AI REVIEW SPECIAL VOLUME ON INTEGRATION OF VISION

AND LANGUAGE

L_0—The First Five Years of an Automated Language Acquisition Project

Jerome Feldman George Lakoff David Bailey
Srini Narayanan Terry Regier Andreas Stolcke

International Computer Science Institute
1947 Center St. Suite 600
Berkeley CA 94704

Email: {jfeldman,lakoff,dbailey,snarayan,
regier,stolcke}@icsi.berkeley.edu.

Keywords: language learning, structured connectionism,
embodied cognition

Abstract

The L_0 project at ICSI and UC Berkeley attempts to combine not only vision and natural language modelling, but also learning. The original task was put forward in (Feldman *et al.* 1990a) as a touchstone task for AI and cognitive science. The task is to build a system that can learn the appropriate fragment of any natural language from sentence-picture pairs. We have not succeeded in building such a system, but we have made considerable progress on component subtasks and this has led in a number of productive and surprising directions.

1 Introduction

The L_0 project at ICSI and UC Berkeley attempts to combine not only vision and natural language modelling, but also learning. The original task was put forward in (Feldman *et al.* 1990a) as a touchstone task for AI and cognitive science in a very simple form:

> The system is given examples of pictures paired with true statements about those pictures in an arbitrary natural language. (See Figure 1.)
>
> The system is to learn the relevant portion of the language well enough so that given a novel sentence of that language, it can determine whether or not the sentence is true of the accompanying picture.

There are a number of attractive features of this general task. It is strictly behavioral and theory-free: nothing has been said about the theory or methodology that should be employed in the task. The problem is closed in the sense that one cannot appeal to some forthcoming result in a related domain that will complete the story—the system has to do the whole job. There is no stipulation of how much should be built into the system and how much learned. And the requirement that the same system should work for equivalent fragments of any natural language rules out *ad hoc* solutions.

As it turns out, a number of other groups were looking at similar tasks and there is now a modest body of work (Suppes *et al.* 1991; Nenov & Dyer 1993; Nenov & Dyer 1994; Siskind 1992) on the Miniature Language Acquisition task. Our own research has (as we expected) developed in ways that we didn't expect. This paper presents a brief reprise of the effort and a description of its current state in late 1994.

From the outset it was clear that we would not be able to attack the fully integrated L_0 task at once. The first round of effort broke it down into three projects, each of which begged a major question. The first project ignored the learning issue and concentrated on building a system that could be reprogrammed readily to solve the L_0 problem for an arbitrary natural language. There were a number of important lessons learned, primarily in the subtle semantics of spatial language, but it took well under a year to build up a hand-adaptable system (Feldman *et al.* 1990b; Weber & Stolcke 1990). Our wizards, Andreas Stolcke and Susan Weber, got to the point where it took only an hour or so to realize L_0 for a new language, given a sophisticated informant. This test-bed project was abandoned, but aspects of it are still appearing in current work.

The other two pieces of the original partition have taken a bit longer. These divided the learning aspects of L_0 into a single-word concept learning task (suppressing syntax) and a syntax learning task that assumes or ignores semantics. These turned into the doctoral theses of Terry Regier and Andreas Stolcke, respectively. We believe that the two efforts could be combined to solve the original L_0 task and we might try this as a student project, but it now seems more important to extend the task and solutions. This is best understood in the context of what has been done already. Although the concept learning system was completed earlier it is convenient to first describe the state of our progress on grammar learning.

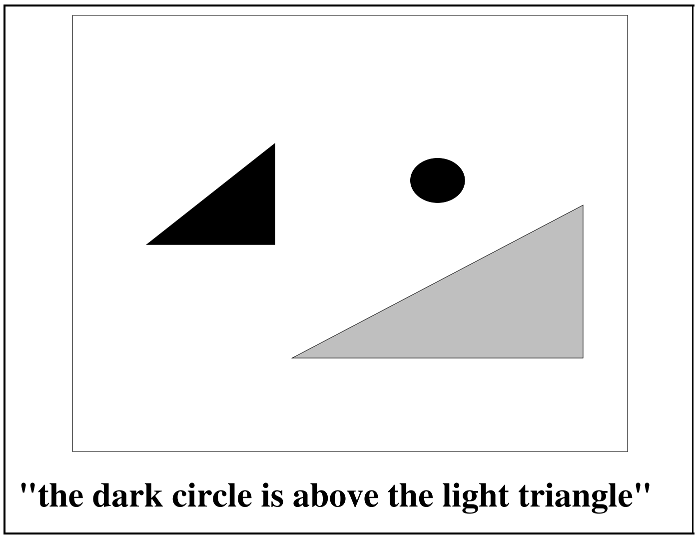

Figure 1: A typical picture–statement pairing.

2 Grammar Learning in L_0

Grammar induction in the context of L_0 has a two-fold function. On one hand it induces a probabilistic, syntactic model of the language, using only the provided positive samples as input. Secondly, it augments the induced syntactic model with rules for semantic features, including those representing the provided 'meaning' of the sample sentences. The resulting combined model is a probabilistic mapping that assigns both syntactic structure and meaning (within the limits of L_0 semantics) to a sentence.

Note that the way the L_0 task is posed, the syntactic aspect of learning is entirely ancillary: in principle the learning agent could construct and represent the correspondence between surface forms and semantics without any notion of syntax. However, it is generally accepted that syntactic structures have an independent motivation and are useful in both describing and processing language, although linguistic theories disagree as to the degree to which syntax should be considered 'autonomous.'

2.1 Induction of context-free grammars

Our approach to syntactic learning can be broadly described as learning of *probabilistic context-free grammars* using a *Bayesian inference* framework and *model merging* as the principal induction operator. We will explain this approach by elaborating each of these terms in turn.

Probabilistic grammars. In this framework, a grammar G is a probability distribution over all possible sequences of words, i.e., it assigns a number $P(x|G)$ between 0 and 1 to a string x, such that the probabilities of all strings sum to 1. Such an approach has several advantages over a simple all-or-nothing concept of grammar. For example, it allows objective comparisons of candidate grammars vis-a-vis a sample corpus or a given target grammar (by evaluating the sample probabilities or the relative entropies, respectively). A particularly appealing aspect of the probabilistic approach is that is provides an intuitive and formally simple solution for an old dilemma: how to learn a grammar (or more generally, any binary classification) in the absence of negative evidence, i.e., sentences labeled as non-grammatical. Under the assumption that the 'true grammar' is also probabilistic and generates the sample data according to the distribution it defines, it becomes possible in principle to identify the correct grammar (or an equivalent one) given sufficiently many data, by choosing the one that maximizes the probability over the data.

Context-free grammars. The particular variant of syntactic probabilistic grammars used in L_0 is the *stochastic context-free grammar* (SCFG). This is an extension of standard context-free grammars (CFGs) in which each production $X \to \alpha$ has an associated probability $P(X \to \alpha)$, the conditional probability of choosing the right-hand side α when nonterminal X is to be expanded. The context-free nature of non-probabilistic CFGs is extended by the stipulation that the expansion of a nonterminal X is conditionally independent of the surrounding context, given X itself. This assumption is clearly violated in natural language, but seems adequate for L_0, and keeps the formalism computationally tractable.

Bayesian inference. Belief about the plausibility of a grammar in the light of data is formalized as a probability distribution over all possible grammars. Thus the *prior distribution* $P(G)$ embodies belief in the validity of grammar G before having observed any samples. The *posterior distribution* $P(G|X)$, on the other hand, gives the revised belief in G after having observed the data X. Both are related by a factor that expresses how well the grammar G explains the data X, which is known as the *likelihood* $P(X|G)$. Bayes' Law relates the three quantities:

$$P(G|X) = \frac{P(G)P(X|G)}{P(X)}$$

Bayesian inference implies finding a grammar with maximal posterior probability given the observed data. The Bayesian framework thus provides a normative approach to the inference of grammars from finite data. (In the limit of infinite data the Bayesian approach degenerates to the simple comparison of likelihoods $P(X|G)$, since that term will come to dominate the prior $P(G)$ in the above equation.)

Bayesian inference also incorporates (a version of) Occam's Razor, or a default preference for simpler over more complex models. The Bayesian version of Occam's Razor comes in two variants: for the discrete parts of grammar (i.e., its rule structure), the total prior probability mass has to be allocated over all possible grammar structures. Since the number of structural choices grows with the grammar size (measured, e.g., by the number of nonterminals, productions, etc.) each of these larger grammars will receive less prior probability than a smaller one. This is just a probabilistic reformulation of our intuition that succinct grammars are to be preferred over profuse ones, other things being equal. The second variant of Occam's Razor, know in the literature as 'Occam factors' (Gull 1988), concerns the comparison of models with continuous parameters (e.g., rule probabilities). Here, models with fewer parameters are inherently preferred by the Bayesian approach since the prior probability over the continuous parameter space has to be spread 'thinner' if that space has more dimensions. Again, this is just a mathematical reflection of our intuitive preference for models that leave less free open choices, provided they explain the data equally well.

Within this broad characterization it is now possible in principle to construct prior distributions for grammars that reflect very specific theoretical biases. For example, we could construct a prior that assigns zero or very low probability to all context-free grammars that do not conform to the stipulations of X-bar syntax. This is known among Bayesians as an *informative prior*. On the other hand we have *uninformative priors* that simply spread the prior probability evenly in a more or less neutral fashion (clearly prior informativeness is a matter of degree).

We have generally taken the latter approach, since one of the interesting theoretical questions at this point is to what extent linguistic knowledge can be inferred with minimal built-in bias. Since the L_0 task is clearly a rather simple linguistic domain we would expect that a 'knowledge-free' approach be successful here if not elsewhere.

One convenient way to construct prior distributions on grammars is to devise ways to encode them. For example, we could encode each nonterminal in an SCFG using a certain number of bits (dependent on the total number of nonterminals), and use those codes to compile a list of all rules. The length of the entire encoding is now a precise version of what we can mean by the *size* $\ell(G)$ of a grammar G. Each coding scheme now defines a prior distribution on grammars by virtue of a standard dualism between encodings and probabilities: a grammar receives a probability proportional to $e^{-\ell(G)}$. A grammar become less probable by a constant factor > 1 for each bit more that it takes to encode it. Conversely, by a theorem of information theory, each prior $P(G)$ corresponds to an encoding in which G is conveyed using $-\log P(G)$ bits. The Bayesian approach thus becomes equivalent to the principle of induction by *minimum description length* (MDL) (Rissanen 1983; Wallace & Freeman 1987).[1]

Model merging. Given the Bayesian framework, it is in principle possible to find the grammar G that maximizes the posterior probability $P(G|X)$ by examining all possible grammars. Horning (1969) has shown that there is an effective procedure for doing this provided that we can enumerate all possible grammars in order of decreasing prior probability. However, such an enumeration would not be practical even for grammatical domains of modest size.

An alternative is to use clever heuristics that guide the search through the grammatical model space in an efficient (but not necessarily optimal) fashion. Intuitively, we would like this search to be *data-driven*, i.e., informed by the observed data, and not just blindly generating and testing hypotheses. Omohundro (1992) has recently made an argument for *model merging* as an efficient solution to this problem, which can be applied in a variety of domains and which has some degree of cognitive plausibility. The model merging approach has since been applied successfully to Bayesian inference in the domain of finite-state probabilistic grammars (Stolcke & Omohundro 1993; Wooters & Stolcke 1994), and the current approach for L_0 is a direct extension of that work.

[1] In this framework, informative priors, or linguistic theories, can be seen as very specialized codes that provide short descriptions for all theory-conforming grammars, leaving the long descriptions for the non-conforming ones.

The two basic ideas in model merging are:

Data incorporation Initially, with little or sparse data, build a model that directly mirrors the data. This corresponds to a 'case-based' approach, were the best one can do is record past experiences hoping that they will be repeated in similar or identical form in the future.

Applied to the domain of language, this means that in the absence of an applicable model (or if the grammar doesn't match the new data), the learner simply adds new words, sentences, etc. to a list of exceptions yet to be explained.

Merging Whenever possible and warranted by a 'goodness' criterion, *merge*, or combine, models or submodels. The merging operation, which is dependent on the domain, usually represents a *generalization* step.

This step will be illustrated below using the L_0 domain as an example.

The incorporation step in the case of grammars is usually straightforward and involves the creation of new, specialized rules to account for the new data. For example, if our new sample is[2]

```
a circle is below a triangle
```

and not yet accounted for by the grammar, we add a top-level production for the sentence, as well as one lexical rule for each word:

```
S --> A1 A2 A3 A4 A1 A5
A1 --> a
A2 --> circle
A3 --> is
A4 --> below
A5 --> triangle
```

The merging operation for context-free grammars comes in two variants. A *paradigmatic merging* operation involves coalescing two nonterminals. Suppose we add to the above rules one for the sentence `a circle is below a circle`:

```
S --> A1 A2 A3 A4 A1 A2
```

A (paradigmatic) merging step on the nonterminals `A2` and `A5` would involve replacing all occurrence of these two nonterminals with a new symbol, say, `A25`.

```
S --> A1 A25 A3 A4 A1 A25
A25 --> circle
    --> triangle
```

[2]The convention used throughout this section is that `typewriter` font denotes concrete example, whereas *italics* are used for generic names. Nonterminals have uppercase names, while lowercase is reserved for terminals (words).

As a result of merging, entire productions may be coalesced as they have become identical.

What causes 'good' merges (as the one above) to take preference over 'bad' merges (such as `A1` with `A2`)? This is where the probabilistic approach becomes crucial. All productions have probabilities reflecting the relative frequencies with which each was observed in (i.e., accounted for) the data. A 'bad' merge is generally one that causes the probabilities to change in such a way as to lower the probability of the data $P(X|G)$ more than other, 'good' merges. The merge in the example above is good because it allows the two `S` productions to be coalesced, thereby lending high probability to the merged rule (and hence to the samples using that rule).

Alas, it is easily seen that the likelihood $P(X|G)$ can never *increase* as a result of merging; the best that can happen is that is stays unchanged. However, the Bayesian posterior probability of the grammar $P(G|X)$ does increase as a result of merging if the likelihood remains the same, since the prior will prefer a grammar with fewer nonterminals (and possibly fewer productions as well). More generally, the prior can compensate for (small) drops in the likelihood. Intuitively, this means that the data justifies a generalization beyond the observed data, due to the simplification this would bring to the grammar.

To infer general context-free grammars a second operation, *syntagmatic merging* or *chunking* is required. This involves replacing nonterminals that occur adjacently in the right-hand sides of productions by a single, new nonterminal. For example, starting with the rule

```
S --> A1 A25 A3 A4 A1 A25
```

we are tempted to chunk the sequence `A1 A25` to derive a new grammar containing

```
S --> A6 A3 A4 A6
A6 --> A1 A25
```

The new nonterminal `A6` can now enter into more merging and chunking.

The combination of description length prior and likelihood as our inference criterion will again distinguish 'good' from 'bad' chunks. For example, the above choice is good mainly because it causes productions to be shortened. Also, chunks can be found beneficial simply because they create nonterminals that feed into merging such that new generalizations can be expressed.

Summary and results The overall picture can now be described as follows. After formulating a prior for context-free grammars, our learner starts incorporating samples into a working grammar. It constantly examines possible merging and chunking operations, evaluating them according to the posterior grammar probability they produce. Operations that give improvements are applied in a best-first manner, until no more changes are warranted (and until new data arrives).

This simple description glosses over a number of non-trivial technical problems, mostly having to do with efficient search, locality of the evaluation metric etc., but we believe to have found adequate solutions to most of these. (For example, when evaluating the goodness of a chunking operation it is necessary to look ahead to the merging operations it enables to assess its true merit.)

Another interesting problem is how the learning process is affected by noise. Due to both its case-based and its probabilistic character the merging methodology handles 'outliers' quite well. They are simply added to the grammar as exceptions, and will eventually receive very low probability, unless supported by similar samples. Such low probability elements of the grammar can then be pruned, either on-line or in a post-processing step.

Using simple non-informative description length priors we have been successful in finding very reasonable syntactic description for L_0 fragments in a number of languages. For example, in the case of English, our procedure groups words into lexical categories for nouns, pronouns, transitive/intransitive verbs, adjectives, as well as the standard NP and PP phrase structures. For languages such as German the merging process correctly identifies lexical subcategories for the purpose of agreement and case assignment, and builds phrase structure rules reflecting these constraints (e.g., gender agreement between determiners and nouns). The Bayesian grammar learner can likewise induce grammars containing recursive productions, such as required to account for embedded relative clauses or chains of adjectives in noun phrases. A more detailed study of grammatical model merging for SCFGs and other probabilistic models can be found in (Stolcke & Omohundro 1994).

We have no illusions that our method can, by itself, derive arbitrarily sophisticated grammars for the more complex kinds of linguistic phenomena. Such a task would require considerable work both on the representations involved and the necessary learning algorithms. But we believe that the Bayesian model merging approach as shown for the L_0 task provides a rather general framework from with fundamental ideas can be applied to more demanding domains. Moreover, in many applications it is acceptable to use probabilistic context-free grammars as mere *approximations* to the 'true' underlying syntactic distributions. For such uses we believe to have found a very useful tool.

2.2 Learning simple feature-based semantics

As in the case of syntax, learning of semantic interpretation rules for L_0 is based on a probabilistic model. Even more than before, we will be interested in the discretized version of this model, but the probabilistic formulation enables us to apply the principles of probabilistic induction laid out earlier.

Probabilistic feature grammars The sentence semantics in L_0 can be adequately represented as a collection of features. A sentence like

```
a circle touches a square
```

has a meaning represented as the following set of features names and values:

```
rel: TOUCH, tr: CIRCLE, lm: SQUARE
```

where `rel` (relation), `tr` (trajector) and `lm` (landmark), as well as `TOUCH`, `CIRCLE`, `SQUARE` are supposedly language-independent constants of the conceptual language interfacing the language learner with the rest of L_0.

The fundamental, simplifying assumption in our approach is that these feature values originate from syntactic or lexical rules, and are combined according to rules of compositional, propositional semantics. For example, the meaning components `TOUCH`, `CIRCLE` and `SQUARE` are thought to be attached to the lexical rules generating the words `touch`, `circle`, `square`.[3] The syntactic rules that give rise to the complete sentence contain similar specifications that cause these values to eventually appear on the observed top-level features.

The rules that propagate the feature values through the syntactic structure are extensions of context-free rules, and similar to those found in many feature-based grammar theories (Shieber 1986). Consider a simple context-free rule involving some nonterminals X, Y and Z:

$$\begin{aligned} X \;\rightarrow\; & Y Z \\ & X.f = Y.g \quad [p_1] \\ & X.f = Z.h \quad [p_2] \end{aligned}$$

This one determines that in a syntactic environment where nonterminal X expands to Y and Z, the feature f on X (briefly, $X.f$) takes on the value of $Y.g$ with probability p_1, or the value of $Z.h$ with probability p_2 (any number of such probabilistic assignments could be specified). In our current formulation, the dependence among features is a directed one: those belonging to the left-hand side nonterminal get to chose probabilistically among the values on the right-hand side, unlike non-probabilistic feature-based formalisms where the relation is a symmetrical constraint.

Feature values may also be specified directly, as is typically done at the level of lexical rules:

$$\begin{aligned} X \;\rightarrow\; & \texttt{circle} \\ & X.f = \texttt{CIRCLE} \quad [p] \end{aligned}$$

Thus, the collection of probabilistic feature rules determines, for a given syntactic structure, the distribution of feature values appearing at the S node of the sentence. At the S node we have certain feature names determined by convention, but in the remaining parts of the structure features name are entirely arbitrary; only their relation to each other is important for determining the sentence semantics.

Feature grammar merging Learning of feature assignment rules follows the model merging framework previously shown for syntactic rules alone. As before, the *data incorporation* step creates rules that are designed to specifically generate the samples observed. Samples now consist of paired sentences and top-level feature bundles. Hence the sample sentence `a circle touches a square` (with the semantic features given earlier) is incorporated by creating productions

```
S --> A CIRCLE TOUCHES A SQUARE
    S.tr = ‘circle’
    S.lm = ‘square’
    S.rel = ‘touch’
A --> a
```

[3] The coinciding spellings are an artifact or our using English as a metalanguage.

```
CIRCLE --> circle
SQUARE --> square
TOUCHES --> touches
```

The *merging operators* for probabilistic feature grammars fall into two groups. The syntactic merging operators (nonterminal merging and chunking) can still be used, and are simply adapted to preserve the feature specifications in the semantic part of the rules involved.. In addition, we require new operators that specifically generalize feature descriptions. The first such operation, *feature attribution*, takes a feature value assignment and replaces it with a feature equality involving a newly created feature. Consider the two productions

```
S --> A CIRCLE IS BELOW A SQUARE
    S.tr = ‘circle’
    S.lm = ‘square’
    S.rel = ‘below’
S --> A SQUARE TOUCHES A TRIANGLE
    S.tr = ‘square’
    S.lm = ‘triangle’
    S.rel = ‘touch’
```

along with the obvious lexical productions expanding the preterminals `A`, `CIRCLE`, `IS`, etc., which are as yet without semantics, e.g.,

```
SQUARE --> square
```

A feature attribution operation can now associate the nonterminal `SQUARE` with a newly created *ad hoc* feature `f1` and the value `‘square’`:

```
S --> A CIRCLE IS BELOW A SQUARE
    S.tr = ‘circle’
    S.lm = SQUARE.f1
    S.rel = ‘below’
S --> A SQUARE TOUCHES A TRIANGLE
    S.tr = SQUARE.f1
    S.lm = ‘triangle’
    S.rel = ‘touch’
SQUARE --> square
    SQUARE.f1 = ‘square’
```

However, this is just one of the possibilities, and we rely on the probability maximization strategy, as well as the interplay with other operators to prevent inappropriate feature attribution (such as `SQUARE` with `‘circle’`).

The second new feature operator is *feature merging* itself, meaning the coalescing of two feature names and the concomitant rewriting and merging of feature assignment rules. For example, after another feature attribution, the first of the `S` productions above gives rise to

```
S --> A CIRCLE IS BELOW A SQUARE
```

```
    S.tr = CIRCLE.f1
    S.lm = SQUARE.f2
    S.rel = 'below'
CIRCLE --> circle
    CIRCLE.f1 = 'circle'
SQUARE --> square
    SQUARE.f2 = 'square'
```

One syntactic merging step (`CIRCLE` and `SQUARE` into `N`), immediately followed by feature merging of `f1` and `f2` into `f`, finally produces

```
S --> A N IS BELOW A N
    S.tr = N.f
    S.lm = N.f
    S.rel = 'below'
N --> circle
    N.f = 'circle'
  --> square
    N.f = 'square'
```

This concludes the overview of the syntactic and semantic merging operators. The operators and their combinations are evaluated in a beam search procedure that tries to find a grammar with high posterior probability. As in syntax-only learning, grammars are evaluated according to a Bayesian criterion that combines data fit (likelihood) and conciseness of the grammar (expressed as a prior distribution derived from the grammar description length).

Results and outlook Using these techniques we were successful in inducing grammars for L_0 fragments of English and German, using only the syntactico-semantic distributional evidence contained in randomly generated samples. These fragments encompass intransitive, transitive, and prepositional object sentences, using simple noun phrases (possibly containing adjectival modifiers). Note that the semantic representation so far allows no recursion, and due to that the syntactic range of these fragments is more limited than in the syntax-only learning scenario. On the other hand, we were able to construct training corpora which demonstrate that the semantic aspect of learning can disambiguate syntactic structure. For example, it is possible to have sample sets of transitive sentences for which a "verb phrase" grouping subjects with verbs has equal posterior probability as the traditional verb-object grouping. However, such devious structures become less favored once features are added to the training corpus. Limited space prevents us from working through a detailed example at this point; the reader is referred to (Stolcke 1994).

Future work on the learning of probabilistic feature grammars aims at relaxing some of the purposely simplified aspects of the current model, with the goal of better approximating the use of feature structures in standard linguistic theories. For example, feature values should be allowed to propagate probabilistically in any direction, as opposed to only bottom-up as in the current formulation. Another important extension would include hierarchical feature structures.

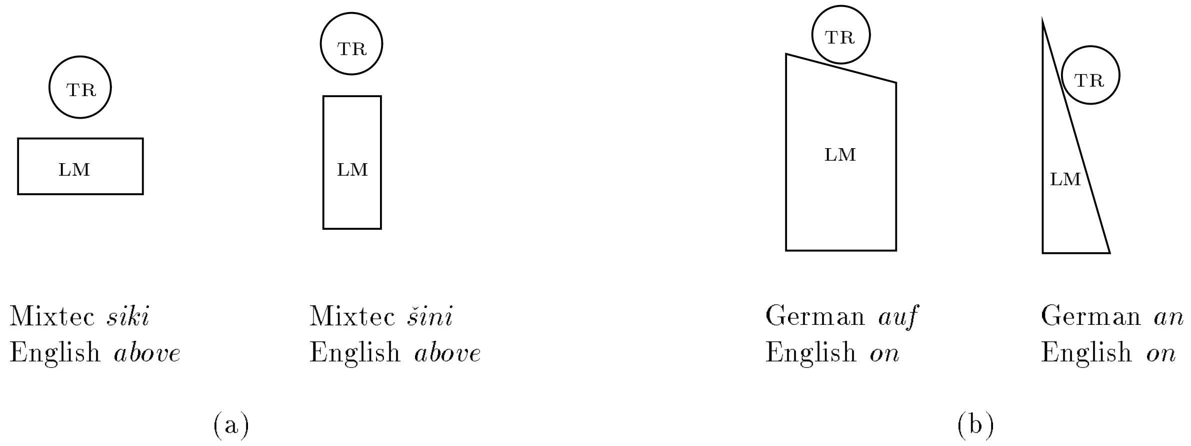

Figure 2: Examples from Mixtec and German

3 Lexical Spatial Semantics

3.1 The Linguistic Categorization of Space

The linguistic categorization of space is a topic that has captured the attention of linguists and other cognitive scientists for a number of years. There is good reason for this: space is an attractive semantic domain in that it is concrete and tangible, and yet still affords a good deal of subtlety in its semantic structure. In this respect it resembles the domain of color, another objectively measurable domain which has attracted work in semantics (Berlin & Kay 1969).

In addition, space (along with time) has a privileged position as a fundamental conceptual structuring device in language, a position which most other domains, color included, do not share. Lakoff (1987) points out that such fundamental concepts as those relating to time and space tend to be used in many other concepts throughout a given linguistic system, often through metaphor, and are not localized to isolated domains of experience, as most other concepts (e.g. *socialism*) are.

Space is also attractive as a domain of semantic inquiry because there is considerable cross-linguistic variation in spatial systems. The notion of cross-linguistic variation in conceptual systems and modes of thought has long been a subject of fascination for linguists (see Whorf (1956) for an early and extremely influential example). Such differences in spatial systems are sometimes quite dramatic, but more often than not they are rather subtle, particularly when one compares closely related languages. provides examples of non-English spatial structuring, from Mixtec (Brugman 1983), a Mexican Indian language, and from German. In Figure 2(a) we see two spatial configurations which would both be categorized as *above* in English, but which receive distinct categorizations in Mixtec, which is sensitive to the major axis orientation of objects. Similarly, in (b) we see two configurations both of

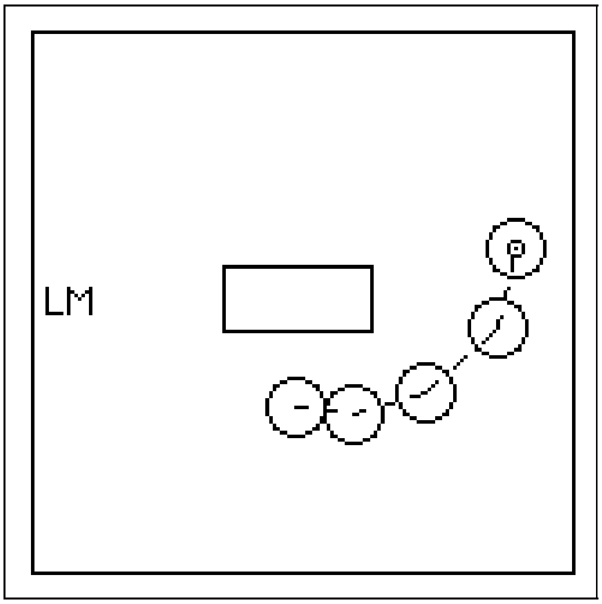

Figure 3: A movie: Russian *iz-pod*

which would be categorized as *on* in English, but which are treated distinctly in German, in which the orientation of supporting surfaces is semantically significant. An exhaustive cataloging of cross-linguistic variation in spatial systems is of course well beyond the scope of this article.

Despite all this variation, there must also exist commonalities across languages by virtue of the fact that all linguistic spatial systems are based on human experience of space, which is in turn constrained by the nature of the human perceptual system and the nature of the world around us. The point of interest here is that the varying spatial systems all derive from the same neural mechanisms and the same experiences with objects, gravity, and the like.

This leads us to ask the following question:

> *What sort of system could adapt itself to so many different structurings of space, as the human capacity for categorizing space can?*

We present a connectionist model which learns the perceptually grounded semantics of lexemes for spatial events and relations from a range of languages. Thus, it provides at least a preliminary model of the human capacity for lexicalizing spatial events and relations, and a preliminary answer to the above question.

3.2 The Model

Imagine a set of movies of simple 2-dimensional objects moving relative to one another, such that each movie has been correctly labeled as a positive instance of some natural language spatial term. The movies may be of arbitrary length. For example, presents such a movie, a positive example of the Russian preposition *iz-pod*, which has no single-word English counterpart, but translates to "out from underneath". The connectionist model presented here takes a set of such movies, each labeled as a positive example of some spatial term from some language, and learns the association between words and the events or relations which they describe. Once such a system has successfully accomplished this task, it should be able to determine which of the spatial terms it has learned would be appropriate for describing a movie not previously seen. The system's task as a whole can be viewed as:

> *Learning how to perceive simple spatial relations, both static and dynamic, so as to name them as a speaker of a particular language would.*

Clearly, if this system is to be taken as a model of the human capacity for lexicalizing spatial relations and events, it must be able to perform this learning task for spatial terms from any language. Furthermore, it should be able to learn without the benefit of explicit negative instances, as children appear to acquire language under those conditions. While it is unlikely that the model as currently constructed can learn an arbitrary human linguistic spatial system, the model does learn systems of spatial terms from a range of languages with disparate categorizations of space. And it learns from positive evidence only. Thus it begins to answer the question posed above: it provides an indication of what sort of system could adapt itself to the various structurings of space found in the world's languages.

The key to the model's design is its incorporation of a number of structural devices which are motivated by neurobiological and psychophysical evidence concerning the human visual system. The model contains motivated structures of three sorts: (a) orientation-sensitive cells, (b) cells with circularly symmetric center-surround receptive fields, and (c) a tripartite representation of motion trajectories which consists of separate subrepresentations for the starting point, ending point, and path taken.

There is evidence for each of these structures in the human visual system. The critical point is that the incorporation of these structures into the model is in keeping with the spirit of inquiry into issues of semantic universality and relativity that motivates the work as a whole. After all, the central tension is between cross-linguistic variation on the one hand, and the knowledge that languages must have some semantic structure in common simply because all humans have essentially the same perceptual system. The model presented here fits in as an element of this line of inquiry, as it is a linguistic learning system whose architecture is inspired by neurobiological and psychophysical insights into the workings of the human visual system.

presents the model's overall architecture, with the non-linguistic structures we have been discussing highlighted. As shown in the figure, the model is configured to learn a set of English spatial terms. Once a movie has been shown to the trained network, the nodes corresponding to those lexemes which describe the event shown should be fully activated. The model is trained using the connectionist error back-propagation training algorithm (Rumelhart *et al.* 1986), while its architecture is informed by the design philosophy of *adaptive structured connectionism*. The fundamental idea behind this philosophy is the combination of elaborate innate structure, such as that found in actual neural structures in the brain, with the remarkable adaptability that characterizes their operation.[4] The model illustrated in Figure 4 is an example of this: the neurobiologically and psychophysically motivated

[4]The general idea of structuring connectionist learning networks is not new; see for example (LeCun 1989; Keeler *et al.* 1991; Mozer *et al.* 1991; Guyon *et al.* 1991), among others. What distinguishes adaptive

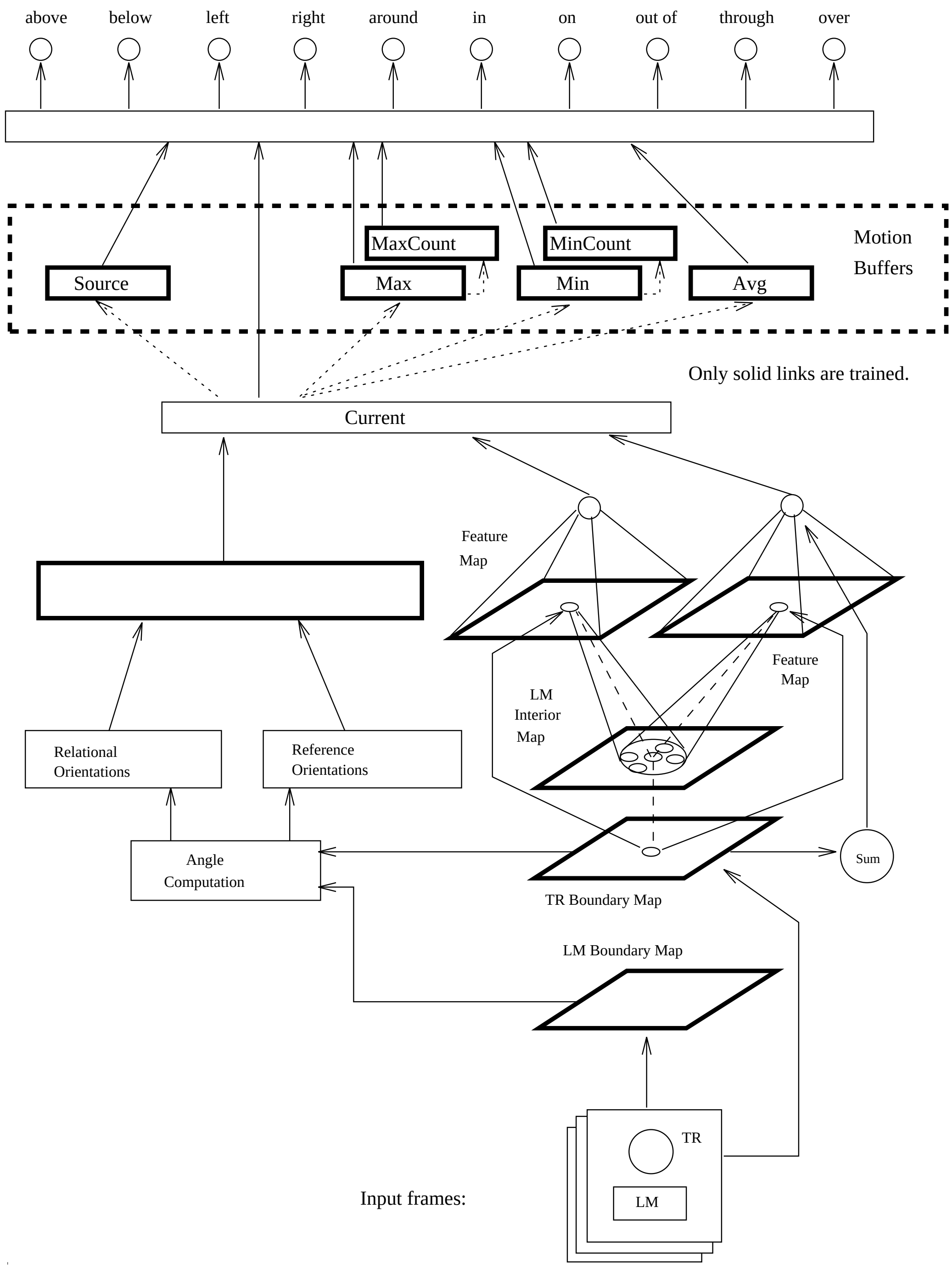

Figure 4: The model's architecture, with non-linguistic structures highlighted.

structures mentioned above are embedded in a trainable network whose parameters are adjusted during training to come to reflect the spatial structuring of a particular language.

Although we shall not be able to describe this network in detail, a few words regarding its operation and ramifications are in order. The model has learned the perceptual grounding for spatial lexemes from English, Mixtec, German, Bengali, and Russian, and preliminary testing has yielded encouraging results for Japanese as well. For all languages, generalization to test sets is excellent. The model is able to learn without the benefit of explicit negative evidence, and as a result of its structure and operation, gives rise to a number of linguistic predictions concerning which semantic features one may or may not expect to find in spatial lexemes in the world's languages. We refer the interested reader to (Regier 1992; Regier 1993) for further details.

4 Current and Planned Efforts

The problem of learning and representing spatial relations terms in the world's language is extraordinarily difficult, and Regier's model is only a beginning. It is limited to 2-dimensional representations, and to artificial scenes with geometrical figures. In addition, it lacks force dynamics, deixis, object grouping, imposed framing, polysemy, and much much more. Like any other fully explicit model, it makes one realize just how difficult the job is.

Regier's model is, of course, limited to spatial relations concepts and does not include any of the metaphorical apparatus necessary to characterize abstract concepts. At present there is no serious model of any significant portion of the system of metaphorical mappings needed to characterize abstract concepts.

Most important, Regier's model does not characterize spatial inferences. To do so, it would need to include some ability to do binding. Such capabilities are available in connectionist models, but until some form of binding is incorporated into some version of Regier's model, this gap remains. There are also a number of other problems with inference in Regier's model discussed below. Both of the current semantic projects focus on extending the representation from layers of units to more active recurrent connectionist networks. It is not clear how far we can go with the existing models, but we are convinced that some version of the schema story presented below is needed for further progress.

4.1 Modelling Verb Semantics for Hand Actions

4.1.1 The Task and Its Motivation

A new project is underway which extends the domain of the original L_0 task. David Bailey's thesis work involves the learning of verbs which apply to activities performed with the hand, such as *push*, *pull*, *get*, *drop*, and *squeeze*, and perhaps harder words such as *open*, *feel*, *stop* and *whoops*. We hypothesize that the semantics of these terms can be expressed (and learned) in terms of suitably-defined, relatively high-level action "schemas"—essentially compiled plans—which the child already possesses when he comes to the language learning task, and which reflect the parameterizations of the underlying neural control mechanisms.

structured connectionism is its emphasis on the use of highly domain-specific structure, such as that focused upon here, rather than very general techniques such as weight-sharing across temporal or spatial locations.

The goal, then, is to build a system, containing a set of schemas for common hand actions, which can learn the semantics of the hand-action verbs of any single natural language, from labelled examples of these schemas executing in various world configurations. The resulting representations must be "active" in the sense that the system, in addition to labelling its activities, must be able to realistically carry out verbal commands in a variety of world configurations.

Once this task is solved, there are several natural extensions to consider. For instance, we can extend the system so that it can learn new, complex tasks by verbal instruction in terms of already-learned verbs (e.g., "to *open*, *grab* then *pull*"). Alternatively, we can model the learning of schemas themselves from goal-based experimentation, as well as the reorganization of schemas based on the language learning.

This project shares many of the motivations of Regier's work, but adds the following:

- A traditional contrast exists in AI between "procedural" and "declarative" forms of knowledge. The procedural form facilitates control, while the declarative form supports flexible reasoning. A challenge presented by this task is that our schemas must support both uses; they must be able to drive (simulated) behavior, but must also support linguistic mappings and compositional reasoning.

- If schemas as well as verbs are to be learned, we have another pair of opposing constraints on their representation: for lexical semantics we prefer a rich, detailed representation; but for schema learning, we prefer the simplest possible representation. A goal is to find a midpoint which is tractable for learning but sufficient for language. Interesting issues also arise from the potential interaction of these two learning tasks. The order of lexical acquisition may depend on and be partially explained by ongoing schema development. Moreover, the language learning task may be able to capitalize on starting out with a smaller schema set. More interestingly, language learning itself may cause re-organization of the schema knowledge, permitting the modelling or prediction of so-called Whorf effects. For example, when a verb becomes associated with two distinct schemas (e.g., *pull* for both PULL-BY-ROPE and PULL-BY-HANDLE), there may follow a consolidation into a single parameterized schema, facilitating the sharing of strategies by the two (say, PULL-BY<GRABBABLE-PART>).

- This project connects with Srini Narayanan's thesis work (described below) involving the mapping of narratives in the "abstract" domain of politics and economics into activities of the body, so that abstract inferences and predictions may be made via this mapping to bodily knowledge. We are hoping that similar schema representations will suffice for both projects.

4.1.2 Schemas

Activity sequences and plans seem to be characterized by "key events" which cause a transition from one action to another. Furthermore, it appears that atomic (though parameterized) primitive actions suffice at this level of representation. For example, a GET-OBJECT schema may work as follows:

> "Concurrently move hand in direction of object and open fist, until contact occurs. Then close fist until all five fingers are exerting pressure p. Then move hand up and toward self."

Thus, we can represent schemas as finite state machines, with a few extensions. First, since multiple primitive actions can take place simultaneously, we need a mechanism for concurrency. Second, we would like our agent's set of schemas to be hierarchical to maximize reuse of simple task strategies when implementing complex tasks, so we plan to utilize the standard augmented transition network (ATN) strategy of allowing actions on transitions to invoke entire (sub-)schemas. Third, we plan to parameterize our schemas along various dimensions, such as object size, speeds, forces, directions, etc. The parameters are used by the schema to compute parameters of its primitive actions and sub-schemas.

4.1.3 Learning Semantics

In many cases, verb semantics will be expressible as schema invocations with particular parameter settings. As an example, consider the following excerpt from Webster's Online Thesaurus:

> "TAKE, SEIZE, GRASP, CLUTCH, SNATCH, GRAB mean to get hold of by or as if by catching up with the hand. TAKE is a general term applicable to any manner of getting something into one's possession or control; SEIZE implies a sudden and forcible movement in getting hold of something tangible or an apprehending of something fleeting or elusive when intangible; GRASP stresses a laying hold so as to have firmly in possession; CLUTCH suggests avidity or anxiety in seizing or grasping and may imply less success in holding; SNATCH suggests more suddenness or quickness but less force than SEIZE; GRAB implies more roughness or rudeness than SNATCH."

If our GET-OBJECT schema were parameterized as GET-OBJECT<speed, force, security, precision>, we might translate the above set of terms as shown in Table 1. (Note we are not able to make all distinctions—more on this below.)

	speed	**force**	**security**	**precision**
take	any	any	any	any
seize	HIGH	HIGH	any	any
snatch	HIGH	LO/MED	any	any
grasp	any	any	HIGH	any
grip	any	any	HIGH	any
grab	any	any	any	LOW
clutch	any	HIGH	LOW	any

Table 1: Verb mappings to the GET-OBJECT schema.

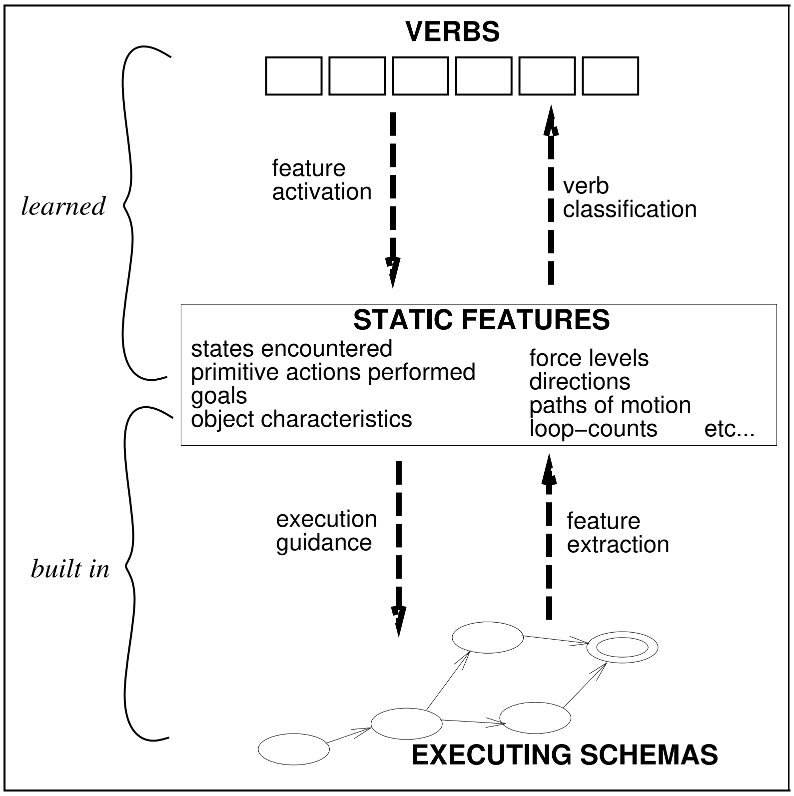

Figure 5: Architecture of the verb-learning system.

However, other verbs (such as those encoding only goal-achievement) may map directly to entire schemas, and still others may depend on details of the execution of schemas such as the number of times a loop is taken. The general system architecture shown in is designed to handle all these cases. A central feature of the architecture is the introduction of a middle layer between schemas and verbs, consisting of a feature vector representing the salient characteristics of a given schema execution. The feature layer is a reduced—and static—description of a dynamic and complex schema execution, and thus provides a strong bias to the learning component, hopefully rendering the learning process tractable. Note that the mappings between the feature and verb layers are bi-directional, a simple consequence of our requirement to be able to drive behavior based on verbal commands. Of course, we must account for the connection between schemas and this feature layer as well. This portion of our system is hand-designed and hardwired, as opposed to learned. Feature *extractors* map from schemas to features, and perform functions ranging from simply passing along the executing schema's goal, to summarizing parameters on the schema's primitive actions, to providing aspectual information based on the pattern of states visited during the execution. In the reverse direction, feature-based *constraints* guide schema execution by either directing control flow or influencing parameter values, in accordance with the current feature vector.

4.1.4 Status

At this point work is proceeding along several fronts. Most importantly, we are implementing a simple version of our architecture for the very limited domain of motions of a cube along a horizontal surface, starting with the hand in contact with the cube. This restriction allows us to focus almost exclusively on parameterization issues, since there is just a single schema and its finite-state structure is trivial. We are working on learning the verb mappings via a clustering algorithm which can handle both numerical and non-numerical parameters simultaneously, and which is also capable of representing multiple prototypes for verbs which require them.

Additionally, we are in a continual process of collecting data on the crosslinguistic variations in hand-action verbs, by eliciting from speakers of various languages their intuitions on the important features behind their verbs, as well as their labellings of various common situations. Besides providing training data, this process also serves to guide our choice of models and to clarify the limits of our schema-based semantics (e.g., for our GET-OBJECT example, we cannot capture the "avidity" of *clutch* nor the "rudeness" of *grab*). From the psycholinguistics literature we hope to extract an informative story on the development of hand-action skills, the order of acquisition of hand-action verbs, and any interesting connections between the two.

Lastly, we are working on a solution to the methodological problem of how to collect training data and how to evaluate the success of our models. Line drawings were an ideal solution for Regier's system; unfortunately, the action-orientation of our task seems to require a robot for this purpose. Furthermore, our claim that details of *human* movement are germane to verb semantics seems to require more robotic sophistication than currently exists. We have found a compromise solution, however, in an existing simulation package tailored to animating human body motions. We are currently working on mapping our proposed primitive actions into simulator commands and success here will legitimize our chosen set of primitive actions as well as provide a suitable environment for judging our models' abilities to both label and generate behavior.

4.2 Metaphor

4.2.1 Introduction

This project attempts to build a computational model of the semantic structure of causal narratives. The input is taken from newspaper stories in the domains of Politics and Economics, and the proposed system should represent the semantics in a way that can generate commonsense inferences about the stories.

Two observations motivate this project. First, causal narratives from abstract domains acquire a significant portion of their semantics from metaphoric mappings to specific schemas in the concrete spatial and temporal domains. We postulate therefore, that an adequate representation of the concrete domain is an essential component of modeling the semantics of these narratives. Recent work in Cognitive Semantics provides good evidence for this view. In particular, there are proposals for schema-based semantics, as in the case of *Image Schemas* (Lakoff 1987). The work in Cognitive Semantics, however, lacks a computational model for such theories.

Second, we note that the deep semantics of the causal narratives are *dynamic* and arise from a *continuous interaction* between input and memory. This enables the high degree of context sensitivity required since changing input context can dramatically affect the correlation between input and memory and thereby the set of possible expectations, goals, and inferences. Any representation of the semantics should therefore be of a fine granularity, be flexible and adaptive in the way context affects interpretation, and be able to support concurrent activities in the memory storage, retrieval, and indexing process.

Our observations suggest the use of *process semantics* (Nielsen *et al.* 1981) to characterize the deep semantics of causal narratives. Of special interest to us is the possibility of modeling such processes in structured connectionist models (Feldman 1989).

This work attempts to develop computational models based on a synthesis from process theory, insights from structured connectionist systems and from linguistic research on Cognitive Semantics to model the commonsense understanding of causal narratives.

We first illustrate these observations by describing their usage in interpreting a recent newspaper story. We then describe the status of this effort and the problems remaining. Finally the scope of the project is summarized.

4.2.2 Example

Consider the following paragraph that appeared in the New York Times on Sunday, August 1, 1993.

> Britain was deep in recession while Germany was booming three years ago. France kept booming long after Germany sunk into recession. Now France is plunging deeper while Germany hopes it has bottomed out. Britain, which was able to cut interest rates after leaving the exchange rate mechanism, has been able to stimulate its economy and has started to emerge from recession.

In interpreting the narrative above, there are several aspects to be noted.

- Germany, France, and Britain are all conceptualized as persons. This is a fairly common occurrence and is based on a conventional metaphoric mapping *Country Is Mapped As Person.*

- Economic recession is treated as a feature in a terrain, namely a *hole* that you can sink into, plunge into, be deep in, emerge from, and a *hole* which can bottom out. The mapping from an abstract economic state (such as Recession) to a concrete spatial location/region is an instance of the metaphoric mapping *Abstract State Is Mapped As Spatial Region.* This particular metaphor belongs to a set of inter-related metaphors called The Event Structure Metaphor (Lakoff 1992).

- There is an implied inference about the *negative* aspects of sinking into holes, and the *positive* aspects of booming or emerging out of them. This kind of directional preference is an instance of a class of *orientational* metaphors. Notice that this preference generates an *implicit goal* for any agent, that constrains the possible interpretations of the input text.

- From the fact that the countries are conceptualized as persons, we note the specific causal events that are used in the article, such as sink into, emerge from, plunge into, stimulate, leave, etc. These are specific *agent motions in space* whose semantics is based on the *manner* of motion, the implied or explicit *path* (such as *into*), the *resultant state* (such as increased energy in stimulate) and *aspect*) (*kept booming* versus *was booming*, *started to emerge*, *has bottomed out*, *is plunging*, etc).

4.2.3 Modeling Approach

Described below are the inputs and outputs of the proposed model followed by a summary of the current status of the effort.

Input/Output

- **Input**.

  Input sentences are parsed using a probabilistic version of the Construction Grammar Interpreter *SAL* developed by Jurafsky (Jurafsky 1992). We are in the process of enhancing the evidential access and selection of constructions in *SAL* to include metaphoric knowledge. Our hypothesis is that metaphors are structures that can potentially increase the information that a linguistic utterance provides about the agent's world model. We are developing a probabilistic theory formally stating and testing this hypothesis.

  The proposed system will also include a lexicon of the relevant action terms and the associated surface semantic role constraints. We hope to test our theory on examples of action terms such as carry, lead, plunge, push, touch, back, cripple, fall, and stumble.

- **Output**.

  The output of the proposed system will be the set of bindings that result from interpreting the abstract narrative as a series of mappings to schemas in the spatial and temporal domain. The mappings make use of metaphors (such as the Event Structure Metaphor) as well as the specific process characteristics resulting from the execution of the concrete domain schemas.

Current Status Our approach to modeling the on-line narrative interpretation system is proceeding along the following dimensions.

- Representation of the set of conventional metaphors that are relevant in the understanding of social causality. We are currently designing our interpretation system using the Event-structure metaphor system since it is well motivated and fits a large portion of our data.[5].

- Investigation of the use of metaphors to guide on-line interpretation. We have developed a probabilistic theory of metaphor selection, and are are implementing our ideas to enhance an existing construction grammar parser.

[5]Our database comprises of English language newspaper stories from around the world. (Lakoff 1992) discusses how the Event-structure metaphor is grounded in our daily experience.

- Representation of the concrete spatial domain. Apart from the general requirement of being able to model states (with real variables) and transitions (including hierarchical transitions) we have identified some specific requirements that pertain to modeling concurrent activation and execution of multiple schemas. These specific features are outlined below.

  1. Should be able to distinguish *enabling* of actions from their *execution*.
  2. Should support specific process primitives such as *temporal order*, *enabling and disabling*, *conflict*, *choice/nondeterminism*, and *covarying actions*.
  3. Should be implemented as a connectionist style of computation (ie. should be based on finite state machines, and should use a generalized spreading activation mechanism for inference).

  The requirements outlined above suggest a net based model of process semantics. The requirements of learnability and connectionist implementation require us to define special modifications to standard concurrency models (such as Petri Nets and Event Logics (Nielsen *et al.* 1981)). We have done some preliminary investigation of such a model by using variant of Petri Nets to model aspectual distinctions made by different languages.

- Mapping the abstract social domain structures into the relevant concrete spatial and experiential domain structures. This includes the selection, initialization and execution of specific concrete domain schemas. To this end, we have divided the domain schemas into the following categories:

  - **The World Model**. This is the overall model of the characteristics of the spatial domain. The world simulation comprises of a model of space called a *spatial map* (that includes relevant terrain features such as holes, bumps, valleys, etc.). The proposed world model also includes an agent that moves through the spatial map affected by various agent initiated as well as external forces (gravity, wind, and friction).
  - **The Body Model**. The body model comprises of a schema that models various body parts, spatial relations between these parts, and various postural constraints (based on relevant joint location, force and motion constraints). We assume a *CSG* model of the body along with a discrete event characterization of the dynamics is sufficient for our purpose.
  - **The Social Cognition Model**. The social cognition model pertains to specific family and social relationship schemas. The various aspects of this model are currently being developed.

  We are currently building a simulation of the agent's *world model* to demonstrate our ideas concerning metaphoric mapping and inference. Consequently, our current focus is on representing the Event-structure metaphor, its use in on-line interpretation, and the inferences obtainable from its source domain (agent motions in space). We believe that a successful model of this mapping can be extended to cover the other concrete domains and their associated abstract domain projections as well.

4.2.4 Summary

This project proposes to investigate some new ideas for modeling the understanding of narratives such as newspaper stories involving political or economic causation. The central novel ideas proposed here are

1. A computational model of narrative understanding by metaphoric mapping from abstract domains, such as politics and economics to the concrete spatial domain. We claim that such a mapping process constrains acceptable parses of an input sentence. We are testing our theory by implementing metaphor integration and selection in an existing construction grammar parser.

2. The translation of the metaphoric interpretation process into multiple, executing models or schemas. This allows for a single system to be query answering, generative and capable of counterfactual reasoning.

3. The grounding of the deep semantics of the abstract causal terms in body-based active models.

5 Synthesis and Conclusion

The theme of this issue is that much can be gained from considering vision and natural language processing problems together. The L_0 project has extended this idea to encompass learning and, more recently, motor control and metaphorical inference. We believe that we have already shown that this broad approach can yield valuable insights and, somewhat surprisingly, practical applications as well. We began by looking hard at the task of first language acquisition in the domain of simple 2D geometric scenes. This task divided cleanly into a concept learning problem and one of inducing grammars from sample sentences. Notice that both sub-problems span issues in learning, language, and spatial representation. Despite considerable initial success, we were unable to extend our paradigm to encompass simple spatial inferences that should be part of knowing the meaning of a term. Current work aims at developing representations capable of supporting learning and inference and also usable as a base for metaphoric mappings from other domains.

Another interesting aspect of the work to date is the range of methodologies used. Although we were originally inclined to attack the L_0 task with connectionist methods on all fronts, we quickly conceded to ourselves that that was not realistic, at least not given the current state of the field. The initial L_0 testbed (Feldman *et al.* 1990b) relied on traditional symbolic AI methods for purposes of rapid prototyping. The grammar learning project started out connectionist (Stolcke 1990), but eventually adopted a probabilistic framework which was thought to provide more versatile representations while still capturing the 'soft computation' aspects of the problem. The spatial semantics learning project has remained the most purely connectionist. The current projects are informed by, but not in any way committed to connectionist solutions.[6]

[6] We might add that the methodological development of the L_0 project is somewhat symptomatic of connectionist research as a whole during recent years.

The entire L_0 project should be viewed as part of a larger movement to develop a grounded and unified cognitive science. We start with the assumption that cognitive science, like other sciences, should be grounded and that base for the grounding must be the physical and biological sciences. The ultimate determiner of our minds is the situation of our existence, co-existence, and evolution on the planet. Most importantly, we claim that taking this stance seriously now can lead to a unified and more productive cognitive science.

The paradigm example of what we have in mind is the work on color-terms in the world's languages and its basis in physiology and psychophysics. Decades of work by people from a wide range of disciplines has produced an understanding of this (admittedly very narrow) aspect of cognition that seems as solid and incontestable as any in the sciences (Berlin & Kay 1969). Our belief is that this paradigm can be extended, albeit with enormous effort, to yield a cognitive science that will establish a continually growing body of established scientific truth.

The central idea is to take very seriously the notion that all of our concepts and cognitive processes have their roots in the nature of our bodies and brains. The color story provides a clear, albeit deceptively simple, version of the fruitfulness of this approach. There are a variety of existing projects that are providing pieces of similar stories in other basic domains. Connectionist models provide a scientific language for linking concrete concepts to the body. The path to treating more abstract domains is being explored in studies of metaphor and other mappings from abstract to concrete representations.

This is not to suggest that everyone should immediately start working at the foundational level. No field has ever developed this way. What is crucial is to recognize that there is an incontestable core of science on which all cognitive theories must eventually be based. It is always relevant to ask the current state of the grounding and it is necessary to be skeptical when no answer is forthcoming.

References

BERLIN, BRENT, & PAUL KAY, 1969. Basic color terms: Their universality and evolution. University of California Press, Berkeley.

BRUGMAN, CLAUDIA, 1983. The use of body-part terms as locatives in chalcatongo mixtec. in Report No. 4 of the Survey of California and other Indian Languages, pp. 235-90. University of California, Berkeley.

FELDMAN, JEROME A. 1989. Neural representation of conceptual knowledge. In *Neural Connections, Mental Computation*, ed. by Lynn Nadel *et al.*, 68–103. Cambridge, Mass.: MIT Press.

——, GEORGE LAKOFF, ANDREAS STOLCKE, & SUSAN HOLLBACH WEBER. 1990a. Miniature language acquisition: A touchstone for cognitive science. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 686–693, MIT, Cambridge, Mass.

——, SUSAN HOLLBACH WEBER, & ANDREAS STOLCKE. 1990b. A testbed for the miniature language L_0. In *Proceedings of the 5th Rocky Mountain Conference on Artificial Intelligence*, 25–30, New Mexico State University, Las Cruces, N.M.

Gull, S. F. 1988. Bayesian inductive inference and maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering, Volume 1: Foundations*, ed. by G. J. Erickson & C. R. Smith, 53–74. Dordrecht: Kluwer.

Guyon, I., P. Albrecht, Y. LeCun, J. Denker, & W. Hubbard. 1991. Design of a neural network character recognizer for a touch terminal. *Pattern Recognition* 24.105–119.

Horning, James Jay. 1969. A study of grammatical inference. Technical Report CS 139, Computer Science Department, Stanford University, Stanford, Ca.

Jurafsky, Daniel, 1992. *An On-line Computational Model Of Human Sentence Interpretation*. Berkeley, CA: Dept. of Computer Science, University Of California dissertation. Available as Technical Report UCB/CSD 92/767.

Keeler, James, David Rumelhart, & Wee-Kheng Leow. 1991. Integrated segmentation and recognition of hand-printed numerals. Technical Report ACT-NN-010-91, Microelectronics and Computer Technology Corporation.

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

——. 1992. What is metaphor? In *Advances In Connectionist and Neural Computational Theory, Vol. 2: Analogical Connections*, ed. by K. J. Holyoak & J. A. Barnden. Ablex.

LeCun, Yann. 1989. Generalization and network design strategies. Technical Report CRG-TR-89-4, Connectionist Research Group, University of Toronto.

Mozer, Michael, Richard Zemel, & Marlene Behrmann. 1991. Learning to segment images using dynamic feature binding. Technical Report CU-CS-540-91, Dept. of Computer Science, University of Colorado at Boulder.

Nenov, Valeriy I., & Michael G. Dyer. 1993. Perceptually grounded language learning: Part 1 — a neural network architecture for robust sequence association. *Connection Science* 5.

——, & ——. 1994. Perceptually grounded language learning: Part 2 — DETE: A neural/procedural model. *Connection Science* 6.

Nielsen, M., G. Plotkin, & G. Winskel. 1981. Petri nets, event structures, and domains, Part I. *Theoretical Computer Science* 13.

Omohundro, Stephen M. 1992. Best-first model merging for dynamic learning and recognition. In *Advances in Neural Information Processing Systems 4*, ed. by John E. Moody, Steve J. Hanson, & Richard P. Lippman, 958–965. San Mateo, CA: Morgan Kaufmann.

Regier, Terry, 1992. *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. Computer Science Division, EECS Department, University of California at Berkeley dissertation. available as Technical Report TR-92-062, International Computer Science Institute, Berkeley.

——. 1993. Two predicted universals in the semantics of space. In *Proceedings of the Nineteenth Annual Meeting of the Berkeley Linguistics Society*. University of California, Berkeley.

Rissanen, Jorma. 1983. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics* 11.416–431.

Rumelhart, David E., Geoffrey E. Hinton, & Ronald J. Williams. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, ed. by James L. McClelland & David E. Rumelhart, 318–362. MIT Press.

Shieber, Stuart M. 1986. *An Introduction to Unification-Based Approaches to Grammar*. Number 4 in CSLI Lecture Note Series. Stanford, Ca.: Center for the Study of Language and Information.

Siskind, Jeffrey Mark, 1992. *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition*. Cambridge, Mass.: Massachussetts Institute of Technology dissertation.

Stolcke, Andreas. 1990. Learning feature-based semantics with simple recurrent networks. Technical Report TR-90-015, International Computer Science Institute, Berkeley, CA.

——, 1994. *Bayesian Learning of Probabilistic Language Models*. Berkeley, CA: University of California dissertation.

——, & Stephen Omohundro. 1993. Hidden Markov model induction by Bayesian model merging. In *Advances in Neural Information Processing Systems 5*, ed. by Stephen José Hanson, Jack D. Cowan, & C. Lee Giles, 11–18. San Mateo, CA: Morgan Kaufmann.

——, & Stephen Omohundro. 1994. Inducing probabilistic grammars by Bayesian model merging. In *Grammatical Inference and Applications. Proceedings Second International Colloquium*, ed. by Rafael C. Carrasco & José Oncina, volume 862 of *Lecture Notes in Artificial Intelligence*, 106–118, Alicante, Spain. Springer Verlag.

Suppes, Patrick, Lin Liang, & Michael Böttner. 1991. Complexity issues in robotic machine learning of natural language. In *Modeling Complex Phenomena*, ed. by L. Lam & V. Naroditsky. New York, N.Y.: Springer Verlag.

Wallace, C. S., & P. R. Freeman. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B* 49.240–265.

Weber, Susan Hollbach, & Andreas Stolcke. 1990. L_0: A testbed for miniature language acquisition. Technical Report TR-90-010, International Computer Science Institute, Berkeley, CA.

Whorf, Benjamin Lee. 1956. *Language, Thought, and Reality*. Cambridge, MA: MIT Press. (ed. John B. Carroll).

Wooters, Chuck, & Andreas Stolcke. 1994. Multiple-pronunciation lexical modeling in a speaker-independent speech understanding system. In *Proceedings International Conference on Spoken Language Processing*, volume 3, 1363–1366, Yokohama.

L_0—The First Five Years of an Automated Language Acquisition Project

Jerome Feldman is director of the International Computer Science Institute and Professor of Electrical Engineering and Computer Science at U.C. Berkeley. George Lakoff is Professor of Linguistics at U.C. Berkeley. The other four co-authors have all been graduate students of Computer Science at U.C.Berkeley, working at the International Computer Science Institute. Terry Regier is now Assistant Professor of Psychology at U. Chicago and Andreas Stolcke is on the scientific staff of SRI International. David Bailey and Srini Narayanan will complete their dissertations in 1996.