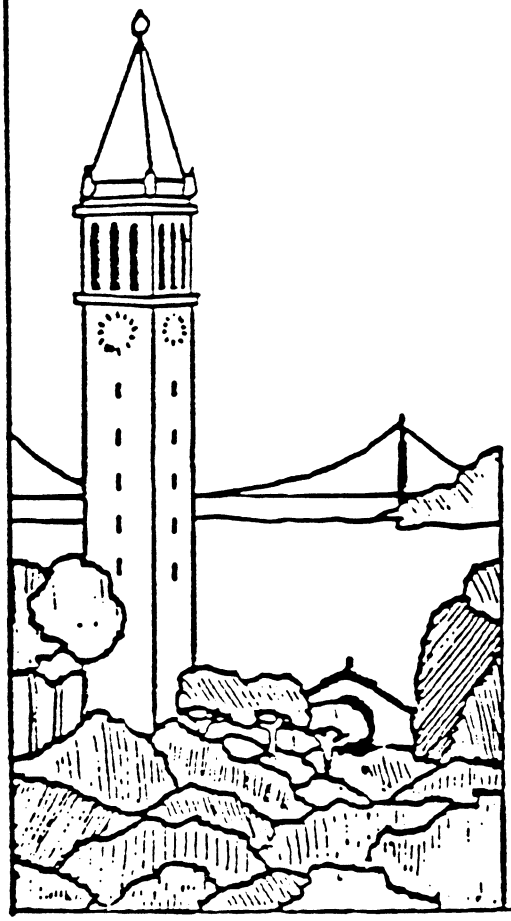
Automatic Inference: A Probabilistic Basis for Natural Language Interpretation

Dekai Wu



Report No. UCB/CSD 92/692

June 1992

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Automatic Inference: A Probabilistic Basis for Natural Language Interpretation

by

Dekai Wu

Automatic Inference: A Probabilistic Basis for Natural Language Interpretation

by

Dekai Wu

Abstract

This work proposes a probabilistic basis for natural language understanding models. It has become apparent that syntax and semantics need to be highly integrated, especially to understand constructs like nominal compounds, but inadequate modelling tools have hindered efforts to replace the traditional parser-interpreter pipeline architecture. Qualitatively, associative frameworks like spreading activation and marker passing produce the desired interactions, but their reliance on *ad hoc* numeric weights make scaling them up to interestingly large domains difficult. On the other hand, statistical approaches ground numeric measures over large domains, but have thus far failed to incorporate the structural generalizations found in traditional models. A major reason for this is the inability of most statistical language models to represent compositional constraints; this is related to the variable binding problem in neural networks.

The proposed model attacks these issues from three directions. First, it distinguishes two fundamentally different mental processing modes: *automatic* and *controlled inference*. Automatic inference is pre-attentive, subconscious, reflexive, fairly instantaneous, associative, and highly heuristic; this delimits the domain of parallel interactive processing. Automatic inference is motivated by both resource bounds and empirical criteria, and is responsible for much if not most of parsing and semantic interpretation.

Second, the nature of mental representations is defined more precisely. The proposed cognitive ontology includes mental images, lexical semantics, conceptual, and lexicosyntactic modules. Automatic inference extends over all modules. The modular ontology approach accounts for a range of subtle meaning distinctions, is consistent with psycholinguistic and neural evidence, and helps reduce the complexity of the concept space.

Third, probability theory provides an elegant basis for evidential interpretation, to model automatic inference in language understanding. A uniform representation for all the modules is proposed, compatible with both feature-structures and semantic networks. Probabilistic, associative extensions are then made to those frameworks. Theoretical and approximate maximum entropy methods for evaluating probabilities are proposed, as well as the basis for a normative distribution for learning and generalization.

Acknowledgements

Whether by design or unavoidable circumstance, an unfortunate consequence of dissertation regulations is the emphasis on sole authorship. Works of this kind are inevitably joint efforts despite having only one direct author, regardless of any claim to independent thought. It gives me true pleasure—and some relief—to have an opportunity to thank the many indirect authors of this dissertation.

First, my teachers. Robert Wilensky first kindled my interest in cognitive science and natural language and has continued to fan the flames throughout my studies. My thinking in general and this work in particular have gained enormous amounts from the insightfulness and persistence with which he would follow through our many discussions and arguments to their foundational assumptions. The numerous comments from his careful readings of draft after draft are only a single case in point. His refusal to overlook the intuitively obvious for the sake of easy formalization has been, and always will be, inspiring to me. Jerry Feldman has likewise been an instrumental influence. Like Robert Wilensky, in spite of an extremely busy schedule he has always found the time to sit down and discuss issues when I needed help. His breadth of perspective and ability to cut through to the key problems have been enormously educational. Over time he has taught me uncountable individual things, but perhaps more important than any of them, he has shown me the subtleties of methodology and the limits of paradigms. Chuck Fillmore and George Lakoff have been wonderfully supportive and their insights, as the reader will see, play a fundamental role in this thesis. Both are unbelievably knowledgeable and they make me wish I could stay another six years at Berkeley to learn more from them. Many, many others have been extremely helpful, among them Steve Omohundro, Stuart Russell, Dan Slobin, Jane Edwards, Steve Renals, and Jitendra Malik. At Berkeley I have been blessed with marvelous teachers who create a fascinating environment, and I am very grateful to my committee and all my teachers for their candor, generosity, insight, comments, criticisms, patience, and support.

Next, my colleagues and cohorts in graduate school. Among the many students and ex-students I'd like to thank for countless discussions, sharing their inexhaustible founts of knowledge, and untold midday cafe and late-night Chinese runs, are, in reverse alphabetical order, Jordan Zlatev, Nigel "dreck" Ward, Andreas Stolcke, Mike "do I really want to do this" Schiff, Terry "most triumphant" Regier, Dana Randall, Erwin Praßler, Peter "world's best landlord" Norvig, Ron Musick, Jim "Fenton's?" Mayfield, Andy Mayer, Jim "world's youngest curmudgeon" Martin, Marc Luria, Erwin Klöck, Dan "that's what she said" Jurafsky, Narciso "bad nj. no thesis." Jaramillo, Marti "neofeminist" Hearst, Othar Hansson, Adele Goldberg, Marie DesJardins, Charlie "big brother" Cox, Dave "aloha" Chin, Mike "funny you should ask" Braverman, Nina "tiny but glamorous" Amenta, and Subutai Ahmad. In particular, I'd like to thank Marti Hearst for implementing code to compute the distributions in chapter 8. Working amidst such bright (and entertaining) folk has been highly productive and pleasurable.

Many have helped negotiate the university obstacle course. Sharon Tague, den mother and moral supporter, has time and again salvaged seemingly disastrous and unnavigable paths. There will never be anyone quite like Sharon. Superb assistance has always been provided by the CS Division staff, in particular Kathryn Crabtree, Liza Gabato, Teddy Diaz, and Jean Root.

And, stimulating as graduate school has been, it is the people one finds in Berkeley, who are some of the best people you'd ever meet, that ultimately makes the difference. I especially

want to thank my bandmates in Nervous for Nigel—Erin Dare, Dan Jurafsky, Terry Regier, Eric Enderton, and Pearl Chow—for providing musical (some might say comedy) relief, and all our friends who put up with us and egged us on, including, besides the folks already listed above, the Hillegass boys and the Theory girls, Seth "so, where is the Nigelmobile anyway?" Teller, Ramon Caceres, Lu Pan, and many others. Plus the great friends outside of CS who have made life life, including Bettina Horster, Bob Chan, Eric Liu, Christiane Henkel, Karl Glaesser, Simon Kao, Trung La, Claudia Schmidt, Jean Shields, the Emmonses, and Mr. and Mrs. Gross. My family have been supportive and loving as always: my parents C.W. and Yvonne, and my sisters Y, H, and T. And, of course, Becky. To everyone I've listed and everyone I could not, thank you from the bottom of my heart.

Two other institutions besides U. C. Berkeley have been very generous. ICSI, the International Computer Science Institute, has given me a second home with highly diverse and brilliant staff and visitors. Also, Christian Freksa and Prof. Wilfried Brauer generously supported me during 1986–87 which I spent at the Technische Universität München, shaping many of my early directions.

Finally, none of this work would have been possible without funding from the Defense Advanced Research Projects Agency (DoD), monitored by the Space and Naval Warfare Systems Command under N00039-88-C-0292, the Office of Naval Research under contract N00014-89-J-3205, and the Sloan Foundation under grant 86-10-3.

Brief Contents

Contents

Chapter 1

Chapter 1

Introduction

What is the connection between language interpretation and rationality? I argue that much of the human language facility follows rational principles—enough, in fact, to warrant studying interpretation as the heuristic approximation of an ideally rational agent.

And though historical divisions would have us believe otherwise, I argue that a heavy part of this burden is borne by an associationist component. Rational behavior inherently demands adaptive forward inference, because we must be able to interpret and act effectively, given limited computational resources. These forward inferences are cued solely by contextual and shallow perceptual features and are made *probabilistically* based on knowledge of syntactic, semantic, and conceptual usage patterns.

Consider the case of nominal compounds, on which I focus throughout this thesis. Examples include *desk drawer handle, cleaner equipment firm, rubber baby buggy bumper, front wheel, high-speed buses,* and *coast road.* People understand nominal compounds easily and effortlessly, and fit them coherently into the surrounding context. For example, the interpretation most people prefer for *coast road* is a road that runs along the seacoast, even if they have never heard the phrase. Yet for *coast wheel* people rarely suggest interpretations having to do with the seacoast at all, and instead prefer the unpowered-movement sense of *coast.* Moreover, even *coast road* can have other interpretations; consider

(1.1) Since the earthquake damaged the only Interstate to the coast, old Highway 17 will temporarily be the main coast road.

Here *coast road* means a road that runs to the coast. Some factors that may enter into one's interpretation of *coast road* in the more usual case are:

- The word *coast* is used slightly more often to mean a seacoast rather than an unpowered movement.
- Nominal compounds are used to express containment relationships about as often as spatial direction relationships.
- Most of the time when one thinks about roads in the context of seacoasts, one thinks specifically of the *coastal road* subcategory of roads.

- Roads that run to the coast are not mentally subcategorized as *coastal roads*.[1]
- Living on the West Coast, Highway 1 is a frequently used concept of the *coastal road* subcategory.

People integrate such factors effortlessly, yet despite the ease with which people process nominal compounds they have defied linguistic analysis. Non-computational theories at most posit intuitive classifications of the nominals' semantic relationships that are too underspecified to implement. On the other hand, computational theories resort to *ad hoc* heuristics, and there is inadequate motivation to expect them to generalize successfully to interestingly large conceptual domains.

The proposed model adopts a three-pronged probabilistic approach. First, it distinguishes two fundamentally different modes of inference—*automatic* and *controlled*—to separate out the associative parts of interpretation. Second, to combat underspecified semantic classifications, it puts forth concrete ontological proposals for handling some finer points of semantic and conceptual representations, including feature *structures* rather than simple feature sets. Third, to avoid *ad hoc* heuristics and yet keep parsing and semantic interpretation highly interactive, it combines evidence using probability theory and maximum entropy.

FRIEZE is an implementation of this theory. Using probability theory and entropy maximization to combine evidence and prior knowledge, it produces the most probable interpretations of nominal compounds. Unlike other approaches to this problem, FRIEZE is capable of integrating constraints from related facts to discriminate interpretations of novel as well as familiar compounds. Figure 1.1 previews the flavor of probabilistic evidence integration for *coast road* (the example, described later, is taken from figure 7.16). Results for two separate runs are shown, to demonstrate the effect of probabilities. Each node is an abbreviation for a feature structure incorporating lexicosyntactic, semantic, and/or conceptual constraints. The internal nodes (those not in the bottom row), like *C:coast:seacoast*, are annotated with marginal probabilities that indicate how often the agent uses structures meeting those constraints. For example, a probability of $1 \cdot 10^{-4}$ (abbreviated "1e-4") is assumed on the structure for a noun compound expressing a containment relationship such as *road in coastal area*. Similarly, in the first run, a marginal of $1.9 \cdot 10^{-7}$ is assumed for "coast" expressing *seacoast*, whereas $1.5 \cdot 10^{-7}$ is assumed for "coast" expressing the unpowered-movement *coasting accomplishment* sense. These are switched for the second run. The nodes at the bottom are hypothesized automatic inferences; their probabilities are estimated using a *maximum entropy* procedure and the hypothesis with the maximum resulting probability is selected. The first row of probabilities beneath them shows that in the first run, *road in coastal area* and *road along coastline* are preferred. The second row shows how the preference switches to *coasting road* (a road on which to coast) by assuming a different usage pattern on "coast". In the same way, the hypotheses' probabilities are sensitive to all the factors listed earlier. Thus, this model of a linguistic agent is *adaptive*, since the marginal probabilities are (in principle) derived from the agent's experience.

[1] The name *coastal road* is just a convenient label for a subcategory; I don't mean to suggest the phrase "coastal road" couldn't also be used to mean a road to the coast.

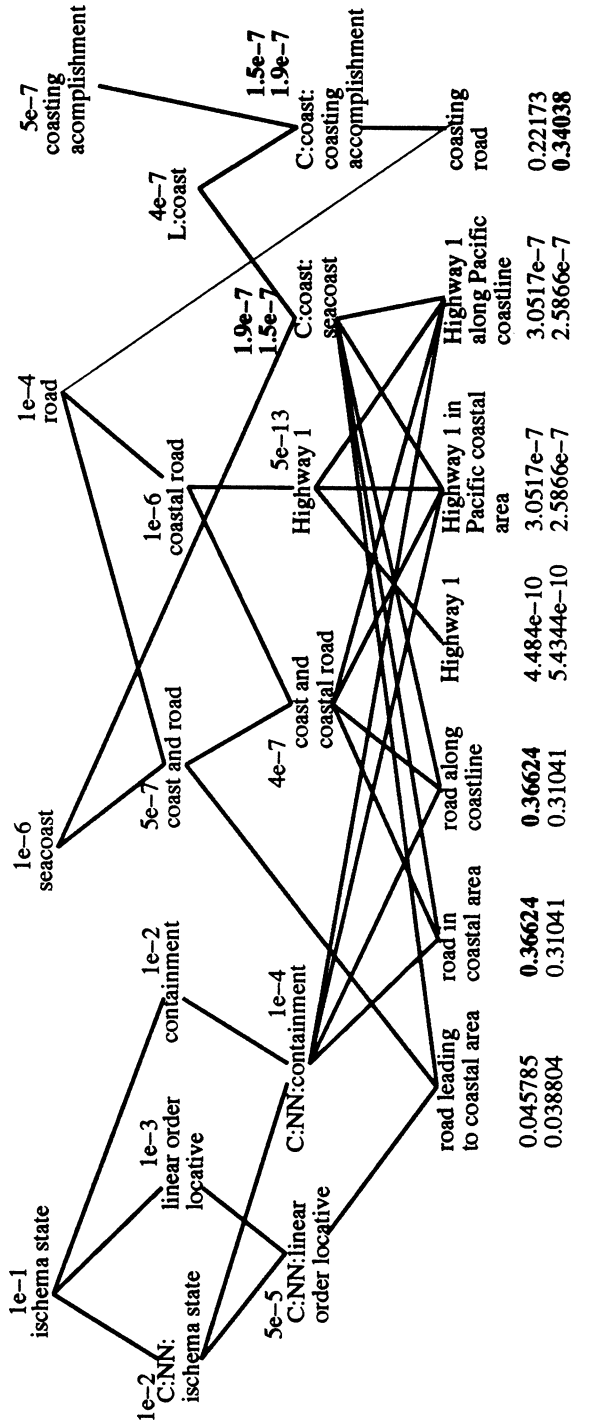


Figure 1.1: A preview of probabilistic evidence integration for interpreting *coast road*.

1.1 Overview of Major Claims

1.1.1 Automatic Inference

The major theoretical paradigm that this work is situated in—and which it aims to sharpen and define—is a distinction between *automatic* and *controlled inference*, particularly in connection with though not limited to language comprehension. Table 1.1 summarizes how assorted characteristics break down between automatic and controlled inference. Automatic and controlled inference are conceived of as the two most general classes of inference, with sufficiently different properties as to warrant two distinct modelling frameworks. The terms "automatic" and "controlled" are psychological terms describing whether attention is required to perform some cognitive process (LaBerge & Samuels 1974; Posner & Keele 1975; Shiffrin & Schneider 1977). Automatic inference is pre-attentive and subconscious, is reflexive and fairly instantaneous, is associative and heuristic, and most likely is implemented by massively parallel processes. Controlled inference involves conscious problem segmentation and sequential chaining. In general a process requires less attention and becomes more highly automized when it is extensively practiced, either by frequent occurrence or by rehearsal.

Besides empirical cognitive arguments for automatic inference, there are theoretical motivations stemming from the fact that an agent's resources are finite. As I argue in chapter 3, rational behavior requires a resource-bounded agent to make fast forward inferences based only on ready information. (Forward or data-driven inferences are automatically triggered by new input, as opposed to backward or demand-driven inferences which are triggered by the agent because of some high-level need.)

The distinction between automatic and controlled inference lies in the type of processing, not in the type of knowledge. It is a horizontal rather than vertical modularization. As the introductory example implied, automatic inference spans lexicosyntactic, semantic, and conceptual domains, and is not confined to any subset of ontological modules.

The inferences produced by automatic inference need not always turn out to be the interpretations ultimately chosen. To be effective, automatic inference merely needs to produce useful intermediate results most of the time. In this model automatic inference gives rise to cognitive default and prototype effects, where subjects reach logically unwarranted conclusions that are plausible but may later need to be rejected. Thus automatic inference is also related to non-monotonic inference, though this is not addressed since it also depends on intervention by controlled inference.

The theoretical analysis makes use of probability and decision theory. Ultimately, chapter 3 derives a probabilistic formulation of the automatic inference task, various aspects of which are addressed by the subsequent chapters. The nature of probabilities and quantitative measures, and justification for their use in a cognitive model, is discussed in section 2.2.5.

Automatic Inference	Controlled Inference
Instantaneous Inference is quick, momentary, spontaneous.	**Arbitrarily prolonged** Inference can be sequentially chained for indefinite periods.
Data-driven Triggered by new input data, either bottom-up perceptual input or top-down conceptual input.	**Goal-driven** Triggered by high level goals.
Primitive Performed by basic mechanisms.	**Derived** Procedures for performing controlled inference built upon the more basic ability to learn and chain action sequences.
Pre-attentive	**Attention required**
Subconscious	**Conscious**
Reflexive	**Reflective, deliberative**
Reconstructive Inherent tendency to reconstruct, from previous experience, the situation that probably gives rise to the perceptual input information.	**General purpose** Problem solving techniques for many different types of tasks can be learned.
Bounded capacity Only relative small conceptual chunks can be handled at a time.	**Segmentative** Complex problems can be segmented into arbitrarily many manageable steps.
Evidential Inferences adhere to some form of probability, likelihood, or plausibility optimization, subject to resource bounds.	**Non-quantitative** Inferences do not necessarily involve weighted comparisons.
Online Sensitive to the timing (and thus, order) of input information. This produces context sensitivity, as prior inputs establish the context.	– Meaningless; too slow and general –
Non-monotonic Commitments to default inferences are retracted to maintain consistency with subsequently acquired contrary evidence.	– Not a defining characteristic –
Heuristic Produces quick results that are useful most of the time.	– Not a defining characteristic –
Supports controlled inference by providing heuristic memory retrieval.	**Depends on automatic inference** for heuristic memory retrieval.
Associative models are the most promising for efficiency reasons, especially massively parallel models.	**Logical models** have produced more successes.

Table 1.1: Differentiating characteristics of automatic and controlled inference.

1.1.2 Modular Ontological Model

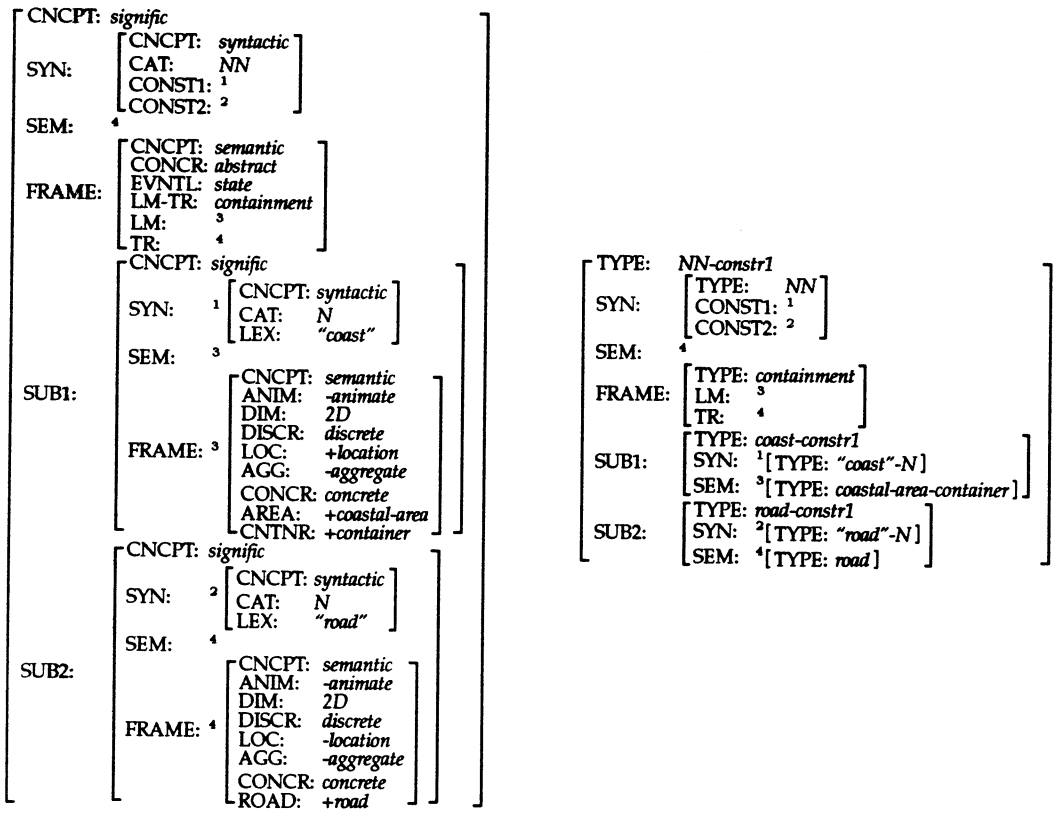
The second major point of attack is to define more clearly what kinds of intermediate semantic representations automatic inference should produce. The term *ontology*, in the sense in which AI has adopted it, means a particular theory of the world and what exists in it. Realists use the term to refer to attempts at representing the "true" nature of the existence ("*the* ontology"). The mentalist sense, taken here, refers to a particular linguistic agent's ordinary everyday representation of the world, which may be inaccurate, incomplete, inconsistent, and biased. The agent may even be consciously aware of this—Newton's laws are not entirely accurate—but such representations can nonetheless be useful for the majority of the agent's interactions with the environment.

The choice of ontology affects the inductive bias in a probabilistic model just as in any logical model. In choosing the representational primitives, both expressiveness and empirical concerns must be taken into account. Chapter 4 observes some representational needs that are often overlooked in AI knowledge representations and linguistic semantics models. It also surveys empirical evidence from various cognitive disciplines, converging towards a modular mental ontology. These desiderata are synthesized into an organization including modules for mental image, lexical semantics, conceptual, and lexicosyntactic structures. Chapter 5 then puts forth a concrete set of representational primitives for these modules, using a formalism that is amenable to both unification grammar approaches and probabilistic modelling.

A Note on Notation. Before discussing how the probabilistic model facilitates evidential interpretation, let me point out some important and potentially confusing terminology and notation that I use.

1. *Frames versus concepts.* In the proposed model, concepts and frames are distinguished primarily as a matter of convenience. When I use the term "frame" I am emphasizing the roles and the relationship between internal structures of a concept. When I use the term "concept" I am emphasizing the gestalt properties of the entire structure. There is one slight difference, namely that frames must always have explicit internal structure, whereas concepts can be primitive, atomic feature bundles.

2. *Feature structures and graphs.* Though a wide range of notations is employed in the literature, the notations have similar expressiveness and are equivalent in many cases. It is sometimes convenient or more intuitive to think of feature-structures using either graph/network notation or predicate logic. Particularly useful are typed notations of either the feature-structure or semantic network variety, which define *types* using an inheritance hierarchy so that conceptual structures can be written (and stored) using pointers to the type rather than by explicit enumeration of every feature value. I will be using typed notation throughout. Since these techniques are well known, rather than giving a formal definition, figure 1.2 simply shows an example of the same structure written using a feature-structure, typed feature-structure, and typed DAG (directed acyclic graph) notation. This figure can be used for later reference; it is not essential to understand the representational details yet as they will be gradually introduced.



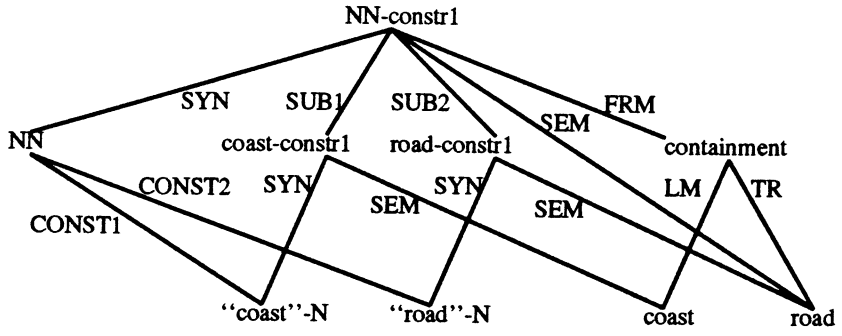


Figure 1.2: An example of the same structure using equivalent feature-structure, typed feature-structure, and typed DAG notations.

1.1.3 Evidential Interpretation

The third major line of attack is to develop a probabilistic basis for integrating evidence and knowledge sources across all the ontological modules. Until now probabilistic models have seen little development in natural language work. However, recent interest in neural, connectionist, and probabilistic models on the one hand, coupled with interest in large electronic corpora on the other, is leading to a resurgence of statistically based approaches to language processing and acquisition.

The primary contribution of the proposed evidential interpretation model is to extend probabilistic modelling to complex structures such as those in figure 1.2. This is particularly important for semantic and conceptual interpretation because simple feature sets or vectors are insufficiently expressive for representing meanings involving relations or roles. Structured representations have been problematic for statistical methods, both in computational linguistics and in AI and neural networks. Most linguistics-oriented methods only analyze probability distributions on feature sets. This includes current large-corpus techniques as well as the statistical approaches of the late 50's and early 60's. Similarly, existing distributed neural network models work on feature vector representations. Proposals for storing compositional structures with roles and relations have been advanced, in effect translating structured representations into feature vectors. However, so far this has turned out to be ineffective for statistical modelling because those models that can successfully handle a wide range of structures end up simply using neural networks as a sequential memory, and thus miss important generalizations over similar structures. Thus the advantage of using a neural network is lost. This problem is known as the *variable binding problem* and continues to be an active area of research.

The proposed model operates directly on structured symbolic representations. In fact, one of its advantages is that familiar unification grammar feature-structure and semantic network formalisms are used. A (theoretical) probability distribution is placed directly over the *concept space* of possible feature-structures, thus avoiding the variable binding problem. Chapter 6 analyzes the relationship of probability to semantic network and feature-structure representations, and proposes probabilistic extensions. I argue that the probabilistic approach stays more faithful than the logical approach to the traditional conception of semantic networks as associational representations.

Throughout this work, attention is paid to establishing the meaning of the probabilities, an issue introduced in chapter 2. A common shortcoming among language interpretation models that employ quantitative measures is that they lack motivation for the numbers used. The proposed model combats *ad hoc* measures in three ways. First, a clearly mentalist stance is taken, so that there is no confusion as to whether probabilities are realist (objective probabilities) or mentalist (subjective or logical probabilities). Toward this end chapter 3 embeds the function of probabilistic automatic inference within a situated linguistic agent, and chapter 4 explains the significance of the intermediate conceptual structures over which probability measures are placed. Second, if the investigator is able to choose *a priori* the patterns believed to be relevant within the domain of interpretation, then the probability distribution over the patterns can be estimated by loading the statistical distribution from a training set, as described in chapter 7. In contrast many quantitative models give no derivation procedure to ground the numbers. Third, a theoretical distribution, defined in chapter 8, is proposed as a normative learning theory in the event that the relevant patterns are not known beforehand. The distribution is motivated by considerations of desirable generalization behavior. It is unlikely that distributions of this kind could actually be stored using

a reasonable number of patterns; instead patterns and probabilities should be chosen so as to approximate the normative distribution.

Chapter 2

Chapter 2

Background: Nominal Compounds and Probability

This thesis proposes a probabilistic basis for language understanding models; the domain of application is the interpretation of nominal compounds. This chapter surveys work relevant to the latter, and presents some basic principles underlying the former. The reader may skip either or both sections without seriously impairing the line of presentation of concepts in this work. However, the latter outlines the philosophy of probability, and familiarity with some of the philosophical issues will be necessary to understand the subtler points of the proposed usage of probabilities.

2.1 Theories of Nominal Compounds

In this section I survey some theories of nominal compounds. Nominal compounds are simple and small, yet provide a fertile ground for studying semantic and conceptual approaches to language. They often involve homonymous nouns—is *dream state* a sleep condition or California?—and this necessitates resolving lexical ambiguity. Nested nominal compounds involving three or more nouns have more than one parse and therefore require resolution of structural ambiguity; for example, consider [*baby pool*] *table* versus *baby* [*pool table*].

I distinguish three major classes of nominal compounds:

1. *Lexicalized*, such as *clock radio*. Such compounds have established conventional interpretations; to handle these, lexicosyntactic biases must be integrated into the interpretation process.

2. *Identificative*, such as *clock gears*. These compounds are novel in Downing's (1977) sense, in that they are not conventional phrases. However, they do identify a conventional conceptual schema, since one knows about gears in clocks beforehand, from experience.

3. *Creative*, such as *clock table*. These are also novel compounds, but interpreting them requires the hearer to create a new conceptual structure, such as *table on which a clock sits*. For creative as well as identificative compounds, semantic and conceptual biases play an important part in guiding composition tasks like frame selection and role/slot binding.

The interaction of biases can become quite complicated. Normally, there is a bias to use the most specific pre-existing categories or schemas possible—syntactic, semantic, or conceptual. Preference is first given to lexicalized forms, then to identificative interpretations, and lastly to creative interpretations. However, this sort of "Maximal Conventionality" principle can easily be overridden by global factors arising from the embedding phrase and context. The first two columns of Table 2.1 show nested compounds where both potential parses involve lexicalized forms. Depending on semantic and conceptual factors, the preferred interpretation can either include the more lexicalized compound as in [*kiwi fruit*] *juice* or *baby* [*pool table*], or it can break the more lexicalized compound as in [*navel orange*] *juice* or *New York* [*state park*]. The rightmost column of Table 2.1 shows even more extreme cases where an identificative compound like *afternoon rest* breaks a lexicalized compound like *rest area*.

PREFERRED PARSE	COMPETING LEXICALIZED COMPOUNDS		COMPETING LEXICALIZED AND IDENTIFICATIVE COMPOUNDS
	First compound more lexicalized	Second compound more lexicalized	
 N N N	kiwi fruit juice LEXICALIZED LEXICALIZED	navel orange juice LEXICALIZED LEXICALIZED	afternoon rest area IDENTIFICATIVE LEXICALIZED
 N N N	New York state park LEXICALIZED LEXICALIZED	baby pool table LEXICALIZED LEXICALIZED	gold watch chain LEXICALIZED IDENTIFICATIVE

Table 2.1: Interacting biases in competing lexicalized and identificative interpretations (see text).

2.1.1 Descriptive, Predictive, and Situated Language Models

Nominal compound theories, and theories of natural language in general, can be structured as *descriptive, predictive,* or *situated* models. These terms are qualitative and there are no exact lines. Descriptive and predictive models of language are primarily theories that characterize aspects of language from the external scientific observer's perspective, without characterizing the precise processes by which a language user operates in real time. In contrast the distinguishing criterion of situated language models is the constraint that the theory explain how the resource-bounded language user processes information under time constraints, in order to function within the environment. In Chomsky's terms a situated language model must incorporate not only language competence but also performance. Situated models are called *agents.*[1] Although there is a good deal of overlap between what is labelled "computational linguistics" versus "natural language processing", computational linguistics more strongly connotes the descriptive and predictive models and natural language processing more strongly connotes situated models.

Situated models have been studied predominantly in simpler domains than natural language. The most basic type of environment is one of survival; one popular mode of research is to construct agents whose purpose is to survive in a video-game micro-world (Agre & Chapman 1987; Russell & Wefald 1991). In the case of natural language modelling the survival goal is more remote and an agent may instead emphasize satisfying cooperation goals. The Unix Consultant (Wilensky *et al.* 1988) is an agent that operates in an environment in which the agent's sensory input is limited

[1] Not to be confused with the case role.

to the user's typing, and the agent's actions are limited to generating textual output.[2] Natural language is a tool that can be used to help it achieve its goal of assisting the user; toward this end both interpretation and generation are useful. The later Wittgenstein (1963) view of meaning as being the *use* of a linguistic utterance or text, both conventional and contextual, is naturally captured by the situated language model.

Grammatical versus stimulus-response models. In a *grammatical* model no function or task is explicitly specified; the grammar is a collection of rules that aim to describe and/or predict linguistic phenomena. Traditionally we conceive of a grammar in terms of being able to generate all the acceptable strings of a language, given unlimited time to apply all the rules. In contrast a *stimulus-response* model associates an input stimulus with an output response, where either or both are strings belonging to a language. A stimulus-response model is therefore functional or task-oriented. As we shall see both grammatical and stimulus-response models can be probabilistic.

Stimulus-response models come in a variety of flavors. The most theory-neutral models are *behaviorist*, attempting only to model the externally observable behavior of a language user without postulating what internal states the user's cognitive mechanisms must pass through. Another predominant class of models are *interpretive* and seek to describe how linguistic inputs cause semantic or conceptual structures to be constructed; of course these structures are highly theory-dependent. Conversely the object of *generative* stimulus-response models (not to be confused with generative grammars) is to transform semantic or conceptual inputs into linguistic structures. Clearly a behaviorist model may contain interpretive and generative components but no claims are made about cognitive correspondence.

Converting grammars into stimulus-response models. For certain classes of grammars, methods are known for generating particular kinds of stimulus-response models. The most obvious examples are the many parsing algorithms for accepting context-free grammars. The stimulus is an input string; the response is an accept or reject signal, or a parse tree.

For situated agents in most environments, merely computing an accept or reject signal is not a very useful response. However the basic technique can be used to produce more interesting response behavior by augmenting the grammars with things like attributes. For example, if the Noun category can have syntactic attributes like "gender" or semantic attributes like "count/mass", then the attributes effectively split the Noun category into a number of subcategories. We can then require the parser to decide not only that an input item is a Noun, but also which specific subcategory of Noun it belongs to. This produces a simple kind of a semantic interpretation as the output response.

The problem is that after adding attributes (or other extensions) to the grammar the old parsing methods no longer apply. In general adding attributes causes the grammar to become ambiguous, especially when the attributes are semantic and have no directly corresponding syntactic surface form realization. Simple parsers have only random or ad hoc means of resolving ambiguities, and more motivated disambiguation methods require additional assumptions about cognitive processing biases; here statistically justified heuristics play a significant role.

[2]Survival goals are present in limited form; the agent has a built-in rudimentary model of its own existence, which says that the agent's own existence depends on the continued well-being of its host computer (Chin 1988). Thus the user will be denied information about means of sabotaging the host computer.

2.1.2 Descriptive and Generative Accounts

Historically, generative grammar has pursued descriptive and predictive aims, while interpretive models have paid more attention to situatedness. Actually there are certain equivalence classes of models that can be arbitrarily transformed between generative and interpretive models. However, because the kind of claims made by theories vary significantly with their descriptive or interpretive orientation, I keep the traditional distinction.

Jespersen. Jespersen's (1946, v. 6) analysis of compounds is restricted to a survey of some general classes of relations that can hold between the constituent elements of a compound. For nominal compounds, which Jespersen calls "substantive-compounds", these classes include subject-action (*nightfall*), location (*garden-party*), destination (*land-breeze*), instrumental (*sabre-cut*), and so on.[3]

It is debatable whether Jespersen claims generalizations can possibly be made about the relations in a compound. On the one hand, Levi (1978, p. 105) reads Jespersen as saying nominal compounds are "inherently idiosyncratic", drawing as evidence the following quotes:

> (a) Compounds express a relation between two objects or notions, but say nothing of the way in which the relation is to be understood. That must be inferred from the context or otherwise. (p. 137)
>
> (b) On account of all this it is difficult to find a satisfactory classification of all the logical relations that may be encountered in compounds. In many cases the relation is hard to define accurately. (p. 137)
>
> (c) No definite and exhaustive rules seem possible. (p. 140)
>
> (d) The number of possible logical relations between the two elements is endless. (p. 143)

On the other hand, the fact that Jespersen even bothers to enumerate classes of relations implies some belief in systematic subregularities. Probabilistic systematicity, as in the model I propose, is possible even if the classes are not exhaustive, and is compatible with quotes (b), (c), and (d). This leaves only quote (a), which is problematic since the constituents of a compound do say something about the likelihood of the possible relations. But Jespersen's "say nothing of" is probably just a careless statement, not intended to rule out probabilistic tendencies.

Similar general classifications are found in other grammars. For example, Quirk *et al.* (1985, pp. 1330–1335) also classify the subclass of nominal compounds they call "premodification by nouns" into "source-result" (*metal sheet*), "part-whole" (*clay soil*), "place" (*top drawer*), "time" (*morning train*), and "whole-part" (*board member*). Nominal compounds are not the focus of such accounts, and are treated purely descriptively and not in depth.

Levi. Levi's (1978) analysis of "complex nominals", a superclass of nominal compounds, is set within a generative semantics framework. The main innovation is a claim that in most compounds the relation between the constituents is one of nine predicates. These are CAUSE (*disease germ, birth*

[3] Jespersen does not give names to the classes.

pains, HAVE (*apple cake, lemon peel*), MAKE (*silk worm, snowball*), USE (*steam iron*), BE (*target structure*), IN (*morning prayers*), FOR (*arms budget*), FROM (*test-tube baby*), and ABOUT (*price war*). The first three predicates have two variants corresponding to the example pairs. In one, the modifying noun is the object of the relative clause implied by the predicate; in the other it is the subject. Because of the generative semantics framework, Levi thinks of compounds as being transformed from longer paraphrases with explicit predicates, like *a germ that causes disease*. The predicates are deleted by the transformation, but can be recovered by the hearer; thus Levi refers to the predicates as "Recoverably Deletable Predicates" (RDPs).

From the point of view of the conceptual system, it is not clear that the set of RDPs actually constrains the range of semantic relations that a surface form might potentially signify. The RDPs are extremely general concepts, and can be seen as the merely the most abstract forms of an infinite number of more complex types of roles. For example, Schank's (1973) Conceptual Dependency representation is limited to a set of relations of the same order of magnitude, but an infinite number of complex structures can be constructed around the relations, in effect representing many of those variations that Jespersen claims cannot be exhaustively enumerated. Semantic network hierarchies make this even more explicit with role hierarchies; in such a conceptual representation one might expect to find roles similar to the RDPs near the top of the abstraction hierarchy.

Another somewhat unclear point is what distinguishes RDPs as semantic rather than syntactic forms. The analysis is clearly conceived from a semantic perspective, but the transformational paradigm of generative semantics makes the process of deleting RDPs from relative clauses resemble a syntactic transformation, as in Lees' (1963, 1970) treatment. The problem is that the predicates, even when made explicit in the paraphrases, are still polysemous; the metaphoric temporal IN in *morning prayers* should be distinguishable from the IN in *field mouse*. Independent justification of the semantic primitives, not connected to their use in nominal compound paraphrases, would be one way to improve this.

Warren. Warren (1978), like Levi, postulates a number of primitive semantic relations that can hold between a nominal compound's constituents. However, Warren's relations are somewhat more detailed in that they actually form a taxonomic hierarchy. At the most abstract level, the primitive relations are Constitute, Belonging-To, Location, Purpose (Goal-Instrumental), and Activity-Actor (OBJ-Actor).[4] The next more specific level includes relations such as Source-Result, Copula Compounds, Whole-Part, Part-Whole, Size-Whole, Goal-OBJ, Place-OBJ, Time-OBJ, and Origin-OBJ. Still more specific are relations like Material-Artefact (*clay bird*) and Matter-Shape (*raindrop*) Though Warren's semantic categories are, like Levi's, not independently justified, Warren's analysis has a more conceptual bent than Levi's, since there is less reliance on paraphrastic transformations with their polysemy problems.

The primary contribution of Warren's work is a statistical analysis of the frequency of semantic patterns over a relatively large set of compounds. Warren's corpus of 4,557 distinct nominal compounds was taken from 180 of the 500 texts in the Brown Corpus (Kučera & Francis 1967), a total of about 360,000 words. Working with standard, widely available data is a great advantage in terms of facilitating comparisons and building upon previous analyses. I have therefore concentrated on compounds from Warren's corpus; most of the examples in this work come from Warren and the sources are footnoted.

[4] Warren also treats proper name combinations.

2.1.3 Interpretive Accounts

Downing. Downing (1977) argues against the paradigm of deriving nominal compounds from underlying structures. Concentrating on experiments in which subjects produced novel compounds in the course of performing naming tasks, Downing argues that the set of potential compounding relationships is infinite. She also argues that the possible interpretations of a compound are constrained by the fact that speakers choose to use a compound construction only subject to pragmatic functions (e.g., naming in context). Another interesting result is variation in the frequency of occurrence of general semantic relationships, depending upon the semantic type of the head noun (human, animal, plan, natural object, synthetic object). Such variation is consistent with Warren's statistical findings.

Leonard. Leonard (1984) describes a computationally implemented algorithm for interpreting nominal compounds, in which input compounds are paraphrased as noun phrases with prepositional phrases (*fir bough* becomes *the bough of a fir*) or relative clauses (*hire-car* becomes *a car that someone hires*. She employs a similar typology to Warren's, but uses her own somewhat smaller corpus of 1,944 compounds, drawn from 305,000 words from sixteen novels.

Each noun participating in a compound is marked with semantic features; Leonard uses four classes of semantic features, which different rules are sensitive to. The 21 "primary features" include "The noun is related to a verb" (*attack*), Locative, Material, Plural, and so on. The 22 "secondary features" apply only to nouns that are related to verbs, and include such features as "Related to a covert verb" (*accident* happens) and "Instrument of an overt verb" (*pick*). There are nine "tertiary features", also called "semantic fields", including Mechanism, Human organization, Part of the body, Plant or tree or part of one, furniture, and so on. The five "quaternary features" really just encode the relative size of an object using a discrete five-level scale. To interpret a compound, a sequence of nine rules is applied; preference is strictly determined by the order of the rules. The rules search for things like "match in semantic fields", which occurs when two nouns share a "tertiary" semantic feature (*plant or tree* for *fir* and *bough*). Another rule searches for a "material head", such as *stone* or *clump*; this is paraphrased as *[head] composed of [modifiers]*.

Leonard reports a 76% accuracy rate for the nine rules applied to her corpus. It is probably true that at least this percentage of compounds found in text can be reasonably paraphrased by such relatively simple techniques. However, as is usually the case in language processing, the last 24% or so is the difficult part that requires integrating syntax, semantics, and complex conceptual structures.

Moreover, it is not entirely clear whether Leonard's program is more an interpreter or more a paraphraser. In some sense, paraphrasing can be easier than interpretation, because it can be regarded as a kind of syntactic transformation, where the burden of interpretation is still unconsciously being performed by the investigator rather than the model. For example, the phrase *the bough of a fir* does not actually specify a very prices semantic relation; it is we, the investigators, who impose the interpretation of *of* as a part-whole relationship. In this particular example, Leonard's program does actually have an internal representation of the part-whole ("Annex") relation, but in general the use of paraphrasing for evaluation is susceptible to this problem.

McDonald. McDonald's (1982) computational model interprets nominal compounds by combining a number of heuristics. Some of these heuristics pre-empt others; others are weighted and

combined. The heuristics can be broken down into two families (my analysis, not McDonald's). One family has to do with filling slots in frames; this is McDonald's method of selecting interpretations for what I called creative compounds (above). The "Slot Verification Heuristic" checks how well fillers match prototype-like expectations that are attached to slots. Also, selectional restrictions are enforced on the head noun, and there are consistency checks against mutliple fillings the same slot. A rough mechanism is suggested to give preference to interpretations matching the contextual topic, but McDonald acknowledges that determining the topic is difficult. The other family of heuristics deals with finding previous knowledge of conventional (lexicalized) nominal compounds and "interpretations" (really schemas or semantic patterns for identificative compounds). The previous knowledge can be in the form of either categories (types) or instances (tokens); in the latter case the interpretation with the most instances stored in the knowledge base is preferred. For compounds involving three or more nouns, preference is given to the interpretation with the most lexicalized sub-compounds or sub-compounds matching stored instances.

McDonald's heuristics derive from many of the same qualitative intuitions addressed by the proposed model. However, they are implemented by heuristics that can run into trouble because they are too coarse. For example, the "Embedded Instances of an Interpretation Heuristic" selects the interpretation involving the most schemas with instances stored in the knowledge base. However, the way the heuristic is formulated makes it insensitive to the relative strength of schemas, which is related to their frequency of occurrence. The proposed model uses probability theory to integrate such factors cleanly.

McDonald performed a hand-simulation on 625 compounds found in newspapers and journals; implementation would have required building a complex real-world knowledge base. He argues that roughly 60% of the compounds would be processed correctly by the model, and that 30% more would be processed somewhat correctly. However, it is difficult to check these results because the coarseness of the heuristics would make them highly sensitive to the exact structure of the knowledge base.

Wermter. An interesting connectionist model is proposed by Wermter (1989b; Wermter & Lehnert 1989), in which the semantic relations between constituents are learned from a training corpus. Seven relations can be predicated between the nouns: BY, FOR, FROM, IN, OF, ON, and WITH. Figure 2.1 shows a small-scale version of the back-propagation network used.[5] Each node in the top layer corresponds to an individual input noun; the nodes are duplicated for the modifier and head noun. The weights from the top layer to the microfeatures layer are hardwired to encode a set of features characterizing each noun. Sixteen features are used to encode a noun, using the NASA thesaurus (NASA 1985). Each node at the bottom corresponds to one of the seven relations; the output activation strength indicates how likely the network judges that semantic relation to hold between the input nouns. In the training mode, an error signal is back-propagated through the network from the node corresponding to the desired semantic relation for the input nouns. In the testing mode, the network is deemed to produce the correct output when the desired semantic relation has the maximum activation. Each of 108 compounds from the NPL (National Physics Laboratory) corpus of abstracts and physical sciences (Sparck-Jones & van Rijsbergen 1976) was judged with respect to each of the semantic relations. The network was trained on 88 of the

[5]For exposition, I have added a layer of nodes at the top of Wermter's net; Wermter actually used the microfeatures layer directly for input, so nouns were translated into feature clusters by a mechanism outside the network.

compounds with every semantic relation deemed plausible for the compounds. Wermter reports 93–98% accuracy on plausibility judgements within the training set. The network was then tested on the remaining 20 compounds. More than one interpretation was permitted; each output node with activation over 0.5 was considered a potential semantic relation for the compound. For the test data, Wermter reports 73–95% accuracy.

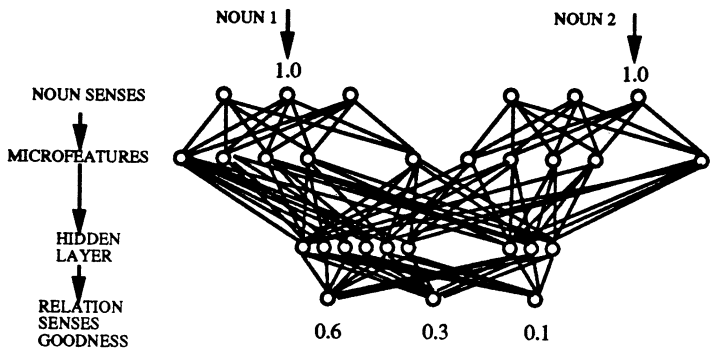


Figure 2.1: Wermter's connectionist nominal compound interpreter (see text).

Wermter's model is intended primarily to demonstrate the feasibility and qualitatively desirable behavior of connectionist modelling. The semantic relations are not as detailed as in some of the linguistic theories, and do not handle compositional structures or nested compounds. Also, the evaluation of success would be more convincing if the criteria were to predict a single most plausible interpretation, rather than a set of interpretations.

Hobbs. The *weighted abduction* framework (Hobbs *et al.* 1988; Hobbs 1990) has been applied to nominal compound examples. In this approach the part of the interpretation task dealing with nominal compounds is to prove an expression like:[6]

$$(\exists o, a, nn)\textit{lube-oil}(o)^{\$5} \wedge \textit{alarm}(a)^{\$5} \wedge nn(o, a)^{\$20} \tag{2.1}$$

The superscripts represent *assumability costs* associated with the various terms. In weighted abduction, theorem proving is augmented by a mechanism that allows unproven terms to be assumed, for a price. The object is to prove the least-cost expression for the input utterance, generating a parse and semantic interpretation as a by-product. Only examples involving simple semantic relations of the Levi sort have been implemented; for instance, one of the kinds of noun-noun relations is FOR:

$$(\forall x, y)\textit{for}(y, x)^{\$a} \supset nn(x, y)$$

If $\$a$, the cost for assuming $\textit{for}(y, x)$, is lower than any other way of proving equation (2.1) then it will be taken.

Realistically speaking, more complexity would be needed than in the above example to distinguish FOR from all the other relations that might hold. To make the selection of the semantic

[6] Hobbs *et al.* subsequently drop the use of *nn* as a predicate variable in favor of treating it as a simpler first-order predicate constant.

relation sensitive to the types of the constituent nouns requires the addition of axioms of the form

$$(\forall x, y)\mathit{function}(x, e) \wedge \mathit{involve}(e, y) \wedge \mathit{extra}(x, y)^{\$b} \supset \mathit{for}(y, x)$$

This says "if the function of x is some eventuality involving y, then FOR may be assumed at a cost $\$b$, which presumably is less than $\$a$". Thus contextual conditions may change the preferred semantic relation.[7]

Weighted abduction has not been applied to nominal compounds in depth, but the framework is intriguing. One version of weighted abduction has been shown to have a probabilistic semantics by Charniak & Shimony (1990). Given this, it will be interesting to see whether probabilistic and weighted abduction models encounter the same difficulties, particularly in how they deal with computational resource bounds.

2.2 Probabilistic Language Models

This section's aims are twofold:

1. To introduce the critical philosophical distinctions concerning how probabilistic models are interpreted.
2. To clarify the theoretical positions of some previous probabilistic approaches to language, by pointing out where they fit into the survey.

Work on applying probability theory to natural language study is experiencing a resurgence after several decades of near non-existence. However, the nature of the relationship between probability theory and language is still under debate. Many different uses of probability are possible; probability theory by itself is merely a calculus built on a few axioms, a metaphor in the hands of the investigator just like any other model. It is up to the investigator to specify how the distribution and event space map onto aspects of the real world. In other words the semantics of a model must be specified, just as in logical, non-probabilistic models. Entirely different claims can be made depending on the model's semantics.

The reader may skip this section or return to it later. However, this section clarifies various foundational points about probability theory that are all too often misunderstood. Such misconceptions potentially lead to unfounded methodological and philosophical objections.

2.2.1 Interpretations of Probability Theory

Probability theory and statistics are intricately related and this is one source of confusion over the nature of claims made by any given model. There are probabilistic versions of the descriptive, predictive, and situated language models discussed in section 2.1.1. With respect to these different modelling goals, probability can be used in various ways. Table 2.2 charts some of the main distinguishing ways of applying probability theory. Each row describes one possible use of probability theory, classified along a number of dimensions discussed in the following paragraphs.

[7] Personal communication with Jerry Hobbs transmitted and elaborated by Robert Wilensky.

Type of probability:	Investigator ascribes probabilistic beliefs to:					
	Nobody/ Agent	Descriptive/ Inductive	Statistical/ Propositional (Propositional => Inductive)	Societal/ Individual	Generated/ Enumerated (Enumerated => Descriptive)	Finite/ Infinite (Infinite => Generated)
Objective [Phys–Obj–RF]		Descriptive	Statistical			
Objective [Propensity]		Descriptive	Propositional			
Subjective	Agent	Descriptive	Statistical			
Subjective	Agent	Inductive	Statistical			
Subjective [Auto. Inf.]	Agent	Inductive	Propositional	Individual	Generated	Finite

Table 2.2: Applications of probability theory.

Subjective versus objective probabilities. The debate over what probabilities really are has a long history. Only a common set of mathematical axioms unifies the many different interpretations of probability.[8] The three major schools are *logical* probabilities, *objective* probabilities, and *subjective* probabilities.

The logical interpretation of probabilities is the most neutral. Probabilities are treated as purely mathematical relations and probability theory is simply a type of logic. Probabilities are attached to propositions, sentences, logical events or atoms; no predication of physical entities is assumed within the logic. Any intended correspondence to empirical physical observations must be explicitly and a priori defined, by giving a semantics that maps the logical sentences to the world. Much of the groundwork in logical probabilities was done by Carnap (1952, 1962).

Objective probabilities are postulated to be physical properties existing in the real world. The simplest case of objective probabilities are relative frequencies, denoted "Phys-Obj-RF" in Tables 2.2–2.4. Another case of objective probabilities are propensities (see footnote on page 22).

Subjective probabilities denote degrees of belief. They are posited to be quantitative measures of certainty in a human (or computational agent). Subjective probabilities are frequently criticized by scientific researchers as being arbitrary. The objection to subjective probabilities is not that they should be avoided for cognitive modelling, but that one should specify how an agent derives its subjective probabilities from experience in its environment. The use of probability in the proposed model is primarily logical probability, but a subjective interpretation can also be ascribed to it. However, we will see that the probability distribution is estimated from observations of certain relative frequencies.

Descriptive versus inductive statistical models. In some sense the most basic type of statistical model, *descriptive* statistical models are simply used to summarize data too numerous be explicitly stored.

[8] Actually several equivalent axiomatizations can be used.

Insofar as a descriptive statistical model employs the probability calculus, those probability values are interpreted in the objectivist sense, that is, the relative frequencies describe the distribution of a set of real-world instances. For example, word counts and category counts (either of pre-tagged corpora or using a real-time parser) are one way to summarize certain surface aspects of large amounts of text. A more subtle kind of application is to count co-occurrences of word pairs or n-grams (Smadja 1991b) or fixed-window word associations (Church & Hanks 1989).[9] In addition descriptive statistical models can be used to capture information about processing times. Example applications are recognition times, semantic retrieval times, or voice-onset times (Macken & Barton 1980).

Inductive statistical models attempt not only to capture and summarize previously seen data but also to predict future data. Statisticians also refer to this mode as *statistical inference*. The most straightforward form of statistical inference is to assume that the relative frequency distribution of a finite sample extends to the rest of the world. For example, in a series of studies concerning the dialectology of New York English, Labov (1966, 1972) uses a sample of 264 employees in Lower East Side branches of Saks, Macy's, and S. Klein department stores to support the conclusion that prevalence of consonantal [r] in postvocalic position is correlated with social status. The extension is clearly only valid if there is no correlation between the sample and the feature being assessed; all the various statistical sampling techniques are means of factoring out, or correcting for, any known feature that could possibly be correlated with the sample.[10]

Hypothesis testing is a branch of inductive statistics that facilitates more powerful checks when the investigator has hypothesized some general rule. In essence these methods check the rule's degree of consistency against a probabilistic model of some other (preferably large) fragment of the investigator's knowledge or assumptions. Though hypothesis testing is prevalent in most scientific disciplines, its use in language studies has been largely confined to external, easily observable phenomena such as phonology (Fasold 1972). In fact Fasold (1972) observes that even Labov's work did not employ hypothesis testing, and Davis (1990, p. 41-5) shows that several of Labov's distinctions are not adequately (i.e., with greater than probability 0.9 or 0.95) supported by the data.

Of greater relevance to computational linguistics are inductive statistical methods in algorithmic form. Automation permits large amounts of electronically stored data to be processed. Among the techniques frequently employed are methods for inducing Hidden Markov Models (HMMs) and probabilistic context-free grammars (PCFGs). HMMs and PCFGs are extensions of regular and context-free grammars, respectively, where the transition or rewrite rules are augmented by values specifying the probability of that rule being applied, as opposed to any alternative rule that could be applied at that point in the expansion (Fu 1974). The Forward-Backward (Baum 1970) algorithm is a standard iterative method in speech recognition that converges on the probability distribution for a Markov model, i.e., when one assumes that the input strings are being generated by a regular grammar. The Inside-Outside algorithm is an extension of the Forward-Backward algorithm for estimating the probability distribution when input strings are assumed to be generated by a context-free grammar. The method has been applied by Fujisaki *et al.* (1991) to learn a context-free grammar from a corpus of about 30,000 Japanese noun compounds taken from

[9] Though these models go on to use the statistics *inductively*.

[10] If however there is a skew but none of the correlated features are known, no sampling technique will help and the only solution is to discover a better model which includes a correlated feature. The ramifications of this are discussed more in section 2.2.2.

machine translation abstracts. Based on a classification of the individual nouns into 46 categories, the model learned to bracket (i.e., assign a tree structure to) nested compounds with 75% accuracy on a test set of 153 compounds. For both Forward-Backward and Inside-Outside algorithms one must provide a priori constraints on the number of rules (or terminals and non-terminals), though methods for dynamically adjusting these constraints have been experimented with (Lari & Young 1990).

When the features fed into an inductive statistical model are the same features a human language user in ordinary life perceives or otherwise has access to, then inductive statistics can be used to model language acquisition. Probabilistic grammar learning approaches of the sort described in the preceeding paragraph are sometimes regarded as models of syntax acquisition.

Statistical versus propositional probabilities. What Bacchus (1990) calls statistical and propositional probabilities have in the past been termed "definite" and "indefinite" probabilities (Pollock 1990; Jackson & Pargetter 1973). Statistical probabilities are probabilities of categories of events (types) whereas propositional probabilities are probabilities of individual events (tokens). A statistical probability is thought of as a relative frequency in a set, regardless of whether the entire set has been sampled in which case the statistical probability is descriptive, or only part of the set has been sampled in which case the estimated statistical probability is inductive. Statistical probabilities can have either objective or subjective interpretations.

A propositional probability predicates an individual event. Since individual events are not repeated one cannot count relative frequencies. Instead, propositional probabilities denote a degree of certainty and are therefore subjective probabilities.[11] Nonetheless the same probability calculus extends consistently to propositional probabilities, and this is the basis of decision theory which is discussed below in section 3.2. It is possible to think of individual events as if they *were* repeatable by imagining a large set of alternate possible worlds and asking how many of the possible worlds the individual event in question occurs in. From this perspective worlds are generated by stochastic application of the same set of non-deterministic probabilistic rules. Note that the only useful form of propositional probabilities is inductive since there is little reason to attribute any probability other than 0 or 1 to an event that has already been sampled.

Other distinctions. Probabilities can be used to describe variation among the members of a society, or they can be used to describe variation in the performance of an individual person. Labov's (1966, 1972) work, described above, is an example of a sociolinguistic application. On the other hand, recognition time studies are usually a mix where samples are taken both over individuals and groups.

A probability distribution can be used to describe a finite set of instances that is explicitly enumerated. Alternatively, a method can be specified for generating or gathering a set of instances. One way to specify a generation procedure is by giving a mathematical definition, e.g., "the set of all even numbers". Another possibility is to give an empirical procedure like "the set of all possible English utterances". Generated sets may be either finite or infinite.

[11] A notable exception to this view is Popper's (1959a, 1959b, 1983) theory of *propensities* which are postulated to be physical properties of individual objects in the world.

2.2.2 Static and Dynamic Probabilities

Good's (1971, 1977) distinction between *static* and *dynamic* probabilities actually applies at more than one level, because the definition of static probabilities is relative to one's point of view. We first examine the distinction from the standpoint of the investigator, then from the standpoint of an agent being modelled.

Static versus dynamic investigator probabilities. The way one usually speaks of static probabilities is from the "physicalist" point of view of an investigator who assumes that probabilities actually exist as physical entities in the real world. However except for purely descriptive probability distributions, none of the (inductive) probability distributions hypothesized in scientific theories are static probabilities themselves; rather they are empirically motivated guesses as to the approximate nature of real static probabilities. Since our time as investigators is constrained the static probabilities can never be determined with absolute certainty. The hypothesized (inductive) probability distributions are referred to as dynamic probabilities.

Given this physicalist view of probabilities, most dynamic probabilities are not true probabilities. A dynamic probability distribution is one that the investigator estimates after having seen some finite number of examples. Because of this sometimes dynamic probabilities are written as conditional probabilities, which are conditionalized on having seen those examples (e.g., Breese & Fehling 1990):

$$Pr_{\text{dynamic}}(x|y) = Pr_{\text{static}}(x|y, i_1, i_2, \ldots, i_n)$$

However as this formula makes clear, this intuitively sensible notation is in reality only formally correct when the inductive process is itself Bayesian. Though dynamic probabilities adhere to the Bayesian axioms within their own distribution, the inductive methods through which entire dynamic probability distributions are postulated and evolved often are not (though adherence to Bayesian learning methods often results in mathematically elegant theories). To reiterate, the physicalist believes that probabilities exist as properties of the real world and considers those static probabilities, which human investigators under time constraints approximate with dynamic probabilities using some separately specified theory of induction.

Static versus dynamic situated agent probabilities. The notion of dynamic probabilities as being time-limited estimates of a static distribution can also be applied to a subjectivist framework. In this framework probabilistic beliefs are ascribed to agents rather than the investigating observer. Again a distinction is made between static probabilities, which are probabilities in an idealized distribution that an agent "believes" to exist in its environment but is possibly intractable, and dynamic probabilities, which are probabilities in the intermediate distributions the agent computes as its working hypotheses.[12] A probability distribution might be impossible for a situated agent itself to compute explicitly, and yet the agent might *act* as though it were trying to learn this distribution ("Unbounded agent" in the tables). As investigators we can describe the behavior of this agent in terms of this distribution.[13]

[12] In this discussion I have attempted to stay within the neutral view of probability theory as merely a logical theory that can be applied either objectively or subjectively. As discussed in the previous section, a subjectivist would argue that all probabilities are held by agents, and that scientific investigators are simply particular agents.

[13] This approach will be taken in modelling automatic inference.

In contrast the agent might be explicitly computing certain distributions, possibly approximations to the intractable ones ("Bounded agent" in the tables). The statistical computations that are part of a situated model must be performed in real time.[14] Situated statistical models are also inductive statistical models to the extent that predicting regularities helps them function in the environment.

The role of situated statistical computations is most likely quite limited. It is entirely conceivable that agents could function effectively in their environments making few explicit statistical computations. For example, consider (possibly Bayesian) learning of (deterministic) decision trees, where the agent adapts its decision tree so that no statistical computations are needed to handle inputs. The problem with situated statistical models is their expense. Above we discussed inducing probabilistic grammars. Unfortunately for situated models, natural language contains too many potential sources of non-determinism. Beyond a certain complexity, parsing a probabilistic grammar takes too much computation time. We must therefore include more structural assumptions to simplify the necessary computation. I will return to this subject in discussing compilation.

The subjective-objective circularity problem. The circularity that arises in trying to define the relative frequency and subjective versions of probabilities is shown in Table 2.3 using the device of a meta-investigator.

The essence of the loop is: I am an investigator who, in this work's line of research, takes the physicalist viewpoint and "believes" in the existence of real world probabilities. However when I introspect on my belief system I find that this position is a subjective one. This is because I am myself an agent whose resources are bounded by my environment.

In this view investigator probabilities are strictly a subcase of situated agent probabilities. The only differences between theories of investigator probabilities and theories of agent probabilities surface at orthogonal points: (1) the types of inputs that are presented to the investigator or agent being modelled, since the range of criteria used by scientific investigators is far broader than that available to the average child who learns to use language (or the average chess player, etc.), (2) the desired type of output: descriptive or reactive; whether there is a utility or loss function, and if so the nature of the function, and (3) scientific investigation commands far greater time and space resources.

It is important to remember that objective probabilities are not in fact entirely objective. In working with probabilities one must always select a model; in other words one always assumes that certain variables are random variables whose values are drawn perfectly randomly from a probability distribution. So far as we know the real world only contains phenomena that can be approximated, but not exactly described, by this assumption (albeit very closely approximated in many cases). One tries to select a model so as to minimize controversiality but in the end a probabilistic theory can only be as correct as its underlying model is.

[14] Situated statistical models are different from statistical descriptions of nondeterministic situated models, which I will also be considering later. The distinction is discussed in section 2.2.4.

Meta–investigator (philosopher–logician) chooses definition of probability to be:	Investigator (bounded by real–world environment) takes probability to be:	Investigator ascribes probabilistic beliefs to:		
		Unbounded agent/ Bounded agent	Descriptive/ Inductive	Statistical/ Propositional
Objective [Phys–Obj–RF]				
Objective [Propensity]				
Subjective	Objective	(Self)	Descriptive	Statistical
Subjective	Objective	(Self)	Inductive	Statistical
Subjective	Objective	(Self)	Inductive	Propositional
Subjective	Subjective	Unbounded agent	Descriptive	Statistical
Subjective	Subjective	Unbounded agent	Inductive	Statistical
Subjective	Subjective [Auto. Inf.]	Unbounded agent	Inductive	Propositional
Subjective	Subjective	Bounded agent	Descriptive	Statistical
Subjective	Subjective	Bounded agent	Inductive	Statistical
Subjective	Subjective [Situated]	Bounded agent	Inductive	Propositional

Table 2.3: The circularity problem in defining subjective and objective probabilities.

2.2.3 Probabilistic Processing Models

The relation of an actual processing model or algorithm to probability theory may be implicit or explicit. The behavior of an implicitly probabilistic model conforms to some known probability distribution, and yet the algorithm itself need not explicitly compute the probabilities in the distribution. The implicit probability distribution may describe different aspects of processing behavior. As mentioned earlier in section 2.2.1, the distribution can describe processing times; for example, online models of phoneme/letter/word recognition, lexical access, or semantic retrieval, even if not explicitly probabilistic, should exhibit the same delay characteristics as humans. On the other hand the distribution can describe the probability of various responses to the same input; in this case we have a model of a stochastic function, such as a Hopfield net (Hopfield 1982) or Boltzmann machine (Hinton & Sejnowski 1986). Note that these can be viewed as implicit subcases of random Markov fields or probabilistic grammars.

2.2.4 Models Using Multiple Applications of Probability

A model can employ multiple systems of probability. These systems can be entirely independent or they can be interrelated. For example, Bacchus's (1990) logic maintains two independent systems of probabilities, one objective and one subjective. The objective probabilities are used for representing descriptive statistics and the subjective probabilities are used for representing degrees of certainties. Bacchus goes on to propose a specific method for deriving subjective probabilities from objective probabilities, thus interrelating them.

We can have a descriptive probabilistic model of a situated statistical agent. The descriptive probabilistic model is used to describe how a situated agent behaves; the situated agent may (or may not) employ statistical methods in its own computations.

2.2.5 Probabilities in the Automatic Inference Model

The "Auto. Inf." label in Tables 2.2–2.4 indicates the intended interpretation of the proposed model of automatic inference. Probabilities are subjective because, as we see in the next chapter, they represent the likelihood of conceptual structures being *useful* to an agent, which does not necessarily correspond to real-world frequencies. In fact, being intentional, different mental structures may not even map to distinct real-world situations.

The probabilities themselves are not ascribed to the agent although the intent is to model a bounded, situated agent. Probabilities are used in describing the behavior, but in humans this behavior may well emerge from heuristic adaptive mechanisms that do not explicitly compute probabilities. In fact, computing rigorous probabilities in the complex domains considered here is likely to be intractable. The approximate maximum-entropy estimation method introduced in chapter 7 offers one possible solution; as neural network techniques improve in representational expressiveness, they may offer other approximation heuristics. Nonetheless probability theory is one of the most powerful and enlightening metalanguages at the investigator's disposal, and will complement more "black box" distributed adaptive neural-network style accounts that may eventually develop.

Probabilities are inductive because they affect the agent's behavior in novel situations according to the degree of similarity to previously encountered situations. Probabilities are propositional because when they are used, it is to judge the chances of conceptual structures being useful

Meta-investigator (philosopher-logician) chooses definition of probability to be:	Investigator (bounded by real-world environment) takes probability to be:	Investigator ascribes probabilistic beliefs to:					
		Unbounded agent/ Bounded agent	Descriptive/ Inductive	Statistical/ Propositional (Propositional => Inductive)	Societal/ Individual	Generated/ Enumerated (Enumerated => Descriptive)	Finite/ Infinite (Infinite => Generated)
Objective [Phys-Obj-RF]							
Objective [Propensity]							
Subjective	Objective	(Self)	Descriptive	Statistical			
Subjective	Objective	(Self)	Inductive	Statistical			
Subjective	Objective	(Self)	Inductive	Propositional			
Subjective	Subjective	Unbounded agent	Descriptive	Statistical			
Subjective	Subjective	Unbounded agent	Inductive	Statistical			
Subjective	Subjective [Auto. Inf.]	Unbounded agent	Inductive	Propositional	Individual	Generated	Finite
Subjective	Subjective	Bounded agent	Descriptive	Statistical			
Subjective	Subjective	Bounded agent	Inductive	Statistical			
Subjective	Subjective [Situated]	Bounded agent	Inductive	Propositional			

Table 2.4: Summary of the classification of probabilistic models.

in a *particular* novel situation. They are applied to certainty judgements in a single individual agent. We assume the data sets are generated by the agent's environment and that at any given point the agent has only encountered a finite number of instances, though this number is not bounded.

Relation to Goldman and Charniak. Goldman & Charniak (1990a, 1990b; Charniak & Goldman 1988, 1989) propose a story understanding model based on Bayesian belief networks Pearl's (1988). For example, figure 2.2 shows the network built to process hypotheses for interpreting

(2.1) Jack got a rope. He killed himself.

Most of the nodes have fairly clear interpretations. An unusual node is the one marked *(patient g3)=r2*, which represents the hypothesis that *r2* fills the *patient* role of *g3*. This is equivalent to a *variable binding hypothesis* because it posits binding the constant *r2* to the unnamed patient variable associated with *g3*.

Variable binding hypotheses are problematic on a larger scale, because they interact. For example, if *(patient g3)* is bound to *r2*, then it can no longer be bound to any other rope *r4* (not shown in the example). This means the conditional probabilities of *(patient g3)=r2* and *(patient g3)=r4* are not independent. In a belief net, any nodes whose probabilities are dependent must either be directly connected or be connected by intermediate nodes whose instantiation makes them conditionally independent. Thus in the general case all the binding hypotheses in a belief net should, properly speaking, be heavily interconnected. This causes loops in the belief net, however, which are particularly expensive to evaluate. As a consequence Goldman and Charniak use as few

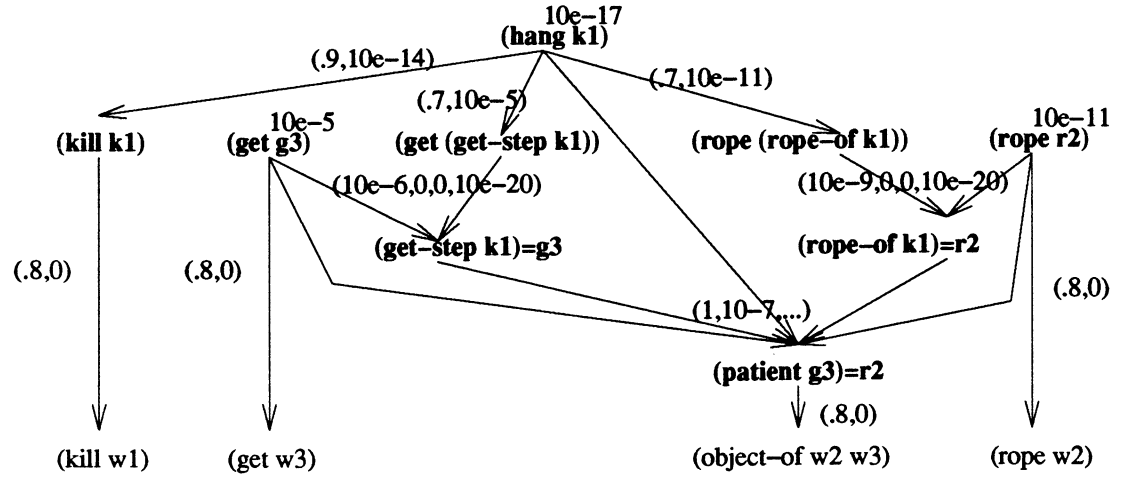


Figure 2.2: Example belief net from Goldman & Charniak (1991). Nodes at the bottom are lexical input events; the others are conceptual hypotheses. Given conditional probability matrices are shown in parentheses; probabilities on the hypothesis nodes are computed.

variable binding hypotheses as possible, and in constructing the belief nets, rule out *a priori* most potential interactions. A universal parameter called the "knob" conditions the probability of all binding hypotheses (they intend this as a first stab at the problem). As we will see, the method proposed in this work takes an *implicit binding* approach that circumvents the problems with explicit variable binding hypothesis nodes, while allowing alternate bindings to interact probabilistically.

Though Goldman and Charniak's model is probabilistic, it does not really acknowledge the difference between objective real-world relative frequencies or subjective belief measures. Goldman and Charniak say the probabilities are objective relative frequencies, but must then introduce the "knob" parameter to increase the probabilities for variable binding. The knob is set relatively low for real-world probabilities, and high for story understanding to account for text coherence. To gloss this, one thing is more likely to be the same as another if the domain is text rather than the real world. However, the resulting conditional probabilities are then no longer relative frequencies; thus interpreting the probabilities is somewhat problematic. This difficulty does not arise in the proposed model since probabilities are interpreted subjectively, as a measure of frequency of *use*.

Relation to Skousen. Skousen (1989) has proposed a framework for language modelling based on an analogical framework. In this model, all previous instances of input-output pairs are stored in a database. To predict the correct response for a new situation, the "analogical set" of relevant previous instances is selected—tokens rather than types—is selected and one member is randomly selected. The output response from that instance is returned as the answer for the new situation. Skousen explicitly rejects the use of probability values:

> One of the most difficult problems in language description has been non-deterministic or probabilistic language behavior. Rule approaches typically account for such behavior by positing probabilistic rules. Even if we suppose that probabilities exist, there is

> still the very difficult question of how those probabilities are actually learned from the statistics and then used to predict behavior. But in an analogical approach no probabilities are directly postulated; instead, an analogical set of examples is constructed and then one of these examples is randomly selected in order to predict stochastic behavior. Thus it may look as if probabilities are learned, but in fact none are. (p. 9)

Nevertheless, Skousen's approach is essentially probabilistic. Mill (1896) recognized the intrinsic connection between analogical induction and probabilities (see section 3.1.1). Here, storing all previous instances as tokens rather than types has the effect of preserving the relative frequency distribution over the space of possible events. Selecting an "analogical set" conditions the event space on the new situation's input event. The frequency of different output values among the members of the "analogical set" are the conditional distribution. The "analogical set" is selected using a distance metric based on the chosen set of *a priori* representational features. Such methods can be reduced to a prior distribution derived from the data set, and are in the spirit of the similarity-based prior discussed in chapter 8 (though in this work we will be considering feature-structures, which are more general than the feature-vectors considered by Skousen). Moreover, just because a random selection is made from the "analogical set" does not mean that probabilities are emergent. To store all previous tokens is more expensive than storing the probabilities of types, and it is really an alternate way of storing the relative frequency distribution. Thus there is really no fundamental difference between Skousen's model and a probabilistic one.

Also note in comparison that the "probabilities" in Skousen's model are statistical rather than propositional; they model stochastic process variation rather than certainty levels. He applies his framework primarily to sociolinguistic data, that is, to variation over societal members rather than individuals.

On human probabilistic judgements. Kahneman *et al.*'s (1982) results demonstrate that humans are poor at making probabilistic and utilitarian judgements. This does not conflict with probabilistic modeling of automatic inference, because Tversky and Kahneman's experiments dealt with controlled inference where conscious, deliberative judgements of probability and utility were sought. Even if automatic inference processes are consistent with some form of bounded probabilistic inference, as I suggest, humans have no direct introspective access to the corresponding probabilities, nor in fact are the probabilities necessarily stored anywhere in explicit form. Any correspondence to probability theory may well be purely epiphenomenal and emergent.

Chapter 3

Chapter 3

Utility, Inference, and Language

In this chapter I construct a Bayesian utility optimization framework for linguistic agents, and situate the automatic inference model therein. The tools of probability and decision theory are used to formulate the situated agent's use of language, interpretation and generation. Decision theory is an extension of probability theory that prescribes the optimal *action* in any given situation, by weighting the utilities of the various possible outcomes by the probability of their occurrence. It forms the basis for analyzing how situated agents function and thrive in their environments. Moreover, as we see, inference and interpretation are themselves computational actions possessing utilities.

The decision-theoretic framework provides an elegant formalism for modelling the relationship between language and its use. It has been an unfortunate characteristic of language understanding research that elegant and simple theories of interpretation are nearly always subject to counterexamples. One of the prime culprits is the pragmatics of language use, whose influences extend far into semantic and lexical levels of inference. For this reason, even when a module can be studied in isolation in great depth, it remains imperative to embed the function of the module within the context of its use.

3.1 Inference in Language Interpretation

Language understanding models are usually formulated in terms of performing some amount of inference upon the input string or utterance. This inference may be deductive, as is the case with many parsers. A good deal of recent work on interpretation casts the understanding process in terms of abductive inference. The purpose of this section is twofold. First, it surveys abductive language interpretation models and the nature of claims made by such models. Second, I argue that abduction is merely a useful conceptual tool that must be supplemented by a statistically-based inductive theory to solve the critical problems in language understanding.

3.1.1 Inductive Inference

An inductive inference is any non-deductive inference. This strict definition of induction, held by Carnap (1962), is more general than many traditional views such as the commonly held idea that induction must extract a universal conclusion from specific observations (the counterexample

of analogical induction is discussed below). The terms *reductive inference* or simply *generalization* will be used to refer to the induction of universals. Inductive inference has sometimes been called "ampliative", since it amplifies knowledge in drawing conclusions that do not logically follow from the data.[1] Not all inductive inferences are useful; it becomes immediately clear that a minimal condition must be met for an inference to be rational, this being that a "Dutch book" bet cannot be made on the agent's resulting set of beliefs.

Subjectivist probabilists call a bet a "Dutch book" when the agent makes two bets based on odds from two of the agent's belief subsets, and the agent is guaranteed to lose regardless of the outcome. For the agent's beliefs to be coherent, it cannot be the case that a Dutch book can be made. Ramsey (1931) showed that avoiding the possibility of a Dutch book is equivalent to the axioms of the probability calculus (I. J. Good has called this the "Dutch Book Theorem").

Enumerative induction and the role of statistics. *Enumerative induction* is defined as the mode of induction in which increasing the number of observed instances of a class lends additional probabilistic weight to that class. The term is used historically in opposition to "eliminative" or "variative induction" in which the number of varieties of classes is the critical weighting factor (see Cohen 1989). As modern statistical techniques are based on enumerative induction, when I say "statistical induction" I will be concerned only with enumerative induction.

Statistical induction is the focal method of inference in the present work. Statistical induction was largely neglected by both AI and linguistic research in the decades preceding the late 1980's, before the resurgence of interest in neural network models and their applications in language processing. However, from the late 1950's to 1960's, statistical induction machines were heavily studied in AI. As we saw earlier, the application of statistical methods in linguistics has been quite limited.

Analogical induction. Analogy is often treated separately from inductive statistics, a tradition perhaps dating back to Hume (1888) who classified probabilities "arising from analogy" as an altogether distinct category from statistical and propositional probabilities, never suggesting using a numerical measure for analogy. However, Mill (1896) recognized the inherent connection, observing that

> If we discover, for example, an unknown animal or plant, resembling closely some known one in the greater number of the properties we observe in it, but differing in some few, we may reasonably expect to find in the unobserved remainder of its properties a general agreement with those of the former, but also a difference corresponding proportionately to the amount of the observed diversity. (p. 367)

Thus Mill anticipates the idea of primitive features upon which a similarity metric can be defined, thus regulating analogical transfer.

[1] The philosophical literature sometimes contrasts "ampliative induction" with "summative induction", which is an oxymoronic term. "Summative induction" produces a universally quantified rule in the special case where the rule can be verified over the entire event space. It is erroneously viewed as a form of induction because its form (producing a universal rule from a set of individual propositions) syntactically resembles generalization, which is a true ampliative induction. In fact "summative induction" merely restates the given data, adding no information; in other words it is a case of descriptive statistics. If universal verification is possible with no assumptions outside of the data, then the rule follows logically from the data, making the "summative induction" actually a form of deduction.

3.1.2 Abductive Inference

Peirce (1931) defined the term *abduction* in his trichotomy of syllogistic reasoning comprised of deduction, induction, and abduction. He exemplifies his typology as follows:

> Deduction
> Rule.—All the beans from this bag are white.
> Case.—These beans are from this bag.
> *therefore* Result.—These beans are white.
>
> Induction
> Case.—These beans are from this bag.
> Result.—These beans are white.
> *therefore* Rule.—All the beans from this bag are white.
>
> Hypothesis (Abduction)
> Rule.—All the beans from this bag are white.
> Result.—These beans are white.
> *therefore* Case.—These beans are from this bag. (v.2 ¶623)

Abduction, unlike deduction, is a synthetic or ampliative form of inference, because the truth of its premises does not preclude its conclusion from being false. It is an operation that generates an explanatory hypothesis:

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true. (v.5 ¶189)

Thus abduction is often glossed as "inference to the best explanation" following Harman (1965). The question is what "best" is. Peirce himself does not address the nature of the hypothesis generation process. Clearly if the hypotheses are always wrong then abduction is of no use. In fact, hypotheses should be generated so that in the long run one expects hypotheses of high average utility.

Logical versus causal interpretations of abduction. Peirce's typology is a purely logical distinction, between modes of logical inference. I follow Peirce in taking the strictly logical interpretation. Note however that in AI literature, "abduction" is often used to convey more than the purely logical sense, and appeals implicitly or explicitly to the investigator's intuitive sense of causality. Causally interpreted, abduction is a process that produces not only correlative explanations for observed events, but causal explanations. Causality is a philosophical morass that, not being directly relevant to the purposes of this work, I attempt to sidestep.

Abductive models of language interpretation. The primary drawback of pure abductive models is that they cannot choose between multiple competing explanations, as many authors have observed (e.g., Pearl 1990; Goebel 1990). This is a severe limitation for language interpretation purposes, as natural

language nearly always contains ambiguities. Thus all practical abductive language models are hybrids, containing additional mechanisms for explanation selection.

Charniak & McDermott (1985, ch. 10) lay out a basic model of abductive plan recognition, story understanding, and speech act analysis. In this model, abduction is used to explain the actions of the characters in a story, or to infer the motivations that underlie a speaker's speech acts. The story understanding direction is pursued in a probabilistic variant in Goldman & Charniak (1990a, 1990b; Charniak & Goldman 1988, 1989) using Pearl's (1988) belief networks. Other models along these lines include the coherency-based story understanding model of Ng & Mooney (1990), Hinkelman's (1990) speech act recognition model, and the question-driven story understanding model of Ram (1990; Ram & Leake 1991). None of these models cast semantic interpretation or parsing in terms of abduction.

The most ambitious application of abduction to language interpretation to date has been the *weighted abduction model* (Hobbs *et al.* 1988; Hobbs 1990; Stickel 1990; Appelt & Pollack 1990). In this model, both semantic interpretation and pragmatics are elegantly integrated into the same abductive framework, which is a theorem prover supplemented with a cost-minimization mechanism. Parsing remains deductive in the Prolog style, but parse rules are augmented with semantic constraints. A cost is associated with each constraint. In interpreting a sentence, semantic antecedents of parse rules can be assumed if they are not known to be false, but the associated cost is charged. The goal is to minimize the total cost of the proof; intuitively, we want to interpret the sentence making a minimum of extra assumptions.

The primary advantage of the weighted abduction framework is that the cost mechanism allows finer-grained coherency judgements than simpler abductive models that use counts of overlapping concepts or other *ad hoc* coherence metrics. However, it remains to be shown that the weighted abduction framework provides tractability gains over probabilistic models, or even that they are expressively different. Charniak & Shimony (1990) have given probabilistic semantics for a very similar cost-based abduction model. The relation between abduction and probability is further considered below.

Induction and abduction. According to Carnap's view that any non-deductive inference is inductive, it follows trivially that abductive inference is inductive. Still, it is instructive to consider the relationship between specific forms of inductive inference.

Firstly, the relationship of abduction to *deduction* is presented in an interesting way by Josephson (1990), who suggests considering abduction as the limiting case of the deductive disjunctive syllogism

$$P \vee Q \vee R \vee S \vee \cdots .$$
$$\text{But} \neg Q, \neg R, \neg S, \neg \cdots .$$
$$\textit{therefore } P.$$

If in fact we assert that all alternative explanations have been ruled out, then we can deduce P. Abduction, then, is the weaker case where we accept that most alternative explanations are unlikely, and heuristically conclude P.

Now let us consider induction. Arguing in favor of abductive models, Harman (1965) argues against giving special status to enumerative induction:

> If we think of our knowledge as based on enumerative induction (and we forget that induction is a special case of the inference to the best explanation), then we will think that inference is solely a matter of finding correlations which we may project into the future, and we will be at a loss to explain the relevance of the intermediate lemmas.

We may re-interpret Harman's argument as a legitimate call to pay close attention to correctly structuring the event space of a probabilistic model. For an AI model this translates to finding the correct conceptual primitives (to which chapters 4 and 5 are devoted). While probabilists have often in the past been guilty of ignoring this issue, it is not an inherent weakness in the probabilistic framework. Moreover the issue of finding the most appropriate conceptual primitives—i.e., determining what in machine learning work is called the *inductive bias* or, more specifically, the *language bias*—applies equally to abductive and inductive models.

Formally, analogical induction (or any non-reductive inference) can always be rewritten as a pairing of a generalization (reductive induction) plus an abduction. In the generalization step we create an abstraction of the source domain that also includes the target domain; subsequently we apply the generalized schema to the target domain, a case of abduction. This goes back to the connection made in the previous section between analogy and induction. The generalization in the intermediate step represents shared features of the source and target domains. (Note that to reformulate a sequence of analogical inductions, it may be necessary to discard the previous generalization before a new instance is handled, if there is any undesired interaction between the genera.)

Any abductive model needs to start from a belief set containing universals or generalizations, which can then be used to explain new observations. Abduction-based theories of language interpretation are fine, but unless some explanation is given for the source of these universals—be that an adaptive, genetically innate, or other source—the theories are ungrounded. In most abductive models the universals assumed clearly derive to a large extent from the agent's experience, so to obtain the universals one needs a non-abductive theory of reductive inference. The best and most neutral theories of reductive inference to date are still the enumerative induction models based on probability theory. Thus, while abduction is a useful conceptual tool in constructing language interpretation models, it does not permit us to avoid the issues with which statistical induction is concerned.

3.2 Utility of Inferential Actions

This section presents a view of inference as an action that has some beneficial or detrimental effect for the agent. The language agent makes inferences in such a way as to optimize its overall effectiveness within the environment, given the constraints on the agent's architecture. Bayesian utility theory is employed as the framework.

3.2.1 Utility and Decision Theory

Bayesian utility and decision theory is only beginning to be applied to language. The most natural application, and the only one I know of so far, is to model the stimulus-response behavior of an individual agent in terms of trying to maximize subjective expected utility. Within this approach there is the same range of modelling options as described earlier for probability

theory alone; one may model stimulus-response behavior from a behaviorist, interpretive, or generative standpoint. Due to the complexity of language use, purely behaviorist approaches are impractical except for the grossest level of detail.

Utility theory prescribes a normative method for choosing a rational course of action. It is a straightforward extension of probability theory where the possible outcomes of each action A_i in a world state W_k are assigned some numeric utility measure $U([A_i, W_k])$. (The square brackets denote "the new world state after action A_i is taken in world state W_k.) The utility of outcomes are weighted by the probability of their occurring. Thus the rational action is the one that maximizes the expected utility

$$E[U([A_i])|\mathbf{e}] \stackrel{\text{def}}{=} \sum_k P(W_k|\mathbf{e}) \cdot U([A_i, W_k])$$

given the constraints that the known evidence **e** already tells us about the state of the world. For more in depth introductions to utility and decision theory see Berger (1985), von Winterfeldt & Edwards (1986), and Luce & Raiffa (1957); the seminal work is von Neumann & Morgenstern (1944).

One might question using a theory of "rationality" to model human language processing. Humans after all do not act rationally and are poor at estimating utilities (Kahneman *et al.* 1982). However, to view utility theory as a direct model of conscious human decision making would be oversimplistic; this is why I prefer the term "utility theory". Utility theory is applicable to cognitive modelling insofar as it provides a simple, analyzable mathematical framework for adaptive agents. In this work utility theory is used as a framework for unifying the contribution of diverse cognitive processes toward helping a language-using agent to function and adapt in its environment. Utility theory has the advantage of being a very simple mathematical model in the sense that all assumptions are made explicit in a small set of axioms. This ensures internal consistency in the model; there are no "hidden surprises" in the processing mechanism.

Moreover, the proposed utility model is resource-bounded, which removes much of the "idealistic" rationality. The power of the bounded model is that one can build in all sorts of resource constraints that force the model to behave in certain ways, for example by restricting decision processes to a coarse granularity to match cognitive limitations. From another, less constrained viewpoint, the model would not be behaving "rationally". (Counterarguments are invalid if they are based on the scientific investigator's (relatively unbounded) evaluation of a human's (highly bounded) actions.) Except for abstract analysis I will only be considering highly constrained models.

3.2.2 Russell and Wefald's Bounded Rationality Framework

In its pure form decision theory ignores the fact that the real world always imposes bounds on the computation time and memory space resources that a situated agent requires to evaluate its expected utilities. In other words the computations to implement a decision-theoretic mechanism themselves constitute actions with a cost. Resource-bounded agents cannot perform an unlimited amount of computation before deciding what further course of action to take.

In this work I treat inference as a commitment to some hypothesis without absolute validation, for reasons of resource limitations. (The commitment to a hypothesis is always retractable though not necessarily without substantial computational cost.) Over the years a number of other theories concerning normative conditions for hypothesis acceptance have been advanced (e.g.,

Kyburg 1961, 1983; Pollock 1990). Though the resource-bounded view of acceptance is not fleshed out with the rigor and detail of those theories, it deals nicely with the value of information and computation using an elegant Bayesian framework.

From this perspective there are two subkinds of inference: (1) The hypothesis may be one that follows deductively from the agent's premises, though the agent is unable to verify deductive consistency within time constraints. (2) The hypothesis may not follow deductively, and may even be inconsistent with other premises held by the agent. In neither case can the agent know with certainty which type of inference is being made, i.e., whether the hypothesis is deductively consistent with all other premises.

To analyze inference I employ Russell & Wefald's (1988, 1989, 1991; Wefald & Russell 1989b, 1989a; Russell 1990, 1991) metalevel framework for resource-bounded computation. As in straight utility theory, define

- $U(W_k)$: a utility measure over the possible states of the world. Typically the utility depends only on aspects of the world that are external to the agent, so we might instead write $U(E_k)$.
- $[X, W_k]$: the world state resulting from taking an action X in world state W_k. Where no ambiguity can result $[X]$ may be used as an abbreviation for taking the action in the current state.

Actions, however, are now divided into

- A_i: an external action that affects the surrounding environment, and
- S_j: an computational action that changes only the agent's internal cognitive state.

In addition define the following (for clarity reasons my notation differs very slightly from that of Russell and Wefald):

- e: the body of evidence available to date, from previous computations and percepts.
- e_S: the new evidence made available by computation S.
- α_S: the external action that the agent estimates to have the highest utility after computation S. α alone denotes the action currently deemed best.
- $\hat{Q}^{\mathbf{e}}$: an *estimate* made by the agent of the quantity Q given the available evidence. Typically, the exact computation of the quantity Q would be intractable.

Computational actions have the effect of revising α, the external action currently deemed optimal. The *net value* of a computational action S_j is defined as

$$V(S_j) \stackrel{\text{def}}{=} U([S_j]) - U([\alpha]) \tag{3.1}$$

However, the analysis of $U([S_j])$ is quite complex when the utility measure is defined only over external aspects of the world (since there can be multiple values of j for which $[S_j] \in E_k$, but where the values of $U([S_j])$ differ nonetheless because they enable divergent future computation chains). Thus we generally employ a simplifying *single-step* assumption where the value of $U([S_j])$

is approximated by $U([\alpha_{S_j}, [S_j]])$, the utility if the "best" external action were taken immediately after computation S_j:

$$V(S_j) \stackrel{\text{def}}{=} U([\alpha_{S_j}, [S_j]]) - U([\alpha]) \tag{3.2}$$

Ideally, the agent would follow the algorithm:

1. Compute $E[V(S_j)|\mathbf{e} \wedge e_{S_j}]$, the expected value of all possible next computational actions S_j given the previous evidence $\mathbf{e}$ plus the additional evidence e_{S_j} made available by the computation.
2. If any S_j has positive net value, perform the S_j with highest expected value and go to step 1.
3. Execute the current α.

The problem is of course that usually step 1 itself requires significant computation; it is potentially impossible to compute without actually performing the actions, and perhaps even intractable. Thus instead of assuming the expected values can be computed exactly, we assume that estimates of $V(S_j)$ are made by the metalevel:

$$\hat{V}^{\mathbf{e}}(S_j) \stackrel{\text{def}}{=} \operatorname*{est}_{\mathbf{e}} E[V(S_j)|\mathbf{e} \wedge e_{S_j}] \tag{3.3}$$

given only the available evidence to date $\mathbf{e}$. The algorithm can then be modified using $V^{\mathbf{e}}(S_j)$ instead, and will approximate the ideal to the extent that the estimation function is accurate.

Often an additional simplification is made, by assuming that the dependence of utility on the time it takes to perform a computation S_j is captured by a cost function C that is independent from the external actions being considered. In this case the normative value of computation is

$$\begin{aligned} V(S_j) &\stackrel{\text{def}}{=} U([\alpha_{S_j}, [S_j]]) - U([\alpha]) \\ &= U([\alpha_{S_j}]) - C(S_j) - U([\alpha]) \\ &= U([\alpha_{S_j}]) - U([\alpha]) - TC(|S_j|) \end{aligned} \tag{3.4}$$

where $TC(|S_j|)$ in the last expression denotes the *time cost* associated with the duration $|S_j|$ it takes to perform S_j. This can be done as S_j only affects the agent's internal state. Correspondingly for the estimated value of computation,

$$\hat{V}^{\mathbf{e}}(S_j) = \operatorname*{est}_{\mathbf{e}} E[U([\alpha_{S_j}]) - U([\alpha])|\mathbf{e} \wedge e_{S_j}] - TC(|S_j|) \tag{3.5}$$

3.3 Language Interpretation as Rational Forward Inference

There are at least two levels at which it is interesting to consider how the utility of forward inference relates to language interpretation: (1) all language interpretation is forward inference, and the mechanism responsible for those forward inferences must be sensitive to the past effectiveness of inferences in helping the agent function in its environment; (2) automatic inference is a type of forward inference within the language-interpretation-forward-inference scheme, whose power

is heavily constrained by architectural (presumably biological) limitations. The latter approach is discussed in subsequent sections.

The former approach has been applied in a dialog model with both interpretive and generative components (Wu 1991a; Wu &.Horster 1989). The use of a decision-theoretic model of inference allowed the model to account for more sophisticated, realistic, and efficient interchanges than in previous dialog models. The example application domain is that of a route consultant system advising a user asking for directions. Let us consider a brief example here, in order to make the idea of a situated language agent more concrete. Consider the dialog:

(1) U: How do I get to the center of the bay?
(2) S: Why do you want to go there?
(3) U: I want to take a picture of the skyline.
(4) S: Is it sufficient to drive to Treasure Island, or is it necessary to take a cruise?
(5) U: No, a cruise isn't necessary.
(6) S: Then you should drive to the Bay Bridge and take the Treasure Island exit.
(7) U: What about Angel Island?

This last query is difficult to analyze, because the plan recognizer is not able to produce a plan that would explain the user's speech act. The system agent at this point either continues by generating an *active acquisition goal* to identify the unknown user goal:

(8a) S: Why do you ask?
(9a) U: I also want to visit Angel Island.

or it generates a goal to identify and/or correct a suspected user misconception:

(8b) S: There is no bridge to Angel Island, you must take a ferry.

The decision-theoretic framework allows the agent to trade off the utility of acquiring further information from the user against the additional conversational burden that would be required to do so. The paradigm is proposed in Feldman & Sproull's (1977) "hungry monkey" example, who used visual ("LOOKAT") rather than linguistic information-acquisition operators in the classic monkey-and-bananas domain. A number of rules for generating active acquisition goals are considered in Wu (1991a), accounting for the other back-and-forth interchanges in the example dialog as well. The absence of such rules in previous dialog systems, which prevented them from asking back appropriate questions, is due to the fact that the active acquisition goals will proliferate rapidly, putting an unnecessarily heavy conversational burden on the user. What is needed is a strong control mechanism that is capable of efficiently pruning the unnecessary goals; this is one application of utility estimation in the model.

The other reason to estimate utilities arises from the need to judge when it is fruitful to continue attempting plan recognition on the user's query. Since the agent cannot possibly search the space of all possible user plans exhaustively, and since the probability of many possible plans is extremely low, it is important that the agent have some means to trade off the utility of further search against the cost of waiting to respond to the user. The estimation of a multi-attribute utility metric is treated in depth in Wu (1991a).

3.4 Compilation and Adaptation

In the above model, utility is estimated for actions at a fairly coarse level. We now consider finer inferential steps, and the role of compilation in resource-bounded agents, in preparation for the treatment of automatic inference.

Russell and Wefald describe compilation as "a method for omitting intermediate computations in the input-output mapping . . . when an entire *class* of computations can be omitted, so that a whole class of decision-making episodes can be speeded up" (p. 41). To classify the ways in which computations can be omitted, four types of "static" knowledge are identified, dealing with (1) conditions predicated of the current world state, (2) conditions predicated of the world state that results from an action, (3) the utility values associated with those states, and (4) the optimal action for the current world state. These are depicted as stages in figure 3.1. The stages are linked by "dynamic" knowledge (the solid edges) of four kinds:

A: $condition(W_k) \Rightarrow condition'(W_k)$

B: $condition(W_k) \Rightarrow condition'([A_i, W_k])$

C: $condition(W_k) \Rightarrow U(W_k) = value$

DT: The decision-theoretic principle takes knowledge of the possible actions' utilities and chooses the one with the highest expectation.

Depending on what regularities the environment exhibits, these uncompiled types of knowledge can sometimes be converted into compiled rules (the dotted edges) that bypass one or more stages without changing the decisions. This results in efficiency gains ranging from marginal to exponential. Russell & Wefald (1991) discuss at greater length the conditions that make a domain amenable to compilation. Of the three types of compiled knowledge only the condition-action rules (the long dotted edge) will concern us.

Russell and Wefald also suggest an architecture for limited rational agents containing one representative from each of the four kinds of execution architecture (uncompiled decision-theoretic; goal-based; action-utility; and condition-action or production systems). All four execution architectures run in parallel. The proposal is preliminary and does not address the difficult issue of how compilation is performed, i.e., how knowledge could be transferred from the uncompiled decision-theoretic execution architecture to one of the others. Compilation is an ongoing area of research, SOAR (Laird *et al.* 1986; Laird *et al.* 1987; Newell 1990) being a notable example of a compilation architecture motivated by cognitive concerns. However, no model has resolved the issue of balancing the compiled operators' speed gain against the number of operators and thus the increased search time (Tambe & Newell 1988; Tambe & Rosenbloom 1989; Minton 1988; Braverman & Russell 1988, 1992). In this study automatic inference is treated as a compilation mechanism that creates a time-bounded condition-action system, whose output inferences need not always be correct but only statistically useful.

3.4.1 Compilation as Adaptation of Computational Action Set

I propose to treat compilation as a process whose effect is to create a new computational action(s). For example, one type of compilation creates condition-action rules. We can view the

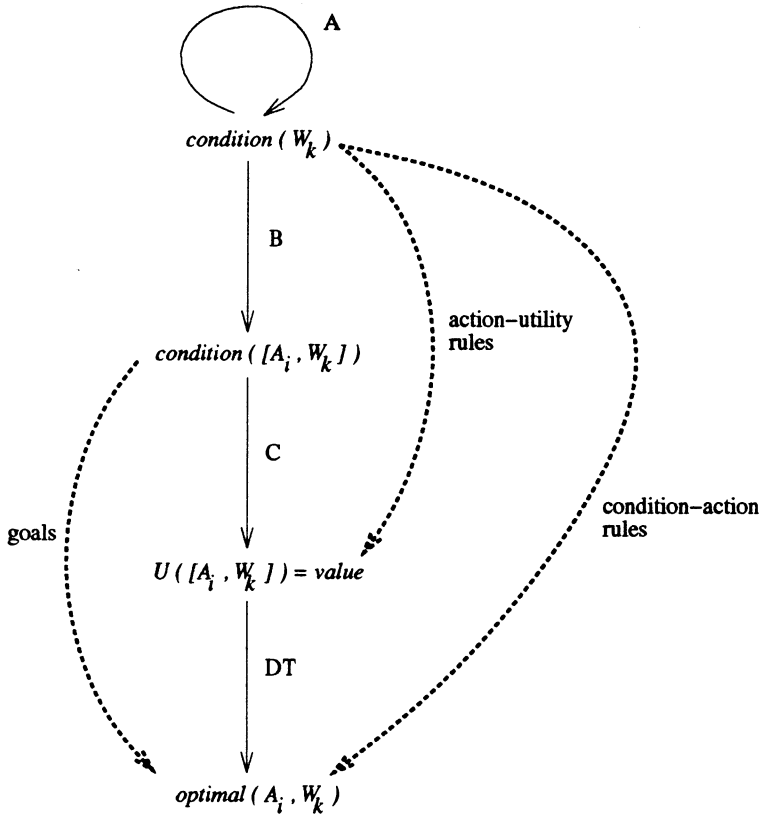


Figure 3.1: Types of "static" knowledge (nodes), uncompiled "dynamic" knowledge (solid edges), and compiled knowledge (dotted edges), based on Russell & Wefald's (1991, p. 45) figure.

creation of a condition-action rule as the creation of a new computational action $S_{j'}$. Even if $S_{j'}$ was created by compiling some sequence of actions including explicit utility estimations, the compiled form is an impenetrable single action. In handling new situations, $S_{j'}$ must be weighed alongside all the old S_j actions.

Note that the condition-matching process for selecting a condition-action rule also has a cost. A compilation method may attempt to reduce this cost as well. For example, a decision tree is sometimes an efficient way to precompile the selection algorithm. In this case the entire decision tree algorithm must itself also be considered a single computational action $S_{j''}$. If, subsequently, a new rule is added into the decision tree, $S_{j''}$ changes and so we think of $S_{j''}$ as an *adaptive* computational action.

In my presentation of compilation I wish to emphasize the fact that there is no clearcut division between execution architectures and computational actions. Russell and Wefald assumed that their condition-action execution architecture contained a built-in, prewired condition-matching process. As pointed out in the preceeding paragraph, this need not be the case. In fact selecting an optimal indexing strategy is equivalent to constructing a new execution architecture on top of the primitive level. The issue of what primitive level is optimal can only be addressed through empirical means. Thus the interaction between compilation and execution architectures is actually quite subtle.

Nonetheless it is pragmatic to assume some small number of primitive execution architectures, as Russell and Wefald have. The issues surrounding the optimality of compilation methods are quite complex and remain unsolved. In general the overarching issue is how an agent can perform compilation incrementally, as it gains experience. It is not yet clear how the set of computational actions can be adapted to maintain optimality, or at least, come close to optimality. Moreover in deciding whether to compile some non-trivial subdomain into, say, a decision tree of rules versus a rule hierarchy, the optimal indexing method will depend heavily on the subdomain's statistical structure. The larger the computational actions, the more complex and infeasible the optimality analysis becomes. If too many adaptive parameters are permitted on computational actions, a resource-bounded agent cannot perform the optimality analysis in real time. It appears that we can only expect an agent to consider a few different execution architectures, possibly prespecialized. Thus for the present, compilation remains mostly a conceptual metaphor with implementations being heavily constrained.

In this work automatic inference is induced by a compilation method that has prewired limitations on the form and complexity of compiled rules. The entire automatic inference facility is treated as a single computational action S_A with an associated utility estimation function. Over time, the results given by S_A adapt, and therefore also the estimated utility. It follows that the automatic inference compilation method is optimal only if its prewired structure fits the environment.

3.4.2 Normative Value of Forward Inference

Assuming that an uncompiled decision-theoretic module is among an agent's battery of execution architectures, let us now consider why forward inference—automatic inference being one instance—is a crucial facility.

In search-theoretic frameworks such as the Russell-Wefald theory, the boundary between forward and backward inference is blurred. One normally thinks of forward inference as some

amount of computation that is automatically performed when input data is acquired, and backward inference as focussed computation driven by a particular goal. This distinction made sense in simple rule-based systems, but in more complex inference models the distinction is entirely relative to one's definition of what constitutes goals. In one sense much of language interpretation is forward inference, including any aspect that is recognition-like in character, as opposed to being driven by a conscious goal to disambiguate or re-interpret the input utterance. (This of course is part of the definition of automatic versus controlled inference.) Yet in a different sense, even the recognition-like forward inference can be considered goal-driven since the only reason to perform forward inference is for the goal of interpreting the input more quickly. The Russell-Wefald framework takes the latter approach and incorporates everything into a teleological metalevel of search, driven by the overall goal of maximizing expected utility. The only place in the Russell-Wefald model where true forward inference occurs is within a single computational action S_j.

This latter view alone, however, is not very illuminating with respect to practical execution architectures. In practical architectures (say, the brain), the need for recognition-like "forward inference" processes arises from the fact that procedures for determining maximum-value computational actions are themselves computational actions with a cost—actions that explicitly compute the formulae described previously, by estimating the expected values of actions, comparing them, and selecting the maximum, using procedures like

1. function EstimateVal $(S_j; \mathbf{e} \wedge e_j; [S]) : \Re$
2. return EstimateUtil $([\alpha_{S_j}, [S_j]]; \mathbf{e} \wedge e_j; [S])$ − EstimateUtil $([\alpha]; \mathbf{e} \wedge e_j; [S])$

A second possibility, on the other extreme, is to use a production system which precompiles the maximum-value selection process into a fast rule such as a decision tree. In this case there is almost no time to perform forward inference; in fact an optimal compilation method will incorporate all possible forward inference into the fast rule itself.

The most likely and practical option, however, is to use some semi-compiled hybrid procedure that is intermediate between the foregoing extremes, but nonetheless employs some amount of explicit decision-theoretic reasoning. Because of this, optimal behavior nearly always requires *some* non-zero amount of forward inference. Intuitively the optimal amount of forward inference is that which is, on the whole, less costly than the average combined cost of estimating utilities, comparing them, and then choosing and executing the maximum-value inference. The only case where no forward inference should be done is when the agent architecture is such that meta-level utility computation is substantially cheaper than the smallest amount of forward inference permitted by a single computational operator.

We can see this more explicitly by "unfolding" the Russell-Wefald model to analyze the metametalevel. Define the following notation:

- $ObjectActs \stackrel{\text{def}}{=} ComputationActs \cup ExternalActs$.
- A_i, as before, denotes a member of $ExternalActs$.
- S_j, as before, denotes a member of $ComputationActs$.
- $MetaActs \stackrel{\text{def}}{=} MaxValComputationActs \cup ObjectActs$, where $MaxValComputationActs$ is the set of metalevel computational actions that determine the maximum-value action from the object level sets $ComputationActs$ and $ExternalActs$.

- $\dot{A}_i$ denotes a member of *ObjectActs*. At the metalevel, both external and "regular" computational actions are considered "external" actions.
- $\dot{S}_j$ denotes a member of *MaxValComputationActs*. At the metalevel, the only "computational" actions are those for choosing among external and regular computational actions.
- $\ddot{A}_i$ denotes a member of *MetaActs*, i.e., any action.
- *ComputationActs* $\subset$ *AtomicComputationActs*^{g} where g denotes the granularity of *ComputationActs*. The idea is that the computational actions whose values are compared by an $\dot{S}_j$ action are generally substantially larger than $\dot{S}_j$ itself. To capture this we conceive of computational actions as molecular entities comprised of smaller atomic computational actions. Accordingly g is set to the size of molecular computations that the metalevel *MaxValComputationActs* can handle.
- s_h denotes a member of *AtomicComputationActs*.

In this case the normative value of a forward inference action F is

$$V(F) = U([\alpha_F, [F]]) - U([\alpha]) \tag{3.6}$$

Optimal forward inference is the process that is sufficiently likely to yield useful results (i.e., of high enough expected utility) that it outweighs the expected utility of explicit computation of meta-level utility judgements. Thus the optimal amount of forward inference can be expressed as the maximal function of the form

$$\check{F} \stackrel{\text{abbrev}}{=} \check{F}(\mathbf{e}) : E \rightarrow \textit{AtomicComputationActs}^g$$

such that for any evidence e given by the situation, it is the case that

$$E[V(F)|\mathbf{e} \wedge e_F] \;>\; E[V(\dot{S}_j)|\mathbf{e} \wedge e_{\dot{S}_j}] \tag{3.7}$$

which means

$$E[U([\alpha_F, [F]]) - U([\alpha])|\mathbf{e} \wedge e_F] \;>\; E[U([\alpha_{\dot{S}_j}, [\dot{S}_j]]) - U([\alpha])|\mathbf{e} \wedge e_{\dot{S}_j}]$$

$$E[U([\alpha_F, [F]])|\mathbf{e} \wedge e_F] \;>\; E[U([\alpha_{\dot{S}_j}, [\dot{S}_j]])|\mathbf{e} \wedge e_{\dot{S}_j}] \tag{3.8}$$

$$E[U([\alpha_F])|\mathbf{e} \wedge e_F] - TC(|F|) \;>\; E[U([\alpha_{\dot{S}_j}])|\mathbf{e} \wedge e_{\dot{S}_j}] - TC(|\dot{S}_j|) \tag{3.9}$$

The last inequality, again, follows from assuming independence between computation cost and the action. It can be rewritten to express the tradeoff as a function of the difference in time costs for F and $\dot{S}_j$:

$$E[U([\alpha_F])|\mathbf{e} \wedge e_F] - E[U([\alpha_{\dot{S}_j}])|\mathbf{e} \wedge e_{\dot{S}_j}] > TC(|F|) - TC(|\dot{S}_j|) \tag{3.10}$$

Thus the slower the agent's facilities for estimating utilities ($\dot{S}_j$) are, the larger the steps that forward inference (F) should make.

3.4.3 Precompiled Compilation Methods

Automatic forward inference can be seen as the effect of condition-action rules that are learned and compiled. To remain strictly within the framework, compilation would itself be treated as a computational action, and the various possible compilations would be weighed against each other. However, utility estimation on such a large scale would be prohibitively expensive to implement in explicit declarative form. It makes more sense to regard the adaptive learning process as one where the entire perception, action, utility optimization, and compilation procedure is itself precompiled. This has the effect of adding strong enough constraints to make compilation tractable. For example, in a neural network employing the Widrow-Hoff learning rule, the Widrow-Hoff rule can be considered a case of a precompiled learning procedure.

Even with strong constraints, humans and other animals are capable of accomplishing quite a lot with precompiled learning and inference mechanisms. This amounts to a claim that the structure of the real world is such that a good deal of useful regularity can be picked out of it using hardwired learning techniques; human language, having presumably evolved to fit the needs of its users, also not surprisingly has a convenient structure. It is unlikely that a completely competent game of chess can be played this way, but in fact the degree of success of neural net game-players such as Tesauro's (1991) backgammon system attests to the importance of a primitive learning mechanism even for such abstract tasks.

The adaptive mechanisms behind human automatic inference presumably evolved under selection pressures, which Russell and Wefald do not address. To explain the evolution of an automatic inference architecture within Russell and Wefald's framework requires extending the compilation paradigm into a teleological view of evolution. This may be stretching the framework to the point of diminishing returns but a brief thought experiment on the interface between rationality and evolution is worthwhile.

In Russell and Wefald's normative framework an individual agent maximizes its expected utility.[2] From the evolutionary perspective, a noisy environment generates "random" agent architectures, and selection pressures cause those agents who perform in a more optimal manner to survive. Some agent architectures in fact turn out to be compiled versions of the decision-theoretic principle as it applies to the structure of the real world, and these survive. Evolution thus functions as a stochastic utility maximization architecture.

For one discussion of the biological constraints within which language facilities evolved see Lieberman (1984). See also Millikan's (1984) discussion of adapted devices and meaning.

3.5 Automatic Inference

According to the automatic inference theory, human cognitive architecture performs a certain amount of forward inference by itself without recourse to explicit decision-theoretic computation. This forward inference nonetheless tends to increase the agent's utility toward optimal. The cognitive architecture, the argument goes, includes a hardwired capability that supports automatic forward inference regardless of other inference capabilities that may be slower, more conscious, more deliberative. The optimal use of the architecture exploits both the automatic

[2] That is, its dynamic expected conditional utility.

and controlled inference capabilities to their limit. We take it that evolutionary pressures are ultimately responsible for this.

As noted in the preceeding chapter, in the model I propose the automatic inference facility is modelled by an engine that attempts to maximize the probability that the inferences drawn are useful. The automatic inference engine can be analyzed from either an *integral* or an autonomous *functional* standpoint. An integral analysis describes how the automatic inference engine helps improve the expected utility of the agent's actions. A functional analysis construes the automatic inference engine as an autonomous agent whose function is to optimize some utility metric of its own. This subagent "resides" in a micro-environment that is actually the rest of the cognitive mechanism.

I will use this distinction to examine automatic inference from two vantage points within the Russell-Wefald approach. In the integral analysis, depicted in figure 3.2(b), what one might ordinarily think of as automatic and controlled inferences are simply construed as being different kinds of computational actions at the object level. At each point in time the choice between alternative actions is mediated by the global utility metric. Denote automatic inference computational actions by S_{A_I}; controlled inference are still denoted by S_j.

However, this view by itself does not sufficiently capture the way in which human cognition appears to be constrained. The large number of alternative automatic inference actions would make prohibitive any implementation based on a decision-theoretic execution architecture that explicitly maximizes utilities to select inferences. Furthermore the level of computational granularity is too fine, in the sense that it is unrealistic to assume that at every step computational resources can be allocated to either automatic or controlled inference. This is a sensible assumption on a conventional computer where each computational step might take on the order of one-millionth of a second, but in the brain each computational step takes on the order of one-hundredth of a second, which is too slow to accommodate much metalevel control.[3] The coarseness of the computational step is a serious constraint in cognitive modelling.

Thus we also examine automatic inference from the autonomous functional point of view, where automatic inference is performed by a subagent. This is depicted in figure 3.2(c). The agent itself only "knows" about the "atomic" computational action S_A which, as explained below, it always deems rational to execute. The task of selecting a particular inference is left to the subagent which chooses its action A_I to maximize its own expected utility. The subagent's decision algorithm *adapts* autonomously of the agent's metalevel. By encapsulating automatic inference within a subagent, we admit the possibility of efficiently compiling its decision algorithm independent of the rest of the agent's computation processes.

I noted earlier that the atomic computation step is the one place where true forward inference occurs in the Russell-Wefald framework. Whereas they assumed that the metalevel could direct forward inference in tiny steps, in human cognition physical and biological constraints make the granularity of forward inference steps so coarse that it becomes important to study the effectiveness and adaptability of single steps. This perhaps is one of the forks at which pursuit of cognitive modelling issues diverges from pursuit of optimal game-playing and problem-solving AI, and because natural language is so closely matched to cognitive processing abilities (presumably, evolved that way) cognitive constraints must also be built into effective language processing

[3] Besides which, as I noted earlier, some researchers (most notably, Kahneman *et al.* 1982) would claim that humans are inefficient or incorrect at utility estimation even at this coarse level.

models.

I will be concentrating on the autonomous functional analysis of automatic inference, as is the norm with language interpretation models. However, in keeping with the philosophy of meaning as use, in this chapter we also consider the integral stance where automatic inference is embedded within a utility-optimizing agent, and examine why automatic inference helps the agent function in its environment.

3.5.1 Automatic and Controlled Processes

We can gain a better idea of the architectural constraints on human inference from empirical studies of online language understanding. There is a substantial body of evidence that forward inference extends deep into the pragmatic level and is not confined to syntactic processing or Logical Form (Potts *et al.* 1988; Till *et al.* 1988; McKoon & Ratcliff 1981, 1986, 1989b, 1989a; O'Brien *et al.* 1988; Anderson & Ortony 1975; Whitney & Kellas 1986; Duffy 1986). However, results at any more specific level have been inconclusive and controversial. For good surveys and discussion of empirical techniques see Keenan *et al.* (1990) and McKoon & Ratcliff (1990).

One important issue in interpreting the results is whether the subject makes the forward inferences at the time of reading, as opposed to making them when probed by the investigator, which would not be forward inference.[4] Keenan *et al.* (1990) argue that the techniques used in most previous studies fail to differentiate these, including cued recall, sentence verification, sentence reading times, recognition, lexical decision, and naming. Not all these methods are inherently unable make the distinction; the weaknesses lie in their particular realizations. Keenan *et al.* also argue against the position McKoon and Ratcliff take in some of their more recent work that the distinction is neither important nor empirically decidable. From the natural language processing standpoint, it is certainly important to have a theory of what forward inferences are made directly upon input.

This methodological problem notwithstanding, the sheer weight of the many studies combined with the results of several more careful studies argue fairly convincingly that *some* significant amount of forward elaborative inference takes place. However, attempts to delimit specific semantic categories in which elaborative inference can take place have been less than successful. One working hypothesis that was accepted for some time was that readers do not elaborate characteristics of instruments, e.g., infer "spoon" from "stir the soup". In fact Lucas *et al.* (1987) showed the actual situation to be far more complex, depending on the context of the task. That the bounds of automatic inference cannot be delineated along crude semantic characteristics is a recurring theme of this work. The fact of failure to identify such bounds is some small evidence against such *a priori* delineations.

A couple of other caveats are in order. First, the distinction many studies make between elaborative inferences (a kind of forward inference) and bridging inferences (a kind of backward inference) is somewhat misleading. It assumes there is such a thing as null context, since bridging inferences are distinguished by virtue of being "drawn in order to establish coherence between the present piece of text and the preceding text" whereas elaborative inferences are "simply drawn to embellish the textual information" (Keenan *et al.* 1990, p. 378). I agree that this is a useful distinction, but it is not necessarily reflected in the sorts of associative mechanisms that presumably play a

[4] At least, not in the same sense.

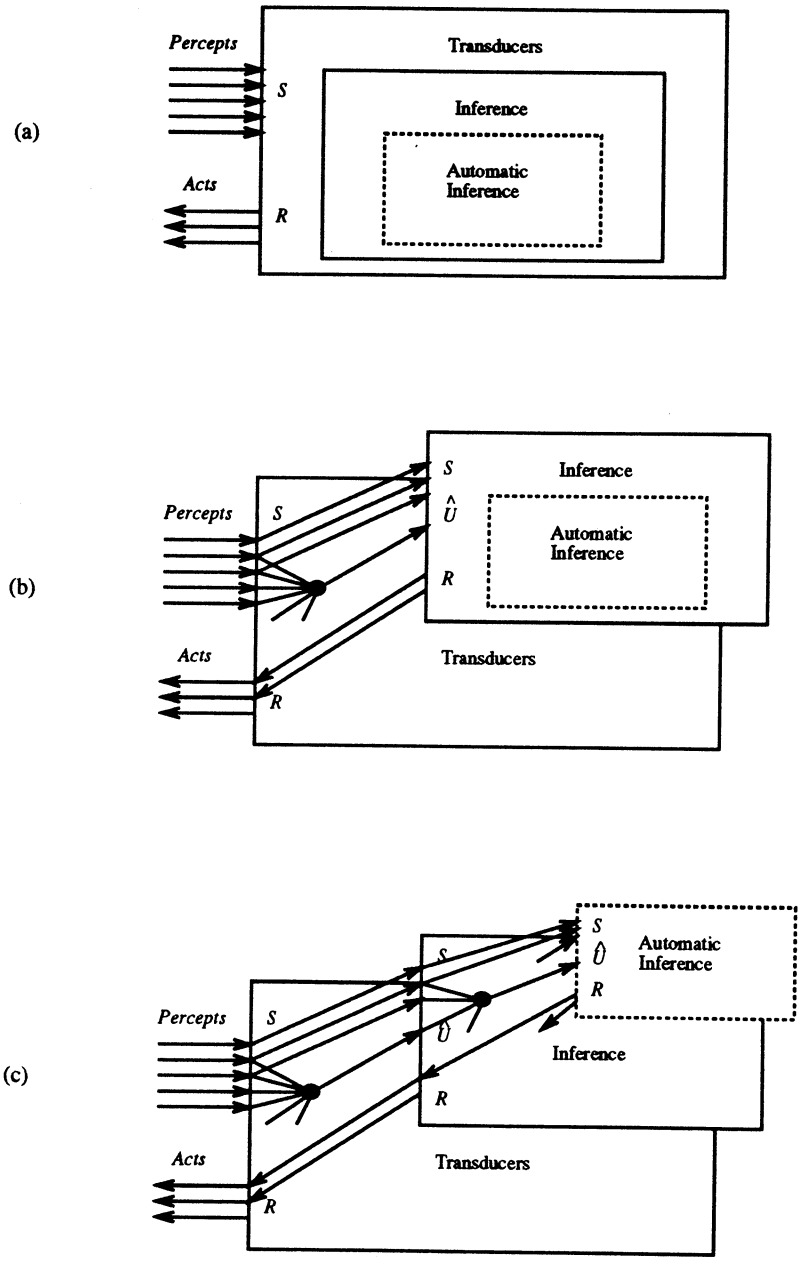


Figure 3.2: (a) Agent's interface with world. (b) Integral view of automatic inference. (c) Functional view of automatic inference as autonomous subagent. $\hat{U}$ is an informal abbreviation for estimated expected utility, a simplification for intuitive purposes; recall that in the analysis we are actually estimating value rather than utility of computations.

large role in comprehension processes. Associative mechanisms are intrinsically context-sensitive and the presence or absence of context is a continuum rather than a binary condition. The degree to which "embellishments" are automatically made is a function of all the preceeding contexts weighted in some way by their recency.

Second, not all forward (elaborative) inferences in these studies need be of the type denoted by the term automatic inference. As noted in section 3.3, all of interpretation is in some sense forward inference, including controlled processes that require attention and are sequentially chained and segmentative. The experiments described above do not necessarily distinguish how the elaborative inferences are made. The speed with which the inferences are made could yield some clues. However, procedures for measuring the speed of forward inference would require a degree of sophistication not found in existing online empirical methods.

Despite these problems, the limited empirical evidence relating to language understanding plus a healthy dose of intuitionism have led a number of researchers to propose dichotomies related to the automatic/controlled inference distinction I outlined in section 1.1.1. Psychologists have long used the informal terms "reflexive" and "reflective" in much the same way. What Marslen-Wilson & Tyler (1980; 1981; 1987) call "obligatory processing" in their language understanding work is precisely automatic processing in the traditional, domain-independent sense (following LaBerge & Samuels 1974; Posner & Keele 1975; Shiffrin & Schneider 1977).

Fodor (1983) made an influential and controversial distinction between modular and central cognitive faculties. In many respects his criteria parallel those distinguishing automatic from controlled inference. Though my conclusions diverge sharply from his, his proposal is worth examining. The overtone of his arguments suggests that Fodor's intuitions are related to processing resource bounds and the forward/backward inference distinction. This is the intuition formalized by the processing arguments of the previous section; yet these arguments are in no way restricted to particular knowledge types and domains.

Fodor identifies nine properties that distinguish what he calls modular systems from central systems:

1. *Domain specificity.* Each module is specialized for dealing with a particular domain, input sensory systems being the most obvious example.

2. *Mandatoriness.* Modular processing must always occur when input is received.

3. *Limited access to intermediate representations.* Modules compute intermediate results that are not made available to other subsequent processes.

4. *Speed.* Modular systems are fast and reflexive.

5. *Information encapsulation.* Feedback from central systems to modular systems is highly restricted.

6. *Shallow output.* The final results (as opposed to intermediate representations) of modules are of a "shallow" form, implying that results can be computed with higher certainty.

7. *Neural localization.* Modules correspond to specific neurally hardwired areas of the brain.

8. *Characteristic breakdown patterns.* Common types of breakdowns are evidence for neural localization.

9. *Ontogenetic uniformity.* Regular patterns in child visual and language acquisition are evidence of innate modularity.

Of these criteria, both "mandatoriness" and "speed" are essential characteristics of automatic processing. Interestingly enough, in his exposition on mandatoriness (p. 53–54) Fodor suggests a caveat on the attentional definition of automaticity, by observing that one can choose to "not attend" to the input speech. Instead, he proposes, the automatic input system continues processing, but their output is ignored by central systems.

Where I and others diverge from Fodor is his conclusion that automized and modular processes are confined within static, *a priori* determined boundaries conveniently situated at Logical Form, between the semantic and conceptual systems. Fodor's other criteria, particularly "domain specificity" and "information encapsulation", are intended to distinguish vertical syntactic faculties. While the criteria themselves are important (in fact these criteria are used extensively in chapter 4), they diagnose *tendencies* rather than rules, and the available evidence indicates that a great deal of horizontal, cross-modular processing occurs as well. Marslen-Wilson & Tyler (1987, p. 39) have argued cogently that automaticity is not confined to processes prior to Logical Form (roughly corresponding to lexical semantics in my model). Various discourse phenomena seem to exhibit both mandatoriness and speed, as well as most of the other modularity properties. It is problematic that on this issue most arguments, including Fodor's and Marslen-Wilson and Tyler's, are largely phenomenological; the search for more direct empirical methods continues.

Shastri (1988b, 1988a, 1989) argued that because agents must act in real time, the correct question to be asking in AI is what forms of tractable "limited inference" are needed.[5] He proposes "inheritance" and "recognition" as forms of limited inference, and implements them using evidential reasoning in a structured connectionist model that operates in sublinear time. Recognition tasks are of the form, "What is something that is red and sweet most likely to be (apple, grape, banana)?" Shastri's recognition is a restricted case of automatic inference, since it only permits answers that are single, predefined concepts. Automatic inference permits the answer to be composed from multiple predefined concepts, so the answer to "What is something that is blue and creamy most likely to be?" could be "blueberry ice cream". Inheritance tasks are of the form, "Which color is an apple most likely to be (red, green, blue, or yellow)?" Posing this query in terms of automatic inference is harder; it requires first asking "What is something that is an apple and colored?" and then extracting the color value from the result. It seems intuitively plausible that inheritance queries should require the extra step.

In Weber's (1989c, 1989b, 1989a) connectionist model for interpreting figurative adjective-noun constructions, "direct inference" refers to the priming of property values by the values of correlated properties. Priming effects of this sort are an excellent example of automatic inference operating in the semantic and conceptual domains.

Rieger (1977) suggested applying the notion of data-driven "spontaneous computation" to a wide range of tasks including plan recognition in story understanding and plan critics for problem solving. The proposed automatic inference theory can be seen as a specific version of this. Note, however, that spontaneous computation also applies to tasks that I would suggest require at least some controlled inference, such as plan recognition.

[5] In a more recent paper, (Shastri 1991, p. 111-2) also adopts the terms "reflexive" and "reflective reasoning" over "limited inference".

3.5.2 A Model of Automatic Inference in an Agent

I will be making a significant assumption, namely, that the automatic and controlled inference computations are independent, at least up to the point that it is possible to optimize the utility of the automatic inference mechanism independently of all else. From the autonomous perspective, a separate subagent performs automatic inference, maximizing its own expected utility by observing how its computations produce (binary) utility feedback from the rest of the cognitive mechanism. Fodor (1987) notes that any modularized agent is inherently irrational in the sense that its "informationally encapsulated" knowledge prevents it from using all its available knowledge for all its computations. In a similar discussion on treating metalevel computation as a separate agent, Russell and Wefald (p. 71) observe that it is difficult to show that what is rational for the separate agent is rational for the agent as a whole. With regard to automatic inference, the desired autonomy property can be realized by adopting the following "loose interaction" assumptions:

- *Fixed computation resource allocation.* Performing automatic inference does not in any way interfere with the progress of other inference mechanisms. That is,

  $$|S_{A_I} \wedge S_j| = \max(|S_{A_I}|, |S_j|)$$

  for all S_j. Thus whenever $|S_{A_I}| \leq |S_j|$,

  $$TC(|S_{A_I} \wedge S_j|) = TC(|S_j|)$$

  implying that

  $$\begin{aligned} & E[U([\alpha_{S_{A_I},S_j}, [S_{A_I} \wedge S_j]])|\mathbf{e} \wedge e_A \wedge e_j] - TC(|S_{A_I} \wedge S_j|) \\ \geq\ & E[U([\alpha_{S_j}, [S_j]])|\mathbf{e} \wedge e_j] - TC(|S_j|) \end{aligned}$$

  and therefore

  $$V(S_{A_I} \wedge S_j) \geq V(S_j)$$

  Note that this does not mean that the cost of an automatic inference step is itself zero, but just that parallelism in the architecture allows it to be performed concurrently.

- *Fixed non-automatic inference.* The agent's non-automatic inference processes do not adapt so as to optimally exploit automatic inference. Thus the utility of an automatic inference action is

  $$U([\alpha_{\sigma,S_{A_I}}, [\sigma \wedge S_{A_I}, W_k]])$$

  where σ is the current best computational action S_j as determined by the metalevel.

- *Binary utility.* The agent's architecture is such that whether it can make use of the result of automatic inference computation is a binary all-or-nothing condition. The agent either can or cannot use the result. Either the result makes no difference on the agent's object level decision, in which case

  $$U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) = U([\alpha_\sigma, [\sigma, W_k]])$$

or the *gain* in utility is a constant "inference value" IV,

$$U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) = U([\alpha_\sigma, [\sigma, W_k]]) + IV$$

One way to intuitively motivate this simplifying assumption is to consider the case where automatic inference produces a speedup ΔT in the time that it takes the agent to reach its object level decision, such that

$$U([\alpha]) - TC(|\sigma| - \Delta T) = U([\alpha]) - TC(|\sigma|) + IV$$

- *Supervised learning.* The only performance feedback that the computational architecture permits the agent to make available to the automatic inference subagent is the "correct" result that automatic inference should have computed, i.e., the result that (1) could have been used by the agent, and (2) could conceivably be computed by the automatic inference architecture. The subagent takes the value of the "correct" result to be IV. Denote each instance seen by the subagent, then, by the pair $\langle A_I, \mathbf{e} \rangle$.

Analyzed from the integral standpoint, any automatic inference that is made should be selected in the course of maximizing the agent's global estimated utility. Recall that to maximize its estimated utility, the agent chooses a computational action with the maximum positive value. Computational actions are now $\langle S_j, S_{A_I} \rangle$ pairs since automatic inference can be performed concurrently with other computations. The agent selects

$$\begin{aligned}
&\langle S_j, S_{A_I} \rangle : \max_{j,I} V(S_j, S_{A_I}) \\
&= \langle S_j, S_{A_I} \rangle : \max_{j,I} U([\alpha_{j,A_I}, [S_j \wedge S_{A_I}, W_k]]) - U([\alpha_{S_j}, [S_j, W_k]]) \\
&= \langle \sigma, A_I \rangle : \max_{I} U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) - U([\alpha_\sigma, [\sigma, W_k]])
\end{aligned}$$

the last following from the fixed non-automatic inference assumption, where

$$\sigma = S_j : \max_j U([\alpha_{S_j}, [S_j, W_k]]) - U([\alpha, W_k])$$

Thus S_{A_I} can be chosen independently by maximizing $V(S_{A_I})$ separately from $V(S_j)$:

$$\begin{aligned}
&S_{A_I} : \max_I V(S_{A_I}) \\
&= S_{A_I} : \max_I U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) - U([\alpha_\sigma, [\sigma, W_k]])
\end{aligned}$$

From the functional standpoint, the automatic inference selection process can now be viewed as an autonomous subagent with its own utility function

$$U_{\text{auto}}(A_I) \stackrel{\text{def}}{=} V(S_{A_I}) \tag{3.11}$$

I will continue to use the integral standpoint for the next several paragraphs to show that the agent's global estimated utility is being optimized.

Of course neither the agent nor the subagent can predict the exact utilities without actually performing the computations and actions. Instead, they must be estimated given only the evidence available so far from previous computations and percepts. The automatic inference must be chosen on the basis of the evidence:

$$S_{A_I} : \max_I V(S_{A_I}|\mathbf{e})$$

This selection can be performed by the previously described engine that maximizes the probability of drawing a useful inference. From the binary utility assumption we know

$$U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) - U([\alpha_\sigma, [\sigma, W_k]]) \in \{0, IV\}$$

and therefore can define

$$\text{Useful}_k(A_I) \stackrel{\text{def}}{=} \begin{cases} \text{true}, & \text{if } U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) - U([\alpha_\sigma, [\sigma, W_k]]) = IV \\ \text{false}, & \text{if } U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) - U([\alpha_\sigma, [\sigma, W_k]]) = 0 \end{cases}$$

The derivation is then:

$$\begin{aligned}
& S_{A_I} : \max_I V(S_{A_I}|\mathbf{e}) \\
&= S_{A_I} : \max_I \left(\sum_k P(W_k|\mathbf{e}) \cdot [U([\alpha_{\sigma,A_I}, [\sigma \wedge S_{A_I}, W_k]]) - U([\alpha_\sigma, [\sigma, W_k]])] \right) \\
&= S_{A_I} : \max_I \left(\sum_{k:\text{Useful}_k(A_I)} IV \cdot P(W_k|\mathbf{e}) + \sum_{k:\neg\text{Useful}_k(A_I)} 0 \cdot P(W_k|\mathbf{e}) \right) \\
&= S_{A_I} : \max_I \left(\sum_{k:\text{Useful}_k(A_I)} IV \cdot P(W_k|\mathbf{e}) \right) \\
&= S_{A_I} : \max_I \left(\sum_{k:\text{Useful}_k(A_I)} P(W_k|\mathbf{e}) \right) \\
&= S_{A_I} : \max_I P(\text{Useful}(A_I)|\mathbf{e}) \qquad (3.12)
\end{aligned}$$

This final expression is of a form that can be sampled empirically by the automatic inference subagent. In section 7.5 I discuss some heuristic approaches to hand-gathering rough statistics of this form. Chapter 8 considers more theoretically sound methods for automated learning of the distribution, and related issues concerning generalization.

3.5.3 Language Bias

Russell and Wefald's rational metareasoning framework provides a powerful and general basis for modelling adaptive intelligent behavior. However, to apply it to natural language interpretation we must also take into account the cognitive biases that cause humans to learn language the way they do. The human cognitive architecture is a concrete entity that cannot alter its execution architecture arbitrarily. The brain constrains computation in very specific ways, and it is difficult to believe that any reasoning and learning model could produce adequate results without

approximating fairly closely the inductive biases implied by those architectural constraints. We should therefore incorporate as much information about the structure of the cognitive architecture as we can determine from psychological and neurological observations. In machine learning terminology, the *language bias* for induction (Utgoff 1986) should correspond to that of humans. Yet another way to look at the position I am suggesting is that it is a decision-theoretic augmentation of the innateness postulate (though the kinds of innate features could be more primitive than typically assumed in linguistics).

In choosing a probabilistic architecture for a cognitive model, one is faced with the decision whether (1) to stick to a framework that is overly general and allow computations that the brain couldn't possibly implement (at least, in polynomial time) and then to claim that the brain operates by heuristically approximating the framework which is normative, or (2) simply to "build in" the brain's architectural constraints by selecting the appropriately corresponding language for the model space, and then claim that the brain actually performs optimally under these constraints. In practice, as in this work, it is difficult to achieve either extreme. As much as possible, a model will incorporate constraints that are intended to correspond to physical constraints in the human cognitive system, but many aspects will remain underconstrained. In the following chapter we consider the ontology that generates the language bias in the proposed model.

Chapter 4

Chapter 4

Modular Ontology

In this chapter I outline a modular framework for ontology, that is, a mental representation or cognitive semantic system for a linguistic agent. I propose to include the following modules:

1. *mental images,*
2. *lexical semantics,* including image schemas,
3. *conceptual system,* and
4. *construction lexicon.*

There are both empirical and representational expressiveness motivations for the modular approach taken. Insights of recent empirical work in cognitive linguistics, semantics, mental representations, and ontology are surveyed. These are synthesized into a framework based on multiple ontological modules (or levels) that, taken as a whole, account for meaning differences. The more concrete semantic modules, like the mental image module, are closely connected to sensory registers, while others are more abstract like the reified image schema level. Modules interact with each other via intermodular associations, so for example, a hierarchy of associations (the "signification constructions") brokers the interaction between the phrasal lexicon and other semantic modules. The semantic system does not come with a direct mapping to the "real world" but interfaces only through the perceptual system.

It is probably the case that no exact formal definition can be given for what constitutes the intermediate conceptual state produced by interpretation. To precisely define a cognitive model (or any model), an empirical testing procedure giving falsification conditions must be given. The model must be defined functionally, with respect to *observable* inputs and outputs. For example, the Turing test defines an intelligent natural language system as that which produces conversational responses indistinguishable from those produced by humans. A conceptual state, however, is not an observable output. Thus, at least at present, it is hard to do better than to characterize conceptual states in terms of intermediate purposes, in the vein of the preceding chapter.

The mentalist ontology can be criticized for its ungrounded semantics which makes falsification difficult, but this is less of an immediate concern in a long-range research program. Realists like Barwise & Perry (1983) and Dowty (1989) would argue for mapping conceptual structures to

their real-world referents, at least in part to make the semantic theories more empirically verifiable. While it is true that a mentalist semantic system is discomfittingly unfalsifiable, realist systems are not well enough motivated to make them adequate alternatives. Realist theories ignore the fact that humans make use of concepts not merely to describe the world, but to facilitate achieving particular goals. There is no doubt that being able to describe the world is an important part of rationality but this is not inconsistent with the intermediate representation levels hypothesis. Significance can be given to the representational symbols via the notion of *associative grounding* rather than mapping to real-world referents (section 4.2.2). The process of discovering the appropriate intermediate conceptual levels is a longer-term proposition than the realist program. Assuming the methodology is correct, verification will eventually come in terms of an agent's ability to succeed in its environment as humans do, with linguistic communicative ability being one of the agent's tools.

In section 4.1 I sketch the modular framework used in the proposed model. Section 4.2 discusses some of the issues that motivated it, having to do with the need for more representational expressiveness. Subsequently, sections 4.3–4.4 survey some of the evidence from different fields converging upon modular approaches of the kind proposed.

4.1 Intermediate Levels of Meaning

Though linguistic work often assumes a single uniform level of semantics, there are different types of intermediate conceptual states, if for no other reason than the fact that humans possess limited processing resources. These intermediate levels, one suspects, are responsible for the multitude of intuitions that researchers have had about "Logical Form", case and thematic roles, lexical semantics, and so on. For many years, linguists have debated whether thematic roles properly belong at the "shallow" or "deep" semantic level, because various pieces of evidence suggest cognitive representations at different levels. Debates between advocates of propositional and image-based representations oversimplify the more plausible case which is that both forms are employed.

The problem with the multimodular approach is twofold: (1) the relationship between modules must be well-defined, and (2) the modularization must be well-motivated and must permit all necessary interaction between modules including bi-directional "feedback" interaction. Traditionally in multimodular approaches, the relationship between two modules is defined algorithmically as a sequential algorithm for mapping the contents of one module to another is given. This satisfies the first condition, but often, at the expense of the second. I propose an alternative which is to define intermodular relationships based on statistical association.

In this approach, modules are broken down according to information type rather than process type. This is a weaker form of modularity than Fodor's (1983). In Fodor's proposal, the "limited access to intermediate representations" and "information encapsulation" constraints combine to prevent mental processes from crossing vertical modular boundaries in either "forward" or "backward" directions. One of the primary benefits of these constraints is that they place restrictions on one dimension of processing complexity, thereby ostensibly improving our understanding of how mental processing can be tractable. However, the model proposed here eliminates the constraints since they conflict with empirical data, raising the question of why we would want to retain a modular approach. The answer is that the statistical paradigm, which Fodor didn't have

recourse to, includes notions like information entropy that can help us use ontological modularity to reduce processing complexity, even without the strong constraints. Connectionists in particular have made it a priority to integrate process and representation: in this paradigm, modularity in representation leads directly to improved modularity in processing complexity, without sacrificing interaction between modules. Thus, even though the data does not bear out strong modularity hypotheses like Fodor's, it seems likely that the structural patterns that have been discovered will turn out to have statistical complexity and computability advantages, and we should therefore keep these modules while relaxing the encapsulation assumptions.

The remainder of this section gives an overview of how the ontology is structured.

4.1.1 Overview of Modules

Figure 4.1 depicts how modules are organized in the ontology, using a topological device. Only modules that share a flat border have direct associations. (We see in chapter 6 how to realize this constraint by restricting the way different modules can participate in the same correlational term.) The nonshaded modules—lexicosyntactic constructions, lexical semantics, spatial images, and the conceptual system—are the ones for which some representation has been attempted here. The shaded modules are not studied here, and are shown only to orient readers familiar with those subfields.[1]

Conceptual structures mediate the relationship between mental images and natural language. In this chapter we will consider two such intermediate-level modules, lexical semantics and the conceptual system. It is possible in these modules to represent abstract conceptual propositions that either have no directly corresponding image, or violate spatiotemporal properties. While it is necessary to ground such things as spatial relations in perception, other intermediate representations can be associatively grounded instead. From the memory processing point of view, additional distinctions need to be made as to the directness of the connection between perception and representation. This is discussed in section 5.2.2.

4.1.2 Common Characteristics of All Modules

Contrary to Fodor's model, automatic processes are permitted in any module. As discussed earlier, Marslen-Wilson & Tyler (1987) argue convincingly that automaticity extends all the way to discourse representations (section 3.5.1). However as with Fodor, controlled inference is permitted only in the conceptual level; in other words, syntax, lexical semantics, and mental images are only accessible to controlled inference via their automatic associations with the conceptual module. Thus far, no study contradicts this hypothesis, including Marslen-Wilson and Tyler's.

Each module adds statistical constraints, representing associations between conceptual structures, that can influence the interpretation of an utterance. Chapter 6 discusses how statistical constraints are captured using *correlational terms*. Correlational terms can be used to assert associations between conceptual structures that are all within one module, or else they can span bordering modules (figure 4.1) to assert intermodular associations.

Structures are always combined in strictly compositional ways; we will see that this is how the representation enforces transitivity relationships. "Non-compositionality" enters the

[1] In particular, the term *audio images* may be confusing. Mental image researchers often use "image" in a generalized sense that encompasses eidetic representations of any sensory mode.

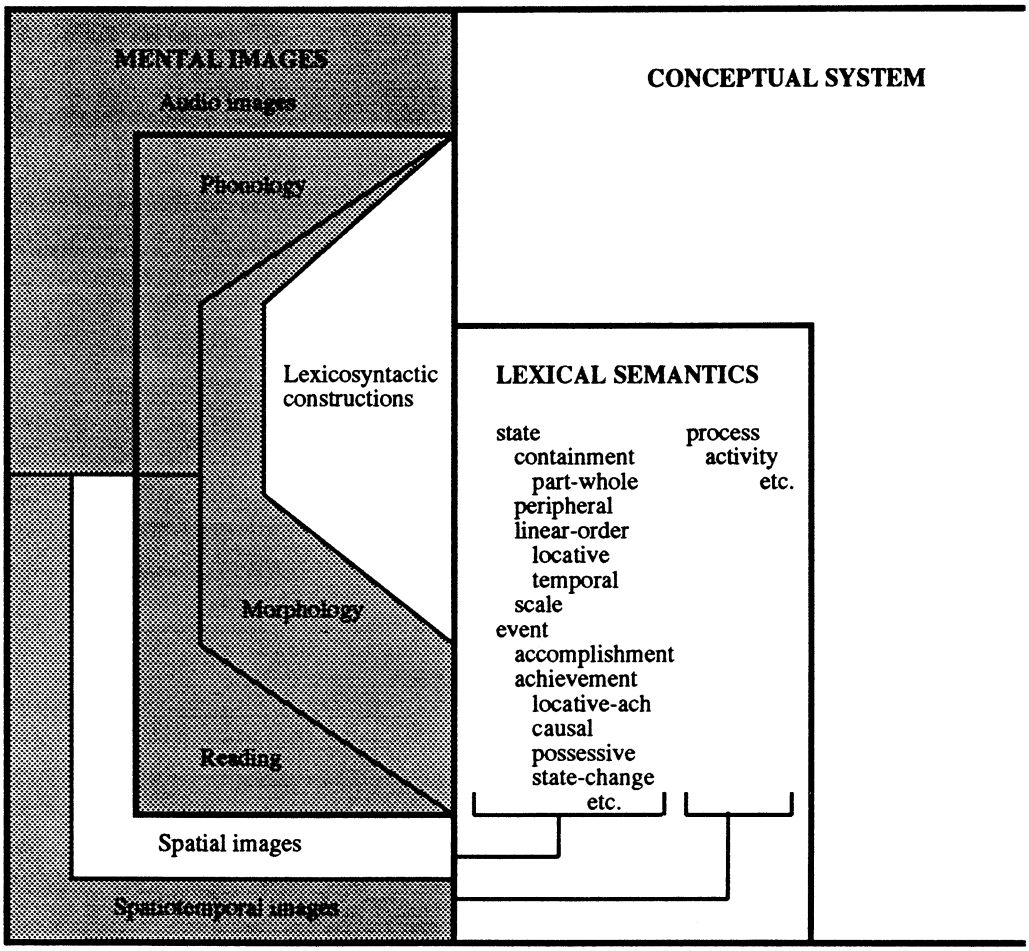


Figure 4.1: Modularization of the ontology.

model by the fact that alternative substructures can be chosen to account for different inputs, so local compositional rules can be overridden. This is analogous to any non-monotonic parser (say, a chart parser): the preferred interpretation for an input sequence can switch when given subsequent input, as in

(4.1) The environmentalist couple put together a baby bottle.

(4.2) The environmentalist couple put together a baby bottle bill.

Intermodular mapping rules, including the rules that map syntactic to semantic structures, are probabilistic, or associative, rather than absolute. Each individual rule is strictly compositional; however, the association it represents can be defeated by other stronger associations. Partee (1984) gives a good survey of compositionality issues; in her terms, the model proposed here is compositional in a weak sense but not in the strong Montague sense.

4.1.3 Differentiating Characteristics of Modules

The primary differentiating characteristic is *ontological expressivity*. Each module has a different set of compositional roles that in effect determines what can be expressed within that module. Thus, although every module expresses statistical associations that influence interpretation preferences, each module is restricted to expressing certain kinds of statistical associations. Though all compositional roles satisfy the transitivity constraint, they capture different abstractions. This means that the representations in different modules intrinsically enforce different relational consistency conditions. Of course, each module also has a different set of primitive features corresponding to its domain. This is essentially Fodor's "domain specificity" condition.

This approach, where different modules hold disparate types of information but are highly interconnected, can be seen as a generalization of Paivio's (1971, 1986) dual coding theory. Figure 4.2 shows a schematic conception of a mental representation system broken into verbal and non-verbal systems. Mental image structures correspond to what Paivio calls "imagens", and the other modules are all lumped into "logogens"; imagens and logogens are highly associated by referential connections. Paivio distinguishes referential connections from ordinary associative connections, but referential connections (i.e., signification) should really be reserved for lexicosyntactic mappings. The connections between imagens and logogens would not all have to be referential if logogens were divided into lexicosyntactic and non-image conceptual modules.

A second condition is *constrained maximum information* (CMI for short). As I have said, the "limited access to intermediate representations" and "information encapsulation" constraints appear to be too strong. The CMI condition is a weaker constraint that addresses qualitatively similar concerns. It is put forth here solely for motivationary purposes, and will not be developed in depth since its parameters are determinable only by long-term empirical study.

Essentially, the frequency distribution of the input data should be summarized using as few correlational terms as possible. However, the set of correlational terms must be chosen subject to resource bounds, which can be viewed as an *a priori* model of neural architectural restrictions. First, the function of certain modules is preset (in particular, those close to the perceptual apparatus). Second, though non-perceptual modules are not preassigned any function, the long-term

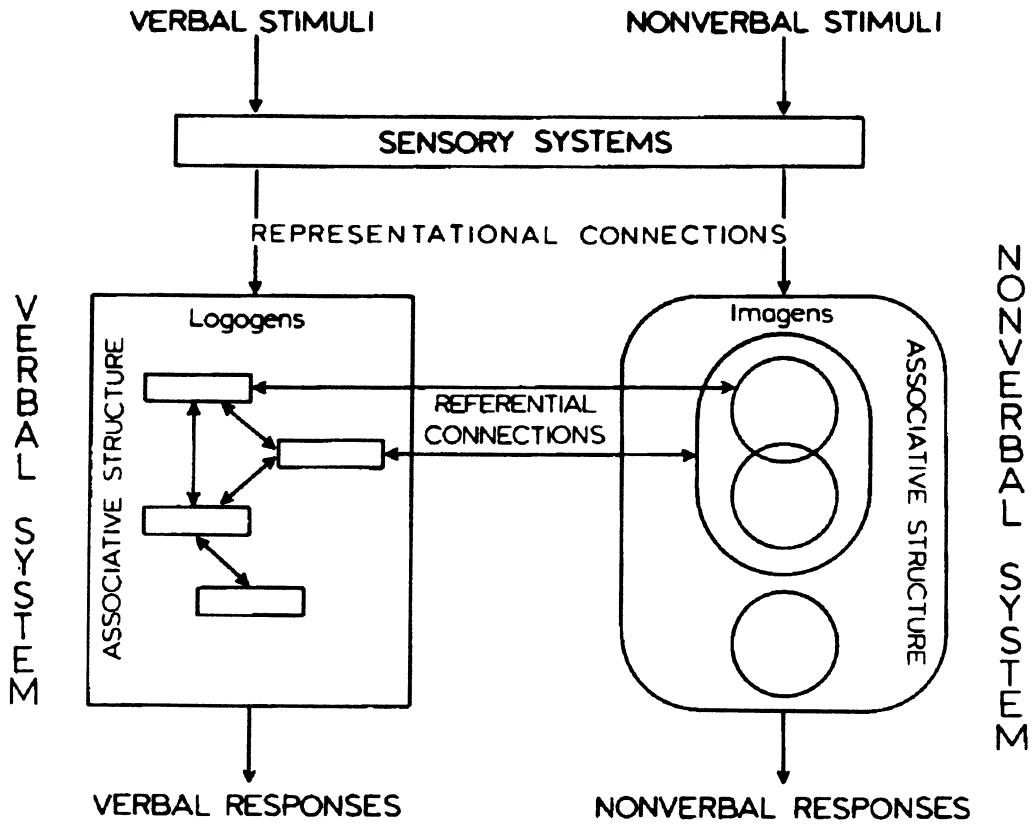


Figure 4.2: The dual coding theory of mental representations (from Paivio 1986, p. 67).

storage capacity of each module is bounded.[2] There is one limit parameter for each module, to be estimated empirically. Third, correlational terms are only permitted to relate structures from bordering modules (figure 4.1); this is a topological constraint. This means associations between non-bordering modules must either (1) emerge from combinations of other, legitimate correlational terms, or (2) be captured by correlational terms that span some shared intermediate module(s). Within the preceding constraints, the CMI principle assigns all non-perceptual primitives[3] to the remaining modules—thus drawing the boundaries of a modular ontology—so as to preserve as much of the input frequency distribution as possible.[4]

Whether the CMI principle will be borne out with respect to the particular modules I have proposed remains to be seen. Certainly the principle is overly strong as stated, and in practice a weaker interpretation should be taken. There is no reason to assume that neural modularity should have evolved perfectly to draw the boundaries at the exact optimal minima. Instead, the condition should be something like "boundaries between modules cause the learnable part of the input distribution to be close to optimal under the resource constraints". We can reasonably assume that the boundaries conform relatively well to visual and linguistic usage patterns, since humans and natural language are tuned for efficacy.

Below I suggest some other characteristics of modules for intuitive purposes. These are rough and imprecise, and are therefore open to question and not very useful as diagnostic criteria for distinguishing modules.

The *size limitation* condition says that a module can only hold active conceptual structures of a particular size. This is a "working memory" restriction rather than a "long-term memory" restriction of the sort given above. A more specific version of this condition, related to Fodor's neural localization criterion, can be postulated by asserting that each neurally localized site has some static size limitation.

The *simultaneous conceivability* condition says that one is more likely to be able to simultaneously hold multiple inconsistent concepts active if they are in different modules. This idea will become clearer below in discussing schematization associations. Note that "inconsistent" is meant here in the intuitive sense. What the condition really means is that alternative ways of schematizing situations are likely to be found in different modules. Formally, the "inconsistent" concepts must be consistent since no active conceptual structure is permitted to be inconsistent.

The *completability* condition says that it is easier to recognize that a pattern is "complete" within a module. An utterance is syntactically complete if a complete parse is recognized. A conceptual proposition is complete if its argument structure is filled in. A complete visuospatial image must have some minimum set of features such as shape and dimensionality. On the other hand, it is much more difficult to say if a complete semantic interpretation has been recognized for an input utterance; instead, one just does the best one can to associate a conceptual interpretation with the input utterance.

---

[2] One way to do this would be to set a limit on the information-theoretic entropy over all correlational terms within a module.

[3] In the proposed model, features and role features.

[4] The discrepancy between the input and stored distributions can again be measured using information entropy.

4.2 Representational Needs

4.2.1 Image Reification

The representation of physical objects in existing AI formalisms usually fails to make clear the way in which the object is conceived. For example, in conceptualizing a *desk*, you inherently conceptualize its *drawers* and the entire picture forms a "gestalt". On the other hand, it is less obvious that you always need to conceptualize the *handles* on the drawers as well, but do so only in situations where you need to focus on the handles, perhaps to open a drawer. Furthermore, the fact that the handle is a COMPONENT of the drawer is normally left implicit in your "drawer gestalt" image, but sometimes you need to be able to reflect explicitly on the component relationship, for example to assemble a new desk or to understand this sentence.[5] As another example where a relationship must be considered explicitly, take

(4.3) Becky says half the fun of eating marrow is that the marrow is in the bone.

where it is the *containment* relationship, rather than *marrow* or *bone*, that is "the fun of eating".

AI formalisms have largely focussed on the *epistemological* rather than *heuristic* aspects of knowledge representation (for readers unfamiliar with this distinction, I will discuss it in greater depth later). The upshot is that knowledge representation work has concentrated on ensuring that things like desk-in-drawer and handle-component are *somewhere* in the knowledge base, but the form and/or location in the knowledge base is less important. Unfortunately, this obscures some equally important issues. For example, suppose there is only one representation of the COMPONENT relation between handle and drawer. Sometimes this relation will have to function as an implicit part of the drawer gestalt; other times the same relation must function as an explicit concept that can be reasoned with, as any other concept. This leads to messy higher-order logics where predicates (or roles or relations) are permitted to apply to other predicates (like COMPONENT).

Alternatively, an operation called *reification* can be used to transform roles (or relations or predicates) to objects (constants).[6] For example, in the representation of *drawer* in a component role of *desk*, COMPONENT might normally be a role that cannot be used as an argument to another role. However it can be reified into an object; as we see in chapter 5 the object is a schema in which both the desk and the drawer play roles. This object can be used as an argument to some other structure, for example to handle

(4.4) In this design, the drawers stand independently on the floor instead of being part of the desk.

One primary purpose of the proposed modularization is to capture structures at different levels of reification. Reification does not transform arbitrary relations into arbitrary objects. Rather, there are conventional types of reification, that associate specific types of relations and objects in regular ways. Below are listed several types of *image reification* that are dealt with by the ontological primitives proposed in chapter 5. Image reification is defined as the relationship between concrete "vivid" conceptualizations and reified schematizations of images.

[5] To reduce ambiguity I will follow the convention of capitalizing names of roles (relations).

[6] Reificiation is related to *paramodulation* and *demodulation* inference rules in logical deduction systems. The exact relationship depends on the interpretation of one's constants and predicates.

The case of *desk drawer* above is an example of a kind of image reification called *constituency reification*. On one hand, the base-level representation, where *drawer* is an implicit component of *desk*, is a mental image representation in which physical objects are conceptualized as gestalts whose spatial properties are largely preserved (as discussed in section 4.3). On the other hand, the mental image module does not hold the reified representation in which *desk* as well as *drawer* play roles connected to an explicit containment-schema object reified from the COMPONENT role. (As discussed later, such reified representations are employed by both the lexical semantics and conceptual modules.) Note that there may still be mental image representations of *desk* and *drawer*. Only the reified object and its associated roles are excluded from the mental image module.

The reified representation permits imposition of explicit foreground/background distinctions. For example, the whole (*drawer*) can be marked as the background (LANDMARK) and the part (*handle*) can be marked as the foreground (TRAJECTOR). Foreground and background distinctions make sense only in the context of an embedding schema. Since there is no embedding schema in the mental image representation, no explicit foreground/background distinction can be made in the base-level representation.

Two other kinds of image reification are state and property reification. *State reification* occurs when a mental image representation of some relational state is turned into a schematic object. Wilensky (1989) calls this qualitization. For example, the mental image representation of an *apple* being *red* is transformed into a state schema to handle something like

(4.5) Adam, who had a fondness for bright things, only ate the apple because it was red.

Property reification occurs when the fact that some concept has a property is itself made a proper concept. For example, the same mental image representation of a red apple can be transformed into a property schema to handle

(4.6) It was a Granny Smith apple that had a red color because it had been spray painted.

4.2.2 Associative Grounding

The notion of modules also facilitates addressing the "symbol grounding" issue, because modules have varying degrees of closeness to the perceptual apparatus. The symbol grounding issue (Harnad 1990) concerns defining the meaning of symbols or structures, particularly at abstract levels like image schemas, state schemas, or property possession schemas. Harnad argues that the meaning of conceptual symbols must be grounded in the agent's perception and use of those symbols. A more concrete version of the issue is: what does it mean to apply a visuospatial relation to a non-physical object? We will see in chapter 5 that image schemas can be used to capture not only visual distinctions, but all sorts of abstract distinctions like *low anxiety* and *high speed*. The source of meaning for an image schema, however, derives from those occasions when it is used to describe an actual image. To begin with, consider the question from the standpoint of traditional frameworks. If some relation is defined with reference to a visuospatial domain, to apply it metaphorically to non-visual objects is to violate the selectional restrictions that form an intrinsic part of its definition. We are left with two choices: (1) non-visual objects can be mapped into a visual representation through some sort of regular transfer protocol, or (2) the relation itself can be abstracted to apply to non-visual objects as well. Humans employ both mechanisms, but we will focus on the latter.

I propose the notion of *associative grounding* as the basis for formal grounding of symbols that have no direct perceptual connection. Associative grounding is a probabilistic version of symbol grounding, where the significance of symbols is determined either

1. by their occasional co-occurrence with symbols in a module closer to the perceptual apparatus, or

2. by the degree of syntactic and structural correspondence to symbols that are grounded as in (1).

In case (1) the notion of absolute selectional restrictions is replaced by probabilistic association (section 7.3.2). In case (2), syntactic correspondence for feature structure representations translates to the amount of overlap in feature values and roles, including recursive application to substructures. The mental image representations in section 4.3 are more directly connected to the perceptual apparatus than image schemas, which are in the lexical semantics module. One direction in cognitive linguistics is to ground the semantics of image schemas perceptually (Talmy 1983, 1985, 1988; Lakoff 1987b; Langacker 1987; Feldman *et al.* 1990). The object here is to formalize this notion so as to apply to abstract uses of image schemas as well. By considering how image schemas are associatively grounded to the mental image module, the reader should be able to extend the general idea to other cases, like grounding the mental image module to more "vivid" visual primitives, or grounding of temporal event and process schemas. In the case of image schemas used to describe perceptual images, the image schema symbols are grounded by their frequent association with the corresponding unreified mental images.

The meaning of an associatively grounded symbol is determined indirectly by the intermodular associations that syntactically similar symbols participate in. This is similar to the notion researchers have had in the past about the meaning of nodes in semantic networks. However, in traditional non-statistical frameworks the problem is to abstract a visuospatial relation without making it altogether meaningless. If in the extreme case all forms of selectional restrictions are removed the resulting relations will be indistinguishable. The associative grounding approach employs a statistical remedy, allowing us to view selectional restrictions not as hard and fast rules, but rather as probabilistic distributions. Because of these distributions, for a particular abstract relation (and visual ones as well), certain image schemas will be preferred over others. Thus there is still a difference between relations even if they can all apply to the same fillers.

4.2.3 Compatible Differentiated Semantics

The modular approach fulfills yet another representational need, similar to image reification, namely to capture both the conceptual commonalities and differences between sentences like

(4.7) Franny, with Zooey, went off to college.

(4.8) Zooey, with Franny, went off to college.

(4.9) Franny, together with Zooey, went off to college.

(4.10) Zooey, together with Franny, went off to college.

In one sense, the sentences have the same logical meaning, roughly *Franny and Zooey went off to college.* In another sense, the sentences express different events, with either Franny or Zooey being the actor. Most existing theories ignore the distinction. A few account for the similarity by positing a "co-agent" case role for the actor of the adverbial, so that Franny is the primary agent and Zooey is a secondary co-agent in sentence (4.9), and vice versa in sentence (4.10). However, this fails to capture the similarity between sentences (4.9) and (4.11):

(4.11) Franny, followed soon by Zooey, went off to college.

The reason it misses the similarity is that the co-agent role cannot be used to handle *Zooey* in sentence (4.11), since it is embedded in a reduced relative. Even if one wanted to use a co-agent role, there would be no place to attach the semantics of *followed soon by*.

Given that a single representation does not effectively capture all needed distinctions and generalizations, clearly the representation should allow *both* the common and different "meanings" of sentence (4.9). However, this raises the issue of which state should actually result from interpretation. The proposed modular ontology addresses this issue using the idea of a cross-modular semantic distinction, which allows interpretation to result in either or both "meanings", but in a clean fashion. In this approach, sentences (4.9) and (4.10) are considered to have different interpretations at the lexical semantics level, while the shared interpretation is held in the conceptual module. The lexical semantics module is able to capture the grammatical and semantic similarities between sentences (4.9) and (4.11); problems with co-agent roles do not arise because having a shared conceptual-level interpretation captures the meaning commonalities without co-agents. The representations at the lexicosemantic and conceptual levels may be held either individually or simultaneously. The correct result of automatic inference depends on which level(s) are most likely to be useful given the particular sentence and context (for this example I would guess that automatic inference should produce both representations, since both seem fairly salient).

I implied above that automatic inference for sentence (4.9) does not produce the concept *Zooey went off to college* at all—neither at the lexical nor at the conceptual level. This claim that might appear implausible since if you were to infer that *Franny and Zooey* went off to college, it seems you would also naturally know that *Zooey* went off to college. The proposed framework does not use static structures to represent this sort of duality, but instead deals with it using the notion of *associatively inferrable conceptual shift*. The notion suggests that one can rapidly shift between two different but closely related concepts. Though the shifts themselves are associative, they are initiated by controlled inference; it is only when the specific fact is needed that you infer that Zooey, independently of Franny, went off to college. The shift is easy since the two views are highly associated, and this makes it difficult to distinguish actual "meaning" from what is merely inferrable. Whereas above we considered the simultaneous holding of multiple (but consistent) interpretations across different modules, now we consider rapid sequential shifts within the same (conceptual) module. One must be careful about introspection here. If one asks whether sentence (4.9) implies that Zooey went off to college, one's instinctive response is yes. However, the intuition is misleading, because the very act of understanding the question creates the need to perform the inference, which otherwise might not have existed; see section 3.5.1.

It would be pointless to take up the debate over distinguishing statically stored knowledge from knowledge that can be inferred upon demand. The argument goes back at least to Plato's assertion in *Phaedo* that everyone is born knowing everything and that it is only a matter of time to

infer the facts one already knew. Chapter 3 argued that in any reasonable intelligent architecture some knowledge will be "cached" by forward inference, while other information will be inferred upon demand, with varying time requirements. We want to model how humans do this. Only empirical tests over the long run can determine the boundaries of cached forward inference. For now I only wish to show how the paradigm avoids problems caused by overconstraining the lexical semantics level, and suggest plausible examples. Moreover, even in the long run with extended empirical observation, the inferential boundaries probably differ significantly from person to person, making it pointless to strive for an exact theory. Rather, the important thing is for the theory to acknowledge the shiftable character of associated meanings.

A similar example is the relationship between *buy* and *sell*, two ways of conceptualizing a commercial transaction. Although both describe the same event, they seem to have different agents. A method of capturing the similarity without overloading the agent role, suggested by Jackendoff (1972), is to have both verb frames share a subcomponent that does not specify agency. This would be a case of cross-modular semantic distinction as discussed above if we thought of the shared subcomponent as being in the conceptual module, and the agentive frame as being in the lexical module. However, for the *buy/sell* example this explanation alone seems implausible. The conceptual notion of buying has a strong intrinsic sense of agency that should not be confined to a lexicosemantic level, and this is motivated by the fact that the concept of agentive buying, not just agentless commercial transaction, is needed in order to plan one's ordinary day-to-day actions. It makes more sense to take the associative conceptual shift approach, whereby one can shift between *buy* and *sell* conceptual frames by virtue of their high association. The two frames still share a common *commercial transaction locative achievement* subcomponent.

4.3 Mental Images

One may argue that it is possible to represent knowledge using different levels of reification without ascribing a particular module to each relation, simply by lumping all the base-level relations together with the relations that apply to reified relations. Indeed I believe this is sometimes unavoidable, and in the proposed modularization the conceptual module permits arbitrary multiple reifications. However, evidence is amassing from linguistics, psycholinguistics, and neurology for the localization of certain kinds of modules that also happen to capture a particular level of reification. We now turn to examine—by no means comprehensively, and purely for motivationary purposes—some of this converging evidence.

In this section we consider the first of the modules, mental images. Only visuospatial mental images are considered. There are three primary motivations for including a visuospatial mental image module. First, it demonstrates how the theoretical framework permits the agent's concepts to be associatively grounded in perception. As mentioned earlier, Harnad (1990) refers to this as the "symbol grounding" problem. Second, it provides a way to represent the distinction between abstract image schemas and more "vivid" images, which, as I discuss below, differ in nature but are nonetheless systematically related (by reification). Third, there is a good deal of evidence supporting a mental image module and thus, it makes sense methodologically to start with the assumption that the language bias for a probabilistic model incorporates mental images.

4.3.1 The Nature of Mental Images

One definition of mental images, proposed by Finke (1989), is "the mental invention or recreation of an experience that in at least some respects resembles the experience of actually perceiving an object or an event, either in conjunction with, or in the absence of, direct sensory stimulation".

"Mental image" does not connote any specific type of representation syntax (array, propositional, etc.). The issue of appropriate representations is taken up later.

Mental versus retinal and iconic images. Though closely tied to perception (of visual, auditory, or any other sensory nature) mental images are neither retinal nor iconic images. Retinal images are impressed upon the visual sensory apparatus. Iconic images hold percepts reflecting current or very recent stimuli. However, when I use "image" it should be understood as short for "mental image". Mental images are one step up the chain and are less vivid than iconic representations,[7] though more vivid than image schemas. They can be generated either from iconic sensory input, or by associative recall in which case they are similar to representations produced by previous sensory stimuli. A recalled mental image can always be distinguished from one being produced by direct sensory input, though as with many types of memories it can be difficult to recall whether some past mental image was generated from imagination or perceptual input (Johnson & Raye 1981; Finke *et al.* 1981).

Principles of mental images. Finke (1989) has proposed five major principles describing the essential properties of mental images, which, broadly interpreted, can serve as a working definition for the purposes of this work. These are:

1. *Implicit encoding.* "Mental imagery is instrumental in retrieving information about the physical properties of objects, or about physical relationships among objects, that was not explicitly encoded at any previous time." (p. 7)

2. *Perceptual equivalence.* "Imagery is functionally equivalent to perception to the extent that similar mechanisms in the visual system are activated when objects or events are imagined as when the same objects or events are actually perceived." (p. 41)

3. *Spatial equivalence.* "The spatial arrangement of the elements of a mental image corresponds to the way objects or their parts are arranged on actual physical surfaces or in an actual physical space." (p. 61)

4. *Transformational equivalence.* "Imagined transformations and physical transformations exhibit corresponding dynamic characteristics and are governed by the same laws of motion." (p. 93)

[7] Eidetikers are supposedly able to recall scenes in exact detail, as if able to recall the iconic representation of a previous sensory experience. This is also not what mental images are. The literature sometimes implies that array-based representations must only be used to model eidetic images. Although I will be using a structured propositional representation of mental images, I would nonetheless admit the possibility of constructing array-based, or other non-propositional, representations to model mental images rather than "vivid" eidetic images (e.g., Paivio 1971; Shepard 1981; Kosslyn 1980, 1983; Kosslyn *et al.* 1984; Farah 1984).

5. *Structural equivalence.* "The structure of mental images corresponds to that of actual perceived objects, in the sense that the structure is coherent, well organized, and can be reorganized and reinterpreted." (p. 120)

I say "broadly interpreted" because the last three principles can be applied to propositional representations of the sort used in this work, although that is not how these principles are usually construed. The paragraphs below discuss the issue of how "spatial" a representation is.

Preserving spatial properties in representations. To encode vivid or semi-vivid spatial images, the representation must ensure that certain important spatial properties are preserved, such as topological proximity. In general there are two approaches: *spatial representations* that intrinsically preserve these properties, and *propositional representations* that explicitly assert the requisite spatial properties.

Although the proposed model does not employ spatial representations, they are frequently used for mental image models. Spatial representations are intrinsically structured in the same way as the real-world situations they represent, and because of this, they inherently preserve spatial properties. The most common type of spatially-structured representation are array-like pixel matrix representations. On the other hand, alternatives to matrix coordinates can be useful. For example, polar coordinates have the advantage of concentrating the most pixels around the origin which can be positioned at the focal point. Thus a number of different, related representations have been proposed (e.g., Shepard 1981; Kosslyn 1980, 1983; Farah 1984). Pinker (1984) suggests that the critical characteristic of all spatially-structured representations is that they (should) satisfy the axioms of metric spaces. The metric axioms are:

> (a) the distance between a point and itself is less than the distance between a point and any other point; (b) the distance between point a and point b is the same as the distance between point b and point a; (c) the distance between point a and point b plus the distance between point b and point c must be greater than or equal to the distance between point a and point c. (p. 40)

Note that (b) is the symmetry axiom and (c) is the transitivity axiom applied to distances.

Spatial representations have a property that is often useful: because their representations are spatially structured, they inherently ensure that only concepts that are consistent with the real world are representable. For example, using a pixel matrix there is no way to represent a square that has only three sides, whereas the same concept is easily coded propositionally (as indeed this sentence just did!). Borrowing Carnap's terms, a spatial representation always preserves analytic truths (concerning space) whereas a propositional representation permits synthetic statements that may be analytically false. A propositional representation must be explicitly prevented by external axioms from allowing ill-formed spatial descriptions. In any spatial representation that satisfies the metric axioms, spatial properties are intrinsic and require no external consistency checks. Thus ultimately it would not be surprising to discover that humans employ both spatial and propositional representations.

As mentioned, if mental images are represented propositionally (as in the proposed model) then intrinsic spatial properties must be marked using some artificial means. One way to do this is to make use of the standard "terminological" versus "assertional" distinction first

proposed in work on KL-ONE style representations (Brachman *et al.* 1983; Vilain 1985).[8] We mark as "terminological" the fact that a square has four sides so that any instantiation of the square concept is automatically constrained by the interpreter to have exactly four sides. In contrast the fact that squares are Mr. Kien's favorite shape is an "assertional" property of squares since a square is still a square regardless of Mr. Kien's inclinations. The interpreter allows such properties to be negated.

KL-ONE theorists sometimes have trouble distinguishing terminological and assertional properties. This problem also arises in trying to say which properties are intrinsic to spatial concepts. I attribute this to the investigator's ability to rapidly and effortlessly map between spatial and propositional representations. So if asked "Is a three-sided square a possible concept?" we answer "No" when using a visuospatial representation and "Yes" when using a propositional one.

Propositional representations for mental images. It is not entirely clear how mental images ought best to be represented. Iconic images are generally modelled using spatial representations, but for purposes of representing mental images these are subject to Kant's[9] objection that they are too specific and cannot capture, for example, the general notion of a triangle that encompasses both acute and obtuse triangles. It seems that mental images should be abstract (and non-vivid) enough to enable conceptualizing a triangle without committing to one or the other subclass.

Similarity-based representations have the property that similar concepts like acute and obtuse triangles are actually represented by similar codes, for example by activation vectors with low Euclidean distance. The extreme case of similarity-based representations degenerates back into spatial representations, but normally by saying that two concepts are similar one is implicitly abstracting away differences on some dimension, e.g., the angles at a triangle's vertices. In the context of mental images Shepard & Chipman (1970) proposed a version of the similarity-based principle called "second-order isomorphism".

The proposed model uses propositional representations, which fail to capture the pictorial quality of visual images. To compensate I suggest that people also have an array store with a spatial representation, and different array capacities account for the fact that people have varying degrees of eidetic visualization ability. This is not a radical move; imagery theories such as Kosslyn's (1983) contain both array and propositional representations.

Most measurable properties of mental images are related to retrieval speed. However, it would be a mistake to use retrieval speed data as evidence either for or against propositional representations. Since propositional representations are entirely neutral with respect to control, Anderson (1978) observed that mental image effects can be obtained from propositional models simply by using the correct retrieval strategy. By the same token, on the other hand, when using a propositional representation one must be careful to avoid the pitfall of implicitly assuming one's own favorite control strategy.

Moreover in the extreme case, a propositional notation does not even prevent one from using it to store pixel-structure information. A pixel structure can straightforwardly be coded as a set of propositions where a variable is assigned to each pixel. Clearly there is no advantage

[8] Using the terminological/assertional distinction this way is subtly different from the way it is used in most KL-ONE systems. This point is discussed in section 6.4.1.

[9] *The Critique of Pure Reason.*

over a dedicated pixel representation if one's imagery theory only uses spatial representations. If, however, one wished to construct a model that employed both complex propositional semantics and spatial representations, this would be one way to handle both uniformly.

In the proposed model, the terminological-assertional mechanism described above is used to enforce spatial consistency conditions. Compositional transitivity is enforced by the semantics of the representation language (not natural language!). For example, figure 4.3 shows the encoding of the image of a *desk drawer handle*[10], where transitivity guarantees that the handle is a part of the desk as well as the drawer. The metric axioms, including transitivity of distances, are not currently built into the primitive semantics, and if needed, must be enforced by making use of compositional transitivity.

$$\left[\begin{array}{l}\text{TYPE: } \textit{desk} \\ \text{COMPONENT: } \left[\begin{array}{l}\text{TYPE: } \textit{drawer} \\ \text{COMPONENT: } [\text{TYPE: } \textit{handle}]\end{array}\right]\end{array}\right]$$

Figure 4.3: The encoding of the mental image for *desk drawer handle* enforces compositional transitivity.

Spatial versus visual images. The difference between spatial and visual processing is still a matter of debate; some argue that most visual imagery experiments can be interpreted in terms of spatial image processing (Neisser & Kerr 1973; Kerr *et al.* 1985). On the other hand, some recent neurological studies describe patients who are able to generate mental images of spatial locations but not visual appearances, and vice versa (Levine *et al.* 1985; Farah *et al.* 1990). Even if spatial and visual mental image processing occurs in separate modules it may well be that many tasks can be performed using either type of processing. The evidence is too sparse to warrant strong assumptions and therefore I will often use the term "visuospatial" to indicate neutrality.

Spatiotemporal images and motion. Real world events occur over time intervals, and impinge upon our sensory registers over time. Humans learn not only static patterns, but also pattern sequences. Unfortunately, at this time I have no dynamic representation of temporal pattern sequences. Models of temporal sequence learning that employ recurrent neural networks are being investigated (Elman 1989, 1990, 1991; Pollack 1988, 1989, 1990; Rumelhart 1991; Schmidhuber 1991), but too little is known as yet of either the mathematical theory or neurological workings to warrant inclusion of any particular model here.[11]

4.3.2 Arguments for Mental Image Semantics

For some time cognitive linguists have been appealing to image-like schemas to explain prepositional and case phenomena (Talmy 1983, 1985, 1988; Lakoff 1987b; Langacker 1987; Jackendoff 1972, 1983). Unfortunately it is not always clear whether the spatial schemas are an actual claim about cognitive representation or merely a convenient research meta-language. Here I am putting forth a more specific hypothesis about representation and cognitive process, one tenet of

[10] From Warren (1978, p. 127).

[11] Regier deals with motion by converting dynamic sequences into static representations. See section 5.2.2, p. 104.

which is that understanding a linguistic input often automatically creates a mental image. Moreover, though it may be intuitively obvious that reasoning about the physical world is often more "picture-like" than verbal, it is still necessary to show that there are corresponding modularities in the cognitive processes. Below, we consider some types of evidence that have been put forth in favor of modularizing mental image representations.

Psychological evidence. An experiment by Brooks (1968) showed that mental image processing is more closely associated with visual processing than verbal processing, by using an interference technique. Subjects were given two similar tasks, with verbal input versus imagery input. In the verbal task, a sentence like "A bird in the hand is not in the bush" is given and the subject is to indicate serially whether each word is a concrete noun: "no, yes, no, no, yes, no, no, no, no, yes". In the imagery task, the subject is to imagine a (previously seen) block letter such as that in figure 4.4 and indicate serially whether each corner is at either the top or bottom: "yes, yes, yes, no, no, no, no, no, no, yes". Both visuospatial and verbal manners of response were tested, for both tasks. The subjects were told to either to point to a response sheet with staggered rows of the letters Y and N, or to say the words "yes" and "no". As figure 4.4 shows, visuospatial responses were faster for the verbal task while verbal responses were faster for the imagery task. Thus Brooks concluded that visuospatial processes interfere with mental imagery more than the verbal processes.

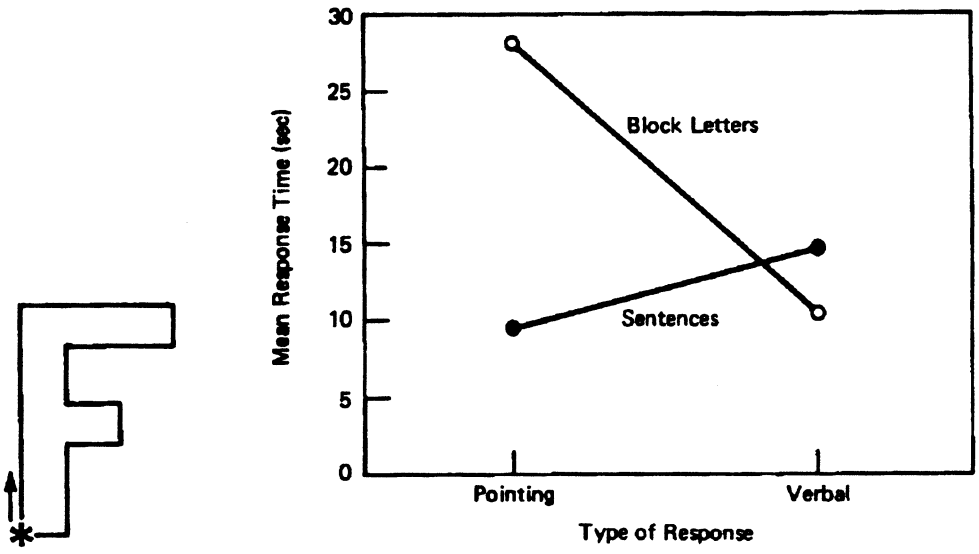


Figure 4.4: Brooks (1968), from Finke (1989).

In her study of prototype effects, Rosch (1975) found that priming effects required less time for visual tasks than verbal tasks. Subjects heard a priming word (*furniture, vehicle, vegetable*) and some interval later were presented with either a picture pair or a word pair. The subject was to answer "yes" if the pictures/words came from the same category or were identical, and "no" otherwise. Rosch found that the priming word facilitated faster response times when the two pictures/words were identical and came from the named category. Moreover, the interval between the priming stimulus and the picture/word pair could be shorter for picture pairs (200 msec.) than

word pairs (at least 300 msec.). Rosch therefore suggested that the internal cognitive interpretation of the priming words is closer to mental images than verbal or linguistic forms.

In a related experiment Potter & Faulconer (1975) discovered that given a priming word naming a superordinate category, subjects were faster at deciding whether a subsequent stimulus belonged to that category when the subsequent stimulus was a picture rather than a spoken word. For example given the word *fruit* subjects could determine more quickly that a picture of an apple, rather than the word *apple*, designated a kind of fruit. This also led Potter and Faulconer to conclude that at least the meaning of words like *fruit* is closer to an image-like representation.

Other evidence for the connection between visuospatial processing and imagery comes from a series of experiments on the *symbolic distance effect* (Moyer 1973; Paivio 1975; Kosslyn 1975, 1976; Moyer & Bayer 1976; Holyoak 1977). Subjects are slower at identifying which of two objects named (e.g., animals) is larger when the two objects are similar in size (beaver, raccoon) than when they are very different (beaver, squirrel). The response speed is inversely proportional to the relative size difference and the effect is invariant with respect to many other parameters. This effect is easy to explain in terms of visuospatial processing but difficult to explain in terms of verbal processing.

Neurological evidence. Neurological evidence for the modularity of mental image processing comes primarily from studies on brain damage. No modular breakdown has clearly emerged and it appears that the data cannot be adequately captured by any simple model. Modularized semantic systems are therefore still controversial, but because of the weaknesses of unified semantic systems new modular models continue to be proposed. Moreover the advances in neural and distributed modelling may provide the means to resolve the problems with modular semantic systems. For an in-depth survey of neurological modularity issues, the reader is referred to Farah (1990), from which the summary here is derived.

To illustrate the difficulty with unified, non-modular semantic models consider the examples in figures 4.5–4.8 based on past proposals to account for optic aphasia (Ratcliff & Newcombe 1982; Riddoch & Humphreys 1987; Farah 1990). Optic aphasics can name objects described verbally and can recognize visually perceived objects (demonstrated by gestures indicating their use), but cannot name visually perceived objects. Though each of the illustrated models were proposed to explain this apparently paradoxical condition, none of them resolve the paradox satisfactorily. If there is only one route from visual input to semantics (figure 4.5) then patients with that route broken should not be able to visually recognize the function of objects. Yet if any unbroken route remains (figures 4.6, 4.7) then the semantic level ought to be able to categorize the object sufficiently well to name it, since being able to recognize its function requires a large amount of semantic knowledge. Optic aphasics cannot do this even given an indefinite amount of time. The idea that two separate damage loci are needed to explain optic aphasia (figure 4.8) obliges one to assume that the conjunction of two partially damaged loci causes worse effects than one would predict by serially concatenating the effects. Farah observes that superadditive impairments do occur in many neural network models, which can restore degraded patterns but fail when the signal is too heavily damaged. However from the computational point of view, this analogy unrealistically assumes that the cognitive system is a flat neural network that operates in one massively parallel shot. In fact it is more plausible that the semantic system can restore degraded input from a damaged visual system before passing on to the naming task, in which case there should be no

superadditive effect.

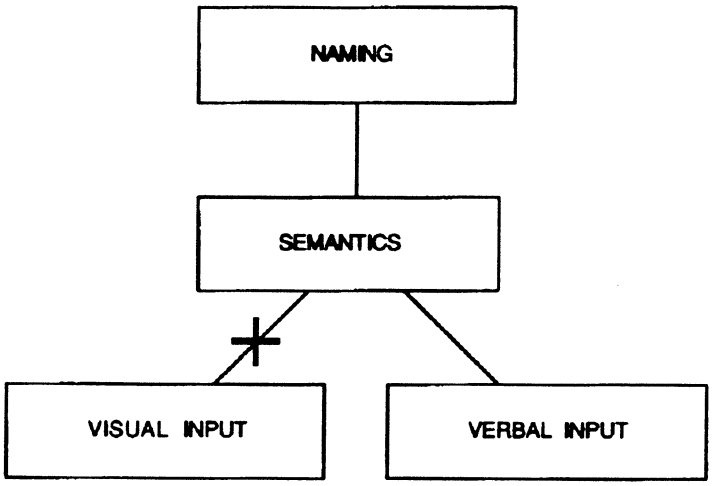


Figure 4.5: Only one route from visual input to semantics.

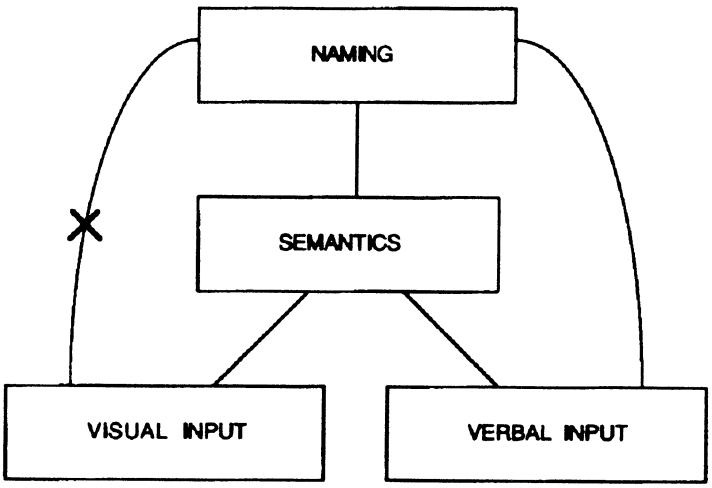


Figure 4.6: Ratcliff & Newcombe (1982), from Farah (1990).

Several interesting modularized semantic models have been proposed. Beauvois (1982) proposed distinguishing visual and verbal semantics as in figure 4.9. This is a version of Paivio's dual coding model, visual semantics being "imagens" and verbal semantics being "logogens". Farah (1990) objects that for this model to account for optic aphasics' ability to pantomime the use of objects, redundant conceptual knowledge would be required in the visual semantics module. However, the type of knowledge that is required for gesturing is strongly associated with spatiotemporal abilities, and it is perfectly plausible that the gesturing abilities derive from an organization where just parts of the conceptual system dealing with spatiotemporal knowledge are replicated. This modularization, highlighted in figure 4.10, is used in my proposed model, where

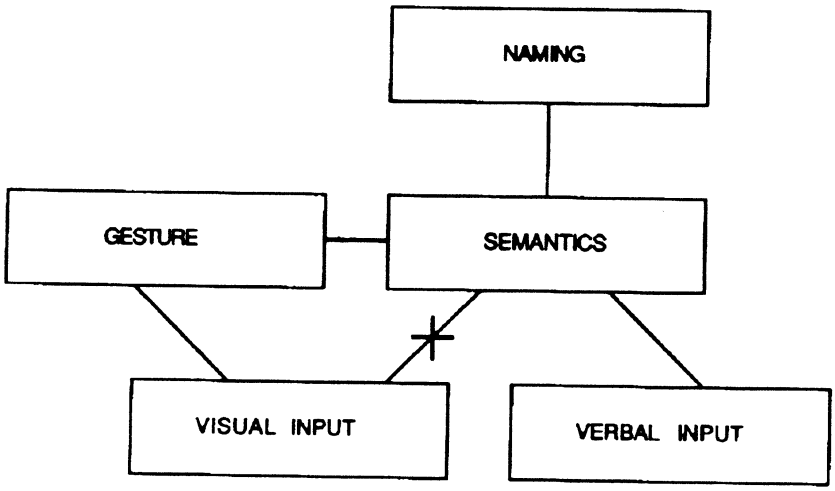


Figure 4.7: Riddoch & Humphreys (1987), from Farah (1990).

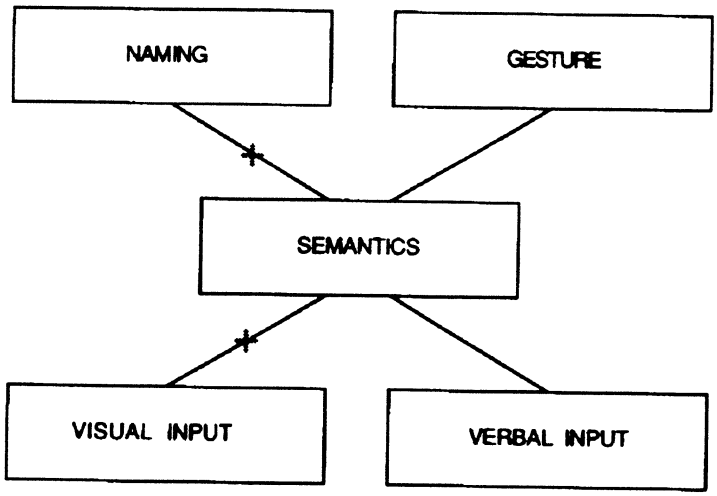


Figure 4.8: Farah (1990).

spatiotemporal knowledge in the lexical semantics and conceptual modules are image schemas (section 5.2.2). Not only is a certain degree of redundancy neurally plausible, but there can also be advantages from the standpoint of computational efficiency, because localizing concepts that are frequently associated (e.g., gestures and other spatiotemporal scenes) is a form of "caching" that speeds retrieval.

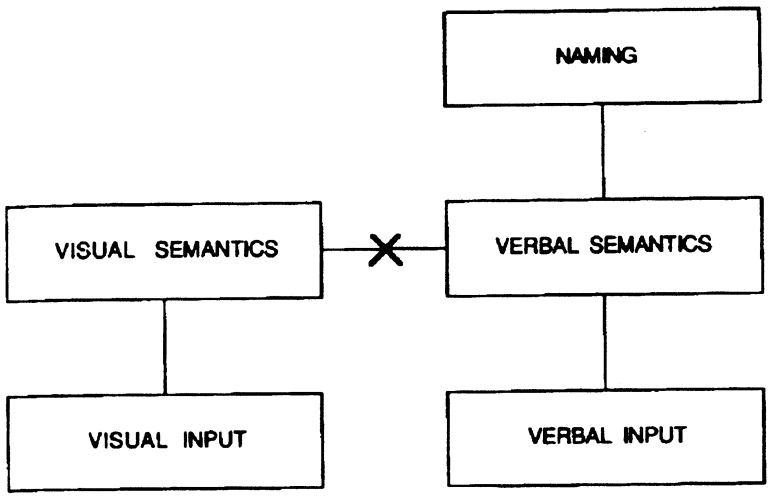


Figure 4.9: Beauvois (1982), from Farah (1990).

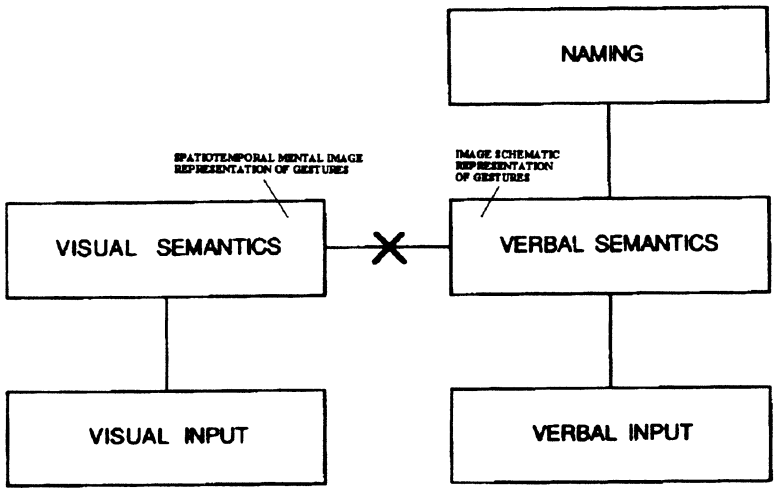


Figure 4.10: Distributing "redundant" spatiotemporal knowledge to account for optic aphasia in Beauvois's model.

An even more radical approach along these lines is proposed by Coslett & Saffran (1989). Figure 4.11 shows the architecture, which differentiates semantic modules by left and right hemisphere. In this model a loose correspondence is assumed between hemisphere and visual/verbal

knowledge and processes, so that one can view it as a relaxation of Beauvois' model. Right hemisphere semantics is coarser than left (Zaidel 1985; Kosslyn *et al.* 1985) and tends to be more directly associated with visual processing; left hemisphere semantics is more closely tied to verbal processing.[12]

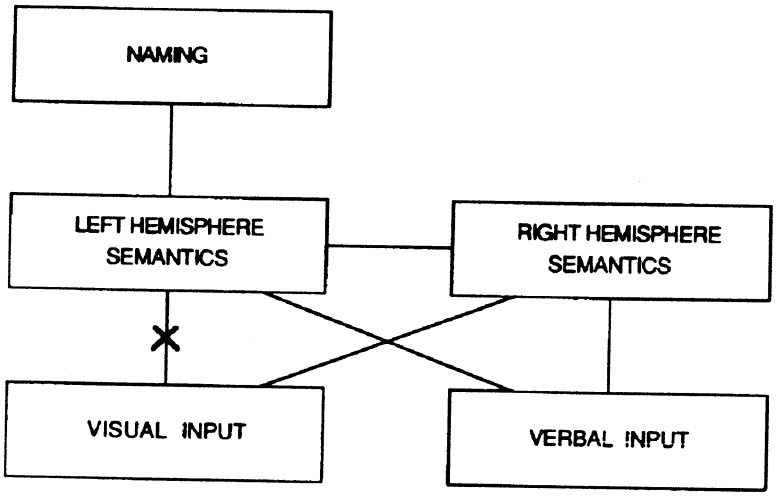


Figure 4.11: Coslett & Saffran (1989), from Farah (1990).

Neurological researchers sometimes interpret findings such as these to mean that non-propositional representations are used for visual mental images since the relevant loci are primarily part of the visual system. Whether this is a valid argument all depends on what one means a proposition to denote; it was noted in section 4.3.1 that propositions could even be used to describe the early visual system, though it would usually not be felicitous to do so.

Spatial properties can be exploited for computation. There is also a computational efficiency argument to be made for the existence of a mental image module in humans. According to the spatial equivalence principle, mental images are encoded using representations that preserve spatial properties. The computational advantage comes if humans use a representation that is *intrinsically* spatial. As discussed in section 4.3.1, spatial representations have the property of only being able to represent concepts that are consistent with real-world spatial properties. Since this eliminates the need for constant consistency checking, any tasks involving spatial reasoning can be more efficiently processed.

Spatial representations may be useful for non-spatial reasoning domains as well, when reasoning by metaphor or analogy. Reasoning in abstract domains that can be mapped to a spatial schema can be speeded up by mapping invariant properties (transitivity, dimensionality) to spatial properties that are intrinsic, again eliminating the processing overhead of enforcing those constraints.

Unfortunately, the propositional representation used in the proposed model is too abstract to enforce intrinsic spatial properties. Instead, explicit meta-level terminological definitions

[12]Interestingly enough, Farah (1984, 1988) found in a survey of patients with brain damage that the process of *generating* images tends to be localized in the posterior *left* hemisphere.

must be used to achieve this effect. It is a weakness of the model that consistency must be enforced by the underlying propositional logic interpreter. The consistency checking must be explicitly excluded from the intended mapping from the model to the human cognitive architecture.

4.4 Lexical Semantics

For reasons mostly having to do with accounting for syntax but also having to do with language acquisition and learnability, a number of researchers have proposed lexically-based theories of semantics. Usually the theories postulate an intermediate level of representation between accepted syntactic categories and pure AI-style conceptual representations. The present linguistic tradition of lexical semantics derives primarily from the work of Fillmore (1968, 1977) and Gruber (1965), who argued for *case relations* and *thematic relations*, respectively, as the primitive structural elements for verb arguments.[13] I will follow Dowty (1989) in using the general term *thematic role*. After 25 years there are still as many variant theories as investigators, yet even work outside the tradition has slowly converged toward a remarkably similar paradigm; broadly construed, the term lexical semantics might be applied to work as diverse as Logical Form and KL-ONE. Within the tradition there are several major lines we can identify.

It is important at the outset to note which of two possible metatheoretical motivations for pursuing lexical semantics I have in mind. There are two schools of research that go by the rubric "lexical semantics", but whose methodological goals are entirely different in character. One motivation is theoretical neutrality: language being the most accessible and observable function of the mind, organizing semantics around the lexicon in some sense makes the least commitment to any cognitive theory. In particular, those who are interested in a highly descriptive style of linguistics have created a strain of atheoretic lexical semantics (Cruse 1986).

The other motivation is, as mentioned above, to account for morphosyntactic phenomena that established morphosyntactic categories fail to explain, but without resorting to a complete dependence on the conceptual system. Current data supports the working hypothesis that many generalizations can be captured at the intermediate lexical semantics level. This is the more usual motivation, and it is also the motivation for including a lexical semantics module here (albeit in a slightly different, information-theoretic sense of "accounting for syntax"). Lexical semantics in this tradition is highly committed to theory and relies on linguistic and psychological data for justification of the structural and ontological elements to be included.

4.4.1 The Boundaries of Lexical Semantics

Syntax versus lexical semantics. It is no surprise that there is widespread disagreement as to the proper lines between syntax and lexical semantics, and between lexical semantics and the conceptual system. With regard to the division between syntax and lexical semantics, Jackendoff (1990) argues emphatically that

> The fundamental point, from which all else proceeds, is that *thematic roles are part of the level of conceptual structure, not part of syntax* [italics in original]. Recall Gruber's (1965)

[13] Similar ideas have been traced back as far as 350 B.C., to the Sanskrit grammarian Pāṇini's *kāraka* theory (Ananthanarayana 1970; Singh 1974; Somers 1987).

> intuitive definition of Theme: the object in motion or being located ... thematic roles are nothing but particular structural configurations in conceptual structure; the names for them are just convenient memonics [sic] for particularly prominent configurations. (pp. 46–7)

On the other hand, it is unclear that categories like "count noun" and "patient" are sufficiently unlike categories like "plural noun" and "accusative" to make it worthwhile to promote lexical semantics as a distinct module from syntax. Taking this view to the extreme, the division is purely historical and syntax should simply incorporate lexical semantics. Interestingly enough, Gentner (1988) argues that verb meaning and grammar may be neurally localized in the same region, since agrammatic aphasics not only suffer impaired syntactic construction use but often also have trouble naming verbs, which, she hypothesizes, reflects damage to verb meaning representations. The *Logical Form* or LF module (figure 4.12) that grew out of Chomsky (1965) represents quantification, scope, and reference information in a form suggesting truth-conditional logic, which seems semantic in nature, and indeed, many semanticians have followed up on this by attempting to explicate the truth conditions underlying LF. Yet by definition LF is a level of syntactic representation since formal transformation rules relate it to the other syntactic modules, and since its purpose is solely to facilitate accounting for syntactically allowable structures.[14]

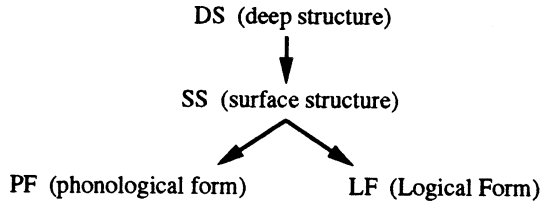


Figure 4.12: Standard placement of Logical Form in grammar.

However, I am proceeding on the hypothesis that there will prove to be a division in the weak sense of the CMI principle. Moreover, since the most complete semantic theories are currently cast within the lexical semantics framework, it is convenient, for expository reasons if nothing else, to retain the distinction between syntax and lexical semantics.

Lexical versus conceptual semantics. The division between lexical semantics and the conceptual system is likewise blurred. Selectional restrictions can require arbitrarily deep conceptual knowledge.

[14] Note that *logical form* (without capitals) is used by semanticians and philosophers to refer not to LF, but to a realist truth-conditional semantic representation as in Montague semantics; see for example Lycan (1984). In fact, this can be taken a step further since "logical form" need not even be restricted to realist representations. The term "logical" describes a type of representation language rather than the content that can be represented using that language. Some aspects of ordinary predicate logic make it well suited as a realist representation for handling phenomena such as quantifier scope. Other aspects are less desirable, especially when looking at associative effects and reflexive automatic inference, since the psychological evidence shows humans to be poor at understanding any but the simplest of logical constructs. Logics can also be used to construct mentalist representations where the only predications made concern an agent's state of mind (indeed, this is my approach). Discussions of the relationship between LF and logical form (in the realist, truth-conditional sense) are found in Chierchia & McConnell-Ginet (1990) from the semantician's perspective, and May (1985) from the syntactician's perspective.

McCawley (1968) argues that *devein* only applies to shrimp, *assassinate* only to political figures, and *diagonalize* only to matrices. Unfortunately, it is probably the case that there is no *a priori* way to absolutely define the elements of the lexical semantics module, so that it is guaranteed to conform to the CMI principle. It is only possible to use linguistic and psychological data as evidence for heuristically demarcating the boundaries. I first consider here some of the criteria that have been proposed for distinguishing lexical from conceptual semantics, but which are unsatisfactory for this purpose.

1. *Equating verb meanings to single logical predicates.* This approach to defining lexical semantics requires that all predicates used in the lexical semantics module correspond one-to-one with lexemes. This form of semantics is non-reductionist in the sense that one cannot define "smaller" predicates to capture common aspects between different lexemes, unless those predicates happen to correspond exactly to some other lexeme. Otherwise, commonalities are only captured by using the same case or thematic roles with different lexemes. The motivation behind pursuing such an approach lies more in describing the semantics of particular lexemes; there is no reason to expect the approach to be flexible enough to capture all pertinent generalizations about morphosyntactic variation.

2. *Requiring one-case-role-per-argument.* A stricter variant of the above approach also restricts the predicate arguments to one-to-one correspondence with case roles. The one-case-role-per-argument constraint is the most important feature distinguishing Fillmore's (1968) case grammar from Gruber's (1965) thematic relations, which permits arguments to fill an arbitrary number of roles. The constraint proved to be too strong, and is relaxed in a number of different ways in subsequent case grammar work, including Somers's (1987) case-grid system which is a major influence in the present proposal. Although I attempt whenever possible to minimize the number of case roles per argument, the principle is contravened by other representational flexibility needs. When, as with Gruber, an argument can have multiple roles, the set of all different role *combinations* effectively forms an abstraction hierarchy of composite roles; for example, the single $(AGENT)$ role is an abstract ancestor of the composite role $(AGENT, PATIENT)$, which could be used to characterize a reflexive action. In the original case grammar, case roles are required to form a flat set rather than a hierarchy, with the goal of minimizing the number of distinct case roles needed to handle the syntactic phenomena.

3. *Including all automatic processes.* As already discussed, a distinction based on automaticity (Fodor 1983) does not fit empirical results (Marslen-Wilson & Tyler 1987). People are equally fast at inferences extending beyond the lexical level into discourse and real-world knowledge.

4. *Defining lexical semantics as a pipeline stage.* In some approaches the distinction between lexical semantics and the conceptual system lies between processing stages in the interpretation pipeline. For example, Allen's (1987) "logical form" (a mix of case grammar and logical quantifiers, closer to LF than the semantic-philosophical sense) is conceived of as an intermediate stage where certain types of case role and scoping interpretation are performed, but prior to resolving anaphoric reference and ellipsis ambiguities, or any kind of stereotype-, script-, or plan-based interpretation. Here logical form has the advantage that it is considered to be an intermediate result rather than the full meaning; further inferences can be made upon

the logical form. Again, the drawback is that the evidence indicates automatic processing extends to discourse and real-world knowledge. The particular choice of constraints on the language of logical form cannot be justified empirically.

5. *Restricting lexical semantics to literal forms.* Another pipeline-based distinction involves literal versus figurative or metaphoric forms (Nirenburg & Levin 1991). The lexical semantics stage comes first, and is responsible for parsing the surface grammatical form into a largely uninterpreted result; for example, the result for *posting growth* leaves the metaphor uninterpreted. The conceptual stage follows and is to produce a fully interpreted canonical form, which for *posting growth* is some sort of *increase* concept. The problem is not with the idea that the purpose of lexical semantics primitives is to capture surface grammatical variation. However, the assumption that processing should take place in corresponding stages is unwarranted, especially given cognitive propensities toward automatization. With bound phrases like *posting growth*, the intended conventional interpretation should be retrieved without necessarily first requiring a compositional interpretation. Even for many novel metaphors, humans appear to perform interpretation quickly by following conventional (and therefore, automatized) patterns (Martin 1988, 1990, 1991).

6. *Including only compositional-monotonic semantics.* Another distinction (Bierwisch & Lang 1989; Herweg 1991; Lang *et al.* 1991) is that lexical semantics is syntax-driven and compositional. Because whether an interpreter is compositional is such a notoriously vague condition (see Partee 1984), we might try to take a slightly more rigid and concrete interpretation of the distinction, namely that inference at the lexical semantics level combines structures purely monotonically—that is, with no backtracking—while conceptual inference permits nonmonotonic composition decisions. Still, it is not clear that this condition has much discriminatory power since it is so dependent upon the particular theorem prover and search ordering rules used, and since the monotonic composition criterion can be circumvented by building structures in parallel, like a chart parser does.

7. *Varying notation.* Sometimes commonalities that may exist between lexical and conceptual modules are obscured by different notations used at each level. For example, Allen's (1987) aforementioned logical form employs a mix of relatively standard constructs from case grammar and quantification theories, whereas the stages that follow employ a variety of other AI knowledge representation formalisms. While there are practical engineering reasons for interfacing a standard linguistic theory to a standard AI theory, it can be difficult to tell whether a theoretically meaningful distinction is being drawn.

Pinker's Grammatically Relevant Subsystem hypothesis. Among recent approaches to lexical semantics, Pinker's (1989) appears to be most promising for the purpose of constructing a lexical semantics module that is likely to approximate the CMI principle. The guiding principle of Pinker's approach is to minimize the specificity of the semantic features and roles, subject to the constraint of handling morphosyntactic variation (but not lexical variation). In other words, what belongs in the lexical semantics module is all and only ontological distinctions that help explain phenomena at the morphological and syntactic surface-level. I regard this as a heuristic, because I do not feel a sharp line can be drawn between knowledge that can and cannot influence syntax, but for cognitive

processing motivations it is nonetheless advantageous to modularize knowledge, and then allow intermodular interactions to influence syntax.

My approach permits arbitrarily complex semantic structures, thus following Jackendoff (1972, 1983, 1990) and Pinker (1989). No one-case-per-argument constraint is enforced, though the use of a case-grid type of thematic role system is an attempt to minimize the ratio of roles to arguments. Jackendoff and Pinker deal with surface cases by using a set of linking rules to map surface cases to open arguments in the complex semantic structures; a similar function can be performed by signification relations in the Construction Grammar framework I use. Pinker proposes the following "Grammatically Relevant Subsystem" hypothesis in the context of his study on verb argument structure acquisition:

> Perhaps there is a set of semantic elements and relations that is much smaller than the set of cognitively available and culturally salient distinctions, and verb meanings are organized around them. Linguistic processes, including the productive lexical rules that extend verbs to new argument structures, would be sensitive only to parts of semantic representations whose elements are members of this set. The set would consist of symbols that have cognitive content, such as "causation" and "location," but not all cognitively meaningful concepts are members of this privileged semantic machinery. (p. 166)

In this kind of approach the distinction between lexicosemantic and conceptual primitives merely boils down to a matter of how specific the representational features are. The artificial intelligence tradition simply employs some of the conceptual primitives for specifying lexicosemantic patterns or rules, with the object of including only the least specific features necessary to account for syntax. Though Pinker denies that the primitives of the lexical semantic system are necessarily primitives of the conceptual system, the thrust of his argument is just to distinguish lexical semantics as a part of the cognitive system with a special function in acquisition, and nothing in the argument actually prevents the conceptual system from embracing lexical semantics as one of its components. Thus Jackendoff (1990) states:

> it certainly turns out that only limited aspects of conceptual structure interact with syntax. This might be seen as motivation for an independent level of θ-structure that encodes only a subset of conceptual information. But there is an alternative account that requires no extra level of representation: one can incorporate the constraints directly into the correspondence rule component. ... As far as I can see without detailed examples, the constraints on the theory and the need for stipulation are exactly the same in either case, and the latter treatment makes do with one fewer level of representation. (p. 49)

Although this approach means that cases and slots are essentially the same, Charniak's (1981) well-known analysis of "case-slot identity" does not hold because verbs are not identified one-to-one with predicates (frames). Charniak observes a number of theoretical implications that arise as a consequence of either using the same role types for both verb frames and conceptual frames (the "minimal hypothesis"), or using exactly the same frames as both verb and conceptual frames (the "maximal hypothesis"). However, as Charniak writes,

> this theory rests on an assumption that underlying most verbs in English is a frame which captures most of the verb's meaning. If this is not the case, but rather a verb plus arguments is represented by a large number of complex statements (as in, say, conceptual dependency theory . . .), then we will again have lost the underlying structure necessary for the theory. (p. 291)

Indeed, verb frames are represented here by structures composed of conceptual primitives that are not necessarily themselves verb frames. Thus the use of thematic roles instead of strict case roles eliminates the case-slot identity difficulties.

I also disagree with Pinker's belief that the sort of distinctions made at the lexical semantics level are insignificant with respect to cognitive categorization, though not because of the question as to whether lexical semantics belongs to the conceptual system (to say that lexical semantics does or does not belong to the conceptual system is just a terminological quibble). The important point is that even if one excludes lexical semantics from the conceptual system proper, it *is* still the case that the lexical semantics representation of an utterance is part of our conceptualization of the situation. How we conceptualize a situation is at least as important to cognition as the situation itself, since we have many alternate ways of conceptualizing the same situation, upon which our responses depend. In other words, we cannot exclude any level of semantics—be it lexical semantics, mental images, detailed conceptual structures—from the "meaning" of an utterance.

4.4.2 Arguments for Lexical Semantics

Linguistic arguments for lexical semantics. Nearly all linguistic arguments for lexical semantics fall into one of two classes. First, the sorts of primitive roles posited in most lexical semantics proposals surface as closed-class morphemes in one or more languages. Since closed-class morphemes have highly restricted functions, this argues for the primitive status of those functions. It is then highly plausible that those same functions are employed wherever possible for verb complements in the same language. Talmy's (1985) survey of semantic categories, from which are derived many of the proposed lexical semantics primitives, showed a remarkable overlap in the kinds of semantic relations that can or cannot be realized both as closed-class "satellite" surface forms, or as incorporated parts of a verb root's meaning. Moreover, when the same roles occur in multiple languages either as closed-class morphemes or in verb argument structures, as is often the case in Talmy's study, it suggests the roles are universal.

Second, the same roles also account for surface grammatical variation in the argument structures over a wide range of verbs, both within and across languages. The case grammar and thematic relation theories discussed above, as well as Talmy's studies, are largely motivated by the need to account for these generalizations.

Psycholinguistic evidence for lexical semantics. Pinker (1989) argues that the verb usage errors made by children in the standard developmental sequence are explained by incorrectly constructed semantic structures, and supports the lexicosemantic module hypothesis by accounting for overregularization phenomena in a wide range of examples involving argument structure alternations in dative, causative, locative, and passive constructions. Such errors involve verbs used in the

correct sense, but with incorrect argument structures; similar errors have also been studied for past tense acquisition (Bybee & Slobin 1982; Rumelhart & McClelland 1986). The three stages are:

1. *Conservative usage.* Initially, children learn the correct usage of constructions involving particular verbs, for example, causatives like *open, break, wet,* and *hurt* (Bowerman 1974, 1982). There is no productive use of the constructions.

2. *Overregularization.* This stage, depending on the child and the type of error, typically occurs within the range from ages 2;0 to 5;5–9;3. Children apply constructions productively to verbs that in adult usage do not accept particular alternations. For example, the causative is overextended in *I'm gonna just fall this on her* and, in a double-object construction, in *Will you have me a lesson?* (Bowerman 1982).

3. *Adult usage.* Eventually overregularization errors drop out, particularly when alternate lexemes (for example, causatives like *knock down* rather than *fall*) are learned.

Pinker posits "broad-range" and "narrow-range" rules that link syntactic forms to semantic structures; these rules are sensitive only to the lexicosemantic level, and not the general conceptual system. He is able to explain the overregularization stage by having the broad-range rules already learned, but the narrow-range rules not yet learned.

Pinker also argues that the adult verb usage data suggest verb meanings are stored and processed as divisible assemblies. Gentner (1981) surveys a number of differences between noun and verb usage including (1) the relative difficulty of remembering the particular verb that was used, (2) frequent change of verb in paraphrasing tasks, and (3) frequent change of verb in double translation tasks. Pinker interprets these findings to indicate that verbs are represented not as cohesive gestalts but as structures that can lose or gain elements.

Learnability arguments for lexical semantics. Pinker (1989) argues for modularization on the basis that it makes more plausible assumptions about child acquisition of verb argument structure. He criticizes the model in figure 4.13, arguing that it requires assuming that (1) children can accurately encode the adult's intended meaning from context alone, (2) a special explanation is needed for acquisition of languages where correlations between syntax and semantics are different, and (3) since languages are full of subregular patterns and rules that hold only part of the time, either parents must filter out the violations (passives, deverbal nouns) or children filter them out using external criteria. Instead, Pinker proposes the model in figure 4.14 which he argues makes simpler and more reasonable linguistic and psychological assumptions: (1) syntax and semantics are related by formal grammatical linking principles that may be universal and fully regular, and (2) parents need not filter their speech, and need merely use semantic structures that the child shares by virtue of common context.[15]

[15]Note that in its pure form, Pinker's argument depends on the assumption that lexical semantics is innate, as discussed below. However, I believe a relaxed version of the argument could be formulated, where lexical semantics evolves rather than being innate, but still facilitates acquisition. Consider a purely intuitive analogy to a back propagation network. Inserting an intermediate hidden layer of the correct restricted size can bias the network to generalize much more accurately, even though the hidden layer weights are not preset but evolve.

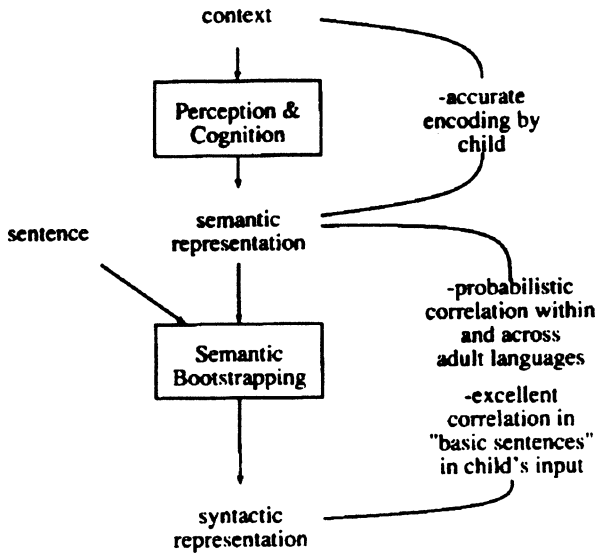


Figure 4.13: Pinker (1989).

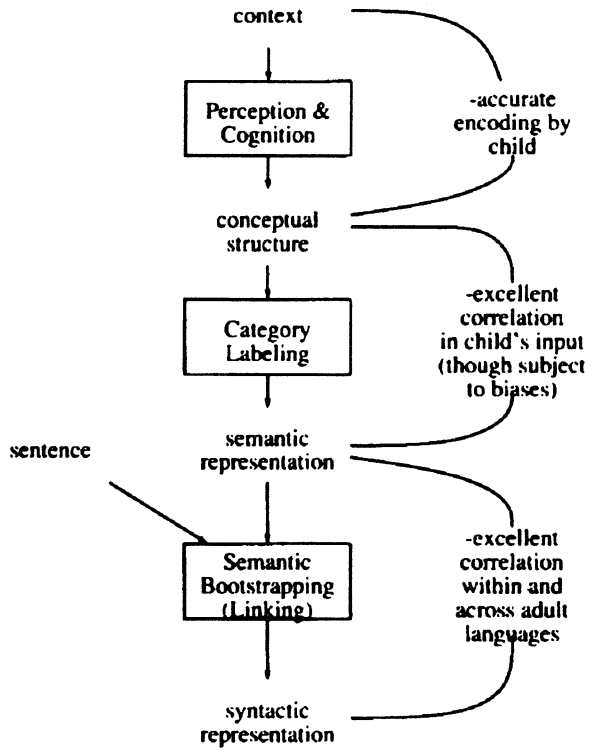


Figure 4.14: Pinker (1989).

Processing motivation for lexical semantics. Aside from arguments that it facilitates verbal argument acquisition, an argument can also be made that intermediate lexical semantics module facilitates faster adult language processing. Suppose that module C contains conceptual categories that the agent formed for non-linguistic reasons. Suppose further that conceptual regularities in module C help account for syntactic phenomena in a purely linguistic module A, but because C is large and complex, it is not possible to index the correlations between A and C for quick retrieval without incurring unrealistically expensive storage costs. In neural terms, this might correspond to topological complexity limitations. If, however, enough regularities can be capture in module B at an intermediate level between A and C, a good part of the processing that would have required interactions between A and C could be replaced by within-module processing in B, speeding up the recognition times. However, these intermediate categories in B cannot fully take over the function of C because conceptual structure within B is too restricted. Thus some correlations must still be indexed all the way from A through to C. Moreover, B occasionally makes mistakes that C must correct, albeit at a slower speed.

On the innateness of lexical semantics. Whether the lexical semantics module is innate is a far-ranging question we cannot hope to answer yet, but I will offer a few comments from the computational point of view. In his arguments from the standpoint of learnability, Pinker argues for the innateness of the lexical semantics module. This seems improbable for the level of detail of his lexical semantics primitives, which include entities like *EVENT, PATH, liquid/semisolid, time-line.* It is more reasonable to take the intermediate position that any innate biases are too primitive to easily pin conceptual tags on, and that children adapt/evolve the exact lexical semantics primitives. Pinker leans away from this position most likely because of over-restrictive assumptions on how children learn in the presence of exceptions to subregular patterns (what in machine learning would be termed "noisy inputs") and on how statistical conceptual clustering methods can be biased to incorporate implicit classes for generalizing on indirect negative evidence. In fact humans are faced with subregular patterns in nearly every aspect of the environment and therefore concept formation and conceptual clustering are fundamental and ongoing areas of machine learning research, and it would be premature to make strong evaluations of the potential behavior of the many competing symbolic, statistical, and neural approaches. Pinker argues that "one would not want to posit a complex ad hoc pruning algorithm [for conceptual clustering] just for this task, as it is not the kind of task that the child has any strong need for" (p. 272), but the "ad-hoc-ness" is more a reflection of the immature state of conceptual clustering models than a convincing argument against the need for them.

We can adapt the processing speed argument above into a more general conjecture in the spirit of statistical and neural AI approaches, in which learning is noise-tolerant and the syntactic, semantic, and conceptual modules evolve in parallel. This is more in line with the proposed model's view of the human cognitive mechanisms' tendency to automatize frequently-used processes, thus adapting to speed up overall average recognition and reaction times. Suppose modules A and C are as before, but instead of prewiring primitive predicates into module B, new predicate categories are formed in response to the usage patterns between A and C. Again, B will be too restricted in size and complexity to absorb all of the regularities between A and C. The more the lexical semantics module B evolves, however, the more it facilitates further learning of the intermodular correlations.

4.5 Signification Mappings

A different kind of representational need from those we have been considering is the need to link lexicosyntactic or grammatical forms to structures in the mental image, lexical semantics, and conceptual modules. The proposed modularization follows the philosophy of Fillmore's (1988) Construction Grammar. Fillmore defines a construction as "a pairing of a syntactic pattern with a meaning structure". Like HPSG, this approach returns to de Saussure's (1966) idea of a *sign*, which "unites, not a thing and a name, but a concept and a sound-image (p. 66)". Note that de Saussure's wording places his approach firmly in the mentalist tradition. de Saussure goes on to replace these terms with *signifié* and *signifiant*, from which my term *signification mapping* derives.

Signification mappings are permitted to map lexicosyntactic constructs to any of the modules. However, those that map lexicosyntactic constructs to the lexical semantics module, but not the conceptual module, are postulated to be especially important in terms of capturing generalizations involving grammatical variation. In many theories similar kinds of mappings are given special status as "linking rules". We will see how signification mappings are encoded in subsequent chapters.

4.6 Summary

In this chapter I have discussed some representational needs that can be elegantly met by a modular ontology, including:

1. *image reification*, the relationship between concrete "vivid" conceptualizations and reified schematizations of images, including imposition of foreground/background distinctions,

2. *associative grounding*, the relationship between perceptually grounded image schemas and metaphoric uses of image schemas,

3. *compatible differentiated semantics*, the relationship between alternative interpretations or "meanings" that are compatible, handled by the notions of cross-modular semantic distinction and associatively inferrable conceptual shift, and

4. *signification*, the relationship between grammatical roles and thematic and conceptual roles.

In addition, to further motivate taking a modular ontology as the representational basis, evidence from various cognitive disciplines in support of mental image and lexicosemantic modules has been surveyed. The following chapter describes a particular approach to representation that fulfills these objectives, and which is amenable to probabilistic modelling of automatic inference.

Chapter 5

Chapter 5

Ontological and Grammatical Primitives

This chapter surveys the types of knowledge structure primitives used in the proposed parsing and interpretation model. The object is to construct a representation for mental images, lexical semantics, the conceptual system, and the phrasal lexicon, adhering to the modular philosophy described in chapter 4. To facilitate evidential reasoning and efficient storage, all knowledge in the model is eventually encoded using MURAL as described in chapter 6, but when thinking about linguistic knowledge a higher-level notation is convenient.

I have chosen to follow the philosophy of Construction Grammar (Fillmore 1988) as a basis for extension. A number of other grammatical frameworks could have been used. The crucial requirements are that (1) it is structuralist rather than transformational, (2) syntactic and semantic information are represented in such a way as to permit a uniform statistical treatment of disambiguation, and (3) complex syntactic and semantic patterns can be defined for the purpose of stipulating statistical information about those patterns. Construction Grammar satisfies the first criterion by providing a uniform notation for syntax and semantics. It satisfies the third by permitting constructions of arbitrary complexity (even if the syntax redundantly mirrors a pattern that could have been derived by composing simpler constructions).

The literature varies widely in notation, partly due to the differences in underlying theoretical paradigms. I have tried to keep the dependence on notation to a minimum in this chapter. However, a bit of formal notation is sometimes unavoidable, when the object is to communicate the differences between the proposed model and others. Where necessary, a feature-structure notation that is as close to "standard" as possible is used. Details of the representation formalism are reserved for chapter 6.

Many modern computational theories of grammar have returned to the structuralist rather than transformational paradigm. Common to most of these models is the assumption that various structures in the database represent constraints on legitimate structures. The most common form of structure is the *feature-structure*, also abbreviated as f-structure, which are structures formed by recursive role-filler or attribute-value composition. Internal substructures are permitted to be co-indexed, meaning that the same substructure fills two or more roles. To denote co-indexed substructures, I use the standard device of placing the same superscript number after each role; any information written under one role applies to all others with the same superscript (to avoid confusion, all information is generally collected under the same role). Especially in more formal discussions, I also refer to a feature structure as a *feature-DAG* to connote the distinction from a

simple feature-vector.[1]

The most widespread method of interpreting the structures employs the *recursive unification* operation, which combines feature-structures by identifying the root of one structure with some node in the other. Each node in the first structure is matched to a node in the second, with the restriction that the role paths must match. The new structure resulting from unification has, at each node, the features from both original structures. The *constraint satisfaction* power of this approach comes from the fact that the unification operation enforces feature consistency at each matched node, i.e., no mutually exclusive features can be present in any two nodes that are to be unified. The most influential models of this type include Functional Grammar (Kay 1979), LFG (Bresnan 1982), and HPSG (Pollard & Sag 1987). Shieber (1986) gives an excellent introductory survey of unification-based approaches.

The ontology described here is comparable in expressiveness to typical conceptual representations. However, besides (or rather, because of) its orientation towards statistical association, the representation is also powerful enough to address the less obvious yet important issues raised in the preceding chapter. With respect to these issues, implementation methods are discussed in the course of introducing the modules' primitives as they become relevant. Primitive mechanisms for several major areas have not yet been developed, including dynamic temporal representations and the certain quantification types. Several other areas that are standard concerns in knowledge representation do not require special mechanisms here, but the appropriate ontological primitives have not yet been defined; these include instantiation for real-world extensional entities, and various modalities and belief predications.

Notational de-emphasis notwithstanding, the primary goal of this chapter is to synthesize and formalize the modular paradigm of chapter 4 into a common representational framework that is amenable to the proposed evidential interpretation methods. Thus particular attention is paid to ensuring the concept space is well-defined, to facilitate constructing an underlying probabilistic event space as described in chapters 6 and 7.

5.1 Primitives for Mental Images

A substantial body of work in cognitive semantics deals with grounding semantics in the human perceptual system (Talmy 1983, 1985, 1988; Lakoff 1987b; Langacker 1987). In the L_0 project (Feldman *et al.* 1990; Regier 1991b, 1991a; Weber & Stolcke 1990), visuospatial primitives are selected by testing their adequacy for learning spatial vocabulary across many languages. My choice of semantic primitives is influenced by such studies, though because these studies are ongoing the selection must be regarded as extremely preliminary.

The representation sketched here uses compositional frames, that is, images are organized as collections of structured hierarchical objects. The representation is propositional rather than array-based, and is conveniently representable using either feature structures or DAGs. As such it is similar to Marr & Nishihara's (1978) 3D representation and to Hinton's (1979b, 1979a) model. However unlike those models no commitment is made here as to the coordinate system, say, whether positions are specified in object-centered or viewer-centered terms.

I have three goals behind laying out a set of mental image primitives. First, they serve as an example of how mental images could be integrated into a theory of language. Second, enough

[1] In graph theory, a DAG is a directed acyclic graph.

primitives are included to crudely represent the "vivid" semantics of the nominal compounds in my corpus (described in section 7.1.2). I am not going to propose a sufficient set of primitive types or relations, but I do consider certain primitives necessary at a minimum. Third, a subgroup of the primitives account for a range of types of physical constituency involving discrete objects, groups, substances and masses, a problematic area of ontology representation that is usually ignored. Under this account, atypical schematizations of concepts are handled through *type coercion*, a relationship between a concept and an alternate, possibly metaphoric, schematization of the concept.

5.1.1 Primitive Features for Visuospatial Mental Images

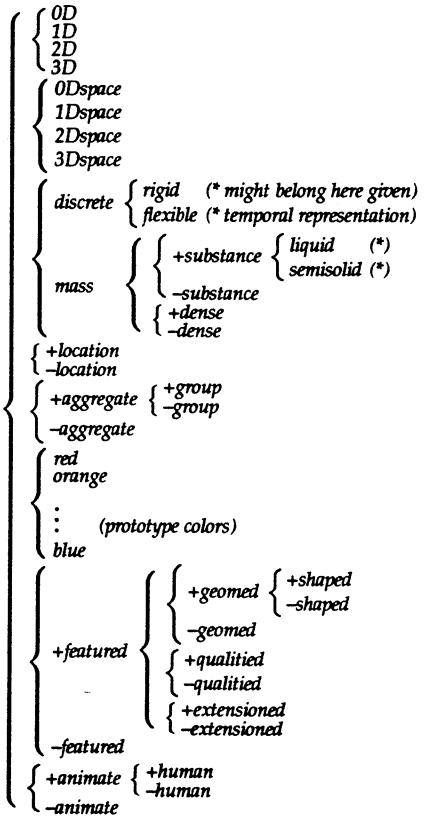


Figure 5.1: Primitive features for visuospatial mental images.

At the atomic level there are a number of features describing primitive characteristics of visuospatial mental image entity. The features used, listed in figure 5.1, are derived in part from Pinker (1989) though they are more flexible with respect to type-coerced constituency, as discussed below. Each bracket indicates a feature dimension, with the enclosed features being mutually exclusive. Except for nested dimensions, each dimension is orthogonal, meaning that any combination of features from separate dimensions is permissible. Features that are listed in a

nested bracket can only apply when the "parent" feature is set; otherwise, they are meaningless. Such features are sometimes referred to as "subfeatures".

The two feature dimensions that are listed first deal with the dimensionality of the mental image entity. For example, the aerial view of a *road* is a one-dimensional object, embedded in a two-dimensional space (assume that the road has curves and that elevation is imperceptible).

The next feature dimension distinguishes discrete and mass entities. An *object* concept represents an entity perceived as discrete and individuable (thus the *discrete* feature is set), such as *Pacific ocean*. In contrast a *mass* concept represents an entity perceived as non-individuable and amorphous, such as *ocean water*.

The *+location* feature marks entities that are perceived as locations. Although the distinction makes intuitive sense, theoretical justification for this feature is primarily linguistic, because not all concepts can equally easily be referred to by surface locative forms. For example, deictic terms like "there" apply more readily to a location like *coast* than to the typically non-location *handle*. Somewhere in between fall the concepts like *road*, which sometimes schematize as locations.

Concepts with the *+aggregate* feature set represent entities perceived as composite collections of other entities. If the *+group* feature is also set, this means the entity is perceived as a collection of like constituent entities; note that this is a form of coarse quantification. The way the constituent roles are specified is described in the next subsection.

The color features are obvious, as are *+/–animate* and *+/–human*. A set of additional *+featured* features are used only to indicate the presence of various roles specifying other properties and qualities of the entity; these are described further below.

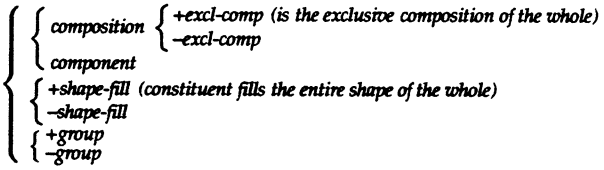
5.1.2 A System of Constituency Roles with Type Coercion

Thus far we have considered only features, which can be thought of as one-dimensional (single argument) predicates on concepts. We now consider compositional roles (which will be formalized in chapter 6). Their first application, a natural one, is to represent the compositional structure of aggregate mental image entities. Aggregates have roles filled by subparts. Such roles are often called "part-whole" roles, but because the term "part-whole" is badly overloaded, I will use the term *constituency*.

Figure 5.2 shows how different kinds of constituency are induced by a set of *role features*, which operate in the same way as the concept features above, except that they are attached to roles. Some common types of roles are also shown, along with their underlying role feature sets. For example, the standard type of constituency in *desk drawer*[2] is denoted by the COMPONENT role between *desk* and *drawer* in figure 4.3. On the other hand, *dirt clod* (p. 47) requires the FILL-X-COMPOSITION role indicating that the entire composition of the clod is constituted by dirt, and the dirt fills the entire shape of the clod. Both the COMPOSITION and PARTICLES role are needed for *seed pods*, shown in figure 5.3.

As the table in figure 5.2 shows, the precise function of the role also depends on the concept types of the whole and constituent part, especially with regard to whether they are mass or discrete. In fact, the first four entries are forms of *type coercion* in which discrete nonmass objects are transformed into substance masses. As a more obvious example of the type coercion relationship, consider:

[2]From Warren (1978, p. 127).



Role	Features				Whole	Part	Example
FILL-X-COMPOSITION	*composition*	*+excl-comp*	*+shape-fill*	*–group*	*mass*	*discrete*	*skunk meat ("grinding", "quantitization")*
FILL-X-COMPOSITION	*composition*	*+excl-comp*	*+shape-fill*	*+group*	*mass*	*discrete*	*skunks meat ("grinding", "substancification")*
FILL-COMPOSITION	*composition*	*–excl-comp*	*+shape-fill*	*–group*			*skunk hash*
COMPOSITION	*composition*	*–excl-comp*	*–shape-fill*	*–group*	*discrete*	*discrete*	*skunk burger*
COMPONENT	*component*		*–shape-fill*	*–group*	*discrete*		*skunk exhibit (i.e., a skunk at the zoo)*
PARTICLES	*component*		*–shape-fill*	*+group*	*mass*	*discrete*	*skunk herd*

Figure 5.2: Constituency role features for visuospatial mental images, along with some common examples.

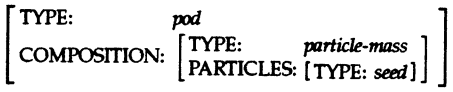


Figure 5.3: Constituency in *seed pod*.

(5.1) There's skunk all over the roadway.

The concept *skunk* is usually a discrete object, but in this case, it is coerced into a substance that one constructs by mentally "grinding" the skunk. The FILL-X-COMPOSITION role performs the coercion when applied to a discrete concept, as in *skunk meat*. Setting the *+group* feature indicates that the meat is derived from some group of skunks rather than an individual skunk. Wilensky (1989) has referred to these as "quantitization" and "substancification".[3]

A similar relationship holds when a group of like discrete objects is treated as a group mass. The PARTICLES role in *seed pod* or *skunk herd* exemplify this transformation.

Conversely, a mass concept can be transformed into a discrete object consisting of some quantity of the mass. The standard COMPOSITION role performs this function for *dirt clod* or *tofu burger*. Less obviously, the same role can be used to transform mass concepts like *water* into some imagined but nameless *body of water*.[4]

A couple of caveats on what constituency roles are *not*: (1) Constituency roles are not meronymy relations. Meronymy is a relationship between two *lexemes* where some sense of one denotes a part of some sense of the other (Cruse 1986). In contrast, constituency is a relationship between two conceptual entities. Moreover, the term "meronymy" is subject to vaguenesses in the definitions of "part" and "sense". (2) Constituency roles are not image schematic, in the sense that

[3]Wilensky's distinction actually differs slightly due to ontological differences. Quantitization transforms a discrete individual object into a mass (*Your skunk is all over the roadway*), whereas substancification transforms a category concept into a mass (*Skunk is all over the roadway*). In Wilensky's ontology, categories are primitive concepts at the same level as objects, which is not true of the analysis I propose. To explicitly represent that the source was possibly more than a single particular skunk, an (arbitrarily sized) set of skunks can be ground into substance.

[4]In Wilensky's ontology this is an individuation function performed by the "AIO" (an instance of) relation.

no figure/ground or landmark/trajector schema is imposed. Although they represent visuospatial images, mental images need to be reified into image schemas such as *containment* in order to impose a choice of foregrounding. This is described in section 5.2.2.

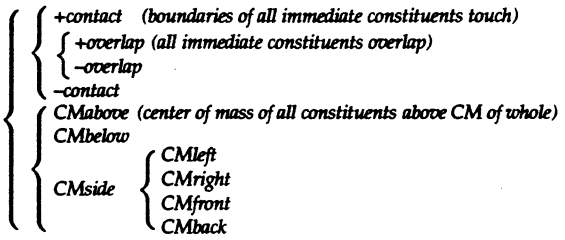


Figure 5.4: Supplementary constituency-related features for visuospatial mental image entities.

In addition to the type of constituency, the location and orientation of constituents needs to be specified. Marr & Nishihara (1978) suggested that these be represented relative to their immediate whole. Hinton & Parsons (1981) argue that all such object-centered positional relations are preserved except for handedness. I will only be making very crude assumptions about the conclusions that the representation makes readily available. Location is specified using various prototypical constituency frame types that describe idealized configurations between multiple constituents, for example, whether one constituent is above or below another, or whether they touch. This is accomplished by the supplementary set of features in figure 5.4. If the *+contact* feature of an aggregate is set, then all of its immediate constituents are understood to be in contact with the whole; *+overlap* is a stronger condition yet. The other features specify where the centers of mass of the immediate constituents are, relative to the center of mass of the whole aggregate.

Though I agree that the way orientation is specified should probably also be relative somehow to the immediate whole, as of now orientations are only specified in viewer-centered coordinates, as described below.

5.1.3 Other Roles for Mental Images

We now consider some non-constituency roles, for holding orientation and other information. The orientation primitives are based on the L_0 project's preliminary results on identifying language-universal primitives (Regier 1990, 1991b, 1991a, 1991c; Feldman *et al.* 1990). Again, it is my intent to give the flavor of the model's representation philosophy; no attempt is being made to address vision issues here.

A sampling of non-constituency mental image roles is shown in hierarchical form in figure 5.5. Features are only worthwhile in the bottom half (where they generate a number of different roles combinatorically). The *property* and *quality* roles are differentiated by the fact that property roles are filled by discrete concepts, while quality roles are filled by nondiscrete values. The only case of nondiscrete values suggested here are continuous-valued scalars.

A GEOM or SHAPE role is filled by a concept representing the geometry, or specifically, the shape of the entity. The EXTENSION role is filled by a concept that is the agent's representation for the real-world extensional *identity* of the imaged entity. It could be argued that the role is conceptual, and not a mental image role.

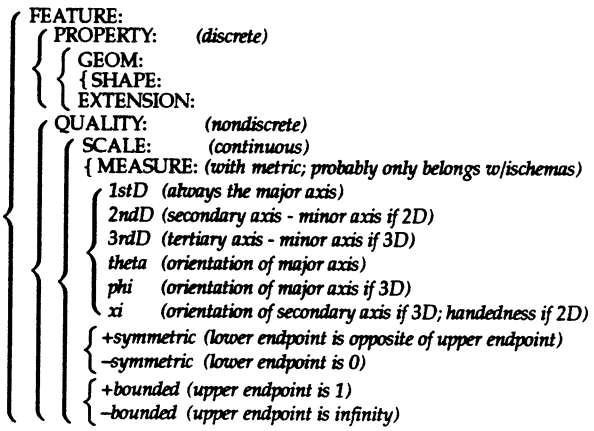


Figure 5.5: Supplementary roles for visuospatial mental images.

Scalar quantities can be of four ranges, depending on whether the *+symmetric* and *+bounded* role features are set. The *+symmetric* feature indicates the upper and lower endpoints of the range are centered around zero; otherwise the lower endpoint is zero. The *+bounded* feature indicates that the upper endpoint is one, rather than infinity. There are three scalar roles specifying the size of the entity along its (1) major axis, the longest axis that can be drawn through the entity, (2) secondary axis, next longest axis that is orthogonal to the major axis, and (3) minor axis, shortest axis, orthogonal to both others. An additional three scalars specify the entity's orientation.

5.2 Primitives for Lexical Semantics

5.2.1 A Feature System

I now outline the proposed lexical semantics module, to flesh out the semantic characteristics I think are important for this level. We begin with the feature system, which is a supersystem of the features employed for mental images. There is a strong methodological motivation and a weak theoretical one for including the mental image primitives. The theoretical motivation is that mental image distinctions like *discrete/mass*, *+/–location*, *+/–group*, *+/–animate*, and *+/–human* do surface as inflectional, morphological, and argument structure variations. The methodological motivation for including all the rest of the mental image distinctions is that arbitrarily excluding the other primitives would seem premature.

There are only a few new features. The *+/–craftable* feature distinguishes substances that are simply conceived of as matter from materials over whose shape the agent has control, at least hypothetically.

The *abstract* feature identifies entities like states, events, processes, and categories. All mental images are *concrete*, so this feature was not needed earlier. Following Bach (1983, 1986), the term *eventuality* is a general category covering states, events, and processes. Barwise & Perry (1983) use the term "situation" to cover the same things, but are more committed to a realist

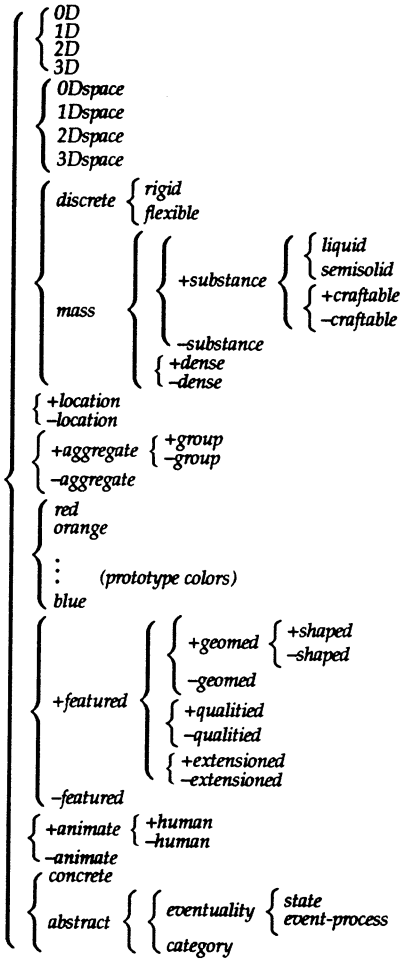


Figure 5.6: A feature system for lexical semantics.

interpretation.[5] There are a substantial number of subfeatures for eventualities; they are omitted here. Instead they are discussed below, in conjunction with their function of identifying thematic roles.

Note that though every semantic substructure type is a category in one sense, here *category* is meant in a different sense. The *category* feature marks semantic structures where the notion of categoryhood is explicit, for instance in the substructure *Interstate highway category* needed to handle

(5.2) The legislature considered designating Highway 17 an Interstate in order to qualify for federal funds.

Contrast this with the non-category structure *Interstate highway* used to handle the compound *Interstate truck stop* meaning a truck stop situated on an Interstate highway, and not a truck stop situated in the category of Interstate highways.[6]

5.2.2 A Thematic Roles System

The thematic role system described in this section is most directly related to the "case grid" systems of Somers (1987) and Ostler (1980), but is also strongly influenced by spatially-based cognitive semantics (Talmy 1983, 1985, 1988; Lakoff 1987b; Langacker 1987; Jackendoff 1972, 1983, 1990; Pinker 1989) and work on aspect (Bach 1983, 1986; Parsons 1985, 1990; Herweg 1991). I do not propose a thematic role system that claims to provide a full argument structure for the meaning of every verb. As with Jackendoff's and Pinker's theories the semantic representation allows nested structures, emphasizing flexibility and expressiveness rather than flat predicate-argument structure. Verb meanings in this approach are permitted to be composed out of multiple case frames. The approach also permits the same concept to fill different roles from multiple frames to help account for instances where an entity appears to be playing multiple case roles, as in Schank's (1973) Conceptual Dependency representation.

Overview. In discussing the lexical semantics features above I omitted the features distinguishing subtypes of eventualities. As can be seen from figure 5.7, the breakdown of eventuality types is reasonably detailed. Rather than explicitly enumerating the features, the most important frame types are listed hierarchically. The distinctions are finer than in Pinker's (1989, p. 195) proposal, which distinguishes events and states using a *+/–dynamic* role feature, and subcategorizes them with a *+/–control* role feature yielding the frame types *have* and *act* for the *+control* case, and *go* and *be* for the *–control* case. The approach of giving verb semantics by combining eventualities follows Parsons's (1985, 1990) subatomic semantics and Dowty's (1979) work on Montague grammar. Many of the frame types are frequently called image schemas. Image schemas are the special case of eventualities that are associatively grounded in mental imagery (and thus indirectly possibly eidetic memory); this is discussed later.

Perhaps the most striking characteristic of the thematic role system is that it has exactly four basic role types. The type of frame that a role is used in determines the precise function of

[5] See Higginbotham (1983) and Vlach (1983) for arguments for event-based approaches over situation semantics.

[6] Section 6.4.1 discusses how explicit categories work together with propositional rather than terminological IS-A relations.

eventuality	SOURCE	PATH	GOAL	LOCAL
state				
lm-tr	-	TRY	LM	TR
containment	-	BOUNDARY	CONTAINER	CONTENT
part-whole	-		WHOLE	PART
isa-category	-		CATEGORY	MEMBER
peripheral	-	BOUNDARY	CENTER	PERIPHERY
possession	-	-	POSSESSION	POSSESSOR
has-property	-	-	PROP	TR
lm-lm	LMONE	-	LMTWO	-
link	LMONE	LINK	LMTWO	-
lm-tr-lm	LMONE	TRY	LMTWO	TR
linear-order	LMONE		LMTWO	
locative				
temporal				
value				
front-back	BACK		FRONT	ENTITY
flank	LEFT		RIGHT	ENTITY
above-below	BELOW		ABOVE	ENTITY
state-schema				
opposition	OPPST:state		ST:state	TR
nonbinary	STSET		ST:state	TR
discretization		MIDPT		TR
scalarization				
bounded	BEGPT		ENDPT	
locative				
vert	DOWNPT		UPPT	
horiz	LEFTPT		RIGHTPT	
quant	LEAST		MOST	
positive	BEGPT		ENDLM	
...				
negative	BEGLM		ENDPT	
...				
unbounded	BEGLM		ENDLM	
...				
quant	LESS		MORE	
event				
accomplishment				
locative-acc	ORIGIN:loc	TRY	DESTINATION:loc	PATIENT
achievement				
locative-ach	ORIGIN:loc		DESTINATION:loc	PATIENT
*acc/ach**				
causal	CAUSER	MEANS	EFFECT:s/e	PATIENT
volative	AGENT	PLAN	INTENT:s/e	PATIENT
instrumental	INSTRUMENT	MEANS	EFFECT:s/e	PATIENT
psychological	STIMULUS	CAUSE	EXPERIENCER	EXPERIENCE
possessive	GIVER	CAUSE	RECIPIENT	PATIENT
value-ev	INITVAL:val		FINALVAL:val	PATIENT
state-change	INITST:state		FINALST:state	PATIENT
loc-st-chg	INITST:loc-st		FINALST:loc-st	PATIENT
val-st-chg	INITST:val-st		FINALST:val-st	PATIENT
process				
activity				EV:event

Figure 5.7: Overview of thematic role system. *An acc/ach frame is an accomplishment if PATH is a process, and an achievement if PATH is an event.

that role. Some roles are used exclusively in conjunction with certain types of frames, while others attach to different frames. One might think of this as a form of context-sensitivity in role function. Two major questions arise as to the motivation for this arrangement.

The first issue is why one would want to avoid simply having arbitrarily many types of roles. The empirical motivation against this is discussed below. There is also a computational motivation related to tractability concerns, namely to reduce the size of the concept space, by keeping the number of features needed for roles to a minimum. This is accomplished as described above by making the interpretation of roles depend as much as possible on the frame. Suppose there is some arbitrary number r of role types, and no restriction is placed on the combination of roles particular frames can have. If each frame can have up to b roles and frames can be nested to depth d, then the concept space includes $O(rcb^d)$ possible structures, where c is the number of different feature combinations. In standard knowledge representations r might typically be on the order of hundreds or thousands, whereas if we know that frames only occur with four role types the concept space is reduced by several orders of magnitude. Even more importantly, roles turn out to be far more difficult than simple feature bundles to represent effectively using vector representations. Although this is an open problem in neural network research, any method of reducing the complexity of roles stands to gain. It would be reasonable to speculate that concept formation also increases in difficulty much more rapidly with role complexity than feature complexity.

The second issue concerns the assignment of specialized roles to general primitive roles. The question is why the specialized roles are subsumed under *four* general roles, and whether there is any significance to labelling the roles SOURCE, PATH, GOAL, and LOCAL. Why not, for example, just label them 1, 2, 3, and 4? For that matter, one could minimize the number of roles by having just 1 and 2, since anything can be represented using binary relations.

In fact, an empirical claim does underlie the particular arrangement chosen. What is significant about the roles is not their intuitive labels, but the patterns of usage of the same roles across different frame types. EXPERIENCE (*grief*) and PATIENT (*roses*) often surface in the accusative as in *She caused him grief* and in *for giving her roses of the wrong color*. Similarly, the fact that EXPERIENCER and RECIPIENT are both GOAL roles surfaces in grammatical subregularities like their use of dative case—*He caused her grief* and *She gave him roses*—or the corresponding *to-PPs*—*He caused grief to her* and *She gave roses to him*.[7] That both EXPERIENCER and LMTWO (the second landmark in image schemas with two landmarks, like *West Coast* in the locative *They moved from the East Coast to the West Coast*) are both GOAL roles captures the shared use of *to-PPs*, though the locative does not permit the dative. The PATIENT in a *location state change* frame is semantically very close to the TR (trajector) role in a *flank* frame. Polarization tendencies like the fact that *down* is typically considered *less* are captured by subsumption under the SOURCE role. Having just two roles would fail to capture all the subregularities across frame types. The correspondence is by no means perfect and there are certainly many pairs of specific SOURCE role types that by themselves predict no surface correlations. However, the fact that they share surface grammatical predictions with other related roles constitutes evidence for an indirect "family resemblance" similarity in the way the roles are cognitively represented. The labels are merely intended to convey some notion

[7] Many people find *He caused grief to her* strange, yet do find the construct acceptable with heavy NPs as in *He caused grief to anyone foolish enough to get involved with him* or *He caused great and severe grief to her* (Chuck Fillmore, personal communication).

close to a "core" significance.

The case grid organization also serves the secondary purpose of providing primitive roles for the conceptual system; it defines a combinatorically restricted AI-style role hierarchy. This is discussed in section 5.3.2.

The interpretations shown of thematic roles for different frame contexts are only some of the major classes. The listing is by no means exhaustive. Fairly broad coverage is intended, however, since I have included all roles needed to interpret the nominal compound subcorpus taken from Warren (1978).

Image schemas and constituency reification. The cognitive status of image schemas with respect to representation and processing has not been clearly laid out in the past. As a first stab toward remedying this, I put forth a more concrete proposal of the relationship between mental images and image schemata, namely that *image schemas are reified mental images where certain components and relations have been foregrounded.* In knowledge representation, to reify is to view an abstract or implicit relation as a concrete concept, node, or object (Wilensky 1989, 1991). We consider the first of three kinds of reification here, *constituency reification*; the others are discussed later.

(a)
$$\left[\begin{array}{l}\text{SEM: } {}^{1 \text{ or } 2 \text{ (see text)}} \\ \text{FRM: } {}^{1}\left[\begin{array}{ll}\text{TYPE:} & \textit{containment} \\ \text{CONTAINER:} & \left[\begin{array}{ll}\text{TYPE:} & \textit{bone} \\ \text{COMPONENT:} & [\text{TYPE: } {}^{2}\textit{marrow}]\end{array}\right] \\ \text{CONTENT:} & {}^{2}\end{array}\right]\end{array}\right]$$

(b)
$$\left[\begin{array}{l}\text{SEM: } {}^{1 \text{ or } 2 \text{ (see text)}} \\ \text{FRM: } {}^{1}\left[\begin{array}{ll}\text{TYPE:} & \textit{part-whole} \\ \text{WHOLE:} & \left[\begin{array}{ll}\text{TYPE:} & \textit{desk} \\ \text{COMPONENT:} & [\text{TYPE: } {}^{2}\textit{drawer}]\end{array}\right] \\ \text{PART:} & {}^{2}\end{array}\right]\end{array}\right]$$

Figure 5.8: Constituency reification on mental images, producing image schemas for (a) *bone marrow* and (b) *desk drawer* (see text).

In constituency reification, an explicit *containment* schema is used to relate some constituent to a whole object for which the agent has formed a mental image. Constituency reification makes it possible for the agent to conceptualize a previously implicit constituency relation as an object, albeit an abstract one. For example, figure 5.8 shows the structures containing both the mental image and the reified image schema for *bone marrow*[8] and *desk drawer*[9]. The SEM role is used to pick out the specific substructure of the FRM that is being signified (see section 5.4.2). With ^{2} in the SEM role, the structure would be used to represent the marrow or the drawer. On the other hand, with ^{1} the structure would be used to represent the containment, for use in cases such as

(5.3) Becky says half the fun of eating marrow is that the marrow is in the bone.

Reification is what makes visual foreground/backgrounding possible. The mental image level representation contains no foreground/background distinction. However, the propositional image-schema level uses the LOCAL role to impose a foreground or trajector function on particular objects, and the SOURCE and GOAL roles to impose a background or landmark function.

[8] From Warren (1978, p. 185).

[9] From Warren (1978, p. 127).

After image schemas have been learned they can be applied to abstract concepts as well as perceptual ones. However only spatial (or other sensory) concepts can be mapped back to the mental image level.[10] This accounts for the fact that though we can think of *law degree*[11] in terms of the containment of the degree by the field of law, and though we can apply the usual operators on containers and talk about *getting out of law*, we cannot visualize either the degree or law field.

In our model image schemas are given fundamental status but not as representational primitives. Lakoff (1987b) considers image schemas the primitive relations out of which cognitive "chunks" are built (what he calls ICMs or idealized cognitive models). However, this would leave open the question of which sense to consider primitive, since primitive relations can be sensed in more than one way. Containment can be perceived visually yet a blind person can also sense it by feel. In fact all the different ways of sensing a relation reinforce the meaning of the abstract primitive (this is what I mean by associative grounding, more formally defined below).

Cognitive semanticists may take exception to the use of "image schema" to mean a non-perceptual conceptual structure. However, to "schematize" something is to abstract it and even perceptual images, once abstracted, can lose their perceptual status. Abstract, spatially-motivated relationships become propositionalized to the point where they can be applied to non-spatial, non-visualizable (non-imageable) concepts as well. Normally the propositional relationship would be thought of without actually visualizing/imaging the composite picture. However if forced to, one can sometimes construct visual interpretations for abstract domains, e.g., the idiom *hump day* is construed (or at least, was construed when initially coined) by visualizing the seven-day week as a hill centered at Wednesday.[12] (Following Levesque (1986), we might call this representation of the metaphor "vivid".) Of course for spatial concepts, one can simultaneously think of the abstract relationship between them, and visualize them.

States. States describe non-changing situations. They do not intrinsically include any information on duration, and can hold for arbitrary lengths of time. A state has no culmination or intrinsic termination.[13]

Events. All events have intrinsic terminatation points, so one can sensibly ask whether an event is "finished". Usually events also have intrinsic, definite culmination points. Events are broken down into two major classes: the *accomplishment* and the *achievement*. An achievement is a happening that is conceived of as an instantaneous event such as *hit*, *stop*, *win*, or *stumble*. One cannot meaningfully ask for the duration of an achievement. In contrast, an accomplishment is an event that is conceived of as an event whose progress spans some temporal interval.

Many types of event frames can be used for either accomplishments or achievements. In all event frames except locative events, the PATH role specifies the means by which a state change

[10] Or, for that matter, further back to array or eidetic store.

[11] From Warren (1978, p. 174).

[12] Nowadays, of course, the term is so entrenched in certain dialects that hearers need not perform the visualization: *hump day* is simply a lexicalized form of *Wednesday*.

[13] This may be confusing since one might argue that *owning a 57 Chevy from 1963 to 1966* is a state. Note that this phrase might mean two different things, however, one of which is an event and one of which is a state with no termination. First, in the terminated sense it is an event, a sense we might reinforce by changing the tense—*owned a 57 Chevy from 1963 to 1966*. Second, viewed as a state, someone who can be predicated *now* as *owning a 57 Chevy from 1963 to 1966* will *always* be predicated as such. In other words, the state sense has no termination, even though the event from which the state is reified contains a termination point.

is brought about. If the semantic substructure that fills the PATH role is a process (see below) then the event frame is taken to be an accomplishment. On the other hand, if the substructure is another event, the event frame is taken to be an achievement.

Locative events are statically schematized versions of spatiotemporal (motion) images. As I mentioned earlier the model currently has no representation for spatiotemporal images. If a representation were developed, the method of schematization would be what Langacker (1987, p. 145) calls "summary scanning". The spatial path taken by the trajector/patient over time is converted to a static, directed curve containing all and only those points passed through. Such a mechanism has been implemented in a connectionist model to handle motion prepositions in the L_0 project.[14]

The notion of agency is accounted for by wrapping a causal event around some effect (state or event). This can be glossed by saying that *skater* is designated the agent in *The skater rolled down the hill* by using the semantic interpretation *The skater caused the skater to roll down the hill.* This is a different resolution of the problem that Jackendoff (1972) noted with sentences whose subject functions both as theme and agent. Jackendoff argued that thematic relations, unlike case systems, do not constrain each noun to only one role, and thus more elegantly capture the similarity of the above sentence to *The tire rolled down the hill.* In Jackendoff's proposal *skater* functions as both the theme and the agent of the same sentence frame. In contrast I distinguish the causal frame, with which the agent role is associated, from the effect frame, with which the theme role is associated. Both the causal frame's agent role and the effect frame's theme role are source roles in the case grid. This more cleanly captures the generalizations in the previous two sentences, as well as *The skater rolled the tire down the hill.*

Processes. I propose using the *discrete/mass* feature to distinguish events from processes when a frame is *abstract* rather than *concrete.* Processes resemble events in being dynamic, changing situations. However, unlike either accomplishments or achievements, they are not intrinsically terminating. Whereas events have culminations giving them the *discrete* property, processes are indefinitely bounded entities, giving them the *mass* property.

The literature does not usually differentiate the terms "process" and "activity", but I define *activity* as a special case of process that is created by schematizing an indefinite sequence of event repetitions as a process. For example, *hurricane season*[15] requires schematizing an indefinite set of *hurricane* events as an activity that occurs during hurricane season. The schematization of events into activities is represented as a special case of *particle mass* constituency, which we saw earlier.

5.2.3 Discussion

The first of several kinds of schematization, type coercion, was brought up earlier in section 5.1 for the case of physical constituency. A second, constituency reification, was discussed in the context of image schemas. Here we consider some other kinds of schematization in the lexical semantics module. First we consider three related kinds of schematization based on the *scale* type, namely discretization, scalarization, and similarity. These are sometimes also considered

[14] Terry Regier, personal communication.

[15] From Warren (1978, p. 246).

as type coercion, though in the proposed model at least, no concepts are coerced. Afterward we will examine two other kinds of reification.

Scales, discretization, scalarization, and similarity. A *scale* concept is one against which relative positions or intervals can be defined, and is the basis for representing degrees such as weight or temperature. Scales may be bounded at both ends (fullness; sometimes health, temperature), unbounded at one end (weight, cost, speed; sometimes health, temperature), or unbounded at both ends (beauty, poverty). They follow the same notation rules as any other feature structure. There are underlying semantics special to *scales* but they are not considered here, being irrelevant for current purposes.[16] Though the current schematization mechanisms are crude, they capture the flavor of the approach. For additional examples involving scales, also see the examples in the discussion on reification.

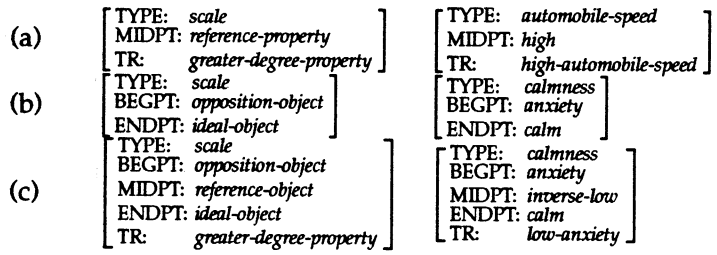


Figure 5.9: (a) Discretization, (b) scalarization, and (c) similarity schemas, using the *scale* type and with examples.

Discretization is the operation whereby a new discrete property is defined, relative to some interval on a scale. The current discretization mechanism operates by using the MIDPT and TR (trajector) roles of a predefined *scale* such as *automobile speed*, as shown in figure 5.9(a). The *reference property* is a predefined point or interval on the scale such as *high*. The *greater degree property* is then defined to be the property ranging over the part of the scale greater than the reference property. The resulting *high automobile speed* property can be used in the semantic representation of, for example, *high-speed buses*[17].

Scalarization occurs when a scale is defined by turning an ordinarily discrete property or object into a matter of degree (Wilensky 1989). For example, normally someone is either pregnant or not, but in the sentence

(5.4) The cover photo of an unclad and very pregnant actress caused a media overreaction.

the property *pregnant* is converted into a degree. Figure 5.9(b) shows the scalarization mechanism's use of the BEGPT and ENDPT roles of a *scale* to specify the endpoints or landmarks of the scale. The *ideal object* is an object that serves as the reference point for a greater degree on the new scale, for example *calm* on the *calmness* scale, and the *opposition object* serves as the reference point for the lesser degree, for example *anxious*. Like primitive scales, scalarized scales can be bounded on both

[16] Various calculi can be employed, including interval logics, mean-variance calculi, and fuzzy logics.

[17] From Warren (1978, p. 148).

ends, as in the example, or they can be unbounded on either or both ends. For an unbounded end, specifying a landmark indicates the directional point of reference, rather than an absolute endpoint of the scale.

The discretization and scalarization operations combine to form the *similarity* mechanism, which places discrete properties or objects on a scale to support a new discretization. In figure 5.9(c) the *calmness* scale from (b) is combined with a discretization operation, producing the new property *low anxiety* which is closer to the ideal end, *calm*. This property is used in the semantic representation for *low-anxiety child*[18], as shown in figure 5.10. Note the polarity of the scale has been chosen to produce the desired property; were the endpoints reversed, the scale would be *anxiousness* and the property *low anxiety* could not be defined. This is not a limitation on expressiveness because scales are always permitted to be inverted.[19]

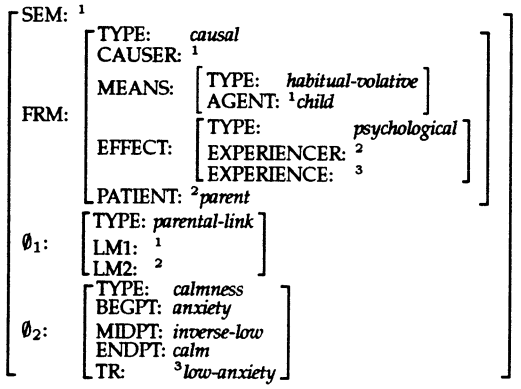


Figure 5.10: Use of the defined *low anxiety* concept in the semantic representation for *low-anxiety child*.

Figure 5.11: Use of the similarity mechanism for peripheralization.

The similarity mechanism can also be used to derive "peripheralized" (Wilensky 1989) concepts like *reddish* which apply to objects that fall in the periphery of the category described by *red*.[20] As shown in figure 5.11, a scale is created between the *any red* and ideal *prototype red* concepts. The *reddish* property is then defined by the interval between *any red* and the reference

[18] To avoid dealing with quantification the plural form *low-anxiety children* from Warren (1978, p. 148) has been dropped. Groups or categories could possibly be used to handle the plural form.

[19] Whatever the underlying calculus is, it should guarantee symmetry of scales.

[20] Note that Wilensky's use of the term "peripheralized" extends the periphery of a concept beyond its normal applicable range. I instead use the term to mean the peripheral area of the applicable range, and would suggest a term like "peripheral extension" for the former.

point, which is a default or average point corresponding to *-ish* forms.[21]

Comparatives like *redder* are handled using the similarity mechanism; several variations are found in the reification examples in figure 5.12. In (a) a *redder red* property is defined to characterize the color of *a redder apple [than apple/2]*, where the *reference red* property is contextually defined in $\emptyset_1$ to be the color of some other apple *apple/2*. (The slash / notation differentiates extension.) An alternate interpretation is shown in (b), where instead of defining a redness scale, we define a "redness-of-apple" scale with a slightly different connotation.

Representing state and property reification. Recall that in knowledge representation terminology, to reify is to view an abstract or implicit relation as a concrete concept, node, or object. We considered how to implement constituency reification in the discussion on image schemas. State reification and property reification are considered here.

(a)
$$\begin{bmatrix} \text{SEM: } 1 \\ \text{FRM: } 1 \begin{bmatrix} \text{TYPE: } \textit{apple/1} \\ \text{COLOR: } \textit{redder-red} \end{bmatrix} \\ \emptyset_1\text{: } \begin{bmatrix} \text{TYPE: } \textit{apple/2} \\ \text{COLOR: } \textit{reference-red} \end{bmatrix} \\ \emptyset_2\text{: } \begin{bmatrix} \text{TYPE: } \textit{redness} \\ \text{MIDPT: } \textit{reference-red} \\ \text{ENDPT: } \textit{proto-red} \\ \text{TR: } \textit{redder-red} \end{bmatrix} \end{bmatrix}$$

(b)
$$\begin{bmatrix} \text{SEM: } 1 \\ \text{FRM: } \begin{bmatrix} \text{TYPE: } \textit{redness} \\ \text{MIDPT: } \begin{bmatrix} \text{TYPE: } \textit{apple/1} \\ \text{COLOR: } \textit{reference-red} \end{bmatrix} \\ \text{ENDPT: } \begin{bmatrix} \text{TYPE: } \textit{apple/1} \\ \text{COLOR: } \textit{proto-red} \end{bmatrix} \\ \text{TR: } 1 \begin{bmatrix} \text{TYPE: } \textit{apple/1} \\ \text{COLOR: } \textit{redder-red} \end{bmatrix} \end{bmatrix} \\ \emptyset_1\text{: } \begin{bmatrix} \text{TYPE: } \textit{apple/2} \\ \text{COLOR: } \textit{reference-red} \end{bmatrix} \end{bmatrix}$$

(c)
$$\begin{bmatrix} \text{SEM: } 1 \\ \text{FRM: } 1 \begin{bmatrix} \text{TYPE: } \textit{state-schema} \\ \text{TR: } 2 \\ \text{ST: } 2 \begin{bmatrix} \text{TYPE: } \textit{apple/1} \\ \text{COLOR: } \textit{redder-red} \end{bmatrix} \end{bmatrix} \\ \emptyset_1\text{: } \begin{bmatrix} \text{TYPE: } \textit{apple/2} \\ \text{COLOR: } \textit{reference-red} \end{bmatrix} \\ \emptyset_2\text{: } \begin{bmatrix} \text{TYPE: } \textit{redness} \\ \text{MIDPT: } \textit{reference-red} \\ \text{ENDPT: } \textit{proto-red} \\ \text{TR: } \textit{redder-red} \end{bmatrix} \end{bmatrix}$$

(d)
$$\begin{bmatrix} \text{SEM: } 1 \text{ or } 2 \text{ (see text)} \\ \text{FRM: } 1 \begin{bmatrix} \text{TYPE: } \textit{has-color-schema} \\ \text{PROP: } 3 \\ \text{TR: } 2 \begin{bmatrix} \text{TYPE: } \textit{apple/1} \\ \text{COLOR: } 3\textit{redder-red} \end{bmatrix} \end{bmatrix} \\ \emptyset_1\text{: } \begin{bmatrix} \text{TYPE: } \textit{apple/2} \\ \text{COLOR: } \textit{reference-red} \end{bmatrix} \\ \emptyset_2\text{: } \begin{bmatrix} \text{TYPE: } \textit{redness} \\ \text{MIDPT: } \textit{reference-red} \\ \text{ENDPT: } \textit{proto-red} \\ \text{TR: } \textit{redder-red} \end{bmatrix} \end{bmatrix}$$

Figure 5.12: Schematizations of *redder apple* with alternate scalarizations and reifications.

State reification is implemented by taking the state represented by some feature structure and making it into a proper concept (i.e., feature structure). Figure 5.12(c) shows how the *redder apple* from (a) is turned into *being a redder apple* by embedding it in a *state schema*. The difference between a mental image state and a state eventuality cannot be represented in the simpler single-level *+/–dynamic* distinction between events and states proposed by Pinker (1989, p. 195) and based on Jackendoff.

Property reification is implemented by making a proper concept out of the a (discrete) property role in a feature structure. Putting [1] or [2] in the SEM role in figure 5.12(d) determines whether it represents *apple having a redder color [than apple/2]* versus *apple which has a redder color*

[21] Consider, however, that sometimes *reddish* is extended to describe objects one would never call red, like *tangerines* in

(5.5) The study showed that consumers shun reddish tangerines.

To handle such cases it is necessary to instead relativize to a scale whose endpoints are *prototype tangerine* and *reddest tangerine*.

[than apple/2]. In either case, the property of having a red color is reified by the *has color schema*. Scalar properties can also be qualitized by first applying discretization.

Other kinds of schematization. Wilensky (1989) defined as "categorization" the coercion of individuals into types, as for example in *an independent Namibia*. I suggest handling this by using a category, group, or set concept whose members are Namibia in alternative possible worlds. Since extension is not a primitive relation in the proposed representation, the intensional feature structure for Namibia (i.e., without the extensional attribute) need not be type-coerced. Figure 5.13 defines a group concept having the intensional Namibia as its constituent structure.

$$\begin{bmatrix} \text{TYPE:} & \textit{intensional-Namibia-group-category} \\ \text{COMPOSITION:} & \begin{bmatrix} \text{TYPE:} & \textit{particle-mass} \\ \text{PARTICLES:} & [\text{TYPE: } \textit{intensional-Namibia}] \end{bmatrix} \end{bmatrix}$$

Figure 5.13: "Categorization" coercion for handling *an independent Namibia*.

Wilensky (1989) also suggests "objectification" which converts properties into discrete objects. For example, an *oddity* is an object with the property of being odd. This seems to be a lexical rather than conceptual operation. From the standpoint of ontology, it is simple to create a feature structure representing *an odd object*.

Associative grounding revisited. Section 4.2.2 discussed the issue of how symbols at abstract levels can be grounded by statistical intermodular associations. The representational primitives we have been studying provide the structural basis for encoding such associations.

In figure 5.12, the relationship between (a) and (d) is the property reification of the COLOR role, which is part of the mental image module, into the *has color schema*.[22] Note that the structure in (d) contains both the reified (lexical semantics) and non-reified (mental image) representations. If structures of the sort in (d) are frequently used by the agent, an association is established between mental image structures with a COLOR role and lexical semantic structures with a *has color schema*. What is important about associative grounding is that the representation must provide a way to capture the co-occurrence frequency correlation; exactly how this is formalized in the proposed model is described in chapter 6.[23]

Similarly, a schema type like the *height* scale is identified by some feature and role combination that frequently occurs in conjunction with a mental image of an object of corresponding height. An abstract image schema type like the *calmness* scale then obtains its grounding through the fact that its representation employs nearly the same features and role structure. Usually in symbolic models, the associative perceptual connection is implicitly assumed but no formal or operational mechanism is provided.

[22] The *has color schema* is a specialization of *has property* from the lexical semantics module, but is itself part of the conceptual system rather than the lexical semantics module. While the fact that something is a reified property may have direct syntactic or morphological bearing, the fact that the property is in particular a color is unlikely to capture any useful generalizations at that level.

[23] A high marginal probability is assigned to the structure containing both the reified and non-reified representations.

Differences from Case Theories. This representation depends more heavily on nested structures than ordinary case systems, thereby reducing the number of thematic roles. A side effect of this is that surface cases do not map straightforwardly onto thematic roles in a one-to-one fashion. For instance, the sentence

(5.6) John moved his furniture from San Francisco to Berkeley.

is analyzed as an outer *volative* frame whose EFFECT role is filled by an inner *locative*. Thus we control proliferation of cases by defining the less immediate cases as internal roles. Similarly,

(5.7) John broke the window with a hammer.

is analyzed as an outer *volative* frame whose PATIENT is the *window*, EFFECT is a *window breaking state change*, and PLAN is an inner *instrumental* frame, whose EFFECT role is in turn filled by the same *window breaking state change*. No single frame can have both AGENT and INSTRUMENT roles; a verb frame involving both must use nested semantic structures. For convenience we can use a chain notation, so that the traditional instrument case role becomes PLAN.INSTRUMENT. Similarly, the beneficiary case role is equivalent to INTENT.PATIENT.

Note that INSTRUMENTs are considered CAUSERs, just like AGENTs except for being non-volative. Thus in our analysis the hammer in sentence (5.7) is an instrumental CAUSER internal to the PLAN of the breaking event, and the internal frame filling the PLAN role is itself the semantic structure for

(5.8) A hammer broke the window.

where the hammer is the non-volative cause. Omitted here is John's precise action upon the hammer; to represent this one would have to add a third *volative* frame in which the AGENT is John and the PATIENT is the hammer. Of course we would expect a high degree of association between this *volative use of tool* frame and the *instrumental* frame.

I discussed motivations for restricting frame roles earlier; I now add another *peripheral semantic role* motivation. There are analogous precedents in syntactic theory for the philosophy of narrowing the number of direct roles of a frame. Various theories of grammatical cases treat clauses as layered structures with a central and peripheral argument structure. In these theories the clause consists of one or two core arguments surrounding the predicate, typically labelled "subjects", "direct objects", "actors", "undergoers", and so forth. In addition a larger number of peripheral arguments express spatiotemporal information or secondary participants such as "beneficiaries". The distinction has variously been termed "core" versus "peripheral" (Silverstein 1976; Foley & van Valin 1984; Foley & Olson 1985), "inside the vp" versus "outside the vp" (Fillmore 1968), "nuclear" versus "satellite" (Dik 1978), "inner" versus "outer" (Halliday 1970; Platt 1971; Somers 1987), "nuclear" versus "peripheral" (Longacre 1976), and "propositional" versus "modal" (Cook 1972). Though some of the predicate argument classes are characterized using semantic notions, the theories are syntactic and do not extend the core-periphery claims to the semantic representation. The proposed representation does exactly that: instrumental roles, for example, are postulated to be semantically more peripheral than agent roles if the verb expresses a volative rather than non-volative causal event. Note that this does not present any verb valency problems since nested structures are still available for specifying the linking between verb clauses and their underlying semantic frame(s). The core-periphery contrast here applies to

semantic frames—Parsons (1990) might say "subatomic"—rather than verb frames. Across many languages, including English, certain roles are more readily marked by word order while others retain case or prepositional markers. (The fact that other languages like Latin or Russian use case markers for a wider set of roles does not contradict the hypothesis that not all the cases encode semantic roles of the same status.) A semantic explanation of why this pattern should evolve would be more satisfying than a purely grammatical one.

	Source	Path	Goal	Local
Active	instigator of action ±volitive ±animate	instrument or means	intended result (−animate) active recipient (+animate)	non-passive patient
Objective	original state (−concrete) material (+concrete)	counter-instrument passive means	result state (−concrete) factitive (+concrete)	undergoing change-of-state
Dative psychological:	stimulus	medium	experiencer ±dynamic	content
possessive:	original owner	medium/price	recipient	thing transferred
Locative	place from where	space traversed	final destination	static position
Temporal	time since	duration	time until	time at which
Ambient	reason	manner	aim (+volitive) consequence (−volitive)	condition

Figure 5.14: Somer's (1987) case grid proposal.

My use of the "case grid" organization differs somewhat from the usual linguistic usage (Ostler 1980; Somers 1987) although they resemble each other. First, the proposed thematic role system is more of a hierarchy than a grid. Somers is unsure whether to interpret his grid as an elaborate model for inner cases only, or a sparse model of both inner and outer roles, but he leans towards the former. As we have just seen, I propose to distinguish outer roles by hierarchical nesting, rendering the question moot. Second, Somers' grid, shown in figure 5.14 employs additional role features (although he qualifies them as "optional"), whereas the restriction to four role types is absolute in the proposed system. Third, in contrast to the case grids, activeness/passiveness is a distinction made orthogonally to the case system. There is no "eventive/agentive" (Ostler 1980) or "active/objective" (Somers 1987) separation within the role system. Somers sees Objectives as "processes" (a different usage than the eventuality sense) as opposed to Actives which are actions. In the proposed model, the *volative* frame type distinguishes actions with volitional agents and all others are considered passive. I see no need to distinguish active and passive patients; Somers suggests handling reflexives by using Active Local as an "active patient", but reflexives can be handled equally well by coindexing the AGENT and PATIENT. Fourth, the function of a role cannot

be ambiguous, as opposed to cells like Somer's Objective Path, which can indeterminately function either as a "counter-instrument" (enabler or non-tool-like instruments), or as the instrument in an agent-less event as in sentence (5.8).

Thematic roles as probabilistic entailment. The semantic significance of thematic roles, I propose, should be couched in terms of Bayesian conditional probabilities, instead of the traditional perspective based on entailment. In model-theoretic semantic proposals like Dowty's (1989), it is argued that thematic roles are properly seen as a cluster of entailments and presuppositions. Instead, we should see them as probabilistic entailments, and see thematic roles as entities developed by the cognitive mechanism in order to facilitate making useful inferences a large percentage of the time. They are an organizational construct that need not be rigorously deductive to increase the agent's efficacy and chances of success. We cannot call thematic roles entailments, because (almost) any condition we try to stipulate on what types are permitted to fill a role can be violated. The filler of an AGENT role is not necessarily "a rational, sentient, animate being", though it usually is; as the AGENT becomes increasingly dissimilar to a rational, sentient, animate being (*The child/seedling/glacier revelled in the cold*), we begin to call it metaphoric or figurative usage. Probabilistic entailment, however, is precisely how we can describe thematic roles: a thematic role entails certain changes to the probability distribution over its potential fillers. Knowing the type of the thematic role between two semantic substructures conditions the distribution over the type of the substructures. Conversely, knowing the types of substructures conditions the distribution over the kinds of thematic roles that might hold between them. Thus, even without rigorous entailment conditions, the thematic role structure is useful to a cognitive agent for evaluating possible inferences in the presence of missing information.

Comparison with Conceptual Dependency Theory. In some respects the philosophy of representation is quite similar to Schank's (1973) influential Conceptual Dependency (CD) Theory. Like CD Theory there is a small set of roles that are conceptually motivated and influenced by case roles. Unlike CD, the roles are intended to be lexicosemantic rather than conceptually canonical since (1) they are intended to help capture surface subregularities, and (2) they are dependent on how a situation is schematized.

Moreover, unlike CD, the set of frames is not limited to a small number of primitives. Concepts like *transfer* have only a couple of variants in CD. Instead, to capture fine conceptual distinctions I encourage proliferation of frame types as described below, and define them using a multiple hierarchy (or set of features; see section 5.3.1).

There is yet another subtler but important difference in my interpretation of both semantic and conceptual structures. A Conceptual Dependency representation is supposed to capture the meaning of an utterance by combining concepts using the primitive relations. In the proposed model, however, a semantic or conceptual structure only captures part of the meaning of an utterance, because of the notion of associatively inferrable conceptual shift. The meaning is entirely captured only by examining the agent's state as a whole, including the lexicosyntactic representation and signification mappings, and also the state of the controlled inference system (which I do not treat here) that takes care of higher-level pragmatic functions and backtracking. Thus in the proposed model, all conceptual distinctions need not be captured by static aspects of representation.

5.3 The Conceptual System

In this section I describe the nature of conceptual structures under the proposed model's representational philosophy. No attempt is made to actually construct any particular comprehensive conceptual hierarchy of theoretical significance. This is not to say that I think such structures don't exist. Rather, they exist but depend on ecological factors which are shared over specific communities, some structures being relatively universal and others being particular to very specific communities. At this level most constraints on concept types appear to be of general functional and storage capacity nature, making the conceptual ontology more flexible, adaptive, and dependent upon experience than the previous levels.

5.3.1 The Conceptual Hierarchy Approach

Shared roles for lexical and conceptual semantics. The conceptual system uses the thematic roles as its primitive roles. In other words, the conceptual system's basic conceptual building blocks (frames) are the eventualities from the thematic role system. In section 4.4.1 we considered motivations for separating the lexical semantics and conceptual modules. No criteria were found for determining how conceptual representations differ from lexical semantics representations. The evidence for lexical semantics comes from linguistic patterns; the evidence for conceptual structures is far less accessible. The kind of conceptual structures we are interested in here are not higher-level structures such as plans, because those are presumably more relevant to controlled inference. Direct empirical evidence for low-level conceptual organization comes from experiments on non-linguistic categorization and prototypes (Rosch 1975; Rosch *et al.* 1976; Smith & Medin 1981; Mervis 1980), but even there much of the evidence is derived through linguistic tasks. In the absence of convincing arguments for particular conceptual systems, having the thematic roles serve double duty as the primitive roles of the conceptual system helps avoid unnecessarily proliferating role types and notations.

Note that this does not imply that an utterance's semantic representation is the same at both levels; in fact I will argue below that the representation is often different. Nor, for that matter, does it impact on lexical semantics' distinguishing claim to close association with surface grammatical form. Still, this is nowadays a somewhat unusual move; contrast for example Nirenburg & Levin's (1991) model and the Logical Forms of May (1985) and Allen (1987), discussed in section 4.4.1, all of which employ different representations for lexical semantics and conceptual levels.

We also saw in section 4.4.1 that it would be incorrect to construe the sharing of primitive role types as a version of the Charniak's (1981) "case-slot identity theory", which says that case roles are identical to slots in frames. Charniak pointed out some major pitfalls with equating cases with slots, but the model I propose avoids them by using thematic rather than case roles. It should be noted that some deep-level case proposals have made similar, though less explicit, assumptions; for example, both Conceptual Dependency (Schank 1973) and Preference Semantics (Wilks 1975a, 1975b, 1982) employ case-like roles as conceptual primitives. Charniak's argument that this merely reduces case roles to a purely syntactic device is true in one sense, since one could equally well name the roles 1, 2, 3, and 4 as I have pointed out in section 4.4.1). However, Charniak does not observe that these roles can *still* be more concise than others for describing surface grammatical distinctions, even if the mapping is not one-to-one.

The shared roles approach is also compatible with the processing motivation given in

section 4.4.2. Since the lexical semantics module mediates between the conceptual, lexicosyntactic, and mental image systems, the conceptual structures most frequently employed in intermodular mappings are the most efficient for the lexicosemantic level. The shallowest, most general features of the conceptual system will obviously participate in more mappings than specific ones, and therefore these should be innate to or acquired by the lexical system.

The case grid system generates the sort of role hierarchy typically found in semantic net representations. For example, the relation between the PATIENT and EXPERIENCER roles is simply role subsumption in the KL-ONE variety of languages. However, there are two important restrictions not always found in semantic networks. First, the absolute number of eventuality roles (i.e., not constituency or property roles, which are part of the mental image level) is at most four. Second, the eventuality roles cannot be duplicated; for example, we cannot define a new frame with two PATIENT roles just by renaming the roles differently. Together these restrictions are intended to reduce the complexity of the concept space, hopefully improving learnability, without sacrificing the expressivity needed to recognize the regularities in surface grammatical forms.

Schemas. Cognitive and AI models typically employ larger memory organization structures or schemas than thematic role frames. Among the most important are *operators*, *plans*, and *scripts*. Such structures can easily be constructed by composing thematic role frames. However, since this is not the focus of the model being proposed, a description of how this is accomplished is omitted.

5.3.2 Discussion

Representing cross-modular semantic distinctions. As discussed in section 4.2.3, the modular mentalist approach cleanly represents fine conceptual distinctions that many logical form and realist semantic systems do not make, by spreading the representation across multiple modules. We considered the problematic notion of "co-agency" sometimes proposed to account for the difference between sentences like

(5.9) Franny, together with Zooey, went off to college.

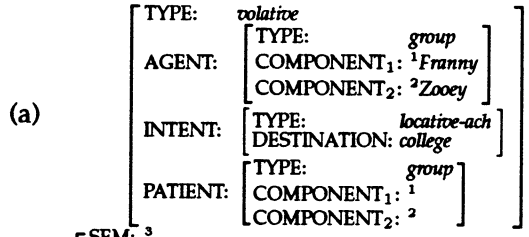
(5.10) Zooey, together with Franny, went off to college.

Instead of trying to force two AGENT roles into the same semantic frame, I distinguish two levels of interpretation: a conceptual group-agent interpretation and a lexicosemantic separate-agent interpretation. The same conceptual group-agent interpretation is automatically inferred for both sentences—*Franny and Zooey went off to college*—in which *Franny and Zooey* is a group AGENT,[24] as the fragment structure in figure 5.15(a) shows. However, the lexicosemantic representations differ, as shown in figure 5.15(b). For sentence (5.9) *Franny* is the AGENT of the sentence's primary event frame *Franny went off to college*, while *Zooey* plays only the LMTWO role of the secondary conjunctive-event frame modifying the primary event. In the case of sentence (5.10) the roles would be reversed. The cross-modular distinction captures the difference between the sentences' meanings, while the common conceptual representation captures their close similarity.[25]

[24] To be more accurate, the *Franny and Zooey* concept serves as both AGENT and PATIENT since the *go* action is reflexive.

[25] Another way to look at this solution is that the conceptual module holds what Norvig (1989) calls the result of "interpretation by compatibility", which is the shared extensional interpretation, while simultaneously using the lexicosemantic module to hold a different, though logically compatible, intensional interpretation.

(a)



(b)

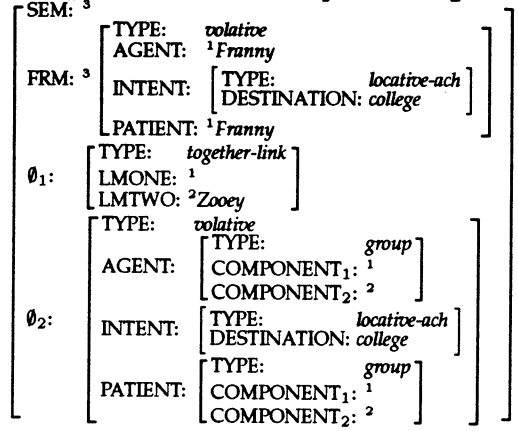


Figure 5.15: (a) Fragment structure for the concept *Franny and Zooey went off to college*, and (b) combined structure with lexicosemantic and conceptual interpretations.

The approach also solves the problem the notion of co-agency encounters in sentences like

(5.11) Franny, followed soon by Zooey, went off to college.

Recall that the co-agent role would not be used to handle *Zooey* because it is embedded in a reduced relative, yet by not using the co-agent role, the similarity between sentences (5.9) and (5.11) is not captured. The problem does not arise with the proposed approach, since *Zooey* plays the agent role of a secondary modifying event frame, similar to the role it played in sentence (5.9). The similarity between the "co-agent" and the reduced relative is thus reflected by a similar lexicosemantic structure.

Observe that, as exemplified by figure 5.15(a), a conceptual structure is not necessarily lexicosemantic just because it consists of primitives that are all permitted in the lexical semantics module. Confinement to lexical semantics primitives is a necessary but not sufficient condition for being a lexicosemantic structure; an equally important criterion is that it should help capture some surface grammatical pattern.

The human ability to simultaneously hold multiple representations of the same real world situations using different modules, I suspect, prevents realist semantic systems from adequately capturing linguistic usage subtleties. It explains the slipperyness of the notion of "semantic well-formedness" that some theorists have proposed in extension of Chomsky's treatment of grammaticality as "syntactic well-formedness". Intuitive judgements of syntactic well-formedness are slippery enough; intuitions on semantic well-formedness are worse yet, particularly if, as I

proposed in section 4.2.3, we can effortlessly shift representations upon demand. Still we cannot altogether throw out intuitive meaning judgements since one does not discard data just because it is noisy. The framework described here proposes to provide for the subtle semantic and conceptual distinctions that influence language use, rather than simplify or dispense with such distinctions on grounds of noisy intuitions.

Comparison with KL-ONE and KODIAK. It is difficult to compare the proposed ontological representation with traditional semantic networks. On the one hand, the structural formalisms—frames, feature structures—are nearly identical. On the other hand, the mentalist approach makes different assumptions about the intensional and extensional significance of concepts. These are discussed in chapter 6. The observations here are confined to a few more direct contrasts.

The small set of primitive roles contrasts with the proliferation of roles encouraged in KL-ONE (Brachman & Schmolze 1985), KODIAK (Wilensky 1986; Wilensky *et al.* 1988), KRYPTON (Brachman *et al.* 1983), KL2 (Vilain 1985), CLASSIC (Brachman *et al.* 1991), and their descendants. Though these representations make a variety of different assumptions, they all employ a second multiple hierarchy for defining roles, in addition to the concept hierarchy. In the proposed representation, the context sensitivity of roles' functions to the frame type effectively also defines a role hierarchy that parallels the concept/frame hierarchy. However, a frame is not permitted to have an arbitrary number of roles, for the reasons discussed in section 5.2.2.

One might be tempted to take an alternative interpretation of KL-ONE type networks, particularly in dealing with formulations like Norvig's (1987), where there are only two primitive roles that link an objectified relation to its source (DOMAIN) and destination (RANGE). In Wilensky's (1991) terms, this is the extreme version of an "aspectualized" representation because each role is made an explicit object, which can then be used to fill arguments of other predicates. For example, the HITTER of a *hit event* is made into an explicit *hitter* concept whose DOMAIN role is filled by the *hit event* and whose RANGE is filled by whatever previously filled the HITTER role. The advantage is that external predications can then also be made on the *hitter* concept, for example *hitters will be punished*, which could not have been applied to the original HITTER relation. Norvig pointed out that this leads to potentially infinite relational regress, and suggested as a rule of thumb that roles only be "aspectualized" up to the point that concrete objects are actually needed in any interpretation.

The problem with this comparison is that aspectualized roles are not thematic roles with regard to their intended motivation. They do not capture a useful generalization for keeping schemas "chunked" together for processing purposes. As I observed in section 5.2.2, just having roles labelled 1 and 2 instead of having four roles—equivalent to an aspectualized representation—does not cleanly account for evidence from cross-linguistic surface grammatical variation.

The notions of cross-modular semantic distinctions and associatively inferrable conceptual shifts are directly related to the KODIAK notion of "views" (Jacobs 1985; Wilensky 1986; Norvig 1987; Martin 1988, 1990; Wilensky *et al.* 1988). Views are as a type of structured association—what some have called "cables", or multiple parallel links relating structures—that impose a particular perspective upon a structure. Jacobs (1985) writes:

> A view is a relationship between two concepts which represents the fact that an instance of one concept may be expressed, or viewed, in terms of the other, without requiring that the instance be a [sic] instance also of the second concept. ... *Actions may be*

> *viewed as transfer-events, with the actor playing the role of the source, the object playing the role of recipient, and the action itself playing the role of object.* (pp. 45–47)

In Jacob's treatment views are intended primarily to model linguistic metaphor, though the line between linguistic and conceptual metaphor and analogy is by no means clear. The later treatments, especially Martin's, use views for entirely conceptual purposes. Views also capture the flavor of multiple, highly associated interpretations for the same utterance (or multiple, highly associated representations of the same situation). Whereas views are purely declarative structural entities, here I am drawing, for both cognitive modelling and computational tractability reasons, a processing distinction between multiple interpretations that are statically held in parallel and those that require sequential "flipping" between interpretations. The notion of views did not make this distinction, so for example when views were used to represent *buy* and *sell*, the issues raised above were not resolved. Norvig (1989) makes a similar processing distinction in a study of ambiguous constructions with multiple interpretations: compatible interpretations can be conjunctively held, while incompatible interpretations are disjunctive and only one can be held at a time. Here we are not concerned with incompatible interpretations; however, even for logically compatible interpretations the agent may switch serially between interpretations rather than holding them in parallel. The most pressing motivation for sequential switching is to stay within storage space bounds. The storage bound explanation is consistent with my hypothesis that simultaneous interpretations are more easily held across multiple modules than within the same module (section 4.1.3).

Cognitive categories and basic objects. Research in cognitive psychology has documented extremely important effects in the kinds of concepts humans develop and use. The work of Rosch and her collaborators (Rosch 1973, 1975; Rosch & Mervis 1975; Rosch *et al.* 1976) showed that cognitive categories are graded, with some members being better or more typical examples than others, the best examples being "prototypes". For example, the prototype *road* might be a *two-lane paved concrete road*, while *gravel alleys*, *freeways*, and *dirt roads* constitute more marginal examples, and *boardwalks*, *freeway exit loops*, and *bicycle paths* are barely members. They also observed that categories come in a range of specificity levels like *conduit*, *way*, *thoroughfare*, *road*, *highway*, *freeway*, *Interstate*, and *Interstate 980*, in the midst of which there are categories like *road* corresponding to the nouns children learn first, that are the most distinct from each other by perceptual and functional criteria, and that are frequently monosyllabic in languages with polysyllabic lexemes. Rosch postulates that a line of "basic objects" cuts across all category hierarchies, establishing "basic level" categories. Categories more abstract or specific than basic objects are called "superordinate" and "subordinate", respectively.

Note that the psychological use of the term "category" differs from the *category* concept/frame type which represents a conscious, explicit conception of category-hood (say, as a group or set) in the mind of the agent. A "category", in the terms of the proposed framework, means a collection of concepts/frames that are associated with each other in the sense that (1) concepts from this category are frequently used by the agent in similar contextual situations, and (2) empirical studies designed to elicit similarity judgements from subjects reveal greater similarity between these concepts.

Categorization effects are not primarily modelled by frames, schemas, or any other declarative structure in the proposed model. This approach departs from traditional AI and psychological models, and is more similar in spirit to connectionist and neural net models. There

is no notion of "prototype links" or "default slot values". Rather, prototype and default effects emerge from probabilistic association between correlational terms. This allows inferences to be more sensitive to the context. Roth & Shoben (1983), for example, showed that even basic object effects are sensitive to the linguistic context of the categorization task. Similarly, Barsalou (1985) found that subjects use different measures of graded membership (central tendency versus ideals) depending on the context. Such results argue against models where each category has a single prototype exemplar, with category membership being a simple function of Wittgensteinian "family resemblance" to the prototype, as originally suggested by Rosch & Mervis (1975), Rosch *et al.* (1976), and Tversky (1975). Instead, categories are better thought of as associated collections of concepts; Lakoff's (1987b, 1987a) "radial category" structure is an informal version of this approach. A similar argument is advanced by Medin & Wattenmaker (1987), who suggest that concepts are embedded in theories, i.e., what we might think of as a concept's "external structure". The relationship of a concept to other concepts makes the categorization effect sensitive to context. Default values are generally not context sensitive; in order to make them so, each default would have to be made an explicit conditional instead of a single value. Normally, if slot A is sensitive to slot B, the converse also holds, so both defaults would have to be made sensitive to the other. A combinatoric storage cost explosion can easily arise from creating conditionals to handle every potentially relevant combination of contextual dependencies. In contrast, as we will see, the probabilistic correlational term representation does not require reflexive interdependencies to be redundantly stored.

I do not draw a clean division between taxonomic and goal-derived categories (Barsalou 1985, 1987), but the contrastive notion is useful. Taxonomic categories are the things we saw above like *bird* and *fruit*; sometimes these are claimed to be "natural" categories. An example of a goal-derived category might be {*California, Hawaii, Hong Kong, Caribbean, ...*} which are all *places to vacation*. Other examples include *things to eat on a diet, birthday presents,* and *things to pack in a suitcase*. There are two ways in which this distinction relates to the proposed model. First, the notion of the "naturalness" of a taxonomic category is captured to a large extent by its closeness to the mental image module (and even more eidetic or vivid modules). Especially for basic-level objects, perceptual and functional features are the most important determinants of category membership (Rosch *et al.* 1976); these categories are the first to be learned. Second, there are two ways to associate concepts: through sharing of features, and by incorporating them as substructures in common frames. Features (and feature combinations) represent a much stronger association because the category membership becomes an intrinsic part of a concept's definition. We can assume that features are evolved by an adaptive mechanism when the frequency of a category's use warrants it. Many "natural" categories are used frequently and thus tend to be represented by features, even outside the mental image module. Frequently-used goal-derived categories are also represented by features; however, goal-derived categories tend to be more sparsely used, the extreme case being what Barsalou (1983) called "ad hoc categories" like *ways to escape being killed by the Mafia* and *things that could fall on your head* which are created in the process of solving a novel goal. The association between members of such categories is not strong enough to warrant designating an explicit feature.

5.3.3 Approaches to Constructing a Conceptual Hierarchy

In the work presented here, actual conceptual features are primarily collected from the corpus. That is, the overwhelming majority of conceptual features are derived in the course of

manually analyzing a corpus of nominal compounds (see section 7.1.2). This approach has proved adequate to demonstrate the ideas being presented. However, in the long run any serious computational model *will* require a conceptual system containing enough general world and cultural knowledge to interpret the sort of nominal compounds found in real text. Short of building an autonomous situated learning agent, there are a couple of possible approaches to constructing a hierarchy:

1. *Lexicographical taxonomies.* The work of lexicographers can be exploited as a starting base for constructing conceptual hierarchies. For example, WordNet (Miller 1990; Miller *et al.* 1990; Gross & Miller 1990) is an online thesaurus resource containing hand constructed cross-indexed entries. The major categories in a thesaurus closely resemble many categories found in conceptual models. Another approach is to use automated methods to extract hierarchies from machine-readable dictionaries (Chodorow *et al.* 1985; Guthrie *et al.* 1990); see section 7.5.2.

2. *Intensive manual analysis.* Hopeless as it seems, the alternative of hand-constructing entire ontologies for all the micro-domains in the world has until this point yielded more results than any other approach. The CYC project (Lenat & Guha 1988, 1990) is currently the largest and best known of these efforts. The major problems one might foresee with this approach are (1) the arbitrariness and variability of conceptual structures, particularly as many different "knowledge engineers" are involved in constructing different pieces of the knowledge base, and (2) it may be impossible to efficiently index the immense number of microtheories being produced, since the "parallel knowledge engineering" approach does not inherently and continuously produce generalizations that might be needed across microtheories.

Also see the discussion in section 7.5.2 of acquiring statistics for semantic and conceptual categories.

5.4 Integrating Syntactic and Semantic Constraints

The primary characteristics of the grammatical approach, which is based on Construction Grammar (Fillmore 1988), are:

1. It is structuralist rather than transformational.

2. The representation for syntactic and semantic structures is uniform.

3. Complex mappings are permitted between lexicosyntactic and semantic structures.

I discussed the first point in the introduction to this chapter. The semantic structures we have been seeing throughout this chapter are examples of the feature-structure notation; as we see below, such structures are usually used to encode syntactic structures. We now examine the second and third characteristics.

5.4.1 Uniform Syntactic and Semantic Representation

Syntactic and semantic constraints are encoded using a common attribute-value notation. The same unification operation thus applies to both. This permits a straightforward probabilistic treatment of tightly interactive syntactic and semantic processing, because the same hypothesis choice methods can be used to simultaneously select the best parse and interpretation. Though I am not concerned with giving a specific theory of search, in the long run the same control mechanism could also be used for both. Jurafsky (1990, 1991, 1992) argues for an interesting approach along these lines, drawing extensively from psycholinguistic data.

We have seen many examples of semantic structures in this chapter; a few examples of syntactic constructions are given here. As with the conceptual system, the emphasis in this work is not upon the specific syntactic structures to be used in a full grammar, but rather upon the nature and form of the structures. The number of abstract syntactic categories needed to process nominal compounds is minimal (though a large number of lexical entries are needed).

(a) [TYPE: *"firm"-N*]

(b)
$$\begin{bmatrix} \text{TYPE:} & NN \\ \text{CONST1:} & [\text{TYPE: } N] \\ \text{CONST2:} & [\text{TYPE: } N] \end{bmatrix}$$

(c)
$$\begin{bmatrix} \text{TYPE:} & NN \\ \text{CONST1:} & [\text{TYPE: } \textit{"business"-N}] \\ \text{CONST2:} & [\text{TYPE: } \textit{"firm"-N}] \end{bmatrix}$$

(d)
$$\begin{bmatrix} \text{TYPE:} & NN \\ \text{CONST1:} & [\text{TYPE: } N] \\ \text{CONST2:} & [\text{TYPE: } \textit{"firm"-N}] \end{bmatrix}$$

Figure 5.16: Some lexicosyntactic constructions, for (a) the lexeme *firm*, (b) the general noun compound, (c) the bound phrase *business firm*, and (d) the lexicalized pattern of a noun followed by the lexeme *firm*.

Lexicosyntactic constructions of varying specificity are shown in figure 5.16. In (a) the syntactic construction for the single lexeme *firm* is shown. Normally there would be other syntactic agreement features, but as this is inessential to our purposes they are omitted. In (b) the general noun compound construction is shown. The CONST roles specify syntactic constituency. The digits 1 and 2 are a notational device for indicating ordering relations. In (c) the construction for the bound phrase *business firm* is shown. It resembles (a) in that all constituents are specific lexemes, and resembles (b) in that it is a composite structure. An intermediate case is shown in (d), where one of the constituents is a fixed lexeme but the other can be any noun. Such constructions are useful for specifying subregularities, where the fact that the head noun is "firm" biases the preferences for the relationship between the head and modifier nouns (see section 7.4.3 for more discussion).

5.4.2 Representing Signification Mappings

In Fillmore's (1988) framework, the general term "construction" applies to both syntactic and semantic structures, as well as structures incorporating both plus signification mappings. To avoid confusion, I will refer to the latter as *signification constructions*. Figure 5.17 shows signification constructions representing different categories of what nominal compounds can signify, again at various levels of specificity.

Neither LFG nor HPSG, nor for that matter de Saussure, permit structured mappings of this form; in those frameworks, signification mappings only relate individual lexemes to concepts.

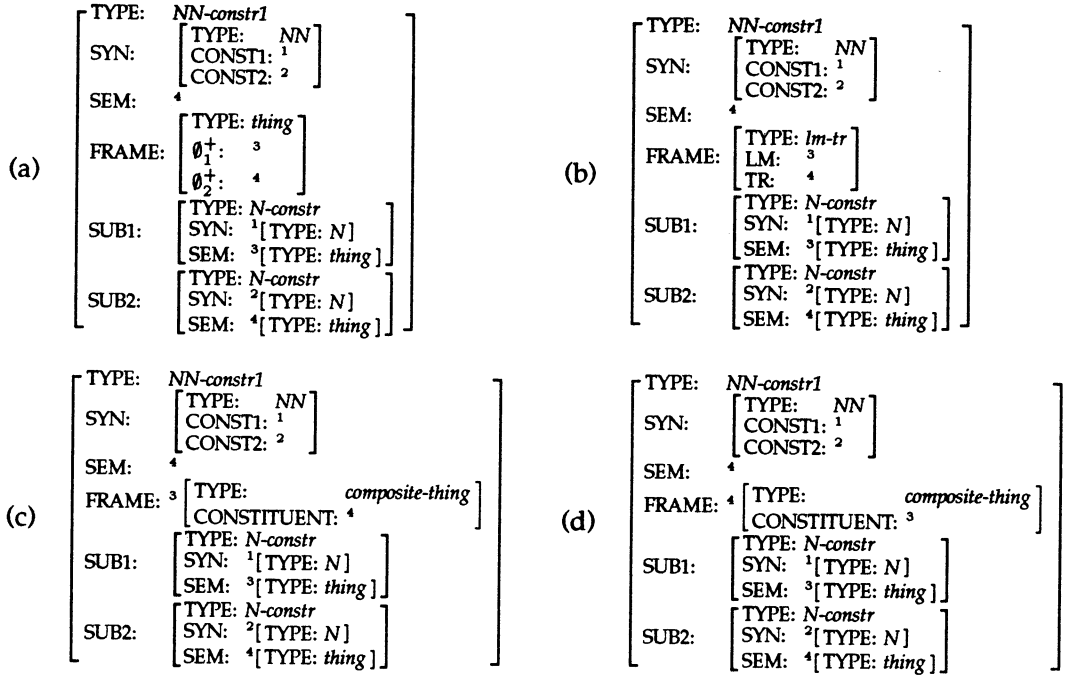


Figure 5.17: Some signification constructions for nominal compounds, where the relation between the nouns is (a) general co-occurrence, (b) co-occurrence in a *lm-tr* image schema, (c) visuospatial constituency, and (d) reverse visuospatial constituency.

In this respect, the proposed approach more closely follows Wilensky & Arens's (1980) approach of "pattern-concept pairs", which also related syntactic and semantic *structures*, though a uniform notation was not employed for both.

Lexical redundancy. Unlike standard transformational approaches, my approach encourages redundant structures when they help capture statistical distribution information, even if the structures parse and interpret utterances that could also be handled by other smaller, compositional linking rules. Because my emphasis is on evidential interpretation, Occam's razor is wielded against grammar complexity in a different fashion than in traditional linguistics. Statistical independence, rather than lexical and linking rule nonredundancy, is my cutting edge. For now I simply point this out; the point is treated in section 7.4.6.

Chapter 6

Chapter 6

Knowledge Representation

This chapter discusses representation issues involved in storing structures of the sort put forth in the previous chapter. The previous chapter proposed various knowledge structures without considering how automatic inference could operate upon them. I now turn to the question of how such structures could be represented and stored so as to facilitate evidential interpretation.

One would like to leverage existing AI representation language techniques to the largest extent possible, for the storage efficiency of inheritance, and because many of their mathematical properties are understood. For this reason, I begin by looking at semantic networks, which, being often called "associational representations", are most likely to fit the bill. However, these existing languages do not store the probabilistic information needed for evidential interpretation. Examination of the way such networks are structured reveals difficulties in augmenting them to encode the requisite information in a clean, manageable manner, and I suggest some network organization principles that (1) make the representation more manageable, and (2) are well-suited to feature-structures of the sort we have been seeing. Moreover, I argue that because existing "associational representations" lack probabilistic information and structure, the traditional hierarchy of semantic network types loses the original associationist motivation.

6.1 The Correlational Level

In this chapter I propose a *metarepresentation* approach to tackle what I call the *heuristic gap* problem in semantic network representations. Roughly speaking, the problem is reflected by the tension between old-style associational networks and the newer KL-ONE style of propositional network. The associative nature of network representations makes them elegant for modelling automatic inference, while controlled inferences are better expressed using propositional languages. The differences between these approaches are examined in this section.

Interestingly, the use of graph notation for KL-ONE has dropped as KL-ONE has become increasingly formalized in propositional terms. Though KL-ONE links originally had the connotation of association between concepts, the nature of associativity was so ambiguous and imprecise that any hint of associativity was eliminated from the various formal semantics that have been proposed. Graphs and links thus because less useful, in the interest of precision.

I believe the propositional approach to be useful but not adequate alone. The popularity of semantic networks over straight propositional languages is due largely to their intuitive appeal

Level	Primitives
Implementational	Atoms, pointers
Logical	Propositions, predicates, logical operators
Epistemological	Concept types, conceptual subpieces, inheritance, and structuring relations
Conceptual	Concepts, semantic or conceptual relations (cases), primitive objects and actions
Linguistic	Arbitrary concepts, words, expressions

Table 6.1: Brachman's (1979) levels of semantic networks.

as associative models, a trait that propositional models do not share. Simply reducing semantic networks to propositional logic by abstracting away their associational character is throwing out the baby instead of the bathwater. The alternative, throwing out the bathwater, is to concede that semantic networks characterize certain kinds of knowledge less cleanly than propositional logic, and to use a more propositional representation to model controlled inference. In exchange, the associativity can be formulated more cleanly, allowing us to maintain an intuitive interpretation of links that, at the same time, is rigorously grounded.

6.1.1 The Heuristic Gap

The propositional network position is laid out most explicitly in Brachman's (1979) influential classification of semantic networks, and most semantic network research of the past decade is based on the assumptions contained therein. Brachman posits the five levels of semantic networks summarized in Table 6.1: implementational, logical, epistemological, conceptual, and linguistic. A semantic network can be classified by the kind of interpretation assigned to its links; the network is implementational if links represent pointers in memory, conceptual if links represent case relations, and so forth. The critical claim is that each level should ideally support any valid system at the next higher level; for example, an adequate epistemological level should be able to support any legitimate conceptual system. Thus Brachman's stratification is reductionistic, in that it claims that any system is reduceable to the lowest (implementational) level by successive transformations through each level.

The stratification follows a tradition of distinguishing "epistemological" and "heuristic" areas of research, first proposed by McCarthy & Hayes (1969; see also Hayes 1979). The former concerns the epistemological power of the knowledge representation language, that is, what types of information are available, and what rules compute legitimate conclusions. The latter determines what heuristic interpretation is given to the language, meaning such things as what search order

and matching algorithms are most efficient. Association is considered a search issue and thereby heuristic.

Brachman's stratification relegates the problem of heuristic associativity to the implementational level, so that heuristic issues are already abstracted out of the picture at the logical level. However, in using semantic network models we customarily interpret links at the conceptual and sometimes linguistic levels to indicate heuristic associativity. Thus the levels in the stratification where links indicate heuristic associativity—the implementational, conceptual, and linguistic levels—are separated by a gap—the logical and epistemological levels—where links are by definition devoid of heuristic interpretation. This heuristic gap makes it difficult to see how the conceptual level can be reduced to the implementational level; something in our intuitions about semantic networks is missing in the classification.

6.1.2 A New Reductionist Classification of Semantic Networks

To address the heuristic gap I propose a reductionist classification that includes *probabilistic* and *correlational* levels. Again, each level can be reduced to the primitives of the next lower level, and each level is an abstraction of the previous one. Table 6.2 shows the stratification along with Brachman's for comparison. Although superficially similar to Brachman's, the new stratification fundamentally shifts the assumptions about the nature of semantic networks.

Stated more explicitly, the assumptions behind the classification are:

1. Associativity between concepts should not be directly encoded by links at the conceptual level, but should rather be an emergent property from the implementational level. This assumption is Brachman's also.

2. A "concept" is something that corresponds to our intuition of what a conceptual category is, i.e., something that exhibits prototype and default effects but is sensitive to the context. I cannot tell whether Brachman agrees or is ambivalent with respect to this assumption.

3. Conceptual knowledge should be reduced to a probabilistic—not logical—set of sentences. Otherwise, there is a heuristic gap in the reduction.

4. The lower-level networks are better models of automatic inference, and the higher-level networks are better suited for controlled inference. Though I believe the logical$_2$ level should in principle be reduceable to an automatic inference level, in practice it is sufficiently stratified to warrant a separate modelling framework.

5. A *correlational term definition* is a tool for explicitly describing what the investigator conjectures are intrinsic representation and access characteristics of internal structures actually used by the cognitive mechanism, as described in this chapter.

6. Two ontologically different sorts of "concept definitions" are distinguished, where Brachman only acknowledged one. The two types of concept definitions correspond to whether the definition is used for automatic or controlled inference. An *reflective definition* is a kind of meta-knowledge about some concept that is available to the controlled inference mechanisms. On the other hand, an *intrinsic categorical definition* is that which accounts for prototype, default, and context-sensitivity effects in recognition tasks. As such, intrinsic categorical definitions are the ones that count for automatic inference.

	Level/Primitives	Brachman(1979)
I	**Implementational** Atoms, pointers	**Implementational** Atoms, pointers
II	**Probabilistic** Propositions, predicates, logical operators probabilities	**Logical** Propositions, predicates, logical operators
III	**Correlational** Correlational terms, types, inheritance and structuring relations, probabilities	**Epistemological** Concept types, conceptual subpieces, inheritance and structuring relations
IV	**Conceptual** Concepts, semantic or conceptual relations (cases), primitive objects and actions	**Conceptual** Concepts, semantic or conceptual relations (cases), primitive objects and actions
V	**Linguistic** Arbitrary concepts, words, expressions **Logical$_2$** Propositions, predicates, logical operators	**Linguistic** Arbitrary concepts, words, expressions
VI	**Epistemological$_2$** Concepts, conceptual subpieces, definition frames	

Table 6.2: A classification of semantic networks with heuristic associativity.

7. The purpose of the epistemological$_2$ level is to represent reflective definitions.

8. The notion of an intrinsic definition is purely for conceptual purposes. Unlike correlational term definitions, intrinsic categorical definitions are not represented explicitly at any of the levels. I suspect intrinsic categorical definitions would be combinatoric if explicitly enumerated for all concepts. Instead, concepts are implicitly defined by their position in the network. Though any concept's intrinsic categorical definition *can* be explicitly extracted from the network structure if desired, what is of greater interest is how the network structure as a whole influences automatic inference, i.e., how placing concepts (and other correlational terms that do not correspond to concepts) in the network and choosing a probability distribution over them influences the output of automatic inference.

9. The intrinsic categorical definition of a concept is extracted by taking the partition of the network containing all correlational terms that are not conditionally independent of the concept, and the probability distribution over that partition. This will become clearer in the course of the chapter.

10. The network should be structured such that the automatic inference mechanism manifests prototype, default, and context-sensitivity effects as emergent properties. This issue is not addressed in any depth in the present work; some of the examples of inference in chapter 7 can be interpreted as minimal default effects.

The approach does not attempt to encode heuristic associativity into links at the conceptual or linguistic levels. As Brachman observes, this would mix levels in a messy way. Heuristic associativity is still ultimately a property of links at the implementational level; associativity at the higher levels are emergent properties.

At the same time, we would like associativity to emerge in such a way that links at the conceptual and linguistic levels are still intuitively associational. To do this the heuristic gap must be closed by stating what the heuristic interpretation is at the intermediate levels. The probabilistic and correlational levels can express heuristic interpretations, unlike the logical and epistemological levels, because they encode a probability distribution describing usage patterns. I propose to apply much of the machinery originally developed for epistemological-level work, like terminological hierarchies and inheritance mechanisms, to the formal study of the heuristic level.

6.2 Marker Passing

6.2.1 Associative Models of Language Understanding

As discussed above, semantic networks have increasingly come to be viewed as a convenient notation for propositions, of lesser or equal expressiveness compared to various logics. The original semantic network representations following Quillian's (1969) and Anderson & Bower's (1973, 1980) were conceived of as parallel associative models. Because of the subsequent proliferation and sloppy interpretation of link types, a movement toward formalizing the meanings of links and nodes occurred in the late 70's and early 80's (e.g., Woods 1975; Brachman 1979, 1983, 1985), culminating in the development of the KL-ONE style languages whose semantics are formalized using set theory. Though such analyses are quite useful, the associational nature of

network representations is sometimes overly de-emphasized. Let us consider the associational use of semantic networks.

Since Quillian, research on spreading activation models and semantic networks has alternately diverged and converged, with investigators concentrating on either the expressiveness of multiple-hierarchy languages or the computational properties of activation networks. *Marker passing* models are the result of one convergence, and were developed after network representations had acquired more well-defined structure than early semantic networks. It is a variant of spreading activation where "markers" instead of activation values are propagated from node to node. Input is given to the model by placing markers on multiple concepts in the smenatic network; I will refer to these initial markers as *origin markers*. Markers are then iteratively propagated outward from the original concepts. When two or more markers arrive at the same node a *collision* occurs. In effect, marker propagation performs a parallel intersection search.

In a sense marker passing is a generalization of spreading activation, since markers can be used to carry numerical values. However, markers can carry symbolic information as well, so that when a collision occurs, the structural relationship between the two concepts from which the colliding markers originated can be diagnosed. Taking advantage of the more precise definitions of concepts and relations, inferences can then be made. In pure spreading activation models, there is no way of knowing where the activation arriving at a node came from; among other things, this leads to "crosstalk" phenomena where activation from an irrelevant node "leaks" into another node.[1] This advantage of marker passing models over spreading activation models, including existing neural networks, can be stated in many ways; another common perspective on the same issue is that marker passing handles *variable binding*.

Marker passing models have evolved substantially since Fahlman's (1979) original use of the term. Fahlman's NETL representation system uses marker passing as a means of finding potential relationships between concepts. Marker passing involves exhaustively propagating each origin marker to every node in a subsumption relationship with the originating node; this is repeated as needed for the type of inference being performed. Massively parallel hardware is assumed, where each concept has a processor of its own, connected to the rest of the network via the semantic network's links. The strategy results in a scaling problem since symbolic representations in interesting real world domains tend to be extremely large. Fahlman did not directly use the model for any natural language problems, and the types of inference in NETL's repertoire are probably not by themselves sufficient for semantic interpretation. More recent marker passing models attempt to solve problems that are closer in line with those addressed by evidential interpretation. Below I survey some of the most pertinent proposals.

Charniak. A marker passing model that does not expect exhaustive search is developed in a series of papers by Charniak (1983, 1985, 1986). In these models markers carry activation strengths. Marker propagation continues as new inputs are placed in the network, rather than operating in single shots over the entire semantic network as in Fahlman's model, in order to model recency-weighted context effects. An exponential time-decay factor on activation levels allows old markers to be removed. To further cut down on the number of markers propagated, "anti-promiscuity

[1] Crosstalk can actually be desirable for modelling certain confusion effects in humans (McClelland 1986, p. 139–142). However, more often it is problematic, especially when dealing with compositional structures where the distinction between disjoint substructures is normally not confused.

rules" prohibit propagation from concepts whose branching factors are above some threshold, the rationale being that such concepts are too general to provide much evidence anyhow (Charniak 1985).

In Charniak's models, the only information that markers carry besides an activation level is the path through the network over which they were propagated. After collisions have occurred a second "path checker" program eliminates erroneous collisions—Charniak calls these "false positives"—caused by markers whose paths, upon examination, show that the relationship between the origin concepts is irrelevant. This is one solution to the crosstalk or variable binding problem. The path carried by a marker essentially encodes a variable binding, since it permits the system to determine whether two markers derive from the same or different origins. Thus, in this model marker collisions generate hypotheses that must be subsequently evaluated.

Hirst. Hirst's (1987) marker passing method is similar to Charniak's. Anti-promiscuity rules are also employed. To reduce the search overhead, it restricts the number of links a marker can be propagated. Intuitively this corresponds to limiting the "semantic distance" between two concepts if they are to be associated with each other. Unlike Charniak, Hirst does not rely on marker passing to do most of the inferential work, and instead relies heavily on built-in structural linguistic knowledge.

Norvig. Marker passing is used in Norvig's (1987) FAUSTUS model to perform story-understanding inferences using a KODIAK knowledge representation. The model follows Charniak's method of checking paths after propagation. There are six types of inferences: elaboration, double elaboration, reference resolution, view application, concretion, and relation concretion. Each inference type is defined by the types of paths of the two markers participating in a collision. The path types (elaboration, ref, view, constraint, and filler) are defined by regular expressions on the link types of KODIAK.

To reduce the number of markers, lower activation is assigned to markers further away from the point of origin. Each KODIAK link type is associated with a decrement value. Origin markers are placed in the net with an initial activation value; as they are passed through the network the activation of each successive marker decreases. Below a preset threshold, markers are not propagated.

Martin and Riesbeck. The DMAP model of Martin & Riesbeck (1986) and Riesbeck (1986) differs from the preceding ones in being entirely driven by marker passing, with no path checking. As such, whenever collisions occur inferences are made, and moreover all inferences are produced by collisions. To eliminate the need for path checking, markers carry explicit variable binding information. This means that when a marker is propagated over a link, the marker's internal structure must be modified to reflect the semantics of the link. In this way when a collision occurs, all information needed to determine whether the participating markers predicate the same instances is available at the node, without tracing back over paths.

DMAP is intended to perform all parsing and inference in parallel, by marker passing alone. Collisions do not represent hypotheses since a collision results in immediate inference. Though the framework is elegant, the goal is ambitious and it is unclear that the structure of the

network can provide enough control over the propagation of markers, so as to ensure no false positives will be found before the actual desired collision occurs.

6.2.2 Problems with Marker Passing

The models described above are associative in nature. They are intended to perform inferences for semantic interpretation, many of which are of the type that automatic inference is concerned with. However, there are a number of difficulties with marker passing that make existing models inadequate for semantic interpretation.

The first, which I have touched on already, is the issue of resources required to propagate markers over an interesting semantic network of substantial size. This problem can be formulated in terms of either space or time resources. In sequential implementations marker passing is a theory of search; if the types of inferences are sufficiently powerful for semantic interpretation, the search space tends to become combinatorically large. Parallelizing the implementation only reduces the cost by a constant factor. Existing models use various heuristics to control the growth of the search space, but because of the need to "over-propagate"—to be overly liberal in propagation in order to have a reasonable chance of finding the desired concepts or paths—it is difficult to reduce the search space much. The heuristic of presetting the number of links a marker can be propagated limits the space to an acceptable size, but in practice the number of links does not reflect semantic distance in any well-defined way, thus leading to unreliable hypothesis generation. A similar caveat applies to anti-promiscuity rules and to Hendler's (1988) method of searching a more constant number of nodes by making the propagated activation value inversely proportional to the branching factor.

Second, existing marker passing methods are overly sensitive to notational variants in the underlying semantic network. Because the same propositional information can be represented in many ways, semantic networks can be varied in many ways without changing their content. An arbitrary number of intermediate abstraction levels, for example, can be introduced between any two concepts. The effect of a notational variant such as this is to alter the knowledge base's indexing. Of course changes in indexing should indeed affect the behavior of marker passing if a semantic network is considered an associational representation. However, semantic networks are generally used in a semi-propositional fashion, and close attention is not paid to ensuring that the links really provide the correct indexing. Indeed, one of the primary advantages of semantic network representations is that they require less attention to the particulars of network connectivity than the more neurally-oriented approaches, so long as concepts are correctly placed in the subsumption hierarchy. Given this, current marker passing approaches do not adequately accommodate the degree of notational variance one should expect in a semantic network.

Third, the significance of activation is not well defined in existing marker passing models, in particular with respect to the distinction between using activation to control propagation (as above) versus as a measure of belief. The mathematical behavior of activation in many types of neural networks has been studied extensively, and interpretations can be given to the numbers. Shastri (1988b, 1988a, 1989) uses activations whose interpretations are Bayesian probability values. Pearl's (1988) belief networks, while not neural networks, also compute probability values. This kind of use of activation, as a certainty level, is also found in marker passing models. Charniak's (1986) measure of "path strength" is computed from the activations of markers on a path, and the inference mechanism is only permitted to consider paths with strengths of the highest order of magnitude. Alshawi (1987) uses a set of "context factors" that influence activation in various ways,

determining in effect the evidence combination function. However, up to this point such methods have been neither theoretically justified nor empirically verified. Moreover, the relationship of certainty to search control has not been well analyzed in any of the marker passing models, and the various activation mechanisms are only justified at an intuitive, qualitative level.

In summary, marker passing in semantic networks is an attractive framework for associative semantic interpretation because variable bindings can be maintained. However, I am not convinced that the information and/or structure in typical semantic networks is adequate either to gauge the certainty of inferences hypothesized by collisions, or to efficiently guide the process of generating collisions and hypotheses. These problems with marker passing led me to propose the more rigorous use of probabilities (Wu 1989), as discussed below.

6.3 The Need for Probabilities in Semantic Networks

Towards addressing the issues of certainty and search control, note that both are related to the idea from chapter 3 that a concept is *useful* to the agent for some class of tasks. That is, the point of inferring some intermediate conceptual structure that was not present in the input is, as stated formulaically by equation (3.12), to maximize the chances of that structure turning out to be useful to the agent in its interaction with the environment. The more certain we are that some concept is liable to be the one that proves useful, the more important that the search method find it, and the higher the certainty of the concept.[2]

Thus some way is needed to incorporate information into the knowledge representation about the conditional probability of some concept being useful, given the other concepts in the input. Note that this goal differs from that of probabilistic logics (e.g., Nilsson 1986), in that it does not concern the conditional probability of some concept being true in the world. In the following sections we explore some of the difficulties encountered in augmenting semantic network representations with the requisite probabilities, due to their underconstrained structure.

6.3.1 Inelegance of Standard Semantic Net Organization

The subsequent discussion makes use of the example micro-domain introduced in figure 6.1. It is a simplified encoding of the knowledge that the steering wheel controls an auto's front wheel, that the front wheel controls the direction of movement, and that the scenario of a curbed wheel includes the front wheel as a component.[3] The network fragment is intended to be representative of typical semantic network organization (and thus does not yet conform wholly to the ontology of chapter 5). I assume for the time being that the network is simply a graphic notation equivalent to a restricted propositional language, with no associational characteristics implied.

The main organizational principle in the typical semantic network might be described as "centralize all information known about a concept". Thus, in the example all knowledge about front wheels is directly linked to the *front wheel* node.

[2]Note that whether a concept is useful to the agent must be evaluated with respect to the entire set of environmental situations that would give rise to the same inputs to automatic inference.

[3]To park a car on a slope safely, the front wheel must be turned against the curb so that the car's weight rests on the curb if the brakes should fail.

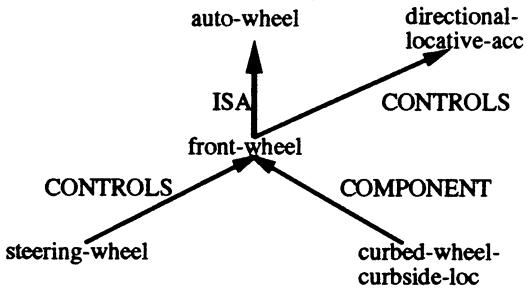


Figure 6.1: A standard semantic network organization.

Now suppose that the *front wheel*[4] concept comes into the agent's context through a linguistic utterance or otherwise, and thereby appears as input to automatic inference. It will be necessary to know the conditional probabilities of the other concepts being useful, given *front wheel*. Suppose that in the agent's past experience when the *front wheel* concept has been in use, it has proved useful 80% of the time to make the inference that the *front wheel* concepts stands in the CONTROLS relation to a *directional locative accomplishment*, 60% of the time to *steering wheel* in the CONTROLS relation, and 35% of the time to *curbed wheel curbside location* in the COMPONENT relation. The obvious way to encode the conditional dependence information is to annotate the relations with the conditional probabilities 0.8, 0.6, and 0.35, as shown in figure 6.2.

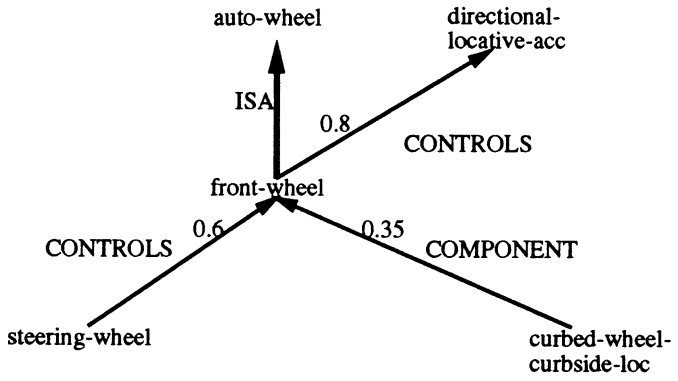


Figure 6.2: Annotating relations with conditional probabilities.

However, suppose in addition that half the time when the agent uses the concept of a steering wheel, it is because the car's front wheel needs to be curbed.[5] Then we need to encode the information that *front wheel* being a COMPONENT of *curbed wheel curbside location* has proved useful to infer in 50% of the cases when *steering wheel* CONTROLS *front wheel* has also been useful. This means that $0.6 \cdot 0.5 = 30\%$ of the time when dealing with *front wheel*, the combination of *steering wheel* and *curbed wheel curbside location* has been useful. The information must be kept explicitly since otherwise nothing would keep us from assuming conditional independence between the

[4] From Warren (1978, p. 165).

[5] Parts of Berkeley are *very* hilly.

probabilities, which incorrectly implies that the combination of inferences is only useful $0.6 \cdot 0.35 = 21\%$ of the time.

There is no single relation we can annotate as we did earlier. The only sensible alternatives, shown in figure 6.3, are to (a) annotate the *front wheel* concept itself, with explicit reference to the CONTROLS and COMPONENT relations, or (b) use a doubly-linked annotation on both the CONTROLS and COMPONENT relations. Neither alternative is particularly elegant, and the problem grows worse when there are many correlated combinations, or when we generalize to correlations of more than two relations.

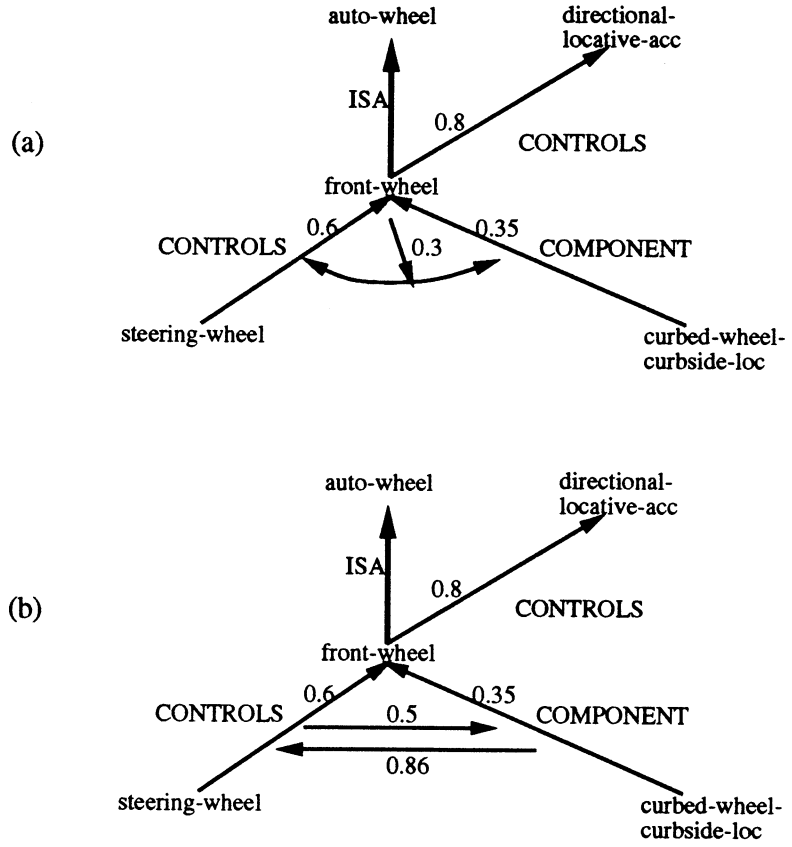


Figure 6.3: Encoding multiple correlations under standard semantic network organization.

6.3.2 Correlational Organization

The problem with the representation above is that not enough nodes, or terms, are defined explicitly, thus making correlations difficult to encode. The *correlational organization* principle says that the set of nodes in a propositional semantic network should be such that for any concept, there is a separate node for every combination of adjacent relations whose utility has been statistically correlated. These nodes are called *correlational terms*.

Continuing the example, a special node should be created to represent the combination

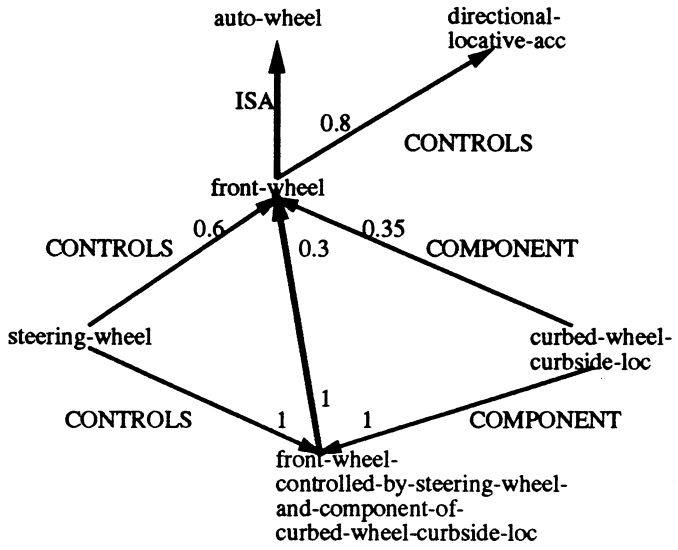


Figure 6.4: A correlational term.

of the *front wheel* concept in the CONTROLS relation to *steering wheel* and the COMPONENT relation to *curbed wheel curbside location*. Figure 6.4 shows one way this can be done. There are three points of note on the network.

First, the correlational term simply makes explicit the internal structure of the annotation in figure 6.3(a).

Second, each relation needs two conditional probability annotations, one for each direction. For example, the network tells us that *steering wheel* CONTROLS *front wheel* is useful 60% of the time when *front wheel* is being used, but it does not tell us how often the same *steering wheel* CONTROLS *front wheel* is useful when *steering wheel* is being used. To encode that conditional probability, another annotation would be needed on the *steering wheel* end of the CONTROLS relation.

The third point concerns the significance of the "1" annotations, which denote absolute conditional dependence. Their presence is an artifact of the notation in some sense. We invent a node for a correlational term purely to make it easier to assert a conditional dependency; thus, the relations that connect a correlational term to its components wholly define the term's meaning. The "1"s arise from the fact that it makes no sense to say that a correlational term is useful unless what one means is that its component concepts and relations are useful. For example, the "1" on the COMPONENT relation between the correlational term and *curbed wheel curbside location* says that *curbed wheel curbside location* is always useful if the correlational term is useful. In this respect, correlational terms are purely *terminological*. A notation that distinguishes more cleanly between terminological and non-terminological uses of nodes and arcs will be proposed in section 6.4.

Correlational terms look deceptively similar to Minksy's (1975) frames, but though correlational terms are motivated in part by similar considerations, there is an important difference. Frames contain knowledge; correlational terms do not, since they merely provide terminology for describing evidential knowledge. Correlational terms are less ambitious in their structure than frames; they have no defaults or procedural attachments. However, they do have a very specific

heuristic interpretation based on the notion of expected utility, namely, that *every* proposition implied by the correlational term is expected to be useful in the context in which the term is used. In the past it has been difficult to build a fleshed-out theory of frame-based inference because frames have only been given intuitive, inexact heuristic interpretations. The heuristic interpretation of correlational terms is more restrictive and more precise.

6.3.3 The Term Decomposition Problem

There are actually several other correlational terms that we could have chosen to define instead of the one in figure 6.4, all of which enable the conditional probability to be attached equally well. Figure 6.5 compares the options. Each alternative employs a correlational term representing a different concept involved in the correlation: (a) the original representation, a *front wheel* that is controlled by a steering wheel and is a component of the curbed wheel scenario, (b) a *steering wheel* that controls a front wheel that is a component of the curbed wheel scenario, or (c) a *curbed wheel curbside location* scenario involving a front wheel that is controlled by a steering wheel.

The problem of choosing which of the representations to use is an instance of what I call the *term decomposition problem*. More examples of the problem arise upon further investigation of figure 6.5(b). Figure 6.6 shows that the same information encoded by the correlational term *front wheel component of curbed wheel curbside location* can also be encoded by further variations. The alternatives shown are: (b) with the original correlational term, (d) with the "inverted" correlational term *curbed wheel curbside location with front wheel component*, and (e) with both correlational terms.

The *equates*, boxed in the figures, encode co-indexing of slot fillers, much as superscripts in feature structures. What the equate says is that the term that fills the CONTROLS role of the bottom node must be the same one (token, not just type) that fills the COMPONENT of the term filling the INVOLVES role.

To control the term decomposition problem, the representation syntax can be more strongly constrained. One possible criterion is *minimality*, which chooses the representation that globally minimizes the number of nodes in the semantic network. Except for the network in figure 6.5(a), all the variants require defining an additional correlational term as an intermediate step toward defining the desired correlational term. According to the minimality criterion representation (a) should be chosen. However, any of the others might have been better if additional probability information had been required concerning the usefulness of the individual inferences that *front wheel* is a COMPONENT of *curbed wheel curbside location*, or that *steering wheel* CONTROLS *front wheel*. Under minimality, the correlational terms cannot be chosen without considering the knowledge base at large since minimality is not a locally determinable criterion. Below we consider a formalism using only locally determinable constraints.

6.4 MURAL: A Metarepresentation Language for Uncertainty

MURAL is a metarepresentation language that makes predications about the agent's *use* of conceptual structures. As I mentioned above, correlational terms do not represent knowledge in the way that frames and semantic networks normally do, even though they have a framelike structure. They encode information about usage correlation patterns which, in a sense, is below the level of the actual concepts and relations chosen for the ontology. MURAL's objectives are (1)

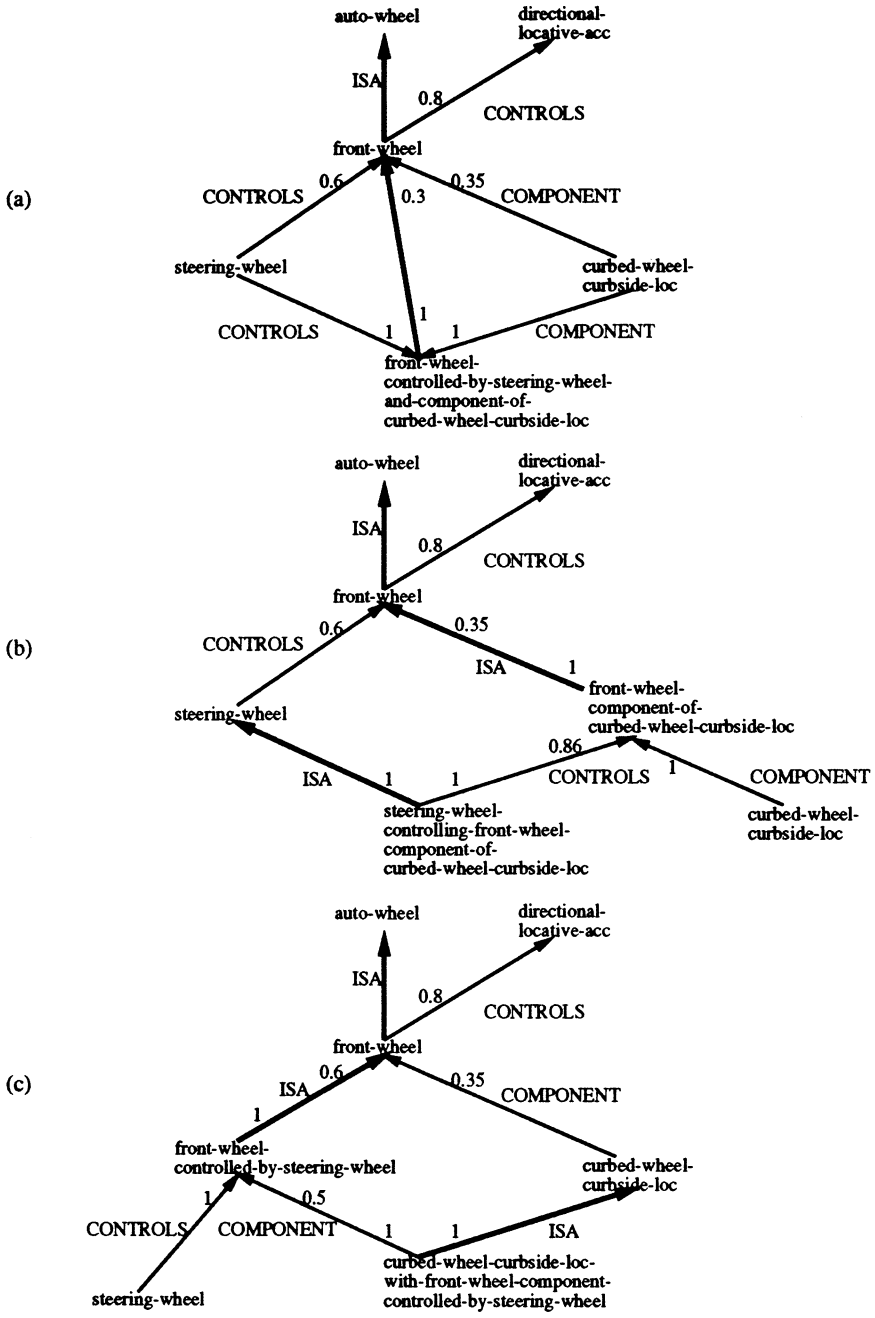


Figure 6.5: Equivalent correlational organizations.

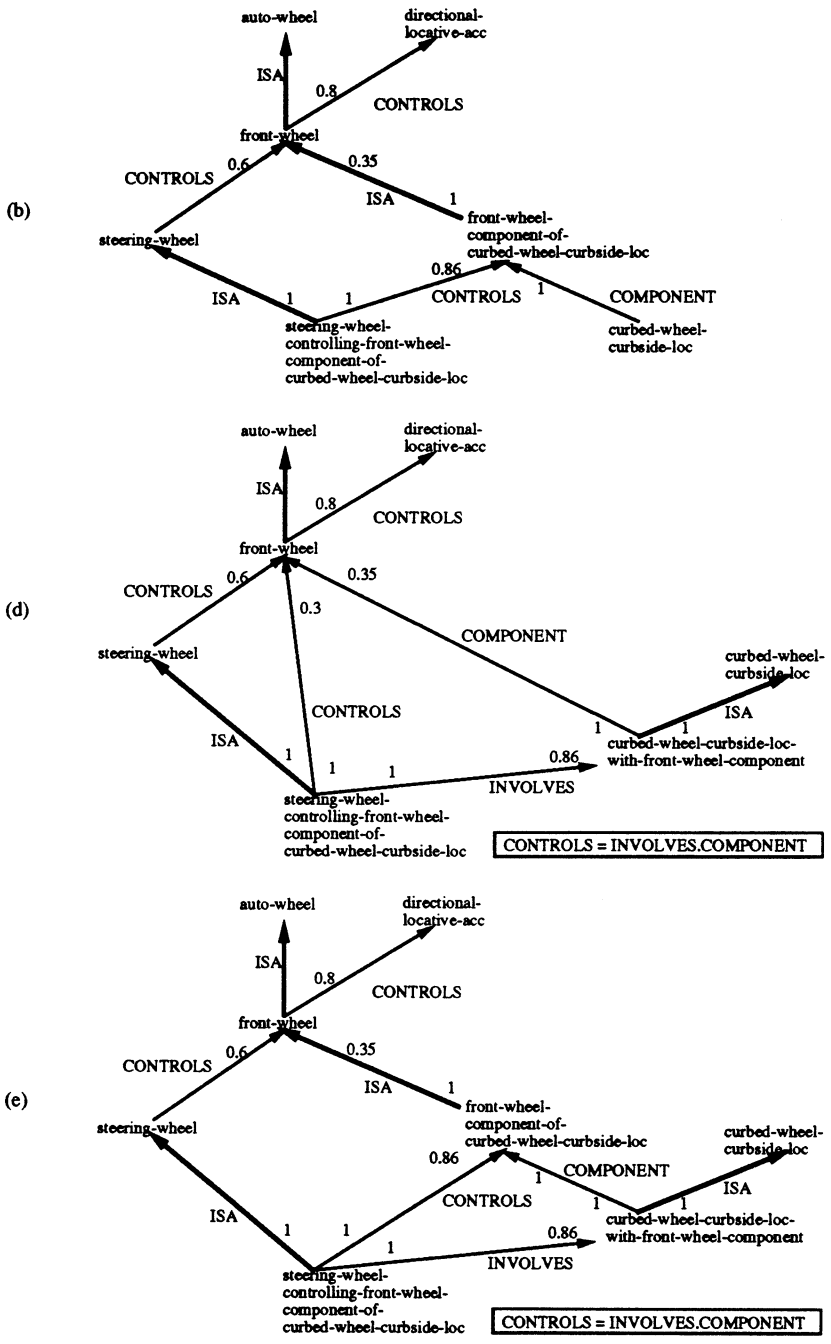


Figure 6.6: More equivalent correlational organizations. See text for explanation of the equate boxes.

to alleviate the term decomposition problem by restricting the allowable network configurations, using locally determinable constraints, and (2) to make the correlational representation reasonably transparent, so that the motivation for any term's existence is not obscured.

6.4.1 The Terminological Hierarchy

MURAL uses a abstraction hierarchy representation similar to that in KL-ONE languages (Brachman & Schmolze 1985). It follows KRYPTON (Brachman *et al.* 1983) and KL2 (Vilain 1985) in using the KL-ONE style hierarchy purely as a term-defining language, rather than as a language for asserting information about anything outside the language itself. A hierarchically-organized set of correlational terms balances storage efficiency and clarity considerations. However, it diverges from KRYPTON and KL2 in that the defined terms have a metalevel interpretation. I will return to this point after explaining how correlational terms are encoded.

Two restrictions are placed on allowable configurations. First, the *definitionality restriction* requires every relation to be a directed asymmetric operator with a "frame end" and a "filler end"—denoted graphically by an arrow on the filler end—with the "1" annotation always on the frame end. The definitionality restriction reduces notational variation by forcing a relation's correlational directionality to agree with the compositional structure. Figure 6.6(b) is not acceptable since there is a "1" on the frame end of the COMPONENT relation. Figures 6.6(d) and (e) almost meet the definitionality restriction; it is only necessary to define correlational terms for *steering wheel controlling front wheel* and *front wheel controlling directional locative accomplishment* as shown in figure 6.7. Unlike the minimality constraint, the definitionality restriction can be enforced by local examination of a concept and its substructures. The trade-off is the extra correlational terms that may be required.

The definitionality restriction allows correlational terms to be defined using a *terminological hierarchy*, for example as shown in figure 6.8. The "1" annotations in previous examples are eliminated since it is known that any relation is definitional in the direction of the arrow; thus each relation bears only one probability. Directed cycles are not permitted, so the network is guaranteed to be strictly hierarchical and to establish a partial order. As in KL-ONE-derived languages, the hierarchy consists of intertwined is-a and has-a hierarchies. I use the more technical terms *subsumption hierarchy* and *composition hierarchy*, respectively, because "is-a" and "has-a" connote claims about cognitive category structure that would be a misinterpretation of this level of metarepresentation. The subsumption hierarchy is defined by the is-a relations, henceforth called *Subsume* relations. All other relations are roles, which define the composition hierarchy.

Note that a basic distinction is being drawn between relations that are 100% correlated with a concept, and those that are 99% correlated, by saying that the former is definitional while the latter is not. There are several motivations for the distinction. First, inferring a 100% correlated relation (and the related concept) is a deductive operation, whereas inferring a 99% correlated relation is an abductive operation. Second, it makes the representation more transparent to the investigator. Correlational terms often look rather unintuitive because they do not always correspond to what we normally think of as "concepts" or senses of lexemes. Since a large number of correlational terms will be needed for any interesting domain, the network should be structured in a way that makes clear the need for any particular correlational term. The terminological hierarchy does this by arranging the correlational terms in a partial order that shows which terms are defined from more primitive terms.

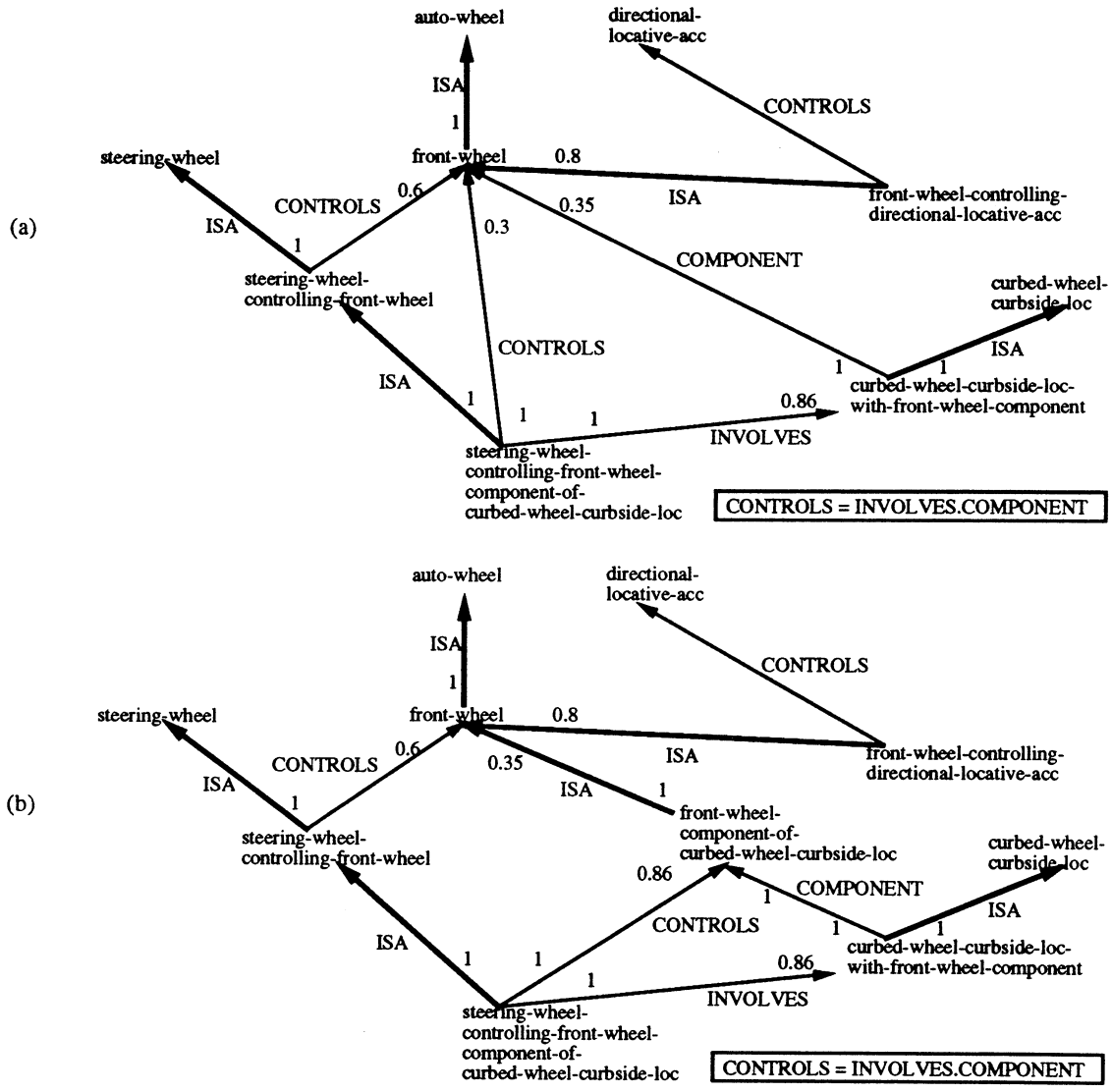


Figure 6.7: Correlational term modified to meet the definitionality restriction.

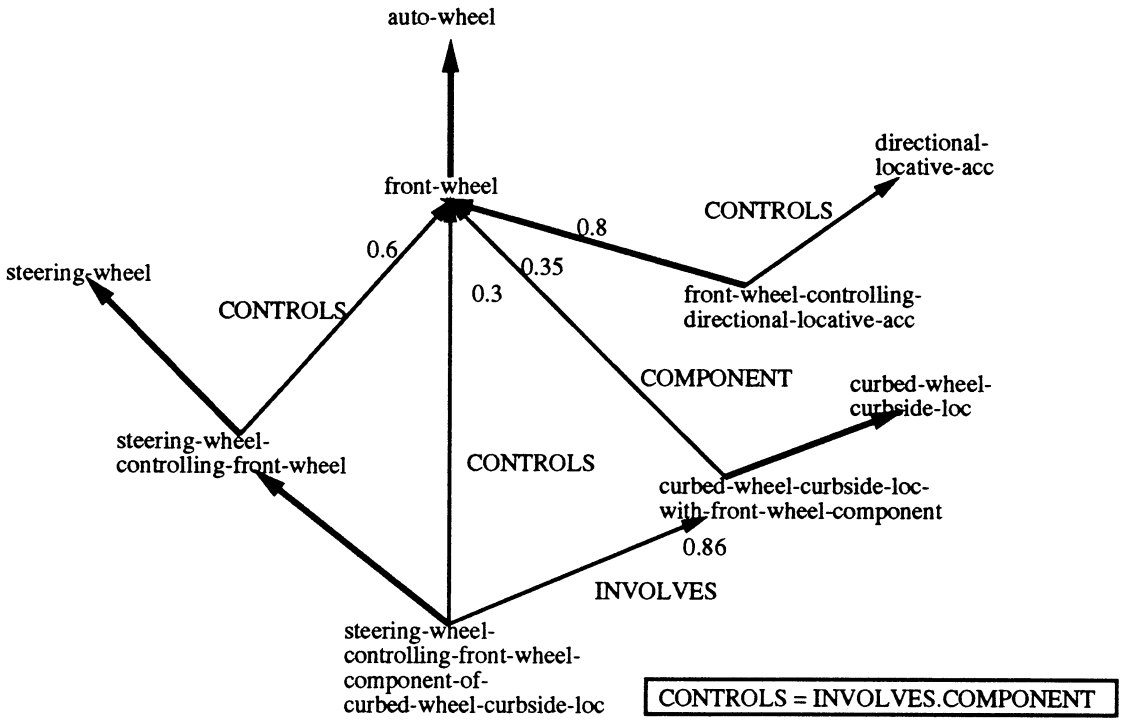


Figure 6.8: The terminological hierarchy method of defining correlational terms.

The second restriction is the *local minimality restriction*, which requires that as few correlational terms be defined as possible, subject to the definitionality restriction. This restriction disallows figure 6.7(b). To enforce this requires decomposing each term into its substructures (equivalent to traversing up the term hierarchy), marking any subterms that are necessary to the term's definition. In figure 6.7(b), *front wheel component of curbed wheel curbside location* would not be marked because it does not carry any non-redundant information necessary to the definition of its descendents (terms below it). After all terms have been processed, any unmarked terms can be eliminated.

Let us digress briefly to consider a possible objection. It might be argued that imposing the definitionality constraint merely pushes the term decomposition problem elsewhere, namely, into the problem of choosing what roles to allow. For example, figure 6.6(b) could be made legal simply by creating an inverse IS-COMPONENT-OF role as in figure 6.9. However, this problem is more easily manageable than without the definitionality restriction for two reasons. First, the definitionality and minimality restrictions together determine a single canonical encoding *with respect to any given set of roles*. Thus instead of dealing with two sources of notational variation—choice of roles and choice of terms—we only have to deal with choice of roles. Second, the choice of roles is usually a more carefully considered decision than choice of terms; for example, figure 6.9 is not likely to be allowed unless there is an independent principled reason for permitting both COMPONENT and IS-COMPONENT-OF roles. In fact, the ontology presented in chapter 5 is very restrictive in the set of allowable roles.

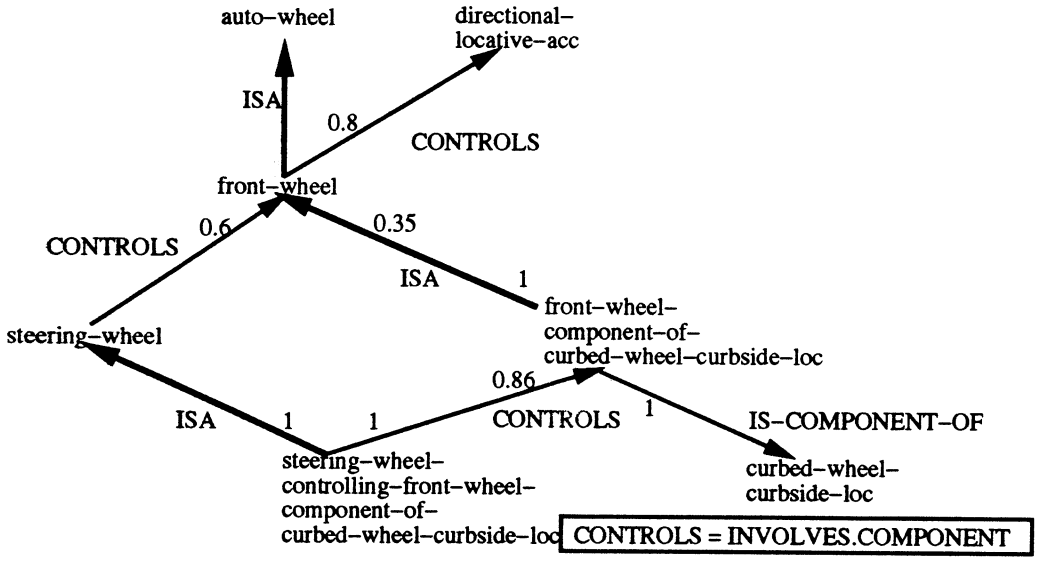


Figure 6.9: Using an inverse role to circumvent the definitionality restriction.

Returning now to the point mentioned earlier, this use of a terminological hierarchy differs from the "TBox" in KRYPTON, KL2, and related languages. The difference lies not in the formalisms, which are essentially identical, but rather in the particular interpretation of the formalism. One way to think of the difference is in terms of what the "instantiation" operation does. In KL-ONE style languages, instantiating a term does not inherently require all the subterms

comprising its definition to be instantiated as well. However, the equivalent operation in MURAL does require it, as a bit of reflection on the definition of the link probabilities reveals. The object of automatic inference is to make forward inferences, and the way a forward inference is represented is by creating an "instance" of a term to be output from the automatic inference submodule.[6] Saying that a term is "instantiated" in MURAL intrinsically means that all its substructures, which are 100% correlated by definition, are also "instantiated". To distinguish this interpretation from KL-ONE's, I refer to an *occurrence* rather than an "instance" of a term.

Note the difference between Subsume relations and non-term-defining IS-A propositions. There is a need to represent explicit IS-A propositions that the agent is considering, but which do not form a part of the agent's intrinsic conceptual representation. For example:

(6.1) The legislature considered designating Highway 17 an Interstate in order to qualify for federal funds.

(6.2) Kids think porpoises are fish.

Such non-term-defining IS-A propositions are handled in MURAL by using "propositional IS-A" frames to describe them, like the *isa-category* frame (see figure 5.7).

MURAL has no ABox. In KRYPTON and KL2 there is a separate "assertional box" or "ABox" module of the knowledge representation, that uses a language distinct from the terminological semantic network hierarchy. Putting a proposition in the ABox means it is a belief held by the agent; in other words, an attitude is expressed toward the proposition. In MURAL, most of this work is done by the probabilities. Most of the time, when an agent needs access to a proposition, it is not so much because it believes the proposition is true, but rather because the proposition is often useful. Thus at least for automatic inference, the ABox distinction is only of secondary importance.

There are indeed cases where it is necessary to be able to express the agent's explicit beliefs about the truth of propositions. This can be handled by using "propositional truth" frames to express the metaproposition that a proposition is true, roughly like saying

$$True(category\text{-}isa(porpoise, fish))$$

This begs the question, how do we know the metaproposition is true? One possible answer would be that we do after all need a special attitude-expressing primitive to make the metaproposition true, but then, why not simply use it on regular propositions? The MURAL position is that no attitude-expressing primitive is needed for the metaproposition, and instead a high prior probability should be placed on the proposition, indicating that the agent uses the frame so often that its truth is not questioned. One can come up with even more pathological cases, but even those few can be handled by more levels of metapropositions.

6.4.2 Storing Prior Versus Conditional Probabilities

Up to this point the discussion has been in terms of conditional probabilities on relations, because this is the intuitive notion we would like to capture: given that some term is being used, how useful is another related term? However, the same information can be encoded by

[6] Or, in other architectures, one could imagine creating an "instance" to be cached in a "working memory" buffer.

annotating the correlational terms with prior probabilities. Figure 6.10 translates the example from figure 6.7(a) into prior probabilities. Note that it is the relative values rather than the absolute values that are significant, because the priors are normalized. Conditional probabilities can be re-derived as needed. The conditional probability on the "frame" (tail) end of an edge is always 1 by the definitionality restriction. The conditional probability on the "filler" (arrow) end of an edge is the ratio of the frame prior to the filler prior.

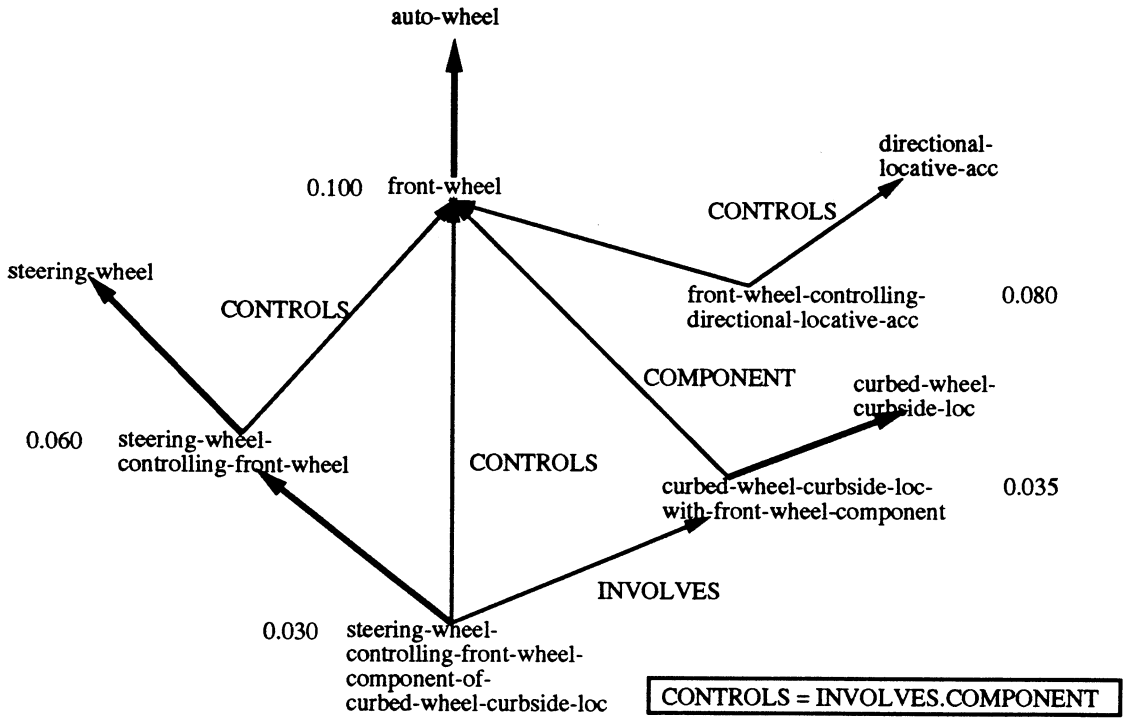


Figure 6.10: Annotating priors instead of conditional probabilities.

There are several reasons for storing priors rather than conditional probabilities. First, for semantic networks storing priors is more concise than storing conditionals, because there are always more relations than terms. If conditional probabilities are used, some annotations will contain redundant information. There are ways to get around the redundancy using inheritance or independence assumptions in conjunction with conditional probabilities, but the added complexity would not buy anything. Second, and more importantly, it provides a simple way to think about acquiring probabilities from experience, as described next.

6.4.3 Storing Relative Frequencies Versus Probabilities

Since all the probabilities in the proposed model are defined as logical or subjective probabilities, the "correct" probability values are never actually known with certainty. What can be sampled from experience are the relative frequencies of particular, *a priori* chosen classes. As a methodological simplification, my working assumption is that the correlational terms in the terminological hierarchy are chosen beforehand by the investigator, and that the agent's adaptation

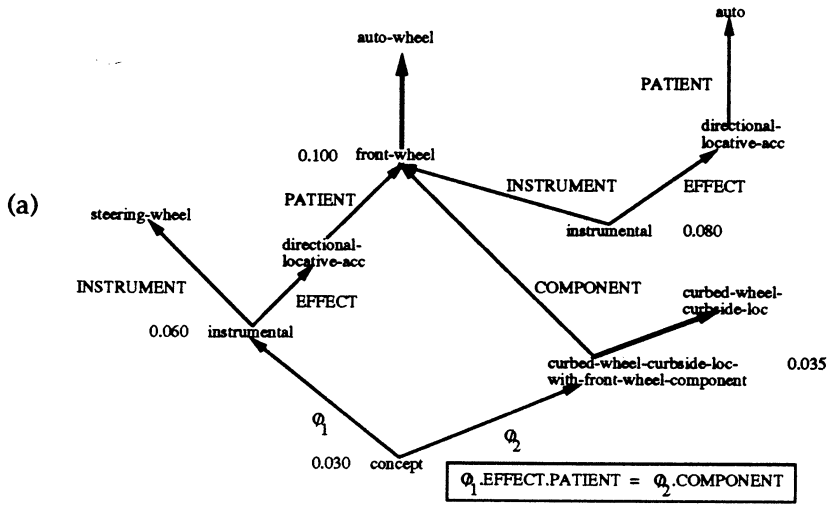


Figure 6.11: (a) Encoding a MURAL hierarchy with proper ontological roles. (b) The feature structure corresponding to the bottom node from (a).

is restricted to sampling the relative frequency distribution over those terms. These relative frequencies are used as *estimators* for the priors for those terms. In the long run, if the underlying terms were chosen correctly, the relative frequency distribution will approach the "real" probability distribution. Thus we think of adaptation in terms of updating relative frequency counts. Note that this methodology essentially bypasses the concept formation problem, i.e., the problem of having the agent itself evolve the correct correlational terms from experience. The problem is examined further in chapter 8.

6.5 Encoding the Ontology and Grammar in MURAL

For the proposed model, the choice of correlational terms will be couched in terms of the ontological and grammatical primitives described in chapter 5. The only relations permitted are subsumption, and the mental image and thematic roles. This means, for example, that the structures in figure 6.10 would be formed as in figure 6.11(a) (which, besides showing how the ontology's roles are mapped into a terminological hierarchy, also introduces the $\emptyset_i$ role type discussed below). The feature structure defined by the bottom node is shown in (b). (The *concept* type is the most abstract term at the root of the entire terminological hierarchy, equivalent to having no specified feature values.)

The terminological hierarchy provides an efficient method of storing feature-structures. First, inheritance eliminates the need to make copies of the same substructure when creating specialized versions of feature-structures. The typed feature structures I have been using take advantage of inheritance. The terminological hierarchy's subsumption relations define a hierarchy of "types", which are bundles of feature value constraints. Specifying a type name in the TYPE pseudo-slot is a way of inheriting all the feature value constraints associated with that type.

Second, the compositional hierarchy allows defining new structures out of existing smaller structures without recopying the smaller structures. For example, there is only one copy in the network in figure 6.11(a) of the *instrumental* f-structure filling the $\emptyset_1$ role in (b), and some other larger f-structure could also be defined by pointing to the same *instrumental* node.

6.5.1 Untyped Roles

The $\emptyset$ roles (pronounced "null-roles") in figure 6.11(a) are *untyped roles*, meaning that all types of roles—COMPONENT, FILL-X-COMPOSITION, GOAL, or any other role type—are subkinds of a $\emptyset$ role. Subscripts serve to differentiate multiple $\emptyset$ roles of the same frame. The $\emptyset$ roles are useful for three purposes.

First, they permit indirect correlations to be encoded. The curbed-wheel scenario is not related in any direct conceptual way to the *instrumental* frame that represents the steering wheel controlling the front wheel; only the fact that they involve the same *front wheel* relates them. Nonetheless, the heuristic association between them needs to be expressed. As we saw earlier, leaving out the bottom node and relying on *front wheel* to connect the two concepts leads to an erroneous probability, calculated by conditional independence. The reason the $\emptyset$ roles extend from an untyped *concept* frame is that, aside from general association, no more specific relation between the two fillers is worth specifying.

Second, $\emptyset$ roles permit highly abstract compositional correlations to be encoded. The example above involved an untyped frame. However, $\emptyset$ roles can also be useful to capture an abstraction over substructures and roles in more specifically typed frames. Figure 6.12 shows their use, for example, in specifying general noun compound signification constructions. The construction in (a) matches compounds where the modifying noun describes a frame structure and the head noun describes some substructure within the frame; for example, a likely interpretation of *spoon handle*[7] is a mental image of a handle that is a COMPONENT of a spoon. The $^+$ in $\emptyset_1^+$ indicates an arbitrary positive number of embedded roles, after regular expression notation; this allows covering, for example, cases of the COMPONENT of an intermediate COMPONENT. The construction in (b) matches compounds where both nouns specify substructures within a frame, for example a landmark and a trajector.

Third, in the formalism used here, $\emptyset$ roles provide the mechanism for encoding unrelated structures in the input to automatic inference. This is most useful for adding secondary structures representing the context to the main input. Section 7.4.5 explains how this could be used to model contextual priming.

[7] From Warren (1978, p. 125).

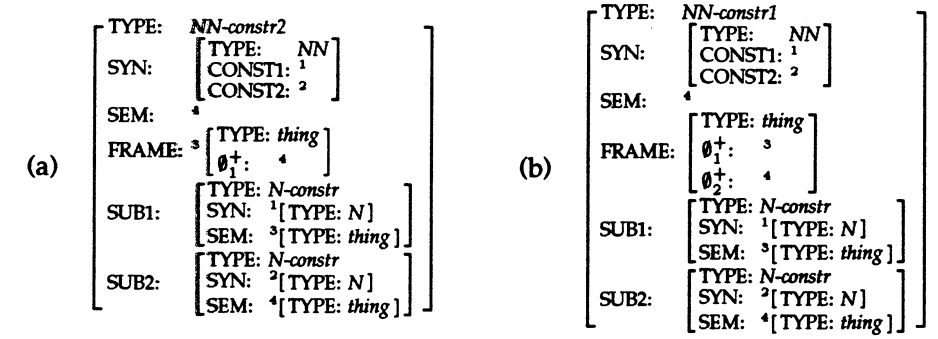


Figure 6.12: Signification constructions for nominal compounds showing use of ∅ roles (see text).

6.5.2 Constituent Ordering

As a notational convenience, order information is encoded directly in attribute labels like CONST1 and CONST2. Mathematically, the ordering can be represented using recursive nesting (much as a LISP list). Thus the abbreviated structure in figure 6.13(a) expands out to that in (b) using the special-purpose SEQ role, which can then be translated directly into MURAL.

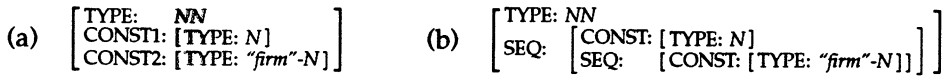


Figure 6.13: Expanding the CONST roles of the lexicosyntactic constructions in (a) to the full form in (b).

Note that this form, unlike the subscripts in section 6.5.1, is only a notational abbreviation. When unifying two structures the ordering indices must match; one can see from the expanded SEQ form that changing the order affects an f-structure's nesting. In contrast, subscripts do not encode order, so for unification the subscripts need not match. Subscripts can be useful with constituency roles, to represent mental images where an object has multiple constituents—say, COMPONENT_1 and COMPONENT_2—for which ordering constraints would be inappropriate. Subscripts are not permitted on thematic roles since there would be no clear interpretation of duplicate roles, and this would only allow roles to proliferate without any apparent gain in expressiveness.

6.6 A Closer Look at the Probability Space

Having seen how the feature structures from the ontology fit into the terminological hierarchy, we now re-examine a bit more closely how the probability distribution relates to feature structures, in preparation for evidential interpretation (chapter 7) and learning (chapter 8) with these structures.

6.6.1 Complete and Abstract Feature-Structures

To begin with it is important to recognize a kind of type-token distinction for feature-structures. Up to this point I have adhered to the convention in unification grammar work of not explicitly differentiating "type" and "token" feature-structures. However, this becomes necessary in order to establish a probability distribution over the space of possible feature structures.

The difference between *complete* and *abstract* feature-structures is not related to the syntactic form of a feature-structure, but to the interpretation given a feature-structure. A complete f-structure is a "token" in the sense that it predicates a particular "state of mind" of the agent. On the other hand, an abstract feature-structure is a "type" that describes a class of states-of-mind. Remember that f-structures (or correlational terms) have a metarepresentational interpretation in the proposed model. Such philosophical niceties are unnecessary in the conventional unification grammar use of f-structures, but here we need to know what the probability distribution over the space of f-structures means.

Different modelling architectures may realize the distinction in various guises. In the proposed model, a complete f-structure is what the automatic inference mechanism outputs. In another model, a complete f-structure might specify the current contents of a working memory buffer. In contrast, abstract f-structures, which denote classes of complete f-structures, are used by the models to store priors.

A unification operation may only be applied to abstract f-structures; to unify a complete f-structure with something else makes no sense. When two abstract f-structures are unified, the result is another abstract f-structure that denotes a narrower class of complete f-structures. Sometimes it is convenient to informally say *complete an abstract f-structure* to mean creating a complete f-structure that has exactly the structure of some abstract f-structure.

To indicate a complete f-structure, the outermost brackets are written with floor brackets as shown in figure 6.14(a). The ABRV pseudo-role is only a notational mark; in text I will often refer to a correlational term by a convenient abbreviation associated with some f-structure, so for example I would refer to the f-structure in figure 6.14(a) by ⌊*steering wheel and curbed wheel*⌋ using the same floor bracket convention. The abstract f-structure in (b) is the same feature-structure, but its interpretation is the class of complete f-structures that includes (b) as a substructure. In other words, unifying any other f-structure with (b) produces another abstract f-structure that, if completed, is in the class denoted by (b). In text (b) would be abbreviated as *steering wheel and curbed wheel* without floor brackets.

6.6.2 Lattice Structure of the Concept Space

We now turn to characterizing the entire space of possible f-structures—we can say "concept space" to stay with conventional machine learning terminology, being careful to note that "concept" here means correlational term rather than categorical concept. This concept space is the *event space* for the probability distribution.

In the basic case, the space of possible f-structures forms a *lattice*. A lattice is a hierarchical directed acyclic graph establishing a partial order over nodes, such that for every pair of nodes there is a unique *least upper bound*—a common ancestor that is closer to both nodes than any other common ancestor—and a *greatest lower bound*—a common descendant that is closer to both nodes than any other common descendant. Intuitively, the space is a lattice because for any two

(a)
$$\left[\begin{array}{ll} \text{ABRV: } \textit{steering wheel and curbed wheel} & \\ \text{TYPE: } \textit{concept} & \\ \emptyset_1: & \left[\begin{array}{ll} \text{TYPE: } \textit{instrumental} & \\ \text{PATIENT: } [\text{TYPE: } \textit{steering-wheel}] & \\ \text{EFFECT: } & \left[\begin{array}{l} \text{TYPE: } \textit{directional-locative-acc} \\ \text{PATIENT: } {}^{1} \end{array}\right] \end{array}\right] \\ \emptyset_2: & \left[\begin{array}{l} \text{TYPE: } \textit{curbed-wheel...component} \\ \text{COMPONENT: } {}^{1}[\text{TYPE: } \textit{front-wheel}] \end{array}\right] \end{array}\right]$$

(b)
$$\left[\begin{array}{ll} \text{ABRV: } \textit{steering wheel and curbed wheel} & \\ \text{TYPE: } \textit{concept} & \\ \emptyset_1: & \left[\begin{array}{ll} \text{TYPE: } \textit{instrumental} & \\ \text{PATIENT: } [\text{TYPE: } \textit{steering-wheel}] & \\ \text{EFFECT: } & \left[\begin{array}{l} \text{TYPE: } \textit{directional-locative-acc} \\ \text{PATIENT: } {}^{1} \end{array}\right] \end{array}\right] \\ \emptyset_2: & \left[\begin{array}{l} \text{TYPE: } \textit{curbed-wheel...component} \\ \text{COMPONENT: } {}^{1}[\text{TYPE: } \textit{front-wheel}] \end{array}\right] \end{array}\right]$$

Figure 6.14: (a) A complete feature structure is indicated by floor brackets. (b) The abstract feature structure denotes a class of complete feature structures that includes (a) but may also include larger or more specific feature structures.

f-structures, (1) there is a largest common substructure containing everything shared by the two f-structures, and (2) there is a smallest common superstructure that is the unification of the two f-structures.[8] Keep in mind that this lattice graph is a conceptual tool used to visualize the event space—I am not proposing to store the entire lattice, which is immense in any interesting domain.

A slightly more complicated concept space than the basic one is needed for evidential interpretation, to facilitate handling (1) complete and abstract f-structures, (2) abstraction over role types, and (3) cases where one f-structure is unified into an internal role of the other. In this formulation, it turns out the concept space forms a *semi-lattice.*[9] An example fragment of a semi-lattice is shown later (figure 6.16).

To handle (1) requires identifying the *leaf* nodes of the semi-lattice (which are the bottom layer of nodes having no descendants) with all possible complete f-structures. Accordingly, all other *internal* nodes of the semi-lattice correspond to the possible abstract f-structures. Note that this space permits only a finite number of possible conceptual structures. There are a number of different ways of bounding the number. I assume that limits are set on the number of internal slots in an f-structure; on the branching factor, which is the number of roles allowed at one level in the f-structure; and on the depth of nesting within an f-structures. Intuitively and qualitatively, such bounds correspond to the idea that the cognitive mechanism has limited resources for statically holding a conceptual structure (along the lines of restricting "working memory" size). Beyond this size, structures must be processed sequentially.

Handling (2) and (3) requires defining a number of *abstractive relations*, or *abstractors* for short, which are operators that transform an f-structure into another one that is more abstract by exactly one "atomic" difference. Abstractors can be seen as the inverse of interpretation operators: interpretation is the process of taking various input structures, hypothesizing additional structures

[8] For more in-depth introduction, see, for example, Shieber (1986, pp. 14–16).

[9] For a semi-lattice only one of the two lattice requirements applies, namely, that any two nodes have a unique least upper bound.

not given in the input (abduction), and unifying them in some appropriate fashion. The abstractors, conversely, start from all possible outputs (i.e., complete f-structures), and generate all the possible "starting points" from which the interpretation mechanism could have produced them.

In the standard f-structure space the only abstractor is *feature deconstraint*, which takes an f-structure and removes a single feature value constraint (not necessarily at the root, but at any level in the f-structure). In effect, this abstracts an f-structure's type by one increment. Consider figure 6.15(a), which employs feature-DAG for compactness instead of bracket notation (refer to figure 1.2 for the correspondence). Each box contains an entire feature-DAG. There are only two types of concepts, *a* being superordinate to *b*. In this example, feature deconstraint transforms the f-structure in the box labelled C into B. It may help to think of the interpreter proceeding in the reverse direction, in which case it is abducing that the feature value which was unspecified in C should be specified as in B.[10]

The proposed model requires six additional abstractors at a minimum, described below, without which certain useful forms of generalizations could not be captured. Other abstractors might turn out to be useful as well, but will not be pursued here.

Role detypification takes an f-structure with a typed primitive role (i.e., a role that is defined by exactly one role feature) and transforms that role into an untyped $\emptyset$ role. Figure 6.15(b) shows an example of role detypification. Again, the inverse operation can be thought of as the interpreter abducing a specific type for the role.

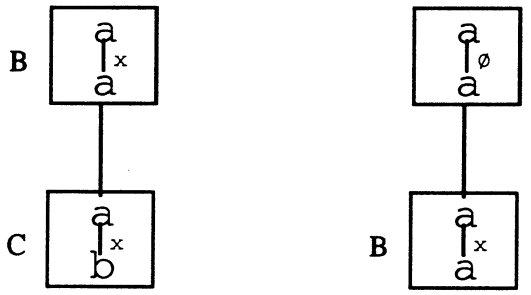
Untyped role conflation takes an f-structure with two directly nested untyped $\emptyset$ roles and collapses them into a single $\emptyset^+$ role, as shown in figure 6.15(c). In the original f-structure there must not be any other roles nested within the outer $\emptyset$ role, besides the inner $\emptyset$ role. This is how the regular expression operator $^+$ I used earlier in various abstract f-structures is formally dealt with.

Substructure partition takes an f-structure and removes a single primitive slot. To be more precise, it removes from the feature-DAG a single untyped node that is only connected to the other nodes by an untyped $\emptyset$ role. Figure 6.15(d) shows an example.

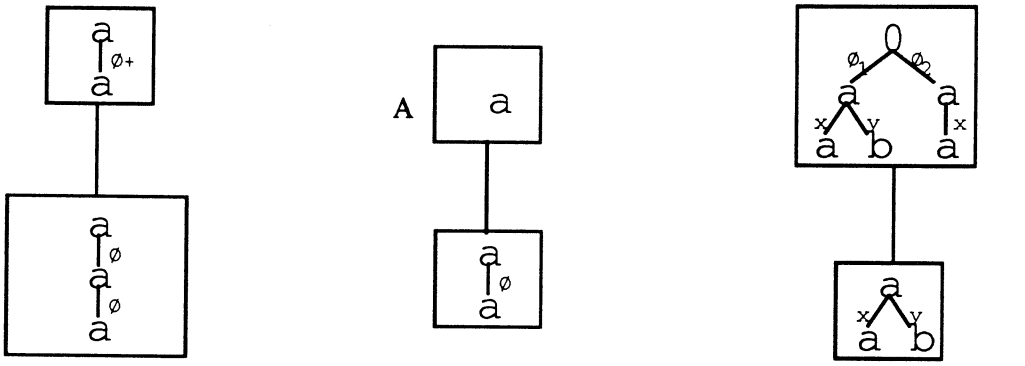
Substructure deunification takes an f-structure and produces a new f-structure containing both the original one and any substructure of the original one. An instance is shown in figure 6.15(e), where two $\emptyset$ roles are created in a new untyped frame (the *0* node) to hold the original f-structure and substructure. This kind of abstract f-structure is used in the proposed model to represent input consisting of several not-yet-connected fragment structures, the task of interpretation being to correctly relate them. The interpretation process is only permitted to unify two substructures if they stem from different $\emptyset$ roles; the only exception to this rule is described next.

Role decomposition takes an f-structure with a typed composite role (i.e., a role that is neither a $\emptyset$ role nor one defined by only a single role feature) and splits the role into two roles; the role features of the original role are assigned to either, but not both, of the new roles. The substructure that filled the original role is duplicated to fill both roles. For example, consider figure 6.15(f). There are two role features x and y, which might be things like *composition* and *-shape-fill* from figure 5.2. This yields three role types: X which would be COMPOSITION, Y which would be SHAPE-FILL, and the composite role Z defined by the combination x, y which would give FILL-COMPOSITION. Role decomposition produces D from F. The two role-filling substructures are marked as being *unifiable substructures*. This is a new notion: in an abstract f-structure, certain pairs of roles are explicitly marked as being potentially unifiable, and no other pairs of roles are

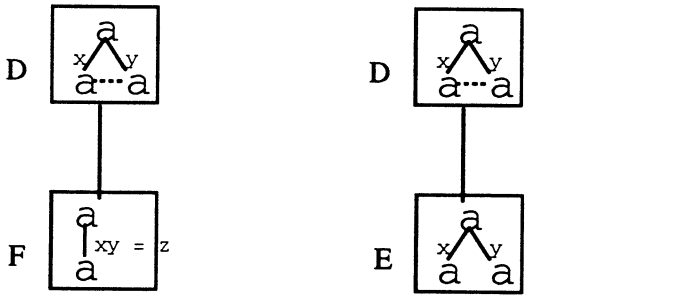
[10] This would be a simple form of concretion inference (Wilensky *et al.* 1988; Wu 1987).



(a) feature deconstraint (b) role detypification



(c) untyped role conflation (d) substructure partition (e) substructure deunification



(f) role decomposition (g) unifiable substructure optionalization

Figure 6.15: An inventory of abstractors (see text).

permitted to be unified unless they stem from different $\emptyset_i$s. There is no f-structure notation for marking unifiable substructures; in feature-DAG notation, dashed lines connect the nodes that can be unified, as in D. This abstractor may not be absolutely essential; it is included for three reasons. First, it provides a mechanism to escape the restriction that generalizations about unification can only involve structures in different $\emptyset_i$s. Second, this escape mechanism may come in useful at some point for modelling crosstalk phenomena. Third, it clearly demonstrates the kind of distortion the unification operation creates on a lattice, while being easier to depict graphically for purposes of examples than the full-fledged substructure deunification abstractor.

Unifiable substructure optionalization complements the role decomposition abstractor. It applies to any f-structure with two unifiable substructures that neither stem from different $\emptyset_i$s nor are explicitly marked, and it adds the link granting the option to unify them. Node E is transformed by this abstractor into D in figure 6.15(g).

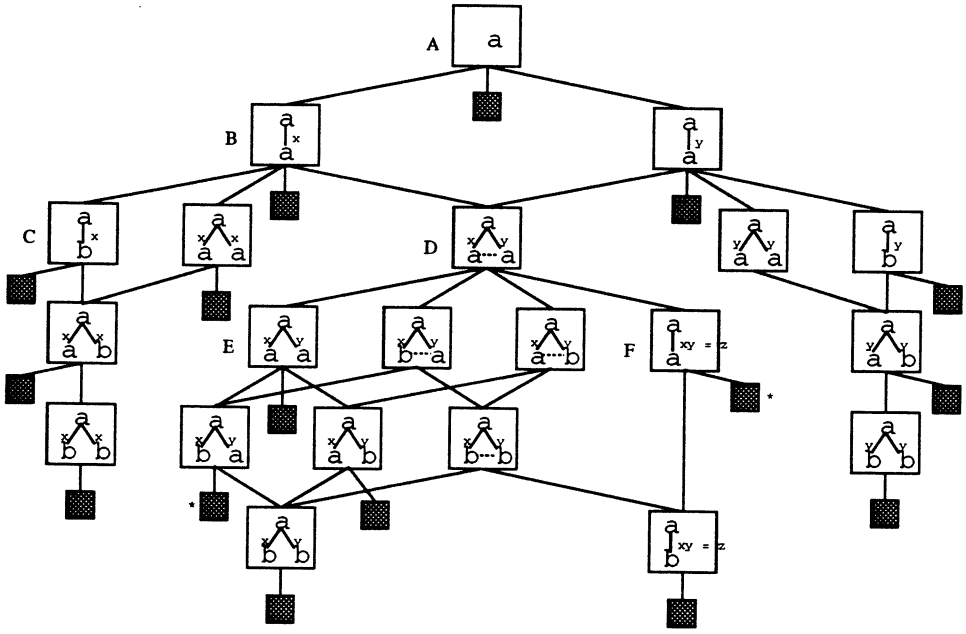


Figure 6.16: Partial semilattice space of feature structures (see text).

Figure 6.16 depicts part of the semi-lattice for a very simplified, strongly bounded f-structure space (the semi-lattice rapidly grows too complex and interconnected to depict otherwise). The feature-DAGs are restricted to depth 2 and branching factor 2. Also the untyped role conflation and substructure deunification abstractors are omitted because they produce too many ancestors, and the role detypification and substructure partition abstractors are combined (e.g., as in the relationship between B and A). For orientation, the labels on the boxes correspond to those in figure 6.16. The shaded boxes denote complete feature-structures that have exactly the structure of their parental abstract feature-structures. The type-token relationship between them is a pseudo-abstractor called *completion*.

6.6.3 Probabilities on Abstract Feature-Structures

The probabilities on abstract feature-structures are *marginal probabilities*. The marginal probability of an abstract feature-structure is the sum of all probabilities of its leaves, i.e., the complete feature-structures it subsumes. The probabilities we store on correlational terms are a special case of marginal probabilities.

As we see in the next chapter, the priors on these probabilities are used for evidential interpretation. Evidential interpretation is formulated in terms of pattern completion where the input is an abstract feature-structure—a partial pattern—and the output is a complete feature-structure.

Chapter 7

Chapter 7

Evidential Interpretation

This chapter brings the concerns and general models of the previous chapters to bear upon the specific task of learning how to interpret nominal compounds. Nominal compounds have been a constant issue in linguistic research and they have defied all but the vaguest of characterizations. Nonetheless humans are able to understand even novel compounds easily and effortlessly, fitting them coherently into the context. The compounds manifest the sorts of multi-way ambiguities whose resolution requires integrating knowledge sources at all levels. Thus, while I do not propose to give a comprehensive account of the knowledge structures needed to interpret English nominal compounds, they serve well as a focal domain for demonstrating the strengths of the probabilistic framework against more traditional approaches. As will become apparent, the approach also attempts to maximize the coherency of interpretations. Note that this work is not concerned with the role of controlled inference in interpreting compounds; that is for the future.

The general picture I will be drawing is a model of automatic inference across various knowledge modules to interpret a nominal compound. Following the analysis of chapter 3, the task of the automatic inference mechanism is to return the interpretation with the highest estimated probability of being useful to the agent. In this chapter we examine how the mechanism can derive these probabilities. A *learning theory* is implicit in the formulation since the probability distributions are derived from the statistics over previous instances. The relevant statistics are kept in a semantic net of correlational terms, as proposed in chapter 6. All correlational terms are built upon the ontological primitives described in chapter 5. The examples are produced by a prototype implementation of the theory called FRIEZE.

7.1 Ranking Interpretations

7.1.1 Probabilistic Integration of Knowledge Sources

In the class of automatic inference models I propose, the agent's syntactic, semantic, and conceptual knowledge sources are integrated probabilistically so as to predict the interpretation most likely to be useful. Recall from chapter 3 that the object is for the automatic inference subagent to select the inference action

(7.1) $$S_{A_I} : \max_I P(\text{Useful}(A_I)|\mathbf{e})$$

Letting q_i represent the interpretation that results from S_{A_I}, this means we want automatic inference to choose the q_i that maximizes $P(q_i|\mathbf{e})$. In the models described here, $\mathbf{e}$ is simply taken to be the nominal compound, that is, the input sequence of lexemes is treated as the evidence that conditions the probability distribution over possible interpretations (section 7.1.3). Since an "interpretation" consists of the entire parse tree and signification mappings as well as semantic and conceptual structures, it follows that parsing and semantic interpretation are integrated in these models.

Parsing has traditionally been viewed as a distinct process from semantic interpretation. Sometimes this is for methodological reasons, as it conveniently delineates an area for closer scrutiny; in other cases a theoretical claim is made for autonomous syntax. Many disagreements result from misinterpreted claims and incompatible goals.

The simplest and oldest view is that the output of the parsing process is fed to the semantic interpretation process as if in a pipeline. The parsing process maps input strings to parse structures such as parse trees; the semantic interpretation process then maps the parse structures to conceptual representations. While the pipeline architecture is easy to implement and theoretically elegant, it is not easily amenable to cases where extensive backward feedback is required. Thus semantic and conceptual information cannot easily be used to guide the parser to the correct nesting of a long nominal compound.

Even in interactive models, the degree of interaction is a frequent point of contention, as are the types of formal machinery best suited to describing and modeling interaction. Indeed, some constructions require no semantic influence whatsoever to parse correctly. There are certainly many types of constructions for which semantic and conceptual preferences are needed but have less influence than for nominal compounds. Because of this it is possible to build low-interaction parsers and interpreters that handle the majority of such constructions in typical, non-anomalous cases. Nonetheless it is significant that there are constructions like nominal compounds, which have defied characterization within traditional low-interaction frameworks and yet occur ubiquitously in ordinary language. A complete theory of language must account for these massive-interaction cases.

I believe a revision is needed in the theoretical concepts we use to characterize massive-interaction phenomenon. Mechanisms that integrate degrees of evidence—especially those that draw from multiple knowledge sources that interact in complex ways—are difficult to analyze, implement, and evaluate. Probability theory should be used to describe their behavior whenever possible because it provides a basis for comparison. This includes connectionist and spreading-activation language models (Cottrell 1984, 1985; Wermter 1989a, 1989b, 1989; Waltz & Pollack 1985) which have contributed a great deal to our intuitions about interacting quantitative preferences. However, the use of *ad hoc* activation functions and network structures engenders skepticism as to large-scale generalizability, and the absence in most cases of any theory of how the networks are learned makes it difficult to compare the adequacy and performance of alternative proposals. Indeed an active area of neural network research is the probabilistic characterization of learning (Haussler 1990; Buntine & Weigend 1991a, 1991b; Mjolsness 1990); unfortunately the networks that yield to analysis thus far are too simple and unstructured to support the level of language inference we are concerned with.

Initially, probabilistic language models like the ones proposed here will be overly simplistic with regard to the kinds of optimization that can be done using sequential processing. The models are restricted to an atomic level of description in which all evidence and all knowledge

sources are taken into account at the same time. In fact, even in a probabilistic model many parsing decisions could still be made using precompiled, sequential, possibly non-probabilistic rules, as in traditional parsers. This would fit well into the compilation paradigm discussed in chapter 3. By the same token, however, choosing the wrong set of rules can lead to unrealistic claims from the point of view of tractability and efficiency, and I regard the analysis as a complex issue for future research. The goal of the present models is to construct parsing and interpretation on a high-interaction probabilistic foundation upon which further sequential optimizations might be applied, without yet making commitments to specific compiled rule sets. When we do eventually discover sequential optimizations, they may be of somewhat different nature than rules typically found in purely syntactic theories, due to the semantic and conceptual factors.

7.1.2 Encoding Warren's Corpus

As noted in section 2.1.2, Warren (1978) categorized each nominal compound in her corpus according to a rough semantic analysis. Starting from this classification, one representative compound was taken from each category and analyzed by hand. This process suggested a number of semantic and conceptual schemata. The semantic schemata were integrated into the ontology described in chapter 5. The conceptual schemata serve as a skeletal representative of what a fully fleshed-out commonsense hierarchy would look like.

Figure 7.1 shows the preferred "correct" analysis of *coast road*, roughly meaning *road contained in coastal area*, both in compact typed feature structure notation and expanded out using explicit features. In a very informal verbal survey, all respondents preferred this interpretation or some more specific version of it such as *road along coastline*. For comparison, figures 7.2–7.4 show several other analyses. The interpretation *road on which an automobile coasts*, shown in figure 7.2, was given by several informants. One informant suggested figure 7.3, in which *coast* is interpreted as the name of the road where the Coast soap operation is located, analogously to *Great America Parkway*. Interestingly enough, figure 7.4, which corresponds roughly to *road leading to the coastal area*, was not suggested by any of the informants, but when asked all agreed that it would be a highly plausible and felicitous use of *coast road*, given a suitable context.

A few words about "correct" analyses are in order. Hand analyses may seem arbitrary and at this point they are. Analysis by hand is merely an intermediate step. It is not the method of hand analysis I am advocating here, but the question of how to proceed given that the situated agent will present better (i.e., more certainly relevant) analyses to its adaptive automatic inference mechanism.

Moreover it is not the correctness of every individual compound that matters here; rather what counts is the statistical distribution over abstract categories in various dimensions. Thus "incorrect" analyses here and there will cancel each other out over a sufficiently large sample, appearing only as a small amount of noise. Of course maintaining a higher degree of accuracy in hand analysis will permit a smaller sample to suffice, and thus we should still try to be as accurate as possible.

7.1.3 Form of the Input

In the abstract case, the form of input to the models is a skeletal feature structure containing just the lexemes and their order. That all the lexemes need to be presented at once is to be

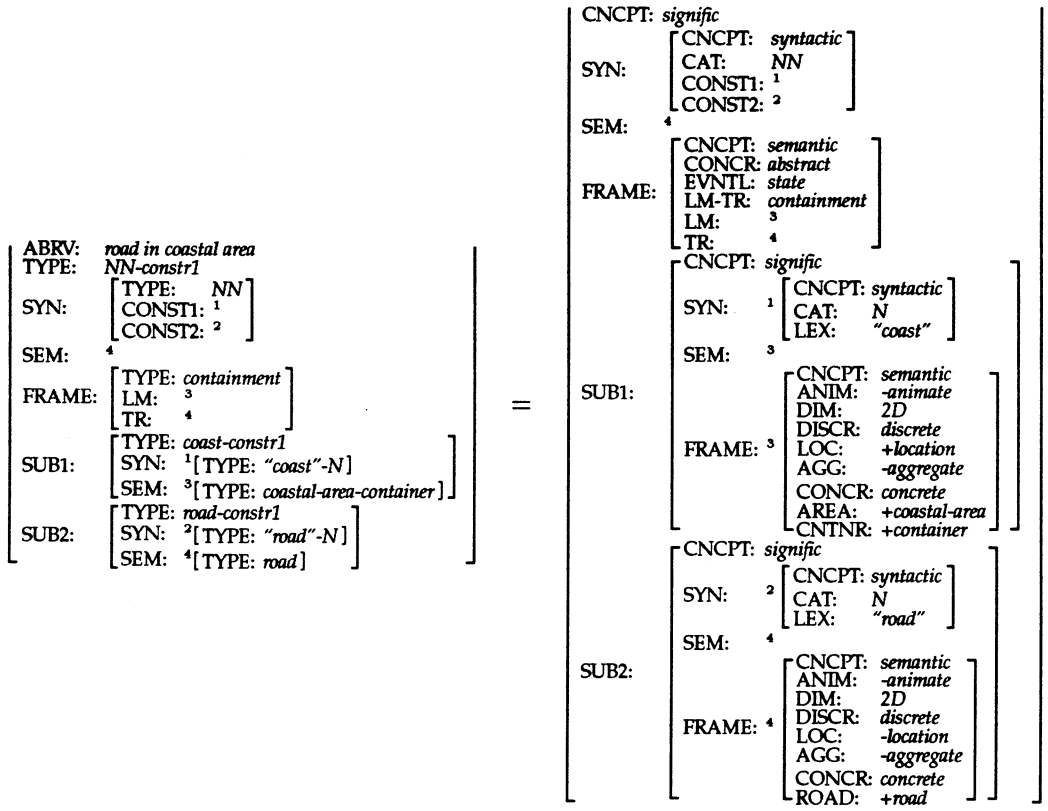


Figure 7.1: Encoding of the "correct" result (output of parsing and interpretation) for *coast road*, namely ⌊*road in coastal area*⌋, in both typed and explicit feature structure notations.

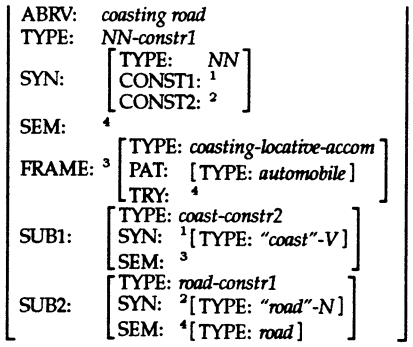


Figure 7.2: Encoding ⌊*coasting road*⌋ meaning a road on which an automobile coasts, an "incorrect" result for *coast road*.

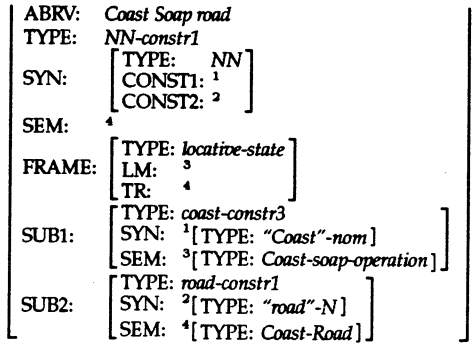


Figure 7.3: Encoding ⌊*Coast Soap road*⌋ meaning the road on which the Coast soap operation lies, an "incorrect" result for *coast road*.

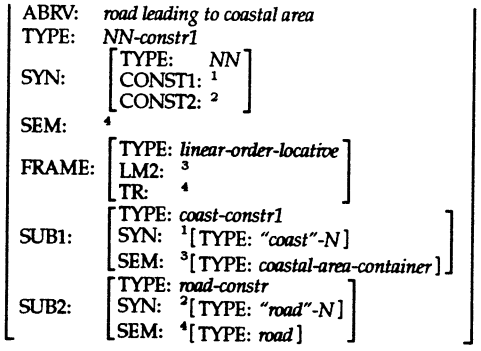


Figure 7.4: Encoding ⌊*road leading to coastal area*⌋, an "incorrect" result for *coast road*.

regarded as a preliminary assumption, made for simplicity's sake. Figure 7.5 shows an example for the input compound *coast road*. In (a) the idealized case is shown, where no structuring information other than the input order of individual lexeme constructions is given. (Recall from section 6.5 that $\emptyset_i$ denotes an untyped attribute and that ordering relations are abbreviated by the suffixes "1", "2", etc.) Since we know that our input is a nominal compound it is more practical to assume the additional structuring information in (b). The notation in the FRAME attribute indicates that the SEM substructure fits somewhere within it.

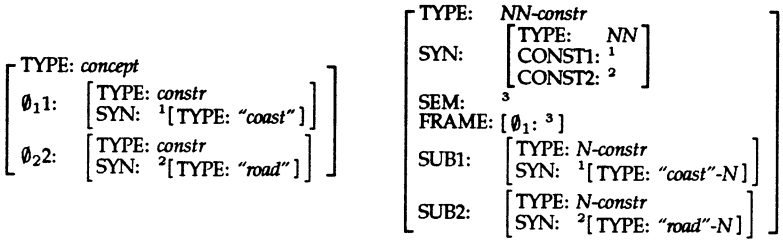


Figure 7.5: (a) Idealized input form, and (b) a practical input form.

7.1.4 Generating Hypotheses

In the current models candidate interpretations are generated by marker passing. Some other possible strategies are proposed in section 7.5.1. The structures in figures 7.1–7.4 are all examples of candidate interpretations. Marker passing methods, as discussed in section 6.2, are subject to a number of severe problems. However, the space of candidate interpretations must somehow be restricted, to make evaluating the conditional probability distribution over the space possible.

Conceived abstractly, marker passing generates all possible interpretations and all known constraints on the probability distribution over them. In fact, to do this given the size of the ontology and conceptual system would far exceed reasonable computational resource bounds. (After all if we had the resources to store the complete set of all alternative interpretations, we could just rank them and never have to commit to an interpretation until a decision needed to be made.) Instead, the hypothesis-bounded model assumes that only the most pertinent hypotheses and constraints will be generated. Here "pertinence" is only meant as some gross, coarse characteristic, indicating that marker passing takes a first crude cut at narrowing down the space of interpretations to be considered.

Depending on one's interpretation of Charniak's (1983) original marker passing proposal, this model can be seen as having one additional intermediate stage. In Charniak's model the hypotheses produced by marker passing feed directly into a "deductive" mechanism. The sorts of inferences made by the deductive mechanism often resemble the high-level reasoned inferences I propose are performed by controlled inference, involving plan chaining, rather than easy reflexive inferences that are automatically inferred. Charniak & Goldman's (1988) later proposal replaces the deductive mechanism with a Bayesian belief network (Pearl 1988) but the inferences still concern chained plans.

For purposes of the discussion in this chapter only a few other simple assumptions about the propagation strategy are needed. Origin markers are instantiated for each of the input lexemes, along with the ordering relations between them. The origin markers are propagated to all nodes that are more abstract. Markers are also propagated to more specific nodes via some heuristic strategy. Such a strategy would need to do something like guaranteeing that the probability of missing a "pertinent" collision will be lower than some acceptable threshold.[1] In other words, the marker passing control mechanism is augmented by some adaptive mechanism that learns to predict the search directions in which "pertinent" hypotheses are likely to be found.

Upon completion, propagation yields a set of candidate interpretations. We denote these by the leaf nodes of a graph, which I will call the *marker graph*. The internal nodes correspond to the nodes of the semantic net to which markers are propagated. Accordingly, the internal nodes denote more abstract classes of possible interpretations, whose marginal probability values are known. This means there is one abstract class for every known constraint on the conditional probability distribution, since a correlational term and its probability exactly specify a marginal constraint.

Finding a sound adaptive marker propagation strategy is an open problem. While many marker passing models have been proposed, a means is still needed to statistically guarantee the reliability of marker passing with respect to finding all relevant interpretations. Some directions are suggested in section 7.5.1.

In summary: the marker graph's leaves specify a set of interpretation hypotheses, and its internal nodes specify some constraints on the probability distribution over the space of interpretation hypotheses. This input is represented as a conditioning event $\mathbf{e}$. The hypothesis q_i that maximizes $P(q_i|\mathbf{e})$ should be selected.[2]

7.1.5 Canonical Distribution Models

From the simplest, most ideal view, the way to choose the best interpretation would be to condition the implicit distribution in the knowledge base on the input event $\mathbf{e}$, and then take the hypothesis (leaf node) with the highest conditional probability. However, because the constraints encoded by correlational terms only partially specify the distribution, whatever method is used to compute the conditional distribution must complete the distribution, either explicitly or implicitly. The issue of how to complete the probability distribution as neutrally as possible is considered in subsequent sections. We first consider an example of a *canonical distribution model* in which unspecified parts of the probability distribution are completed by assuming that the combination of incoming edges to any node is guaranteed to be in a canonical form. What this means is that the full distribution can be reconstructed by applying a small set of standard approximation rules. The canonical form assumption is needed because of the loops in the marker graph, corresponding to the multiple inheritance loops in the semantic net which give rise to tangled distributional

[1] In a similar spirit to the PAC (probably approximately correct) formal learning theory (Valiant 1984; Haussler 1990; Kearns 1990).

[2] Conditioning on $\mathbf{e}$ is performed by marker passing, since it is assumed to generate only hypotheses consistent with the input event. Because any hypothesis q_i generated is compatible with the input evidence, it is sufficient simply to maximize

$$P_i \stackrel{\text{def}}{=} P(q_i) \tag{7.2}$$

over the set of q_i's.

constraints. An algorithm of this type is shown in figure 7.6, based on an early version of the model (Wu 1990). The primary advantage of this model is its simplicity and tractability.

- Let g_i be all parents of h in the marker graph, and let $\mathbf{g}^j$ be all the possible combinatoric states of g_i. Then

$$\begin{aligned} P(h|\mathbf{e}) &= \sum_j P(h|\mathbf{g}^j, \mathbf{e}) P(\mathbf{g}^j|\mathbf{e}) \\ &= \sum_j P(h|\mathbf{g}^j) P(\mathbf{g}^j|\mathbf{e}) \end{aligned}$$

- To estimate each of the $P(\mathbf{g}^j|\mathbf{e})$ terms, break down $\mathbf{g}^j$ into its component node states $g_1^A, g_2^B, \ldots, g_n^Z$. Make a *conditional independence assumption* yielding

$$P(\mathbf{g}^j|\mathbf{e}) \approx P(g_A^1|\mathbf{e}) P(g_B^2|\mathbf{e}) \cdots P(g_Z^n|\mathbf{e})$$

- Estimate each term of the form $P(g_A^1|\mathbf{e})$ by recursive application of this algorithm.

- To estimate each of the $P(h|\mathbf{g}^j)$ terms, again break $\mathbf{g}^j$ down. Make a *disjunctive interaction assumption* yielding

$$P(h|\mathbf{g}^j) \approx 1 - P(\neg h|g_1^A) P(\neg h|g_2^B) \cdots P(\neg h|g_n^Z)$$

- If $P(h)$ and $P(g_1^A)$ are known, then

$$P(\neg h|g_1^A) = 1 - \frac{P(h)}{P(g_1^A)}$$

Otherwise find all known "most specific subsuming" conditionals in the knowledge base $P(b_k|a_k)$ such that $h \subset b_k$ and $g_1^A \subset a_k$, and a_k and b_k are as specific as possible. Make an *independent conditional assumption* yielding a weighted average conditional probability

$$P(\neg h|g_1^A) \approx 1 - \frac{\sum_k P(a_k) P(b_k|a_k)}{\sum_k P(a_k)}$$

Figure 7.6: An algorithm that estimates $P(h|\mathbf{e})$ for any event h subsumed by $\mathbf{e}$.

Methods like these are subject to severe inconsistency problems, since canonical distribution models only ensure local consistency at the node. When all hypothesis nodes are interpreted using the canonical distributional assumptions, the combined constraints are too strong and conflict with each other, so no solution can be found. It would be extremely difficult to choose the right marginals to store in the knowledge base, and to design a marker passing method that would generate a graph guaranteeing the canonical distribution assumptions. Any realistic implementation of this model would require some additional conflict mediation function. Such a function, for example, might include a penalty term for the strength of each conflict. The distributional con-

straints are thus made "soft" and the problem becomes to globally minimize the conflict function. However, a solution of this sort introduces a degree of computational complexity that eliminates the advantage of the canonical distribution model. The maximum entropy model, described next, also requires global maximization but is theoretically more sound.

7.2 Model I: Maximum Entropy Completion of the Distribution

7.2.1 Constrained Maximum-Entropy Distributions

The maximum entropy principle (Jaynes 1979) is a global canonical method that yields a unique completion of a partially constrained distribution. The principle says to select the complete distribution that maximizes the information-theoretic entropy measure

$$H = -\sum_{i=1}^{C} P_i \log P_i \tag{7.3}$$

To explain the maximum entropy model, an oversimplified set of hypotheses and constraints is used here to keep the example small. Also, only semantic and conceptual constraints will be used for now. In subsequent sections more realistic hypotheses and constraints are considered, and lexical and constructional constraints are added.

Suppose we had the following interpretation hypotheses generated by marker passing:[3]

$$\begin{aligned} q_1 &= \lfloor \textit{Highway 1 in Pacific coastal area} \rfloor \\ q_2 &= \lfloor \textit{road along coastline} \rfloor \\ q_3 &= \lfloor \textit{Highway 1} \rfloor \\ q_4 &= \lfloor \textit{Highway 1 along Pacific coastline} \rfloor \end{aligned}$$

so we want the associated probabilities:

$$\begin{aligned} P_1 &= P(\lfloor \textit{Highway 1 in Pacific coastal area} \rfloor) \\ P_2 &= P(\lfloor \textit{road along coastline} \rfloor) \\ P_3 &= P(\lfloor \textit{Highway 1} \rfloor) \\ P_4 &= P(\lfloor \textit{Highway 1 along Pacific coastline} \rfloor) \end{aligned}$$

and suppose the following constraints are known from correlational terms in the semantic net:

$$\begin{aligned} P_{24} &= P(\textit{road along coastline}) = P_2 + P_4 = 0.6 \\ P_{124} &= P(\textit{road contained in coastal area location}) = P_1 + P_2 + P_4 = 0.84 \\ P_{134} &= P(\textit{Highway 1}) = P_1 + P_3 + P_4 = 0.69 \end{aligned}$$

Then the entropy function

$$H = -P_1 \log P_1 - P_2 \log P_2 - P_3 \log P_3 - P_4 \log P_4$$

[3]Recall that $\lfloor$*road along coastline*$\rfloor$ denotes the occurrence of that exact feature structure with no extra roles or feature values, whereas just *road along coastline* without floor brackets denotes the abstract category of all feature structures containing but not restricted to the *road along coastline* substructure.

is maximized by the distribution

$$\begin{aligned} P_1 &= 0.24 \\ P_2 &= 0.31 \\ P_3 &= 0.16 \\ P_4 &= 0.29 \end{aligned}$$

as shown below, and thus we select hypothesis q_2 as the best interpretation.

Some intuitive insight about maximum entropy can be gained from the special case where the constraints only partition the hypothesis space orthogonally. In this case the maximum entropy principle reduces simply to a conditional independence assumption. Suppose the hypotheses being considered were instead:

$$\begin{aligned} q_1 &= \lfloor \textit{road in coastal area} \rfloor \\ q_2 &= \lfloor \textit{road along coastline} \rfloor \\ q_3 &= \lfloor \textit{Highway 1 in Pacific coastal area} \rfloor \\ q_4 &= \lfloor \textit{Highway 1 along Pacific coastline} \rfloor \end{aligned}$$

and the constraints were:

$$\begin{aligned} P_{24} &= P(\textit{road along coastline}) = P_2 + P_4 = 0.6 \\ P_{34} &= P(\textit{Highway 1}) = P_3 + P_4 = 0.65 \end{aligned}$$

Then entropy is maximized by the assignment

$$\begin{aligned} P_1 &= 0.14 \quad (= \bar{P}_{24} \cdot \bar{P}_{34}) \\ P_2 &= 0.21 \quad (= P_{24} \cdot \bar{P}_{34}) \\ P_3 &= 0.26 \quad (= \bar{P}_{24} \cdot P_{34}) \\ P_4 &= 0.39 \quad (= P_{24} \cdot P_{34}) \end{aligned}$$

which can be graphically depicted as in figure 7.7. This is the same distribution that results from taking $P(\textit{road along coastline})$ and $P(\textit{Highway 1})$ to be conditionally independent. However, if we added the other hypotheses back in, conditional independence would no longer hold. In this case the maximum entropy distribution cannot be graphed so transparently.

The "neutrality" of the maximum entropy principle is controversial. Without delving into the philosophical arguments, I will simply observe that whether the principle is "truly" neutral depends upon the correspondence of the event space to the domain being modelled. With regard to automatic inference, the more accurately we structure the ontology, the closer maximum entropy comes to neutrality.

An entropy-maximizing assignment of probability values is found using Lagrange multipliers, following Cheeseman (1987) except that the underlying events here are compositional, non-flat feature structures whereas Cheeseman assumed them to be flat feature vectors.[4] There are

[4] As mentioned earlier, we drop the requirement that probabilities sum to unity since the ultimate goal of finding the maximum-probability hypothesis is independent of the normalization factor. By the same token, we could condition the probabilities on the input evidence by normalizing to $P(e)$, but this step too is superfluous. In Cheeseman's presentation there is one additional constraint (and therefore one additional ω_0) arising from the normalization requirement.

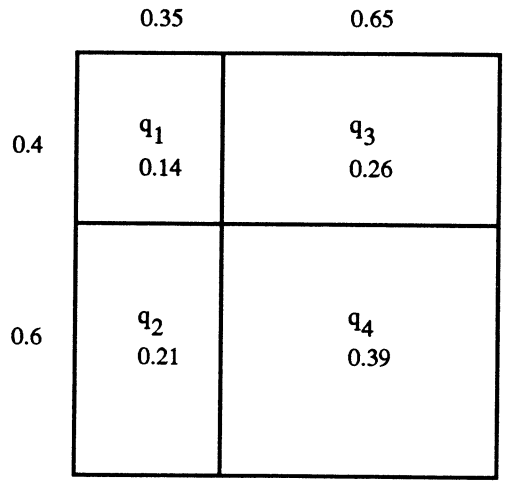


Figure 7.7: The special case of maximum entropy as conditionally independent features, depicted as orthogonal cuts of a unit square (generalizing to a hypercube for n dimensions).

more P_i values to be determined than constraints (in real cases there are typically many more), with the remaining degrees of freedom determined by maximum entropy. In the exposition that follows I return to the original example. First define a new function G to be minimized, incorporating both the entropy and the constraints in a new form:

$$\begin{aligned} G &= -\sum_{i=1}^{C} P_i \log P_i + \lambda_1(P_{24} - P_2 - P_4) \\ &\quad +\lambda_2(P_{124} - P_1 - P_2 - P_4) + \lambda_3(P_{134} - P_1 - P_3 - P_4) \\ &= -P_1 \log P_1 - P_2 \log P_2 - P_3 \log P_3 - P_4 \log P_4 + \lambda_1(P_{24} - P_2 - P_4) \\ &\quad +\lambda_2(P_{124} - P_1 - P_2 - P_4) + \lambda_3(P_{134} - P_1 - P_3 - P_4) \end{aligned} \tag{7.4}$$

We then require the gradient with respect to the Lagrange multipliers λ_j to be zero:

$$\nabla_\lambda G = \mathbf{0} \tag{7.5}$$

or in other words, for all j.

$$\frac{\partial G}{\partial \lambda_j} = 0$$

This condition expresses all the constraints, since

$$\begin{aligned} \frac{\partial G}{\partial \lambda_1} = 0 &\Rightarrow P_{24} = P_2 + P_4 \\ \frac{\partial G}{\partial \lambda_2} = 0 &\Rightarrow P_{124} = P_1 + P_2 + P_4 \\ \frac{\partial G}{\partial \lambda_3} = 0 &\Rightarrow P_{134} = P_1 + P_3 + P_4 \end{aligned}$$

To maximize the entropy we also require the gradient with respect to the probabilities P_i to be zero:

(7.6) $$\nabla_{\mathbf{P}} G = \mathbf{0}$$

But this also implies

$$\begin{aligned}
\frac{\partial G}{\partial P_1} &= -\log P_1 - \lambda_2 - \lambda_3 = 0 \\
\frac{\partial G}{\partial P_2} &= -\log P_2 - \lambda_1 - \lambda_2 = 0 \\
\frac{\partial G}{\partial P_3} &= -\log P_3 - \lambda_3 = 0 \\
\frac{\partial G}{\partial P_4} &= -\log P_4 - \lambda_1 - \lambda_2 - \lambda_3 = 0
\end{aligned}$$

and so

$$\begin{aligned}
\log P_1 &= -(\lambda_2 + \lambda_3) \\
\log P_2 &= -(\lambda_1 + \lambda_2) \\
\log P_3 &= -(\lambda_3) \\
\log P_4 &= -(\lambda_1 + \lambda_2 + \lambda_3)
\end{aligned}$$

Defining for convenience

(7.7) $$\omega_j \stackrel{\text{def}}{=} e^{-\lambda_j}$$

we have

$$\begin{aligned}
P_1 &= \omega_2\omega_3 \\
P_2 &= \omega_1\omega_2 \\
P_3 &= \omega_3 \\
P_4 &= \omega_1\omega_2\omega_3
\end{aligned}$$

Thus we have redefined the P_i values in terms of ω_j values, of which there are only as many as constraints. The constraint equations can then be rewritten as a system of equations of the form

$$\begin{aligned}
P_{24} &= \omega_1\omega_2 + \omega_1\omega_2\omega_3 = 0.6 \\
P_{124} &= \omega_2\omega_3 + \omega_1\omega_2 + \omega_1\omega_2\omega_3 = 0.84 \\
P_{134} &= \omega_2\omega_3 + \omega_3 + \omega_1\omega_2\omega_3 = 0.69
\end{aligned}$$

which contains exactly as many variables as constraints.

7.2.2 Approaches to Implementation

To solve the above system of equations a gradient descent method can be used to find the $\langle \omega_1, \omega_2, \omega_3 \rangle$ that minimizes

$$\begin{bmatrix} |P_{24} - (\omega_1\omega_2 + \omega_1\omega_2\omega_3)| \\ |P_{124} - (\omega_2\omega_3 + \omega_1\omega_2 + \omega_1\omega_2\omega_3)| \\ |P_{134} - (\omega_2\omega_3 + \omega_3 + \omega_1\omega_2\omega_3)| \end{bmatrix}$$

Assuming that a zero minimum

$$\begin{bmatrix} P_{24} - (\omega_1\omega_2 + \omega_1\omega_2\omega_3) \\ P_{124} - (\omega_2\omega_3 + \omega_1\omega_2 + \omega_1\omega_2\omega_3) \\ P_{134} - (\omega_2\omega_3 + \omega_3 + \omega_1\omega_2\omega_3) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

is found, the P_i values are determined by the values of $\langle \omega_1, \omega_2, \omega_3 \rangle$ and we can simply choose the interpretation hypothesis q_i with the highest corresponding P_i.

The feed-forward network depicted in figure 7.8 performs the gradient descent procedure. The mathematics here are the same as for back-propagation networks, but instead of using iterative convergence for adaptive learning, iteration is used to solve the single entropy-maximization problem. In this respect the net is more like a simple recurrent Hopfield net since it settles into a maximum entropy solution. The nodes in each layer compute the activation function O_i shown in the figure, where h_i is the sum of all input activations. A forward pass computes an estimate of the hypothesis probabilities and marginal probabilities (the reader can easily verify this). The difference between the actual marginal constraints and the estimated marginal probabilities is computed and an error measure is propagated backwards. This process is repeated until the ω values converge.

A C implementation of this net revealed great sensitivity to initial conditions. Empirically there seem to be strong local minima that the net settles into given an initial condition that is not sufficiently close to the actual zero solution. Traditional numerical techniques for speeding up convergence, such as Newton's method or conjugate gradient descent, would not help since they would encounter the same local minima.

The net's sensitivity to initial conditions appears to be caused by the back-propagation of error in one relatively high-magnitude constraint to all ω values in parallel, which has the effect of drowning out the ω values corresponding to smaller-magnitude constraints. To eliminate this effect we can solve for only one ω at a time, holding the others constant and back-propagating just the error for the corresponding constraint. Unfortunately some of the processing parallelism is lost this way, and this approach becomes essentially similar to the largely sequential method described below.

Figure 7.9 shows the iterative method used in all the subsequent examples, derived from Cheeseman (1987). For conceptual ease I continue to use the neural net description. We build the net incrementally one constraint at a time, solving for the maximum entropy distribution each time before adding the next constraint link, until the full constraint system is reached. For each such entropy-maximization pass, only one ω and its corresponding constraint are dealt with at a time, as suggested earlier. Conceptually, the other ω values are held constant and only the error term for the corresponding constraint is back-propagated, until the exact zero-error solution is found. In actual fact the value of ω is obtained analytically rather than by error propagation since the constraint with all other ω's held constant becomes simply linear. From this it should be clear that the method is a successive line minimization procedure. Given consistent constraints, this method has always converged successfully.

The algorithm has been implemented in C and serves as the core of FRIEZE, described below. Sample traces from FRIEZE are given in appendix A.

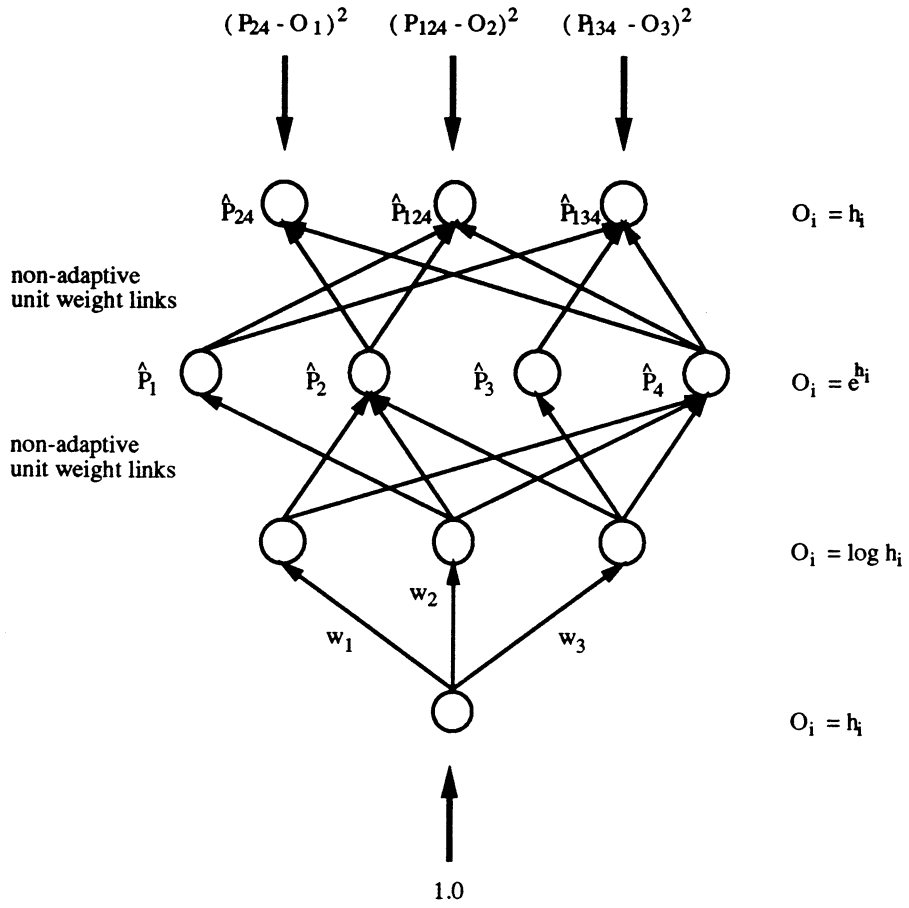


Figure 7.8: A neural net for solving the constrained maximum entropy system.

1. Start with a constraint system $Q \leftarrow \{\}$ and an estimated ω vector $\langle\rangle$ of length zero.
2. For each constraint equation,
   (a) Add the equation to Q and its corresponding ω_i term to $\langle \omega_1, \ldots, \omega_{i-1}, \omega_i \rangle$.
   (b) Repeat until $\langle \omega_1, \ldots, \omega_i \rangle$ settles, i.e., the change between iterations falls below some threshold:
      1. For each constraint equation in Q, solve for the corresponding ω_j assuming all other ω values have their current estimated values.

Figure 7.9: Algorithm for finding the maximum-entropy distribution subject to constraints.

7.2.3 The Combinatoric Event Space

A real problem with using full-fledged maximum entropy to supply missing constraints is the computational cost. Even the methods given above are still expensive in the general case.

In the above examples I oversimplified the kinds of distributional constraints for expository purposes:

$$\begin{aligned} P_{24} &= P(\textit{road along coastline}) = P_2 + P_4 = 0.6 \\ P_{124} &= P(\textit{road contained in coastal area location}) = P_1 + P_2 + P_4 = 0.84 \\ P_{134} &= P(\textit{Highway 1}) = P_1 + P_3 + P_4 = 0.69 \end{aligned}$$

In fact the marginal probabilities in the knowledge base will be far smaller:

$$\begin{aligned} P_{24} &= P(\textit{road along coastline}) = 2 \cdot 10^{-7} \\ P_{124} &= P(\textit{road contained in coastal area location}) = 3 \cdot 10^{-7} \\ P_{134} &= P(\textit{Highway 1}) = 5 \cdot 10^{-8} \end{aligned}$$

Moreover, marker propagation will actually reveal constraints in the knowledge base that are relevant in much more complex ways, like

$$\begin{aligned} &P(\textit{containment}) = 10^{-2} \\ &P(\textit{linear-order-locative}) = 10^{-3} \\ &P(\textit{Pacific}) = 5 \cdot 10^{-7} \end{aligned}$$

These cannot be characterized as simple sums of the interpretation hypotheses alone, since they summarize marginal probabilities over other potential structures as well. For instance, $P(\textit{containment})$ is a normalized count of all previous occurrences of the containment schema, regardless of whether they involved roads and coastal areas. In fact this point even applies to the earlier constraints, albeit to a lesser degree: $P(\textit{road along coastline})$ includes previous occurrences of *Jersey Turnpike*.

In this model the maximum-entropy distribution must be computed over the entire space of possible hypotheses. Entropy cannot be maximized just over the space of interpretation hypotheses because the relevant marginals are defined over a finer event space. The marginals are relative frequencies that were observed in the space of all possible hypotheses, which is a combinatoric space of all feature structures that can be constructed using the ontological primitives. If we attempt to apply the marginal constraints in just the hypothesis space, at best we are misinterpreting the relative frequencies, and at worst we produce an inconsistent set of constraints. Consider for example the smaller, more realistic marginal constraints given above for P_{24}, P_{124}, and P_{134}. The hypotheses that fall under *containment* are q_1, q_2, and q_4. If we misinterpret $P(\textit{containment})$ to be $P_1 + P_2 + P_4$, no possible assignment of probabilities $P_1, \ldots, P_4$ can bring $P(\textit{containment})$ up to 10^{-2}.

In theory, of course, we are guaranteed that the event space is consistent with the marginals if we do compute the maximum-entropy distribution over the entire space rather than just the interpretation hypotheses. This amounts to computing the entire probability distribution

instead of just the conditional distribution.[5] Afterwards, we choose the hypothesis with the highest conditional probability, ignoring the other non-hypothesis events even if their probabilities are higher.

Given that a knowledge base with enough conceptual information to disambiguate nominal compounds will contain thousands of marginals, over a far more enormous space of hypotheses,[6] the amount of computation required by this model is impractical. In the following section we consider a modified version of this approach.

7.3 Model II: Approximate Maximum-Entropy Estimation

7.3.1 Approximation Strategy

The entropy-maximizing methods described above take reasonable computation times only when the space can be kept small. For hypothesis spaces of the size we are dealing with the methods are likely to be inadequate. This section presents an approximation method that uses the same maximum-entropy mechanisms, but in a coarser space, to estimate the true maximum-entropy distribution for the full combinatoric space. A prototype implementation of this theory, FRIEZE, was used to compute all the examples here; traces are in appendix A. For speed FRIEZE is implemented in C, but it does include an interactive symbolic user interface.

The essence of the approximation is to discard the details of how the event space is structured outside the immediate hypothesis space. The major assumption here is that the marker passing process successfully constructs all hypotheses with non-negligible probability given the input. Providing this assumption is valid, the hypothesis space equals the *conditional space*, that is, the space of events conditioned on the input event, containing those events consistent with the input evidence. Actually, the hypothesis space only exactly equals the conditional space if marker passing produces precisely the set of events with *nonzero* conditional probability. Otherwise the hypothesis space is a subset of the conditional space.

We are only concerned with ranking the hypotheses within the conditional space. However unless the full space is considered we cannot apply the marginal constraints in the knowledge base, because entropy cannot be maximized with inconsistent constraints. Thus in this model we approximate the remainder of the event space outside the conditional space—which I will call the *complement space*—with "dummy" compound events. We pretend these "dummy" events are simple events when maximizing entropy, and so there are few enough events to be tractable. Yet the presence of the "dummy" events makes the marginals consistent, because they provide nonzero subspaces for those events that have been counted into the marginals but are inconsistent with the hypothesis space.

One dummy event is used for each constraint; the dummy event represents all the feature structures that are consistent with the constraint, but not with the conditional space. Returning to

[5] Recall that the conditional distribution is simply the part of the distribution that covers the set of hypothesis events consistent with the input evidence, normalized to unity.

[6] A rough magnitude estimate might be obtained as follows. Assume that a static feature structure never gets larger than 25 positions, including lexico-syntactic, semantic, and conceptual structures. Further assume, extremely conservatively, that each position is encoded by 100 binary features. Then there are already 2^{2500} events, and we have not yet even factored in the number of permutations of roles.

the previous example, we define q_a through q_d corresponding to each of the constraints

$$\begin{aligned} P_{124a} &= P(\textit{containment}) = 10^{-2} \\ P_{24b} &= P(\textit{road along coastline}) = 2 \cdot 10^{-7} \\ P_{124c} &= P(\textit{road contained in coastal area location}) = 3 \cdot 10^{-7} \\ P_{134d} &= P(\textit{Highway 1}) = 5 \cdot 10^{-8} \end{aligned}$$

In addition we define a *null event* $q_\emptyset$ representing all other events. Entropy can now be maximized subject to the marginal constraints over this space, yielding

$$\begin{aligned} P(\lfloor \textit{Highway 1 in Pacific coastal area} \rfloor) = P_1 &= 7.65229 \cdot 10^{-17} \\ P(\lfloor \textit{road along coastline} \rfloor) = P_2 &= 6.12183 \cdot 10^{-16} \\ P(\lfloor \textit{Highway 1} \rfloor) = P_3 &= 0.000000025 \\ P(\lfloor \textit{Highway 1 along Pacific coastline} \rfloor) = P_4 &= 1.54592 \cdot 10^{-23} \\ P_a &= 0.01 \\ P_b &= 0.0000002 \\ P_c &= 0.0000003 \\ P_d &= 0.000000025 \\ P_\emptyset &= 0.98999 \end{aligned}$$

The event in the conditional space with the highest probability in $P_1, \ldots, P_4$ is selected, which is q_3.

Though consistent with the constraints, the distribution is skewed oddly. The reasons have to do with inadequate configuration of the constraints. Two configuration principles are particularly important: *proper subsumption* and *most-specific constraint*.

The proper subsumption principle is that any subsumption relation between marginal constraints in the full event space should be reflected in the approximate space. Consider the example depicted graphically in figure 7.10(a), in which $\lfloor$*road leading to a coastal area*$\rfloor$ overly dominates the conditional distribution. Without proper subsumption, *road* and *coastal road* are mistakenly treated as if they were orthogonal features. In (b) every event under *coastal road* is also under *road*, including the dummy event for *coastal road*. This makes *coastal road* a proper subcategory of *road*. The probability of $\lfloor$*road leading to a coastal area*$\rfloor$ falls because the marginal probability on *road* must now also be partly allocated to the dummy event for *coastal road*. An additional more specific intermediate constraint category is added in (c), subsuming $\lfloor$*road leading to a coastal area*$\rfloor$ as well as the other hypotheses. Since this constrains the total amount of probability assignable to the hypotheses, much of the probability weight is shifted to the dummy events. In effect the new constraint says that out of the 10^{-4} marginal on *road*, only $1.3 \cdot 10^{-6}$ comes from co-occurrences with the *coast* concept. This makes a dramatic correction in the resulting maximum-entropy distribution.

The most-specific constraint principle is that for every hypothesis, the most specific applicable marginal constraints that are available should be used. In figure 7.11(a) the hypothesis $\lfloor$*coasting road*$\rfloor$ has been added (short for the interpretation $\lfloor$*road on which an automobile coasts*$\rfloor$). Even with proper subsumption $\lfloor$*coasting road*$\rfloor$ inappropriately overwhelms the legitimate interpretations. What is happening is that $\lfloor$*coasting road*$\rfloor$ "steals" the weight from all the unspecified

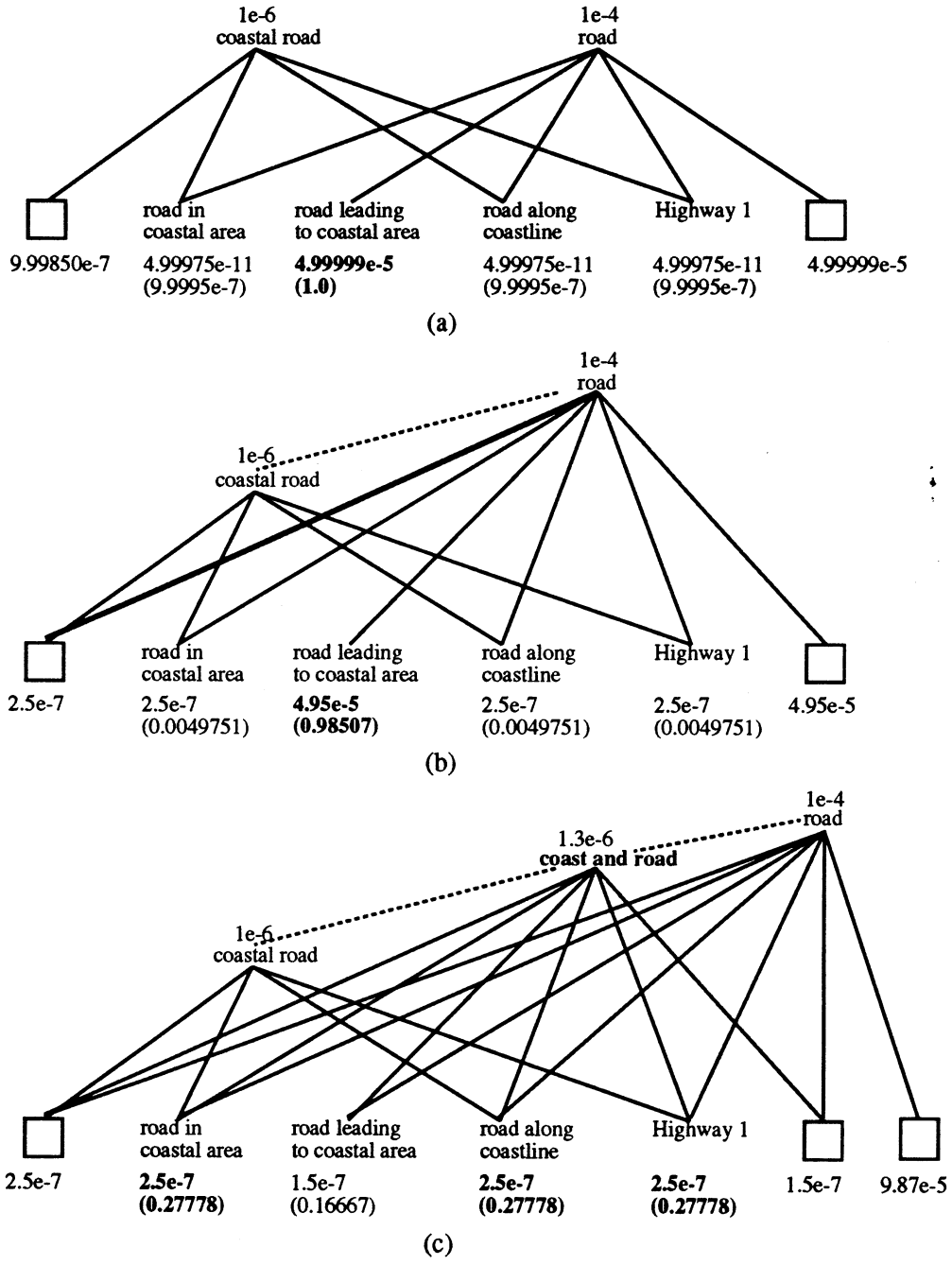


Figure 7.10: Dummy events are represented by the square boxes. The numbers in parentheses are the conditional probability estimates, obtained by normalizing over the hypotheses. (a) The maximum-entropy distribution resulting from not subsuming *coastal road* by *road*. (b) Proper subsumption, represented by the dashed edge, is obtained by adding *coastal road*'s dummy event into the space for *road* as well. (c) Having properly subsumed intermediate constraints has a significant effect.

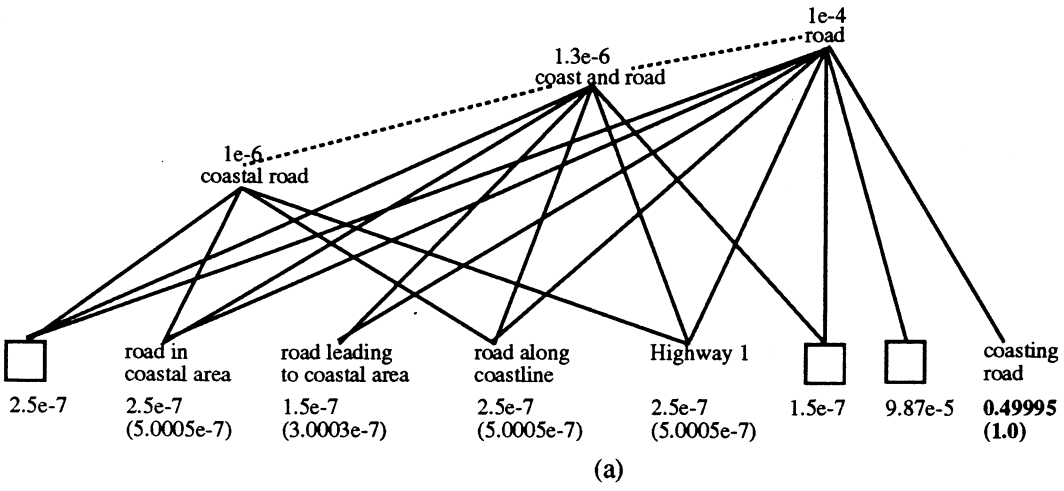


Figure 7.11: (a) The incorrectly skewed maximum-entropy distribution resulting from missing most-specific marginal constraints. (b) The correct skew, obtained by adding the relevant specific marginals.

categories of *roads*, because the maximum entropy principle is given no information that distinguishes ⌊*coasting road*⌋ from all the other kinds of roads there might be. In fact we mean ⌊*coasting road*⌋ to be specifically a kind of road involving a coasting accomplishment event. By simply taking into account the corresponding marginal probability of the *coasting accomplishment* event schema as in (b), the probability of ⌊*coasting road*⌋ drops to a reasonable level.

7.3.2 Selectional Preferences, Explanatory Coherence, and Abduction

The maximum entropy model generalizes the notion of selectional restrictions (Katz & Fodor 1963) to a more realistic probabilistic notion. Traditionally, selectional restrictions enforce the types of structures that can fill a role (usually a semantic or conceptual role). This overconstrains the applicability of schemata, making non-literal or metaphoric interpretations problematic. For example, one cannot characterize in a context-independent way the types of concepts that can play the container role in a *containment* schema since nearly any concept can be schematized as a container (try *There is happiness in peace*). Nonetheless we want to capture the fact that some concepts function more readily as containers than others, and during interpretation, if structural cues fail to eliminate all choices, we want the roles to be filled as coherently as possible. The probabilistic approach recognizes selectional *preferences* rather than restrictions, acknowledging usage tendencies and patterns but not banning atypical role fillers outright. These preferences are quantitative, as opposed to binary preferences (Wilks 1973, 1975a). Wilk's Preference Semantics approach is similar to mine in that it combines templates (albeit without lexical or constructional templates) whose slots have preferred rather than restricted filler types, but as there as no strengths to the preferences, competing interpretations with the same number of preference violations cannot be weighed against one another. Although quantitative preference approaches are not new, the maximum entropy model gives a consistent, non-*ad hoc* method of combining preferences, and grounds the numbers through probability theory to the agent's experience.

Figure 7.12 demonstrates this for the *containment* schema. (From this point on, for constraints in subsumption relationships, redundant links are omitted, as are all "dummy" events. Only the conditional probabilities are shown. For exact details consult the traces in appendix A.) Usually the container role is filled by some object that typically functions as a container, such as a *location*. This tendency is represented by the high marginal values for *locative containment* and similar schemata in (a). However, other concepts are still permitted to act as containers. Thus in (b) the interpretation *road in a coasting event* has a nonzero probability. In fact the probability is far too high, even when a specific marginal for *eventive containment* is added as in (c). This occurs because of another constraint omission, which brings us to a third principle.

The *disjoint-sibling summary* principle is that when one constraint subsumes another, any other siblings of the child that are in the knowledge base should be summarized in an extra marginal. In (d) the *other container-type containment* constraint summarizes the marginal probabilities of all the subkinds of containment in the knowledge base that marker passing rejected. This is the probabilistic equivalent of an or-node; since none of the simple events are shared by the siblings, the sum of their marginals must be less than or equal to that of the subsuming constraint. The resulting conditional probabilities are much more reasonable. In fact, even when the most-specific constraint principle is violated, as in (e) where the *eventive containment* marginal has been removed, the resulting probabilities can still be reasonable if the disjoint-sibling marginal brings the sum close enough to the subsuming marginal.

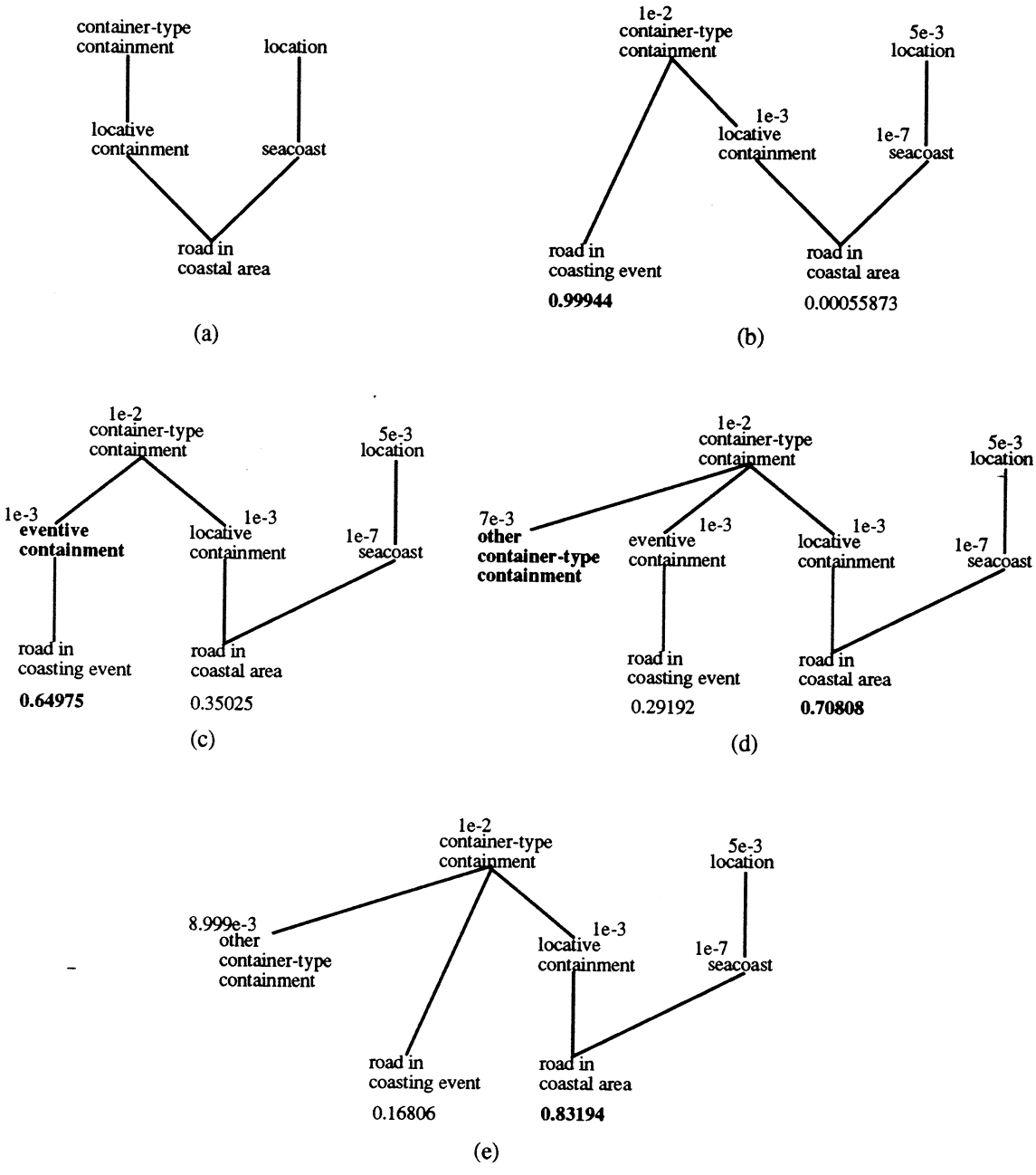


Figure 7.12: Selectional preferences (see text).

When atypical role fillers are used, the absolute probabilities of the feature structures will be lower. However, if there are no other hypotheses under consideration—the more typical role fillers having been ruled out by structural cues—then the atypical structures will have the highest conditional probability.

The probabilistic model tries to strike an balance between choosing oft-encountered substructures and having the substructures cohere. I suspect any correctly formalized theory of coherence necessarily reduces to a probabilistic theory. Theories of coherence tend to be intuitive collections of principles with conflicting pulls. Most coherence theories that have been put forth in the past have been in incomplete states of formalization. When examined with care to remove paradoxical inconsistencies, they are seen to be probabilistic, as for example Thagard (1991; see Thagard 1989, 1990 for background) has discovered. As mentioned in section 3.1.2, Charniak & Shimony (1990) have similarly shown a wide class of cost-based abduction models to be probabilistic. The axioms of probability are so uncontroversial that any reasonable non-deductive framework meets them at some level of reduction.

The probabilistic model can also be viewed in terms of abduction. The hypotheses produced by marker propagation are abductive hypotheses in that they are candidate explanations for the input evidence **e**. As Pearl (1990) and others have observed (see section 3.1.2), the weakness of purely abductive models is their inability to discriminate between alternative explanations. Here, such a means is provided on the basis of statistical experience.

Note that when I say "probabilistic theory" I mean a method of evidence combination. There is no substitute for having the correct structural model underlying the distribution. In concrete terms, this means we cannot get away with, for example, probabilities on only conceptual categories. Norvig & Wilensky (1990b, 1990a) observe that a weakness of Goldman & Charniak's (1990b) probabilistic story understanding model is that it does not take into account patterns of conventional *usage*. Knowledge of what discourse usage patterns are typical, as represented in the present model by constructions, is crucial.

7.3.3 Approximation Accuracy

The approximate maximum-entropy model works by sacrificing accuracy in the distribution outside the hypothesis space. Consider the example of figure 7.11(b). The two dummy events corresponding to the *road* and *coasting accomplishment* constraints should not really be disjoint, because many conceptual structures that are not *road*s are also not *coasting accomplishment*s. We are essentially betting that by using specific enough constraints *within* the hypothesis space, the effect of ignoring jointness conditions in the complement space will be negligible. The preliminary investigations have indicated that this strategy is feasible and yields reasonable behavior.

7.4 Interaction Among Knowledge Domains

The models described above are adaptive since they compute different results depending upon previous instances, and bias their interpretations towards more commonly encountered substructures. We now examine some specific examples demonstrating how previous usage patterns in syntactic, semantic, and conceptual domains can influence the preferred interpretation. In this section again the approximate maximum entropy model is employed, though much of the

discussion also applies to other similar models like the canonical distribution model.

7.4.1 Mental Images, Lexical Semantics, and Conceptual Biases

Semantic and conceptual biases have been used throughout the previous examples. Some more complex, realistic examples are presented here.

The influence of a conceptual category—*Highway 1*—is seen in figure 7.13. Two sets of probabilities are shown. If the *Highway 1* concept is used with relatively low frequency as in the top row, the preferred interpretations are ⌊*road in coastal area*⌋ and ⌊*road along coastline*⌋. If, however, *Highway 1* is a more frequently used concept as in the bottom row, the preferred interpretation switches. The dramatic swing in the conditional distribution is caused by a change in $P(Highway\ 1)$ of only a factor of ten. This sort of sensitivity appears to fit the informal survey, in which all native and most longtime Californians preferred the Highway 1 interpretation to the generic road in a coastal area, but shorter-term residents chose the generic interpretation.

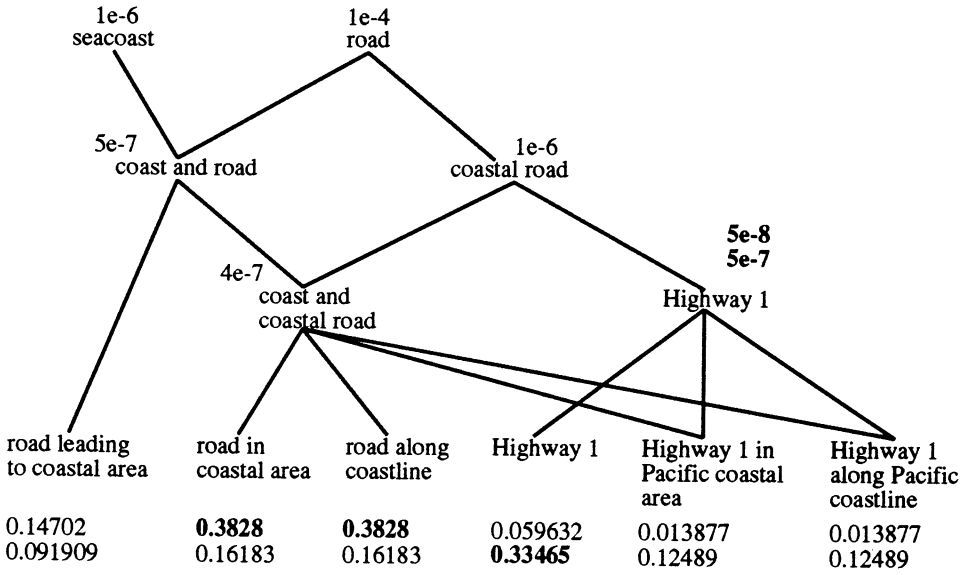


Figure 7.13: Influence of *Highway 1* with low- and high-frequency marginals (see text).

Figure 7.14 shows how the relative typicality of image schemas can influence the preferred interpretation. In the top row the marginal probability on the *containment* schema is ten times that on the *linear order locative* schema, with the result that the ⌊*road in coastal area*⌋ and ⌊*road along coastline*⌋ interpretations are selected. If the marginal probabilities on the image schemata are switched, ⌊*Highway 1*⌋ becomes the preferred interpretation, with ⌊*road leading to coastal area*⌋ second, strongly dominating the two originally preferred interpretations. That the Highway 1 interpretation is preferred shows the influence of other competing conceptual factors. Here we see the tight integration of semantic and conceptual constraints provided by this model.

Deep conceptual knowledge comes into play when there are no strong syntactic or semantic constraints. The compound *championship dollar dinner* (from Warren 1978, p. 204) is difficult to interpret, but *championship check banquet* is far easier.

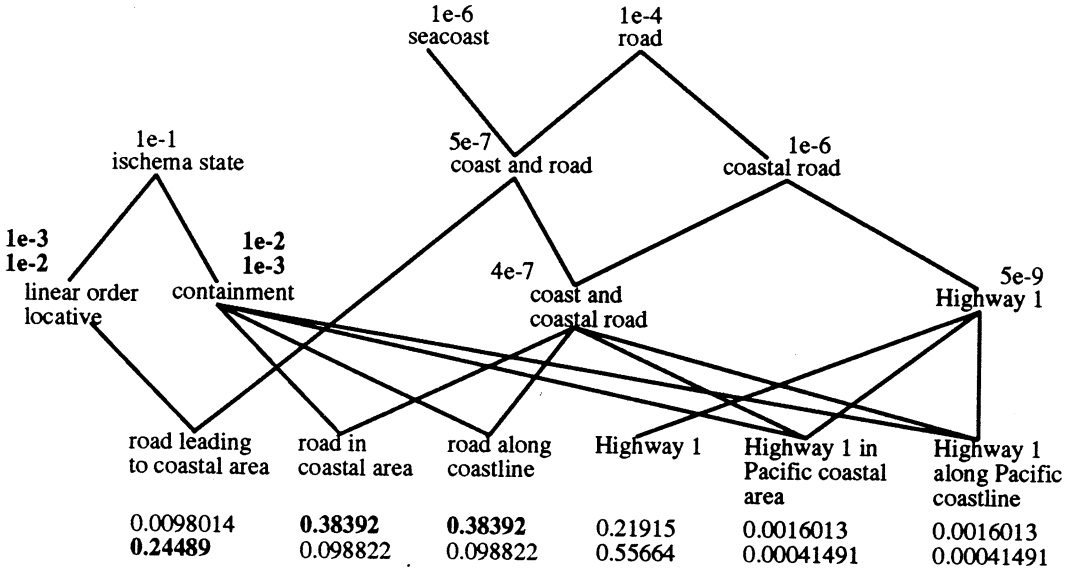


Figure 7.14: Influence of image schema marginal constraints, for the cases when *containment* has a higher marginal, and when *linear order locative* has a higher marginal.

7.4.2 Construction Biases

Thus far we have been concentrating on examples where semantic and conceptual constraints are integrated. We now see that the method extends straightforwardly to the integration of lexico-syntactic constraints as well.

The computation of a hypothesis' probability also factors in the frequencies of the constructions involved in the same way as semantic and conceptual constraints. In figure 7.15 several constructional constraints are added to the semantic and conceptual constraints from the previous example. For example, P(*C:NN:ischema state*) is the probability of a noun compound being used to express a static image-schematic relationship. Three sets of probabilities are shown, corresponding to the three different marginal values shown for P(*C:NN:containment*). As the marginal value decreases—i.e., the less typically the construction is encountered—the more the preferred interpretation shifts from strongly disposed towards ⌊*road in coastal area*⌋ to strongly disposed towards ⌊*road leading to coastal area*⌋.

Figure 7.16 shows an example where the marginal constraints for the different uses of *"coast"* have also been added. P(*C:coast:seacoast*) and P(*C:coast:coasting accomplishment*) are the marginal probabilities of *"coast"* being used to express *seacoast* and *coasting accomplishment*, respectively. The two sets of probabilities correspond to a reversal in the values of the marginal constraints.

The change in preference results from a relatively small difference in usage frequencies. Remember that we are modelling automatic rather than controlled inference; in cases where the usage frequencies and other factors are closely balanced, it is reasonable to believe that multiple interpretations occur automatically to the listener, some of which are subsequently rejected through controlled inference processes. Even though the final interpretation chosen by informants in the informal survey was not ⌊*coasting road*⌋, this interpretation did occur to many of them. If indeed

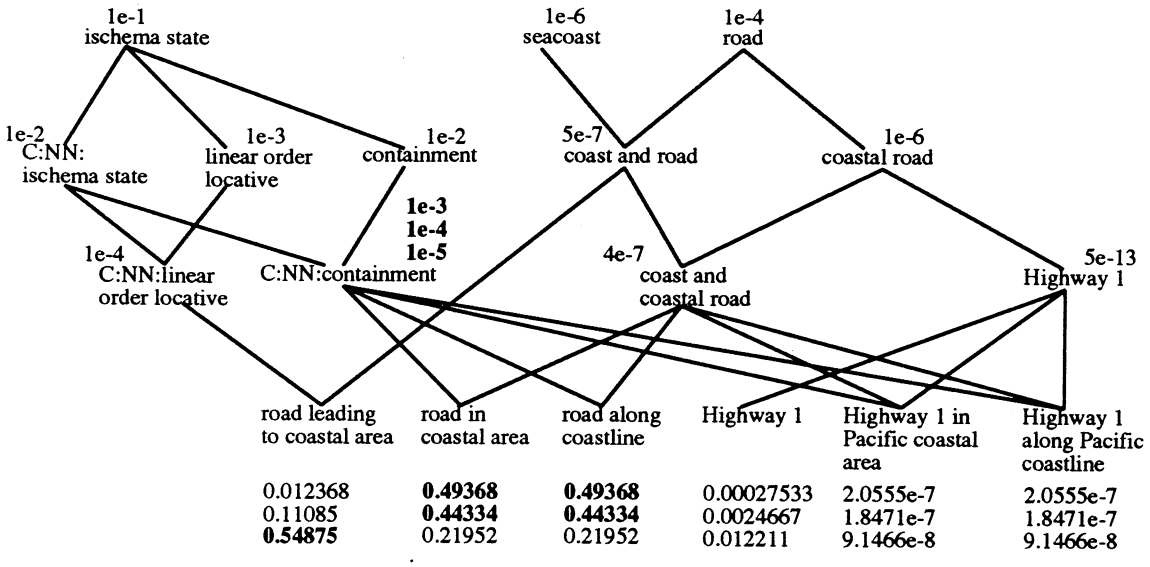


Figure 7.15: Influence of constructional marginal constraints (see text).

"coast" were more frequently used to express the accomplishment of coasting, it would not be surprising if the hearer automatically generated the incorrect interpretation. This would be a form of what has been called a "semantic garden path", in which the hearer recognizes a conceptually faulty initial interpretation and subsequently corrects it. Thus the sensitivity to the frequency of *"coast"* appears to model the phenomena well.

7.4.3 On Collocations and Word Co-occurrence Patterns

Collocations are lexical sequences whose elements co-occur more often than predicted by purely random combination. There is no sharp boundary between collocations and constructions. In the most general sense, the elements of a collocation can be categories like *N* rather than particular lexemes. A distinction should be made between *conventional collocations* that are associated with a conventionalized use, and *co-occurrence collocations* that simply indicate statistical frequency non-uniformities that might happen for less obvious reasons. Unfortunately, while most linguists mean the former, definitions of the term "collocation" usually imply the latter.[7] In thinking about the probabilities on various constructions, however, it is important to decide whether a constraint value should go on a signification construction, indicating a conventional collocation, or on a purely syntactic construction, indicating a co-occurrence collocation.

With regard to nominal compounds, knowledge of conventional collocations of nominals weigh heavily in deciding the interpretation of a compound. For example, consider *county unit system* (from Warren 1978, p. 133), which no one in my informal survey was able to interpret easily. In contrast *narcotics unit system*, which has a semantic structure that *county unit system* could have paralleled, contains the (weakly) conventional collocation *narcotics unit* facilitating interpretation. However, as we saw in chapter 1, the conventional construal of a compound can

[7] Actually, linguists often mean a third class of *non-transparent collocations* whose meaning cannot be accounted for by standard (e.g., "compositional") theories, like *bobby pin* or *sack tobacco* (from Warren 1978, p. 260).

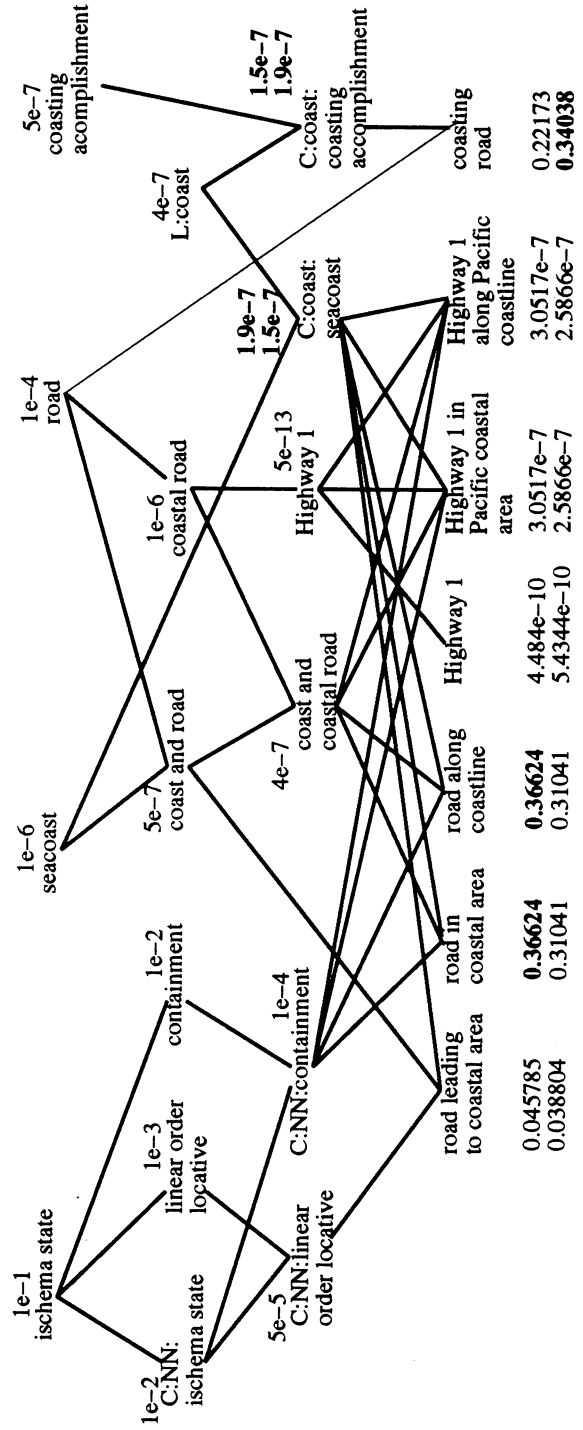


Figure 7.16: Influence of marginal constraints on constructions for individual lexemes (see text).

still be broken by a sufficiently discriminating context. Thus a rule of the form "always take the conventional interpretation if one exists" is inadequate. The probabilistic model integrates knowledge of conventional collocations with the other considerations, and, all other things being equal, the conventional interpretation will prevail.

The collocation *narcotics unit* is a lexeme sequence. Constructions allow us to specify marginals not just on lexeme sequences, but on pattern combinations involving lexemes, syntactic categories, and semantic and conceptual roles. For example, Warren (1978, e.g., p. 212) observed a high frequency of certain nouns, and subclasses of nouns, appearing in particular kinds of compounds. The nouns *department, company, firm, institute,* appearing as the head noun are a good indication that the modifying noun serves to indicate some purpose. We encode this using a relatively high marginal on constructions that either include the specific lexemes or constrain the conceptual structure, as shown in figure 7.17. This helps disambiguate *cleaner equipment firm* by providing additional coherency for the *cleaner equipment* substructure. This method would help formulate many other patterns like *pet* ⟨*animal*⟩, ⟨*nationality*⟩ *restaurant*, ⟨*topic*⟩ *committee*, and so forth.

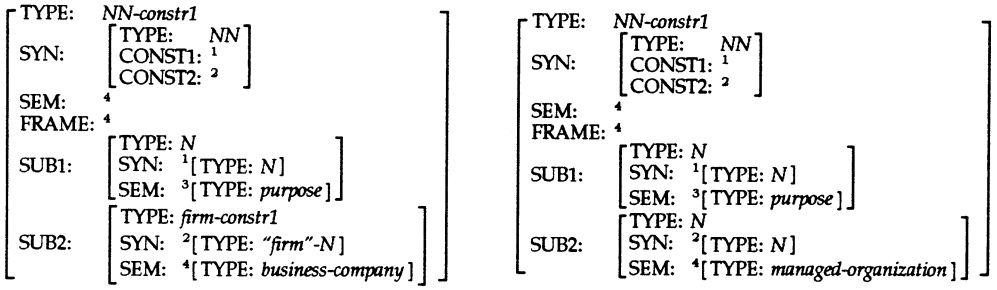


Figure 7.17: Constructions on which to place high marginal values (see text).

Statistical data-gathering techniques use large text corpora to produce co-occurrence rather than conventional collocations; such methods are discussed in section 7.5.1. This is a valuable paradigm for identifying large numbers of collocations that might be conventional. Church & Hanks (1989) have suggested using a mutual entropy measure to judge whether word pairs are collocations (in the co-occurrence sense). If those sequences with the highest measure are stored for use during evidential interpretation, then maximum entropy is the most appropriate way to compute the conditional distribution.

7.4.4 Patterns of Nesting

Another type of structure for which it is might be useful to constrain marginal values are longer nominal sequences. Such constructions specify the nesting of the compounding along with salient syntactic, semantic, and conceptual factors. In a sense these are probabilistic context-sensitive parsing rules. Although I have not investigated nesting patterns, at the very least it should be necessary to include preferences for left-nested [[*N N*]*N*] compounds, as shown in figure 7.18, and right-nested [*N*[*N N*]] compounds.

In postulating more specific patterns care should be taken not to include constraints that

add no information (as measured by the entropy of the distribution) to what the simpler existing constraints already say. I suspect not many specific patterns will be justifiable. For a specific constraint to be useful, it must change the marginal value from what it would otherwise be if the only available constraints were on the substructures. If we already have marginal values on both the inner and outer compounds that are specify syntactic, semantic, and conceptual contexts, there is a great deal of information already present.

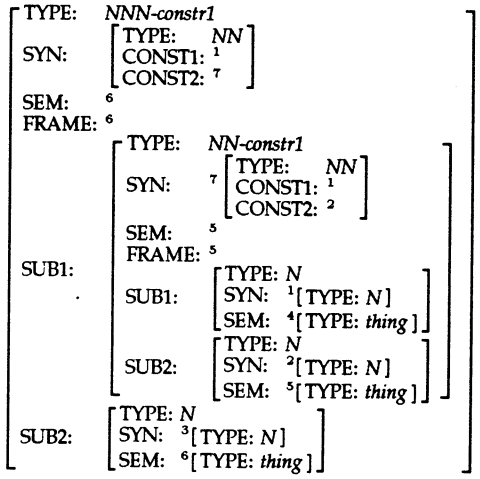


Figure 7.18: A generic construction for constraining the marginal probability of left-nested structures.

7.4.5 Contextual Priming

Contextual priming is modelled by presenting the context as additional evidence. Priming effects occur when the results of recent perceptual, automatic inference, or controlled inference processes are still active when new input is presented, thus biasing automatic inference processes. This corresponds in the present model to a form of recurrency, where part of the input to automatic inference is the previous context. In probabilistic terms, the conditioning event is a feature structure that now incorporates the combined input and context. Because it is not known how the context might be structurally related to the new input, untyped attributes $\emptyset_i$ must be used as sketched in figure 7.19. The contextual structures in those positions can be thought of as fragments of conceptual structures that must be present in any hypothesis interpretation, but whose ultimate preferred position within the interpretation is not yet known. The set of hypothesis interpretations includes various potential positions for the contextual structures (including the "non-attached" interpretation, which simply leaves the contextual structures in their untyped roles). To determine the preferred position, the conditional distribution is computed over the hypotheses as before, and the maximum probability interpretation is taken. If any hypothesis relates the contextual fragment to the new input compound in any typical (high marginal value) configuration, that hypothesis will be preferred to the non-attached interpretation.

Any type of contextual structure is permitted, allowing lexical and constructional as well as conceptual biases. Clearly only a limited amount of structure can be carried over from one

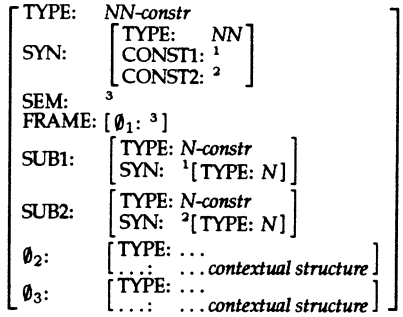


Figure 7.19: Incorporating contextual structure into the input evidence.

inference pass to the next, since otherwise the context would rapidly grow to an enormous size. No theory is attempted here as to what contextual structures to retain and recurrently present to automatic inference.

7.4.6 The Need for Lexical Redundancy

As we have seen, it is useful to store redundant constructions when their marginal values cannot be computed from more general constraints. Such constructions are only redundant in the sense that the strings they parse could also be parsed using other constructions in the knowledge base. However, this is not quite the sense of lexical redundancy meant by the typical linguistic parlance. Here inheritance hierarchies eliminate actual storage redundancy. What I mean is that the statistical patterns captured by the *marginals* on redundant constructions weigh into the preference decision, both for recognizing conventionalized collocations, and for integrating general and specific rules to handle novel constructs.

Of course there are alternative ways to encode the same distribution using different constructions. The general problem of finding the minimum number of marginal constraints needed to encode a distribution is difficult. Moreover, the problem of deciding which distribution best characterizes a data set is an open issue. There are problems having to with avoiding under- and over-generalization, some of which are discussed in chapter 8.

In general, redundancy is an area where AI goals sometimes diverge from the linguistic goal of finding elegant characterizations of aspects of human language processing (e.g., acceptability of a string) without necessarily producing computationally tractable models. What constitutes an elegant characterization is an open question. Parsimony is often taken to be a deciding factor, because the shorter the grammar, the more generalizations presumably captured by the grammar. A parsimonious grammar leads to storage efficiency; however, it does not necessarily lead to time efficiency. Both kinds of efficiency are important to natural language processing. Humans use language effectively despite the finite bounds on their processing resources, and to account for this requires a different notion of parsimony. A parsimonious grammar can increase time efficiency since there are fewer grammar rules for the parser to consider. It can also decrease time efficiency since more time is needed to expand a series of general rules than to apply a pre-stored equivalent specific rule. To maximize time efficiency, not only the most general rules should be stored, but also specific rules that are frequently used. Such a set of rules may not transparently characterize

language. The exact boundaries of the trade-off are still under study in machine learning research; in the present work both general and specific rules are assumed to exist, but the particular choice of rules is representative rather than theoretical.

7.5 Non-Adaptive Sources of Statistics

The primary emphasis of this work is to ground quantitative measures in language modelling by relating them to the agent's experience. Nonetheless we are a long way off from constructing a language agent capable of autonomously learning all the semantic and conceptual concepts needed. In this section we consider some alternative sources of statistics that could be used in the probabilistic models as a practical intermediate step.

7.5.1 Lexico-Syntactic Categories

Automated large-corpora analysis techniques are proving useful in gathering reliable statistics on lexical and syntactic categories. In one of the best-known efforts, Francis & Kučera (1982) compiled a comprehensive frequency list and ranking of lexemes in the Brown corpus of American English (Kučera & Francis 1967). A tagged version of the corpus was used, in which an automated technique (Greene & Rubin 1971) produced 77% of the tags and the remainder was performed manually. Since Warren's (1978) nominal compound corpus was also drawn from the Brown corpus, this data should be invaluable. An example of the sort of information available is shown in figure 7.20. Unfortunately, no collocational counts were performed in the study.

Eegs-Olofsson (1990) has applied a newer automatic tagger and phrase parser to the Brown corpus, that uses the more sophisticated tags from the Text Segmentation for Speech, or TESS, project on the London-Lund corpus (Svartvik 1990). Johansson & Hofland (1989) have performed a similar study on the Lancaster–Oslo/Bergen (LOB) corpus (Leech & Leonard 1974), which was designed to collect British English data in as similar a fashion as possible to the Brown corpus. While their study includes collocations beginning with verbs and adjectives, again unfortunately there is no data on compounds.

Of course, automated tag-and-parse methods cannot determine the proper bracketing for nested nominal compounds, as this would require deep semantic and conceptual analysis. However, they can be used to find frequently occurring collocations as described below, which can then in turn be used to increase the accuracy of automatic bracketing attempts. If a collocation that has previously occurred frequently is subsequently found in a nested compound, its past typicality is evidence for bracketing it in the new compound. Naturally, text bracketed in this manner cannot be used directly for gathering statistics on compound bracketing, as this would be circular, but the automated bracketing could make hand-analysis faster and less tedious. After hand-checking, the bracketed compounds could then be statistically summarized.

Less restricted extraction techniques have been suggested for bigrams, trigrams and n-grams. The most general methods are used to obtain fixed-window word associations (Church & Hanks 1989, 1990; Smadja 1990, 1991b, 1991a; Hindle & Rooth 1991).[8] In the latter, the words between which associations are measured do not have to occur in sequence; rather, they are

[8]Note that the term "bigram" technically implies sequence, although it is used in a relaxed sense by some authors (e.g., Smadja 1990) who use it to denote fixed-window pairwise associations between words.

Lexeme	Tag	Frequency	No. of genres found in	No. of samples found in
coast	**noun**	**67**	**14**	**44**
coast	singular/mass	32	11	22
coast	in headline	2	2	2
Coast	in title	26	10	21
Coast	in title and headline	1	1	1
coasts	plural	6	4	5
coast	**verb**	**3**	**2**	**2**
coasted	past tense	2	2	2
coasted	past participle	1	1	1
coastline	**noun**	**1**	**1**	**1**
road	**noun**	**262**	**14**	**102**
road	singular/mass	179	14	76
rd.	singular/mass	0	0	0
Rd.	in title	3	1	3
road	in headline	1	1	1
Road	in title	15	6	10
Road	in title and headline	2	1	1
road's	possessive	4	2	2
roads	plural	53	10	26
roads	in headline	2	2	2
Roads	in title	3	2	2

Figure 7.20: Some statistics on the Brown corpus (see text).

within a predetermined distance from each other (thus "fixed-window"). This corresponds to not using ordering relations within a construction. The window size may be a syntactic structure, for example, a sentence, or it may be some number of words. Additionally, more sophisticated distance gradations have been proposed. Church & Hanks (1989, 1990) defined a mutual information version of word association between two words w_1 and w_2:[9]

$$\text{(7.8)} \qquad \textit{association}(w_1, w_2) = \log \frac{P(w_2 \textit{ follows within five words of } w_1)}{P(w_1)P(w_2)}$$

Other measures like "depth" and "height" have also been suggested (Smadja 1990). I suspect there is a certain value to such constructions, but that much of their disambiguation power arises from rough correlations to semantic and conceptual regularities. Since the goal of the models I propose is to incorporate real semantic and conceptual evidence, most word association constructions would probably become information-theoretically redundant in the long run.

I would suggest an interesting extension of the entropy approach, again with the goal of generating large lists of bracketed compounds for human checking. The approach uses Hindle's (1990) noun-similarity metric. The metric estimates the mutual information from a large corpus between a verb v_i and its subject or object n_j:

$$C_{subj}(v_i, n_j) = \log \frac{P(v_i, subj(v_i) = n_j)}{P(v_i) \sum_{m \neq i} P(subj(v_m) = n_j)}$$

$$C_{obj}(v_i, n_j) = \log \frac{P(v_i, obj(v_i) = n_j)}{P(v_i) \sum_{m \neq i} P(obj(v_m) = n_j)}$$

The "subject similarity" between two nouns is then defined as in terms of minimum overlap:

$$\text{SIM}_{subj}(v_i, n_j, n_k) = \begin{cases} \min(C_{subj}(v_i, n_j), C_{subj}(v_i, n_j)) & \text{if } C_{subj}(v_i, n_j) > 0 \text{ and } C_{subj}(v_i, n_k) > 0 \\ |\min(C_{subj}(v_i, n_j), C_{subj}(v_i, n_j))| & \text{if } C_{subj}(v_i, n_j) < 0 \text{ and } C_{subj}(v_i, n_k) < 0 \\ 0 & \text{otherwise} \end{cases}$$

and the "object similarity" is defined analogously. Finally, noun similarity is defined as

$$\text{(7.9)} \qquad \text{SIM}(n_j, n_k) = \sum_i \text{SIM}_{subj}(v_i, n_j, n_k) + \text{SIM}_{obj}(v_i, n_j, n_k)$$

In Hindle's initial study, the nouns judged most similar by this syntax-based metric show surprising semantic similarity. This metric could be useful for automatic bracketing of nested compounds. If there is no relevant collocational information, the correct bracketing can be guessed at by grouping the most similar nouns first.

[9] In this section I am taking some liberties in transcribing formulae to make their meaning more transparent. In information theory $\log_2$ is used, but as the base is insignificant to a constant factor I will omit it. The original concept of mutual information as defined by Fano (1961) is

$$I(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

in which $P(x, y)$ does not impose linear ordering on the events x and y.

As mentioned earlier, collocations consisting of particular lexemes that occur non-randomly can be identified using entropy to measure the amount of distributional information added by the frequency of the collocation. This technique could be extended to constructions that specify syntactic categories rather than particular lexemes, thus yielding statistics on constructions of varying specificity.

Using large-corpora techniques for hypothesis filtering. Another interesting potential use of large-corpora techniques has been pointed out by Hearst (1991), who proposes an automated method for coarse disambiguation of noun homographs. The method is based on learning orthographic, syntactic, and lexical features near a noun by sampling an online encyclopedia. Such a technique could be used for nominal compound interpretation as a preprocessing filter for generating interpretation hypotheses. This would act in tandem with marker passing processes to find more probably pertinent hypotheses. In fact, I conjecture that if adaptive marker propagation techniques were pursued they would rapidly converge with this sort of approach, since an adaptive marker search would similarly prune on the basis of crude surface-level features like orthograph, syntactic, and lexical features.

The same idea could also be applied to the large-corpora techniques suggested above for automated bracketing. Instead of being used to preprocess training set material for hand parsing, they could be used to prune the set of interpretation hypotheses either by setting a lower threshold on bracketing certainty, or by restricting the size of the hypothesis set.

7.5.2 Semantic and Conceptual Categories

Semantic and conceptual categories present much more of a problem than lexico-syntactic categories. Existing automated techniques cannot extract information from large corpora other than what is explicitly tagged or what can be parsed. Until we have an ontology with reliable enough statistics to automate semantic interpretation of large corpora, we cannot bootstrap the statistics gathering process.

Hand analysis of a corpus. The investigator may manually analyze a large corpus of compounds, producing for each compound the feature structure representing the correct parse and interpretation. Statistics on the frequencies of all categories with a sufficiently high information value are then collected. This appears to be the best intermediate approach. It is expensive and time-consuming; however, short of constructing an actual agent, there is no existing method that yields comparable accuracy.

A semi-automatic method could be developed using existing coarse nominal compound interpreters as an aid. The investigator would simply tag each compound with enough semantic features to make it feasible to apply an interpreter such as Leonard's (1984). The main advantage of Leonard's system is that the manual overhead is relatively low compared to other hand-analysis methods. On the other hand, the accuracy rate and level of conceptual detail of the interpretations produced by the existing version is probably too low to be of much assistance in generating detailed feature structures.

Borrowing approximate statistics. Studies like Warren's (1978) contain relative frequency counts on coarse semantic categories. An example of the kind of statistics available is shown in figure 7.21.

These can be hand-massaged to fit the ontological hierarchy. The most immediate difficulty is to mediate the differences in ontology.

Subgroups	Press			Informative			Imaginative			Total
	A_1	A_2	B	E	G	J	K	L	N	
Obj-Part	20	19	7	32	16	10	29	23	37	193
Group-Member	96	45	51	6	4	8	8	8	4	230
Whole-Outline	10	11	9	34	2	14	7	7	11	105
Residual cases	11	17	11	3	3	4	1	5	2	57
Obj-Quality	7	3	7	16	7	34	4	-	3	81
Whole-Extension	7	9	6	10	1	14	2	-	4	53
Whole-Abstract Shape	4	3	2	1	3	14	3	-	1	31
Possessor-Legal Belonging	17	7	13	5	2	1	11	17	5	78
Possessor-Habitat	7	6	5	3	5	1	8	5	1	41
Authority-Subordinate	25	6	13	1	3	-	2	2	-	52
Total	204	126	124	111	46	100	75	67	68	921

Figure 7.21: Example of Warren's (1978) statistics on nominal compound semantic relations. The letters denote different subsamples.

Unfortunately, Warren's classification does not include subdistributions for different semantic classes of nouns. In other words, the relative frequencies for semantic relations are not sensitive to the types of constituent nouns. This shortcoming is attributable to Warren's non-interpretive paradigm.

Another difficulty with using Warren's statistics is that her analysis assumes each compound has exactly one interpretation. In fact, many, if not most, compounds can be interpreted many ways. Though usually the interpretations are incompatible and the context can be used to select one, sometimes multiple interpretations can be compatible (Norvig 1989). For example, it does not really matter whether *vaudeville tapdance routine* (from Warren 1978, p. 136) is taken to mean *routine for a vaudeville tapdance* or *tapdance routine for vaudeville* since they both mean the same thing. One possible way to make the statistics reflect this would be to update the marginals as though half an instance of each had been seen.

Automatic acquisition from machine-readable dictionaries. For some time ontology builders have been attempting to construct hierarchies from machine-readable dictionaries. Although dictionaries do not yield statistics on semantic and conceptual usage frequencies, ontology-building is still in such a primitive state that the most pressing issue is the structure of the hierarchy. Chodorow *et al.* (1985) describe a semi-automatic method for constructing multiple hierarchies of hypo- and hypernyms under human supervision, by reading *Webster's Seventh New Collegiate Dictionary* (WEBSTER 1963).

More recently, Guthrie *et al.* (1990) have been using more sophisticated techniques to extract relations between word *senses* in the *Longman Dictionary of Contempory English* (LDOCE 1978). Although lexical relations such as hyponymy are quite different from conceptual primitives, the results could be used as a starting point for the ontology.

Machine-readable dictionaries are also unique in that they could potentially be useful for automatically extracting significative constructions. Hindle & Rooth (1991) found that their bigram approach to finding associations useful for determining prepositional phrase preferences missed 30–62% of the preposition-noun and preposition-verb associations listed in the machine-readable version of the COBUILD dictionary (Sinclair *et al.* 1987), depending on the significance threshold used. This suggests that dictionary extraction is not entirely supplanted by large-corpora techniques. Moreover, unlike automated collocation extraction methods, a dictionary relates constructions to a semantic ontology.

Hand estimation. Finally, the most practical current method is still, unfortunately, to manually estimate probabilities that appear reasonable. Obviously there are major problems with this approach and it is difficult to see how one could take it very seriously in the conceptual domain. However, hand experimentation could yield reasonable initial guesses at the more coarse lexical semantics levels.

Chapter 8

Chapter 8

Learning and Generalization

The model constructed in chapters 6 and 7 assumes that (1) the automatic inference mechanism tries to record the distribution of input-output conceptual structures, but due to resource constraints is only able to record partial information in the form of marginal constraints, (2) maximum entropy is used to complete the parts of the distribution that could not be stored, and (3) the correlational terms on which marginal constraints are recorded are chosen by the investigator before any training data is seen. Up to a point, the last assumption is sensible from the engineering standpoint; I discussed some approaches to collecting statistics for prespecified categories in chapter 7. Beyond that point, however, detailed semantic and conceptual constraints would probably elude hand-engineered approaches. From the theoretical standpoint the assumption is a reasonable interim approach, but again it would be unsatisfactory in the long run since it is unlikely that linguistic agents are born equipped with any but the most abstract innate categories. It would be more desirable to have a motivated theory explaining what correlational terms should be adaptively selected given the agent's experience.

As an initial step toward such a theory, this chapter constructs a framework that defines an ideal normative distribution over the space of feature-structures. The framework's hypothesis is that, given predetermined storage bounds on correlational terms, the adaptation mechanism should choose the ones that yield closest approximation to the normative distribution, when completed by maximum entropy. The framework is also closely related to the equally important issue of *generalization* in learning.

In machine learning terms, a concept formation theory is needed for correlational terms. Coming up with a general, operational, tractable, online method for concept formation is the major unsolved machine learning problem. Clearly, no attempt to give such a procedure will be made here. Indeed, most machine learning models deal with extremely simple kinds of concepts like feature-vectors, and even feature-DAGs are beyond their expressive capacity. Perhaps for this reason, little if any work in language acquisition has attempted even to characterize the functional desiderata for a statistical learning model. The formal language acquisition approaches of, for example, Wexler & Culicover (1980) and Pinker (1984, 1989) are nonstatistical. Consequently, the models do not degrade gracefully with noisy data—i.e., when some percentage of training instances are incorrect—nor can they make likelihood judgements for disambiguation. On the other hand, the most powerful statistically-based approaches are currently grammar induction methods restricted to context-free or nearly-context-free underlying models (e.g., Fujisaki *et al.* 1991; Brill *et al.* 1990;

Magerman & Marcus 1990; Marcus 1991; Finch & Chater 1991) that cannot handle feature-structures or complex semantic and conceptual frames. What I would like to do, thus, is to take the first step of setting out a formal functional characterization of statistical pattern completion learning with enough expressive power to handle feature-structures.

8.1 Concept Formation, Generalization, and the Normative Prior

Concept formation is closely related to the issue of obtaining desirable generalization behavior. Generalization is necessary for several reasons.

First, whatever pattern completion algorithm is learned by the adaptive agent, it should not be restricted to only classifying or recognizing exact duplicates of training exemplars. Natural language is full of novel constructs signifying novel conceptual interpretations—novel nominal compounds being an excellent example—and to be able to construct novel interpretations at all requires some degree of generalization. Thus, if the training set contains *early morning patrons*, *fair-weather friends*, *weekend guests*, and *summer people*, this should bias the interpretation of a novel compound *10-o'clock scholar*[1].

Second, it takes many training examples to acquire reliable probability estimates, and yet interpretation sometimes needs to be done before enough data has been seen to accurately estimate the probabilities of every possible feature-structure. This is related to the issue in statistical pattern recognition of *overfitting*, where too stringent a distribution is created, with not enough data to warrant it. It is a general statistical principle that the more parameters there are in a system, the more sampling needs to be done. Generalization has the effect of "loosening" the stringency of the distribution.

Third, given finite bounds on storage resources, some incoming information from observed training data must be discarded and this inherently results in generalization. The proposed maximum entropy model actually performs generalization, because only storing partial constraints rather than a complete distribution has the side effect of generalizing. Consider the extreme case where the exact relative frequency distribution of feature-structures in the training set is stored. If this distribution is used to estimate the priors, no generalization is performed since zero probabilities are assigned to any feature-structure not in the training set. On the other hand, suppose only the marginal prior for an abstract correlational term is set, and none of the relative frequencies for any of its subordinate abstract or complete terms are recorded. Then maximum entropy replaces all the priors for the subordinate complete terms with equal probabilities, even if they originally had different training set frequencies. Effectively, a single generalization about all the complete terms subsumed under the recorded marginal is made. Finally, at the other extreme, if all the relative frequencies are discarded then maximum entropy makes all complete terms equiprobable. This can be seen as extreme overgeneralization because it generalizes every training instance to every complete term.

Since generalization with maximum entropy depends on the choice of correlational terms, the concept formation issue must still be resolved. To prescribe the choice of correlational terms, the approach of defining a function to learn the *normative prior* is proposed in this chapter. The normative prior is a theoretical prior distribution on the concept space that is not subject to resource bounds. It is defined by a mathematical function that maps the entire training set to to a

[1] All from Warren (1978, p. 179).

prior distribution. As we will see, generalization is built into the function. The concept formation hypothesis is that some finitely bounded number of correlational terms should be chosen, such that in conjunction with maximum entropy completion, they minimize the discrepancy with the normative prior.

8.2 Completion and Generalization in Vector Spaces

Before addressing the normative prior for feature-structures let us consider the problem for the simpler case of flat feature-vectors. Recall from section 6.6.2 that the shape of the concept space is determined by the set of abstractors. Suppose the only abstractor used is feature deconstraint, which can be thought of as substituting "don't care" or "x" bits for specific feature value constraints. Figure 8.1 shows the space, which is a lattice, for both (a) feature pairs and (b) feature triples, i.e., 2-vectors and 3-vectors of binary-valued feature dimensions. The row of shaded boxes at the bottom denote the complete terms, which have the same feature values as their parental abstract terms. The distinction between them and the row immediately above them is not very important for feature-vectors, but they are included here to make the transition to feature-structures easier later. As another example, if instead the abstractor used is *position-insensitive feature deconstraint*, the lattice shown in (c) results. The lattice shown in (d) is the space under the *final feature deconstraint* abstractor.

Pattern completion is most clearly viewed as an *autoassociation* task where given a partial input pattern, the most plausible completion of the pattern should be constructed. In our case, the partial input pattern is an abstract feature-vector (or feature-structure), corresponding to an internal node of the lattice. The output pattern is the best complete feature-vector under that node. In the rest of this chapter, I will often use the probabilistic terms *event* to describe any node of the lattice, *simple event* to describe any leaf node, and *compound event* to describe an internal node. Remember that the leaves or simple events form a disjoint partitioning, and their priors sum to unity. Also, the prior of a compound event is a marginal prior, and is the sum of all the priors of its leaves.

We now ask what probability distribution over the concept space would be normative. Consider an w-vector space such as those of figure 8.1(a) and (b). As discussed above, the relative frequency distribution over a training set is inadequate as a normative prior; zero priors should not be assigned to any of the potential complete feature-vectors even if they are not in the training set because novel inputs might require them. The major contribution of Carnap's (1952, 1962) classic work on logical probability is an elegant general solution to the zero-probability problem, which he calls the "λ-continuum of inductive methods". It is a family of methods for inducing a prior distribution from a sample or training set, parameterized by λ:

$$P(q_i) \stackrel{\text{def}}{=} \frac{f(q_i) + \lambda/2^w}{f_{\text{total}} + \lambda} \tag{8.1}$$

One way to think of λ is that it weights the tendency toward equiprobable priors. If we set $\lambda = 0$ the priors are exactly the relative frequencies and can include zero probabilities, but if $\lambda > 0$ then all priors are positive. At $\lambda = \infty$ equiprobable priors are assigned to all simple events (complete terms, as opposed to abstract terms which are compound events), and as with unconstrained maximum entropy there is no sensitivity to the sample (Dias & Shimony 1981).

Setting an intermediate $\lambda = 1$ yields the widely used form[2]

$$P(q_i) = \frac{f(q_i) + 2^{-w}}{f_{\text{total}} + 1}$$

This is a step toward solving the zero-prior problem, but unfortunately, none of the methods perform generalization. An instance in the training set only lowers, and never raises, the probability of other similar events.

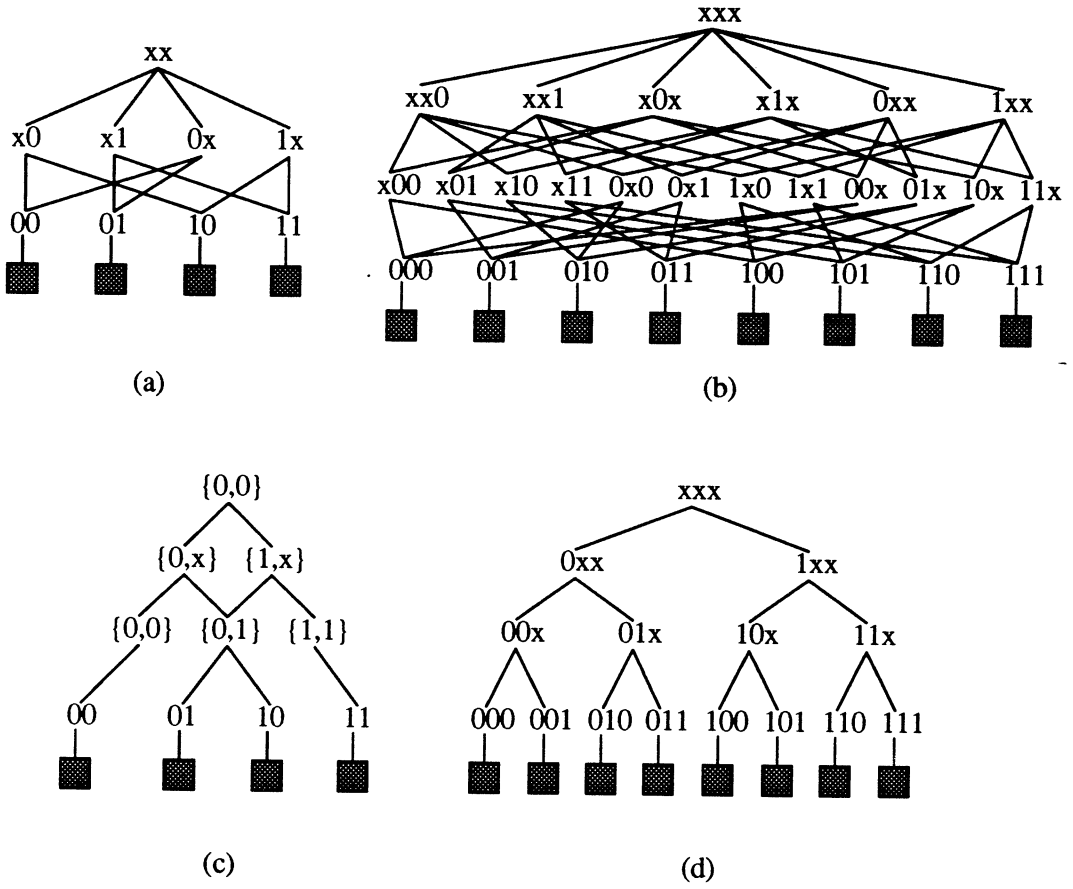


Figure 8.1: Concept space lattices for w-vectors, using (a–b) feature deconstraint, (c) position-insensitive feature deconstraint, and (d) final feature deconstraint.

[2]Interestingly, $\lambda = 2$ yields a corrected version of Laplace's famous "rule of succession". Laplace's own formulation

$$P(q_i) = \frac{f(q_i) + 1}{f_{\text{total}} + 2}$$

turns out to contradict the axioms of probability when $n \neq 1$. However, Carnap's method yields a consistent prior even when $n \neq 1$:

$$P(q_i) = \frac{f(q_i) + 2^{1-w}}{f_{\text{total}} + 2}$$

What I propose is a different continuum of methods for generating priors, parameterized by a value γ that controls how much generalization the priors contain. The γ prior is suggested as the normative prior (assuming something like the ontology of chapter 5 is the underlying event space of possible correlational terms). The extreme ends of the continuum are the same as Carnap's λ-continuum. However, whereas λ dictates the degree of sensitivity to the training set, γ dictates the degree of generalization from the training set. At $\gamma = 0$ no generalization is done, so the priors are simply the relative frequencies. At $\gamma = \infty$ all priors are equiprobable due to extreme overgeneralization.

The full form of the γ prior is simply stated here, and detailed explanation is left to the following section. Denote the set of simple events (complete terms) by $Q = \{q_1, q_2, \ldots, q_C\}$, and let X be a random variable with values ranging over Q. Given a training vector $T = (t_1, t_2, \ldots, t_N)$ where $t_i \in Q$,

$$P_i \stackrel{\text{def}}{=} P(q_i) \stackrel{\text{def}}{=} Pr(X = q_i) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} norm[2^{-d(q_i, t_n)/\gamma}] \tag{8.2}$$

where *norm* means normalization to 1 as follows:

$$P_i \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \frac{2^{-d(q_i, t_n)/\gamma}}{\sum_{k=1}^{C} 2^{-d(q_k, t_n)/\gamma}} \tag{8.3}$$

and $d(q_a, q_b)$ is the *logical distance* between two simple events. Logical distance is discussed below; for now it is sufficient to think of it as a similarity metric between two complete terms.

We first examine two simple examples dealing with flat feature-vectors. To begin with, let us consider the qualitative behavior of the γ prior on the example of figure 8.1(d). I will give the results without explicit calculation. Figure 8.2 shows the effect of a training set with just one instance 010 upon the γ prior, for different values of γ. In (c), the greatest proportion is $P_{010} = 0.4$. A lesser proportion $P_{011} = 0.2$ is assigned to the closest other simple event (complete term); a still lesser proportion for $P_{000} = P_{001} = 0.1$; and finally $P_{100} = P_{101} = P_{110} = P111 = 0.05$. Intuitively, what has happened is that the single instance has distributed its effect among all the simple events. In the straight relative frequency prior, only the training instance's simple event is affected, as shown in (a) for the case $\gamma = 0$. On the other extreme as shown in (e), when $\gamma = \infty$ the effect of the instance is generalized equally to every simple event. Figures 8.2(b) and (d) show how the value of γ controls the degree to which the effect is "smeared" toward progressively dissimilar families of events; the more "smear", the more generalization.

Now consider the full lattice of figure 8.1(b). This example is particularly good for demonstrating an interesting perspective on the γ prior, namely, its relationship to statistical *kernel estimator* methods. Kernel methods are most commonly used for estimating probability density functions in continuous domains, but have also been extended to estimating probability distributions on discrete categorical variables. The idea is to smooth out a distribution as in the previous example, by treating each input instance as a "kernel" K that is spread out over the distribution. The general form of a discrete kernel estimator is

$$P_i \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} K_{in} \tag{8.4}$$

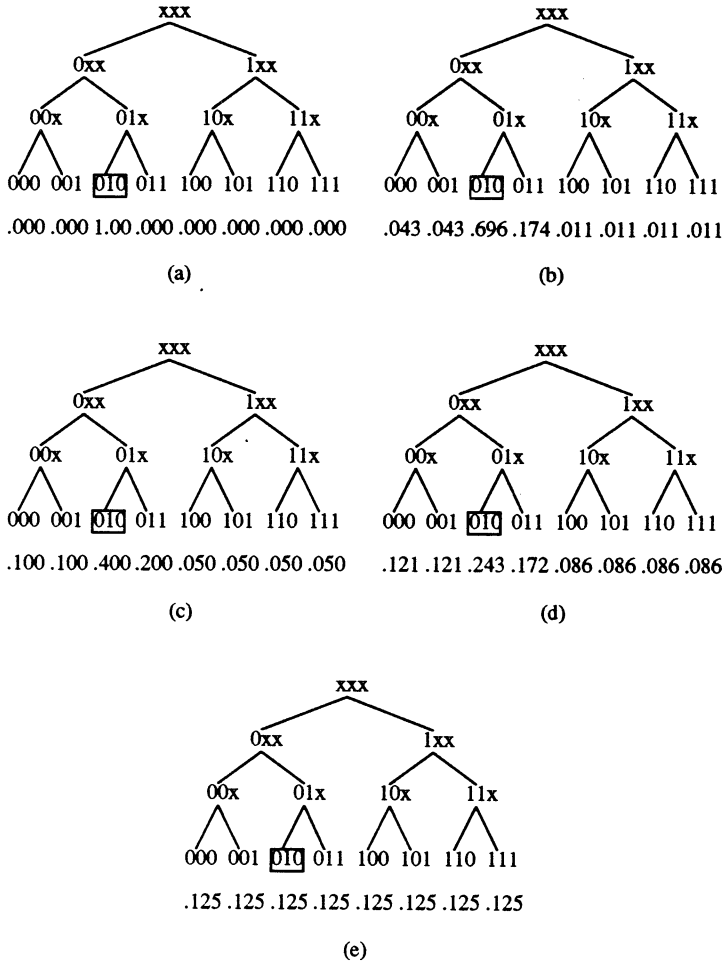


Figure 8.2: Effect of a single training instance 010 for (a) $\gamma = 0$, (b) $\gamma = 0.5$, (c) $\gamma = 1$, (d) $\gamma = 2$, and (e) $\gamma = \infty$. To avoid clutter the simple events (leaf nodes) are not shown.

In the discrete case, binomials are typically used for the kernel function K (in the continuous case, similarly bell-shaped Gaussians are typical). The γ prior is a kernel estimator in the general sense, as can be seen by substituting

$$K_{in} = \text{norm}[2^{-d(q_i,t_n)/\gamma}] \tag{8.5}$$

For the special case of figure 8.1(b), the logical distance between two w-vectors turns out to be simply the number of features in which they differ. This time we will explicitly derive the effect of a single training instance on the prior. The contribution of the training instance to the prior on its own term q_i is, from equation (8.3),

$$\begin{aligned} P_i &= \frac{2^{-d(q_i,q_i)/\gamma}}{\sum_{k=1}^{C} 2^{-d(q_k,q_i)/\gamma}} \\ &= \frac{1}{\sum_{k=1}^{C} 2^{-d_{ik}/\gamma}} \end{aligned} \tag{8.6}$$

Since we know the binomial structure of the $C = 2^w$ simple events (complete terms), we can rewrite the normalization factor as

$$\begin{aligned} \sum_{k=1}^{C} 2^{-d_{ik}/\gamma} &= \sum_{d=1}^{w} \binom{w}{d} 2^{-d/\gamma} \\ &= (1 + 2^{-1/\gamma})^w \end{aligned} \tag{8.7}$$

Thus the contribution of the training instance to any term q_j is

$$\begin{aligned} P_j &= \frac{2^{-d_{ij}/\gamma}}{\sum_{k=1}^{C} 2^{-d_{ik}/\gamma}} \\ &= \frac{(2^{-1/\gamma})^{d_{ij}}}{(1 + 2^{-1/\gamma})^w} \end{aligned} \tag{8.8}$$

If we now define

$$\kappa = \frac{1}{1 + 2^{-1/\gamma}} \Longrightarrow 1 + 2^{-1/\gamma} = \kappa^{-1} \Longrightarrow 2^{-1/\gamma} = \kappa^{-1} - 1$$

then equation (8.8) can be rewritten as

$$\begin{aligned} P_j &= \frac{(\kappa^{-1} - 1)^{d_{ij}}}{\kappa^{-w}} \\ &= \kappa^w (\kappa^{-1} - 1)^{d_{ij}} \\ &= \kappa^w \kappa^{-d_{ij}} \kappa^{d_{ij}} (\kappa^{-1} - 1)^{d_{ij}} \\ &= \kappa^{w - d_{ij}} (1 - \kappa)^{d_{ij}} \end{aligned} \tag{8.9}$$

This is exactly the form used in Aitchison & Aitken's (1976) kernel estimator method.

Despite the fact that the γ prior is technically a kernel method—and despite the fact that it yields a very typical kernel estimator in the regular w-vector domain—there are important differences in the way γ priors are interpreted, versus what the term "kernel method" connotes. First, the

γ method is used here to estimate a prior distribution to be used for MAP (maximum a posteriori) completion—a form of autoassociation, with unsupervised learning. In contrast, kernel methods are generally used to estimate class-conditional distributions that are subsequently used in a much lower-dimensional Bayesian classifier—a form of heteroassociation, with supervised learning. In order to view γ priors in these terms, one must treat pattern completion as a heteroassociation task where the task is to classify an input pattern (which happens to be an "incomplete pattern") into one of a set of output classes (which happen to be "complete patterns"). However, the number of classes is combinatoric in the dimensionality of features, which is opposite from the usual classification paradigm. Second, kernel methods are usually applied to flat feature-vector spaces with simple Hamming or Euclidean distance metrics, as in the above example. The feature-structures require a substantially more sophisticated distance metric than is typically found in statistical kernal methods, in order to handle semi-lattice irregularities as discussed next. Third, under our interpretation of the distance metric, the goal is to model innate cognitive representational biases, not to search for maximum neutrality as in typical statistical applications.

8.3 Completion and Generalization in Feature-DAG Spaces

The simple Hamming distance metric does not work for feature-DAGs with their distorted semi-lattice spaces. This problem is related to the difficulty of representing compositional structures in neural networks, which employ feature-vector representations. Distributed feature-vector representations rely on the intrinsic similarity metric based on Euclidean or Hamming distance. They exploit this property to perform incremental, online learning, because the property has the consequence that small changes in the weights (say, in discriminant functions) correspond to small changes in the category partitioning. It would be highly desirable to have this ability in feature-DAG spaces; however, since there is no obvious mapping from feature-DAGs to feature-vectors that preserves the similarity structure implied by the semi-lattice, there is currently no good method of learning to cluster and discriminate feature-DAG categories.

In this section I first describe the distance metric used in the general case for feature-DAGs. Afterwards we examine a small example of the γ prior's behavior in a feature-DAG semi-lattice.

8.3.1 Logical Distance

The logical distance $d(q_a, q_b)$ between two simple events derives from the bias given by the choice of abstractors. Again, one can see this as the choice of K in a kernel method, though the approach of having explicit abstractors for compositional structures is not the sort of application connoted by "kernel methods". The logical distance is defined in terms of the *logical class cardinality*

$$lcc(q_a, q_b) \stackrel{\text{def}}{=} |\, leaves(lub(q_a, q_b))| \tag{8.10}$$

where *lub* is the least upper bound, or the most specific ancestor common to q_a and q_b. The logical distance is then defined as

$$d(q_a, q_b) \stackrel{\text{def}}{=} \log_2(lcc(q_a, q_b)) = \log_2 |\, leaves(lub(q_a, q_b))| \tag{8.11}$$

Note that because the logical distance metric depends on the least upper bound but not greatest lower bound, the concept space need only be a semi-lattice rather than a lattice.

Logical distance reduces to Hamming distance for the case of the regular w-vector lattice. For example, in figure 8.1(b) there are four leaves under the least upper bound $lub(001, 100) = \text{x0x}$, so the logical class cardinality $lcc(001, 100) = 4$. Thus the logical distance $d(001, 100) = \log_2 4 = 2$, which is the number of components in which 001 and 100 differ.

The other abstractors yield different distance metrics. In the case of figures 8.1(d) and 8.2, $d(000, 001) = 1$ but $d(000, 100) = 3$. Note that in spaces other than regular w-vectors, logical distances are usually non-integers. The logical distance between the two leaves marked with asterisks in figure 8.3 is $\log_2 6$. However, the logical distance between a node and itself is always $\log_2 1 = 0$.

8.3.2 The γ Prior for Semi-Lattice Spaces

For each training instance the γ method distributes a "unit" kernel among all simple events, in a proportion that depends on logical distance. Given a training instance t_n that is a simple event q_j, if the proportion of the "unit" given to the simple event is u_j, then the proportion u_i given to any other simple event q_i satisfies the constraint

$$\frac{u_i}{u_j} = 2^{-d(q_i, q_j)/\gamma}.$$

Let us first examine the extreme-case behavior. At $\gamma = 0$, $u_i/u_j = 0$ and so $u_i = 0$ for all $i \neq j$ and $u_j = 1$, thus degenerating into the relative frequency method. At $\gamma = \infty$, $u_i/u_j = 1$ and so $u_i = u_j$ for all i, thus incrementing every simple event equally, regardless of what the training instance is.

With this general method of distributing the "unit" kernels based on logical distance, distorted or irregular semi-lattice spaces can be handled straightforwardly. Figure 8.3 repeats the semi-lattice we examined in figure 6.16. A training set containing 100 instances randomly generated from the set of simple events (leaf nodes) was created. The distribution of these instances can be seen from the relative frequency distribution shown in figure 8.4(a), which is obtained at $\gamma = 0$. The smoothing effect of setting $\gamma = 0.8$ and $\gamma = 1$ is shown in (b) and (c).

The marginal probability for any compound event is just the sum of all its simple events' probability. If a compound event is comprised of a set of simple events $\{s_1, s_2, \ldots, s_r\}$ where $s_i \in Q$, then

$$\begin{aligned} Pr(X \in \{s_1, s_2, \ldots, s_r\}) &= \sum_{i=1}^{r} \left[\frac{1}{N} \sum_{n=1}^{N} \frac{2^{-d(s_i, t_n)/\gamma}}{\sum_{k=1}^{C} 2^{-d(s_k, t_n)/\gamma}} \right] \\ &= \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{i=1}^{r} 2^{-d(q_i, t_n)/\gamma}}{\sum_{k=1}^{C} 2^{-d(q_k, t_n)/\gamma}} \end{aligned} \tag{8.12}$$

8.4 Directions

In choosing a set of abstractors, the goal is to model innate biases deriving from whatever representational primitives the cognitive mechanism actually employs. Ultimately the choice of

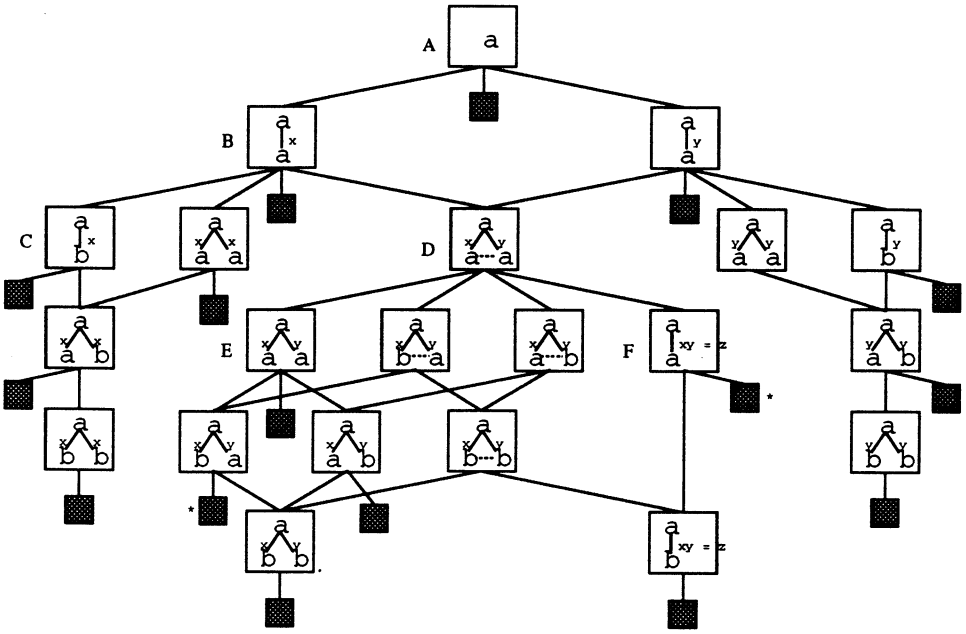


Figure 8.3: Semi-lattice from chapter 6.

abstractors is, as with ontological primitives, validated only by fit to empirical data. Moreover, only empirical tests will tell what values of γ are optimal.

Further flexibility for modelling cognitive biases can be obtained by augmenting the logical lattice structure determined by the abstractors by a weighting scheme. Different abstractors, in other words, might imply greater or lesser similarity distances. A weighted notation would increase the expressiveness of the declarative language for specifying an abstractive bias.

A second important issue is the nature of the *discrepancy function*. A form of cost or error function, the discrepancy function is the function that should be minimized when choosing the correlational terms to approximate the normative prior. Since there is a finite bound on the number of correlational terms, there will always be discrepancies between the γ distribution and that produced by maximum entropy on the marginals of the chosen correlational terms. Different ways of weighting the various discrepancies will produce different discrepancy functions to minimize, leading to different sets of correlational terms. In a sense the choice of discrepancy function is the other innate bias parameter in this model and accordingly, it must be set by further empirical study.

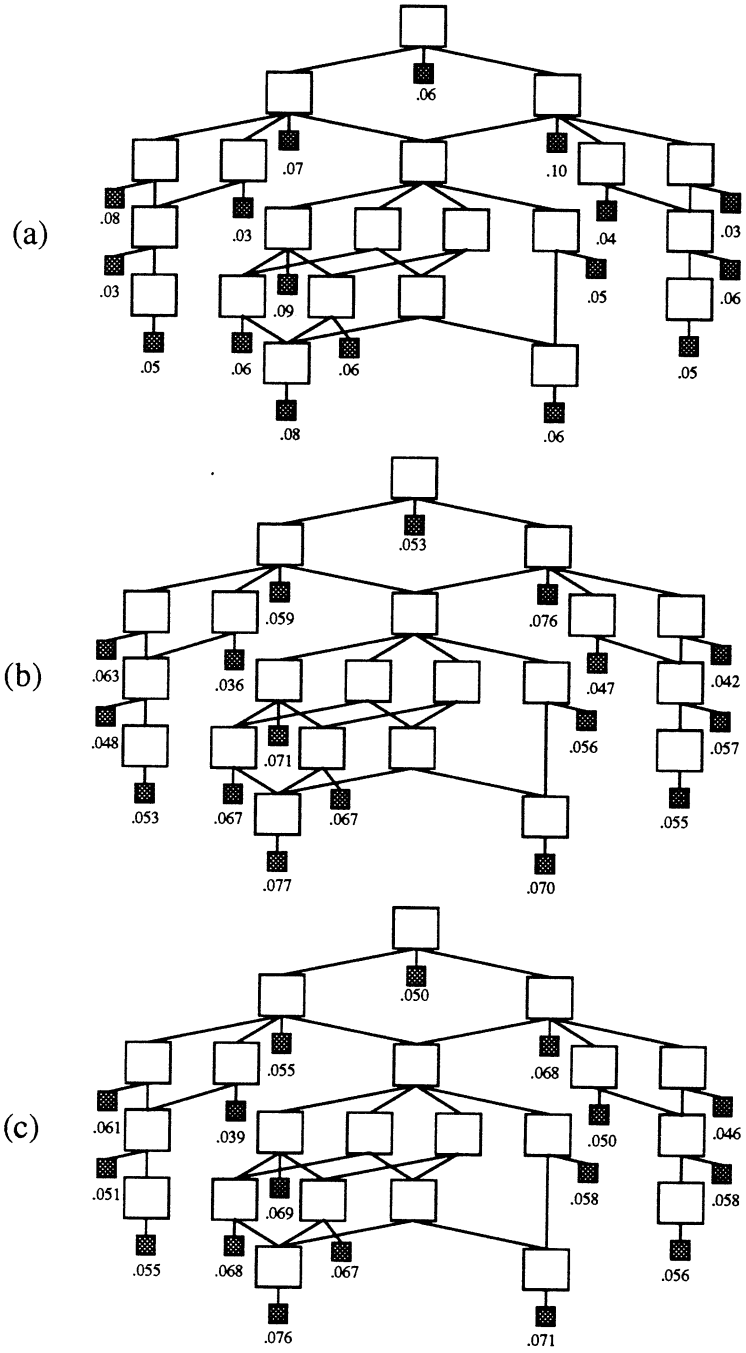


Figure 8.4: The smoothing effect of generalization in the semi-lattice for (a) $\gamma = 0$, (b) $\gamma = 0.8$, and (c) $\gamma = 1$.

Chapter 9

Chapter 9

Conclusion

To recapitulate the claims I have argued for in this work:

- Language understanding, like other mental reasoning and recognition processes, breaks down into two modes: *automatic inference* and *controlled inference*. Automatic inference is pre-attentive, subconscious, reflexive, fairly instantaneous, associative, and highly heuristic. There are both empirical and processing motivations, stemming from resource bounds, that make automatic inference desirable. Much if not most of parsing and semantic interpretation is performed by automatic inference.

- The cognitive ontology includes mental images, lexical semantics, conceptual, and lexicosyntactic modules. Automatic inference extends through all of these modules. The modular ontology approach accounts for a range of subtle meaning distinctions, is consistent with psycholinguistic and neural evidence, and helps reduce the complexity of the concept space.

- Probability theory provides an elegant basis for evidential interpretation, as a model of automatic inference in language understanding. To employ probability theory, the ontology and concept space must be cleanly defined. Toward this end a uniform representation for all the modules is proposed that is compatible with both feature-structures and semantic networks, and makes probabilistic, associative extensions to those frameworks. Theoretical and approximate maximum entropy methods for evaluating probabilities are proposed, as well as the basis for a normative distribution for learning and generalization.

Much of the model remains to be fleshed out. A particularly interesting direction is hypothesis generation. In some ways, hypothesis generation is to automatic inference what automatic inference is to controlled inference. Hypothesis generation must operate on the basis of even shallower cues—say, in the orthographic, morphemic, phonemic, syntactic, lexical domain—but does not have to pinpoint the single best hypothesis. Marker passing strategies have been among the most powerful associative retrieval techniques for generating candidate sets in structured representations. However, their mathematical basis is poorly understood, and they cannot be expected to scale up unless given a more motivated statistical basis. As mentioned in section 7.5.1, other non-marker-passing statistical methods, like those based on analysis of large text corpora, can be used to generate hypotheses. Such techniques, however, do not necessarily produce hypotheses

with detailed conceptual structure. Thus we need to synthesize structured hypothesis generation and adaptive statistical techniques.

Another important issue is incremental learning. Using the device of a normative prior I have suggested *which* marginal constraints (and thus, conceptual terms) should theoretically be stored, but not *how* they could be incrementally chosen by an situated adaptive agent. This is a version of the concept formation problem: given limited storage resources to keep track of past experience, how can the agent maximize the odds of adopting a set of marginal constraints that is close to optimal?

Incremental learning is related to the long-term goal of constructing models with emergent probabilistic behavior. Models may exhibit global probabilistic behavior without explicitly computing probabilities or only computing local probabilities (take for example Markov random fields, Boltzmann machines, or even some variants of backpropagation). Behavior may converge in the long run to probabilistic behavior, or may only be a resource-bounded approximation. Such models may turn out to better implementations of a probabilistic theory, even if they encode knowledge in a less decipherable "black box" representation.

Finally, there is the issue of the innate inductive bias. As many have suggested, quite possibly the only way natural agents can overcome the intractability of concept formation is by being prewired to consider only certain parts of the space, or adapt only along certain paths. Here evidence from cognitive studies will prove invaluable in suggesting architectural and representational restrictions.

In the meantime, the model I have proposed based on approximate maximum-entropy estimation solves some more immediate problems in current natural language modelling. It provides a feasible method for hypothesis discrimination, that permits constraints and evidence from multiple knowledge domains to be integrated. Its probabilistic basis is more sound than *ad hoc* quantitative approaches, and motivates the possibility of scaling up.

Natural language processing is at a crossroads. After three decades of research, the various subfields still often eschew or fail to exploit the powerful analytical and modelling tools of one paradigm and the painstaking empirical data of another.

In the proposed model of parsing and semantic interpretation I have applied the tools of probability theory to a semantics model derived from insights of linguists, psychologists, neurologists, philosophers, and computer scientists. Much remains to be done in synthesizing a cognitive ontology adequate to meet more precise and detailed semantic, conceptual, and dynamic temporal representational demands. However, I hope to have argued persuasively that complex problems in natūral language understanding systems can benefit from adopting the representational organization suggested by converging empirical evidence. Much remains to be done in improving the tractability, incremental learning, and generalization abilities of probabilistic agents. However, I hope to have demonstrated that probability and utility theory are powerful tools in mentalist approaches to overcoming problems in semantics.

Appendix A

Appendix A

Trace Output

The following traces were obtained from FRIEZE, a prototype implementation of the approximate maximum-entropy estimation method. FRIEZE is implemented in C under UNIX. Traces for all examples in chapter 7 are included here.

A.1 Trace for Figure 7.10

```
SIMPLE EVENTS.
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1

CONSTRAINTS.
1e-4    X_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
          }
1e-6    X_coastal_road {
          road_in_coastal_loc
          road_along_coastline
          Highway_1
          }

FIRST WITHOUT SUBSUMPTION.
Best hypothesis: road_leading_to_coastal_loc
Probabilities:
[  0]           9.99899e-01  NULL_EVENT
[  1] 9.9995e-07  4.99975e-11  road_in_coastal_loc
[  2] 1         4.99999e-05  road_leading_to_coastal_loc
[  3] 9.9995e-07  4.99975e-11  road_along_coastline
[  4] 9.9995e-07  4.99975e-11  Highway_1
[  5]           4.99999e-05  CT: X_road
[  6]           9.99850e-07  CT: X_coastal_road

SUBSUMING X_coastal_road BY X_road.
Best hypothesis: road_leading_to_coastal_loc
Probabilities:
[  0]           9.99900e-01  NULL_EVENT
[  1] 0.0049751   2.50000e-07  road_in_coastal_loc
[  2] 0.98507     4.95000e-05  road_leading_to_coastal_loc
```

```
[  3] 0.0049751   2.50000e-07  road_along_coastline
[  4] 0.0049751   2.50000e-07  Highway_1
[  5]            4.95000e-05  CT: X_road
[  6]            2.50000e-07  CT: X_coastal_road

CONSTRAINTS.
1e-4    X_road {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           }
1.3e-6  X_coast_and_road {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           }
1e-6    X_coastal_road {
           road_in_coastal_loc
           road_along_coastline
           Highway_1
           }

SUBSUMING X_coast_and_road BY X_road.
SUBSUMING X_coastal_road BY X_coast_and_road.

NOW WITH X_coast_and_road AS INTERMEDIATE CLASS.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]            9.99900e-01  NULL_EVENT
[  1] 0.27778     2.50000e-07  road_in_coastal_loc
[  2] 0.16667     1.50000e-07  road_leading_to_coastal_loc
[  3] 0.27778     2.50000e-07  road_along_coastline
[  4] 0.27778     2.50000e-07  Highway_1
[  5]            9.87000e-05  CT: X_road
[  6]            1.50000e-07  CT: X_coast_and_road
[  7]            2.50000e-07  CT: X_coastal_road

Frieze ready
(frieze)
```

A.2 Trace for Figure 7.11

```
SIMPLE EVENTS.
       -   road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           coasting_road

CONSTRAINTS.
1e-4    X_road {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           }
1.3e-6  X_coast_and_road {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
```

```
          Highway_1
          }
1e-6    X_coastal_road {
          road_in_coastal_loc
          road_along_coastline
          Highway_1
          }

SUBSUMING X_coast_and_road BY X_road.
SUBSUMING X_coastal_road BY X_coast_and_road.
FIRST WITHOUT ADEQUATE DIRECT ANCESTRAL CONSTRAINTS.
Best hypothesis: coasting_road
Probabilities:
[  0]           4.99950e-01  NULL_EVENT
[  1] 5.0005e-07  2.50000e-07  road_in_coastal_loc
[  2] 3.0003e-07  1.50000e-07  road_leading_to_coastal_loc
[  3] 5.0005e-07  2.50000e-07  road_along_coastline
[  4] 5.0005e-07  2.50000e-07  Highway_1
[  5] 1           4.99950e-01  coasting_road
[  6]             9.87000e-05  CT: X_road
[  7]             1.50000e-07  CT: X_coast_and_road
[  8]             2.50000e-07  CT: X_coastal_road

CONSTRAINTS.
1e-4    X_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
          }
1.3e-6  X_coast_and_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
          }
1e-6    X_coastal_road {
          road_in_coastal_loc
          road_along_coastline
          Highway_1
          }
1e-8    X_coasting_acc {
          coasting_road
          }

SUBSUMING X_coast_and_road BY X_road.
SUBSUMING X_coastal_road BY X_coast_and_road.

NOW WITH ADDITIONAL DIRECT ANCESTRAL CONSTRAINT.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]           9.99900e-01  NULL_EVENT
[  1] 0.27624    2.50000e-07  road_in_coastal_loc
[  2] 0.16575    1.50000e-07  road_leading_to_coastal_loc
[  3] 0.27624    2.50000e-07  road_along_coastline
[  4] 0.27624    2.50000e-07  Highway_1
[  5] 0.0055249  5.00000e-09  coasting_road
[  6]            9.87000e-05  CT: X_road
[  7]            1.50000e-07  CT: X_coast_and_road
[  8]            2.50000e-07  CT: X_coastal_road
[  9]            5.00000e-09  CT: X_coasting_acc

Frieze ready
```

```
(frieze)
```

A.3 Trace for Figure 7.12

```
SIMPLE EVENTS.
          road_in_coastal_loc
          road_in_coasting_event

CONSTRAINTS.
1e-2    X_container_type_containment {
          road_in_coastal_loc
          road_in_coasting_event
          }
5e-3    X_location {
          road_in_coastal_loc
          road_in_coasting_event
          }
1e-7    X_seacoast {
          road_in_coastal_loc
          }
1e-3    X_locative_containment {
          road_in_coastal_loc
          }

SUBSUMING X_locative_containment BY X_container_type_containment.
SUBSUMING X_locative_containment BY X_location.
SUBSUMING X_seacoast BY X_location.

FIRST WITHOUT DIRECT ANCESTRAL CONSTRAINTS.
Best hypothesis: road_in_coasting_event
Probabilities:
[  0]           9.86036e-01  NULL_EVENT
[  1] 0.00055873  2.01452e-08  road_in_coastal_loc
[  2] 0.99944     3.60353e-05  road_in_coasting_event
[  3]           8.96396e-03  CT: X_container_type_containment
[  4]           3.96388e-03  CT: X_location
[  5]           7.98548e-08  CT: X_seacoast
[  6]           9.99980e-04  CT: X_locative_containment

CONSTRAINTS.
1e-2    X_container_type_containment {
          road_in_coastal_loc
          road_in_coasting_event
          }
5e-3    X_location {
          road_in_coastal_loc
          road_in_coasting_event
          }
1e-7    X_seacoast {
          road_in_coastal_loc
          }
1e-3    X_locative_containment {
          road_in_coastal_loc
          }
1e-3    X_eventive_containment {
          road_in_coasting_event
          }

SUBSUMING X_locative_containment BY X_container_type_containment.
SUBSUMING X_locative_containment BY X_location.
SUBSUMING X_seacoast BY X_location.
```

```
NOW WITH DIRECT ANCESTRAL CONSTRAINT, STILL INADEQUATE.
Best hypothesis: road_in_coasting_event
Probabilities:
[  0]           9.85000e-01  NULL_EVENT
[  1] 0.35025    2.00001e-08  road_in_coastal_loc
[  2] 0.64975    3.71022e-08  road_in_coasting_event
[  3]           8.99996e-03  CT: X_container_type_containment
[  4]           3.99988e-03  CT: X_location
[  5]           7.99999e-08  CT: X_seacoast
[  6]           9.99980e-04  CT: X_locative_containment
[  7]           9.99963e-04  CT: X_eventive_containment

CONSTRAINTS.
1e-2    X_container_type_containment {
           road_in_coastal_loc
           road_in_coasting_event
           }
5e-3    X_location {
           road_in_coastal_loc
           road_in_coasting_event
           }
1e-7    X_seacoast {
           road_in_coastal_loc
           }
1e-3    X_locative_containment {
           road_in_coastal_loc
           }
1e-3    X_eventive_containment {
           road_in_coasting_event
           }
7e-3    X_other_container_type_containment {
           }

SUBSUMING X_other_container_type_containment BY X_container_type_containment.
SUBSUMING X_locative_containment BY X_container_type_containment.
SUBSUMING X_locative_containment BY X_location.
SUBSUMING X_seacoast BY X_location.

NOW WITH ADDITIONAL OR-SIBLING CONSTRAINT.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]           9.85000e-01  NULL_EVENT
[  1] 0.70808    2.00000e-08  road_in_coastal_loc
[  2] 0.29192    8.24523e-09  road_in_coasting_event
[  3]           1.99999e-03  CT: X_container_type_containment
[  4]           3.99991e-03  CT: X_location
[  5]           8.00000e-08  CT: X_seacoast
[  6]   -       9.99980e-04  CT: X_locative_containment
[  7]           9.99992e-04  CT: X_eventive_containment
[  8]           7.00000e-03  CT: X_other_container_type_containment

CONSTRAINTS.
1e-2     X_container_type_containment {
            road_in_coastal_loc
            road_in_coasting_event
            }
5e-3     X_location {
            road_in_coastal_loc
            road_in_coasting_event
            }
1e-7     X_seacoast {
            road_in_coastal_loc
            }
```

```
1e-3      X_locative_containment {
           road_in_coastal_loc
           }
8.999e-3  X_other_container_type_containment {
           }

SUBSUMING X_other_container_type_containment BY X_container_type_containment.
SUBSUMING X_locative_containment BY X_container_type_containment.
SUBSUMING X_locative_containment BY X_location.
SUBSUMING X_seacoast BY X_location.

WITH OR-SIBLING, AND EVEN WITHOUT DIRECT ANCESTRAL CONSTRAINT.
CONVERGENCE IS VERY TIME CONSUMING, BUT THEORETICALLY CONSISTENT.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]            9.86000e-01  NULL_EVENT
[  1] 0.83194     2.00000e-08  road_in_coastal_loc
[  2] 0.16806     4.04032e-09  road_in_coasting_event
[  3]            9.95960e-07  CT: X_container_type_containment
[  4]            3.99992e-03  CT: X_location
[  5]            8.00000e-08  CT: X_seacoast
[  6]            9.99980e-04  CT: X_locative_containment
[  7]            8.99900e-03  CT: X_other_container_type_containment

Frieze ready
(frieze)
```

A.4 Trace for Figure 7.13

```
SIMPLE EVENTS.
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline

CONSTRAINTS.
1e-4     X_road {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
         }
1e-6     X_seacoast {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
1e-6     X_coastal_road {
           road_in_coastal_loc
           road_along_coastline
           Highway_1
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
5e-8     X_Highway_1 {
           Highway_1
```

```
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
5e-7    X_coast_and_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
4e-7    X_coast_and_coastal_road {
          road_in_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }

SUBSUMING X_coastal_road BY X_road.
SUBSUMING X_Highway_1 BY X_coastal_road.
SUBSUMING X_coast_and_road BY X_road.
SUBSUMING X_coast_and_road BY X_seacoast.
SUBSUMING X_coast_and_coastal_road BY X_coastal_road.
SUBSUMING X_coast_and_coastal_road BY X_coast_and_road.

WITH X_coastal_road, X_Highway_1 AND X_coast_and_road AS
INTERMEDIATE CLASSES.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]           9.99900e-01  NULL_EVENT
[  1] 0.3828     1.30187e-07  road_in_coastal_loc
[  2] 0.14702    5.00000e-08  road_leading_to_coastal_loc
[  3] 0.3828     1.30187e-07  road_along_coastline
[  4] 0.059632   2.02805e-08  Highway_1
[  5] 0.013877   4.71948e-09  Highway_1_in_Pacific_coastal_loc
[  6] 0.013877   4.71948e-09  Highway_1_along_Pacific_coastline
[  7]           9.89000e-05  CT: X_road
[  8]           5.00000e-07  CT: X_seacoast
[  9]           5.59439e-07  CT: X_coastal_road
[ 10]           2.02805e-08  CT: X_Highway_1
[ 11]           5.00000e-08  CT: X_coast_and_road
[ 12]           1.30187e-07  CT: X_coast_and_coastal_road


NOW ASSUMING X_Highway_1 IS FREQUENTLY USED CONCEPT CLASS.
Best hypothesis: Highway_1
Probabilities:
[  0]           9.99900e-01  NULL_EVENT
[  1] 0.16183    8.80367e-08  road_in_coastal_loc
[  2] 0.091909   5.00000e-08  road_leading_to_coastal_loc
[  3] 0.16183    8.80367e-08  road_along_coastline
[  4] 0.33465    1.82055e-07  Highway_1
[  5] 0.12489    6.79449e-08  Highway_1_in_Pacific_coastal_loc
[  6] 0.12489    6.79449e-08  Highway_1_along_Pacific_coastline
[  7]           9.89000e-05  CT: X_road
[  8]           5.00000e-07  CT: X_seacoast
[  9]           2.35890e-07  CT: X_coastal_road
[ 10]           1.82055e-07  CT: X_Highway_1
[ 11]           5.00000e-08  CT: X_coast_and_road
[ 12]           8.80367e-08  CT: X_coast_and_coastal_road

Frieze ready
(frieze)
```

A.5 Trace for Figure 7.14

```
SIMPLE EVENTS.
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline

CONSTRAINTS.
1e-1    X_ischema_state {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-2    X_containment {
          road_in_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-3    X_linear_order_locative {
          road_leading_to_coastal_loc
          }
1e-4    X_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-6    X_seacoast {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-6    X_coastal_road {
          road_in_coastal_loc
          road_along_coastline
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
5e-9    X_Highway_1 {
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
5e-7    X_coast_and_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
4e-7    X_coast_and_coastal_road {
          road_in_coastal_loc
```

```
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          )

SUBSUMING X_containment BY X_ischema_state.
SUBSUMING X_linear_order_locative BY X_ischema_state.
SUBSUMING X_coastal_road BY X_road.
SUBSUMING X_Highway_1 BY X_coastal_road.
SUBSUMING X_coast_and_road BY X_road.
SUBSUMING X_coast_and_road BY X_seacoast.
SUBSUMING X_coast_and_coastal_road BY X_coastal_road.
SUBSUMING X_coast_and_coastal_road BY X_coast_and_road.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]           8.99900e-01  NULL_EVENT
[  1] 0.38392    4.34790e-09  road_in_coastal_loc
[  2] 0.0098014  1.11000e-10  road_leading_to_coastal_loc
[  3] 0.38392    4.34790e-09  road_along_coastline
[  4] 0.21915    2.48187e-09  Highway_1
[  5] 0.0016013  1.81349e-11  Highway_1_in_Pacific_coastal_loc
[  6] 0.0016013  1.81349e-11  Highway_1_along_Pacific_coastline
[  7]           8.90000e-02  CT: X_ischema_state
[  8]           9.99999e-03  CT: X_containment
[  9]           1.00000e-03  CT: X_linear_order_locative
[ 10]           9.89000e-05  CT: X_road
[ 11]           5.00000e-07  CT: X_seacoast
[ 12]           5.95036e-07  CT: X_coastal_road
[ 13]           2.48187e-09  CT: X_Highway_1
[ 14]           9.98890e-08  CT: X_coast_and_road
[ 15]           3.91268e-07  CT: X_coast_and_coastal_road

NOW WITH P(X_containment) = 1e-3 AND P(X_linear_order_locative) = 1e-2.
Best hypothesis: Highway_1
Probabilities:
[  0]           8.99900e-01  NULL_EVENT
[  1] 0.098822   4.43504e-10  road_in_coastal_loc
[  2] 0.24489    1.09902e-09  road_leading_to_coastal_loc
[  3] 0.098822   4.43504e-10  road_along_coastline
[  4] 0.55664    2.49814e-09  Highway_1
[  5] 0.00041491 1.86206e-12  Highway_1_in_Pacific_coastal_loc
[  6] 0.00041491 1.86206e-12  Highway_1_along_Pacific_coastline
[  7]           8.90000e-02  CT: X_ischema_state
[  8]           9.99999e-04  CT: X_containment
[  9]           1.00000e-02  CT: X_linear_order_locative
[ 10]           9.89000e-05  CT: X_road
[ 11]           5.00000e-07  CT: X_seacoast
[ 12]           5.95004e-07  CT: X_coastal_road
[ 13]           2.49814e-09  CT: X_Highway_1
[ 14]           9.89010e-08  CT: X_coast_and_road
[ 15]           3.99109e-07  CT: X_coast_and_coastal_road

Frieze ready
(frieze)
```

A.6 Trace for Figure 7.15

```
SIMPLE EVENTS.
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
```

```
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline

CONSTRAINTS.
1e-1    X_ischema_state {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-2    X_containment {
          road_in_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-3    X_linear_order_locative {
          road_leading_to_coastal_loc
          }
1e-4    X_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-6    X_seacoast {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-6    X_coastal_road {
          road_in_coastal_loc
          road_along_coastline
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
5e-13   X_Highway_1 {
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
5e-7    X_coast_and_road {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
4e-7    X_coast_and_coastal_road {
          road_in_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
5e-4    C_NN__ischema_state {
          road_in_coastal_loc
          road_leading_to_coastal_loc
```

```
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-4   C_NN__containment {
          road_in_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
5e-5   C_NN__linear_order_locative {
          road_leading_to_coastal_loc
          }

SUBSUMING X_containment BY X_ischema_state.
SUBSUMING X_linear_order_locative BY X_ischema_state.
SUBSUMING X_coastal_road BY X_road.
SUBSUMING X_Highway_1 BY X_coastal_road.
SUBSUMING X_coast_and_road BY X_road.
SUBSUMING X_coast_and_road BY X_seacoast.
SUBSUMING X_coast_and_coastal_road BY X_coastal_road.
SUBSUMING X_coast_and_coastal_road BY X_coast_and_road.
SUBSUMING C_NN__containment BY X_containment.
SUBSUMING C_NN__containment BY C_NN__ischema_state.
SUBSUMING C_NN__linear_order_locative BY X_linear_order_locative.
SUBSUMING C_NN__linear_order_locative BY C_NN__ischema_state.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]           8.99550e-01  NULL_EVENT
[  1] 0.46934    4.44568e-11  road_in_coastal_loc
[  2] 0.058678   5.55803e-12  road_leading_to_coastal_loc
[  3] 0.46934    4.44568e-11  road_along_coastline
[  4] 0.0026391  2.49981e-13  Highway_1
[  5] 1.9554e-07 1.85223e-17  Highway_1_in_Pacific_coastal_loc
[  6] 1.9554e-07 1.85223e-17  Highway_1_along_Pacific_coastline
[  7]           8.90000e-02  CT: X_ischema_state
[  8]           9.90000e-03  CT: X_containment
[  9]           9.50000e-04  CT: X_linear_order_locative
[ 10]           9.89000e-05  CT: X_road
[ 11]           5.00000e-07  CT: X_seacoast
[ 12]           5.99999e-07  CT: X_coastal_road
[ 13]           2.49981e-13  CT: X_Highway_1
[ 14]           9.99944e-08  CT: X_coast_and_road
[ 15]           3.99911e-07  CT: X_coast_and_coastal_road
[ 16]           3.50000e-04  CT: C_NN__ischema_state
[ 17]           9.99999e-05  CT: C_NN__containment
[ 18]           5.00000e-05  CT: C_NN__linear_order_locative

NOW WITH P(C_NN__containment) RESET DOWN TO 5e-5.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]           8.99500e-01  NULL_EVENT
[  1] 0.44223    2.22321e-11  road_in_coastal_loc
[  2] 0.11056    5.55834e-12  road_leading_to_coastal_loc
[  3] 0.44223    2.22321e-11  road_along_coastline
[  4] 0.0049727  2.49991e-13  Highway_1
[  5] 1.8426e-07 9.26304e-18  Highway_1_in_Pacific_coastal_loc
[  6] 1.8426e-07 9.26304e-18  Highway_1_along_Pacific_coastline
[  7]           8.90000e-02  CT: X_ischema_state
[  8]           9.95000e-03  CT: X_containment
[  9]           9.50000e-04  CT: X_linear_order_locative
[ 10]           9.89000e-05  CT: X_road
[ 11]           5.00000e-07  CT: X_seacoast
```

```
[ 12]           5.99999e-07  CT: X_coastal_road
[ 13]           2.49991e-13  CT: X_Highway_1
[ 14]           9.99944e-08  CT: X_coast_and_road
[ 15]           3.99956e-07  CT: X_coast_and_coastal_road
[ 16]           4.00000e-04  CT: C_NN__ischema_state
[ 17]           5.00000e-05  CT: C_NN__containment
[ 18]           5.00000e-05  CT: C_NN__linear_order_locative

NOW WITH P(C_NN__containment) RESET DOWN TO 1e-5.
Best hypothesis: road_leading_to_coastal_loc
Probabilities:
[  0]           8.99459e-01  NULL_EVENT
[  1] 0.30246    4.44701e-12  road_in_coastal_loc
[  2] 0.37807    5.55858e-12  road_leading_to_coastal_loc
[  3] 0.30246    4.44701e-12  road_along_coastline
[  4] 0.017004   2.49998e-13  Highway_1
[  5] 1.2603e-07  1.85291e-18  Highway_1_in_Pacific_coastal_loc
[  6] 1.2603e-07  1.85291e-18  Highway_1_along_Pacific_coastline
[  7]           8.90000e-02  CT: X_ischema_state
[  8]           9.99000e-03  CT: X_containment
[  9]           9.50000e-04  CT: X_linear_order_locative
[ 10]           9.89000e-05  CT: X_road
[ 11]           5.00000e-07  CT: X_seacoast
[ 12]           5.99999e-07  CT: X_coastal_road
[ 13]           2.49998e-13  CT: X_Highway_1
[ 14]           9.99944e-08  CT: X_coast_and_road
[ 15]           3.99991e-07  CT: X_coast_and_coastal_road
[ 16]           4.40000e-04  CT: C_NN__ischema_state
[ 17]           9.99999e-06  CT: C_NN__containment
[ 18]           5.00000e-05  CT: C_NN__linear_order_locative

Frieze ready
(frieze)
```

A.7 Trace for Figure 7.16

```
SIMPLE EVENTS.
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          coasting_road

CONSTRAINTS.
1e-1    X_ischema_state {
          road_in_coastal_loc
          road_leading_to_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-2    X_containment {
          road_in_coastal_loc
          road_along_coastline
          Highway_1_in_Pacific_coastal_loc
          Highway_1_along_Pacific_coastline
          }
1e-3    X_linear_order_locative {
          road_leading_to_coastal_loc
          }
```

```
5e-7    X_coast_acc {
           coasting_road
           }
1e-6    X_seacoast {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
1e-4    X_road {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           coasting_road
           }
1e-6    X_coastal_road {
           road_in_coastal_loc
           road_along_coastline
           Highway_1
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
5e-13   X_Highway_1 {
           Highway_1
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
5e-7    X_coast_and_road {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
4e-7    X_coast_and_coastal_road {
           road_in_coastal_loc
           road_along_coastline
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
4e-7    L_coast {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           coasting_road
           }
1.9e-7  C_coast__seacoast {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
1.5e-7  C_coast__coast_acc {
           coasting_road
           }
```

```
5e-4    C_NN__ischema_state {
           road_in_coastal_loc
           road_leading_to_coastal_loc
           road_along_coastline
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
1e-4    C_NN__containment {
           road_in_coastal_loc
           road_along_coastline
           Highway_1_in_Pacific_coastal_loc
           Highway_1_along_Pacific_coastline
           }
5e-5    C_NN__linear_order_locative {
           road_leading_to_coastal_loc
           }

SUBSUMING X_containment BY X_ischema_state.
SUBSUMING X_linear_order_locative BY X_ischema_state.
SUBSUMING X_coastal_road BY X_road.
SUBSUMING X_Highway_1 BY X_coastal_road.
SUBSUMING X_coast_and_road BY X_road.
SUBSUMING X_coast_and_road BY X_seacoast.
SUBSUMING X_coast_and_coastal_road BY X_coastal_road.
SUBSUMING X_coast_and_coastal_road BY X_coast_and_road.
SUBSUMING C_coast__seacoast BY L_coast.
SUBSUMING C_coast__seacoast BY X_seacoast.
SUBSUMING C_coast__coast_acc BY L_coast.
SUBSUMING C_coast__coast_acc BY X_coast_acc.
SUBSUMING C_NN__containment BY X_containment.
SUBSUMING C_NN__containment BY C_NN__ischema_state.
SUBSUMING C_NN__linear_order_locative BY X_linear_order_locative.
SUBSUMING C_NN__linear_order_locative BY C_NN__ischema_state.
Best hypothesis: road_in_coastal_loc
Probabilities:
[  0]           8.99549e-01  NULL_EVENT
[  1] 0.36624    2.72367e-11  road_in_coastal_loc
[  2] 0.045785   3.40494e-12  road_leading_to_coastal_loc
[  3] 0.36624    2.72367e-11  road_along_coastline
[  4] 4.484e-10  3.33470e-20  Highway_1
[  5] 3.0517e-07 2.26952e-17  Highway_1_in_Pacific_coastal_loc
[  6] 3.0517e-07 2.26952e-17  Highway_1_along_Pacific_coastline
[  7] 0.22173    1.64898e-11  coasting_road
[  8]           8.90000e-02  CT: X_ischema_state
[  9]           9.90000e-03  CT: X_containment
[ 10]           9.50000e-04  CT: X_linear_order_locative
[ 11]           3.50000e-07  CT: X_coast_acc
[ 12]           3.10058e-07  CT: X_seacoast
[ 13]           9.89000e-05  CT: X_road
[ 14]           5.99999e-07  CT: X_coastal_road
[ 15]           4.99955e-13  CT: X_Highway_1
[ 16]           9.99966e-08  CT: X_coast_and_road
[ 17]           3.99946e-07  CT: X_coast_and_coastal_road
[ 18]           6.00000e-08  CT: L_coast
[ 19]           1.89942e-07  CT: C_coast__seacoast
[ 20]           1.49984e-07  CT: C_coast__coast_acc
[ 21]           3.50000e-04  CT: C_NN__ischema_state
[ 22]           9.99999e-05  CT: C_NN__containment
[ 23]           5.00000e-05  CT: C_NN__linear_order_locative

NOW WITH P(C_coast__seacoast) RESET TO 1.5e-7
AND P(C_coast__coast_acc) RESET TO 1.9e-7.
Best hypothesis: coasting_road
```

```
Probabilities:
[  0]            8.99549e-01  NULL_EVENT
[  1] 0.31041     1.90480e-11  road_in_coastal_loc
[  2] 0.038804    2.38117e-12  road_leading_to_coastal_loc
[  3] 0.31041     1.90480e-11  road_along_coastline
[  4] 5.4344e-10  3.33479e-20  Highway_1
[  5] 2.5866e-07  1.58723e-17  Highway_1_in_Pacific_coastal_loc
[  6] 2.5866e-07  1.58723e-17  Highway_1_along_Pacific_coastline
[  7] 0.34038     2.08871e-11  coasting_road
[  8]            8.90000e-02  CT: X_ischema_state
[  9]            9.90000e-03  CT: X_containment
[ 10]            9.50000e-04  CT: X_linear_order_locative
[ 11]            3.10000e-07  CT: X_coast_acc
[ 12]            3.50040e-07  CT: X_seacoast
[ 13]            9.89000e-05  CT: X_road
[ 14]            5.99999e-07  CT: X_coastal_road
[ 15]            4.99968e-13  CT: X_Highway_1
[ 16]            9.99976e-08  CT: X_coast_and_road
[ 17]            3.99962e-07  CT: X_coast_and_coastal_road
[ 18]            6.00000e-08  CT: L_coast
[ 19]            1.49960e-07  CT: C_coast__seacoast
[ 20]            1.89979e-07  CT: C_coast__coast_acc
[ 21]            3.50000e-04  CT: C_NN__ischema_state
[ 22]            1.00000e-04  CT: C_NN__containment
[ 23]            5.00000e-05  CT: C_NN__linear_order_locative

Frieze ready
(frieze)
```

References

AGRE, PHIL & DAVID CHAPMAN. 1987. Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, WA. Morgan Kaufmann.

AITCHISON, J. & C. G. G. AITKEN. 1976. Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420.

ALLEN, JAMES. 1987. *Natural language understanding*. Menlo Park, CA: Benjamin/Cummings.

ALSHAWI, HIYAN. 1987. *Memory and context for language interpretation*. Cambridge: Cambridge University Press.

ANANTHANARAYANA, H. S. 1970. The kāraka theory and Case grammar. *Indian Linguistics*, 31:14–27. Cited in Somers 1987.

ANDERSON, J. R. 1978. Arguments concerning representations for mental imagery. *Psychological Review*, 85:249–277.

ANDERSON, JOHN R. & GORDON H. BOWER. 1973. *Human associative memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

ANDERSON, JOHN R. & GORDON H. BOWER. 1980. *Human associative memory: A brief edition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

ANDERSON, R. C. & A. ORTONY. 1975. On putting apples into bottles–a problem of polysemy. *Cognitive Psychology*, 7:167–180.

APPELT, DOUGLAS E. & MARTHA E. POLLACK. 1990. Weighted abduction as an inference method for plan recognition and evaluation. In *Proceedings of the Second International Workshop on User Modeling*, Honolulu, HI.

BACCHUS, FAHIEM. 1990. *Representing and reasoning with probabilistic knowledge: A logical approach to probabilities*. Cambridge, MA: MIT Press.

BACH, EMMON. 1983. On time, tense, and aspect: An essay in English metaphysics. In *Meaning, use, and interpretation*, ed. by R. Bauerle, C. Schwarze, & A. von Stechnow. New York: de Gruyter.

BACH, EMMON. 1986. The algebra of events. *Linguistics and Philosophy*, 9:5–16.

BALOTA, DAVID A., GIOVANNI B. FLORES D'ARCAIS, & KEITH RAYNER (eds.). 1990. *Comprehension processes in reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

BARSALOU, LAWRENCE W. 1983. Ad hoc categories. *Memory & Cognition*, 11:211–227.

BARSALOU, LAWRENCE W. 1985. Ideals, central tendency, and frequency of instantiation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11:629–654.

BARSALOU, LAWRENCE W. 1987. The instability of graded structure: Implications for the nature of concepts. In (Neisser 1987), 101–140.

BARWISE, JON & JOHN PERRY (eds.). 1983. *Situations and attitudes*. Cambridge, MA: MIT Press.

BAUM, L. E. 1970. A maximization technique occuring in the statistical analysis of probabilistic functions in Markov chains. *The Annals of Mathematical Statistics*, 41(1).

BEAUVOIS, M. F. 1982. Optic aphasia: A process of interaction between vision and language. *Philosophical Transactions of the Royal Society*, B 298:35–47.

BERGER, JAMES O. 1985. *Statistical decision theory and Bayesian analysis*. New York: Springer-Verlag. Second edition of *Statistical Decision Theory: Foundations, Concepts, and Methods*.

BIERWISCH, MANFRED & EWALD LANG (eds.). 1989. *Dimensional adjectives: Grammatical structure and conceptual interpretation*. Berlin: Springer-Verlag.

BOWERMAN, MELISSA, 1974. Learning the structure of causative verbs: A study in the relationship of cognitive, semantic, and syntactic development. Papers and Reports on Child Language Development, 8. Stanford, CA: Stanford University Department of Linguistics. Cited in Pinker 1989.

BOWERMAN, MELISSA. 1982. Evaluating competing linguistic models with language acquisition data: Implications of developmental error with causative verbs. *Quaderni di Semantica*, 3:5–66.

BRACHMAN, RONALD J. 1979. On the epistemological status of semantic networks. In *Associative networks: Representation and use of knowledge by computers*, ed. by Nicholas V. Findler, 3–50. New York: Academic Press.

BRACHMAN, RONALD J. 1983. What IS-A is and isn't: An analysis of taxonomic links in semantic networks. *IEEE Computer*, 16(10):30–36.

BRACHMAN, RONALD J. 1985. I lied about the trees: Or, defaults and definitions in knowledge representation. *Artificial Intelligence Magazine*, 80–93.

BRACHMAN, RONALD J., RICHARD E. FIKES, & HECTOR J. LEVESQUE. 1983. KRYPTON: A functional approach to knowledge representation. *IEEE Computer*, 16(1):67–73.

BRACHMAN, RONALD J., DEBORAH L. MCGUINNESS, PETER F. PATEL-SCHNEIDER, & LORI ALPERIN RESNICK. 1991. Living with CLASSIC: When and how to use a KL-ONE-like language. In (Sowa 1991), 401–456.

BRACHMAN, RONALD J. & JAMES G. SCHMOLZE. 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216.

BRAVERMAN, MICHAEL, 1992. (ph.d. dissertation). Forthcoming.

BRAVERMAN, MICHAEL S. & STUART J. RUSSELL. 1988. IMEX: Overcoming intractability in explanation based learning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, 575–579.

BREESE, J. S. & M. R. FEHLING. 1990. Control of problem-solving: Principles and architecture. In *Uncertainty in artificial intelligence 4*, ed. by R. D. Shachter, T. Levitt, L. Kanal, & J. Lemmer. Amsterdam: North-Holland.

BRESNAN, JOAN (ed.). 1982. *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

BRILL, ERIC, DAVID MAGERMAN, MITCHELL MARCUS, & BEATRICE SANTORINI. 1990. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 275–281.

BROOKS, L. R. 1968. Spatial and verbal components of the act of recall. *Canadian Journal of Psychology*, 22:349–368.

BUNTINE, WRAY L. & ANDREAS S. WEIGEND, 1991a. Bayesian back-propagation. To appear in *Complex Systems*.

BUNTINE, WRAY L. & ANDREAS S. WEIGEND, 1991b. Calculating second derivatives on feed-forward networks. Submitted for publication.

BYBEE, JOAN L. & DAN ISAAC SLOBIN. 1982. Rules and schemas in the development and use of the English past tense. *Language*, 58:265–289.

CARNAP, RUDOLF. 1952. *The continuum of inductive methods*. Chicago: University of Chicago Press.

CARNAP, RUDOLF. 1962. *The logical foundations of probability*. Chicago: University of Chicago Press.

CHARNIAK, EUGENE. 1981. The case-slot identity theory. *Cognitive Science*, 5(3):285–292.

CHARNIAK, EUGENE. 1983. Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science*, 7(3):171–190.

CHARNIAK, EUGENE, 1985. A single-semantic-process theory of parsing. Unpublished manuscript. Department of Computer Science, Brown University.

CHARNIAK, EUGENE. 1986. A neat theory of marker passing. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, 584–588.

CHARNIAK, EUGENE & ROBERT GOLDMAN. 1988. A logic for semantic interpretation. In *Proceedings of the 26th Annual Conference of the Association for Computational Linguistics*.

CHARNIAK, EUGENE & ROBERT GOLDMAN. 1989. A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1074–1079.

CHARNIAK, EUGENE & DREW MCDERMOTT. 1985. *Introduction to artificial intelligence*. Reading, MA: Addison-Wesley.

CHARNIAK, EUGENE & SOLOMON E. SHIMONY. 1990. Probabilistic semantics for cost based abduction. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, 106–111, Boston.

CHEESEMAN, PETER. 1987. A method of computing maximum entropy probability values for expert systems. In *Maximum-entropy and Bayesian spectral analysis and estimation problems*, ed. by Ray C. Smith & Gary J. Erickson, 229–240. Dordrecht, Holland: D. Reidel. Revised proceedings of the Third Maximum Entropy Workshop, Laramie, WY, 1983.

CHIERCHIA, GENNARO & SALLY MCCONNELL-GINET. 1990. *Meaning and grammar: An introduction to semantics*. Cambridge, MA: MIT Press.

CHIN, DAVID NGI. 1988. Intelligent agents as a basis for natural language interfaces. Technical Report UCB/CSD 88/396, Univ. of Calif. at Berkeley, Comp. Sci. Div., Berkeley.

CHODOROW, MARTIN S., ROY J. BYRD, & GEORGE E. HEIDORN. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics*, 299–304, Chicago.

CHOMSKY, NOAM. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

CHURCH, KENNETH WARD & PATRICK HANKS. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Conference of the Association for Computational Linguistics*, 76–83, Vancouver.

CHURCH, KENNETH WARD & PATRICK HANKS. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

COHEN, L. JONATHAN. 1989. *An introduction to the philosophy of induction and probability*. Oxford: Oxford University Press.

COOK, W. A. 1972. A set of postulates for case grammar analysis. In *Languages and linguistics working papers* 4. Georgetown University. Reprinted in Cook 1979.

COSLETT, H. B. & E. M. SAFFRAN. 1989. Preserved object recognition and reading comprehension in optic aphasia. *Brain*, 112:1091–1110.

COTTRELL, GARRISON W. 1984. A model of lexical access of ambiguous words. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, 61–67.

COTTRELL, GARRISON WEEKS. 1985. A connectionist approach to word sense disambiguation. Technical Report TR 154, Univ. of Rochester, Dept. of Comp. Sci., New York.

CRUSE, D. A. (ed.). 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

DAVIS, LAWRENCE M. 1990. *Statistics in dialectology*. Tuscaloosa, AL: University of Alabama Press.

DE SAUSSURE, FERDINAND. 1966. *Course in general linguistics*. New York: McGraw-Hill. Translated by Wade Baskin from the French edition of 1915. Reprinted from the 1959 edition (New York: The Philosophical Library).

DIAS, PENHA MARIA CARDOSO & ABNER SHIMONY. 1981. A critique of Jaynes' maximum entropy principle. *Advances in Applied Mathematics*, 2:172–211.

DIK, SIMON C. 1978. *Functional grammar*. Amsterdam: North-Holland.

DOWNING, PAMELA. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.

DOWTY, DAVID. 1979. *Word meaning and Montague grammar*. Boston, MA: D. Reidel.

DOWTY, DAVID. 1989. On the semantic content of the notion of 'thematic role'. In *Properties, types, and meaning, vol. 2: Semantic issues*, ed. by Gennaro Chierchia, Barbara H. Partee, & Raymond Turner, volume 39 of *Studies in Linguistics and Philosophy*. Dordrecht: Kluwer.

DUFFY, S. A. 1986. Role of expectations in sentence integration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12:208–219.

EEGS-OLOFSSON, MATS. 1990. An automatic word class tagger and a phrase parser. In (Svartvik 1990), 107–136.

ELMAN, JEFFREY L. 1989. Representation and structure in connectionist models. Technical Report CRL-8903, University of California at San Diego, Center for Research in Language, San Diego, CA.

ELMAN, JEFFREY L. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

ELMAN, JEFFREY L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

FAHLMAN, SCOTT E. 1979. *Netl: A system for representing and using real-world knowledge.* Cambridge, MA: MIT Press.

FANO, R. 1961. *Transmission of information: A statistical theory of communications.* Cambridge, MA: MIT Press.

FARAH, MARTHA J. 1984. The neurological basis of mental imagery: A componential analysis. *Cognition*, 18:245–272.

FARAH, MARTHA J. 1988. The neuropsychology of mental imagery: Converging evidence from brain-damaged and normal subjects. In *Spatial cognition: Brain bases and development*, ed. by Joan Stiles-Davis, Mark Kritchevsky, & Ursula Bellugi, 33–56. Hillsdale, NJ: Lawrence Erlbaum Associates.

FARAH, MARTHA J. 1990. *Visual agnosia.* Cambridge, MA: MIT Press.

FARAH, MARTHA J., K. M. HAMMOND, D. N. LEVINE, & R. CALVANIO. 1990. Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive Psychology.*

FASOLD, RALPH W. 1972. *Tense marking in Black English: A linguistic and social analysis.* Washington, D.C.: Center for Applied Linguistics.

FELDMAN, JEROME A., GEORGE LAKOFF, ANDREAS STOLCKE, & SUSAN HOLLBACH WEBER. 1990. Miniature language acquisition: A touchstone for cognitive science. In *Program of the Twelfth Annual Conference of the Cognitive Science Society*, 686–693. Also available as technical report TR-90-009, International Computer Science Institute, Berkeley, CA.

FELDMAN, JEROME A. & ROBERT F. SPROULL. 1977. Decision theory and artificial intelligence II: The hungry monkey. *Cognitive Science*, 1:158–192.

FILLMORE, CHARLES J. 1968. The case for case. In *Universals in linguistic theory*, ed. by Emmon W. Bach & Robert T. Harms, 1–88. New York: Holt, Rinehart, and Winston.

FILLMORE, CHARLES J. 1977. The case for case reopened. In *Grammatical relations*, ed. by P. Cole & J. M. Saddock, volume 8 of *Syntax and Semantics*, 59–81. New York: Academic Press.

FILLMORE, CHARLES J., 1988. On grammatical constructions. Unpublished draft, University of California at Berkeley.

FINCH, STEVEN & NICK CHATER, 1991. A hybrid approach to the automatic learning of linguistic categories. University of Edinburgh, (1) Centre for Cognitive Science and (2) Department of Psychology.

FINKE, RONALD A. 1989. *Principles of mental imagery*. Cambridge, MA: MIT Press.

FINKE, RONALD A., MARCIA J. JOHNSON, & G. C.-W. SHYI. 1981. Memory confusions for real and imagined completions of symmetrical visual patterns. *Memory & Cognition*, 16:133–137.

FODOR, JERRY A. 1983. *The modularity of mind*. Cambridge, MA: MIT Press.

FODOR, JERRY A. 1987. Music, frames, fridgeons, sleeping dogs, and the music of the spheres. In (Garfield 1987), 25–36.

FOLEY, WILLIAM A. & MICHAEL L. OLSON. 1985. Clausehood and verb serialization. In *Grammar inside and outside the clause*, ed. by Johanna Nichols & Anthony Woodbury. Cambridge: Cambridge University Press. Cited in Foley & van Valin 1984.

FOLEY, WILLIAM A. & ROBERT D. VAN VALIN, JR. 1984. *Functional syntax and universal grammar*. Cambridge: Cambridge University Press.

FRANCIS, W. NELSON & HENRY KUČERA. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin. With the assistance of Andrew W. Mackie.

FU, K. S. 1974. *Syntactic methods in pattern recognition*, volume 112 of *Mathematics in Science and Engineering*. New York: Academic Press.

FUJISAKI, T., F. JELINEK, J. COCKE, E. BLACK, & T. NISHINO. 1991. A probabilistic parsing method for sentence disambiguation. In *Current issues in parsing technology*, ed. by Masaru Tomita, 139–152. Boston: Kluwer.

GARFIELD, JAY L. (ed.). 1987. *Modularity in knowledge representation and natural-language understanding*. Cambridge, MA: MIT Press.

GENTNER, DEDRE. 1981. Some interesting differences between verbs and nouns. *Cognition and Brain Theory*, 4:161–178.

GENTNER, DEDRE. 1988. Cognitive and linguistic determinism: Object reference and relational inference. In *Thirteenth Annual Boston University Conference on Language Development*. Cited in Jackendoff 1990.

GOEBEL, RANDY. 1990. A quick review of hypothetical reasoning based on abduction. In *Working Notes from the Spring Symposium on Automated Abduction*, 145–149, Stanford University, Stanford, CA. AAAI.

GOLDMAN, ROBERT P. & EUGENE CHARNIAK. 1990a. Incremental construction of probabilistic models for language abduction: Work in progress. In *Working Notes from the Spring Symposium on Automated Abduction*, 1–4, Stanford University, Stanford, CA. AAAI.

GOLDMAN, ROBERT P. & EUGENE CHARNIAK. 1990b. A probabilistic approach to text understanding. Technical Report CS-90-13, Brown Univ., Providence, RI.

GOOD, I. J. 1971. Twenty-seven principles of rationality. In *Foundations of statistical inference*, ed. by V. P. Godambe & D. A. Sprott. Toronto: Holt, Rinehart, and Winston.

GOOD, I. J. 1977. Dynamic probability, computer chess and the measurement of knowledge. *Machine Intelligence*, 8.

GREENE, BARBARA B. & GERALD M. RUBIN, 1971. Automatic grammatical tagging of English. Department of Linguistics, Brown University. Cited in Francis & Kučera 1982.

GROSS, DEREK & KATHERINE J. MILLER. 1990. Adjectives in WordNet. *Journal of Lexicography*, 3(4):265–277.

GRUBER, J. S., 1965. *Studies in lexical relations*. Cambridge, MA: MIT dissertation. Reprinted by Indiana University Linguistics Club, Bloomington, IN. Also reprinted in *Lexical Structures in Syntax and Semantics*, 1976, North-Holland, Amsterdam.

GUTHRIE, LOUISE, BRIAN SLATOR, YORICK WILKS, & REBECCA BRUCE. 1990. Is there content in empty heads? In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 138–143, Helsinki.

HALLIDAY, M. A. K. 1970. Language structure and language function. In *New horizons in linguistics*, ed. by John Lyons, 140–165. Harmondsworth, Middx: Penguin.

HARMAN, GILBERT. 1965. The inference to the best explanation. *Philosophical Review*, 74:88–95.

HARNAD, STEVAN. 1990. The symbol grounding problem. *PhysicaD*, 42(1–3):335–346.

HAUSSLER, DAVID. 1990. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Technical Report UCSC-CRL-91-02, Baskin Center for Computer Engineering and Information Sciences, University of California at Santa Cruz. Revised version of September 1989 report.

HAYES, PATRICK J. 1979. The logic of frames. In *Frame conceptions and text understanding*, ed. by Dieter Metzing, volume 5 of *Research in Text Theory*, 46–61. New York: de Gruyter.

HEARST, MARTI A. 1991. Noun homograph disambiguation using local context in large text corpora. In *Seventh Annual Conference of the University of Waterloo Centre for the New OED and Text Research: Using Corpora*, 1–22, Oxford.

HENDLER, JAMES A. 1988. *Integrating marker-passing and problem-solving: A spreading activation approach to improved choice in planning*. Hillsdale, NJ: Lawrence Erlbaum Associates. Revision of 1986 thesis, Brown University.

HERWEG, MICHAEL. 1991. Aspectual requirements of temporal connectives: Evidence for a two-level approach to semantics. In (Pustejovsky & Bergler 1991), 152–164.

HIGGINBOTHAM, JAMES. 1983. The logic of perceptual reports: An extensional alternative to situation semantics. *Journal of Philosophy*, 80:100–127.

HINDLE, DONALD. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, 268–275, Pittsburgh, PA.

HINDLE, DONALD & MATS ROOTH. 1991. Structural ambiguity and lexical relations. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, 229–236, Berkeley, CA.

HINKELMAN, ELIZABETH ANN. 1990. Abductive speech act recognition. In *Working Notes from the Spring Symposium on Automated Abduction*, 23–25, Stanford University, Stanford, CA. AAAI.

HINTON, GEOFFREY E. 1979a. Imagery without arrays. *Behavioral and Brain Sciences*, 2:555–556.

HINTON, GEOFFREY E. 1979b. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 3:231–250.

HINTON, GEOFFREY E. & L. M. PARSONS. 1981. Frames of reference and mental imagery. In *Attention and performance ix*, ed. by A. Baddeley & J. Long. Hillsdale, NJ: Lawrence Erlbaum Associates.

HINTON, GEOFFREY E. & TERRENCE J. SEJNOWSKI. 1986. Learning and relearning in boltzmann machines. In *Parallel distributed processing*, ed. by David E. Rumelhart, James L. McClelland, & the PDP Research Group, volume 1, 282–317. Cambridge, MA: MIT Press.

HIRST, GRAEME. 1987. *Semantic interpretation and the resolution of ambiguity*. Cambridge: Cambridge University Press.

HOBBS, JERRY R. 1990. An integrated abductive framework for discourse interpretation. In *Working Notes from the Spring Symposium on Automated Abduction*, 10–12, Stanford University, Stanford, CA. AAAI.

HOBBS, JERRY R., MARK STICKEL, PAUL MARTIN, & DOUGLAS EDWARDS. 1988. Interpretation as abduction. In *Proceedings of the 26th Annual Conference of the Association for Computational Linguistics*, 95–103, Buffalo, NY.

HOLYOAK, KEITH J. 1977. The form of analog size information in memory. *Cognitive Psychology*, 9:31–51.

HOPFIELD, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA*, 79:2554–2558.

HUME, DAVID. 1888. *A treatise of human nature*. Oxford: Clarendon Press.

JACKENDOFF, RAY S. 1972. *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

JACKENDOFF, RAY S. 1983. *Semantics and cognition*. Cambridge, MA: MIT Press.

JACKENDOFF, RAY S. 1990. *Semantic structures*. Cambridge, MA: MIT Press.

JACKSON, FRANK & JOHN PARGETTER. 1973. Indefinite probability statements. *Synthese*, 26:205–215.

JACOBS, PAUL. 1985. A knowledge-based approach to language generation. Technical Report 86/254, University of California at Berkeley Computer Science Division, Berkeley, CA.

JAYNES, E. T. 1979. Where do we stand on maximum entropy. In *The maximum entropy formalism*, ed. by R. D. Levine & M. Tribus. Cambridge, MA: MIT Press.

JESPERSEN, OTTO. 1946. *A modern English grammar on historical principles*, volume 6. London: George Allen & Unwin.

JOHANSSON, STIG & KNUT HOFLAND. 1989. *Frequency analysis of English vocabulary*, volume 1 & 2. Oxford: Oxford University Press.

JOHNSON, MARCIA K. & CAROL L. RAYE. 1981. Reality monitoring. *Psychological Review*, 88:67–85.

JOSEPHSON, JOHN R. 1990. On the "logical form" of abduction. In *Working Notes from the Spring Symposium on Automated Abduction*, 140–144, Stanford University, Stanford, CA. AAAI.

JURAFSKY, DANIEL S. 1990. Representing and integrating linguistic knowledge. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki.

JURAFSKY, DANIEL S. 1991. An on-line model of human sentence interpretation. In *Program of the Thirteenth Annual Conference of the Cognitive Science Society*, Chicago.

JURAFSKY, DANIEL S., 1992. *An on-line computational model of human sentence interpretation: A theory of the representation and use of linguistic knowledge.* Berkeley, CA: Univ. of Calif. at Berkeley dissertation.

KAHNEMAN, DANIEL, PAUL SLOVIC, & AMOS TVERSKY (eds.). 1982. *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

KATZ, JERROLD J. & JERRY A. FODOR. 1963. The structure of a semantic theory. *Language*, 39(2):170–210.

KAY, MARTIN. 1979. Functional grammar. In *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society*, 142–158, Berkeley, CA.

KEARNS, MICHAEL J. 1990. *The computational complexity of machine learning.* Cambridge, MA: MIT Press.

KEENAN, JANICE M., GEORGE R. POTTS, JONATHAN M. GOLDING, & TRACY M. JENNINGS. 1990. Which elaborative inferences are drawn during reading? a question of methodologies. In (Balota *et al.* 1990), 377–402.

KERR, B., S. M. CONDON, & L. A. MCDONALD. 1985. Cognitive spatial processing and the regulation of posture. *Journal of Experimental Psychology: Human Perception and Performance*, 11:617–622.

KOSSLYN, S. M. 1975. Information representation in visual images. *Cognitive Psychology*, 7:341–370.

KOSSLYN, S. M. 1976. Can imagery be distinguished from other forms of internal representation? Evidence from studies of information retrieval times. *Memory & Cognition*, 4:291–297.

KOSSLYN, S. M. 1980. *Image and mind.* Cambridge, MA: Harvard University Press.

KOSSLYN, S. M. 1983. *Ghosts in the mind's machine.* New York: Norton.

KOSSLYN, S. M., J. BRUNN, K. CAVE, & R. WALLACH. 1984. Individual differences in mental imagery ability: A computational analysis. *Cognition*, 18:195–243.

KOSSLYN, S. M., J. D. HOLTZMAN, MARTHA J. FARAH, & M. S. GAZZANIGA. 1985. A computational analysis of mental image generation: Evidence from functional dissociations in split-brain patients. *Journal of Experimental Psychology: General*, 114:311–341.

KUČERA, HENRY & W. NELSON FRANCIS. 1967. *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

KYBURG, JR., HENRY E. 1961. *Probability and the logic of rational belief.* Middletown, CT: Wesleyan University Press.

KYBURG, JR., HENRY E. 1983. *Epistemology and inference.* Minneapolis: University of Minnesota Press.

LABERGE, D. & S. J. SAMUELS. 1974. Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6:293–323.

LABOV, WILLIAM. 1966. *The social stratification of English in New York City.* Washington, D.C.: Center for Applied Linguistics.

LABOV, WILLIAM. 1972. *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press.

LAIRD, JOHN, ALLEN NEWELL, & PAUL ROSENBLOOM. 1987. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64.

LAIRD, JOHN, PAUL ROSENBLOOM, & ALLEN NEWELL. 1986. *Universal subgoaling and chunking: The automatic generation and learning of goal hierarchies*. Boston: Kluwer.

LAKOFF, GEORGE. 1987a. Cognitive models and prototype theory. In (Neisser 1987), 63–100.

LAKOFF, GEORGE. 1987b. *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

LANG, EWALD, KAI-UWE CARSTENSEN, & GEOFF SIMMONS. 1991. *Modelling spatial knowledge on a linguistic basis: Theory-prototype-integration*, volume 481 of *Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag.

LANGACKER, RONALD W. 1987. *Foundations of cognitive grammar*, volume 1. Stanford, CA: Stanford University Press.

LARI, K. & S. J. YOUNG. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56.

LDOCE. 1978. *Longman's dictionary of contemporary English*. Harlow, Essex: Longman.

LEECH, GEOFFREY & ROSEMARY LEONARD. 1974. A computer corpus of British English. *Hamburger Phonetische Beiträge*, 13:41–57. Cited in Francis & Kučera 1982.

LEES, ROBERT B. 1963. *The grammar of English nominalizations*. The Hague: Mouton.

LEES, ROBERT B. 1970. Problems in the grammatical analysis of English nominal compounds. In *Progress in linguistics*, ed. by Manfred Bierwisch & Karl Erich Heidolph, 174–186. The Hague: Mouton.

LENAT, DOUG & R. V. GUHA. 1988. The world according to CYC. Technical Report ACA-AI-300-88, MCC Artificial Intelligence Laboratory, Austin, TX.

LENAT, DOUGLAS B. & R. V. GUHA. 1990. *Building large knowledge-based systems: Representation and inference in the CYC project*. Reading, MA: Addison-Wesley.

LEONARD, ROSEMARY. 1984. *The interpretation of English noun sequences on the computer*. Amsterdam: North Holland.

LEVESQUE, HECTOR. 1986. Making believers out of computers. *Artificial Intelligence*, 30:81–108.

LEVI, JUDITH N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

LEVINE, D. N., J. WARACH, & MARTHA J. FARAH. 1985. Two visual systems in mental imagery: Dissociation of "what" and "where" in imagery disorders due to bilateral posterior cerebral lesions. *Neurology*, 35:1010–1018.

LIEBERMAN, PHILIP. 1984. *The biology and evolution of language*. Cambridge, MA: Harvard University Press.

LONGACRE, R. E. 1976. *An anatomy of speech notions*. Lisse: Peter de Ridder Press.

LUCAS, M. M., M. K. TANENHAUS, & G. N. CARLSON. 1987. Inferences in sentence comprehension: The role of constructed representations. In *Program of the Ninth Annual Conference of the Cognitive Science Society*, 566–574, Hillsdale, NJ. Lawrence Erlbaum Associates.

LUCE, R. DUNCAN & HOWARD RAIFFA. 1957. *Games and decisions: Introduction and critical survey.* New York: Wiley.

LYCAN, WILLIAM G. 1984. *Logical form in natural language.* Cambridge, MA: MIT Press.

MACKEN, M. & D. BARTON. 1980. The acquisition of the voicing contrast in English: A study of voice onset time in word-initial stop consonants. *Journal of Child Language,* 7:41–74.

MAGERMAN, DAVID M. & MITCHELL P. MARCUS. 1990. Parsing a natural language using mutual information statistics. In *Proceedings of the Ninth National Conference on Artificial Intelligence.*

MARCUS, MITCHELL. 1991. The automatic acquisition of linguistic structure from large corpora: An overview of work at the university of pennsylvania. In *Working Notes from the Spring Symposium on Machine Learning of Natural Language and Ontology,* 123–125, Stanford University, Stanford, CA. AAAI.

MARR, DAVID & H. K. NISHIHARA. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London,* 200:269–294.

MARSLEN-WILSON, WILLIAM & LORRAINE KOMISARJEVSKY TYLER. 1980. The temporal structure of spoken language. *Cognition,* 8:1–71.

MARSLEN-WILSON, WILLIAM & LORRAINE KOMISARJEVSKY TYLER. 1981. Central processes in speech understanding. *Philosophical Transactions of the Royal Society,* B 295:317–322.

MARSLEN-WILSON, WILLIAM & LORRAINE KOMISARJEVSKY TYLER. 1987. Against modularity. In (Garfield 1987), 37–62.

MARTIN, CHARLES E. & CHRISTOPHER K. RIESBECK. 1986. Uniform parsing and inferencing for learning. In *Proceedings of the Fifth National Conference on Artificial Intelligence,* 257–261.

MARTIN, JAMES H. 1988. A computational theory of metaphor. Technical Report 88/465, University of California at Berkeley, Computer Science Division, Berkeley, CA.

MARTIN, JAMES H. 1990. *A computational model of metaphor interpretation.* San Diego: Academic Press.

MARTIN, JAMES H. 1991. Conventional metaphor and the lexicon. In (Pustejovsky & Bergler 1991), 56–66.

MAY, ROBERT. 1985. *Logical form: Its structure and its derivation.* Cambridge, MA: MIT Press.

MCCARTHY, JOHN & PATRICK J. HAYES. 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine intelligence,* ed. by B. Meltzer & D. Michie, volume 4, 463–502. Edinburgh: Edinburgh University Press.

MCCAWLEY, J. D. 1968. The role of semantics in a grammar. In *Universals in linguistic theory,* ed. by Emmon Bach & R. T. Harms. New York: Holt, Rinehart, and Winston.

MCCLELLAND, JAMES L. 1986. The programmable blackboard model of reading. In (McClelland *et al.* 1986), 122–169.

MCCLELLAND, JAMES L., DAVID E. RUMELHART, & THE PDP RESEARCH GROUP (eds.). 1986. *Parallel distributed processing,* volume 2. Cambridge, MA: MIT Press.

MCDONALD, DAVID B. 1982. Understanding noun compounds. Technical Report CMU-CS-82-102, Carnegie-Mellon Univ., Dept. of Comp. Sci., Pittsburgh, PA.

MCKOON, GAIL & ROGER RATCLIFF. 1981. The comprehension processes and memory structures involved in instrumental inference. *Journal of Verbal Learning and Verbal Behavior*, 20:671–682.

MCKOON, GAIL & ROGER RATCLIFF. 1986. Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12:82–91.

MCKOON, GAIL & ROGER RATCLIFF. 1989a. Inferences about contextually defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15:1134–1146.

MCKOON, GAIL & ROGER RATCLIFF. 1989b. Semantic association and elaborative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15:326–338.

MCKOON, GAIL & ROGER RATCLIFF. 1990. Textual inferences: Models and measures. In (Balota *et al.* 1990), 403–421.

MEDIN, DOUGLAS L. & WILLIAM D. WATTENMAKER. 1987. Category cohesiveness, theories, and cognitive archeology. In (Neisser 1987), 25–62.

MERVIS, CAROL B. 1980. Category structure and the development of categorization. In *Theoretical issues in reading comprehension*, ed. by R. Spiro, B. C. Bruce, & W. F. Brewer. Hillsdale, NJ: Lawrence Erlbaum Associates.

MILL, JOHN STUART. 1896. *A system of logic ratiocinative and inductive.* London: Longmans, Green. Cited in Cohen 1989.

MILLER, GEORGE A. 1990. Nouns in wordnet: A lexical inheritance system. *Journal of Lexicography*, 3(4):245–264.

MILLER, GEORGE A., RICHARD BECKWITH, CHRISTIANE FELLBAUM, DEREK GROSS, & KATHERINE J. MILLER. 1990. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

MILLIKAN, RUTH GARRETT. 1984. *Language, thought, and other biological categories.* Cambridge, MA: MIT Press.

MINKSY, MARVIN. 1975. A framework for representing knowledge. In *The psychology of computer vision*, ed. by Patrick Winston. New York: McGraw-Hill.

MINTON, STEVE, 1988. *Learning effective search control knowledge: An explanation-based approach.* Carnegie Mellon University dissertation.

MJOLSNESS, ERIC. 1990. Bayesian inference on visual grammars by neural nets that optimize. Technical Report YALEU-DCS-TR-854, Dept. of Computer Science, Yale University.

MOYER, R. S. 1973. Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics*, 13:180–184.

MOYER, R. S. & R. H. BAYER. 1976. Mental comparison and the symbolic distance effect. *Cognitive Psychology*, 8:228–246.

NASA. 1985. *Nasa thesaurus.*

NEISSER, ULRIC (ed.). 1987. *Concepts and conceptual development: Ecological and intellectual factors in categorization.* Cambridge: Cambridge University Press.

NEISSER, ULRIC & N. H. KERR. 1973. Spatial and mnemonic properties of visual images. *Cognitive Psychology,* 5:138–150.

NEWELL, ALLEN. 1990. *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

NG, HWEE TOU & RAYMOND J. MOONEY. 1990. The role of coherence in constructing and evaluating abductive explanations. In *Working Notes from the Spring Symposium on Automated Abduction,* 13–17, Stanford University, Stanford, CA. AAAI.

NILSSON, NILS J. 1986. Probabilistic logic. *Artificial Intelligence,* 28:71–87.

NIRENBURG, SERGEI & LORI LEVIN. 1991. Syntax-driven and ontology-driven lexical semantics. In (Pustejovsky & Bergler 1991), 9–19.

NORVIG, PETER. 1987. A unified theory of inference of text understanding. Technical Report 87/339, University of California at Berkeley, Computer Science Division, Berkeley, CA.

NORVIG, PETER, 1989. Non-disjunctive ambiguity. Unpublished draft, University of California at Berkeley.

NORVIG, PETER & ROBERT WILENSKY. 1990a. A critical evaluation of commensurable abduction models for semantic interpretation. In *Proceedings of the Thirteenth International Conference on Computational Linguistics,* 225–230, Helsinki.

NORVIG, PETER & ROBERT WILENSKY. 1990b. Problems with abductive language understanding models. In *Working Notes from the Spring Symposium on Automated Abduction,* 18–22, Stanford University, Stanford, CA. AAAI.

O'BRIEN, E. J., D. M. SHANK, J. L. MYERS, & K. RAYNER. 1988. Elaborative inferences during reading: Do they occur on-line? *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 14:410–420.

OSTLER, NICHOLAS. 1980. *A theory of case linking and agreement.* Bloomington, IN: Indiana University Linguistics Club.

PAIVIO, ALLAN. 1971. *Imagery and verbal processes.* New York: Holt, Rinehart, and Winston.

PAIVIO, ALLAN. 1975. Perceptual comparisons through the mind's eye. *Memory & Cognition,* 3:635–647.

PAIVIO, ALLAN. 1986. *Mental representations.* New York: Oxford University Press.

PARSONS, TERENCE. 1985. Underlying events in the logical analysis of English. In *Actions and events: Perspectives on the philosophy of donald davidson,* ed. by E. LePore & B. P. McLaughlin. Oxford: Basil Blackwell.

PARSONS, TERENCE. 1990. *Events in the semantics of english: A study in subatomic semantics.* Cambridge, MA: MIT Press.

PARTEE, BARBARA H. 1984. Compositionality. In *Varieties of formal semantics,* ed. by Fred Landman & Frank Veltman, 281–311. Dordrecht: Foris.

PEARL, JUDEA. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

PEARL, JUDEA. 1990. Probabilistic and qualitative abduction. In *Working Notes from the Spring Symposium on Automated Abduction*, 155–158, Stanford University, Stanford, CA. AAAI.

PEIRCE, CHARLES SANDERS. 1931. *Collected papers of charles sanders peirce.* Cambridge, MA: Harvard University Press. Edited by Charles Hartshorne and Paul Weiss. Published over 1931–1935.

PINKER, STEVEN. 1984. Visual cognition: An introduction. *Cognition*, 18:1–63.

PINKER, STEVEN. 1989. *Learnability and cognition.* Cambridge, MA: MIT Press.

PLATT, J. T. 1971. *Grammatical form and grammatical meaning: A tagmemic view of Fillmore's Deep Structure Case concepts.* Amsterdam: North-Holland.

POLLACK, JORDAN B. 1988. Recursive auto-associative memory: Devising compositional distributed representations. Technical Report MCCS-88-124, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

POLLACK, JORDAN B. 1989. Implications of recursive auto associative memories. In *Advances in neural information processing systems*, ed. by David Touretzky, 527–536. San Mateo: Morgan Kaufmann.

POLLACK, JORDAN B. 1990. Recursive distributed representations. *Artificial Intelligence*, 46:77–105.

POLLARD, CARL & IVAN A. SAG. 1987. *Information-based syntax and semantics: Volume 1: Fundamentals.* Stanford, CA: Center for the Study of Language and Information.

POLLOCK, JOHN L. 1990. *Nomic probability and the foundations of induction.* New York: Oxford University Press.

POPPER, KARL R. 1959a. *The logic of scientific discovery.* New York: Basic Books.

POPPER, KARL R. 1959b. The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10:5–56.

POPPER, KARL RAIMUND. 1983. *Realism and the aim of science.* Totowa, NJ: Rowman and Littlefield.

POSNER, M. & S. KEELE. 1975. Attention and cognitive control. In *Information processing and cognition*, ed. by R. L. Solso. New Jersey: Lawrence Erlbaum Associates.

POTTER, M. & B. A. FAULCONER. 1975. Time to understand pictures and words. *Nature*, 253:437–438.

POTTS, G. R., J. M KEENAN, & J. M GOLDING. 1988. Assessing the occurrence of elaborative inferences: Lexical decision versus naming. *Journal of Memory and Language*, 27:399–415.

PUSTEJOVSKY, JAMES & SABINE BERGLER (eds.). 1991. *Lexical semantics and knowledge representation: Proceedings of a workshop*, University of California, Berkeley. ACL Special Interest Group on the Lexicon.

QUILLIAN, M. ROSS. 1969. The teachable language comprehender: A simulation program and theory of language. *Communications of the Association for Computing Machinery*, 12(8):459–476.

QUIRK, RANDOLPH, SIDNEY GREENBAUM, GEOFFREY LEECH, & JAN SVARTVIK. 1985. *A comprehensive grammar of the English language.* New York: Longman.

RAM, ASHWIN. 1990. Goal-based explanation. In *Working Notes from the Spring Symposium on Automated Abduction*, 26–29, Stanford University, Stanford, CA. AAAI.

RAM, ASHWIN & DAVID LEAKE. 1991. Evaluation of explanatory hypotheses. In *Program of the Thirteenth Annual Conference of the Cognitive Science Society*, 867–871, Chicago.

RAMSEY, FRANK PLUMPTON. 1931. *Foundations of mathematics*. New York: Harcourt.

RATCLIFF, G. & F. NEWCOMBE. 1982. Object recognition: Some deductions from the clinical evidence. In *Normality and pathology in cognitive functions*, ed. by A. W. Ellis. New York: Academic Press.

REGIER, TERRY. 1990. Learning spatial terms without explicit negative evidence. Technical Report TR-90-057, International Computer Science Institute, Berkeley, CA.

REGIER, TERRY. 1991a. Learning object-relative spatial concepts in the l_0 project. In *Program of the Thirteenth Annual Conference of the Cognitive Science Society*, 191–196.

REGIER, TERRY. 1991b. Learning perceptually-grounded semantics in the l_0 project. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, 138–145.

REGIER, TERRY. 1991c. Learning spatial concepts using a partially-structured connectionist architecture. Technical Report TR-91-050, International Computer Science Institute, Berkeley, CA.

RIDDOCH, M. J. & G. W. HUMPHREYS. 1987. Visual object processing in optic aphasia: A case of semantic access agnosia. *Cognitive Neuropsychology*, 4:131–185.

RIEGER, CHUCK. 1977. Spontaneous computation in cognitive models. *Cognitive Science*, 1:315–354.

RIESBECK, CHRISTOPHER K. 1986. From conceptual analyzer to direct memory access parsing: An overview. In *Advances in cognitive science 1*, 236–258. Chichester: Ellis Horwood.

ROSCH, ELEANOR H. 1973. On the internal structure of perceptual and semantic categories. In *Cognitive development and the acquisition of language*, ed. by T. E. Moore. New York: Academic Press.

ROSCH, ELEANOR H. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

ROSCH, ELEANOR H. & CAROL B. MERVIS. 1975. Family resemblences: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

ROSCH, ELEANOR H., CAROL B. MERVIS, W. GRAY, D. JOHNSON, & P. BOYES-BRAEM. 1976. Basic objects in natural categories. *Cognitive Psychology*, 3:382–439.

ROTH, EMILIE M. & EDWARD J. SHOBEN. 1983. The effect of context on the structure of categories. *Cognitive Psychology*, 15:346–378.

RUMELHART, DAVID E. 1991. (position paper abstract). In *Working Notes from the Spring Symposium on Connectionist Natural Language Processing*, p. 242, Stanford University, Stanford, CA. AAAI.

RUMELHART, DAVID E. & JAMES L. MCCLELLAND. 1986. On learning the past tenses of english verbs. In (McClelland *et al.* 1986), 216–271.

RUSSELL, STUART J. 1990. Fine-grained decision-theoretic search control. In *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence*, Cambridge, MA. Morgan Kaufmann.

RUSSELL, STUART J. 1991. An architecture for bounded rationality. In *Proceedings of the AAAI Spring Symposium on Integrated Architectures for Intelligent Agents*, Stanford, CA.

RUSSELL, STUART J. & ERIC H. WEFALD. 1988. Multi-level decision-theoretic search. In *Proceedings of the AAAI Spring Symposium on Computer Game Playing*, Stanford, CA.

RUSSELL, STUART J. & ERIC H. WEFALD. 1989. On optimal game-tree search using rational metareasoning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI.

RUSSELL, STUART J. & ERIC H. WEFALD. 1991. *Do the right thing: Studies in limited rationality.* Cambridge, MA: MIT Press.

SCHANK, ROGER C. 1973. Identification of conceptualizations underlying natural language. In (Schank & Colby 1973).

SCHANK, ROGER C. & KENNETH MARK COLBY (eds.). 1973. *Computer models of thought and language.* San Francisco: W. H. Freeman.

SCHMIDHUBER, JÜRGEN. 1991. Neural sequence chunkers. Technical Report FKI-148-91, Institut für Informatik, Technische Universität München, Munich.

SHASTRI, LOKENDRA. 1988a. A connectionist approach to knowledge representation and limited inference. *Cognitive Science*, 12(3):331–392.

SHASTRI, LOKENDRA. 1988b. *Semantic networks: An evidential formalization and its connectionist realization.* San Mateo, CA: Morgan Kaufmann.

SHASTRI, LOKENDRA. 1989. Default reasoning in semantic networks: A formalization of recognition and inheritance. *Artificial Intelligence*, 39:283–355.

SHASTRI, LOKENDRA. 1991. Why semantic nets? In (Sowa 1991), 109–136.

SHEPARD, R. N. 1981. Psychophysical complementarity. In *Perceptual organization*, ed. by M. Kubovy & J. Pomerantz. Hillsdale, NJ: Lawrence Erlbaum Associates.

SHEPARD, R. N. & S. CHIPMAN. 1970. Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1:1–17.

SHIEBER, STUART M. 1986. *An introduction to unification-based approaches to grammar.* Stanford, CA: Center for the Study of Language and Information.

SHIFFRIN, R. & W. SCHNEIDER. 1977. Controlled and automatic human information processing: 2. perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84:127–190.

SILVERSTEIN, MICHAEL. 1976. Hierarchy of features and ergativity. In *Grammatical categories in Australian languages*, ed. by R. M. W. Dixon. Canberra: Australian Institute of Aboriginal Studies. Cited in Foley & van Valin 1984.

SINCLAIR, JOHN, PATRICK HANKS, GWYNETH FOX, ROSAMUND MOON, & PENNY STOCK. 1987. *Collins COBUILD English language dictionary.* London and Glasgow: Collins. (COLLINS Birmingham University International Language Database).

SINGH, J. D. 1974. Pāṇini's theory of kārakas. *International Journal of Dravidian Linguistics*, 3:287–320. Cited in Somers 1987.

SKOUSEN, ROYAL. 1989. *Analogical modeling of language.* Dordrecht: Kluwer.

SMADJA, FRANK A. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, 252–259, Berkeley, CA.

SMADJA, FRANK A., 1991a. *Extracting collocations from text. an application: Language generation.* New York: Columbia University dissertation.

SMADJA, FRANK A. 1991b. From n-grams to collocations: An evaluation of Xtract. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, 279–284, Berkeley, CA.

SMITH, EDWARD E. & DOUGLAS L. MEDIN. 1981. *Categories and concepts.* Cambridge, MA: Harvard University Press.

SOMERS, H. L. 1987. *Valency and case in computational linguistics*, volume 3 of *Edinburgh Information Technology Series.* Edinburgh: Edinburgh University Press.

SOWA, JOHN F. (ed.). 1991. *Principles of semantic networks: Explorations in the representation of knowledge.* Series in Representation and Reasoning. San Mateo, CA: Morgan Kaufmann.

SPARCK-JONES, KAREN & C. J. VAN RIJSBERGEN. 1976. Information retrieval test collections. *Journal of Documentation*, 32(1).

STICKEL, MARK E. 1990. A method for abductive reasoning in natural-language interpretation. In *Working Notes from the Spring Symposium on Automated Abduction*, 5–9, Stanford University, Stanford, CA. AAAI.

SVARTVIK, JAN (ed.). 1990. *The London-Lund corpus of spoken English: Description and research.* Lund: Lund University Press.

TALMY, LEONARD. 1983. Spatial orientation: Theory, research, and application. In *How language structures space*, ed. by Herbert Pick & Linda Acredolo. New York: Plenum Press.

TALMY, LEONARD. 1985. Lexicalization patterns: Semantic structure in lexical forms. In *Language typology and syntactic description*, ed. by T. Shopen, volume 3: Grammatical Categories and the Lexicon. Cambridge: Cambridge University Press.

TALMY, LEONARD. 1988. Force dynamics in language and cognition. *Cognitive Science*, 12:49–100.

TAMBE, MILIND & ALLEN NEWELL. 1988. Some chunks are expensive. In *Proceedings of the Fifth International Conference on Machine Learning*, 451–458.

TAMBE, MILIND & PAUL ROSENBLOOM. 1989. Eliminating expensive chunks by restricting expressiveness. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 731–737.

TESAURO, GERALD J. 1991. Practical issues in temporal difference learning. Technical Report RC 17223 (#76307), IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY.

THAGARD, PAUL. 1989. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–502.

THAGARD, PAUL. 1990. Explanatory coherence and naturalistic decision making. In *Working Notes from the Spring Symposium on Automated Abduction*, 125–129, Stanford University, Stanford, CA. AAAI.

THAGARD, PAUL, 1991. Probabilistic networks and explanatory coherence. Cognitive Science Laboratory, Princeton University, Draft of February 14.

TILL, R. E., E. F. MROSS, & W. KINTSCH. 1988. Time course of priming for associate and inference words in a discourse context. *Memory & Cognition*, 16:283–298.

TVERSKY, AMOS. 1975. Features of similarity. *Psychological Review*, 84:327–352.

UTGOFF, PAUL E. 1986. Shift of bias for inductive concept learning. In *Machine learning: An artificial intelligence approach*, ed. by Ryszard S. Michalski, Jaime G. Carbonell, & Tom M. Mitchell, volume 2, 107–148. San Mateo, CA: Morgan Kaufmann.

VALIANT, L. G. 1984. A theory of the learnable. *Communications of the Association for Computing Machinery*, 27(11):1134–1142.

VILAIN, MARC B. 1985. The restricted language architecture of a hybrid representation system. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles. Morgan Kaufmann.

VLACH, FRANK. 1983. On situation semantics for perception. *Synthese*, 54:129–52.

VON NEUMANN, JOHN & OSKAR MORGENSTERN. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press. Second edition: 1947.

VON WINTERFELDT, DETLOF & WARD EDWARDS. 1986. *Decision analysis and behavioral research*. Cambridge: Cambridge University Press.

WALTZ, DAVID L. & JORDAN B. POLLACK. 1985. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, 9:51–74.

WARREN, BEATRICE. 1978. *Semantic patterns of noun-noun compounds*. Gothenburg, Sweden: Acta Universitatis Gothoburgensis.

WEBER, SUSAN HOLLBACH. 1989a. Connectionist models and figurative speech. Technical Report TR 289, Deutsches Forschungszentrum für Künstliche Intelligenz, Kaiserslautern, Germany.

WEBER, SUSAN HOLLBACH. 1989b. Figurative adjective-noun interpretation in a structured connectionist network. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, 204–211.

WEBER, SUSAN HOLLBACH. 1989c. A structured connectionist approach to direct inferences and figurative adjective-noun combinations. Technical Report TR 289, Univ. of Rochester, Dept. of Comp. Sci., New York.

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.

WEBSTER. 1963. *Webster's seventh new collegiate dictionary*. Springfield, MA: G. & C. Merriam.

WEFALD, ERIC H. & STUART J. RUSSELL. 1989a. Adaptive learning of decision-theoretic search control knowledge. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, NY.

WEFALD, ERIC H. & STUART J. RUSSELL. 1989b. Estimating the value of computation: The case of real-time search. In *Proceedings of the AAAI Spring Symposium on AI and Limited Rationality*, Stanford, CA.

WERMTER, STEPHAN. 1989a. Integration of semantic and syntactic constraints for stuctural noun phrase disambiguation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1486–1491.

WERMTER, STEPHAN. 1989b. Learning semantic relationships in compound nouns with connectionist networks. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, 964–971.

WERMTER, STEPHAN & WENDY G. LEHNERT. 1989. Noun phrase analysis with connectionist networks. In *Connectionist approaches to language processing*, ed. by N. Sharkey & R. Reilly. In press.

WEXLER, K. & P. CULICOVER. 1980. *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

WHITNEY, P. & G. KELLAS. 1986. Processing category terms in context: Instantiation and the structure of semantic categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:39–48.

WILENSKY, ROBERT. 1986. Knowledge representation—a critique and a proposal. In *Experience, memory, and reasoning*, ed. by Janet L. Kolodner & Christopher K. Riebeck, 15–28. Hillsdale, NJ: Lawrence Erlbaum Associates.

WILENSKY, ROBERT, 1989. (untitled). Unpublished draft, University of California at Berkeley.

WILENSKY, ROBERT. 1991. Sentences, situations, and propositions. In (Sowa 1991), 191–227.

WILENSKY, ROBERT & YIGAL ARENS. 1980. PHRAN – a knowledge-based approach to natural language analysis. Technical Report UCB/ERL M80/34, University of California at Berkeley, Electronics Research Laboratory, Berkeley, CA.

WILENSKY, ROBERT, DAVID CHIN, MARC LURIA, JAMES MARTIN, JAMES MAYFIELD, & DEKAI WU. 1988. The Berkeley UNIX Consultant project. *Computational Linguistics*, 14(4):35–84.

WILKS, YORICK. 1973. An artificial intelligence approach to machine translation. In (Schank & Colby 1973), 114–151.

WILKS, YORICK. 1975a. An intelligent analyzer and understander of English. *Communications of the Association for Computing Machinery*, 18(5):264–274.

WILKS, YORICK. 1975b. Preference semantics. In *Formal semantics of natural language*, ed. by Edward L. Keenan, 329–348. Cambridge: Cambridge University Press.

WILKS, YORICK. 1982. Some thoughts on procedural semantics. In *Strategies for natural language processing*, ed. by Wendy G. Lehnert & Martin H. Ringle, 495–516. Hillsdale, NJ: Lawrence Erlbaum Associates.

WITTGENSTEIN, LUDWIG. 1963. *Philosophical investigations*. Oxford: Oxford University Press.

WOODS, WILLIAM A. 1975. What's in a link: Foundations for semantic networks. In *Representation and understanding*, ed. by Daniel G. Bobrow & Allan M. Collins, 35–82. New York: Academic Press.

WU, DEKAI. 1987. Concretion inferences in natural language understanding. In *Proceedings of GWAI-87, 11th German Workshop on Artificial Intelligence*, ed. by K. Morik, volume 152 of *Informatik-Fachberichte*, 74–83, Geseke. Springer-Verlag.

WU, DEKAI. 1989. A probabilistic approach to marker propagation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 574–580, Detroit, MI. Morgan Kaufmann.

WU, DEKAI. 1990. Probabilistic unification-based integration of syntactic and semantic preferences for nominal compounds. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, volume 2, 413–418, Helsinki.

WU, DEKAI. 1991a. Active acquisition of user models: Implications for decision-theoretic dialog planning and plan recognition. *User Modeling and User-Adapted Interaction*, 1(2):149–173. Special issue on plan recognition.

WU, DEKAI. 1991b. A continuum of induction methods for learning probability distributions with generalization. In *Program of the Thirteenth Annual Conference of the Cognitive Science Society*, Chicago. Lawrence Erlbaum Associates.

WU, DEKAI & BETTINA HORSTER. 1989. Active acquisition for user modeling in dialog systems. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, 987–994, Ann Arbor, MI. Lawrence Erlbaum Associates.

ZAIDEL, E. 1985. Language in the right hemisphere. In *The dual brain*, ed. by D. F. Benson & E. Zaidel. New York: Guilford Press.