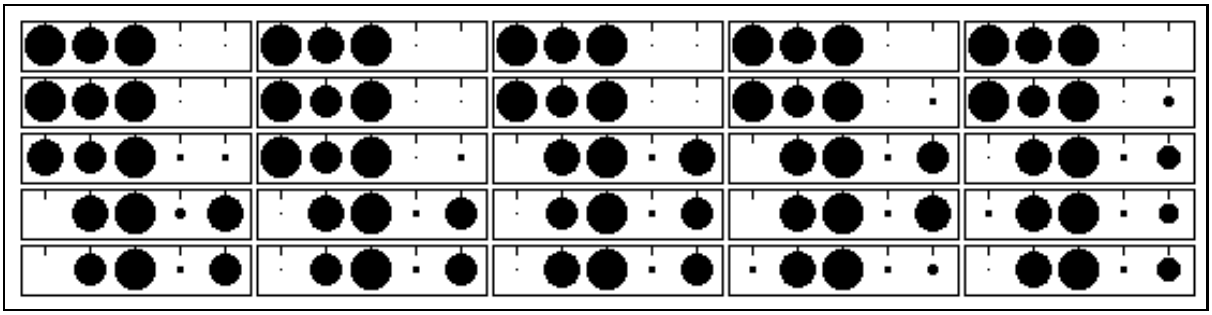

Figure 27: A display of the activations in the 5-node 'over' layer for the examples in Figure 25.

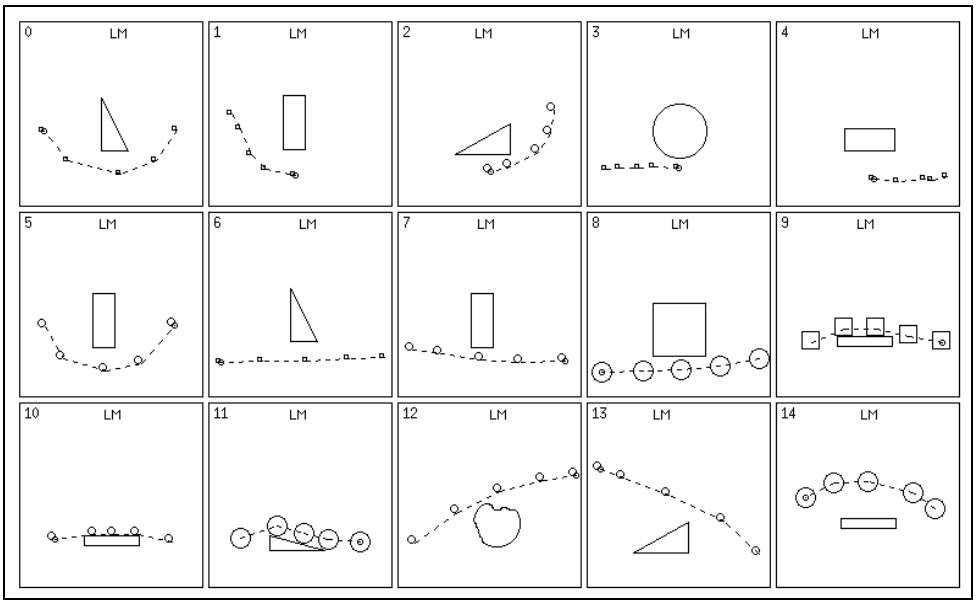

Figure 28: A test set of "go-over" and "go-under" examples.

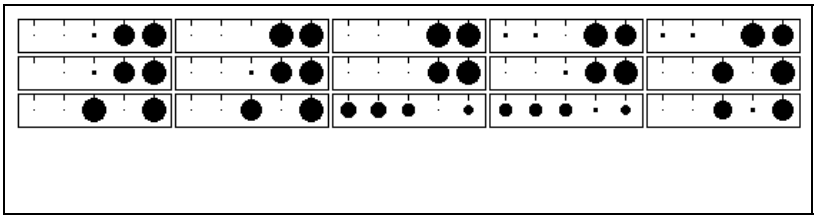

Figure 29: A display of the activation of the 5-node 'go' layer for the examples in Figure 28.

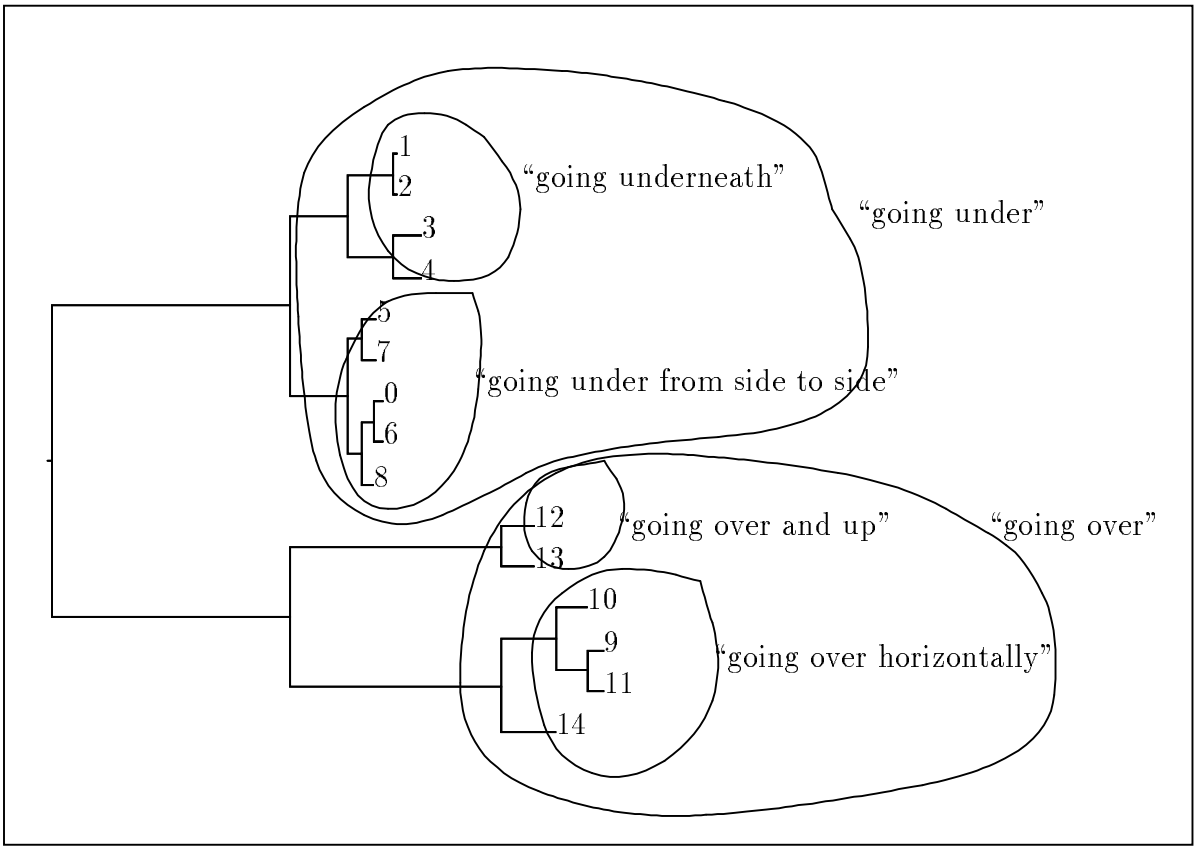

Figure 30: Graph displaying HCA of the activations in of the 5-node layer reserved for ‘go’, tested on the examples in Figure 28.

"going-underneath"... The analysis of the behavior of the system and the structures that it evolves, presented so far, does not allow us to discriminate between the case of *over* and that of *going* by saying, for example, that the first involves polysemy, while the other generality (as is usually assumed). If we decide to treat the variations in meaning reflected in the analyses of the hidden layers as instances of "polysemy", at least two conclusions are suggested, namely that (a) polysemy is even more ubiquitous than generally supposed, and (b) that it is much more of a contextual, than a purely lexical phenomenon.

We could for example call the types of *over* and *going* reflected in Figures 26 and 30, different "senses", but we must be aware that these senses emerge only as a result of simultaneous learning of 'over' and 'go' and the words they combine with. And these "other words" serve the function of, among other things, correlating the appropriate senses with the appropriate events, so that the "subscript problem" from the end of the previous section does not arise.

There are also some implications of the analysis for compositionality, which will be briefly discussed in the next section.

4 Some experiments regarding compositionality

I believe that the structured hidden layer architectures, described above, despite their simplicity come a good way towards showing how the meaning of words can vary according to context — without that implying that this variation can be random and all the negative consequences for understanding and communication that philosophers have been concerned about. There is a *systematicity* in the meaning variations, as the analyses of the hidden layers in the last section tried to demonstrate.

It is tempting to say that compositionality, as formalists have characterized it — e.g that the meaning of each word should be (roughly) the same in any phrase it participates in — is simply an unnecessary hard constraint on semantics: "meanings" can combine in a systematic way without being invariant, as the experiments with architecture in Figure 23 showed. In the end of section 3.2, I used the term *composability* to suggest this variant, context-sensitive kind of meaning-combination.

However, there still remains a problem before the requirement of compositionality can be dismissed from semantics: an alternative explanation has to be given to the cognitive capacity of dealing with *novel types*. In the given domain, that would mean that not only are "novel scenes" categorized as belonging to a certain established category, e.g. 'be over' or 'go over', but on the basis of a number of existing categories, novel ones can be "constructed". Some suggestion that this might not be impossible for a system like Regier's was provided by the HC analyses of the 'over' and 'go' layers described in the previous section. The hierarchical clustering showed a "sub-category" for 'hover over' examples, even though no such examples where presented during training.

Given the way that the SHL architectures that were most successful with polysemy were meant to associate separate layer-segments with separate words, and output nodes as word-pairs, to deal with the phenomenon of novelty of type would imply to suggest how in having learned e.g. the "meaning" of 'fly over' and 'be under', the system would be able to

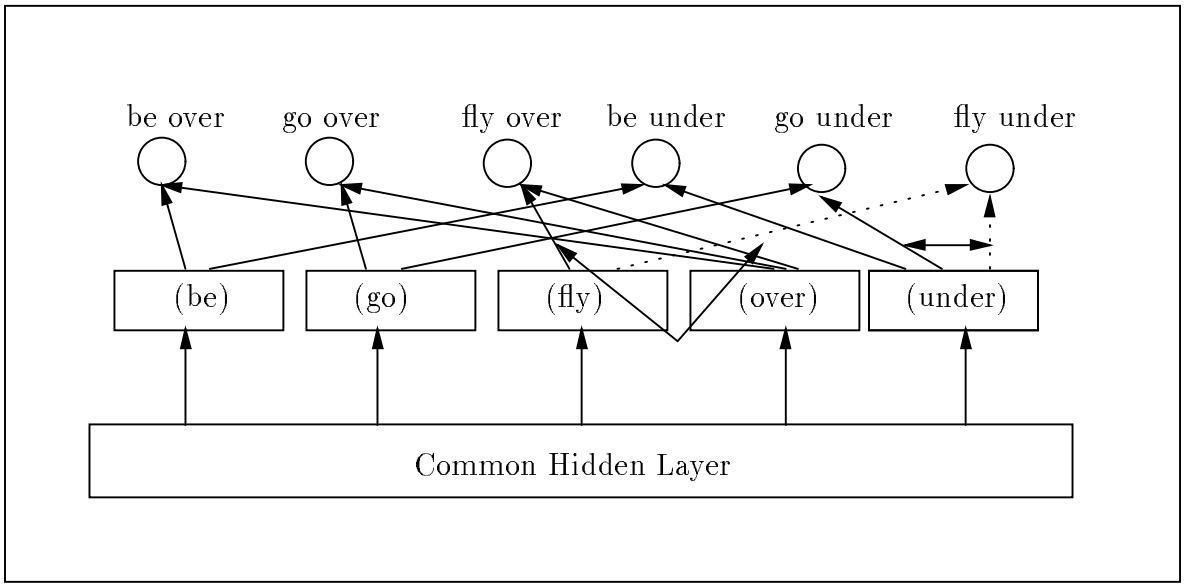

Figure 31: Sharing of the weights from the (fly) to 'fly over' and (fly) to 'fly under', and from (under) to 'go under' and (under) to 'fly under'.

---

provide a reasonably good classification of a scene-type involving a trajector moving under the landmark, one that it has *not* been trained on, and describe it as 'fly under'.

Some experiments along these lines have been performed with a version of the architecture from Figure 23. Training was performed on only five sets of examples: those described as 'be over', 'go over', 'fly over', 'be under' and 'go under'. No 'fly under' examples were presented. However the node for 'fly under' received input from the (fly) and the (under) layers in an analogous way as the 'fly over' node received from (fly) and the 'go under' node received from (under): the weights were shared (or mapped) between the relevant links. This is shown in Figure 31.

The results so far are rather encouraging: the newly established category classifies appropriately "fly-under" scenes of the type it hasn't so far encountered, (cf. Figure 32) and rejects, for example, "fly-over" scenes: (cf. Figure 33). However, when a more comprehensive training set is presented, it becomes clear that this category is general with respect to the different kinds of *under*, without any preference for the kind of motion we have described as 'fly under'. This is shown in Figure 34.

The way to improve these results is first of all to learn the sets using mutual negative evidence, which was not done in the cases described above, but ultimately the technique of weights-sharing should be substituted with some less *ad hoc* operation. One possibility is that it would simply become redundant if instead of the structuring in space that is done in SHL architectures, "structuring in time" is performed, using a recurrent connectionist

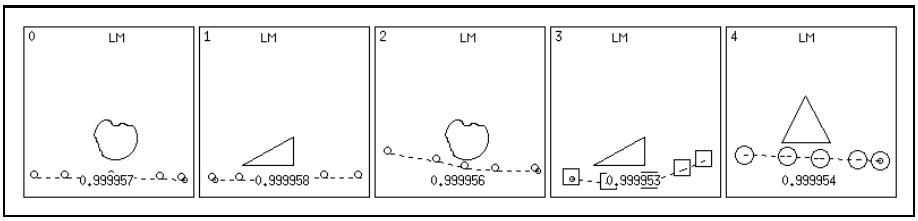

Figure 32: Appropriate classification of "fly-under" scenes, by the 'fly under' node, after the weights-sharing operation.

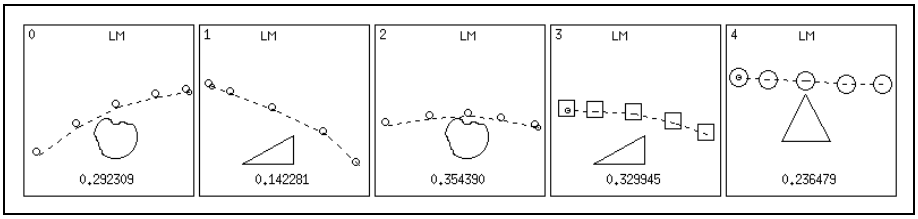

Figure 33: Appropriate classification of "fly-over" scenes, by the 'fly under' node, after the weights-sharing operation.

network such as those of [Elman, 1988], where a "context layer" holds a copy of the activation of the hidden layer from the previous step, so that it can be provided as input together with the new input at the next step. In this case, rather than a segmented hidden layer, there would be a single one, and the activations corresponding to the different words, would be allowed to evolve during different time steps. This idea remains to be tested empirically, and if successful, will be elaborated in a future paper.

5 Conclusions

The empirical studies described in this paper have attempted to shed light on the phenomenon of polysemy in a limited spatial domain, using Regier's system for perceptually grounded semantics. They have involved three different types of architectures. By judging the performance of the system, each type seems to suggest a given "treatment" of polysemy in this domain:

The uniform hidden layer (UHL) architectures could not learn more diverse training sets, and the examples of those that were learned were treated much more in the manner of generality, than polysemy as standardly described: the activations of the hidden layer tended towards uniformity, with relatively few non-distinct deviations.

The no hidden layer (NHL) architectures, on the other hand, could learn different kinds of *over* (with any consistency), only by reducing the phenomenon to homonymy, i.e. different non-related "senses" with (almost) totally disjoint sets of weights. This, like most formalist approaches to the phenomenon, leaves no clue as to how the "senses" are separated during

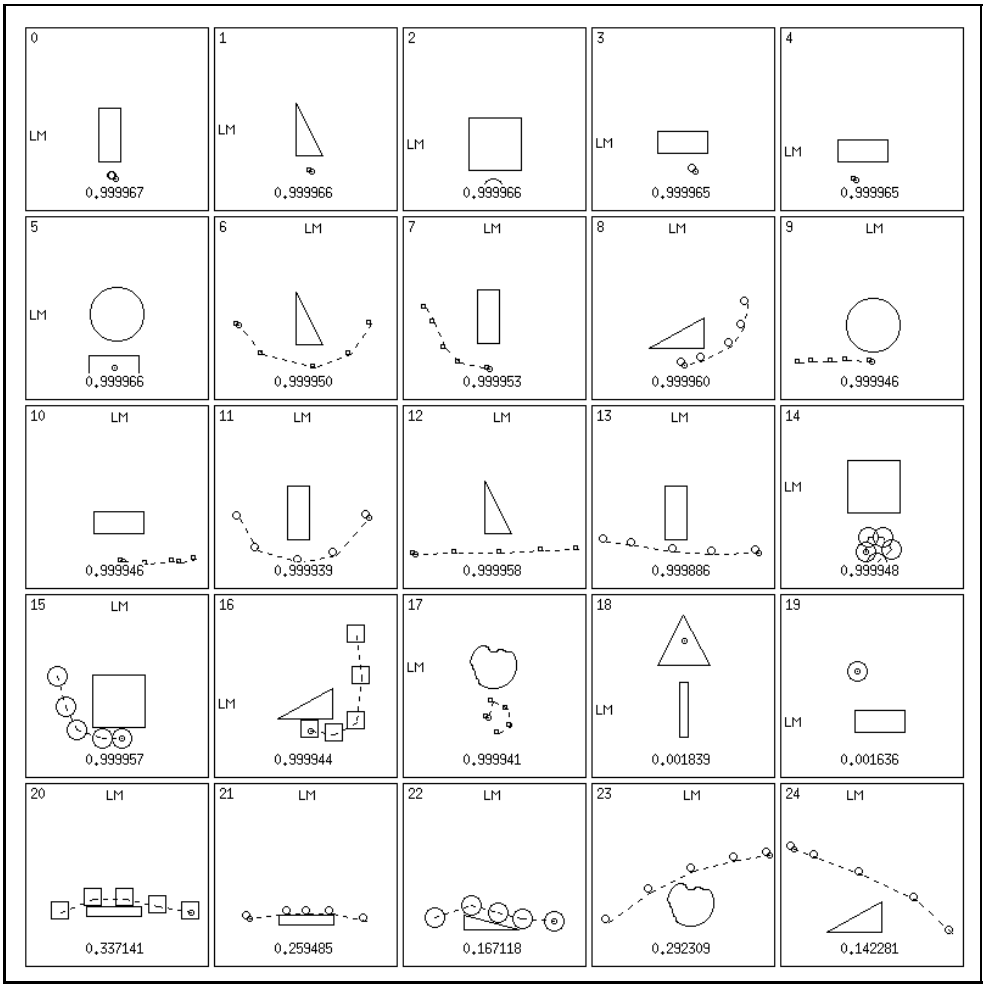

Figure 34: A test set of "over" and "under" examples, showing that the constructed category is actually general *under*.

learning and would require heavy efforts of "disambiguation" during comprehension.

In the third, and most promising kind of architecture, the structured (segmented) hidden layer (SHL) reflected part of the linguistic structure of the utterance, which allowed polysemy to be regarded not simply as a property of single lexical items, but as a *contextual phenomenon*. This means that the "multiplicity of meaning" that words in human languages display is seen as a function of the way words combine with each other and are used in actual situations. Even though in the presented study the linguistic context was limited to some rudimentary semantics of simple verbs, this provided considerable advantages in both learning and categorization of the different "senses" of words like 'over' and 'under'. It also suggested how the different senses are learned and picked out in comprehension: words are like puzzle pieces that restrict and further specify each others' meaning.

In the introduction, I mentioned a number of questions about meaning in general, that have motivated much of the research of the phenomenon of polysemy. Even though this study is performed in a very limited domain and most of the results are preliminary, I will summarize the theoretical implications of the present work by providing answers of the kind this study of perceptually grounded polysemy seems to suggest:

What is a "sense"/"meaning" of a lexical item (word)?

Words do not have a clear meaning out of context. They can be given a "denotation" only when placed within an utterance. If the activations of the hidden layers are regarded as indicative of the meaning of words — after all they perform what "intensions" are supposed to: mediate between the linguistic level and "the world" — this supports views of meaning that are *non-reificational*, i.e. see meaning as context-sensitive and changing, rather than as an object to be manipulated in thought and "put in words", as for example implied by "the conduit metaphor" (cf. [Reddy, 1979]).

In which ways can "distinct" senses be "related"?

First of all, it can not be said *a priori* what constitutes a "distinct sense" and what is an "interpretation" of a word suggested by the context. This is in line with the position argued for in [Geeraerts, 1992] that it is not clear which ambiguity is "intercategorial" and which "intracategorial", since the distinction between generality and polysemy is unstable. Since the meaning of a word like 'over', will be inherently contextual, ("the meaning of the part is a function of the meaning of the whole") by modifying this context, *more or less distinct* senses will appear.

As to relatedness, the analyzed activation patterns tended to cluster in sometimes looser "family resemblance" types of structures, other times in more "hierarchical representations". In both cases the structures were based on similarity, but due to the emergent qualities of perceptually grounding systems, this similarity is not in any simple correspondence with "objective similarities" in the environment. What will be considered *relevant* to base the categorization on will depend on factors such as the particular learning session ("history of structural coupling") and the topology of the architecture. Given a richer environment, perceptually grounding systems could learn to pick up on "features" such as *access*, or *end-pont focus*. It is thus not clear that e.g. the "image-schema transformations" claimed by

[Lakoff, 1987] to be necessary for the formation of "natural categories" are required as more than a convenient descriptive notion for certain recurrent meaning correlations.

Why should the words of human languages display polysemy at all?

One reason might be that some kind of clustering / family-resemblance / radial category structure would help in learning the meaning of individual lexical entries like 'over', and this independently of the other words in the utterance, and of context in general. This amounts to saying that the interrelated senses of a polysemous concept correspond to something like natural kinds (e.g. biologically determined "image-schemas") and if we have the right "feature detectors" in the network, it should somehow automatically "home in" on these senses; the motivation for polysemy is thus *internal*. The UHL architectures with a single output node, described in 3.1, are closest to this view. However, they did not perform very well as models of polysemy: they could not learn training sets with more than two different types of a spatial concept, without leveling the differences in the examples. If one holds to the internal motivation view of polysemy, one must thus conclude that Regier's system lacks the right structures to enable the formation of intuitive "polysemous structures".

There is, however, another view on the motivation for polysemy in human languages, which we might call *interactional*. According to it there is nothing inherently polysemous to be found in concepts like *over*, if one abstracts from the rest of the context in which the word 'over' appears. The SHL architectures were a (partial) reflection of this view. It was observed in them how elaborate polysemous structures emerge as bi-products when the activation patterns associated with the preposition combine with the activation patterns for the appropriate verb (linguistic context) in order to fit the appropriate scene (extralinguistic context). According to the interactional view, the reason for polysemy is at least not entirly internal to the system, but rather has its toots in the tendency of languages to relate to "the world" in a coherent, though contingent, manner. This "coherence" is not a simple property of words, but rather distributed throughout the entirety of the lexicon and grammar of a language. More studies in perceptually grounded semantics could perhaps explicate this view further.

As for "compositionality", the preliminary results from section 5 suggest that such a contextual, non-reified conception of meaning, as the one suggested by the answers to the last three questions need not be incompatible with the ability to interpret novel events and expressions, as is most often claimed. Much more experimentation is, however, warranted before this can be demonstrated in a convincing manner.

6 Acknowledgements

My warmest thanks to Terry Regier, without whom this work would not only have been impossible, but even inconceivable — in more than one way. If it hadn't been for the support of Jerry Feldman, it would have remained on the level of "project report". All the participants in the L_0 seminars, contributed with useful comments and ideas, but special thanks to Srini Narayan. On a broader scale, I must acknowledge the Swedish Institute, for the generosity of sending me — a Bulgarian — to Berkeley to expand my views of Language

and Meaning, as well as Peter Norvig and Bob Wilensky for hosting my stay there during the 1991-92 academic year. Plus, of course, the rest of the Berkeley Artificial Intelligence Research (BAIR) group for the patience of interacting with an often opinionated linguist and benefiting him in the process. Without the "home support" of my Swedish fairy-god mother and (future) advisor Gunnel Kaellgren, this year would have been less than pleasant and without the courses of George Lakoff, Mark Johnson, Hubert Dreyfus, Eleanor Rosch and Jerry Feldman that I was allowed to audit — less than insightful. Finally, a BIG thanks to my son Kamen, who gave me most of my intuitions about "perceptually grounded semantics" as well as the bulk of my inspiration.

References

[Bartell and Cottrell, 1991] Brian Bartell and Garrison Cottrell, "A Model of Symbol Grounding in a Temporal Environment," In *The AAAI Spring Symposium Workshop on Connectionist Natural Language Processing*, March 1991.

[Bates and Elman, 1992] Elizabeth A. Bates and Jeffrey L. Elman, "Connectionism and the Study of Change," Technical Report CRL-TR-9202, Center for Research in Language, University of California, San Diego, February 1992.

[Brugman, 1981] Claudia Brugman, "Story of *Over*," M.A. Thesis, Available from the Indiana University Linguistics Club, 1981.

[Chalmers *et al.*, 1991] David J. Chalmers, Robert M. French, and Douglas R. Hofstadter, "High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology," Technical Report CRCC-TR-49, Center for Research in Concepts and Cognition, Indiana University, March 1991.

[Chapman, 1990] David Chapman, *Vision, Instruction, and Action*, PhD thesis, Massachusetts Institute of Technology, 1990.

[Dreyfus, 1991] Hubert Dreyfus, *Being-in-the-world: A commentary on Heidegger's 'Being and Time', Division I*, MIT Press, 1991.

[Elman, 1988] J. L. Elman, "Finding Structure in Time," Technical Report 8801, Center for Research in Language, University of California, San Diego, 1988.

[Feldman *et al.*, 1990] J. Feldman, G. Lakoff, A. Stolcke, and S. Weber, "Miniature Language Acquisition: A Touchstone for Cognitive Science," Technical Report TR-90-009, International Computer Science Institute, Berkeley, CA, 1990, also in the Proceedings of the 12th Annual Conference of the Cognitive Science Society, pp. 686–693.

[Fodor and Pylyshyn, 1988] Jerry A. Fodor and Zenon W. Pylyshyn, "Connectionism and cognitive architecture: A critical analysis," In S. Pinker and J. Mehler, editors, *Connections and Symbols*. MIT Press/Bradford Books, 1988.

[French and Hofstadter, 1991] Robert M. French and Douglas R. Hofstadter, "Tabletop: An Emergent, Stochastic Model of Analogy-Making," In *Proceedings of the 13th Annual Meeting of the Cognitive Science Society*, 1991, pp. 708-713.

[Geeraerts, 1992] Dirk Geeraerts, "Vagueness's Puzzles, Polysemy's Vagaries," Presented at the 2'nd conference of the International Cognitive Linguistics Association, Santa Cruz, July 1991 (forthcoming), 1992.

[Harris, 1989] Cathy Harris, "A Connectionist Approach to the Story of 'Over'," In *Berkeley Linguistics Society*, volume 15, pages 126–138, 1989.

[Jackendoff, 1990] Ray Jackendoff, *Semantic Structures*, MIT Press, 1990.

[Kamp, 1984] Hans Kamp, "A theory of truth and semantic representation," In T. Janssen J. Groenendijk and M. Stokhof, editors, *Truth, Interpretation and Information*. Foris, Dordrecht, 1984.

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

[Lakoff and Johnson, 1980] George Lakoff and Mark Johnson, *Metaphors We Live By*, University of Chicago Press, 1980.

[Langacker, 1987] Ronald Langacker, *Foundations of Cognitive Grammar I: Theoretical Prerequisites*, Stanford University Press, 1987.

[Nenov, 1991] Valeriy Nenov, *Perceptually Grounded Language Acquisition: A Neural/Procedural Hybrid Model*, PhD thesis, University of California at Los Angeles, 1991.

[Nenov and Dyer, 1988] Valeriy Nenov and Michael Dyer, "DETE: Connectionist/Symbolic Model of Visual and Verbal Association," Technical Report UCLA-AI-88-6, University of California, Los Angeles, 1988.

[Reddy, 1979] Michael Reddy, "The Conduit Metaphor," In A. Ortony, editor, *Metaphor and Thought*. Cambridge University Press, 1979.

[Regier, 1990] Terry Regier, "Learning Spatial Terms Without Explicit Negative Evidence," Technical Report TR-90-057, International Computer Science Institute, Berkeley, California, November 1990.

[Regier, 1991a] Terry Regier, "Learning Object-Relative Spatial Concepts in the L_0 Project," In *Proceedings of the 13th Annual Meeting of the Cognitive Science Society*, 1991, pp. 191-196.

[Regier, 1991b] Terry Regier, "Learning Perceptually-Grounded Semantics in the L_0 Project," In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 138-145.

[Regier, 1991c] Terry Regier, "Learning Spatial Concepts Using a Partially-Structured Connectionist Architecture," Technical Report TR-91-050, International Computer Science Institute, Berkeley, California, October 1991.

[Regier, 1992] Terry Regier, *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*, PhD thesis, University of California at Berkeley, December 1992, (to appear).

[Rumelhart *et al.*, 1986] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning Internal Representations by Error Propagation," In James L. McClelland and David E. Rumelhart, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pages 318–362. MIT Press, 1986.

[Schank and Abelson, 1977] Roger C. Schank and R. P Abelson, *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum Associates, 1977.

[Talmy, 1987] Leonard Talmy, "Force Dynamics in Language and Cognition," Technical Report TR-49, Institute of Cognitive Studies, University of California at Berkeley, November 1987.

[Varela, 1992] Francisco J. Varela, "Whence Perceptual Meaning? A Cartography of Current Ideas," In Francisco J. Varela and Jean-Pierre Dupuy, editors, *Understanding Origins*. Kluwer Academic Publishers, 1992.

[Weber and Stolcke, 1990] Susan Hollbach Weber and Andreas Stolcke, "L_0: A Testbed for Miniature Language Acquisition," Technical Report TR-90-010, International Computer Science Institute, Berkeley, CA, 1990.

[Winograd and Flores, 1987] Terry Winograd and Fernando Flores, *Understanding Computers and Cognition: A New Foundation for Design*, Addison-Wesley, 1987.