1

MAX. RES. 100 LINES/INCH

METAL

MAX. RES. 100 LINES/INCH

**National Library
of Canada**

**Bibliothèque nationale
du Canada**

Canadian Theses Service

Service des thèses canadiennes

**Ottawa, Canada
K1A 0N4**

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

**A Decompositional Semantic System
for Natural Language Apprehension**

by

Terry Copeck, B.A., B.Sc.

**A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of
Master of Computer Science**

**School of Computer Science
Carleton University
Ottawa, Ontario**

April 21, 1989

**© copyright
1989, Frank Terrance Copeck**

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-54400-7

The Undersigned recommend to the
Faculty of Graduate Studies and Research
acceptance of the thesis

**A Decompositional Semantic System
for Natural Language Apprehension**

submitted by Terry Copeck, B.A., B.Sc.
in partial fulfillment of the requirements for
the degree of Master of Computer Science

Thesis Supervisor

Chairman, School of Computer Science
Carleton University

Date

Abstract

The sense of a natural language expression can be described by decomposing it into simpler elements of meaning. The decompositional approach taken here identifies 309 semantic elements called features on the basis of their salience or obviousness. Features are organized into hierarchies called domains through sortal and compositional relationships with one another. Domains are joined together by other relationships to form a single system. Word senses can be described by sequences of features selected from different domains.

To demonstrate the computational utility of this approach a number of texts have been composed from words described in terms of features. A program performs syntactic and semantic parses of these texts and produces network representations of their meaning. It is argued that these representations can support a variety of different cognitive artificial intelligence applications.

Acknowledgements

Prof. Franz Oppacher first shared his encyclopedic knowledge with me in the classroom and subsequently provided many hours of stimulating discussion during the period when these ideas were being formulated. As the thesis moved towards completion he remained always available to proofread it and suggest improvements.

Helen and Jack are valued for reasons they each know very well.

Table of Contents

Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vi
List of Figures	vii
List of Appendices	viii
Overview	1
Introduction	3
Designing General Systems	10
Elements of a Feature System	
Features	13
Domains	16
Building Systems	
A Feature System	19
A Linguistic System	19
Specifying Particular Systems	22
Relationships	
Kinds	24
Contexts	26
Domains and Features	38
Utterance Networks	68
Implementing a Demonstration Program	
Data Structures	74
Processing	81
Performance	96
Implementation	107
The Language Representation	116
Uses	118
Limitations	125
Remarks	127
Notes	214
References	238

List of Tables

TABLE	DESCRIPTION	PAGE
1.	Processing Modifiers Within Phrases	98
2.	Interpreting Auxiliary Verbs	99
3.	Interpreting Simple Prepositional Phrases	101
4.	Interpreting More Difficult Prepositional Phrases	104
5.	Recognition of Identical Entity Symbols	106
6.	Winograd's Adaptation of Halliday's Case Roles	236

List of Figures

FIGURE	DESCRIPTION	PAGE
1.	Types of Feature System Symbols	75
2.	Types of Dictionary Symbols	78
3.	Types of Network Symbols	80
4.	Pseudocode for Processing Input	82
5.	Pseudocode for Expanding Contractions	83
6.	Pseudocode for Matching Inflected Forms	83
7.	Pseudocode for Permuting the Sense of Inflected Forms	84
8.	Pseudocode for Disambiguation	90
9.	Pseudocode for Identifying Noun and Pronoun Referents	91
10.	Pseudocode for Attaching Autonomous Modifier Phrases	93
11.	Pseudocode for Determining Positional Case Tenancies	93
12.	Pseudocode for Determining Prepositional Attachment	95
13.	An Initial Classification of Verbs	177
14.	A Classification of Verbs of Motion	179

List of Appendices

APPENDIX	DESCRIPTION	PAGE
A.1	Glossary of Terms and Conventions	129
A.2	Arguments for A Universal Semantic Representation Based on Features	144
A.3	Linguistic Problems and Complexities	151
A.4	Issues Surrounding the Selection and Description of a Computational Lexicon	170
A.5	Classification of Semantic Theories Used in Computational Linguistic Systems	181
A.6	A Theoretic Context for the Feature System	185
A.7	Semantic Features and Words Defined in the System	195
A.8	Thematic Evaluation and Remarks on the Feature and Linguistic Systems	199
A.9	Feature Definitions of Nine Prepositions	211

Overview

Consider 'pages of the book', 'pages of paper', 'city of Ottawa' and 'city of young people'. Readers readily understand that 'of' communicates a different relationship in each of the four phrases; pages are the parts of a book which are made of paper, Ottawa is the name of a particular city where a large number of young people happen to be.

Because it is not yet certain how people come to understand these and other linguistic expressions correctly, all of the relevant factors have not been successfully expressed in a computational paradigm. The program documented here deals with such problems. It can distinguish the correct sense of each phrase by using aspects of the meanings of the constituent words. It can further discriminate between these and fourteen other kinds of relationships that recur in language on the basis of their semantics. A reader most interested in its details can turn to the section titled Implementing a Demonstration Program.

A number of existing computational linguistic systems have equivalent capabilities. Where this project goes further is in performing its operations on elaborate network representations which can be scaled up indefinitely in a coherent fashion and which are based on a particularly highly detailed and well-defined semantic model. As a result the representations of utterances which it produces are well suited to a wide variety of artificial intelligence applications such as script generation, user modelling and the abstraction of commonsense knowledge. How these and other tasks might be accomplished is described in some detail in The Language Representation section.

Although the main purpose of this thesis is to develop a program which can apprehend the meanings of natural language, doing so depends on the specification of a number of other factors. Computational semantics is not a mature field of study and in

order to resolve these issues satisfactorily I was forced to back up all the way to the point of identifying a proper set of words from which to abstract features to use in a semantic model which would be expressed in the utterance representation scheme implemented in the program.

The practical result is that the project's theoretic side grew to overshadow the practical issues of designing and implementing a program. This manuscript reflects that fact; the program documentation is preceded by two lengthy sections setting out the conceptual model which it implements. That organization may be necessary but it also tends to be somewhat misleading; the foundation in theory needed to make the program's operation intelligible is definitely subordinate to the results the program is capable of demonstrating.

Introduction

Language is unlike any other sphere of mental life. It is less central to the nature of human intelligence than other cognitive functions but it is employed more frequently than any other external expression of consciousness. Although the same principles govern the shape and substance of all utterances, each natural language has a unique lexicon and grammar. Language itself is universal however; a culture may lack writing but none lacks speech.

We are not born possessing language* so we learn a particular one at an early age. Grammars testify to the effort involved in doing this; no other mental process has a counterpart to these codifications of rules. Because language is an adopted rather than an intrinsic mental function, utterances are voluntary, arbitrary[§], and more consciously executed than other mental processes. This autonomy makes it an intersubjective phenomenon (Lacey, 1986), shared among all its users and dependent on none.

All of this suggests that cognitive scientists should find natural language a most suitable subject for study. This has not been the case. Computational research into natural language processing and understanding has advanced fitfully over the thirty year course of formal cognitive science[‡] and recent surveyors of the scene judge it to be one of the less active areas of study (Brittain, 1987; Winograd, 1987; Chapnick, 1988). Furthermore, of the funds being spent in this area, 80% are devoted either to machine translation or to applications and tools for restricted-domain interfaces (Batelle Institute in Obermeier, 1988). No doubt there are many reasons for this situation; a few are pertinent to this project.

* some argue that a universal subjective language or 'mentalese' exists (Fodor, 1975) or that language is latent within us because of our particular propensity or aptitude to learn it (Chomsky, 1965).

§ ie there is more than one way to express any thought in words.

‡ it is hard to identify exactly when a new field emerged from the shadow of its parent disciplines. Johnson (1986) dates the beginning of cognitive science from the 1956 Dartmouth Conference.

Computational research into language is frequently undertaken within studies of larger scope: formal and commonsense reasoning at Yale and SRI, question-answering (Joshi and Webber), even modelling the entire mind (Schank). The desire of these investigators to confront the particular issues that interest them leads them to take a depth-based approach to linguistic processes. Syntax, semantics and pragmatics are all subsumed into a single research design.

When they do study language processing *per se*, cognitive scientists usually investigate syntax before semantics or pragmatics. As a result the parsing process is relatively well understood but the language levels following it have received less attention. Language levels interdepend¹ and a theory of grammar imposes a view on other language phenomena. What this means is that semantics is not often the primary subject of computational research*, a situation this thesis is intended to address.

The goal of investigating semantics without assuming any particular grammatic approach[§] dictates the overall organization of the thesis: the semantic feature system[‡] it proposes is discussed prior to and separate from the demonstration program which must incorporate some particular parsing strategy. Because this research produced a substantial volume of original material which does not lend itself to summarization, some topics which would normally be found in the body of the text have been moved to appendices. Care has been taken to ensure that the main discussion can be read independently of its appendices.

A number of technical terms which are stipulated and used throughout the thesis are discussed in thematic groups in Appendix 1, together with descriptions of various

* aside from traditional sources like Roget (1933/1979) which produce classification schemes as byproducts, Jules Laffal's Concept Dictionary of English (1973) is the only detailed decompositional semantics *per se* I am aware of.

§ Although not all researchers would agree that the general-purpose or universal meaning representations implied by this approach are possible (King, 1987, 1988), the idea does have adherents and my approach can be adopted without any further justification. Because of the general lack of space some remarks that do serve to justify it—two hypotheses that articulate the rationale behind this thesis and a model of human cognitive functioning suited to it—have been relegated to Appendix 6.

‡ what is labelled a feature system here would be called a representation language in other theories.

notational and typographic conventions. These groups of related definitions summarize the main structural elements in the thesis concisely and provide an excellent introduction to the main discussion. Because the body of the text presumes a familiarity with these terms, the reader is encouraged to turn to Appendix 1 before proceeding.

The body of the text is composed of two parts. The first presents the system of semantic features developed during this project in the abstract; it is itself divided into halves. One half discusses the key elements required by a decompositional semantic model capable of describing the real world: elementary meaning tokens called features which are organized into hierarchical and linear structures called domains, and the relationships used to assemble features into domains and domains into a single system. A linguistic system based on this model describes word senses in terms of features and represents utterances as networks composed of these sense descriptions. This section also describes the analytic methodology used to identify the particular features which compose word senses. Simple ontological, psychological and psycholinguistic models supporting this analysis appear in Appendix 6.

The system presentation's second half describes the particular feature and linguistic systems which have been built in accordance with this model during this project. The discussion is couched in general terms so that the system can be presented clearly and succinctly, independent of the idiosyncrasies of an actual implementation. This section also serves as part of the program documentation because the systems it describes are essentially the same as those used in the demonstration program.

Discussion of these particular systems begins with the relationships they incorporate. Three types of relationships link features to one another in the feature system, in word meanings and in utterance representations: hyponymy, the sortal relationship; meronymy, the relationship between part and whole; and association, the relationship which governs whether features may or must occur together. Association is the

fundamental relationship of semantics.

The presentation of individual features is organized around the 25 domains* to which they belong. Although domain organization represents hyponymic relationships succinctly, features also engage in other relationships which are not expressed structurally and which are thus not as self-evident. Furthermore, the choice and interpretation of particular features needs explanation; although it has been a design goal to select self-evident feature names, a single word cannot usually express a feature's entire meaning. For these reasons features are not simply listed in their domains but are in most cases discussed at some length. This discussion constitutes the largest single section of the text.

The second main part of the thesis describes a Lisp program which produces representations of the meaning of input texts using word senses described in terms of features. The program parses sentences using an augmented transition network methodology which uses word morphology, syntax and semantics in its computations. It also attempts to disambiguate the senses of words in the utterance. These word senses are consolidated into networks that express a sentence's meaning and which are in turn amalgamated into a single network that represents the meaning of an episode or series of related sentences.

This section documents the program design and implementation in some detail and illustrates its processing capabilities with examples. Program documentation summarizes the knowledge/data structures used, the functional and actual organization of processing, and implementation particulars. Supporting materials which provide greater detail appear in appendixes. Appendix 7 lists the 305 features and 203 words defined in the program. Appendix 9 reproduces the Lisp definitions of 9 of the 86 prepositions defined in the program. Prepositions are a (relatively) closed class of words which express a wide

* although aspects (the semantic counterparts of perceptual data) constitute small but distinct domains, for simplicity they are treated as belonging to the domain of whatever feature they depend on.

variety of semantic relationships. This makes them a good candidate for performance testing, and in fact the program's most important set of test sentences distinguishes between eighteen different types of relationship realized by prepositional phrases.

Any program that processes language must embody a grammar, and elements of the syncretic grammar used in this program are scattered throughout its code. Because syntactic processing is not a main interest here, the parser implements a limited subset of English grammar; coordinating (logical) conjunctions and subordinate clauses are not recognized.

The semantic networks produced by the demonstration program would be the main interface* between the feature system and an external computational environment containing independent applications to model other cognitive functions. The final section of the thesis describes the structure of the representation which would ultimately result from processing a large volume of text in this framework. It also includes a number of suggestions about ways in which some of the knowledge structures commonly used in artificial intelligence could be generated from the representation.

Two additional appendices contain other materials related to the thesis topic which are not sufficiently germane to the feature system to require integration into some particular part of the body of the thesis. I have often wished that someone would produce a systematic classification of linguistic problems and complexities; no article seems to be complete without a few examples of problematic statements, but no one appears to have assembled a comprehensive list of these chestnuts. Appendix 3 presents a classification produced as a byproduct of this project.

Any research in which it is appropriate for the author to posit an ontological model is going to be influenced by factors other than previous work in the field². The set of features used here was developed by grouping the words in a vocabulary list on common

* unless an application program uses the information embodied in the feature system directly.

semantic factors and then generalizing from those factors. The key to obtaining feature sets which offer good conceptual coverage is selection of a suitable vocabulary. Analyzing most frequently used words proved to be a blind alley because such words have little descriptive range. Educational word lists are more satisfactory, and the one used here is reproduced with permission in Appendix 4 in order to bring it to the attention of interested parties. Not only do school vocabularies provide condensed raw material for semantic analysis, they also identify sets of words that are good candidates for inclusion in the dictionaries of general-purpose natural language processing programs.

Two systems are employed in the body of the thesis to make reference to materials which support the discussion but digress from it somewhat. Footnotes contain parenthetical marginalia which are directly relevant to the discussion but interrupt its flow. They are keyed to non-numeric symbols on a page-by-page basis and appear at the foot of the page to which they refer.

Other notes are too long to be footnoted but too short or too closely integrated with the text to be removed to an appendix. These appear in a second sequence of notes. Another reason for using an alternative scheme is that the presentation of the body of the text is interfered with if too much space on the page is devoted to text below the line. These notes are numbered sequentially from the beginning of the thesis body and proceed through the appendixes; they appear in a single section at the rear of the document. Keys for notes referring to a particular item appear within sentences adjacent to the item, while keys for more general notes appear between sentences. The holistic nature of developing an integrated semantic system is reflected in the fact it sometimes proved useful to reference the same note more than once in the text.

Most citation and discussion of related research appears in notes because of the limited space in the thesis. Secondary sources are cited through the original reference in the body but appear independently in the reference list to more clearly identify them. With the

exception of the handling of notes, the style of the American Psychological Association has been followed throughout this document.

A word about the presentation of material. The text identifies a number of distinct conceptual levels in discussing natural language: semantic features, word senses and utterance networks. Many of the elements composing each of these levels are highly similar. Setting aside whatever this might suggest about the nature of meaning, this similarity makes a repeated use of the same names and similar concepts in different connections almost inevitable. Thus hyponymy is not only a relationship between features but also the basis of a thing's sort. Essentially the same concept of a role is used in relations between features, in the role feature and in relations between words. Reuse risks confusing the reader, but it also means that once a term is understood it is understood in all contexts. Discussions of recurring elements try to provide enough detail to remind the reader of important points without belaboring them and boring him.

Designing General Systems

Any new attempt to formulate a computational semantics is well advised to begin by considering how previous researchers has undertaken the same task ³. Although observers of the scene acknowledge that computational theories about semantics have hitherto been relatively primitive ⁴, it is possible to identify differences in the ways that natural language systems deal with semantics and to classify them accordingly. Five types of semantic theories used in computational systems are summarized in Appendix 5. They are model-theoretic, decompositional, procedural, knowledge and object-oriented semantics.

Procedural, knowledge and object-oriented theories can immediately be eliminated from consideration because they focus primarily on the different ways meanings can be processed and not on how they are represented. This is not to say that these theories may not be suitable, but rather that their representational adequacy has not been evaluated. This leaves only model-theoretic and decompositional semantic theory as potential candidates. Model-theoretic semantics deals with formal rather than factual truth, ie the conditions that make a proposition true regardless of its content. This theory does not emphasize the structures required by a descriptive semantics; semantically simple and complex situations are both rendered as opaque propositions and treated as functional equivalents*. For these reasons semantic model theory must be discarded.

This seems to leave the computational semanticist no option but to adopt a theory based on decomposition. However, discussion to this point may have been framed too narrowly; perhaps the theories used in actual computational systems do not capture the

* the same practice can be found in field semantic studies in linguistics; -red- and cooked-in-a-pot-in-oil are treated as if they were of equivalent semantic order (Lehrer, 1974). As the name 'field' suggests, the emphasis is on discovering local differences which are broadly significant rather than on exhaustive analysis and systematization. Interesting as these explorations may be, they only postpone the primary task.

essential nature of semantics. Hirst (1987) addressed this rather obscure issue, identifying three alternative characters which meanings might have: 1) compositional, in which 'the meaning of the whole is a systematic function of the meaning of the parts'; 2) contextual, where statements take some part of their meaning from the extrasentential context); and 3) gestalt, where particular collocations have compound senses that do not follow systematically from the meanings of the individual words.

The gestalt possibility can be immediately rejected as unproductive. If each utterance's meaning is uniquely constructed, systematic analysis of any kind is impossible and meanings can only be catalogued. I believe (and attempt to establish in Appendix 6) that utterances always have a literal sense which can be understood from their constituents in isolation even if they have a more important pragmatic meaning in a particular context ⁵. Searle (1969) calls a statement's literal meaning its locutionary sense. Taking this position allows the contextual view to be finessed; the very act of acknowledging that an utterance's context may be important (and thereby the contextual approach) also serves to establish that every utterance has a simpler literal reading. According to Hirst's definition given above, literalness implies compositionality.

A compositional perspective imposes an agenda on computational processing. Before anything else can be done with it, an utterance must first be translated from a linguistic to a machine representation. If this is done correctly, cognitive functions distinct from those which perform this processing can successfully infer an utterance's appropriate contextual meaning by emulating whatever process a human recipient of the utterance uses to achieve his own understanding.

Even if this reasoning seems a little pat and although the decompositional approach remains controversial ⁶, the reader will probably agree that taking a compositional view of meaning is at least tenable. This perspective may be made yet more palatable by defining decompositional semantics quite broadly as 'any theory that analyzes and classifies meaning on some basis'. This definition covers a range of approaches from ones which

recognize only the most obvious differences in meaning* to others which push the analysis further in search of more fundamental structures. It does not imply any particular set of assumptions about the nature of the decompositional process and its results⁷.

The first part of this section presents features and domains, key elements of meaning in this analysis, and describes them in some detail. The second part considers how these and other elements can be integrated in a semantic feature system, and how that system can in turn be employed to produce representations of natural language utterances. Appendix 6 presents a conceptual framework that encompasses the feature system, ie models of ontology, psychology and psycholinguistics with which it is consistent.

* decomposition of this simple sort seems to pervade language processing. Examples of the ubiquity of decomposition in grammar are given in Appendix 2.

Elements of A Feature System

Features

Researchers in this field, Wilks (1977) in particular, suggest that well-chosen sets of semantic primitives are characterized by finitude, comprehensiveness, independence, primitiveness and non-circularity⁸. However the letters of the alphabet satisfy four of these five criteria, and no one would contend that the alphabet qualifies as a useful set of semantic primitives or that spelling constitutes understanding. This suggests that these criteria are not jointly sufficient to characterize whatever properties primitive sets must possess*. In fact, no matter what other demands are made of them, a set of primitives must first of all express meaning well. A measure of this characteristic is semantic salience[§], a token's capacity to convey a sense obviously or self-evidently, to require little explanation, to be routinely understood correctly by the largest possible number of people.

I therefore tried to select features which meet the implicit key criterion of salience rather than the five secondary criteria Wilks lists. Secondary criteria were only used to evaluate the results of selection on the primary criteria. Notwithstanding this the set of semantic features detailed here do in fact satisfy some of these additional criteria.

Certain consequences flow from the decision to select features on the basis of salience. A single criterion is less restricting than five. Judgments of salience are also inevitably subjective; people will always disagree on the importance of things. These

* the listed criteria could be defended on the grounds that a primitive's possession of salience is so obvious as not to need mention. However failure to identify a required characteristic outright tends to diminish the importance of that characteristic in the larger scheme of things, and to this degree the list is misleading.

§ in addition to any practical reasons for using it there is a theoretical argument for choosing salience. Features are the linguistic counterpart of aspects. Aspects package sensory experience (and generalizations based on it) and convey concepts; features play the same role in expressing word meanings. Efficacy in doing this is grounds for inclusion in the system.

considerations suggest that the feature system is inherently underdetermined and that certain aspects of its design lie permanently in the realm of craft rather than technology^{*}, 9. And that observation in turn suggests the current system should be considered a prototype whose final design will only emerge after a lengthy iteration of modification and evaluation.

A second consequence of using salience is to emphasize the names given to features. Salience is self-evidence, and a well-chosen name is an obvious mechanism to specify and convey a feature's meaning. For this reason features should be given names which accurately evoke the character which they are understood to have in the system so that users automatically tend to interpret them correctly. A good word sense description would be one that could be understood by itself, ie without reference to explanatory material.

A final result of using salience is to offer an oblique response to Lewis' criticism of compositional systems (1972, in Lakoff 1987) on the grounds that they simply translate utterances from one language to another without considering the truth conditions which give language meaning[§]. Although the most obvious answer to this objection is a successful program (ie such translations may simply be empirically useful), salience itself constitutes a response; it is a sort of truth-assertion. The things we find meaningful are not arbitrary and meaning elements do seem to have an inherent organization. As Lakoff says, "meaning is via real experiences in a very real world with very real bodies" (p. 206).

Applying the notion of semantic salience to the semantic theory as a whole suggests that it should employ terms meaningful to the nonspecialist, and that issues should be presented in a degree of detail which is neither microscopic nor macroscopic but rather

* craft employs judgments which spring from intuition and are apparently arbitrary, rather than consciously reasoned and demonstrably logical 10. This sort of judgment occurs here when decisions are made about which elements of meaning to include in the system (by identifying them with features) and which to leave as compounds (described by two or more features), how far to push decomposition in different areas of semantic space, and how to organize the resulting set of elements.

§ Goebel (1988) sees the issue as one of determining the nature of the necessary correspondence between symbols and the objects in the domain of interest which they identify.

familiar or intuitive. People are the final arbiters of what is meaningful and what is not, and one way to measure a semantic theory's descriptiveness is to see if people find it intuitive or at least easy to understand*.

Human cognitive processes seem to incorporate a natural scale or grain that can help to guide the selection of features. Rosch (1978) identifies 'basic level' categories from which higher and lower level categories are abstracted; computer animators who want to describe things' changing appearances seek intermediate terms between the too-simple elements of computer vision and the too-complex anatomy of many real-world objects (Pentland, 1986); and Cherniak's (1985) ballpark figurings of cognitive science resource requirements are predicated on first establishing the scale of the elements involved.

In language the phenomenon of scale can be seen in the varying degrees of difficulty people have explaining word meanings[§]. Words can be hard to describe concisely if they are very simple ('matter', 'idea') or if they are very complicated ('dawdle', 'carburetor'). In the case of vague, abstract words the speaker gropes for something meaningful to say, while in the case of highly concrete words he is faced with so much significant detail that he becomes verbose. In either case the same set of words tends to be used in explanations, and these words are candidates to be semantic features[‡].

The task of selecting features is similar to that faced by linguists studying semantic fields (Lehrer, 1974) and the same primary analytic technique of introspection will be used here. A top-down analysis based on pure introspection was rejected as too unstructured and an approach based on generalization from word senses was adopted instead. A group

* one could go further and say that a salient semantic theory should be easy to understand because it encodes the same distinctions that people make when they choose between different word meanings. Insofar as the theory draws different distinctions than popular ones, it is wrong.

§ this circumstance provides a way to check the quality of a set of features. Well-chosen features become less immediately meaningful when they are decomposed or amalgamated with other features. A good feature is at a local peak of meaningfulness.

‡ a correspondence can be found between semantic salience and the grain of features used in a system. Too fine-grained features produce empty or obscure definitions ('man' is .a1 + .a2 + .b1) and too coarse-grained ones produce tautologies ('man' is .humanbeing).

of common words were selected (Appendix 4) which were varied enough to allow a reasonably broad range of statements to be composed. These words were then grouped by part of speech. Within these groups words were categorized into sets on the basis of their most salient common aspect(s).

This classification was accomplished by making pairwise comparisons; as soon as a class has two members it becomes reasonably easy to determine if a third individual shares their common nature. In the event that all existing classes do not match the tested individual it is placed in a new class. Once all individuals have been classified, the set of cases can be reinspected as a group for consistency. This may result in individuals being moved from one class to another or entire classes being resorted. These processes were applied repeatedly to each categorial set until its subsets had a single member. The particular aspects used as criteria in the classification compose the set of features ¹¹.

Domains

The world would be almost unintelligible if our concepts of it were haphazard compositions of some sort of element. The contents of our mind only makes sense to us after we impose order on it. At first glance the methodical process of identifying features might seem to provide this order. However each stage of classification in the feature identification phase is based on whatever aspect(s) are currently most salient and makes no reference to aspects selected in other stages. Thus a group of individuals could be classified first on color, then on possession of a particular part, and finally on the shape of that part. Although the part criteria could be related to each other, neither would be related to color. At the same time other features which are related to color might be used as criteria elsewhere in the system. The process of identification produces a heterogeneous collection of features, not an organization.

A second analytic procedure is required to discover the inherent organization of this

collection, if any. It proceeds by grouping features together on the basis of greatest similarity until all have been classified. Irregularities in the structure formed by these classes or an overall lack of veridicality among them prompts reorganization until a relatively regular and descriptive system of features is produced. Thus the overall organizational form of the feature system is determined by the differing degrees of similarity between features, high within classes and low between them, on which this process capitalizes. The most striking thing about the procedure is that it succeeds. There is no inherent reason why features should form a coherent organization*.

The organizational classes into which features fall are called domains and each feature belongs to one and only one. Domains are structured internally as hierarchies or linear series. The great similarity between domain members based on their common possession of the categorical aspect(s) means they hold some relationship with one another. The nature of this relationship is unspecified; it varies from domain to domain and sometimes within a domain. A domain's categorical relationship (ie sortal or compositional) may impose restrictions of implication or contraindication (negative implication) on its members' capacity to occur together.

Orthogonality identifies the low similarity shared by members of different domains and expresses their mutual incommensurability. This characteristic of domains lays the groundwork for association, the basic relationship in decompositional semantics. It is significant when two features from different domains appear together in a word sense because not all features can do so meaningfully[§]. Because domains are dissimilar, orthogonality also means that when two features co-occur in a word sense, the second feature extends the sense expressed by the first one in a direction different from that already

* this rules out gestalt semantics; an organization of gestalt features would be impossible because each would be unique. It also suggests that indirectly through concepts the feature system is an isomorph of the natural world. In a reflection of the 'natural kind' name for physical objects, independent (ie non-contingent) domains are called 'natural'.

§ only a few of the feature tuples from the power set defined on domains are veridical, actually descriptive of the real world ⁶⁹.

established. When .measure is added to .color, the nature of color is not elaborated further; instead, it is looked at from an entirely different perspective.

Grammar offers a concrete example of orthogonality. Consider the difference between verb forms and kinds. A verb may have as many as five different forms: two tenses, two participles and an infinitive. It may also be able to play a main or auxiliary role in expressing the action occurring in a sentence. These two verb classifications employ different criteria and operate independently and at cross-purposes to each other. Adding .auxiliary to a verb described as a .participle extends the description in a way that adding .present would not. This example illustrates another characteristic of orthogonality: it can (but need not) partition a set of individuals effectively.

Building Systems

A Feature System

The previous parts of this section have described how semantic features are organized on the basis of greatest similarity into structures called domains. Although a key feature of domains is their independence from one another, features participate in other relations which serve to link them, domains and all, into a single moderately connected network in which features are the nodes and the relationships between them (especially those on which domains are based) are the arcs. This network is the feature system.

By itself the feature system constitutes a formalism of sufficient power to describe concepts; this capacity is independent of it being used in any linguistic connection. The relationships inherent in its structure are static and permanent properties of the real world expressed in language by analytic statements about universals, ie "Men are material beings possessing gender". Its few internal restrictions* on what and how features can appear together produce a very flexible representation system which is well suited to the construction of word sense descriptions.

A Linguistic System

This thesis seeks to explore linguistic semantics, and the feature system is no more than a curiosity until it is installed in a larger environment where it can be used to that end. This goal establishes a new list of tasks. A dictionary must be constructed in which word senses are described in terms of features; this serves to connect the feature system to the

* as noted, the criterial relationships on which domains are based establish logical relations of implication and contraindication between certain feature pairs. To adjoin implied features is tautologous; to adjoin contraindicated ones produces a conceptual contradiction or type clash: a meaningless construction.

language context. A second representational formalism must be designed which is adequate to express the meaning of utterances extending over a number of sentences. Finally, an application program must be written which will use a suitable natural language grammar to produce representations of utterance meanings in the specified formalism.

The feature system can be viewed as a part of this larger environment (call it the linguistic system). In the linguistic system words join features as key elements. Words are distinct from their meanings; where word senses in natural language are described in terms of other words, in the system they are described in terms of features.

A word sense description* should contain all and only the features which its users would assume of it. Determining what degree of unanimity is required before a feature is added to a sense description is difficult; Lehrer's findings (1974) are typical of linguistic analytic studies in showing that people include and exclude features from a word's definition on a probabilistic rather than an absolute basis. This suggests that a statistical factor should be associated with each feature used in a word sense description. In practice, the features used in word senses in the linguistic system are ones which in my judgment occur in every instance of the sense. Less canonical features occurring with the sense irregularly and giving rise to probability factors can be identified once enough text has been processed to provide an adequate population of instances for statistical calculations.

Utterances are composed of words in sentences. Although at first glance the sentence seems to be the most important structural unit of language, conceptual expression is actually organized around the clause. Series of clauses operate through sentences to compose episodes which are semantically related wholes. In the linguistic system, an episode's meaning is represented by a network constructed out of the senses of the words which constitute the utterance. Nouns and verb senses are the nodes in this network linked by the senses of function words and through roles which are incorporated in verb senses.

* having made the point in the previous paragraph that words differ from their senses, I will feel free to substitute 'word' for 'word sense' and 'word sense description' from now on.

The dictionary is the set of all words known to the linguistic system. In it word entries are organized first by part of speech and then by meaning. Because each sense description is essentially independent of every other one* (being only indirectly related through common features), the dictionary can be structured as a simple list. It possesses the same permanence, immutability and generality as the feature network and word sense descriptions are also context-free expressions§.

In contrast, both the raw material processed by this mechanism and the representational product it produces are specific and relatively mutable and transitory. Although utterance texts would be discarded once processed, their representations should be added to a monolithic store of all previously processed instances to compose a single interconnected network of truly brathtaking size (call this the language representation; it represents all that the linguistic processor knows about and through language). The linguistic system is therefore envisioned to monotonically increase its 'knowledge' of words and the common sense that can be inferred from their use in utterances. The language representation would be permanent but subject to local revision and constantly changing as it accommodated new instances of utterances.

* actual dictionary implementations (including the one in the demonstration program) make references between the entries for synonyms and inflected words to reduce the size of the total data structure.

§ words do have different meanings in different contexts, but a dictionary entry must be completely unambiguous about which sense is intended before its description can be specified.

Specifying Particular Systems

Now that the general character of systems in this framework has been considered, details of the particular feature and linguistic systems defined in this project can be presented. Although features are the more important* element, relationships are discussed first; they are fewer and serve to define a generalized form for nodes in the network that constitutes the feature system's structure. Discussion opens with descriptions of the general nature of hyponymy, meronymy and association, the most important kinds of relationship occurring in the feature and linguistic systems. This is followed by details of the twelve instances of relationships employed in the three different contexts that obtain here.

Features are logistically awkward. They are too important to summarize and too numerous to discuss thoroughly. The compromise presentation is organized around the 25 natural domains; each domain's presentation begins with a notational expression of its members, their relations and the qualities associated with them (details of the notational conventions appear in Appendix 1). Each populated lower level of a domain is described in a similar fashion. Discussions of all domains and many higher-level subdomains follow immediately after these notational formulas.

Features and the relationships between them constitute an abstract model of semantics in terms of which particular utterances express meanings. Although both are networks, the configuration of the feature system and utterance representations structures are substantially different. All utterance networks are similar however and the networks representing episodes are completely consistent with and a logical extrapolation of the representations of the individual utterances that constitute them. The generalized form of

* here as elsewhere relationships depend on the entities that hold them.

the representations of particular utterances is presented in the last part of this section. Appendix 8 contains a number of specific observations about theoretical and practical issues suggested by the development process.

Relationships

Kinds

Hyponymy, meronymy and association are the three most important relationships used in the linguistic and feature systems. Although hyponymic and meronymic relationships only occur in the feature system, association is also used in word descriptions and utterance representations.

Hyponymy

Hyponymy is the sortal relationship. It classifies a set of distinct and heterogeneous individuals into subsets on the basis of a criterion deemed significant by the classifier which each possesses and which partitions them adequately*. Subsets are mutually disjoint and together exhaust the set to which they belong.

Hyponymy is composed of directed transitive relations. As a result classification can be recursively applied to each of a set's subsets until sufficient organization is imposed on the collection to suit one's purposes. Each stage of classification increases the number of terminal cells in the structure, decreases the population in each cell, and adds more specification to the senses of the individuals it affects.

Unlike relationships which have a particular character, hyponymy is a general principle that focuses on the structure rather than the character of categorization. We classify things as being of different sorts for unspecified reasons. Thus the categorical criterion changes from level to level within a hyponymy as it is articulated. One result is that hyponymy can act as an umbrella organizational relationship, using other relationships such as meronymy as the criterion in particular cells without internal contradiction. The

* ie producing subsets of approximately equal cardinality.

reverse could not occur. A second result is that, as hyponymy has no intrinsic character, in most cases all individuals in one are found in terminal cells and interior cells are empty.

Taxonomies have traditionally ¹² been based on things' sortal relationships and the generality of the sortal relationship suggests that it might be the preeminent classification mechanism. This primacy is evident in the feature system where the only named organization of features, the domain, is based primarily on hyponymy.

Meronymy

Meronymy is the compositional relationship. Two kinds are employed in the feature system: ontological and semantic. Ontological meronymy is the directed transitive relationship by which wholes are made up of parts. Semantic meronymy, a postulate of the model presented here, is the directed intransitive relationship between an individual feature and the qualities by which it is embodied.

Ontological meronymic relations can be further distinguished by the degree of the intrinsic autonomy of their parts. There is a significant difference in the individuality of a bolt and the end of a lever, or that of a finger and the palm of a hand. Allowing for inherent differences between animate and inanimate objects, the former part in each pair is more distinct from its whole than is the latter; a machine can be disassembled into its constituent nuts and bolts, but can one remove the end from a lever? An equivalent difference exists between finger and palm.

Ontological meronomies are not necessarily exhaustive or exact. It is infrequent for all of a thing's parts to be labelled with names. Part limits are arbitrary (where exactly does the palm end and the wrist begin?). As a result of this lack of rigor, a thing may give the appearance of supporting different and overlapping meronymic organizations. In addition, certain problems connected with domains based on meronymy have yet to be resolved. How should agglomerations be treated when their scale seems to change their nature?

Forests are composed of living trees, but because of their size seem to be a feature of the landscape like rivers and valleys. What is the correct domain for snow? It comes from the sky but we have more experience of it on the ground.

Association

Association expresses empirical contingency—that one individual's appearance with another can qualify or extend the other's sense in a meaningful way. This relationship is the fundamental semantic connection, the building block of a compositional semantics. Unlike the preceding relationships, association has an unspecified character; it is not set, not part, not quality, simply an identification of significant co-occurrence. Association has a strength characteristic. Individuals which can but do not need to occur together are weakly associated; strong association asserts that a particular directed link always exists between two individuals.

In either type of association, the particular values which co-occur in actual word senses must belong to the domain named in the relationship. Weak association generally tends to treat a co-occurring feature as a domain, while strong association usually views it as a opaque feature specifying one particular aspect of a sense. A strongly-associated feature is therefore rarely the head of a natural domain; there would be no constraint.

Contexts

Although every feature has a unique character, a great deal of semantic significance arises from interrelationships between features in the feature system. The linguistic system is structured by various types of accidental association, perhaps the most important semantic relationship. One type specifies the exact manner in which features are combined to describe words, two others govern how individual word descriptions are combined into a single network to represent a sentence's meaning, and a final kind determines how

individual sentence networks are consolidated into a single network to represent the meaning of an entire episode.

Between Features in the Feature System

Eleven relations have been defined between features in the system. These relations are permanent connections which inhere in the features themselves and are independent of their use in particular word descriptions. Eight of these relations express the opposite sides of the four relationships of hyponymy, meronymy, qualification and association; because these relationships are all directed, each is composed of a pair of inverse relations embodying its two different orientations. Of the remaining three relations, one expresses a direction of necessary association and the other two identify important metasemantic connections between features which play a role in understanding meanings.

Other relationships between features ¹³ have not been implemented in the feature system for reasons such as their having too high a degree of connectivity, their unproductivity or unimportance in cognitive processes (a matter of opinion). The amount of information a relation conveys is inversely proportional to the number of individuals which are joined by it, so a relation which connects too many individuals conveys little information. An unproductive relation is one that fails to incorporate new features into the system or to add significantly to the sense of existing features ¹⁴. Practical reasons also limited the number of defined relations; because there was a desire to develop both systems approximately equally in all areas, the complexity of other areas played a role in limiting the number of defined relations. So did the limited quantity of computational resources available.

Notwithstanding this, it is desirable to increase the number of defined relations between features. This would increase the interconnectedness of the system and the amount of information it conveyed and could only increase its usefulness.

'members'/'is' (hyponymy): Most features in the system hold a position in a

hyponymic structure. By virtue of this position one feature may have other features as its 'members'; a feature's members partition its semantic space exhaustively and exclusively. Conversely, one feature 'is' also able to be a member of another feature.

Because hyponymy is a transitive relationship, members of a feature's members are also its members; taken altogether, these are a feature's descendants. Similarly the feature to which the current feature belongs may itself be a member of another feature; and these are the given feature's ancestors. As one moves up a line of ancestors, features have increasingly large memberships and decreasingly specific content; the reverse is true as one moves down a line of descendants.

A feature together with its descendants (its members and their members) is called a semantic domain. Each domain bears the name of the feature that heads it. Confusion about whether a name denotes an opaque feature or a domain with a membership does not arise because whatever is true of a feature is equally true of its domain membership (unless overridden in the relations of particular features).

The root and the leaves of the tree representing a hyponymic hierarchy are special cases. A root feature heads a domain but does not belong to any other domain. Such cases seem to arise out of the nature of semantic space itself; domains which are not members of other domains appear to be primary or fundamental. A leaf feature has no members. At present the decision to treat a feature as a leaf and not further subdivide the semantic space it superintends is dictated by space and time considerations rather than by any intrinsic limit. It is unclear whether certain features are natural hyponymic primitives.

'parts'/'partof' (*ontological meronymy*): Features may hold positions in meronymic hierarchies through which they have other features as 'parts' or are 'partof' another feature. A feature's parts give its actual composition. Although ontological meronymy is a relationship of equal power to hyponymy and therefore worth considering as the feature system's primary categorical principle ¹⁵, it has in fact been subordinated to hyponymy and is only employed as a criterion when it is a set's most salient relationship.

The subordination of meronymy to hyponymy in the system means that when a new feature is installed in the system, it is usually placed in a domain on the basis of some relation other than meronymy. Meronymic linkages are therefore usually shorter than their hyponymic counterparts and only the most salient meronymic relations between features are represented in the feature system, ones which are either necessary or likely (default or expected). Thus a .mouth has a .tooth and a .machine does not (although a comb is a sort of .machine which does have .teeth, albeit in a collateral sense). To most people, a .tooth is first of all a thing in the mouth, and a .nail is a sort of fastener rather than a part of a finger ¹⁶.

'has'/'qualifies' (semantic meronymy): Semantic meronymy is the relationship between an individual feature and the attributes which embody it. A feature in a meronymy of this sort 'has' other features which specify certain of its material and immaterial characteristics. 'Qualifies' is the inverse relation; a feature 'qualifies' another feature by specifying some dimension of its particular embodiment. Only certain .types of features can participate in semantic meronymy; .entities and .acts alone have qualities, and a feature which 'qualifies' must be a .quality. Thus we know a red rubber ball by its -redness- and -spherical-shape-. Its surface is -smooth- with a small amount of -lateral-tack- and when we squeeze it is -elastic-. If bounced it makes a distinctive -slapping-sound-, and after sitting in the sun on a hot day it may give off a ripe -burnt-odor-.

Although semantic meronymy is an intransitive relationship, features inherit qualities and other relations from their hyponymic and meronymic ancestors.

'affects'/'accepts' (accidental association): Features may have accidental or weak associative relationships with one another whereby one feature 'accepts' values from another domain, and one domain supplies a value to another and 'affects' it. Features related by weak accidental association can but do not need to affect one another. Thus for a feature to accept .measure in a particular instance of a sense does not necessarily mean that

it will be receive a value from that domain in some manner, it simply means that the feature is measurable. Accidental features of a red rubber ball would be its -size-, the -white-stripe around the middle (and its -dimensions-, and the -deep-nick- in it that it got one day when someone threw it onto a piece of broken glass.

`implies`/`implied` (*necessary association*): In this strong variant of association, a feature that `implies` another asserts that the two are linked in all cases. When the first appears the second must follow necessarily. The inverse relation to `implies` has not been defined in the feature system on grounds of its semantic insignificance; many features can cause a feature to be implied. Thus when .animate implies .discrete it means that no living thing has an .amorphous .substance (amoeba are the rule-proving exception). As this example suggests, an `implies` relation usually restricts the co-occurring feature to a value from a subdomain. It can be argued that restrictions which a feature places on those co-occurring with it, implications about their nature, are part and parcel of the feature's meaning. Association is a directed, intransitive relationship.

`character`: **`Character`** (unimplemented) is a monadic relation expressing the intrinsic nature of the area of semantic space subtended by a quality, ie of the domain associated with an attribute. An attribute can be structured in a complex manner or as a line; a line can be continuous or discontinuous; and a discontinuous line can be binary or multiplex ¹⁷. The most striking characteristic of features in linear domains is the fact that their sense can be completely expressed by giving their position on the line relative to a norm in which most of the semantic content inheres. Speed is a continuous linear quality, while among discontinuous linear qualities, rank is multiplex and mobility is binary.

Complex structured domains are composed of distinct features which each have a unique character (sweet, sour, sharp, salty). They are organized as graphs or shallow graph-based hierarchies in which top level features may combine in new values on second and possibly lower levels. The membership of complex domains is usually identified by

enumeration and members have an incommensurable character (how does green differ from red?). Smell and taste are examples of graph-based hierarchic qualities in that particular tastes or smells can be compounds of more than one feature.

'correlate': This binary undirected relation expresses the relationship between a feature and its semantic mirror image and is limited primarily to .linear .qualities. A feature in a 'correlate' relationship has a counterpart feature which is a mirror image of it seen from the other end of the connection. Although correlates are most typically the polar extremes of their line (hot and cold), certain instances have interior positions on the line (warm and cool). 'Correlate' is unique among semantic relations in having an operation defined on it. Exclusive negation converts a feature into its correlate: not warm is cool. Inclusive negation permits not warm to be hot or cold as well.

Between Features in Word Senses

The feature system defined by the preceding set of relationships is an abstract model of semantics adequate to characterize concepts *per se*. Because the focus of interest here is in how concepts are embodied in and understood from utterances, this abstract model must be embedded in a larger system capable of representing linguistic expressions. Such a system is in turn predicated on a formalism by which the elements of the abstract system can be used to describe word senses.

Most of the relations between features presented in the previous section are necessary. In such relations only the implying feature need appear in a word sense description; all features 'implied' (partially undefined) by it are understood to have their existence automatically entailed or contraindicated by its assertion.

The one relation between features which can be used to build a compositional semantics is the weak or accidental form of association. Nothing can be concluded simply from the appearance in a word description of a feature which can 'accept' values from another domain; just what value if any associated with the accepting feature must be

indicated explicitly. In other words, although it is redundant to list an implied feature together with its implier, the occurrence of two accidentally-associated features together communicates information. Precisely because of this accidental association is the relationship by which word meanings are constructed.

Because all relationships between features in word sense descriptions are accidental associations, they do not need to be named (as the links expressing relationships between features in the feature system must be). The existence and direction of accidental association is shown simply by the concatenation and order of features in word descriptions. Each feature in a group is extended or qualified by the one that follows it*. Qualifying features form a single semantic unit together with the feature or features they apply to; extending features, which move a sense in a new direction, are placed in a semantic unit of their own.

Qualification (*accidental association*): This relation is limited to domains expressing degree or existence. A feature that 'accepts' qualification does not change its essential nature as a result. Qualification only adjusts a feature's measure[§] or its existential basis[‡]. Examples of analogous linguistic expressions are 'quite happy' and 'possibly present'; 'quite' adjusts the degree of 'happy' and 'possibly' specifies the likelihood of 'present' existing. In neither case is the fundamental nature of the affected word altered.

Because a feature is presumed to exist by default but cannot have a default measure the two varieties of qualification are represented differently in word senses. Qualification of degree applies to whatever domain the last feature of the subsequent word belongs to, while qualification of existence introduces the domain .existence and specifies its value.

* from which the corollary follows that any further modification of a group of features is applied to the last item and therefore must be in terms of it.

§ when children string tokens of a qualifier of degree together repeatedly in a statement ('very, very, very ...'), little sense is communicated after the two or three occurrences of the word.

‡ .existence includes possibility, appearance, and subjective experience: 'likely', 'seeming' 'apparently' etc.

As a result degree qualifiers add features to the last pair in a sense and existence qualifiers add one or more pairs to an sense.

Words incorporating qualifying features are typically adverbs. Such words can be applied to each other; 'quite possibly happy' or 'possibly quite happy' are both meaningful phrases. The fact that these phrases have slightly different senses illustrates the importance of features' order of occurrence in a word sense.¹⁸

Extension (*accidental association*): Extension is the paradigmatic form of association. The importance of extended accidental association can be seen in the fact that it is the most common relationship between pairs of features occurring together in sense descriptions. When one feature 'affects' another through extension, it specifies one of the multitude of alternative or additional aspects that their common possessor might have. Specifying .male .gender of an individual described as a .human .sort adds to our knowledge of that individual as well as eliminating the possibility that it is .female (or something .human to which .gender was immaterial). Thus, 'a hard green ball' has separate specifications of hardness, color and denotation, each of which can be changed or removed without affecting our understanding of the others.

Most word descriptions contain a number of features. As the set of described words increases, it is reasonable to assume that each description would contain more features as finer differences are registered to distinguish between similar senses.

Between Word Senses in Utterance Networks

Grammatical rules determine which combinations of words are meaningful in a natural language. It is not surprising to find a relationship in the system that performs an analogous task and governs how different word senses may be combined to represent a sentence. This is accidental association. Thus the same relationship that connects features in word senses, also serves to link together the varied words found in utterances :⁹.

Sentences are represented in the system by network structures composed of nodes

linked by arcs. Verbs and nouns^{*} provide the nodes for these networks. Links are formed between them when nodes representing things (nouns) fill case roles in the nodes representing formal relationships (verbs). Relationships expressed by prepositions and conjunctions also have cases and are brought into the network in a similar way. The process of entities filling roles in relationships is called role tenancy. Both sorts of node can be modified by adjectives, adverbs and articles. Modifiers have their senses amalgamated with that of the node to which they apply. This is called sense consolidation.

Concepts are frequently too large or too complex to be communicated by single sentences. As a result they are expressed by a series of related sentences called an episode. The sentences which compose an episode[§] are related to one another through common tenants of their roles. This same relationship is found in sentence representations; entity nodes are the elements that span individual sentence networks and permit them to be combined into a single integrated network that expresses the sense of a series of sentences. The relationship by which the same entity fills roles in different sentences is called thing identity[‡] and is a third type of accidental association occurring in the linguistic context.

Sense Consolidation (*accidental association*): Whether it occurs between different words or between different aspects of the same word[¶], accidental association accidentally associated words as if they were a single word and add the sense of the word which 'affects' to that of the word which 'accepts'; this is sense consolidation. Since the former is an adverb or adjective and the latter a noun, verb or equivalent, the effect of this operation is to fold modifier senses into the symbols that represent the heads of their phrases. This consolidation takes place in a manner consistent with the particular type

* and pronouns, at least temporarily until their referents are found.

§ whose meaning by definition is greater than that of the sum of its parts.

‡ awkward though it may be, 'thing identity' is more euphonious than 'entity identity'. A key process in constructing episode networks is identifying which entity nodes in a new sentence match existing ones in the episode network.

¶ the same sense can be often realized in a single word or a phrase composed of a number of words.

of modification (qualification or extension)*.

Entities which participate in a number of verbs (.acts) in different sentences typically have many modifier senses consolidated with their nodes. Senses which pertain to the same attribute may contradict one another because things change over time[§] and also because language utterances cannot be guaranteed to be correct. Here is one area where inter- and intra-word accidental associations differ; features included in a node's dictionary description are generally[‡] expected to remain veridical because they are after all what the word names literally.

Role Tenancy (*accidental association*): Sentences are composed of words that name things and words that name relationships. Things do not associate with one other directly but only through relationships which bind them together to form coherent sentences. Relationships arise out of generalizations on the multitude of associations that things in the real world enter into and are thereby dependent on the things embodying them. It is frequently useful to ignore this fact and to imagine that semantic relationships have the independent existence that our abstractions of them suggest. According to this pretence, their component relations exist between generalized things called roles.

Roles are different than things. They are semantic variables which exist independent of any entity and have only limited meaning. The representation of an actual utterance is built by filling the roles in its particular relations with suitable instances of things²⁰. However the sense of any tenant of a role must be consistent with the role's own sense. Therefore no matter how limited it may be, a role's sense does restrict the sorts of things that can fill it; the more specific its sense, the greater the restriction. This effect

* if the type of accidental association is qualification, the word must be an adverb; otherwise it can also be an adjective or article.

§ some aspects are intrinsic (red is always red) while others are contextual (early, cheap). Quirk and Greenbaum (1973) call this distinction inheritance.

‡ even this is problematic; eventually 'boy' becomes 'man' and the aspect of -young- is no longer true of the individual. See Appendix 8 for a more detailed discussion of change.

is generally known as selection restriction ²¹. Although distinct entities and roles usually match one-to-one in a relation, it is possible for the same entity to fill more than one role in a single relation, ie in reflexive constructions.

Thing Identity (*accidental association*): The way in which sentences are associated with one another to form episodes is most easily understood from an ontological perspective that views things as persisting over time and entering into sequences of relationships of varying but shorter durations. Syntactically, this perspective views each sentence* in an episode as expressing a different relationship between new and existing entities, each of which can fill roles in more than one sentence. Some entities play peripheral roles and appear only in a single sentence but for a series of sentences to constitute an episode[§], every sentence must be linked to another by at least one entity. The network must be connected. I have a spatial image of this situation: a group of strands of varying lengths (things) tied to each other laterally every so often by threads (relationships). The resulting braid is an episode.

Episode construction poses the problem of recognizing identical entities[‡] in different sentences. Identity can exist between nouns or a pronoun and its referent. Identity between nouns is complicated by the fact that the same thing can be referred to by synonyms of greater ('boy' for 'child'), equal ('lad' for 'boy') or lesser ('child' for 'boy') specificity. An effective synonym can even be from a different syntactic category ('young man' for 'boy'). The problem of recognizing the same entity in different sentences is complicated by the fact that an utterance can nominalize relationships on the fly in the course of its development and subsequently refer to them as if they were entities. A sign of

* strictly, each clause. Words other than verbs can also convey relationships: participles, headless adjectives etc.

§ usually episodes are larger than a single paragraph. Although paragraph structure can mirror episodic structure, it cannot be guaranteed to do so. For this reason the discussion ignores paragraphing.

‡ from a network perspective, the task is to determine the point or points at which a new sentence network is to be attached to the existing set of nodes.

this practice is the presence of causal relationships in an utterance's semantics and the use of demonstrative pronouns and abstract words like 'fact' in its syntax ("That John worked on Monday is a fact." etc).

Domains and Features

Although natural domains are equivalent in theory, in practice their individual significance and utility varies. For instance .type and .class are key domains, having the most members and the greatest frequency of use in word senses. They appear to be natural primary indexes to the way we organize concepts.

On the other hand certain domains only make sense in terms of other domains or features; only .animate things have .gender. These contingent domains have a highly specified sense because they share that of the domain on which they are predicated. Attributes are all contingent because they depend on the things which possess the qualities that they represent.

Other domains are the exact opposite. Although they have relatively little meaning and what little they do have is interpreted differently depending on the circumstances, domains like .orientation, .perspective and .existence are employed in many different semantic contexts. These protean domains seem either to structure senses in some fundamental way or to provide a framework that allows meanings to be formulated.

.Syntax and its contingent domains .mood, .Person, .voice and .constituent are relatively distinct from the rest of the system. They exist only to express the morphological and syntactic particulars of how words and sentences realize meanings and are not used once parsing is concluded.

The system recognizes 25 natural domains containing 305 features (the number changes frequently). These domains do not exhaust semantic space by any means, nor do individual domains completely express the detail of the particular parts of it they map to. In some cases categorizing has been carried out unevenly, deeper in one area than another, in order to develop the semantics of particular test cases or areas of interest.

The size of the feature system precludes anything other than a summary presentation. Only the upper levels of natural domains are discussed, and although they are in fact domains, attributes are treated as features and listed in the notational expressions that summarize the semantics of the domain they refer to. Discussions have a degree of detail that varies with the importance of the domain and the obscurity of its semantics.

203 words are currently described in terms of features. A number of these words are synonyms and an approximately equal number have multiple senses, verbs and prepositions in particular. Many words give rise to inflected forms with somewhat altered senses. These factors tend to cancel each other out and the ratio of coded words to coded senses is approximately 1:1. Because features are more perspicuous than words, the situation of the set of describing features being larger than the set of described words can be expected to reverse itself as both sets grow.

.type: .thing, .relationship {ABSTRACT}

This is the domain of ontology which classifies features according to their essential nature. Types are simply distinguished individuals; they can be either .things or .relationships. These two classes express a very common model which has analogues in all fields of knowledge: finite and distinct entities (.things) are joined to one another by connections (.relationships).

a) *.thing: .sort, .part, .ingredient, .alias, .naming, .quality NATURE
(DENOTATION)*

.Things are individuals which persist and have a coherent and unitary organization which can be perceived or imagined. They correspond quite closely to nouns and pronouns. Hyponymic .things (.sorts), the most important kind, are things which are independent members of some class. Three features in this domain have a meronymic relation to .sorts; two of these express .sorts' composition in the real world (.parts, .ingredients) while .qualities are the semantic constituents of .sorts. .Namings and .aliases are the respective

conversions of the non-thing .relationships and .qualities to .things in order to expand their linguistic utility by allowing them to be treated as entities.

1) .sort: *DISCRETE*

Things which are .sorts are entities considered to be discrete and independent. Because hyponymy is the fundamental categorical relationship, a .thing's .sort provides its canonical definition and identifies its position in a preeminent organization or taxonomy. Hyponymy is the fundamental categorical relationship because of its flexibility and rigor, and because it does not impose semantic constraints on the things it classifies. Hyponymies are structurally defined; the criterion on which they are based is unspecified and varies from level to level depending on whatever particular aspects are the most semantically salient in a given context. As classification proceeds further and further, classes accumulate aspects by inheritance and their members acquire semantic descriptions.

2) .part:

A thing's most important dimension may not be its essential being but how it serves to make up another thing (.part). This is the relationship of ontological meronymy. Parts have their natures subsumed in that of the whole they compose. Thus whatever applies to 'leaves' and 'branches' is understood to apply to 'plant'. Notwithstanding this a .part has its own distinct and separate nature and it can always be treated as a .sort by isolating it from its natural context. .Parts can themselves have .parts. Although a handle is always a part of some thing, we can talk about pulls, knobs and other sorts of handles, and a handle might have an escutcheon, a bar for gripping and bolts to attach it to something.

3) .ingredient: *AMORPHOUS*

This subdomain identifies the amorphous material which ultimately makes up all things. A thing's .ingredients can be construed on more than one level; .ingredients can have .ingredients. Animals are composed of bone, muscle, fat and meat which in turn are composed of hydrocarbon compounds and of subatomic particles. Although .ingredient is compositional like .part and has a related nature, these two subdomains do not contradict;

.parts are composed of **.ingredients** and **.ingredients** are embodied in **.parts** (and **.sorts**). However as is the case with many semantic descriptions, this one does not bear too close scrutiny. Sand is an ingredient although its grains are discrete; a crowd is composed of unspecified but certainly individual people.

4) *.alias: DISCRETE*

Things frequently take on new meanings in particular situations. These adopted meanings are called **.aliases**. An **.alias** is like a role in an abstract relationships; it is a position in a mesh of semantic relations which exists independent of any individual thing and which can be occupied only by particular sorts of things. We can imagine and speak of a King of France although no living person fills that role, and we can also be certain that any King of France would be a human being. Aliases are very numerous.

A thing which assumes an **.alias** takes on an additional sense without surrendering any part of its canonical (literal) sense. Although an **.alias** cannot contradict any part of a canonical sense, it does not necessarily make use of all of it. 'Participant' is a **.alias** which does not even require its holder to be human.

.Aliases once assumed can be temporary or permanent. A sleeper is only one as long as he sleeps, but a mother is a mother forever. Because they are contingent on very particular situations, aliases frequently have highly specified senses expressing accidental rather than necessary aspects of the thing that assumes them. Thus not all women are mothers. In other words **.aliases** differ from **.sorts** in communicating extrinsic rather than intrinsic meaning. **.Roles** are a similar but more restricted kind of **.social .class alias**.

5) *.naming:*

Relationships are not things, and we must reify them in order to treat them as if they were. The result is a nominal in linguistics and a **.naming** in the feature system. When it is reified, the **.relationship** 'walk | walks' (as in "John walks" and "We all walk regularly"), becomes 'walking' or '(a) walk'. These two forms differ in their denotation. 'Walking' is unambiguously **.universal**, while 'walk' can be either **.universal** or **.specific**, ie "A walk is

pleasant on a sunny afternoon" and "I think I'll go for a walk"*. Sometimes no verb or only an infrequently used one denotes a relationship that give rise to a more commonly used .naming, ie `birthday`, `breakfast`.

6) *.quality: (.MEASURE)*

Attributes, our concepts of the components of the sensory impressions by which we perceive things, are a type of relationship and so must be reified like .namings to be treated as things (.qualities). Semantic meronymy is the relation between a .thing and the particular .qualities it possesses. Like other aspects .qualities can be either necessary or accidental; necessary ones are part of a thing's description, and accidental .qualities are attached by .specification. .Qualities are usually manifested as adjectives and adverbs.

A .quality differs from a .naming in its universal and permanent inherence in the things to which it pertains. Compare `dog` and `black` (.specification) to `dog` and `barks` (.act). A black dog is so always and everywhere, but a dog barks only at particular times and places. .Qualities may be only relatively permanent; a brunette may whiten as time passes and a hot drink eventually cools.

b) *.relationship: .act, .causation, .set, .specification, .position, .possession, .component, .arithmetic, .comparison (.is, .has)*

Although we can imagine them in isolation, things in the real world are embedded in surroundings and must be considered in conjunction with other things. .Relationships between things are as important and as diverse as .things themselves (more important, if we accept the linguist's view that the verb, expressing relationship, is the most important word in a sentence).

The nine primary sorts of .relationship which knit the linguistic system together all possess the general characteristics of contingency on things, roles, and restrictions§ on their tenants etc. Although most relationships are expressed by verbs, prepositions and

* the cognate form `walker` is an .alias for the agent of a walking .act.

§ inappropriate tenancies called type clashes such as *"green clouds sleeping furiously" are nonsensical.

conjunctions, a few have other realizations; .possession can be shown by the suffixed contraction ``s`.

- 1) *.act: .condition, .motion, .process, .action, .change { .activity, .energy, .potency, .risk, .constraint, .capability, .readiness, .effort, .speed, .force} .TIME (.CASE, .CLASS)*

.Acts are the largest, most important and most overt class of .relationship. An .act's sense conveys or implies a good deal of concrete semantic detail. Consider the verbs `walk`, `write`, `help`. Each suggests images of various aspects of an act or situation: the parties involved, their attitudes, the instruments used, movements, situations, details. These images are significantly richer and more numerous than those suggested by any of the eight other types of .relationship. .Acts are also distinguished by their capacity to incorporate other relationships such as .position, .causation, .specification in their senses.

.Acts are expressed by verbs. They are called formal relationships because verbs are the focus of clausal syntax and give a sentence its form.

- a> *.condition: (.EXISTENCE, .CAUSATION)*

.Conditions focus on a thing's state and have an unspecified duration. Although static rather than active, they are classed here with .acts because they employ .cases. A .condition highlights the existential aspect of a situation, and words of .condition .type include the copulative verbs which qualify that aspect. `Seems` and `feels` are examples; `seems` focuses on a situation's appearance rather than its essence while `feels` conveys subjective experience than objective reality. A significant number of static acts are simply syntactic vehicles for .specification relations: "the boy is tall" is semantically equivalent to `the tall boy`.

- b> *.motion: { .mobility, .propulsive} .MATERIAL*

.Motions involve material things and occur in tridimensional space. Although other .acts may involve movements in a subordinate way, .motions highlight a thing's spatial displacement.

c> .process: .transfer, .creation, .destruction, .transformation

.Processes focus on modifications effected on things by external agents. .Creation and .destruction describe increases and decreases in a thing's size or organization, while .transformations modify a thing's internal character. .Transfers highlight changes in a thing's external situation; if this involves physical motion, it is in a subordinate capacity. 'Give' and 'sell' are examples of .transfers.

d> .action: .search, .organize, .select (.PURPOSE, .ORIENTATION)

An .action (unimplemented) differs from a .process in not affecting an object in a material way, and from a .motion in not emphasizing transition through space. .Acts of this sort often involve an intellectual component and require a human .agent as a result. Most actions have partial or exact opposites, ie searching is opposed to hiding and to losing.

e> .change

.Changes are modifications in a thing caused by hidden agents which may be factors operating within in the thing itself. Words incorporating .change include copulative verbs like 'become' and those expressing the homeostatic activities of living creatures such as eating, sleeping, sweating etc.

2) .causation: .voluntary, .involuntary (.cause, .effect, .inhibition, .goal, .instrument){.sufficiency, .capability, .obligation, .necessity} (.EXISTENCE) { .EXISTENCE}

.Causation considers contingency in the real world—how things cause or are caused by each other—and provides an alternative to .acts as a basis on which to organize a narrative. .Involuntary causal relationships are necessary happenings between inanimate things, ie ice melting in sunlight. .Voluntary .causation is deliberate action by conscious subjects.

This domain's meronymic values are more salient than its hyponymic ones. .Cause and .effect are correlates expressing the two sides of causation; .inhibitions are negated causes, factors which prevent effects which might otherwise occur. .Goals are effects sought by human agents; .effects follow directly and immediately from .causes, while a

.goal may be the outcome of a lengthy series of linked .causes and .effects. .Instruments are intermediate things in causal relationships; ie the wind may break a window by blowing a branch against it.

Note that this domain has a bidirectional accidental association with .existence. This expresses the different conceptions of asserting a cause's existence (.cause .existence) and of something's existence being caused (.existence .cause).

3) *.set: .conjunct, .disjunct, .negjunct (.EXISTENCE)*

Set relationships mediate how things fill roles in relationships. .Conjunct installs all of its tenants, who are multiple, in a role. It is exactly equivalent to 'A & B & C ...'. The .disjunct .relationship indicates that one and only one of its plural tenants occupies a role. Natural language routinely employs the exclusive (strict) sense of 'or' rather than its inclusive (relaxed) one; when one of a series of disjointed individuals is selected to hold a role, it is understood that no other individual may hold that same role. .Negjunct incorporates an expression of deMorgan's law. It is equivalent to the negation of a .conjunct, and 'neither A nor B' is the same as 'not A & not B'. Set relationships are undirected, one of the few instances of an undirected relationship in the system.

4) *.specification: .topic, .identity, .style, .substance (.EXISTENCE, .MEASURE)*

.Specification ²² is semantic meronymy, the relationship between individuals and the attributes which make them up. Qualities are the vehicle by which we know things and relationships, so it would be reasonable to call semantic meronymy qualification. However that term is used in a different connection to describe relations between features in word meanings. More important, .specification is more descriptive of the role that this relationship plays. The relationship between individuals and their attributes is so intimate that it is not surprising to find a number of different kinds of .specification. Because any .type has so many attributes, the mention of one in an utterance is assumed to be significant.

a> .topic

.Topic is the theme of some form of communication, whether written or uttered, drawn or photographed.

b> .identity

.Identity maps from classes to individuals. Although a thing's .identity is usually successfully indicated by its proper name, this may be less straightforward than appears at first sight ²³.

c> .style

An act's .styles are whichever of its accidental aspects are significant in a given context, in particular ones mentioned in an utterance. Style is usually realized by modifiers and to a lesser extent by prepositional phrases. Thus in 'the prince's haughty speech' or 'the prince spoke in a haughty manner', the fact that haughtiness is referred to is expected to have some bearing on one's understanding of the rest of the utterance.

d> .substance

This is the relationship by which .ingredients compose .things. As phrases like 'the man of steel' and 'the wooden box' suggest, .substance is expressed by modifiers and prepositional phrases.

5) *.position: .spatial, .temporal, .state, .metric*
(.DIRECTION, .PROXIMITY, .MEASURE)

Relationships locating things relative to one another are a significant part of many utterances. Usually both things in a positional relationship are mentioned, but in some cases a thing's position is measured against an understood thing rather than an explicit one. In "that conduct is beneath John", 'conduct' is measured against a concept of what John's honor should be. .Position relationships are realized primarily by prepositions and conjunctions; they are in fact the largest and most elaborate class of prepositional relations. Four sorts have been identified.

a> .spatial: .MATERIAL (.COMPARATIVE, .PLANE)

These relationships locate things in space. Prepositions distinguish three spatial contexts in terms of which things are located. A context is implied by the particular values which a .position accepts from other domains.

The first context is relative to the thing itself and has aspects of interiority/exteriority and distance. Because it views space as unidimensional, it is not surprising that this context reappears in other linear domains like time, number and general measure where it is expressed by .proximity.

The second and third contexts both operate in tridimensional space. The second context locates things relative to one another by adding an external coordinate frame against which both objects are measured. The third context shifts the coordinate frame outward to include the onlooker and locates things relative to each other by locating them relative to him. It is invoked when the utterance fails to provide enough information to allow the second context to be employed. Unless otherwise indicated, the onlooker is assumed to be a person directly concerned with the utterance, its author or recipient.

Examples are the best way of make these abstractions understandable ²⁴.

b> .temporal: IMMATERIAL (.COMPARATIVE)

.Temporal relationships are perhaps the most complex case of a linear relationship^{*}; a thing's temporal extent or duration can be represented as an extent on the time line. The first spatial context is applicable here. Things occur within and outside of the extent of other things ('while'); if outside, they can be in contact with ('as soon as') or near ('before', 'after') the extent of other things. More complex temporal relationships[§] like partial overlap cannot be communicated in a single prepositional phrase.

c> .state: IMMATERIAL (.EXISTENCE)

Although .state relationships have as much similarity to .style as they do to .position, they

* in fact the most easily understood modelling of time relationships represents them as a series of parallel lines rather than a single line.

§ see (Allen & Kautz, 1985) for a nice workup of temporal relationships.

have been included here because they share much of the positional spatial metaphor. States are imagined to be things which one can enter or leave and be inside or outside of. Thus a person is 'on duty' and 'in love', and can move 'out of risk' or 'into danger'.

d> .metric: (.ABSTRACT, .COMPARATIVE)

.Metric is a general domain that incorporates all other positional relationships based on scalar attributes such as weight, temperature, cost etc. It includes space and time when these dimensions are simply being measured; 'at nine o'clock' is a temporal relation while 'for three hours' is a metric one. .Metric expresses comparative .position in non-physical spaces such as organizational and social structure ('staff beneath me', 'down to a lower class'). It is also used to locate its object relative to norms in particular situations ('on the average', 'above the passing grade') and in general ones involving honor, legality, popularity etc ('actions beneath him', 'high in public esteem').

6) .possession: (.EXISTENCE)

The .possession relationship includes the distinct but related aspects of physical possession and ownership. This coverage permits three interpretations: physical possession without a suggestion of ownership, ownership without a suggestion of physical possession, and ownership with physical possession. Contextual information is required to distinguish which sense is intended.

7) .component: (.EXISTENCE)

The relationship between individual things and their parts is ontological meronymy. It has already been discussed at length under .part above and the 'part'/partof' relationship in the previous section. .Component allows explicit expression of a compositional relationship.

8) .arithmetic: .+, .-, ., ./*

This relatively minor relationship is nonetheless a distinct one. .Arithmetic relations are limited to .numerals, linguistic symbols understood as numeric data.

9) .comparison: (.EXISTENCE)

Two things may be compared either in general or in particular ways. 'Like' and 'as' are

used in global comparisons in which the likeness is unspecified (but considered to be obvious to the both parties to an utterance). Comparative and superlative senses of particular modifiers invoke .comparisons of particular dimensions of their two objects (unimplemented).

.class: .material, .immaterial { .age } (.TIME, .MEASURE) (.ABSTRACT)

This is the most important and most comprehensive* hyponymy, the fundamental classification of things[§] by sort. It divides .things into those that have a tangible, physical basis (.material) and those which exist only in our minds (.immaterial). The former are natural kind things; we form concepts of them directly from our sensory experiences. The latter are things which have no physical basis in the real world; we form concepts of immaterial things by generalizing from our perceptions.

Semantic consistency—the existence of some sort of similarity between two individuals—seems to be rooted in an individual's .class in the natural kind hierarchy. Individuals which are consistent in domains that have a physical expression (like appearance or manner) invite employment in figurative speech.

a) .material: .animate, .inanimate { .axis } (.PLANE)

.Material things are subdivided into those which are alive (.animate) and those which are not (.inanimate).

1) .animate: .animal, .plant

.Animate things are subdivided into .plants and .animals. These entities are then further classified according into the traditional botanical or zoological taxonomies. Animals have a larger set of potential relations than do inanimate things because they are alive, and human beings in turn have more varied relations than other animals because they are conscious. 25

* through its .abstract subdomain class includes all other features.

§ although strictly speaking they are all .immaterial, .relationships partake of .class specifications indirectly through the restrictions they apply to the tenants of their roles.

*a> .animal: .mammal, .bird, .insect, .reptile, .fish, .MOBILE,
.PROPULSIVE (.GENDER)*

1> .mammal: .human

b> .plant: .IMMOBILE

2) .inanimate: .natural, .artificial

.Inanimate things divide most obviously into ones which exist in nature (.natural) and those which are produced by conscious human effort (.artificial).

a> .natural: (.space, .star, .planet)

.Natural things (unimplemented) are not as distinct as .animate and .artificial ones. A living thing's vitality generally defines its limits, the creature ending where its vitality does. The deliberateness of manufacturing processes provides most .artificial things with exact edges which serve to separate them from the natural environment.

Boundaries in the natural world are often subtle and hard to discern. It is difficult to tell where a valley ends and a mountain begins, and the exact margins of clouds and lakes are only appearances. When boundaries are exact (ie a rock's edges), they seem random and therefore insignificant. These facts promote the folk concept of Mother Nature*, the inanimate world seen as a single thing.

However viewing the .inanimate as a single thing does not leave anything to classify, so meronymy rather than hyponymy must be applied to subdivide this unity. Classifying .natural things on a meronymic basis gives us both a starting point—the entire natural universe—and an categorical mechanism—parts. 26

1> .planet: .world, .moon

a: .world: (.land, .sky, .ocean)

1: .land: .island, .continent (.mountain, .river, .field, .pass)

2: .sky: (.wind, .rain, .cloud, .snow)

b> .artificial: .mechanism, .construction, .product (.PURPOSE)

.Artificial things are made by human beings directly through hand labor or indirectly by

* a unitary if not anthropomorphic view of the inanimate environment is pervasive in our thinking. When we talk about weather, who is the agent in "It rained again today"?

machine. *.Mechanisms* and *.constructions* (unimplemented) are things we make in furtherance of our goals, while *.products* are things that meet our goals directly. *.Mechanisms* are simple or complex tools frequently made of metal and with moving parts. *.Constructions* are static assemblies; buildings, bridges, roads are good examples.

1> *.mechanism: .vehicle, .tool, .machine*

This domain is not exhaustively specified. *.Vehicles* move things from place to place. Tools are inert unless used by people and they can generally be held in ones' hands. *.Machines* are frequently active (motorized), of indefinite scale, and often operate unattended.

2> *.product: .food, .drink*

a: *.food: .fruit, .sweet*

3> *.construction: .abode, .roadwork, .container*

a: *.abode: .house, .town*

b: *.container: .box, .bottle (.door)*

3) *.axis: (.PLANE)*

b) *.immaterial: .abstract, .person, .social*

Some things are not physically embodied in the real world but exist only in our minds. These notions are not known through direct experience but only through introspection; because *.immaterial* things are abstractions they cannot be guaranteed to be veridical. They express objective reality (*.abstract*), subjective reality (*.person*) and the reality which is shared with other people by convention (*.social*). The first sort of concept is typically acquired by ratiocination or formal study; ie we can learn them from books. The second sort are mental experiences which impinge directly on our consciousness. Concepts about society and *.social* relations are both abstracted from our perceptions and informally taught.

1) *.abstract: { .METRIC } (.TYPE, .CLASS, .CASE, .GENDER, .TIME, .MEASURE, .SUBSTANCE, .NATURE, .DENOTATION, .EXISTENCE, .ORIENTATION, .NUMERAL, .PURPOSE, .DIRECTION, .PROXIMITY, .PLANE, .PERSPECTIVE, .CHARACTER, .PERSON, .SYNTAX, .CONSTITUENT, .VOICE, .MOOD)*

.Abstract is the domain of impersonal intellectual abstractions, ideas which can be communicated to and understood by other people. As the list of accepted domains suggests, all other members of the feature system can be seen as things-in-themselves rather than as descriptive of individuals in the real world.

2) **.person:** (*.mind, .emotion, .will, .sense, .body*)
{.life, .consciousness, .vitality, .appetite, .health}

For simplicity it is assumed that only human beings have personality and consciousness. Although many higher animals act in ways that suggest they have at least limited consciousness, the extent and complexity of such a consciousness must be markedly inferior to our own.

.Mind, .emotion, .will etc are considered to be parts rather than sorts of **.person**. This presumes an organic model of the human psyche; an alternative view would see some of these things as behaviors which a person conducts serially.

a> **.mind:** (*.comprehension, .attention, .recognition, .reasoning, .memory*)
{.rationality, .intelligence}

The same organic presumption makes our various mental faculties **.parts of .mind**.

b> **.emotion:** *fear, .anger, .happiness, .anxiety, .love, .pride*
(.ORIENTATION, .DEGREE)

.Emotions are exemplary attributes. The **.character** of each is a continuous line including one or more pairs of polar opposites and a neutral or quiescent central state. For brevity, only one pole of each (unnamed) sentiment is given ²⁷.

c> **.sense:** *.sight, .taste, .smell, .touch, .hearing, .visceral, .esthetic*

This domain categorizes human sensory modes. Each accepts values from the **.substance** subdomain it superintends. **.Visceral** expresses sensations that arise from within our own bodies; we are describing visceral sensations when we speak of feeling well, ill, hungry etc. Although vision is our preeminent sense, **.touch** seems to distinguish far more of a thing's attributes than **.sight** but this is only because features primarily superintended by

vision (ie those expressing a thing's shape or physical configuration) are so numerous that they are considered separately from *.substance*.

This classification suggests that the appreciation of beauty and the aversion from ugliness is primarily a sensory experience to which an emotional component is attached.

1> *.sight*: (*.COLOR*, *.BRIGHTNESS*, *.OPACITY*)

2> *.taste*: (*.FLAVOR*)

3> *.smell*: (*.ODOR*)

4> *.touch*: (*.TEXTURE*, *.PLASTICITY*, *.ELASTICITY*, *.HARDNESS*,
.HEAT, *.WEIGHT*, *.HUMIDITY*)

5> *.hearing*: (*.SOUND*)

3) *.social*: *.institution*, *.role*, *.occasion*, (*.reputation*)

This is the subdomain of interpersonal relations. Features here exist by convention, ie by common agreement among people. *.Institutions* are social entities like corporations considered as if they existed independently of people. *.Institutions* emulate physical individuals in possessing *.universal* or *.specific* denotation ('companies in the forest industry', 'General Motors Corporation') and in using capital letters to individuate their name. A *.role* is a guise that a person can adopt. *.Occasions* are occurrences involving groups of people like parties, meetings or sports events.

a> *.reputation*: (*.MEASURE*, *.ORIENTATION*)

b> *.role*:

People play *.roles* when they become doctors, policemen and teachers. A *.role* differs from an *.alias* in applying only to human beings and in tending to persist between different contexts (thereby usually lasting longer than a *.alias*). *.Roles* are established by social convention; *.aliases* by than empirical evidence. Because *.social* matters are already quite elaborate abstractions, *.roles* convey greater semantic content. 'Doctor' is a *.role*, 'audience', 'lever' and 'obstruction' are *.aliases*. In other respects *.role* and *.aliases* operate in the same way.

c) *.age*: (*.TIME*, *.MEASURE*)

.case: .subject, .object, .indobj {ABSTRACT}

Cases are roles which abstract things play in relations and which are filled by instances of things. Alternatively cases are semantic variables* which have the minimum sense required to embody a particular relationship²⁸. In contrast to some designs, this semantic theory keeps the number of cases to an absolute minimum²⁹. Thus, because cases are attached to relationships and relationships operate across domains, the tenant of an 'above' relationship occupies the same case whether temporal or spatial superiority is being expressed.

A .case's portion of meaning is added to that of whatever instance fills it in the same way that an .alias's sense supplements the canonical meaning of the thing that occupies it. Cases play a role in linguistic processing by enforcing semantic restrictions on their tenants, ie the subject of 'think | thinks' must be .human[§]. Although many relations hold between binary tuples of things, and others such as the relationships expressed by copula and intransitive verbs are unary, a significant number of relations permit or require a third participant. .Acts for instance can have .subject, .object and .indobj (indirect object) cases realized positionally or through prepositional phrases. Although almost all relationships expressed by prepositions are binary, 'across' and 'between' require ternary tuples³⁰. Infrequently a single instance of a thing may hold two different cases; an example is a formal relationship which accepts a reflexive pronoun, ie "he harmed himself".

a) .subject

The .subject in a formal relationship or .act mandatory and may be filled positionally or through a prepositional phrase ('by'). Semantically, it is the agent of the action, the one who carries it out in the most general sense. Other relationships realized by prepositional

* the only variable defined in the system.

§ literally .human, metaphorically otherwise; ie 'the computer thinks'.

phrases also have mandatory *.subjects* which are the thing which is external to the relation, ie 'boy' in 'the boy on the bench'. Unary relationships have no subject.

b) .object

Depending on the nature of the *.act*, an *.object* can be optional or required. This case can only be filled positionally. In semantic terms, it is the thing with which the action is concerned. Reflexive relationships involve a single tenant of both *.subject* and *.object* cases. In other types of relations realized by prepositional phrases, the *.object* is mandatory and the surface object of the preposition, ie 'bench' in 'the boy on the bench'. *.Objects* are the one case in unary (qualifying) relationships: 'the red apple'.

c) .indobj

The third participant in a formal relationship can be either an optional or required case and is filled either positionally or through a prepositional phrase ('to' or 'for'). Although an *.indobj* is in general the thing with which an *.act* is indirectly concerned, its reading is inherently ambiguous; in particular, positional indirect objects cover two entirely different relationships: as 'to', the spatial *.to* relation; as 'for', the causal *.effect* or *.goal* relation.

.slot: .is/.has (.CASE)

Slot is a metasemantic domain used to indicate the direction of a relation. *.Is* expresses the direct sense of a relationship, *.has* the inverse sense. Taking *.composition* as an example, *-partness-* is the direct sense and *-having-parts-* the inverse one; thus the component in a *.composition* relationship 'is' a part, while the whole 'has' parts.

.gender: .male, .female ANIMAL {.ANIMAL, .ABSTRACT}

.Gender indicates the sex of an *.animate* entity. It is usually used of things which are *.class .animal* or higher ³¹. Rather than specifying a *.neuter* feature for sexed things which have an unspecified gender, *.gender* appears in such word's definitions without specification, ie as a feature rather than a domain head.

.time: *.past, .present, future {frequency, .duration}*

IMMATERIAL {ABSTRACT} (.MEASURE, .PERSPECTIVE)

.Time indicates an individual has a temporal aspect. It is used most frequently to specify an .act's position on the time line relative to the moment of interest (usually the current moment). An .act's temporal sense can be extended by its progressive aspect, represented by .perspective. .Intensive .perspective indicates that the moment of interest lies within the duration of the given .act and that the .act is therefore viewed as ongoing or progressive. .Acts with an .extensive feature are viewed from the outside and are considered perfect or temporally complete in themselves. "He is walking" and "he walks", "he was walking" and "he walked" illustrate the difference.

The three temporal features of .past, .present, .future have been allowed to double up in a few cases (.past .past, .past .present) to express the perfective aspect; this is one of the few places where instances of the same feature qualify one another usefully. The perfective aspect considers an .act's temporal position relative to some point in time other than the current moment. In "he has been gone an hour", 'gone' is considered relative to the current moment; in the perfective "he had been gone a hour when they came", 'gone' is considered relative to 'came' and the current moment is at some unspecified point in the future of both acts.

a) frequency: (.MEASURE)

.measure: *.count, .degree, .comparative (.unit){exactness} {ABSTRACT}*

.Measure is the domain of quantity. Its three values deal with measure within unity (.degree), measure within multiplicity (.count) and the measure of two things relative to one another.

a) .count: .number, .sequence (.NUMERAL, .COMPARATIVE) {.DEGREE}

Situations involving counting are divided into ones where things are ordered (.sequence) and ones where they are not (.number). Cardinal and ordinal numbers express this

distinction in natural language, ie 'one, two, three ...'; 'first, second, third ...'.

1) .number: .single, .multiple

.Multiple expresses the sense that an individual is more-than-one. Sometimes language further distinguishes between cases involving just two individuals (.pair) and ones with more than two (.plural). The prepositions 'between' and 'among' key on this difference.

a> .multiple: .pair, .plural

b) .degree: .exact, .approximate (.COMPARATIVE)

.Degree deals with proportional measure, which is the measurement of some part of a thing 32. Two kinds of .degree are distinguished. .Approximate .degree deals with inexact proportion and .exact .degree deals with numerical proportion, ie fractional or percentage measure.

1) .exact: (.COUNT)

2) .approximate: .none, .minimal, .little, .some, .half, .substantial, .most, .great, .total

Although the assignments are arbitrary, typical percentages for the eight values of .approximate .degree might be: .none, 0%; .minimal, 5%; .little, 15%; .some, 30%; .half, 50%; .substantial, 60%; .most, 75%; .great, 90% and .total, 100% 33. Gaps between values are spanned by the use of qualifiers like 'almost' and 'very' (converting imprecise terms to numerical values is awkward). Applying outward-ranging qualifiers to the end values in this line (ie 'less than none', '?very total') results in logically inconsistent readings; constructions of this sort must be interpreted figuratively.

c) .comparative: .greater, .greatequal, .equal, .lessequal, .lesser

Things which are evaluated in terms of each other have a relationship of .comparison. As already noted, .comparisons can be made either in terms of global similarity (ie "he is not like me") or on the basis of some common factor (ie "he is taller than me"). In the case of a particular .comparison, the relative measures of the two individuals are indicated by a feature belonging to .comparative.

.substance: *.solid, .liquid, .gas, .radiation (.sound, .odor, .flavor, .weight, .color, .brightness, .opacity, .heat, .humidity, .texture, .plasticity, .elasticity, .hardness)* **MATERIAL (.ABSTRACT)**

This domain classifies material things on their state of matter. *.Radiation** has been added to the more familiar states of *.solid, .liquid* and *.gas*. *.Substance* provides an alternative to categorizing material things on their *.class* or *.nature*. These three domains (and others like them) are fundamental, comprehensive and view their subjects from a unique perspective. *.Substance's* own nature is overshadowed by the many attributes which pertain to it. These attributes express the actual substance of sensation—color, brightness, warmth. They are numerous; thirteen are listed above and without a doubt others could be added. Although most are sufficiently obvious to pass without comment, others may benefit from a little explicatio...

.Brightness measures the intensity of light reflected or emitted from a thing, not its hue. The next four attributes imply that their tenants are solids: *.humidity* describes the presence of fluids of all sorts, ie 'greasy', 'damp', 'sopping'; *.texture* refers to the tactile sensation arising from a thing's surface, ie 'rough', 'smooth', 'slippery', 'granular'; *.plasticity* and *.elasticity* express an entity's capacity to be deformed permanently or temporarily; and *.hardness* describes the rigidity of a thing's surface. Note the different characters of these qualities. Some are *.continuous* (*.hardness, .brightness*), others are *.enumerated* (*.texture, .color*), still others are more complex (*.humidity, .sound*).

Vision seems to be the paramount sense. Although all qualities here appear on the same level, in an earlier version of *.substance* a distinction was made between attributes which can be perceived from a distance (ie by vision; 'appearance') and those which require contact or close proximity to be sensed.

* it is questionable whether *.radiation* should be considered a substance because the question of its fundamental nature is unresolved and because so few of the attributes of material things apply to it.

- a) *.color: .green*
- b) *.texture: .SOLID*
- c) *.humidity: .SOLID*
- d) *.plasticity: .SOLID*
- e) *.elasticity: .SOLID*
- f) *.hardness: .SOLID*

nature: *.discrete, .amorphous { .ABSTRACT }*

A thing's *.nature* describes the structural character of its material; *.parts* are *.discrete* and *.ingredients* are *.amorphous*. A *.discrete* thing has a structure and a typical size and shape which an *.amorphous* thing lacks. Ingredient names tend instead to evoke the textural quality of their material. Thus 'stones' are roundish *.discrete* things which I can hold in my hand, while *.amorphous* 'stone' is rough-surfaced, dun-colored and of heterogeneous composition. The difference between the two kinds of *.nature* is reflected in the grammatic distinction between mass and count nouns; we can count and measure 'stones', but 'stone' is a single thing of unspecified measure.

All *.discrete* things are ultimately composed of *.ingredients*. A house is made of individual studs and bricks and sheets of wallboard, but studs are composed of wood, bricks of clay and wallboard of gypsum. Each of these is an *.ingredient*.

.denotation: *.specific, .interrogation, .demonstration, .universal*
{ .ABSTRACT, .ARTICLE }

.Denotation expresses an entity's quantificative and deictic aspects. Two of the four kinds of *.denotation* correspond to universal (*.universal*) and existential (*.specific*) quantification. Universal quantification addresses all members of a class without suggesting that any actually exist while existential quantification denotes the existence of particular members of a class. To these are added features that assert the possible existence of particular individual(s) (*.interrogation*) or point to particular individual(s) (*.demonstration*). Examples of clauses and phrases using words that incorporate these features are "all men

are beasts" (.universal), "the men came today" (.specific), 'which men?' (.interrogative) and 'these men' (.demonstration).

a) .specific: .definite, .indefinite

It is possible to know that some thing exists without knowing which particular one it is. .Definite and .indefinite .specific .denotation express this distinction. Examples are "he gave it to a man" (.indefinite) and "he gave it to the man" (.definite).

1) .definite: .common, .proper

A particular entity may or may not be individually known to a statement's utterer. .Common and .proper identify these two cases. Examples are "he gave it to the man" and "he gave it to John Doe" ²³. 'John Doe' is interpreted here as an unambiguous identifier.

b) .demonstration: .near, .far

A peculiarity of English allows us to point to two things at the same time. The two are distinguished by their distance from the utterer in a space of indeterminate character^{*}; thus, .near and .far. Examples are 'these men' and 'those men'; "this point is more important than that one".

.existence: .assertion/.negation { .reality, .integrity, .perfection, .primacy, .universality, .typicality, .normalcy, .significance, .possibility, .concordance } (.CAUSATION) (.ABSTRACT)

All features have an existential aspect which is implicitly asserted by their presence in a word's description. Notwithstanding this, there are occasions when particular aspects of .existence—ten have been identified to date—are significant factors in a word's sense and appear there explicitly.

.Existence is also the domain of .negation. .Negation identifies the operation by which the sense of one or more features is inverted (unimplemented). The question of interpreting this sort of scope is problematic.

* distance may be expressed in terms of an utterance's sequence. In the last example, 'this point' refers to nearest prior one.

.orientation: *.positive/.negative, .neutral* {*.ABSTRACT*}

.Orientation is a protean domain. It expresses the presence and direction (.positive/.negative) or absence* of orientation (.neutral) in an aspect. People routinely draw inferences based on .orientation which are not explicit in an utterance, ie that .positive .emotion is good and .negative .emotion is bad.

.numeral: *<numeric value>* {*.ABSTRACT, .COUNT*}

One of the minor things people that do with language is express arithmetic relations. Most programming languages (including Lisp) treat numbers as a different type of data than symbols. .Numeral (unimplemented) is an undifferentiated feature which signals that the following token is numeric rather symbolic so that it can be processed correctly[§]. Directly accommodating numeric data in the system removes the need to define an open set of numeric feature symbols. Notwithstanding the functionality that .numeral introduces, it has proven useful to define feature symbols in .measure that distinguish counts of one, two and more-than-two.

.purpose: *.habitation, .transportation, .furnishing, .clothing, .nourishment, .communication, .amusement, .economic* *.ARTIFICIAL* {*.ABSTRACT*}

.Artificial things have .purpose; ie .purpose is a domain contingent on the .class subdomain of .artificial. This domain expresses the fact that absolutely everything that people make deliberately is made for some purpose. This domain is likely to be extended in future by the addition of new features and the subdivision of existing ones. Its members are ultimately based on the basic human needs for shelter, warmth, food etc. .Purpose provides a clear demonstration of canonicity. Many things are used for purposes other than

* .absence means either that an aspect has no orientation, or that its orientation has no bearing on the given sense.

§ and thereby plays a metasemantic role in the system.

their canonical one for which they were designed. A coffee cup makes a good paperweight.

.direction: *.across, .along, .around, .beyond, .pervasive, .so, .from*
{ .ABSTRACT, .POSITION } (.PROXIMITY, .PERSPECTIVE)

Relations between two things in tridimensional space are expressed in part by the enumerated attribute of *.direction*. This domain also applies to static *.conditions* and to other *.acts* which have a limited duration. *.Beyond* implies an *.extensive .perspective*; what is beyond a thing to you may be before it to me.

.proximity: *.interior, .exterior { .ABSTRACT, .POSITION, .DIRECTION }*

Unlike *.direction*, *.proximity* does not imply that the space in which it operates has a particular dimensionality. The first question arising about the proximity of one thing to another is whether it is inside the other.

a) .exterior: .contact, .vicinity

Two *.exterior* things that are touching have *.contact** *.proximity*. Things which are not touching may be in the *.vicinity* of one another; strict negation of *.vicinity* suggests that in the view of the utterance two things are not near each other. *.Vicinity* gives an excellent illustration of how we see some relations as significant and others as not. In any given situation a thing is probably in the vicinity of a number of other things, many of which may be closer to it than the one whose *.proximity* is noted. Mention is made of that particular relationship because it alone is deemed significant. *.Exterior* is an undirected *.binary* attribute (because it does not employ an coordinate frame) and *.vicinity* an undirected *.continuous* one (because it does).

* although it is possible to imagine interior contact—when one thing contacts the surface of another from the inside—this concept is infrequently used and no word exists for it.

.plane: *.horizontal, .vertical* {*ABSTRACT*}

.Plane is a case of a global coordinate frame; whatever is .vertical to me is .vertical to you. But although .vertical unambiguously identifies one axis of tridimensional space, .horizontal names a plane which contains the two axes of length and width. This domain provides good examples of the organization which we unconsciously project on our perceptions. Although it is true by definition that the axes of tridimensional space cannot all lie in the same plane, there is no reason why they must all be normal to one another, nor why objects must be oriented in space along their own major axes of symmetry. It is simply easier for us to deal with things in terms of this particular viewpoint.

a) *.horizontal: .breadth, .depth*

Although the most common names for the two horizontal axes are 'length' and 'width', these terms have not been used because they are best suited to an .intensive .perspective: a thing's length and width are respectively its longer and shorter dimensions* in the horizontal plane regardless of how the thing is actually oriented in an external context. .Breadth and .depth shift the point of view to a .comprehensive (therefore .extensive) context by involving the viewer. A thing's .breadth is its horizontal dimension in whichever of its planes most nearly faces the viewer, and its .depth is its horizontal dimension in whichever of its planes is most nearly oriented along the viewer's gaze.

.perspective: *.intensive, .extensive* {*ABSTRACT*}

.Perspective identifies the context or coordinate frame within which an individual is viewed. It is a protean domain whose values appear in many places in the dictionary. .Perspective§ first distinguishes between contextual and acontextual views, cases where a thing is viewed in isolation or on its own terms (.intensive), and cases where it is seen in

* 'height' is the corresponding vertical dimension.

§ 'subjective' and 'objective' are counterparts for .intensive and .extensive with just slightly different senses.

relation to other things outside itself (.extensive). 'Scope' is a synonym for perspective; although the two domains are homomorphic, .perspective is not the same as .proximity.

a) *.extensive: .circumstantial, .global*

.Extensive heads a subdomain identifying the particular context within which a thing is considered. A thing seen in a .global .perspective is viewed against a universally-understood reference frame like the points of the compass. When the context is centered on some aspect of the thing or its immediate surroundings, the .perspective is .circumstantial.

1) *.circumstantial: .local, .comprehensive*

A .circumstantial .perspective is centered in a thing and may include its immediate surroundings. It is .local (undefined) if it excludes the viewer; otherwise it is .comprehensive (undefined). This apparently minor distinction is a source of real confusion in everyday life; if you and I are facing one another and I speak about 'your right hand' it is unclear whether I mean right-hand-as-I-see-it or right-hand-as-you-see-it.

.character: *.complex, .simple {ABSTRACT}*

.Character conveys information about the structure of semantic space for purposes of processing and is therefore a metasemantic domain. The .character domain and the 'character' relation are limited to each other. Attribute domains, ones expressing qualities, have a structural .character (unimplemented) which can affect their interpretation.

Attributes may be .complex graphs whose members all have unique natures and can only be identified by enumeration, or .simple lines whose members all have a similar character and can therefore be identified giving their position relative to one another or to known positions on the line (ie the poles or the midpoint). Although .qualities belonging to a .simple attribute can be compared to one another meaningfully, complex features are incommensurable. Smell and taste are .complex attributes.

a) *.simple: .continuous, .discontinuous*

Linear or scalar values may be distinct from one another (motion, order, integer numbers)

or they may shade into one another imperceptibly (speed, real numbers). .Continuous attributes can always be subdivided; .discontinuous ones cannot.

1) *.continuous: (MEASURE)*

2) *.discontinuous: .binary, .multiplex (.EXISTENCE)*

.Binary (undefined) attributes are a large and important sort of the .discontinuous quality subdomain. The most salient characteristic of a .binary attribute is its unequivocality. Because a .binary attribute has only two values, an individual either does or does not possess it. .Multiplex (undefined) attributes have at least three values. Values in .continuous and .binary quality domains can be negated, ie have their positions on the scale of interest reversed.

.narrative: .preceding, .succeeding

.Narrative expresses the sequence of sentences in an episode. This relationship is directed and transitive; the linkage that it establishes is called a history. Because the syntactic organization of an episode's utterances are unrelated to the meaning they convey, .narrative is a metasemantic domain.

.person: .first, .second, .third (ABSTRACT)

The preexistence of a .person feature in .class causes this domain's name to be case-shifted. .Person (undefined) expresses the grammatic distinction between the two participants in a dialogue and other people, and also between those participants (speaker, hearer) themselves. It is expressed by direct address (statements incorporating 'I', 'you' or 'we') and through indirect or quoted discourse.

.syntax: .entity, .modifier, .relator, .punctuation (ABSTRACT)

This domain is used solely for parsing. Its organization is dictated by the grammar used in the demonstration program's augmented transition network parser; a more powerful parser would require a finer and more extensive classification. A number of the members of

.syntax are not used in the parsing process. These include most **.conjunctions**, many **.articles**, and all **.pronouns** other than **.personal** ones. No use is made of punctuation other than as a signal of the end of a sentence.

The syntactic categorization used here is eclectic but conventional. As a result the fundamental division of individuals into entities and relationships which is used throughout the rest of the system is modified somewhat here. Relationships are divided into **.modifiers** and **.relators**. Possibly because of its monadic nature, most grammatic theory does not view a **.modifier** as a type of relationship. **.Relators** on the other hand are recognizably connections between things. **.Punctuation** is considered to be a legitimate syntactic element. The three types of punctuation distinguished here delimit and identify sentences and reported discourse and structure sentential components.

a) .entity: .noun, .pronoun

1) **.pronoun: .personal, .reflexive, .relative, .possessive, .demonstrative, .interrogative**

b) .modifier: .article, .adjective, .adverb

1) **.article: (DENOTATION)**

2) **.adverb: .quantifier, .extender**

c) .relator: .verb, .preposition, .conjunction

1) **.verb: .infinitive, .tense, .participle (.CONSTITUENT, .VOICE)**

a> **tense: .prtense, .patense**

b> **participle: .prpart, .papart**

2) **.preposition**

3) **.conjunction**

d) .punctuation: .terminal, .quotational, .structural

1) **.terminal: .period, .question, .bang (.MOOD)**

2) **.quotational: .openquote, .openquotes, .closequote, .closequotes**

3) **.structural: .comma, .semicolon, .colon**

Among **.verb** features the prefixes **`pa-`** and **`pr-`** stand for **`past`** and **`present`**. **.Adverbs** are divided into **.quantifiers** and **.extenders** on the basis of scope and character. **.Quantifiers** specify the **.measure** of the sense to which they apply, usually that of the very

next word, and can be applied to other quantifiers; thus 'very little'. .Extenders specify other aspects of the sense to which they apply, which may be the next word and is usually a near neighbor.

.constituent: *.main, .auxiliary* {*.ABSTRACT, .VERB*}

.Constituent is contingent on .syntax. This domain distinguishes between senses which can express an .act relationship by themselves (.main), and ones which extend an .act's sense (.auxiliary) by specifying one of its existential aspects such as .possibility, .capability etc.

.voice: *.active, .passive* {*.ABSTRACT, .VERB*}

.Voice is another domain contingent on .syntax. .Voice shifts a sentence's focus from .subject to .object with the practical side effect of changing the sequence of positional .cases. .Voice is thus not a feature of any particular word in a statement but is instead determined by the pattern of words conveying the statement's .act.

.mood: *.statement, .question, .exclamation, .order* {*.ABSTRACT, .TERMINAL*}

.Mood (unimplemented) is contingent on .syntax and conveys a statement's illocutionary aspect, ie the sort of linguistic act the statement performs (Austin, 1962). Illocutionary acts have traditionally been classified as declarations (.statement), interrogations (.question) and exhortations. Exhortations are further subdivided here into .exclamations or .orders; .exclamations are utterances made simply to ventilate the author's strong emotions, but although orders may be uttered with great force of will and strong emotion, they are primarily intended to accomplish goals.

Utterance Networks

Networks* are graphs, structures of unspecified organization composed of nodes linked by arcs. Nodes are capacious data objects which can be structured internally in any way desired; in this case, they are repositories of the meanings of the words which constitute utterances. Arcs represent relationships or bidirectional[§] links between nodes. They are composed of pairs of relations and are expressed in the network by a label in each participant node which names its counterpart and specifies the relationship's character.³⁴

This character can vary from one end of a relationship to the other. Although .set and a few others are symmetric, most relationships are directed. 'The boy on the bench' is not the same as 'the bench on the boy'. -On- relates 'boy' and 'bench' to each other in different ways. This directional aspect must be expressed in some way in any representation.

Of the eight parts of speech, only nouns[‡] and verbs are represented by nodes in linguistic networks. Nouns may occur either in unmediated or prepositional phrases. All unmediated noun phrases and some prepositional ones[¶] are linked to verbs by .case relationships. .Case links constitutes the primary skeleton of the network. Articles, adjectives and adverbs modify nouns or verbs; their senses are transferred to the node of their .object and they are not directly represented in the network. Prepositions and

* a substantial literature on the use of networks to represent meanings has developed since Quillian (1968) coined the term 'semantic network'.

§ although the direction of .thing-.act and .modifier-.type relations is implicit in the different .types of participants, .thing-.thing relations which arise from informal relationships need expression of their direction.

‡ a pronoun initially causes a node to be generated temporarily, but as soon as its referent is discovered any sense and relations the pronoun has are transferred to the referent and the temporary node is deleted from the network.

¶ 'to', 'for' realize the .indobj case; 'by' realized the .subject in passive sentences.

conjunctions play a similar role in sentences*, with prepositions taking phrasal objects and conjunctions generally taking clausal ones. These two parts of speech attach peripheral entities to an utterance network by adding optional case relationships to verbs (.position, .specification etc) or by specifying particular aspects of entities (.component, .possession etc).

Nouns do not have case. Rather, they fill cases in relations. Modifiers are monadic and applicative and have a single case slot for the .object to which they apply. Prepositions and conjunctions have two case slots, one for the .object of the relationship they embody and one for the .subject with which that relationship is held. In 'the boy on the bench', the -on- relation has the object 'bench' and the subject 'boy'. Verbs can have a number of case slots, including three which are expressed positionally²⁰ in English and others which are realized through prepositions and conjunctions.

Networks expressing the meanings of a number of different but related sentences can be integrated into a single one that represents a coherent semantic whole or episode by merging the nodes of the entities they have in common[§]. Merging proceeds incrementally; as construction of each sentence network is finished an attempt is made to merge it with the most recent previous ones. If this is unsuccessful the sentence is simply left floating in the representation as a possible beginning of a new episode. The history list, a linked list that records the sequence of all sentences in an utterance, is used to identify different episodes and to establish recency within an episode.³⁵

Episode network structure is completely consistent that of sentence networks. Once sentences have been joined into an episode, all trace of their original syntax vanishes and it is structurally impossible to tell whether a particular formal relationship was realized as a subordinate clause or as a simple sentence etc.

* reflected in a high degree of category overlap between their word sets.

§ nodes are uniquely numbered; when two are consolidated it would be nice to recover the discarded number so that the sequence count is not broken and so that access processes could always rely on finding immediately adjacent neighbors.

The demonstration program documents how the commonest syntactic forms are processed and represented. However for the representation formalism to be adequately descriptive, it must be able to represent the sense of all possible syntactic constructions. For this reason the representations of a number of important syntactic forms, most of which are unimplemented, are discussed in the following subsections.

Complex Sentences

The beginning of a subordinate clause is signalled by a variety of syntactic mechanisms: relative pronouns, infinitives, possessive pronouns with present participles, present participles, conjunctions, coordinating punctuation (',':^;'), complex verb forms ('wanted to do') and without any mediating word or phrase at all³⁶. Although processing language on the clause level is an operation* of some complexity, the representations that result from it are quite straightforward: a subordinate clause's verb is treated as if it were a noun phrase[§] and allowed to fill positional cases in superordinate clauses or modifying noun phrases etc, while any absent roles within the subordinate clause itself are filled with the appropriate imports from the superordinate clause.

Thus in the network representing "The man who broke the bank at Monte Carlo had dinner", |man-nn| would have (.specification |broke-nn|) added to its list of relations, and the |broke-nn| would have its .subject case filled by |man-nn|. Because this structure is identical to the representation of "The man broke the bank at Monte Carlo. The man had dinner" interpretation would be straightforward. Cases in which the subordinate clause plays the role of an unmediated noun phrase as in "When he comes we will go" are more complex. In this case the subordinate clause has a .position .temporal relationship with the main clause and no tenants need to be inherited from above. When user processes

* it has not been implemented here because it is peripheral to this project and because the additional code would occupy too much scarce memory. The particular method in which subordinate clauses would be processed is detailed in Appendix 8.

§ mediated or unmediated as the case may be.

discover that the symbol named in the temporal relation is a verb rather than a noun, they would interpret it as a nominal whose relationships would specify its particulars.

Set Relationships

The logical relationships of conjunction, disjunction and negated conjunction of individuals are expressed by logical conjunctions*. These relationships can occur between sets of two or more modifiers, entities and formal or informal relationships.³⁷

A series of set members is often preceded by an introducer: 'both' before 'and', 'either' or 'whether' before 'or', 'neither' before 'nor'. Parsing the words between a particle and its counterpart conjunction (or significant punctuation like a comma or semicolon) serves to identify the syntactic form of the series members; once this is known an introducer can be dropped from the parse without affecting the sense of the statement in any way.³⁸

"The boys and girls laugh and play together" illustrates logical relationships well. In this utterance's representation each noun node would contain a .conjunct feature with pointers to the two verbs rather than a atomic pointer to a single verb: (.conjunct |laugh-nn| |play-nn|). Verb nodes would receive similar additions (.conjunct |boy-nn| |girl-nn|). Thus the siblinghood between 'boy' and 'girl', 'laugh' and 'play' could only be established in a two stage process: an added link would need to be followed to one of the nodes it points to, and the appropriate .conjunct pair accessed for the names of its members.

Nested conjunctions would be processed in a similar fashion. Once the scope of each conjunction was resolved, a phrase like 'boys and girls or women' could be represented as (.disjunct (.conjunct |boy-nn| |girl-nn|) |women-nn|)[§].

* the words expressing set relationships are uneasy members of the set of conjunctions. They are the only conjunctions which take phrasal objects and their semantic role is quite unlike that played by other conjunctions.

§ (.conjunct |boy-nn| (.disjunct |girl-nn| |women-nn|)) would be the alternative representation.

Complex Modification

There is a small but significant subclass of adjectives and adverbs which do not fall into the neat division between qualification and extension. For instance, although in 'bright green student' the autonomy of the senses expressed by two adjectives can (and perhaps should) be illustrated by inserting a comma between them, in 'bright green apple' the two senses form a compound unit. 'Bright' acts as a qualifier in specifying a particular characteristic of 'green'.

This distinction can be handled by giving 'bright' a .qualifier sense in addition to its .extender one, and subsequently disambiguating senses in the context of their object. Qualifiers have hitherto been easy to handle because they have been unambiguous, and the double reading of 'bright' makes it awkward. However it is subject of the same syntactic constraints as other .qualifiers; it must directly accompany the word whose sense it qualifies. This means that the disambiguation of the .relation category can be limited to adjoining pairs of meanings.

Other cases of complex modification are simpler. In 'monkey wrench handle', the two modifiers would be recognized as a cognate form of the compound noun (which is already recorded as such in the 8th edition of Webster's New Collegiate Dictionary) and treated as a single token in all processing from dictionary lookup onwards. The example of 'large Pacific Ocean liner' is similar. As its capitalization suggests, the compound here identifies a specific body of water rather than a class of wrenches.

Changing the last example to 'large ocean liner' presents the problem in a nutshell. The reference of 'large' is ambiguous about precisely the aspects that .extension and .qualification express. Is the liner intended for large oceans, or is it simply a big ship? Commonsense suggests the latter reading but a good deal of additional context would be needed for a sure choice to be made between these very similar meanings (a liner for large oceans probably would be a big ship). In practice the program would fail to disambiguate

between the two readings and leave 'large' ambiguous^{*}. A user process accessing this information might be able to resolve the issue but is equally likely to just ignore it on the assumption that if the distinction were important the utterance would have pointed it out.

Complements

Phrases following copulative verbs are considered to be semantically identical to modifiers. Their sense is folded into that of their object (usually the .subject, occasionally the .object) and the copula itself is not represented in the network. Copula have limited but specific existential senses[§] which are added to the complement to qualify its own meaning before sense consolidation occurs.

Negation

Negation is represented by the .existence value of that name. It is understood to invert[‡] a sense's meaning. The scope of what is negated can be understood to be a particular pair, in which case it appears within the affected pair, or an entire sense, in which case it appears alone in a pair. This is straightforward; the problem lies in determining which scope and which symbol is to be negated.

Empty Words

Certain words appear in utterances, speech in particular, simply to establish a rhythm or to introduce a pause in the communication. Tags, filler words or phrases, and interjections have no inherent meaning and can be ignored without any consequences to an utterance's interpretation. Examples are 'isn't it', 'hey', 'as a matter of fact', 'well,'.

* .extension is to be preferred because it can be reduced to .qualification but not vice versa.

§ see the Syntax section in Appendix 3 for a list of the 20 copulative verbs.

‡ depending on how large a set of possibilities it labels negation can be strict or relaxed. For instance, 'not every' strictly means -some- and but can also mean -none-; 'not warm' strictly mean -cool- but can also be interpreted as -hot-.

Implementing a Demonstration Program

This section documents a program to demonstrate use of the feature system in a linguistic context which has produced results both satisfactory in themselves and encouraging in their implications. Following a brief conceptual profile of the program's data structures and the processes that operate on them, four sets of test data and their results are discussed and summarized in tables. Particular attention is paid to the disambiguation of eighteen different relationships expressed by prepositions. A final section describes installation and program particulars and some design and implementation practices.

Data Structures

The program's data structures are modelled quite closely on those presented in the general discussion; its three major knowledge bases are a lexicon of semantic features, a dictionary of English words, and a network representing the semantic content of processed utterances. Structurally each representation is defined in a separate environment in which data objects are independent entries recorded on association lists. Conceptually each is a network composed of nodes called symbols linked by arcs called relations. Symbols in the feature system are unlike those in the two linguistic data structures which resemble each other sufficiently that they can be described most easily in terms of variations from a general form. Two auxiliary data objects which play a role in processing are the input string of text and a list that gives the sequence or history of sentences in an utterance.

Lexicon

Objects in the LEX environment are symbols representing semantic features. Each bears the name of the feature it signifies (prefixed by a dot to distinguish it from the English word of the same name) and is organized as an association list containing an indeterminate number of selections from a set of eleven pairs. The key* in each pair identifies the nature of a relation between the feature and others, and the values in the pair identify those others. Figure 1 presents a generalized feature symbol and the particular ones for .thing, .mind and

1. feature symbol schema

((members .value .value [.value])	immediate members of a domain
(is .value [.value ...])	domains of which a feature is a member
(parts .value [.value ...])	parts of typical members of this domain
(partof .value [.value ...])	domain of whose members a feature is a part
(affects .value [.value ...])	domains to which this domain may apply
(accepts .value [.value ...])	domains which may apply to this domain
(has .value [.value ...])	qualities of members of this domain
(qualifies .value)	domain to which a quality applies
(character .value)	intrinsic nature of a quality
(implies .value [.value ...])	features implied by this domain
(correlate .value))	counterpart or negation

2. .thing

```
((members .sort .part .ingredient .role .naming .quality)
(is .type)
(implies .nature)
(accepts .denotation))
```

3. .mind

```
((partof .class .immaterial .person)
(members .comprehension .attention .recognition .reasoning .memory)
(has .rationality .intelligence))
```

4. .reputation

```
((qualifies .society)
(character .continuous)
(accepts .measure .orientation))
```

FIGURE 1: Types of Feature System Symbols

* since these keys are processed disjointly from dictionary symbols, they need not be and are not dot-prefixed.

.reputation, which are respectively a .sort, .part and .quality*. Relations in the generalized symbol are annotated. Besides inherently possessing the semantic content suggested by its name, a feature conveys additional meaning through its relationships with other features. Links for the eleven types of relation defined in the feature system convert the symbols in the LEX environment list into the nodes of a moderately connected network.

Canonical Linguistic Symbol

A linguistic symbol is a list structure containing the senses, relations and housekeeping data for one or more words. It is composed of a number of Categories keyed to features such as .sense or .relation. Categories vary in structure; some contain atoms naming features or symbols for other words and others are composed of complex lists called Meanings. A Meaning can have either a single Sense or more than one., and corresponds to the totality[§] of the semantic content intended by a natural language word. Words are frequently polysemous, and each Sense in a Meaning is understood to express an alternative semantic reading. A Meaning which has multiple Senses is therefore ambiguous; -social-event- and -spherical-object- are two senses of the meaning of the ambiguous noun 'ball'. The individual Senses that represent particular specifications of meaning are structured as association lists containing pairs.

Pairs are the lowest-level entry in a symbol. They are flat lists that concatenate features to express relatively independent semantic components. Taken in conjunction, pairs are intended to provide an adequate description of a Sense's semantic content[‡]. Each is composed of a key identifying the domain that the pair specifies and none or more

* a feature must have two members to justify subdivision, therefore 'members' implies a minimum of two entries. 'Qualifies', 'character' and 'correlate' take single values; all other relations take one or more. Only ancestry lineage ('is', 'partof') is fully articulated; other linkages must be pursued step by step.

§ after semantic processing a single symbol may represent the senses of a word's modifiers as well as the sense of the word itself.

‡ suggesting that it is always possible to add new pairs to a sense if required.

values; the first value in a pair specifies a particular feature within the domain. Second and subsequent values extend or qualify the meaning of the feature they follow. Values name individual features or the roles held by entities that participate in relations. Roles are structured like Meanings and can therefore contain alternative Senses. Negation illustrates the difference between features in one pair and in different pairs. Within a pair, it inverts the sense of the rest of the pair; when in a pair by itself*, it affects the entire sense.

Dictionary

Objects in the DICT environment correspond to individual literal or idiomatic common and proper English words or phrases. Symbols bear the names of their item and have a .syntax Category organized first by syntactic part and then by the different Senses of each part. Synonymous and nearly synonymous symbols have Categories containing pointers to the root words to which they refer (`.see <name>`) which for near synonyms include the particular differences between their Sense and that of the root§. Words beginning idiomatic phrases have Category lists of the rest of the phrases (`.phrase <rest-of-name>`), which when concatenated with the current word, name other DICT symbols.

Figure 2 presents a number of types of DICT symbols. The first is a generalized version which corresponds closely to the canonical form presented in the last section. The next three show entries for particular words and are intended to provide examples of the various structures possible.

Network

Objects in the NET environment represent individual occurrences of entities (nouns)

* that is, without any features within the scope of the pair to capture it.

§ this practice is intended to save space should entries become large; at present it is hardly necessary.

1. generalized dictionary symbol

```
((category (meaning ((.key [{ .feature | role } ... ]) [(key [value ... ]) ... ])
              [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
              ((meaning (pair [pair ... ])
                        [(pair [pair ... ]) ... ])
              (meaning sense [sense ... ]) ... ))
((category {string | symbol} ) ... ))
```

2. |an|, a synonym

```
((.sec |a|)))
```

3. |on|, an preposition with six senses which heads an idiom

```
((.phrase "top of")
```

```
(.syntax (.preposition ((.position .spatial .exterior .contact .vertical .greater)
                        (.object .oriented .horizontal)))
          ((.position .spatial .exterior .contact))
          ((.position .temporal))
          ((.means))
          ((.focus))
          ((.qualification .state))))))
```

4. |on top of|, an idiom

```
((.syntax (.preposition ((.position .spatial .exterior .contact .vertical .greater))))))
```

5. |move|, a verb with three alternate parts of speech and two roles

```
((.syntax (.noun ((.type .naming |move|)(.class .abstract)))
          (.prtense ((.type .motion)(.constituent .main)(.time .present)
                    (.subject ((.propulsive))(.object ((.mobility))))))
          (.infinitive ((.type .motion)(.constituent .main))))))
```

6. |have|, a verb with two alternate parts of speech, each with two senses

```
((.syntax (.prtense ((.type .possession)(.time .present)(.constituent .main)
                    (.subject ((.number .multiple))))
          ((.type .possession)(.time .present)(.constituent .auxiliary)
          (.subject ((.number .multiple))))
          (.infinitive ((.type .possession)(.constituent .main))
          ((.type .possession)(.constituent .auxiliary))))))
```

FIGURE 2: Types of Dictionary Symbols

and formal relationships (verbs) in the input text. NET symbol names are based on the root DICT word that corresponds to an input word*, suffixed with a numerical count that identifies which instance of the word it is.

Linguistic symbols differ slightly from the canonical symbol form in having one

* because of synonymy and inflection these names may differ; 'an' receives its meaning from |a|; 'sufficient' from |enough|, 'walks' from |walk|.

level less nesting; it is as if the Meaning level were removed and one goes directly from Category to Sense. This simplification is due to the fact that linguistic symbols are syntactically unambiguous and do not need to represent part-of-speech alternatives. Depending on their use, linguistic symbols can contain a number of Categories: 1) `.sense` contains one or more word senses (more if the word has not been fully disambiguated) which are direct copies of the appropriate entries in the root DICT word's `.syntax` Category; 2) `.relation` contains a list of the word's relations in the generally chronological order in which they were added during processing. Relations are semantic senses which may contain links to other symbols^{*}; 3) `.part` identifies the word's syntactic category; 4) `.word` identifies its DICT root; and 5) `.symbol` identifies the symbol's own name.

Figure 3 contains examples of linguistic symbols. The generalized form of a linguistic symbol is followed by the two types of symbols which can appear in sentence and episode networks (shown by their possession of a `.symbol` category). The fourth and fifth 'symbols' are temporary variables used during processing to express relational parse parts: a modifier does not require a `.relation` category because its sense added to its object's as soon as that individual has been determined. Prepositions and conjunctions have more autonomy; they initially incorporate their object[§]. Because the object's meaning is stored in their `.sense` category, a new category called `.link` has been added to hold the meaning of the relation itself.

Input

Input is expected to be in the form of a character string of unspecified length in the

* relations contain links if they express binary or plural relationships (verbs, conjunctions, prepositions) and simple senses if they express monadic ones (adjectives, adverbs, articles).

§ this can be an entity (noun or pronoun) or a formal relationship (verb). In the latter case the relationship is nominalized so that it can occupy roles. In "It happened because John wanted it to", [want-nn] acts as the object of the `.cause` relation expressed by 'because' and involves [John-nn] and [it] through its relations with those symbols.

1. generalized network symbol:

```
((category ((.key [{ .feature | role } ... ]) [(key [value ... ]) ... ])
      [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
      (pair [pair ... ])
      sense ... ])
[(category {string | symbol} ) ... ])
```

2. verb / formal relationship:

```
((sense ((key [{ .feature | role } ... ]) [(key [value ... ]) ... ])
      [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
(relation (((preceding <verb-nn>))
          ((succeeding <verb-nn>))
          ((.type .act)(.has .subject <noun-nn>))
          ((key [value ... ]) [(key [value ... ]) ... ]) ... ))
(name <root word>)
(part { .prtense | .patense | .papart | .prpart | .infinitive })
(symbol <symbol name>))
```

3. noun / thing:

```
((sense ((key [{ .feature | role } ... ]) [(key [value ... ]) ... ])
      [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
(relation (((key [value ... ]) [(key [value ... ]) ... ])
          [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
          [((key [value ... ]) [(key [value ... ]) ... ]) ... ]) ... ))
(name <root word>)
(part { .noun | .pronoun })
(symbol <symbol name>))
```

4. modifier / informal relationship:

```
((sense ((key [.feature ... ]) [(key [.feature ... ]) ... ])
      [((key [.feature ... ]) [(key [.feature ... ]) ... ]) ... ])
(part { .article | .adjective | .adverb | .quantifier | .extender })))
```

6. prepositional or conjunctive phrase or clause / informal relationship:

```
((sense ((key [{ .feature | role } ... ]) [(key [value ... ]) ... ])
      [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
(relation (((key [value ... ]) [(key [value ... ]) ... ])
          [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
          [((key [value ... ]) [(key [value ... ]) ... ]) ... ]) ... ))
(link ((key [value ... ]) [(key [value ... ]) ... ])
      [((key [value ... ]) [(key [value ... ]) ... ]) ... ])
(name <root word>)
(part { .preposition | .conjunction })))
```

FIGURE 3: Types of Network Symbols

variable TEXT. This string can be in free format with multiple spaces, tabs and carriage returns between word substrings. If the program finds TEXT to be null at the beginning of the session it assumes that interactive operation is expected. It messages the screen and

waits for keyboard entry, processing input as soon as it is able to determine that it has received a complete sentence.

History

Once sentence representations have been merged with the episode in NET it is impossible to determine the order in which formal relationships (verbs) occurred in the utterance. Because temporal and causal sequence are understood to follow utterance order (unless explicitly indicated otherwise) it is important to record this information permanently. This is done through the history list, pairs of pointers in the .relation Category of the symbols representing the main clause verbs* in sentences which implement a bidirectional list[§] that links these verbs in utterance order.

Processing

So long as input is available, the system reads sentences and parses each one into its phrasal constituents. These elements are assembled into a network on the basis of syntactic (positional) and semantic information. After the network for a single sentence has been constructed, it is merged with an existing episode on the basis of common entity symbols. Ambiguity is accommodated by carrying forward alternative word senses until sufficient semantic information is available in the context to allow inconsistent ones to be discarded. Attempts to disambiguate occurs as soon as parsing is complete and subsequently whenever structural changes occur that alter the semantic context of an ambiguous symbol.

* the .acts in subordinate clauses have their position in the utterance sequence established through the roles they play in the main clause.

§ during development (ie currently) the History list is implemented as an explicit list separate from the network.

The processing discussion is organized on a thematic rather than a strictly sequential or grammatic basis because the program's operations are not compartmentalized and do not follow any consistent order*. Its activities are also comprehensive enough and its individual operations complex enough that the presentation must be made on a level that precludes discussing individual functions.

Input

This vague term covers the tasks of tokenizing, morphological processing and dictionary lookup, and ends when the appropriate meaning for the current input token (in WORD) is assigned to SENSE. The program processes words and idiomatic phrases within sentences within a text of indeterminate length. Sentences are delimited on terminal punctuation; words are sequences of printable symbols separated by white space or punctuation. High-level pseudocode for input processing appears in Figure 4.

```

while get-sentence
  while get-word
    if word first-capitalized & unbound
      shift to first-uncapitalized
    if word bound
      replace with any idiom
    else if root bound
      replace with any idiom
    else if good alternative bound
      replace with good alternative
    else report failure

```

FIGURE 4: Pseudocode for Processing Input

When a sentence is read from input carriage returns are converted into spaces and leading and trailing white space characters are trimmed. Word tokenizing ignores multiple white space characters and expands contractions according to the rules in Figure 5 39.

* although the system generally follows the usual practice of computational linguistic systems in beginning with morphology and moving through syntax and semantics to the discourse level.

```

if the word is 'won't', replace it with 'would not'
if the word is 'shan't', replace it with 'shall not'
if the word ends with 'n't', replace the substring with a space and 'not'
if the word ends with 'm', replace the substring with a space and 'am'
if the word ends with 're', replace the substring with a space and 'are'
if the word ends with 'd', replace the substring with a space and 'would'
if the word ends with 've', replace the substring with a space and 'have'
if the word ends with 'll', replace the substring with a space and 'will'
if the word ends with 's' and no main verb occurs in the clause
    replace the substring with a space and 'is', otherwise
    treat the phrase in which the word occurs
    and the following phrase as a possessive relationship.
if the word ends with "'",
    treat the phrase in which the word occurs
    and the following phrase as a possessive relationship

```

"John's here"
 "John's car broke"

 "The class' papers"

FIGURE 5: Pseudocode for Expanding Contractions

If a word is not identified as an exact match or a case-inflected form of an entry in DICT, the program attempts to decompose the word using a few of the simpler and more common rules of word formation.

```

if the root is not matched
    and it does not end with 'e', add 'e'
    and it ends with 'e', subtract it
    and it ends with 'i' replace 'i', with 'y'
    and it ends with a double consonant pair, delete one of the pair

```

FIGURE 6: Pseudocode for Matching Inflected Words

There is often more than one way to account for a word's derivation, and the order in which strategies are evaluated is critical to ensure that a word is tested for near kinships before more remote ones (Appendix 8, Organization).

Dictionary lookup replaces a synonym with its equivalent root word. Cases where words are close but not exact synonyms are handled by modifying or extending the root word's sense as appropriate. Synonym processing is straightforward. Encountering a category keyed on .see in a DICT entry directs the lookup process to retrieve the sense of the entry following the key; and if any pairs follow that name, they are merged with each

```

if the word ends with `s`
    and its root is a noun,
        set the word's syntax to noun
        set the word's sense to the root's noun sense
        add (.number .multiple) to the word's sense
    and its root is an infinitive,
        set the word's syntax to prtense
        set the word's sense to the root's infinitive sense
        add (.number .multiple) to the word's sense
if the word ends with `ed`
    and its root is an infinitive
        set the word's syntax to patense and papart
        for each syntax part
            set the part's sense to the root's infinitive sense
            add (.time .past) to the part's sense
if the word ends with `ing`
    and its root is an infinitive
        set the word's syntax to prpart
        set the word's .sense to the root's .infinitive sense
        add (.time .present) to the part's sense
if the word ends with `ly`
    and its root is an adjective
        set the word's syntax to adverb
        set the word's sense to the root's adjective sense

```

FIGURE 7: Pseudocode for Permuting the Sense of Inflected Forms

sense entry in the synonym's meaning, supplementing them or replacing the values of whatever pair(s) have matching key(s).

Multiple-word idioms are indexed in the DICT entry of their first word*. When an entry is discovered to begin one or more idioms the input string is tested against each possibility and the longest idiom beginning with the current word is returned. If no idiom or direct match for the current word is found, the program tests whether it is an inflected form of a root word. The rules for analyzing inflect permutations appear in Figure 7. The dictionary must be checked for exact matches before inflected forms otherwise root words which resemble inflected forms would be incorrectly coerced into them.

* this design is intended to reduce searching the input string. Instead of looking in DICT for word sequences of indefinite length each time a word is tokenized, the input string SENTENCE is searched only occasionally for those few strings mentioned in a word's .phrase Category.

Idiom lookup is handled in much the same way as synonym lookup. A DICT entry that begins an idiom has a .phrase Category containing strings with the second and subsequent words of every idiom or subidiom that it heads. Each of these strings is concatenated to a copy of the string coerced from WORD. These complete phrases are tested against the input string. If a match is found, the contents of the symbol it names are stored in SENSE. The program can handle exactly matching idioms and those whose first word is inflected. Thus if the phrase 'kicked the bucket' appeared in SENTENCE and the idiom's uninflected form was defined, after 'kicked' was read and found to be an inflection of 'kick', 'kicked the bucket' would be looked up and the sense of 'die' would be selected. If no idiomatic match is found the word is meant literally and the meaning of the symbol named by WORD is stored in SENSE. Since the meanings of idioms and subidioms by definition differ from the literal senses of the words constituting them, no capacity to add or replace sense elements has been implemented here.

Syntax

Although the primary use of syntax is in the parsing operation (where it is of paramount importance), it also plays in a supporting role in morphological processing before parsing and in network processing after it. Thus a word's part of speech must be identified before its sense can be adjusted to account for inflection during DICT lookup. In a similar way, syntax category initially identifies which words can occupy roles, have their senses merged, take positional cases etc. The continuing relevance of syntax to subsequent processes is shown by the fact that permanent symbols have a .part category to record this information (Figure 3, nos. 2 & 3).

However the main use of syntax is in parsing ⁴⁰. The functions which perform this operation compose a 15-state finite state machine based on the augmented transition network processing methodology (ATN). The grammar this mechanism implements is capable of analyzing simple sentences which include sequences of adjectives and adverbs,

prepositional phrases and autonomous modifier phrases^{*}. It can also parse the modal and auxiliary verb system completely, producing appropriate semantic output⁴¹.

The parser operates by reading a word from input and moving to a different state within the ATN depending on the its capacity to meet the syntactic and frequently semantic restrictions that govern entry to that state. If the attempt to move to a new state is successful, the word's syntactic category is known and a new word can be read after the current one has been added to the existing parse structure. If all transition alternatives in the current state fail, a non-deterministic or backtracking ATN[§] (like the one implemented here) will return to a previous state and try to advance via an alternative path of states.

The parser used here is described as based on an ATN model rather than being an ATN because it does not strictly implement the sort of state machine generally understood by the that phrase, ie one which is memoryless. In a traditional ATN the path taken to reach the current state is not known to that state and the path beyond the immediately following state is also unavailable to the current state. Progress through such a state machine is also usually understood to proceed from state to state based on the purely local circumstances prevailing in each state when control passes to it.

This parser does not fully meet those criteria[‡]. It employs a stack of symbols which allows it to look ahead indefinitely (in practice just a word or two) in order to correctly parse words like adjectives and adverbs which make forward reference. It does this by calling subnetwork fragments of itself recursively and making state transitions on the basis of the values they return ¶. This often includes the entire environment produced

* coding the processing of ambiguous sets of logical conjunctions ("ham and eggs or toast") is well-advanced.

§ all ATNs are deterministic to the degree that they evaluate different transition alternatives in a fixed order.

‡ this is not a bad thing and is probably even a good one since the differences are enhancements; it simply suggests the term 'ATN' is not a completely accurate description of processing.

¶ This is more than is usually understood by ATN exit conditions. Exit conditions usually accomplish side effects when the next state is not in doubt; in this case the side effects determine what that next state will be.

by the called state. At other times the symbol being parsed in the current state is restored to its present value in the new environment returned by a call; this allows processing to be interrupted so that an unrelated phrase can be processed and added to the parse and then to be resumed where it left off. Although this technique is used at present only to parse noun phrases embedded in extended verb phrases ("Is the little boy coming?") the same technique could be used to parse other subordinated constructions.

Non-determinism is ensured by establishing local scope in every state which can change the environment through the capture of a known set of global variables*. Within a state, each block of code representing an environment-modifying transition is also scoped locally so that a transition's failure will not modify the environment inadvertently through side effects[§]. And because the entire parsing operation occurs within its own environment, it is transparent to other processes (the global PARSE variable must be explicitly set to its local value in the final state if parsing concludes successfully).

The parser follows the ATN model in being a network of state nodes connected by transition arcs. Each state is a large OR-statement in which each branch is a code block that executes a state transition. OR-ing ensures that blocks will be executed sequentially until one succeeds or all have failed, in which case the current state also fails. Code blocks are composed of entry conditions, pre-transition side-effecting functions, the transition function itself, and post-transition side-effecting functions.

The parser includes a metasyntactic or heuristic function called Suitable? that directs global processing. An example may show how this is used. The utterance "the boy moves" can be read as either a noun phrase or a complete sentence. As a phrase, 'boy moves' would be moves that boys make, perhaps when dancing with girls who would

* Catch-Env. Throwing (returning environments) and catching are employed frequently in the parser.

§ if this seems to involve a lot of environments (and it does) be aware that environment management appears to be optimized. Environments are extended incrementally by new definitions and are copied only when explicitly requested. In a manner analogous to tail recursion optimization, one copy of a variable's value can serve for many environments.

2

NEWMARKET
MicroD
NEWARK, N.J. 07102

have their own moves. *Suitable?* tests the current state of the environment for specified conditions and returns a boolean value. In the case of a phrasal parse it would determine that the 'sentence' could not end at this point because both a noun and verb have not yet been processed ⁴².

The product of parsing is a list of noun, verb, prepositional, and modifier phrase symbols in PARSE in the order they were processed. Parsing attaches the meanings of local modifiers to their head's symbol so these are not represented independently.

Semantics

As is the case with syntax, semantic information is used throughout the system. Within the parser fully half of the state entry conditions test semantic features and almost all side-effecting code modifies word senses in some way*. Notwithstanding this, semantics' greatest importance is to the operations which follow parsing. These include the construction of a network representation for the sentence from the phrasal parse parts and the consolidation of that sentence representation with the network for the rest of the episode. Word meaning disambiguation is an equally important task; discussion will begin with it because it is the first process that occurs after parsing.

Selection between alternative senses in a word's meaning can only be accomplished in a linguistic context. Thus the process of disambiguation starts with all of the dictionary senses of a word or phrase and discards them one by one as inconsistencies are discovered between them and the senses of words associated with the given word in some particular utterance. Disambiguation can thus occur incrementally, with consistency being reevaluated each time the linguistic context is elaborated further. Although the process could be undertaken as soon as modifier meanings are attached to their objects, to simplify program architecture it is postponed until parsing is completed and is then executed on all parts in PARSE. It is thereafter reinvoked whenever possible.

* dictionary lookup does not use semantic information.

Two word senses are consistent if their .class pairs agree. A sense's .class can be specified explicitly or by implication. Pairs disagree when they identify siblings within a cell; ie .material and .immaterial within .class. In terms of pair structure, consistency equates to the set of features in one pair's ancestry in the domain being the same or a subset of that of the other pair. The elementary evaluative procedure is run on pairs of word senses until an unambiguous result is determined, with each inconclusive iteration causing one more stage of the necessary relational linkages of the features in each sense to be activated. Each sense's accumulated role-required and implied features are tested against the other sense's values until no more activation occurs between successive stages or inconsistency is found⁴³. Sense consistency checking is packaged in functions that apply it to each Meaning in two Categories, to all (appropriate) different Categories within a symbol, and to every symbol in PARSE. Because disambiguation is invoked frequently, it is designed to trap simple cases* and avoid unnecessary computations.

The overall direction of processing moves from word lookup through parsing to the construction of a network representation of a sentence's meaning and the amalgamation of that network with the existing episode network based on previously parsed sentences. This process can be serialized on the basis of its inherent functional dependencies (ie, a phrase cannot be attached to the tenant of the .act's .subject case on the sentence until that tenant has been determined). In the following list of the stages of network construction, conceptually insignificant operations are parenthesized: 1) disambiguate symbols in terms of their own categories; 2) (produce .symbol, .name and .available categories to be used during phrase attachment (role tenancy) operations); 3) replace nouns with referents in the episode network; 4) replace pronouns with referents in the episode network; 5) consolidate logically conjoined sets (partially implemented); 6) attach autonomous modifier phrases to their referents; 7) install unmediated noun phrases in verb positional roles; 8) link mediated

* neither sense or category ambiguous, only one sense or category ambiguous.

consistent-meanings, consistent-categories:

if neither element ambiguous

return elements

if one element ambiguous

test it against each component of other element

test each component of each element against other

test:

loop until return

case test-consistent nil,

return nil

pair,

return alist2

else loop inner

for each alist

accumulate a working set of features

retrieve implied feature

reconsolidate the class ancestry

activate *qualifies, part-of & implies* nodesif no *qualifies, part-of & implies* nodes activate *is* nodes

case test-consistent nil,

return nil

pair,

return alist2

if activated nodes

loop inner

case test-consistent true, return alist2

'?',

return nil

test-consistent:

if both flists are not nil

if either flist is a subset of the other

if both flists meet each other's requirements

return flist2

return '??

return nil

if the non-nil flist meets the other's requirements

return true

return '?'

FIGURE 8: Pseudocode for Disambiguation

noun (prepositional) phrases into the network; 9) (delete used conjunctions (partially implemented)); 10) generate network symbols and update the history list.

This sequence begins by replacing nouns and pronouns with any symbol in the episode that appears to match them* on the basis of the identity of co-referring things.

* therefore this operation is more exactly one of adding a sentence network onto the episode network rather than constructing it independently and attaching it. The latter conceptualization will be maintained because it emphasizes the syntactic integrity of the sentence.

Nouns are attached first because it is assumed that, except for the reflexive case, each entity in a sentence refers to a different individual and nouns have more restrictive senses than pronouns. An attempt is first made to find an exactly matching episode symbol, and if that fails, a semantically equivalent one. Failure here means the sentence symbol introduces a new entity into the utterance. Episode symbols matched to nouns are marked, and when all nouns have been processed, pronouns are tested on .gender and .number against unused episode nouns. This is done in two passes, the first seeking a pronoun's best possible match and the second trying to find minimally acceptable matches for any pronouns which remain unreferenced.

```

make list of referents
for each noun
    if exactly matched in referents
        substitute and mark used
    test approximate match
for each pronoun
    if exactly matched in referents
        substitute and mark used
for each unmatched pronoun
    if approximately matched in referents
        substitute and mark used

matcher:
loop until return
    case no referent, return
    symbol and test-referent consistent,
        if ambiguous, disambiguate
        if symbol1 is an { exact | approximate } subset
            merge senses
        if symbol2 is an { exact | approximate } subset
            merge senses
        return
    loop next referent

```

FIGURE 9: Pseudocode for Identifying Noun and Pronoun Referents

Reference substitution is quite robust. Between nouns it operates in both directions, substituting a more highly specified (semantically richer) sentence noun for a less specific episode symbol. Thus [man-nn] in the episode would be replaced by [father-nn] in the

sentence because the latter's meaning encompasses that of the former. A similar capacity to forward-reference pronouns introduced in a prior sentence to a noun in the current sentence is unimplemented. The scope within which referents are sought should not be unbounded. In practice people do not recognize entities as identical* if too many sentences intervene, so it is reasonable to conclude that the likelihood of a entity's acting as a reference is inversely proportional to its distance from the referring thing. And in practice efficiency considerations suggest that the search for a referent cannot be allowed to range over a large network unrestrainedly. The program limits its inspection to the three prior sentences.

Autonomous modifier phrases (AMP) are ones which are not directly adjacent to their object, ie 'very quickly' in "Very quickly he reached the shore". Based on the order of the sentence's three positional cases and its verb, the syntactic structure of a sentence is partially responsible for determining AMP attachment. When fully occupied, the positional cases define five locations in a sentence where an AMP could appear: <1>+subj+<2>+verb+<3>+indobj+<4>+obj+<5>. Permuting the grammatically - permissible order of cases with phrase positions identifies 18 possible sequences of phrases which can be reduced to five valid cases[§]. The grammar which determines to which phrase an AMP should be attached is based on its syntactic type (.adjective or .adverb), the number of positional cases occupied, the relative positions of the AMP and the occupied cases and the .type of verb (.condition, ie complementary, or otherwise). It correctly attaches complements to the .object if present and to the .subject otherwise and accommodates postpositioned modifiers ('the president elect')⁴⁴. Figure 10 details the rules used to accomplish this.

* unless enough explicit information is provided to ensure that a reference is made over a long distance. In fact the very presence of detailed specification in the body of an utterance is a clue to the existence of a referent for an entity.

§ and one invalid one; verb+[indobj]+obj is invalid because it lacks a subject.

```

for each modifier
  make list of referents
  case adverb,      attach to verb
       verb .condition, eliminate prepositions from referents
                           case verb+entity+adj, attach to entity
                              verb+adj+entity, attach to entity
                              entity+verb+adj, attach to entity
                              adj+entity+verb, attach to entity
                              entity+adj+verb, attach to entity

```

FIGURE 10: Pseudocode for Attaching Autonomous Modifier Phrases

Noun phrases are installed in a verb's positional cases on the basis of their position in the sentence and the sentence's .voice. Figure 11 presents the relevant fragment of grammar.

```

if active voice
  if entity+verb
    link entity subject
  if verb+entity1+entity2
    link entity1 indobj
    link entity2 object
  if verb+entity
    link entity object
else
  if entity1+verb+entity2
    link entity1 indobj
    link entity2 object
  link entity object

```

FIGURE 11: Pseudocode for Determining Positional Case Tenancies

The next stage of network construction involves linking prepositional phrases (PP) with their .subjects (their PARSE symbols already contain their .objects). Two options are possible; a PP can be attached to the verb or to the immediately prior noun phrase if one exists. Mediation of the prior noun is immaterial; PPs can modify each other's objects. Choosing between these two options involves factoring in whatever semantic restrictions a verb's positional case occupants might add to its sense. The rather elaborate pseudocode for this operation appears in Figure 12. Basically evaluation begins by determining that a particular PP indeed has a preceding noun phrase (if not, attachment to the verb is

automatic); if it does and if both possible linkings are semantically consistent, the program invokes a conceptually-dubious suite of functions which evaluates to which of the two candidates for attachment the preposition's .object is most semantically similar. Similarity evaluation is quite complex because the senses of both candidate .subjects and the preposition's .object may all be ambiguous. Consistent sense pairs must be selected, their similarity evaluated and the tournament decided in favor of the overall winner.

Similarity evaluation is done to demonstrate that even in the absence of sufficient semantic context to permit absolute disambiguation, a well-chosen algorithm for evaluating semantic similarity can produce impressive results. The activity is dubious because the program should really not be forced to make a choice between alternative relationships without good cause. Theoretically correct behaviour would consist of attaching a PP in a default manner and reevaluating that attachment in future if necessary, hopefully in a context that makes enough additional available to resolve the issue unequivocally ⁴⁵.

Linking prepositional phrases into the network completes its construction. The final tasks are to actually transfer senses between referents, delete the categories used to aid construction, and update the history list with the current verb.

Network

An Episode network requires housekeeping as new sentence networks are added to it. First of all its boundaries must be determined; ie the single representation of all utterances processed in a session must be broken down into its conceptual chunks. Network analysis techniques used to evaluate telecommunications systems can be employed to determine the points of low connectivity within the larger network which serve to indicate an episode's limits. Episodizing the network involves cutting off particular chunks of nodes that express related concepts from the rest of the structure when it becomes unwieldy and writing these chunks to secondary storage. Inactive episodes would

```

for each prep
  get consistent senses for prior and prep
  get consistent senses for verb and prep
  if no consistent prior
    if consistent verb
      link verb prep
      if consistent verb-subject
        link verb prep
        link most-similar verb prep
    if no consistent verb
      link prior prep
      if prep has unique noun relation
        if prep has unique verb relation
          if prior+verb+prep
            link most-similar verb prep
          if verb+prior+prep
            link most-similar prior prep
            link most-similar verb prep
        link most-similar prior prep
      if prep has unique verb relation
        link most-similar verb prep
      if prior+verb+prep
        link most-similar verb prep
      if verb+prior+prep
        link most-similar prior prep
        link most-similar verb prep

most-similar:
for each sense in prep
  for each sense in subject
    for each sense in object
      case no subjects or objects,
        return subject, prep, object
          unambiguous subjects and objects
        return extender1 subject, prep, extended object
          unambiguous subject
        return best extended subject, prep, extended object
          unambiguous object
        return extended subject, prep, best extended object
        return best extended subject, prep, best extended object

best:
set sum1 to zero
set sum2 to zero
for each pair in prep
  if pair1 is matched in subject
    add to sum1 ( agreeing features / average number of features )
  if pair1 is matched in object
    add to sum2 ( agreeing features / average number of features )
average sum1 and sum2

```

FIGURE 12: Pseudocode for Determining Prepositional Attachment

be moved to archival storage, their space freed up and their addresses added to the dictionary entries of their nodes' roots*. Other housekeeping tasks include promotion of the most specific sense to name a symbol (typically a proper name) and separation of a single symbol into two different ones§ when context renders them inconsistent.

It has been noted that network management includes traversing the history list to determine points of low connectivity where the utterance network could be cut apart into episodes when it becomes unwieldy. A backward-looking episodizing process cannot catch forward references among sentences; this would have to be done subsequently. Forward reference occurs when a sentence is connected to the episode through those that follow it rather than those that precede it. Although the technique of referring to something before it has been properly introduced emphasizes the fense of the forward-referring sentence (it often heads a paragraph), it also tends to confuse readers and is not frequently used. Network management functions have not been implemented.

Performance

The program's performance can be demonstrated in four areas which are described in their order of occurrence in the processing cycle. Modifier processing opens a window on the way in which the constituents of a word's sense can be affected by qualifying and extending modifiers. Auxiliary verb interpretation shows how a limited amount of semantic information allows the syntax of auxiliary and modal verbs to be correctly

* the notion of short-term and long-term memory immediately comes to mind, but short-term memory here could potentially be far larger than seven plus or minus two chunks of information.

§ this sort of structural garden-pathing can be hard to undo once a sentence's network has been merged with the episode network. Although relations added by merging would occupy independent and neighboring sequences in the merged symbol's relation category, it can be hard to know which particular relation constitutes the boundary between them. This rather abstruse problem is more complicated than this summary can suggest.

interpreted. This discussion also demonstrates how the features in an actual sense are interpreted. Because prepositional phrase interpretation makes the greatest use of semantic information it is the most important gauge of the program's capabilities and is discussed in somewhat greater detail than the other results. Attempts to recognize the identical symbols in sentence and episode networks demonstrate episode network processes in a limited way.

Additional functionalities which are not documented are the methods used to generate word definitions from the dictionary ⁴⁶, the attachment of auxiliary modifier phrases to the nearest suitable noun or verb, the (simplified) evaluation of sentence well-formedness, and word sense disambiguation. Although the latter operation is particularly significant, it is not easily documented and is subsumed in the interpretation of prepositional phrases in any case.

Processing Modifiers Within Phrases

The representation formalism recognizes two types of modifier scope, qualification and extension, and two sorts of qualification, degree and existence. Extending modifier senses are added to the .relation Category of the modified symbol. Modifiers which qualify .existence are added as a final pair or pairs to the Sense of the word they modify; degree qualifiers are inserted into the last pair of the Sense of the word they modify.

Table 1 shows how the system can discriminate between examples that combine these different sorts of modification and correctly attach qualifying features to the senses qualified. The first group of entries shows the dictionary descriptions for the words used in the examples.

Interpreting Auxiliary Verbs

Almost all senses of the auxiliary and modal verb systems can be expressed using three domains in the main verb's .sense category and one .relation. The domains are .voice (.active, .passive), .perspective (.intensive, .extensive), .time (.past, .present, .future).

PHRASE	REPRESENTATION
boy	((.type .sort)(.class .human)(.gender .male)(.age .little))
quick	((.type .specification)(.speed .great))
very	((.measure .great))
probably	((.existence .substantial))
quick boy	((sense ((.type .sort)(.class .human)(.gender .male)(.age .little)) (relation ((.type .specification)(.speed .great))))
very quick boy	((sense ((.type .sort)(.class .human)(.gender .male)(.age .little)) (relation ((.type .specification)(.speed .great .measure .great))))
probably quick boy	((sense ((.type .sort)(.class .human)(.gender .male)(.age .little)) (relation ((.type .specification)(.speed .great) (.existence .substantial))))
probably very quick boy	((sense ((.type .sort)(.class .human)(.gender .male)(.age .little)) (relation ((.type .specification)(.speed .great .measure .great) (.existence .substantial))))
very probably quick boy	((sense ((.type .sort)(.class .human)(.gender .male)(.age .little)) (relation ((.type .specification)(.speed .great) (.existence .substantial .measure .great))))
very probably very quick boy	((sense ((.type .sort)(.class .human)(.gender .male)(.age .little)) (relation ((.type .specification)(.speed .great .measure .great) (.existence .substantial .measure .great))))

TABLE 1: Processing Modifiers Within Phrases

.Voice is interpreted straightforwardly, but .perspective and .time require explanation. .Perspective in a verb expresses its simple progressive and perfective aspects, ie whether the verb's action is continuing or completed at a given time (the difference between 'is moving' and 'moves'). In addition to representing occurrence in past, present and future, .time interacts with .perspective to distinguish the most common tense-aspect combinations recognized in traditional grammar. Thus "he moves" is (.extensive .present), "he moved" (.extensive .past), "he was moving" or "he has been moving" (.intensive .past) and the pluperfect "he had been moving" (.intensive .past .past). Table 2 gives a set of examples which exhaustively exercise the finite state machine that represents this grammatic element..

PHRASE	STATES	.EXTENT	.VOICE	.TIME
moves	1	E	A	Pr
moved	1	E	A	Pa
is moving	1, 2	I	A	Pr
was moving	1, 2	I	A	Pa
has moved	1, 3	E	A	Pa
had moved	1, 3	E	A	Pa Pa
has been moving	1, 3, 6	I	A	Pa
had been moving	1, 3, 6	I	A	Pa Pa
is moved	1, 2	E	P	Pr
is being moved	1, 2, 5	I	P	Pr
was moved	1, 2	E	P	Pa
was being moved	1, 2, 5	I	P	Pa
has been moved	1, 3, 6	E	P	Pa
had been moved	1, 3, 6	E	P	Pa Pa
had been being moved	1, 3, 6, 7	I	P	Pa
had been being moved	1, 3, 6, 7	I	P	Pa Pa

I: .intensive E: .extensive; A: .active P: .passive; Pr: .present Pa: .past
 Numbers in STATES refer to nodes in the transition table representing this syntactic element.

TABLE2: Interpreting Auxiliary Verbs

Interpreting Prepositional Phrases

The subdomain of .relationship contains eighteen classes other than .act on its first two levels, all of which can be realized by prepositional phrases. The program attempts to distinguish between most of these types of relationships on the basis of their semantic constituents. It is not completely successful. One of the eighteen relationships, .arithmetic, cannot be processed because the program cannot handle numeric data at this point (although adding this capability would not be difficult.) It also fails to distinguish .set relationships from .possession because no way could be found to easily characterize the difficult distinction between the case when a thing with a person is his possession, and

when is it autonomous and simply in his company^{*}. It must be mentioned again that the program was deliberately forced to make choices between alternative senses with inadequate information; in practice processing would be naturally more cautious, the decision postponed and errors of comission would not occur.

Notwithstanding its success at recognizing 16 of the 18 relationships, further discrimination among more specific types of .position could not accomplished without the addition of more features to the system. This bottoming out is shown in the Table 3 summary of results by the suffixing of a count in four cases of .spatial and .temporal relations which cannot be distinguished by the program's present logic.

The seventeen tested cases[§] employ nine prepositions which have been selected to cover the set of relationships while including the greatest number of alternative senses. This was done to ensure that the discrimination logic was honestly exercised. Two of the prepositions, 'despite' and 'like', are unambiguous. Few prepositions express .inhibition or .comparison and none of those that do offer alternative senses. Because both relationships accept only verbs as subjects, identifying their attachments is trivial.

The other seven prepositions offer the program more scope and will be presented in some detail. The presentation is best begun however with a discussion of the contents of Tables 3 & 4 which summarize results. The first and third column identify the relationship and the test sentence. Although all relationships realized by prepositions take .entity objects[‡] by definition, some can only play roles in .acts, others only with .things and a final group of six relationships can be attached to either .class of symbol. Category restrictions

* The difference is between "He went to the show with his friends" and "He went to the show with a snack". I think that members of a set all belong to approximately the same .class, while possessions and their possessors are in different .classes (with possessions being in the less complex of the two). This is a distinction the network logic is not yet capable of making because it would require the explicit invocation and interpretation of a test on .class.

§ .arithmetic appears in the table for completeness.

‡ .syntax .entity and .class .thing are alternative ways of describing nouns.

RELATION	SUBJECT	TEST SENTENCE	INITIAL OPTIONS	AFTER DISAMBIG.	AFTER SIMILARITY	AFTER ATTACH.
arithmetic	N	3 by 4				
comparison	V	That man is like a boy	comparison	comparison	comparison	is
inhibition	V	He walks despite anger	inhibition	inhibition	inhibition	reads
spatial	N, V	The girl moves around the city	spatial (2) metric	spatial (2)	spatial (2)	girl
cause	V	She did it from love	spatial time cause	cause	cause	did
topic	N, V	The boy read a book about cities	spatial (2) topic	spatial (2) topic	topic	book
state	N, V	The boy is in love	spatial time (2) state style	state	state	is
temporal	N, V	He walks in the morning	spatial time (2) state style	time (2) state	time (2)	walks
set	N, V	The man walks with a dog	style set component instrument possession	set possession	possession	dog
style	N, V	He walked with enough speed	style set component instrument possession	style set possession	style	walked
instrument	V	He moves with his legs	style set component instrument possession	set component instrument possession	instrument	moves

TABLE 3: Interpreting Simple Prepositional Phrases

are shown in the second column in Table 3. The remaining categories on the right hand side of the table summarize the result of processing. Column 4 shows the original set of relationships attached to the preposition. Thus any sentence using 'for' will show .spatial,

.metric, .goal and .effect senses. Column 5 indicates what senses remain after symbol disambiguation at the conclusion of parsing. In some cases this does the lion's share of disambiguation (.state), in other cases it has little effect at all (.instrument).

In theory this is as far as processing should go. The current lack of sophistication of the logic does not allow it to make further eliminations with certainty. However to force the current version of the program to demonstrate its maximum capabilities, additional functions were written to select between alternative senses on the grounds of greatest semantic similarity. The relation sense was selected whose selection restrictions agreed best with both some sense of both subject and object (which of course could themselves be ambiguous). This relation is shown in column 6; column 7 shows the subject of this final selection. Eliminating the unprocessed and unambiguous cases leaves 15 relationships to discuss. Each preposition's relations will be discussed as a unit and presentation will move from the least to greatest number of relations.

Thus we begin with 'around'. Disambiguation eliminates its .metric sense because 'town' is not a .unit (of measure). In the case of 'from', only .cause of its three relations can accept an .orientation (ie .positive/.negative) in its object. Note that the objects of both .temporal and .cause can be .immaterial, so that is not an adequate criterion to discriminate on. Disambiguation does not help clarify the correct sense of 'about'. However because .topic requires a .communication subject, it is more semantically similar to 'book' than any .spatial reading. Note that .spatial requires a .material object while .topic can accept any class whatsoever.

'In' is used in two different ways in the examples. Both 'love' and 'morning' are .immaterial, causing the .spatial senses to be eliminated from both examples. Neither is a .quality, so .style is removed. 'Morning' has .time, so disambiguation leaves it in the .temporal sentence while removing it from the .state one. There is no need to test semantic similarity in the .state case, and in the .temporal one the .time feature tips the balance in the correct direction.

'With' is used in three instances here: .set, .style and .instrument. Disambiguation does a good deal of work to distinguish these cases. In the .set case, it eliminates .style on the grounds of its object being .material, .component because its object is not a .part, and .instrument for the same reason and because the object is also not .artificial or a .naming. Similarity incorrectly selects .set over .possession purely on their order in the list of senses; neither relationship has any restrictions whatsoever so their similarity is exactly equal.

Disambiguation removes .component and .instrument from the .style example because 'speed' is not a .part or a .naming, or .artificial. .Style's requirement for a .quality makes it more similar to the .quality 'speed' than either .set or .possession (which have no requirements). In the .instrument example disambiguation eliminates only the .style sense; 'legs', the object, are .material. Similarity does select .instrument out of .possession, .component and .set because it needs the .part object which 'legs' are and the .act subject which 'moves' is. .Set has no requirements; and although .component also needs a .part object, it takes an .entity subject and 'he' is less specified than 'moves'.

'For' appears in examples for .metric, .goal and .effect. Disambiguation serves to identify the .metric sense immediately; 'miles' are a .unit and only .metric looks for a .unit object. 'Anger' and 'love' are both emotions and disambiguation treats them similarly in the next two cases. It eliminates .spatial and .metric readings because emotions are not .material or .units. The .effect reading is chosen where the .goal one is not because .goals look for a object with .positive .orientation which 'love' has and 'anger' does not.

There are four different readings for 'of': .possession, .substance, .component and .identity. Disambiguation helps to only a limited degree, removing only one or two senses from each example. In a choice between .component, .identity and .possession, 'Ottawa' is most semantically similar to .identity because it has the required .proper .denotation. In the next example 'pages' are the .part which .component's subject requires; the .possession requires a .sort subject. .Substance's .ingredient object is more similar to the .ingredient 'concrete' than are .component or .possession. Finally, 'city' is identified

RELATION	SUBJECT	TEST SENTENCE	INITIAL OPTIONS	AFTER DISAMBIG	AFTER SIMILARITY	ATTACH
metric	V	The boy walked for miles	spatial metric goal effect	metric	metric	walked
goal	V	He did it for love	spatial metric goal effect	goal effect	goal	did
effect	N	It is reason for anger	spatial metric goal effect	goal effect	effect	reason
possession	N	Ottawa is a city of young men	component identity possession substance	component possession	possession	city
substance	N	It is a city of concrete	component identity possession substance	component possession substance	substance	
component	N	The pages of the book are green	component identity possession substance	component possession	component	pages
identity	N	She moved to the city of Ottawa	component identity possession substance	component identity possession	identity	city

TABLE 4: Interpreting Prepositional Phrases

as the subject of .possession because it is a .sort whereas .component requires a .part subject.

In conclusion it should be noted that both the feature system and word senses require a great deal of further specification to allow processes (which must themselves be improved) to draw finer distinctions with greater certainty. Notwithstanding this the descriptions of word senses that supported recognition of 16 out of 18 cases generally on theoretically inadequate data were not chosen for the purpose of processing but instead

were based on the objective and independent analysis presented in the earlier pages of this thesis. Although for lack of space a great deal must go unsaid about the meaning and interpretation of particular features and senses, it is on this level that the feature and linguistic systems are intended to be understood and used.

Recognizing Identical Entity Symbols

Those entities which individual sentence networks have in common with the episode network must be identified in order that the elements of the sentence's network be attached to the proper episode nodes. Recognizing the identity of two symbols can be quite difficult. Although in a majority of cases the same word or a synonym appears both in the episode and the sentence, sometimes synonymy is semantic rather than syntactic and a word must be matched with a phrase which has the same sense. Pronouns must be merged with referents of the proper number and gender. Finally, unless a reflexive pronoun is present in the sentence, each of its entities should match an episode entity uniquely. The effect of this is to preclude the same entity in the episode from matching more than one entity in the sentence*.

The difficulty is compounded when one realizes that references to the same thing can sometimes be quite allusive, basing a recognition of identity on function, appearance or some factor inferred using common sense rather than on a word's literal meaning. In addition, utterances that are intended to persuade frequently nominalize the concepts that they introduce in order to facilitate reference to them in future stages of argument.

Table 5 profiles the program's current capabilities in this sphere.

* this suggests that if one set of sentence-episode associations is satisfactory, the process should back up and try another set. One easy way of doing this would be to reorder the sequence of sentence nodes that it tries to match (unimplemented).

SENTENCE	NETWORK	MATCH
boy-2	boy-1, girl-1	boy-1
boy-1	lad-1, girl-1	lad-1
boy-1	(young) man-1, girl-1	man-1
boy-1	student-1	
(young) man-1	boy-1, girl-1	
student-1	boy-1, girl-1	
he	boy-1, girl-1, city-1	boy-1
she	boy-1, girl-1, city-1	girl-1
it	boy-1, girl-1	boy-1
it	boy-1, city-1	city-1
they	boy-1, (multiple) man-1	man-1

TABLE 5: Recognition of Identical Entity Symbols

Implementation

Site

The program was developed on an IBM-PC AT compatible with 640K main memory, 512K extended memory and a 20 mbyte hard disk with 65 ms average track access time. The machine's Norton system information (SI) speed index is 7.1. Program performance was not measured systematically since this was not a design criterion; however, operational execution speed and system response was not an irritant during development*.

* the decision to implement EDWIN, the included EMACS editor, in Scheme is controversial. Scheme uses 246K memory itself and, after EDWIN uses its additional 97K, only 232K is left for program development. High-level languages are not particularly well-suited to delivering text editors; this point is repeatedly driven home by the editor's sluggish performance. Acknowledging the resource-hogging aspect of this problem, TI added environment management functions in Scheme 3.0 and moved EDWIN to a discardable environment, allowing it to be discarded and reloaded as required.

Language

The program was written in Texas Instruments' PC Scheme, Ver. 3.0. Although PC Scheme 2.0 and later versions support both expanded and extended memory, there is a substantial performance penalty for using either*, so the program was designed to run within the 640K of main memory. The expanded memory version of the language also appears to be very fragile.

Scheme is a variant of LISP that approaches but does not exactly meet the emerging Common Lisp standard. Scheme's own integrity is regulated by revisions to the 'Report on the Algorithmic Language Scheme' published and updated by MIT. The current revision is the third or 1986. The Texas Instruments implementation 'adheres faithfully to the Scheme standard' (Texas Instruments Data Systems Group (TI-DSG) 87) and, where extensions have occurred, these have made been in the direction of conformance to Common Lisp.

PC Scheme is implemented on top of a 64-register virtual machine§ that executes byte codes. This intermediate code is produced by incremental compilation rather than interpretation and the compiler can optimize its output in a number of ways. The language itself is composed of Core Scheme, a set of primitive operations written in byte code, and a larger number of functional applications and special forms which are reduced to core operations through compilation or macro expansion. System-provided and user-defined

* For main/extended/expanded memory, compiled program loads took 166/263/242 seconds; runs took 10.5/27.7/77.7 seconds. The second and third figures were obtained using extended memory to emulate expanded. Because the emulation technique involves handling calls to expanded memory by swapping 16k pages into a main memory buffer instead of switching address lines to actually access different banks of memory, the expanded memory values are quite suspect. The latter two memory configurations also proved fragile.

§ interrupt the initial bootstrapping of the PC Scheme programming environment at the right time and you'll exit to the 'TIPC Interactive Debugger', a undocumented language development object that TI continues to deliver with version 3.0. Run the debugger in trace or execute mode for 1K-instruction blocks and it is possible to watch the virtual machine load the rest of the byte code that executes PC Scheme.

functions are indistinguishable after this process. The implementation is generally well-designed and quite robust .

Development Tools

The 20-month period over which the feature system, the utterance representation formalism and the demonstration program were developed provided ample opportunity to devise tools to aid development. These can be divided into six classes: managers for debugging, execution, the fileset and memory; a utility to enforce consistency on the feature set and functions to facilitate escapes to DOS.

Breakpoints: *Alert, Unalert.*

In addition to offering two of its own implementations, TI-Scheme documents a general-purpose execution control facility embodied in a pair of functions called Advise-Entry and Advise-Exit thereby making it available to users. Alert and Unalert take advantage to implement a flexible shorthand tracing mechanism which displays single linedepth-indented messages signalling entry and exit from targeted functions. Alerting is useful for tracking execution on a high level.

Tracing: *Switch.*

Switch and its auxiliary formatting functions implement a control panel which allows custom 15 different execution-tracing functions to be turned on and off so that particular areas of interest in the program can be monitored dynamically. Messages displayed by these functions show the values of different variables used in parsing, word sense consistency and similarity testing, network construction etc. A number of messages are edited to eliminate nil arguments.

Execution Management

Loop Toggles: *Ptoggle, Ntoggle, Etoggle, Dtoggle*. The program's fundamental operation is to loop through a series of operations so long as a complete sentence is available at input. These operations are parsing, sentence network construction and integration of the current sentence into the episode network. Frequently the current focus of development is on only one of these operations, and it is tedious to execute the full loop each time the apprehending program is run.

These three toggles implement memoizing. They allow sections of the loop to be switched into and out of the execution cycle; a known environment (the one produced the last time the segment was actually executed) is returned when a switched-out segment is invoked. *Ptoggle* switches parsing and *Ntoggle*, network construction. The regular TI-Scheme error-handler breaks execution and enters an Inspector function whenever encountered. Sometimes this is desirable and sometimes not; *Etoggle* switches between the system error-handler and one that simply signals the error and exits to the top level. All the advising procedures documented in this section sometimes overwhelm the developer with information; *Dtoggle* turns off debugging information display.

Portable Trap: *Capture, Uncapture*. These functions operate like portable toggles. When *Capture* is put inline in code the values of its argument variables are stored in *Captures* the next time the code is executed. A subsequent *Uncapture* allows execution to be restarted deep in the loop with known values. Variables can be repeatedly uncaptured.

System Resets: *Reset-Flags, Reset-Variables, Reset-Parse-Indexes etc.* These functions clear or reset system variables to known values. *Reset-Flags* deals with the flags that tell the main program loop where input is coming from and whether a run is beginning or already in process; *Reset-Variables* does the same and clears 12 variables used in parsing. The functions which convert a parse to a network need to know their current

location in the list of parse parts; *Reset-Parse-Indexes* the appropriate indexes. *Reset-Net* performs a similar task on the *Count* property of words in the Dictionary which is used to name the instance symbols based on a word.

File Management: *Boot, Freshen, Update-All etc.* TI-Scheme compiles all code before executing it, and users have the option of manually compiling source code (*name.s*) to object (*name.so*) or fast-load formats (*name.fsl*). The *Update* functions automate this process for one, a group or all files. *Freshen* implements a rudimentary make facility which checks the sizes of all files against a backup copy and allows the user to *Update* changed ones. It is up to the user to be aware of syntactic extension dependencies and file modification status. *Boot* loads the fileset from the *\fsl* directory.

Memory Resource Management: *P.Dump, P.Load, E.Restore, L.Restore, D.Restore.* Coding artificial intelligence applications in Lisp on a 640K machine is a good way to be taught empirically that the packing problem is of non-polynomial order. By the later stages of development, compiling the fileset took 30 minutes (most of which was spent in garbage collection) and 'out of memory' messages appeared with distressing regularity at the precise moment when execution was at its most convoluted and demanding of resource.

A developer quickly adopts a defensive strategy to minimize the risk of such interruptions. In this case that meant writing functions to *Unbind* unused sections of code and *Load* them after the critical execution was complete*. *P.Dump* and *P.Load* do this for chunks of the parser; *L.Restore* and *D.Restore* reload the *Edwin*, *Dictionary* and *Lexicon* environments if they have been unbound§.

* this makes sense in conjunction with toggling. An attempt to extend this idea to the Scheme language itself was unsuccessful.

§ unbinding environments is flaky; all references to an environment's contents must be expunged from the system before its space can be reclaimed by the garbage-collector. However unbinding is only irregularly successful even if the command history in *%STL-History* is reset; and once this occurs the ceiling of available memory is effectively lowered because the pointers that keep the environment recent are unknown.

Feature Consistency Aide: *Checklex*. Human memory being what it is, there were practical problems during the design of the feature system in ensuring that a desired feature name was not in use elsewhere in the system, and in carrying a feature's renaming through all of its instances in the feature system and in word descriptions. *Checklex* is designed to help this by scanning all 'dict'-prefixed files in the current directory and all files in the \lex subdirectory and counting all instances of each particular data object. The resulting list is sorted and written to output in a columnar format to ease reading. It is then a developer's task to correct misspellings, unmatched pairings and duplicate uses.

Executing DOS Processes: *Dos, Dir, Gal*. TI-Scheme recognizes DOS' capability to spawn child processes. However the language is such a memory hog that the function that escapes to DOS takes an argument identifying how much memory should be freed up for the child process. *Dos* frees up just 80k, *Dir* abbreviates the DOS Directory command and *Gal* invokes the Galaxy word processor with a file and 320k of memory.

System Organization

The demonstration program is composed of 270k of code and data occupying 21 files. 16 files (212k) contain executable functions and syntax forms; those remaining contain the dictionary (2 files, 31k) and lexicon (4 files, 28k) linguistic knowledge bases. One program file, *Trim.s* (33k), contains functions and syntax forms which although useful were removed from the bootstrapped fileset to conserve space; another, *Checklex.s* (4k), is a self-contained utility used to check feature names for duplication. The stripped-down version of the program actually in use occupies 179k of memory.

Viewed conceptually, the lexicon contains definitions for 307 semantic features and the dictionary describes 203 English words. Ignoring *Trim.s*, the program adds 36 data objects to the four major ones already discussed. It also defines about 145 syntax forms*

* 94 Syntax, 30 Alias and 21 Define-Integrable forms. Syntax and Alias are used in preference to the more powerful Macro because of their simpler syntax.

and 220 executable functions, all of which are heavily and consistently commented^{*}. The function set is distinguished by the large number of general purpose utilities it adds to PC-Scheme. Eleven rectify shortcomings or idiosyncrasies of Scheme and this implementation⁴⁷. 19 functions generalize the mapping process⁴⁸ and single and multiple execution timers and a programmable alarm clock have their purposes.

More than 60 other functions which access or change flat or association lists, and restructure or decompose strings or lists are sufficiently general in purpose to be useful to others than the developer. Some of the syntax formats and most of the 21 Define-Integrable forms define recognizable contractions[§] of function names which serve to shorten code that applies numerous functions. Certain of the 30 Alias macros take advantage of an alias' automatic evaluation to eliminate the need to type parentheses around functions like GC which require no arguments. The remaining 150-odd functions and almost all Syntax forms were written for the system's peculiar knowledge structures or symbol conventions and therefore unlikely to be useful elsewhere.

Complexity

Stream processors have two relevant space statistics: the size of output as a function of input, and the size of workspace required in best, average and worst case. Due to the expansion of single words into multiple senses and the storage of more than one sense for ambiguous words, output size is a low order integer multiple of input. It is hard to evaluate workspace requirements because garbage collection trades speed for size. All that can be said is that it is a function of syntactic complexity and in any event a fraction of the

* the dictionary (38k) is larger and the program code (165k) about as large in fast-load as in source format, suggesting that what was gained from eliminating comments was lost in expanding syntax forms.

§ & (And), ~ (Access), App (Append), T-E (The-Environment), M-E (Make-Environment), U-I-E (User-Initial-Environment), St->Sym, Sym->St, St-App, St-Len, St-Ref, St-Rep, Subst, St=? and the alltime champion Find-Prev for (Substring-Find-Previous-Character-In-Set).

size of code, data and language; in the cases investigated here, 100k barely permitted execution and 200k provided relatively quick processing.

Apprehending a text's meaning has quadratic average time complexity. Lower bound is linear, upper bound is cubic and is determined by the number of nodes through which backtracking occurs. In practice the average can be expected to be skewed towards the lower bound, in which event complexity is determined by the routinely quadratic functions which include traversing the lexicon during semantic processing (disambiguation and similarity-testing) and searching the episode for identical symbols when sentence networks are merged with the episode network. No attention has been paid to the design of heuristics. Time complexity cannot be estimated for the use of the representation in applications (comprehension).

Design and Programming Practices

Programming Style. The demonstration program has been written in what is hoped to be an uncomplicated style which has been intentionally limited to a subset of LISP which really does seem to be common to all implementations of the language. It should be easy to port. PC Scheme's particular syntax extensions* are the one idiosyncratic feature in which the author has indulged unrestrainedly, however these could easily be replaced by the functions they repackage..

Embedded Grammar Defended. Systems implementing natural language grammar are frequently designed to accept the grammar in a English-like style which is translated into executable code by a macro compiler. That approach has not been used here because I believe it restricts the designer's capacity to specify flow of control too greatly. Instead I have written a number of syntax forms which can be strung together to make directly executable grammar statements. These statements have the same style as ones in a

* functions from this group are the ones that ended up in Trim.s.

translated grammar but avoid the straitjacket of a 'entry-condition', 'code', 'exit-condition' sequence dictated by a translator macro. As a side benefit the designer is always aware that it he and not an intervening translator must code the syntax form.

Memory Limits Affect Coding Style. In addition to the other prices that it exacts, memory constraints have an impact on coding style. Macros which expand code rather than restructure it tend to be avoided in favor of functions since the full bodies of the latter are not inserted in the code at each place where they are invoked.

Top-Down and Bottom-Up Design. Although applications of this size are not particularly hard to manage (ie the development aids were not hard to write), the open-ended design of this particular system made its own demands on the development process. In particular in order to avoid reinventing the wheel in various circumstances, an attempt was made to identify common operations and design general-purpose functions to perform them*. These functions do well-defined tasks on data objects of known structure; their small size and limited scope means that they can be tested exhaustively and thereafter used with confidence. This bottom-up process of design and coding functions of limited scope went ahead simultaneously with a top-down effort to identify and program the task-specific procedures to execute language meaning apprehension using general-purpose functions.

Environments Speed Accessing Long Association Lists. Environments figure greatly in Scheme as a mechanism of organizing and encapsulating objects. They also permit performance to be enhanced in an applications which uses many objects and performs many lookups. Since object lookup is a linear function of the number of objects in the context in which lookup occurs, partitioning the total set of objects by defining related ones in appropriate environments results in significantly quicker execution occurs. Environments act as an indexing mechanism.

This has been reflected in the demonstration program by defining each knowledge base as an environment in the user global environment. All user-defined symbols, literals

and procedures are also bound in this environment*. Environments are closures which organize their objects on association lists. Unless it is not noted to the contrary, this default organization (with linear access time) has also been used in user-defined environments.

* as TI has done with EDWIN.

The Language Representation

The representation that would result from processing a number of texts⁴⁹ would gradually develop a different character from that of the individual episode networks that compose it⁵⁰. For one thing, as it increased in size it would begin to represent an instance* of a general understanding of the language expressed by the words in its dictionary (call it therefore a language representation). The dictionary is paramount in this language representation. Because the names of symbols in episode networks are based on their dictionary entry (plus an unique instance number), the dictionary serves as its primary index. And because storage addresses of all symbols based on a particular entry are recorded in its dictionary entry, the dictionary would also be the main access path into the representation⁵¹.

As a result meanings in a dictionary entry would be overshadowed by the large number of addresses of instance symbols attached to it. A computational directory would differ from a printed one in another respect; it would have more entries for proper nouns than for common words. These would include all the individual things we know: the names of people, companies, brands, towns, streets and many other less general identifiers such as books, records, events, even part and phone numbers. Most proper nouns will however probably have fewer instance symbols than most common words.

To recall a particular instance of a word (an infrequent occurrence) its dictionary entry would first need to be inspected to find pointers to the symbols based on the word. A typical dictionary entry's symbol list would likely identify hundreds, thousands or even more instances where the word has been used in an episode. This manner of accessing

* an instance because a different understanding would arise from different texts.

apprehended text is psychologically plausible. People also begin to understand utterances by looking up the meanings of their constituent words in a printed or a mental dictionary.

Assume we have retrieved an instance symbol. It represents the sense of either a thing or a formal relationship (.act) ⁵². The context of a thing's symbol is established by its monadic (ie modifier) relations together with whatever binary or ternary relations connect it to other things. Each of these relationships has a character: componential, positional, specifying, .act case etc. Because they persist, things usually participate in many relations, all of which would appear in a symbol's .relation category in approximate chronological order. History lists (.act linkages) which give the sequence in which events occurred in the original narrative could be accessed for any .act in which the symbol holds a case.

Symbols representing a formal relationship or .act also accept monadic relationships which specify their particulars. Formal relationships contain cases, and the things that occupy an .act's cases are recorded in its .relation category together with any other relationships such as position, causation or specification in which it participates and which establish its context. From the perspective of an utterance, acts have shorter lives than things; each sentence introduces a new one. For this reason acts have fewer relations than things and the order of their .relation category's contents is less important. One particular entry in this category deserves mention. Relationships with its predecessor and its successor in the narrative establish a linkage which preserves the order in which the narrative's sentences occurred. This linkage is the history list.

Given this structural configuration, it is possible to see how the language representation could be used to address a number of traditional cognitive science tasks.

Uses

Typical Word Use

If a word has more than one sense, we can readily determine (with a confidence factor) how likely a particular sense is to be used by computing its proportion of the total number of occurrences. If we want to know which sense is most frequently used in a particular context (say with a particular sense of another word), we can produce the same statistics after filtering the instances with the test case.

To facilitate future use it might be desirable to memoize these statistics once computed; on the other hand, the representation is continually changing so the numbers will gradually age and may become less representative*. An $N \times N$ array is a logical data structure for this information (if it proved worth collecting); if recursive frequencies are to be stored (ie given that 'boy' occurs with 'walk' what is the likelihood it will also occur with 'girl'), some kind of hashing mechanism§ on the array might be suitable. Since each recursive step squares the amount of information to be stored, it is most likely that a better route to go is to use highly optimized processes to generate data by operating on the representation itself. At least then information would always be current.

The senses in a word's dictionary entry incorporate what are presumed to be its canonical features. If practice shows that features which are not part of a sense's description always co-occur with it, a good argument could be made for redefining the entry to add them. More practically, in the absence of conclusive evidence, statistics on the

* but not necessarily if initial trends continue. The larger the representation becomes, the more stable it is—and the more time consuming frequency computations will become.

§ could a generating formula be envisioned?

frequency of isolated and contextual usage* of a word's different senses could guide choices between them during disambiguation.

Understanding Figurative Use

In addition to their literal meanings and usage, words can have figurative senses or be used in figurative contexts. Some figures of speech are realized syntactically, and the structural symmetry that makes an epistrophe pleasant and a zeugma arresting would not be overtly recorded in the representation. Others like idioms would have their own dictionary entries, as would whichever onomatopoeia were recognized as words (ones that were not would be discarded as nonverbal utterance). Sound symbolism would be lost, embodied in the particular phonemes of words as it is. With the exception of idiom, however, these figures are not semantic.

Semantic figures are more interesting. Synecdoche is the substitution of a part for a whole or vice versa; metonymy is the substitution of an association or aspect for a thing. These and the more general metaphoric use of language (whereby an inappropriate name or modifier is applied to an object) would be viewed[§] as inconsistent or meaningless (how can an error glare? why would a pot be put on the stove?). However in each case the structure of the utterance could not be gainsaid and because figures are not arbitrary, in each case there would be relationships along which to pursue meaning. For instance stoves have burners as parts (on their tops), burners heat things, and pots are containers used for a variety of purposes including cooking which involves heating. Therefore 'put a pot on the stove' actually means to put it on a burner on top of the stove. Errors are bad things and people glare when they have bad feelings; more to the point, when people glare they

* assess contextual usage first. Use statistics on isolated frequency only if there are few or no contextual instances.

§ the parser as currently designed would attempt to disambiguate their meanings and throw away correct senses accidentally until it was reduced to a single sense which it recognized it could not discard. This argues for a reinitialization and redisambiguation of words if they are found to be inconsistent in terms of their current meanings.

become noticeable, so a glaring error is a noticeable one whose sense of badness is redoubled by this fact.

The more general effects of including figurative usage in utterances would be to force inference mechanisms to distinguish between literal and figurative links in particular episodes so that inference and statistics could be based solely on veridical instances. Because figures deliberately break the semantic rules, they define a very large set of possible utterances. It is therefore probably inappropriate to attempt to base generalizations on figures; the resulting observations would be too diffuse and probably contradictory.

Word Acquisition

Given the great deal of semantic structure available in the feature system and the language representation, it should be possible for an artificially intelligent agent to interact efficiently with an interlocutor to define new words by providing him with well-chosen alternatives based on his previous responses. Word senses would be treated as a discrimination tree to be traversed to a terminal during the description process.

The dialogue would begin by determining the new word's .class and .type because all senses have these values. It would then proceed as guided primarily by the necessary (but also by the accidental) relationships between features; contradictory associations would not be accepted). It is well within current technique to assist the definer by showing him menus of features and their definitions, and the synonyms and antonyms of the sense as currently defined. Before a new description was accepted it would be tested against existing ones; if it matched, the interlocutor would be asked to confirm it as a synonym or to attempt to identify the difference between it and its nearest neighbor. The results of defining a new word could be quite far-reaching and include introducing new domains to the feature system.

On-the-Fly Scripting

A script is a mechanism frequently used to package habitual sequences of actions (Schank & Abelson, 1977) for purposes of planning or understanding. One of the major weaknesses of scripts are their inflexibility and the consequent requirement for great numbers of them to handle activities of any variety (Dreyfus, 1985). Contrariwise, they have also been criticized for ignoring explicit detail (Graesser & Goodman, 1985). The language representation addresses both these problems; every pertinent symbol in it participates in shaping scripts which are developed only as needed*.

Take the well-known example of going to a restaurant. A planning process would look up instances of 'restaurant' things or the 'dining' act. This would immediately provide information on all the types of restaurant which have ever been visited and how many of these visits have occurred, and also on how greatly dining out figures in the general dining activity. If the planner wanted a particular ambiance or a cuisine it could use these requirements to map over either set of instances. If it knew that it sought the script for a trip to MacDonald's, it could recall every instance in which that visit has been described; and should it know physical site of the subject for which it is planning, it might even be able to select a subset of occurrences which would tell how to get to the nearest location of that restaurant chain.

This is a static situation. Assume that the planner is being provided with feedback on the execution of its plan and that we are somewhere in the middle of the plan; the food has been placed on our tray and we have paid for it. What to do next? If we have moved the working set of instances into main memory, we can very quickly determine that in previous cases people have sat down at a table. If it is lunchtime and the planner finds that every table in the restaurant is full, past experiences might not offer any guidance because

* only for planning. The representation's very structure guides how word senses are understood.

this circumstance has never happened before. In such a case it could search the general set of restaurant or dining cases for a similar occurrence or it might investigate alternate paths in its McDonald's scenarios. The first course of action might show that standing and waiting for a table to open up is an option; the second, that on other occasions food has been taken out of the restaurant in a bag and eaten elsewhere.

This representation's organization makes it particularly useful for developing scripts, so useful in fact that it would be hard to improve on it for that purpose. This is understandable because all scripts are generalized sequences of behavior, and particular sequences of behavior are precisely what this knowledge base is designed to represent.

Understanding Cause and Effect

Explicit consequences expressed in language by words like 'because' appear in the representation as .causation relationships. User processes can recall whatever .causation relations occur in the context of the current .thing or .act of interest and build causal models by moving along the appropriate linkages.

Unfortunately cause and effect is frequently implicit in the sequence of acts that compose an utterance, ie "The storm blew fiercely. The shingles flew off the roof". It is difficult to distinguish a case like this from the more common situation of successive events which are not intended to be understood causally, ie "The storm blew fiercely. The man walked down the street". Sequences of acts like this can be recognized as causal if other instances pairs are joined in a causal relationship a statistically significant number of times. These occurrences can be identified by traversing the shorter of the two words' symbol lists and matching on .causation. This assumes that causal relationships have been explicitly mentioned in previous utterance. This is psychologically plausible; in addition to learning by seeing (which the computer cannot do), children also learn causal relationships by being told about them, especially if the relationships are not overt.

In addition to inanimate sequences of proximate cause and effect another class of

causal relationship expresses how people consciously act in pursuit of their goals. Since goals are achieved as a result of the effects of actions, human behaviour can be usefully viewed from a causal perspective whenever it is deliberate*. In other words, if a person's plan has been disclosed and a sequence of his actions are observed to result in it (or to fail), we have witnessed a new instance of a script. Alternatively, series of human actions that result in good outcomes can be inferred to be goal-oriented§; getting a cheque at the end of the week because you go to work every weekday is an exemplary script.

Acquiring Commonsense Models

After an extended period of recognizing causal relationships it should be possible to integrate and summarize the results in models or systems of commonsense knowledge. Areas in which results are likely to be forthcoming are: naive physics (gravity, state kinetics), natural processes (fire, weather, aging), vital processes (homeostatic needs), human motivation (values, goals), social structure and process. Because these structures become elevated to the level of objective laws once they have been validated, an argument can be made for expressing them as autonomous knowledge structures. On the other hand every general rule has a host of particulars that affect its use; we don't carry umbrellas when it's snowing. So to use these commonsense rules with some finesse, the instances which gave rise to them will be needed to particularize them.

User Modelling

Among the symbols of things in the representation are ones for individual human beings. Every reported act in which a person has been involved as subject or object or through a more peripheral relationship, appears in the representation and is recorded in the

* however not all human actions are consciously directed. On one level people can be treated just like inanimate objects, ie "I pushed him and he fell over". The falling here is not consciously willed.

§ with somewhat less confidence, both because the existence of a plan has not been explicitly stipulated and also because sometimes good things do happen by chance.

.relation category of the person's symbol. This is also the case for all monadic relationships which specify an person's particulars. Given this prepackaging of information it is a short step to construct a model of an individual person's details, behaviors and circumstances. In addition to knowing that John-21 is six feet tall, 27 years old and lives in an apartment on Riverside Drive, we could identify that he waterskis three Sundays out of four in summer months and goes to MacDonald's far less frequently than the average.

Adopting Viewpoints

By adding certain new features to .existence, the representation can not only support an objective model of an individual person but also a subjective model of that person's viewpoint insofar as this was reported in utterance. The new values of .existence required are .utterance, .receipt, .belief, .knowledge and .disbelief (unimplemented).

.Utterance and .receipt apply to communications, statements or propositions which an individual transmits and receives through a medium, rather than ones which he experiences directly. A communication has no inherent truth value; a person can tell a lie himself and disbelieve what he is told* or he may be in unwitting error. The remaining three features express the truth values a person assigns to propositions once he has received them. .Knowledge asserts the inarguable truth of propositions which a person has experienced directly§; it is the default value. .Belief and .disbelief‡ apply to propositions which someone learned in time past either by report or by inference. These assessments of truth could be qualified by degree.

These existential features would be concatenated with those which express the sense of a proposition's .act. For example, the sense of a statement that a person is

* "I never believe what I read in the papers."

§ "I saw it with my own eyes."

‡ .disbelief could be subdivided into cases which could be true but are believed to be false, and cases which are thought to be imaginary in all contexts.

reported to have heard would be tagged as a .receipt and a truth value (.belief, .disbelief) would be added if this is explicitly mentioned in the utterance. Indirect communications would have repeated attachments for each stage of report.

A person's subjective viewpoint would be modelled* by evaluating the propositions which a person knows in terms of whatever known or inferred qualifications he has about their truth. In addition to explicit mention in the utterance, a person's .belief or .disbelief in a statement could be inferred from his actions after its receipt (or utterance). Once a viewpoint has been created it should be then possible to make gross predictions about the person's beliefs or acts with a limited but measurable degree of confidence§.

Limitations

Absence of Concrete Experience

Word sense definitions in the system can only describe and not fully define what people mean when they use words. A lifetime of sensory experience gives people knowledge of something which cannot be articulated but only pointed to by objective language symbols. Until machines possess equivalent sensory modes, particularly vision, they will not have comparable experiences.

It is hard to overstate the pervasiveness and importance of concrete knowledge. Here is an attempt to express one small and familiar bit of it: we know which physical things can be made to stand on end unattended and which will fall over. Of the things that

* creating a viewpoint would be an expensive operation, and once it has been done the results should be retained. One interesting way to do this would be to copy the person's subjective fact base to an environment and spin off a child process (or autonomous processor) to create and operate the person's viewpoint. The child process would be allowed to use objective information in the representation but would not be given access to the viewpoints of other people.

§ part of the limitation on confidence is due to the fact that most representations will contain far fewer subjective statements than objective ones, and therefore offer less information on which to base predictions.

fall over, we know which will never stand up under any circumstances and under what circumstances others can be made to stand. We know the signs things make when they are about to fall over and what can happen to them when they do. We know what can cause things to fall over and how to conduct ourselves so as to avoid this happening. We know what to do when something is falling over, whether we can stop it or only minimize the effects of the fall. We know these things through sight and touch and even hearing*. We learned them at an early age and we use them quickly, routinely and unconsciously every day.

Structural Rigidity

Modifier senses are incorporated into the symbols of the things or acts they modify, the representation could not be efficiently accessed on the basis of a symbol's attributes. Each symbol would have to be inspected for instances of the sought attribute on its relation category.

One possible strategy to improve access via attributes would be to produce and regularly update statistics on the frequency of modifiers co-occurring with symbol classes as a byproduct of other searches. Although searches for attributes would be undirected during the period before these statistics are first accumulated, co-occurrence frequencies could be used in heuristics to speed searches thereafter.

Since it is unlikely that acts or things correlate significantly with articles, conjunctions, prepositions or pronouns except in idioms, the strategy for modifiers cannot be used here. It remains to be seen whether anyone would want to investigate isolated uses of 'because' or 'the'.

* try to describe in detail your sensory experiences when balancing a pencil on its end.

Remarks

Because the semantics in the system are explicit, processes can use the representation directly or create inferential structures from it with a substantial degree of confidence about just what its tokens mean. For processes other than linguistic ones to use and modify the representation, it must satisfactorily express not only linguistic meanings but also some form of whatever knowledge planners, reasoners and other user applications require and produce. In effect, it must serve as a universal knowledge representation.

As the above remarks suggest, user processes would not necessarily employ the representation's contents directly in their own computations. Reasoning about spatial location might begin by abstracting the relevant elements from matched cases and into an *ad hoc* knowledge base which would differ from the representation in being temporary and purpose-built; its organization would not be known by other cognitive processes and it would be discarded once its owner was done with it. The representation on the other hand is permanent and has an agreed-upon structure that all processes understand.

For all its size and complexity, the representation is no different than any other data structure. Tools are required before any intelligent use can be made of it. The first ones must be dedicated low-level processes to access, search and manipulate its contents. These include the already-assumed functions to access instance symbols by translating dictionary pointers into their storage addresses.

A powerful pattern matcher would be needed to evaluate the complex networks in which instances are embedded. Because the instances it would test and the case it would test for are both articulate structures, a matcher could operate in parallel effectively. Its processor nodes which would be assigned each element of the test case; individual senses if possible, symbols if not. The stream of cases supplied by the access functions (which

could be steered by heuristics) would be filtered through the nodes of this matcher in a synchronous operation*.

It would also be useful to have an inference engine automatically set to work to construct a representation of the temporal and causal dependencies in any fragment of network as soon as it is retrieved (Temporal relationships would be inferred from acts' .time and .perspective values and the history list). The resulting representation would be available to whatever process invoked the retrieval, and if inferencing occurred independent of pattern matching in a distributed environment there would be little overhead so long as the time complexity of the two processes was comparable.

Even with a hardware pattern matcher and inference engine, the representation's ever-increasing size will make large scale searches costly. Good design principles dictate that the results of each search and use should be memoized in some way that would shorten a repetition of the same or a similar search.

The wide variety of useful intelligence-modelling processes that could use the representation produced by the kind of linguistic system envisioned here (a system which is not the only one that could be based on the semantic feature system) demonstrate the advantages of developing an elaborate model of real-world semantics. This is not a revelation. However what may be particularly noteworthy about the present project are the notions of 'feature' and 'domain' (and the perceptions of salience, similarity and orthogonality that underlie them) and the idea of treating decompositional semantic elements not as simple tokens, but as nodes in a complex network that represents the enduring relations of semantic implication, contraindication and possibility as well as sort and composition in the real world.

* it is possible that the complexity of the text case would exceed the capacity of the pattern matcher, ie the number of significant nodes (or senses) would exceed the number of matcher nodes. Screening would then have to be done on a serialized basis; should a tested case match the first assignment of nodes in the test case, the nodes would have to be reassigned and both cases retested symmetrically to determine the ultimate judgment.

A.1

Glossary of Terms and Conventions

The discussion employs a number of terms with exact but unconventional meanings. Only terms used throughout the entire discussion are presented here; others used only in particular sections are defined where used. Terms like 'phrase' and 'clause' with exact and widely used meanings are not defined when their conventional meanings are intended.

Alphabetical order is usually random order, and it is difficult to assimilate definitions from a randomly ordered list, particularly if most of its entries are unfamiliar. People also absorb complete concepts more easily than individual facts. For these reasons terms have been gathered together into nine related groups and are discussed here as a single unit. These discussions provide the skimmer with a synopsis of the entire system. In each heading the title is followed by the terms introduced in alphabetic order. A term's synonyms appear in parentheses behind it, and a term with a suffixed numeral has more than one sense.

When first used, terms defined in a discussion appear in upper case and terms defined in a different discussion are marked '(defined elsewhere)'. If only one sense of a term with multiple senses is used in a discussion, that sense can be assumed when the term appears without a suffix.

In addition to terminology, a number of conventions are used to organize material succinctly in the body of the text and its appendixes, and in the systems that are presented therein.

Terms

Ontology: *entity (thing), formal relationship, informal relationship, individual, natural kind, relationship(2)*

An **INDIVIDUAL** is simply something which can be distinguished from its background sufficiently to be named. This most encompassing and least descriptive term includes both entities and relationships. **ENTITIES** or **THINGS** are individuals which persist and have a coherent and unitary structure. **RELATIONSHIPS(2)** are connections between two or more things which derive from them and share their character (and sometimes their names).

Both of these notions can be meaningfully subdivided. **NATURAL KIND THINGS*** are material things in the real world. They are tangible and are known directly through perception. Immaterial things like abstractions and imaginary or hypothetical entities are derived indirectly from our perceptions and concepts of natural kinds through some mental process such as generalization or reification. **FORMAL RELATIONSHIPS** are expressed linguistically by verbs. They can incorporate other relationships in their senses and usually convey a wealth of concrete semantic detail[§]. **INFORMAL RELATIONSHIPS** are realized by conjunctions and prepositions; they are more generally used and have less detailed senses. Possession, substance and spatial position are typical relationships expressed informally.

Structures: *category(1) (class), graph, criterial aspect (criterion), distance, head(1) (root), hierarchy, inheritance, inverse relation, level, line (linear relationship, scalar relationship), linkage, member(1,2), network(1), peer, pole, relation, relationship(2), sibling, slot (role), tenant*

A **RELATIONSHIP(1)** names one or more meaningful connections between pairs of individuals⁵³. Each connection is expressed through **RELATIONS** which are attached to each participant. The components of a directed relationship are **INVERSE RELATIONS** and its participants experience each end differently. Individuals connected by relationships are **MEMBERS(1)** of various organizations or structures. A **CELL** is a single unit in a

* relationships between natural kind things are also called 'natural kind'

§ the parties involved in an act or situation, their attitudes, the instruments used, the movements, situations etc.

structure, one instance of a relationship together with the individuals that participate in it.

Examples may help clarify these notions. Composition is a directed binary relationship composed of two relations which can be called **-part-of-** and **-has-parts-**. Conjunction names an undirected n-ary relationship whose relations can be called **-together-with-**; as this example demonstrates, an undirected relationship cannot be distinguished from its relation parts. Between is a directed ternary relationship whose constituent relations can be called **-at-end-** and **-intermediate-**. One participant has **-intermediate-** relations with each of the other two; they each have one **-at-end-** relation with it.

The **HIERARCHY** is a common organizational form composed of one or more cells of three or more individuals in which one individual has the same directed relation with the others. The identified individual is the **HEAD(1)** of the hierarchic relationship; individuals which have the associated inverse relation with it are its **MEMBERS(2)**, and **SIBLINGS** of one another. Members(2) in a single hierarchic relationship together compose a **CLASS** or **CATEGORY(1)** which is identified by its head. 'Cell' suggests a hierarchic relationship's structural aspect and 'class' suggests its semantic character.

When hierarchic relationships occur repeatedly within a group, certain individuals are members(2) in one relationship and heads in another*. In a multiple hierarchy, **LEVEL** counts the number of similarly-directed connections between one individual and another (by default the head of the entire hierarchy), and **DISTANCE** counts the number of links between two individuals without regard to their direction. Serial directed relations establish **LINKAGES**, series of individuals connected by the same or a collateral relation, and transitive linkages permit one individual's characteristics to be passed on to its members(2) by **INHERITANCE**. **PEERS** are members(2) of different cells on the same level, and

* this leads to different perspectives on such individuals; seen as single in one hierarchy, they are viewed as lasses in the other.

TERMINAL MEMBERS do not head other relationships. In addition to its structural dimension, every hierarchic relationship has a semantic characteristic which all of its members(2) possess in varying ways. Thus color is a **CRITERION** of material things and different material individuals are red or yellow or blue. This characteristic is the **CRITERIAL ASPECT** used to select and distinguish among the individuals in a category.

A **LINE** is a simpler structure than a hierarchy. All individuals in a **LINEAR** relationship possess the same semantic characteristic and are connected serially to one another in an order determined either by relative degree of the criterial aspect or by arbitrary convention. Terminal positions in a line are called **POLES**. Linear relationships are also called **SCALAR** to emphasize they incorporate relative measurement.

COLLECTIONS are structures of connected individuals such as tastes, smells and tactile sensations which are not organized in any regular way. They are infrequent; it seems to be part of human nature to impose some sort of order however arbitrary on things so that they can be more easily manipulated: days of the week, signs of the zodiac, letters of the alphabet. All three sorts of structures occur alone in simple instances and together in composite ones where, for example, a structure's top level might be a line and each of its value might head a hierarchy. They are universal structures with known and predictable characteristics. A fourth kind of structure arises from specific cases or **INSTANCES** of relationships like those between features in a word sense (discussed elsewhere) or word senses in a sentence. **NETWORKS**(1) are regularly formed but unique constructions of instances of individuals (nodes) and relationships (arcs).

Networks are used to represent the entire feature system and the meanings of particular utterances. **ROLES** are the mechanism by which otherwise unrelated things associate with one another in utterance networks. Relationships depend on things* and roles are generalized things which abstract relationships incorporate to express themselves.

* although we know things directly, we know relationships only through things that embody them.

Roles are semantic variables which have only the minimal character required to express the relationship they embody. In a particular sentence instances of things fill relational roles and become TENANTS of the corresponding relationships. Roles can be found among things as well as relationships; pronouns express roles which are filled by the thing which is the pronomial reference. In frame terminology SLOTS are analogous to roles.

Semantics: *aspect, attribute, character (nature), consistency (semantic consistency), quality, similarity*

Every concept possesses a semantic CHARACTER or NATURE which is a unique selection and organization of ASPECTS. Aspects are the raw materials from which we construct meanings. They are unique, distinct and coherent elements of meaning and in the absence of a finished analysis must be considered divisible. Aspects function as the semantic alphabet and concepts are the words which they spell*. Features and utterance senses (discussed elsewhere) are the counterparts of these two mental phenomena in the linguistic system.

ATTRIBUTES are a particularly important sort of aspect. QUALITIES such as color, pitch and texture are the components of actual perceptual data. Attributes are the permanent records that we make of qualities before they are replaced by the onrush of new perceptions. Qualities and attributes have similar roles in their respective spheres; qualities are the direct or indirect basis for all mental constructs§, and all other aspects are ultimately derived from attributes by mental processes such as generalization (discussed elsewhere).

Two features' capacity to occur with one another without type clash is called SEMANTIC CONSISTENCY (CONSISTENCY for short). Consistent features have the same linkage; inconsistent ones diverge somewhere along the line. Consistency seems to

* the fact that 26 letters compose many more than 50,000 words gives me hope that the current ratio of three features defined for every two words will eventually change.

§ the relationship between a concept and its qualities is what Searle (1959 in Lacey, 1986) calls determinacy. I call it specification; hyponyms specify a head's value in a particular situation although we know things directly, we know relationships only through things that embody them.

be primarily based on an individual's position in the natural kind hierarchy, and individuals which are consistent in domains that have a physical expression (like appearance or manner) invite employment in figurative speech. The degree to which a pair of consistent features agree with one another can be compared with that of other consistent pairs. This is SEMANTIC SIMILARITY or SIMILARITY ⁵⁴. If a convention is established, similarity can be measured against an objective standard. ⁵⁵

Kinds of Relationships: *accidental association, association, hyponymy, meronymy, necessary association, part, sort (kind), whole*

A HYPONYMY is a hierarchy of individuals organized by their sort; its members(2) (described elsewhere) are particular SORTS or KINDS of their categories. This organization does not employ a consistent criterion throughout; instead each cell (described elsewhere) classifies its members on whatever aspect is most striking or useful under the circumstances. I.e. a zoological instance physical form could be used on one level and behavior on another. Hyponymys emphasize structure because they have no inherent semantic character. They treat individuals as discrete and independent, and categories as sets which are exhaustively partitioned by their subset members(2). As a result interior cells are empty and individuals are only found in terminal cells, where they acquire the criterial characteristics of all superordinate cells by inheritance.

A MERONYMY is a hierarchy organized on the basis of the part-whole relationship. It uses the same criterial aspect in all of its cells; the head of each cell in a meronymic hierarchy is a WHOLE whose members(2) are its PARTS. From the perspective of the relationship the members of a meronymy are not independent of one another (by definition a whole is its parts). Therefore cells on interior levels of meronymic hierarchies can have actual individuals as members(2). Inheritance operates selectively rather than rigorously; parts do not share all the characteristics of the wholes they compose. Because of an inherent vagueness in what constitutes a part, a meronymy can fail to

classify its members exclusively; they can sometimes be empirically identified as belonging to more than one class.

Hyponymys are more flexible and powerful classifiers than meronymys because their organizational criterion is semantically unspecified. One result is that composition can be treated as a particular type of sortal relationship. Another is that hyponymic hierarchies are so frequently used that they have become the default taxonomic structure⁵⁶.

Associative relationships are the mechanism by which features are combined in particular word senses and words senses in particular utterances. ASSOCIATION is simply the meaningful co-occurrence of particular individuals, where meaningful stands for congruent with sensory experience and therefore with the concepts we base on it. This relationship has a strength characteristic; associations can be either strong or weak, necessary or accidental.

NECESSARY ASSOCIATION occurs when the existence of one feature is an inevitable consequence of the existence of another feature. It is a directed relationship. ACCIDENTAL ASSOCIATIONS are contingent; a feature or sense can but need not occur with another feature or sense. Thus .color co-occurs necessarily with .flower because all material things including flowers have color*, while .red co-occurs accidentally with .flower because not all material things or flowers are red. .Red is a value of the .color domain and the example correctly suggests that accidental associations frequently exist between a feature and members(2) of a domain with which it has a necessary association; the member(2) specifies the particular value of the necessary association.

Feature System: *contingent domain, domain, feature, key domain, natural domain, orthogonality, protean domain, subdomain, value(1).*

A FEATURE is an element of meaning. Features are linguistic objects whose mental

* the argument that some things like clear water have no color can be defeated by noting that we treat transparency as if it were a color, routinely contrast water's clarity with, say, the blackness of ink.

counterparts are aspects (described elsewhere) and whose most important characteristic is their salience or obviousness of meaning*. We cannot heighten a well-chosen feature's meaningfulness by amalgamating it with others or by dividing it into parts.

Features are organized into hyponymic and linear structures called **DOMAINS**. A domain's members are called its **VALUES**(1). Each value that heads a cell (described elsewhere) in a domain is also the root of a **SUBDOMAIN** whose members are all of its descendants (described elsewhere). Viewed from another perspective, domains are consolidations of subdomains of increasing size. This view correctly suggests that domains are transparent structures; each head of a subdomain is directly accessible to association with values from other domains, and each subdomain can be treated as a domain offering linkages and inheritances to co-occurring features.

The process of embedding subdomains within one another cannot go on indefinitely or all features would belong to a single domain. **NATURAL DOMAINS** are headed by a feature which is not a member of another domain. Natural domains seem to express the intrinsic structure of semantic space. **CONTINGENT DOMAINS** on the other hand are headed by features which belong to some other domain. For this reason they usually have relatively specific senses and can only be applied to certain values in the domain on which they depend. Often they are attributes. Contingent domains are not subdomains; they are truly separate domains and as such have a character which their members share and which is not intelligible in terms of another domain. Domains, natural or contingent, are independent and interact orthogonally. **ORTHOGONALITY** describes how a domain qualifies or extends the sense of domains with which it associates in a way that is radically different from the way that the domain's own members express its character.

Two kinds of natural domains deserve mention. **PROTEAN DOMAINS** are the opposite of contingent. Instead of a highly defined sense and limited application, these

* a feature can be expressed in writing, an aspect cannot.

domains are employed in many contexts and are usually interpreted somewhat differently in each. They seem to provide a structural skeleton or framework for the expression of concepts. **KEY DOMAINS** are another type of natural domain. These highly populous and well elaborated hierarchies seem to be natural indexes to our mental organization.

Cognitive Functions: *abstraction (generalization), apprehension, categorization, comprehension, concept, conceptualization, impression (perceptual data, sensory data), perception (sensation, sensory experience)*

We constantly receive **IMPRESSIONS** of the world around us. These transient **PERCEPTUAL EXPERIENCES*** are the raw material from which permanent mental representations or **CONCEPTS** are constructed. Concepts of natural kind things and relationships are based on **ABSTRACTIONS** or **GENERALIZATIONS** from significant consistencies found in different perceptual experiences; other sorts of concepts are subsequently derived from natural kind ones by further generalization. All mental constructs are thus ultimately based on perception[§]. The entire process of creating mental constructs is called **CONCEPTUALIZATION**. On the basis of the particular significant consistencies which they inherit from their instances, concepts can be **CATEGORIZED** or organized into classes. It is significant for semantics that natural kind categories seem to inhere in the world rather than the individual.

Apprehension and comprehension are two mental activities involved in processing language. **APPREHENSION** identifies the literal senses of the words constituting an utterance and constructs an understanding of the entire utterance from them. This involves selecting the correct sense if a word has more than one and discarding arbitrary factors like syntax. **COMPREHENSION** is the use of apprehended (or other) information for some

* studies (Goldfarb, 1962) have shown that sensory deprivation limits the development of a child's personality and attacks the integrity of an adult's consciousness.

§ sensation, perception and their many synonyms are all vague terms, sometimes naming a process, sometimes its result.

particular cognitive purpose. The two operations often occur in tandem because a person usually receives utterances in course of performing some task.

Linguistics: *author (speaker, utterer), category(3), dictionary, autonomy, episode, grammar, head(2), lexical gap, lexicon, locality, linguistic expression, reader (hearer, recipient), sense(1)(meaning(1)), statement, text (narrative), token, unmediation, utterance.*

Natural and other languages are composed of a LEXICON or set of words for which meanings have been defined, and a GRAMMAR or set of rules specifying how these words can be combined to produce meaningful constructions. A DICTIONARY differs from a lexicon in attaching meanings to the words it lists.

Sentences are composed of TOKENS, which are instances of words. The same word can be used as the basis for more than one token in a sentence. A word may have more than one meaning and may be used as more than one part of speech. Directly or through a root, words label dictionary entries which contain their meanings. Each SENSE(1) or MEANING(1) is a individual semantic construct differing from other meanings to a greater or lesser degree. The set of meanings attached to words do not begin to exhaust our conceptual set; many senses can be expressed discursively for which no single word or phrase exists, ie back-of-the-neck. Such cases are called LEXICAL GAPS and highlight the difference between words and meanings.

The HEAD(2) of a phrase or clause is its most important word or the word on which others in the expression depend. The other words in a phrase are LOCAL modifiers of the head. Modifier phrases without a head are AUTONOMOUS. An UNMEDIATED phrase or clause is one which is not preceded by a preposition, conjunction or relative pronoun. Unmediated noun phrases can only fill a verb's positional cases (subject, object and direct object). The eight parts of speech are called CATEGORIES(3) of syntax.

A STATEMENT is a single sentence, an UTTERANCE one or more sentences, and a TEXT or NARRATIVE more than one sentence. A LINGUISTIC EXPRESSION is a

sequence of words of any length or syntactic structure. An EPISODE is a series of sentences which have related content. UTTERERS produce utterances which are received by RECIPIENTS. Although there are significant differences between the character of spoken and written language, in practice the differences are not important. SPEAKER and AUTHOR can therefore replace utterer, and HEARER and READER replace recipient.

Research Fields: *artificial intelligence, cognitive science, computational linguistics, computational semantics, natural language processing, natural language understanding.*

NATURAL LANGUAGE PROCESSING (NLP) studies the production of computational representations of utterances. NATURAL LANGUAGE UNDERSTANDING (NLU) is a broader term covering the use as well as the production of language representations. COMPUTATIONAL LINGUISTICS is an equivalent term which also identifies computer-based research into issues of library science and literary criticism. All three terms are used interchangeably, with Europeans rather than North Americans tending to use computational linguistics to name the entire field. COMPUTATIONAL SEMANTICS is a subfield which deals with semantic analysis and representation *per se*. ARTIFICIAL INTELLIGENCE (AI) encompasses computational study not only of language processing but also of other cognitive and perceptual functions. It is the subdiscipline of computer science whose investigations of intelligence are carried out in symbolic terms meaningful to people rather than machines. COGNITIVE SCIENCE is an interdisciplinary crossroad between computer science, philosophy, psychology and linguistics. Although cognitive science emphasizes theory and artificial intelligence emphasizes applications, both fields employ the scientific method and use the computer as their primary research instrument.

Demonstration Program: *alist (association list), category(2), clist, DICTIONARY, elt, flat list (flist), key, LEXICON, list, link, lyst, meaning(2), mlist, NETWORK, network(2), pair, sense(2), symbol, value.*

SYMBOLS are Lisp data objects used as components in the program's permanent databases and as temporary variables for processing. The most important databases are the **LEXICON**, a semantic network representing the feature system, the **DICTIONARY**, a list of words described in terms of features, and the **NETWORK**, a semantic network expressing the sense of the episode currently being processed, where symbols embody instances of entities and formal relationships, (ie .things and .acts, nouns and verbs).

Different sorts of symbols are identified in the program by parallel sets of terms, one emphasizing structural form and the other semantic content. Structural terms incorporate the root '-list' while semantic terms are descriptively named. These terms can be used to describe both entire symbols and components within symbols.

The simplest kind of symbol or component is the **FLAT LIST** or **FLIST**, a list composed of atomic data objects. Its semantic counterpart is the **PAIR**, a series of related features. The first atom in an flist is called the **KEY** because it is the item usually inspected in searches. None or more additional atoms can follow the key; these are the pair's **VALUES**. A **LINK** is a particular type of flat list used to embody relationships between symbols. It names a relation and another symbol in the environment. A link is always matched by another link in the named symbol pointing back to it. Links in directed relationships usually include a .slot value specifying the relation's direction.

A list of flat lists is called an **ALIST** or **ASSOCIATION LIST***. The semantic equivalent is a **SENSE(2)**, which is the series of pairs defining a single word sense. A list of alists is called an **MLIST**. Its semantic counterpart is the **MEANING**, which contains one or more alternative senses for a given part of speech. Mlists are the highest synthetic symbol component. When analyzed from the top down, the **CATEGORY** or **CLIST** is a symbol's first level of substructure. It is often equivalent to an mlist although some categories are simpler.

* 'association list' has its usual Lisp meaning and is included here simply for symmetry.

In general structural terms a symbol is a nested object of unspecified depth, a LIST is a series of items which may be ordered in some way, and a semantic NETWORK(2) is a graph composed of nodes (defined elsewhere) joined by arcs (defined elsewhere). LYST and ELT are terms that apply to data objects whose structure is unknown. A lyst is a list of unspecified depth, while an elt is a data object of indeterminate structure, frequently an atom. The structure of symbols representing relationships is complicated by a need to express restrictions on their tenants. Restrictions are clists which are keyed to the .case to which they apply. The net effect of this extension is to produce symbols which have six levels of nesting rather than the usual three.

Conventions

Textual. Examples are enclosed in quotation marks to show they should be considered as linguistic tokens rather than a part of the discussion. Single words and phrases appear inside single quotes; sequences of words understood as clauses or sentences are enclosed in double quotes with initial capital letters but without terminal punctuation. Within an example missing words which must be supplied for a paraphrase to make sense appear in angle brackets. ?, ?? and * have their usual meanings of denoting increasingly ill-formed expressions. These symbols are used primarily in Appendix 3.

Occasionally a sense must be identified directly rather than through its realization in a word. One and two word senses are embedded in hyphens, ie 'ball' is -spherical-object-, while a sense composed of more than two words is strung together with hyphens, ie .multiple means more-than-one. An utterance's syntax can be expressed by concatenating with plus signs the conventional contractions of the parts of speech involved. 'The little boy' becomes art+adj+noun.

Concordance. The second half of the system discussion is intended to present an abstract specification of a feature system as well as to document the particular system implemented in the demonstration program. The abstract and implemented systems do not agree in all particulars. It is easier to describe a functionality than to code it so not all capabilities important enough to require discussion actually appear in the program. Features which have not been defined are marked (undefined) and processes which have not been coded are marked (unimplemented).

Feature Names. A feature's name is chosen to evoke its character as exactly and as strongly as possible. Feature names are unique and, with a few exceptions, synonyms of common English words. Exceptions arise when no common word expresses a feature's character adequately or when the most appropriate common word is polysemous and more than one of its meanings is used in the feature system. These two cases are handled* respectively by affixing syllables to coin names suggestive of the correct sense (.disjunct, .indobj, .prpart), and by changing the case of letters in the name (.person, .Person). Although desirable it proved impossible to keep the part of speech (adjective or noun) used to name features consistent within a cell.

Because so many feature names are common words, each is prefixed with a dot (ie .class, .type) to avoid confusion with literal use in the text. Although features in the program are always singular and almost always lower case, capitalized or plural forms are used in the discussion as appropriate.

Feature Notation. Relationships between features can be expressed in a notational form. In the abstract system discussion, each subsection presenting a particular domain begins with a formal expression of the relationships between the features involved.

* should these techniques prove inadequate at some future date, another alternative would be to suffix a number identifying the intended sense.

The notational system used follows standard conventions. The colon : symbolizes relationship and indicates that the feature or features following it are related to the feature preceding it in one of a variety of ways. The nature and scope of each of these relationships is shown by marking the features affected by it in some way: either by changing their case, or by delimiting them with opening and closing symbols, or both.

Among lower case features, unenclosed ones are hyponyms of the preceding feature; ones enclosed in parentheses () are its meronyms; features in braces {} are its qualities or attributes. Capitalized features all have an associative relationship with the preceding feature. Unenclosed capitalized features are necessarily associated with it, while ones enclosed in some way are associated with it accidentally. A feature enclosed in parentheses 'accepts' accidental association from the preceding feature and one in braces 'affects' the preceding feature accidentally. Features separated by a slash / are correlates of one another in addition to whatever other relationship they have with the preceding feature. Because 'character' is metasemantic and monadic, it is not expressed notationally.

Program Symbol Names. Network instance symbols are suffixed with an index which identifies their position in the semi-chronological sequence generated from a given word (boy-01, run-01). Angle brackets identify variables. Symbols produced from input tokens are enclosed in horizontal bars ([the], [man]). Lisp data objects appear in upper case (DICTIONARY, LEXICON). Lisp function names always begin with a capital letter (Apprehend, Instance?); if they are composed of more than one word, second and successive words are connected to one another with hyphens and may be capitalized (Reset-Symbols, Get-Longest-Phrase).

Figures. The notation for linguistic symbol structure follows simplified railroad chart conventions. { ... } stands for none or more occurrences of the equivalent structure that would occur between a pair of parentheses on the same level. { aaa | bbb | ... } stands for one occurrence from a set of alternates. < name > stands for any one of a class of tokens.

A.2

Arguments for A Universal Semantic Representation Based on Features

Two Hypotheses

These two hypotheses articulate a theoretical foundation for the thesis. The first states that a single well-founded representation can support a variety of artificial intelligence tasks, and the second, that a set of semantic primitives with rules for combining them can be the basis of such a representation. Although the second hypothesis is contingent on the first, the reverse is not true; there is no reason why a universal representation must be based on semantic primitives.

For this reason the hypotheses are dealt with independently and in order. The implications of each are discussed together with whatever grounds for its acceptance I can marshal. It has been noted (Minsky, 1975) that assertions in artificial intelligence are difficult to prove formally. Informal methods of proof such as demonstration and persuasive argument must usually be employed exclusively or as a supplement to an attempt at formal reduction and that is the case here.

Arguments take for granted that human practice is significant insofar as it shows one way to do a cognitive task², and that appeals can be made to the principles of simplicity and symmetry in nature. There is not much doubt that human evidence is pertinent to computational language processing*. It is up to the reader to determine if natural principles can be invoked convincingly.

natural language can be translated into a universal or canonical representation usable by other artificial intelligence processes

Support for this sort of hypothesis cannot be demonstrated easily. The only convincing

* King (1988, p.13) suggests that a universal representation must be "psychologically plausible".

demonstration would be for a number of systems in various artificial intelligence subfields to use the same language representation. Alternatively, the representations used by various applications could be shown to be reducible to a single form. Both tasks are too time consuming; Only when a substantial community of users take up an idea like this is it subjected to the kind of thorough and comprehensive use that permits a judgment to be made about it. The arguments that follow can be summarized as: people use a universal meaning representation, languages are already symbolic so the issue is technique, and current trends in computational linguistic research are supportive.

Different Languages, Universal Meanings. English and French are different languages but they have the same expressive power. Nothing stated in one cannot be expressed in the other; translations demonstrate this equivalence. Languages from other cultures which are incapable of the same range and delicacy of expression simply have impoverished lexicons. Different languages demonstrate that words are arbitrary and meanings are universal. It is a small and step from there to the idea of a universal representation which transcends language differences.

Universal Nature of Language Generation. Introspection suggests that language is generated in the same way whether we are describing the results of ratiocination or an experience arising from memory or perception. If this is so, principles of simplicity and symmetry support the idea of a common analytic process and a universal rather than task-specific representation as a result. Supporting evidence can be found in the fact that although sensory modes are perceptibly different, all kinds of sensations can talked about regardless of their mode.

Utterances Have Literal Meanings. I believe that the semantics of written natural language are self-contained or acontextual on the discourse level (Appendix 6). This

implies that sufficient meaning lies in an utterance's words to permit its literal meaning to be understood even if the utterance is considered in isolation. This suggests that a representation of literal meaning is also acontextual and therefore universal.

Language can be Symbolized. The alphabetization of spoken words to produce writeable equivalents demonstrates that a natural language can be rendered in symbols with predictable and acceptable information loss (Appendix 6). Within this constraint, the sense of a natural language utterance can be said to be represented by symbols of the words that compose it. A universal representation would be a type of symbolic representation. Therefore the hypothesis is proved. What seems at first glance to so obviously beg the question—the whole point is that artificial intelligence programs cannot make use of current 'symbolic' linguistic representations—proves on reflection to highlight the fact that the issue is one of translating the current sort of character-string utterance representation into one usable by the computational processes that interest us.

Problems In Current Research Approaches. The Introduction makes the point that computational linguistics is not now a particularly active research area. This suggests that previous efforts have not been fruitful enough to maintain levels of funding in competition with other fields. This in itself argues for new research approaches.

One reason why previous research might not have been too productive is because of the magnitude of the task investigators have set themselves. This hypothesis is consistent with the model of language processing described below which decomposes language processing into distinct tasks which are more tractable and more easily defined, and therefore more likely to yield satisfactory results.

Weak Counterarguments for Task-Specific Representations. One line of argument advanced in favor of custom representations is that general-purpose

representations would be harder to construct (King, 1988). Given the inability of research based on task-specific representations to advance the field in theory or practice, this is a weak argument. Perhaps the effort to produce a general-purpose knowledge base will be offset by some serendipitous advantage elsewhere.

Current Practice in the Research Community. Although many programs that deal with natural language issues take original language utterances as their input, a number of research teams (Schank, Hobbs, Wilensky) already reuse programs in-house, supplying current projects with the parse trees or other intermediate products produced by parsers written for previous projects. The gravitation to standardized representations seems to be an ad hoc policy, since the literature contains no discussion of this as a theoretically significant practice.

*the meaning of natural language can be adequately expressed
using a semantics based on primitive features.*

Unlike its predecessor, the second hypothesis is more susceptible to incremental demonstration and an attempt is made to do this with the demonstration program. The key word in the hypothesis is 'adequately'; if there are no theoretical reasons why the required changes cannot be made, today's inadequacies drive tomorrow's improvements. The following arguments can be summarized as: the issue is word selection because a language is already a representation of semantics, composition has adequate formal power, and syntax and word and utterance semantics are pervaded by composition.

Lexicons are Arbitrary. A natural language is a hermetic system whose kernel is given initially meaning by deixis ⁵⁷. It has been noted as evidence for the first hypothesis that this kernel (and the rest of the lexicon) is arbitrary because it varies with the language. One can therefore imagine a lexicon whose denotations are such that they could be used directly as a set of primitives. Artificial languages like Esperanto are indicative in this regard.

Equal Expressive Power. A compositional semantics has the same power as natural language to elaborate detail without limit, to extend structure endlessly and to generate infinite alternatives. Moyne paraphrases Frege to make the same point in the opposite direction: '[a semantic theory must] be compositional because it must in some way assign a meaning to all sentences in a natural language. Because this is an infinite set, the meanings cannot simply be listed but must be built up in some way.' (Moyne, 1985, p. 208ff).

Ubiquity of Decomposition in Grammar. Syntax distinguishes differences in person, tense, number and gender; pragmatic (illocutionary) purpose is specified by mood; and distinctions of number and gender can be seen to be semantic. If overt language structure contains this degree of composition it may very well contain more.

Utterance Meanings are Compositional. The meaning of a sentence is given by the meanings of its constituent words. We decompose a mental representation to generate utterances and we compose a mental representation to understand utterances. It should be possible to emulate these processes in objective linguistic terms. As Putnam (1975) notes, Tarski's beliefs that decomposition can be separated from composition and that meanings in utterances are *gestalts* defeat the very idea of systematic language study.

Words are Not Opaque Tokens. Whether used in isolation or in a linguistic context, a word's meaning can be described or elucidated verbally. The fact that the equivalent of a single word's sense can be given by a number of words suggests that word senses are decomposable.

One Model

A conceptual framework well suited to the thesis distinguishes linguistic processing from

other cognitive functions and divides it into separate functions for apprehending and comprehending text. This view arises from a model of the human mind in which mental capacities are expressed by distinct processes that cooperate with one another (for our purposes the nature of the mechanism that controls process interaction need not be specified*). This model offers the practical benefits of imposing a structure on cognition§ and of decomposing a single large task into more tractable parts.

Apprehension describes the initial supply-driven process by which the recipient of a text creates an internal representation of its meaning. Because apprehension is limited to the immediate discourse context, it should take linear space and low order polynomial time. Comprehension is the subsequent demand-driven use of this representation by some other processes modelling different cognitive functions. As such it has a more heterogeneous character. In the worst case comprehension could be quite expensive computationally.

The psychological separateness of these two operations and the boundary between them can be illustrated with an example. Imagine a high school student who has just read and understood a algebra problem but not yet undertaken the pencil-and-paper figuring necessary to solve it. At such a moment the student could demonstrate his understanding of the problem by describing its elements and answering questions about it without advancing towards the solution.

Not only linguistic but also general sensory processing appears to proceed on this basis. We have all had the experience of trying to recall a detail of a scene which was unimportant at the time it was perceived but is now significant, ie "where did I put that glove?". There is no guarantee that the data is recoverable, however, because perception by its nature abstracts only a limited amount of the information available in the sensory field.

* control could be decentralized throughout the mind or resident in a metaprocess that focuses attention.

§ compartmentalizing language function means that those who wish to can pay attention to the problems of recognizing and correctly processing difficult and obscure grammatic structures, while others focus on using the resulting representation to support particular cognitive tasks.

If mental functions are separate processes, they must interface through shared data structures to perform joint tasks. The representation produced by apprehension must therefore be well-defined and capable of supporting a wide variety of uses. A number of different linguistic products could be used for this purpose. Parse trees are an obvious choice but not the only one possible; Winograd (1983) identifies three other types of linguistic structure produced by parsers: trees of deeper syntactic structure, systematic semantic structures, and arbitrary structures. The first two are points to which an apprehender could carry the task of processing natural language.

Linguistic problems and complexities can be classed as lexical, syntactic or semantic depending on their context. Lexical problems involve the interpretation of particular words or phrases rather than entire classes. These problems exist independent of context although they are only evident in it. Syntactic problems involve structural relations between words or phrases; they are often expressed semantically (ie in ambiguity) but this is a symptom rather than a cause.

Semantic problems are subdivided into lexical, syntactic and discourse types. Although both lexical and semantic lexical problems involve single words or phrases, semantic problems are syncategoric rather than specific. Semantic problems in syntax occur when a construction's syntax is correct but its sense is ill-formed or ambiguous, or vice versa. Discourse semantic cases involve an utterance's discourse context and, because utterances of this type are both syntactically and semantically well-formed, describe language complexities rather than problems.

Some of these cases can be handled by existing processing methods while others remain intractable. A few examples are not problems at all but are included simply because they give insight into the nature of the English language. Most are borrowed or adapted from other sources (Hirst, 1987; Leech, 1981; Marcus, 1978; Milne, 1986; Pereira & Warren, 1980; Perrault, 1985; Rettig, 1988; Ross, 1967; Winograd, 1983). Discussion are preceded by examples to which reference may be made by a parenthesized index.

Lexical cases

1. Lexical Ambiguity

in syntactic category (*part of speech*) `sink', `saw', `club', `ring'.

A sink is a -plumbing-fixture- noun as well as a verb that means to -disappear-underwater-. Syntactically ambiguous words can in addition be semantically ambiguous within a given category. As a noun, 'club' is a homonym for both -bludgeoning-weapon- and -recreational-association-; within the latter sense, it is polysemous because it can mean both recreational-social-group and -recreational-building-. Taken as a set the examples suggest that syntactically or semantically ambiguous words are often common and short.

In word denotation 'Baker', 'Quebec'.

These words denote a man or a role (1), or a city or province (2). Denotational ambiguity can affect quantification; a baker is a general role but Baker is a specific person.

In abbreviation denotation 'Jan', 'MA', 'BC'.

Respectively, these particles denote a name or a month, a degree or a state, and a province or a period of historical time. Abbreviations differ from words in being able to incorporate punctuation and capitalization which can help distinguish between their different meanings. These aids however may be missing, partial or nonstandard.

2. Lexical Gaps 'oppression', 'aggression'; 'oppress', *'aggress'.

The first two words both appear to be nominalizations but no form of 'aggression' exists to match the verb 'oppress'. This phenomenon of a lack of a word to express a concept is called a lexical gap. Lexical gaps are generally understood to apply to easily understood concepts having collateral concepts which are realized in words.

3. Unrecognized Words 'blivit', 'snood', 'chela'.

People routinely encounter words which they don't know or which are meaningless.

4. Idioms 'redeye', 'tie one on', 'carry coals to Newcastle', 'yuppie', 'ayatollah of popular culture', 'xerox'.

Idioms come in all sizes — words, phrases and clauses. They can include coinages (1, 4) and proper names (3) and may have relatively short lifespans (4, 5). Idioms that

survive long enough pass into the language permanently (6).

5. Subidiomatic Expressions 'just about', "We were just about to leave"; 'enough' +inf, "old enough to know better but young enough to do it".

In addition to idioms based on nouns and verbs or those with substantial concrete sense, there exists a class of subidiomatic expressions which act as facilitators of expression or linguistic shortcuts. These expressions are usually based on modifiers (3) and may have co-occurring parts of speech (4). A subidiom's sense limited and is homologous to the sense of the modifier or the part with which it occurs.

Syntactic Cases

6. Center Embedding "The wallpaper the man your son suggested put up is crooked".

This well-formed sentence is difficult to understand because encountering a subordinate clause forces a reader to suspend his partial understanding of the entire sentence. The possibility of confusing contexts arises as soon as two are stacked.

7. Garden Path Sentences "The horse raced past the barn door fell", "The man who whistles tunes pianos", "The old man the boat", "The cotton clothing is made of grows in Mississippi", "Have the students who failed the exam take the supplementary".

Garden path sentences employ words which can be understood as more than one part of speech. These categorically-ambiguous words appear both within clauses (3 & 5) and between clauses (1, 2 & 4). The punctuation and relative pronoun introducers which would unambiguously identify a clausal reading are elided. Taking the examples in order, 'raced' can act as the head of an adjective phrase and as a past participle, 'tunes' and 'man' can be noun or verb, 'cotton' noun or adjective, and 'have' main or auxiliary verb in imperative or interrogative constructions.

Garden paths often contain intervening phrases or clauses which postpone a reader's revision of his misinterpretation and therefore encourage misunderstanding. Removing an intervening construction can sometimes eliminate an alternative reading and always makes the main clause easier to comprehend. When this is performed on the four examples of this sort, "Have the students take the supplementary", "The man tunes pianos" and "The cotton grows in Mississippi" immediately becomes unambiguous. "The horse raced fell" becomes ungrammatical; only the presence of the interjected prepositional phrase 'past the barn door' allows the adjective to be postposed legitimately. After paraphrasing it becomes "The raced horse fell", which is unambiguous.

8. Extensive Clause Ordination (*run-on sentences*) "This is the cat that bit the rat that ate the cheese that Jack bought", "They came for dinner and they stayed all night and they left right after breakfast but they were still late and were punished when they got to school".

There is no formal difficulty in processing these sentences but in practice they become increasingly onerous to handle as they go on. It appears that we use sentence terminal punctuation as a signal to chunk our understanding. When a sentence ends, we know that we can stop watching out for an entire class of possible syntactic relationships redeploy our attention to the next series of words.

9. Syntactic Ambiguity

in modifier scope 'the old men and women'.

Only context can tell whether this phrase means 'old men and old women', or 'old men and women of unspecified age'.

in phrase reference "I saw him on the hill with a telescope"; "He worked on the table on the paper", "He worked on the paper on the table".

The first sentence has three readings which depend on how the last prepositional phrase is interpreted. If 'with a telescope' is read as accompaniment, it can modify 'him'

or 'hill' and produce "he and a telescope were seen on the hill" or "he was seen on the hill which has a telescope on it". If the phrase is taken as identifying the instrument used for seeing, it produces "he was seen on the hill by use of a telescope".

The second and third sentences show how ambiguity arises when a sentence contains more than one phrase which can play more than one available role. Selecting one role for the first phrase automatically assigns the second to the other phrase. Thus either sentence can be interpreted as saying that a paper is being written by someone sitting at a table, or that a table is being painted or sanded on top of spread paper. 'At' can be substituted for 'on' anywhere.

in coordinated conjunction 'the men and women or boys'.

Depending upon how the scope of the conjunctions is interpreted, this phrase can mean 'either men and women, or men and boys', or 'either men and women, or boys'.

in indirect object sense "John wrote Mary a letter", "We envy you your wealth".

The indirect object is inherently semantically ambiguous. It almost always expresses senses of -directed-towards- and on-behalf-of and is paraphrased as 'to' +indobj or 'for' +indobj. Although context within a sentence usually establishes what sense is intended, this example proves that it is possible to make ambiguous statements; it is not clear whether the letter was written to Mary or on her behalf.

The second sentence is an example of a sentence which cannot be paraphrased in either of the two standard ways. The 'to' reading makes no sense. "We envy you for your wealth" makes sense but is not quite correct. Envyng someone for wealth is not the same as wishing to possess the wealth itself, which is what the second example really means.

in contracted possessives "I have John's picture".

This suggests either that I have a picture belonging to John, or that I have a picture of him. Ambiguity arises if a form expresses two different senses by virtue of contraction, and disappears when the contraction is expanded. Thus "I have a picture of John's" and "I

have a picture of John" each unambiguously express one of the example's readings.

In part of speech "John performed the girl moves in the folk dance after he taught the girl moves in the opening number. See how quickly the girl moves now"; "They're cooking apples"; "Time flies like an arrow"; "I know that boy is bad", "I know that boys are bad"; "What boys did it?", "What boys do is not my business"; "The trash can be smelly", "The trash can was smelly"; "I told the girl that I liked the story", "I told the girl who I liked the story", "I told the girl the story that I liked".

This is essentially the same problem as lexical ambiguity in syntactic case. However these examples go beyond identifying the problem to demonstrate it. Each 'girl moves' collocation in the first sentence is composed of different parts of speech; ?adj+noun, noun+noun, noun+verb. 'Cooking' acts as a present participle and an adjective (2). The third example is a classic with five readings, only four of which I can identify: flies should be timed a certain way, a type of fly likes arrows, time passes quickly, and ?it's time that flies liked arrows.

10. Ellipsis "She dances better than he does", "I did it; he didn't".

A literal reading of either sentence asks the question: than he does what? Ellipsis within a sentence can frequently be identified by the presence of coordinating conjunctions or punctuation between an elliptical clause and its referent. Elliptical reference between sentences frequently relies on structural symmetry between the two related expressions.

11. Cross-serial Agreement "This land and these woods can be expected to rent itself and sell themselves respectively", "The last two people in this picture live in Columbus and Chicago".

Sentences using 'respectively' explicitly or implicitly coordinate on a cross-serial basis (rather than on a center-embedding one). In cross-serial coordination, each conjunct in a series is linked with its counterpart in the same position in another series. The construction is confusing when the number of conjuncts does not agree. More than two

series can be coordinated; (1) is still easily understood if 'in 1989 and 1990' is appended.

12. Anaphora 'do this', 'such things', 'doing so'; "Jane gave Joan candy because she was hungry", "Jane gave Joan candy because she wasn't hungry"; "A woman and a girl each had a hat and the woman gave hers to the girl".

Although anaphora usually involve references to nouns ('this'), they can also apply to adjectives ('such') or adverbs ('so'). Reference in the next example can only be determined by pragmatic knowledge of who was hungry. In terms of the last example, which hat was transferred in each of the following sentences: 1) "The girl gave hers to the woman"; 2) "The girl gave hers back to the woman"; 3) "The girl gave it back to the woman"?

13. Multiple Word Nouns and Merged Noun Compounds 'pipe wrench', 'putty knife', 'shoe horn', 'property tax office furniture', 'horse collar'.

Phrases which contain more than one noun in their head can be divided into two types. In multiple word noun phrases, the last noun supplies the dominant elements of the phrase's sense and each preceding noun specifies and extends it (1, 2, 4). Merged noun compounds are concatenations of nouns which give rise to a new sense largely unrelated to those of its components (3, 5).

14. Preposed Phrases "For each person there, I had a gift".

This slightly unusual construction emphasizes the opening phrase. There are no problems in processing, particularly since the presence of structural punctuation signals the likelihood of distance dependency (between 'had' and 'for').

15. Syntactically Ill-formed Input "The boys is here". "Hey. You. Yes, you. Do it. That's right. Right. No. Okay".

Ungrammatical utterances will always be with us. They are found particularly in speech, where the parallel visual and aural modes act to integrate the entire communication

and to substitute for broken phraseology and incomplete expressions. Ill-formed statements can be divided into ones which are completely but incorrectly formed and those which are much less complete than the usual elliptical statement.

16. Nominals "I regretted his going", "His going surprised me", "After his going we hired an athlete.", "His going's main effect was to end the feud".

Nominals become more awkward as the construction in which they appear becomes more complex. Ross attributes this awkwardness to the 'nouniness' which a nominalized form acquires as a result of its transformation (Ross, 1967 in Winograd, 1983).

17. Invalid Transformations

interrogative "Merry wanted to see Barbara", "Who did Merry want to see?"; "Merry has a book that pleased Barbara", "Who does Merry have a book that pleased?"

Transforming the first sentence into the second poses no problem, however executing a parallel transformation on the third statement generates an ill-formed sentence. The third sentence must be paraphrased to merge the subordinate clause into the main clause, generating "Barbara was pleased by Merry's book", before it can be transformed into an equivalent interrogative construction "Who was pleased by Merry's book?".

reflexive "I screamed for Harry to open the door", "I tried for me to open the door", "I tried for myself to open the door", "I tried to open the door for myself".

The first sentence is legitimate but a similar construction incorporating a reflexive relationship is ill-formed. The remaining sentences are paraphrases of the second statement that move toward correct syntax while retaining the reflexive element and the same sense. These examples seem to highlight an unmotivated complexity of English syntax.

18. Ergative Verbs "The man is opening the door", "The man is laughing at the joke"; "The man is cooking dinner", "The man is cooking", "The meat is cooking"; "The water tastes bad", "I taste the water".

To establish a context for evaluating these unusual verbs, consider the first two non-ergative sentences which are respectively transitive and intransitive. Truncation produces one satisfactory outcome, "The man is laughing" and a dubious one, ?"The man is opening" (unless a context has previously been established).

Now consider the next three sentences. Although the surface syntax of these examples suggests that 'cook' is bitransitive, taking or leaving an object equally well, in reality the verb seems to have an inherent transitivity. Performing the same truncation on (3) to produce (4) changes its sense; (4) is ambiguous, its different senses being -cooking-something- and -being-cooked-*. This change in meaning did not occur with the first two truncations.

Verbs of this sort are called ergative. They seem to have such a great need for an object that if none is supplied syntactically, the action can be interpreted as rebounding on the agent. A ergative verb can meaningfully fill the two frames "Something is <prpart>" and "Something is being <papart>"§. Ergative sentences like (4) are therefore often ambiguous. Some instances require discourse context, while others can be disambiguated using only the features of the constituents. There is no ambiguity about (5) because -cooking-something- requires a human agent and 'meat' is not a human being.

'Taste', 'smell' and other verbs describing perceptual processes are ergative. The mental sphere in particular the emotional is rife with instances of the sort of situation that ergative verbs express. Consider 'fear' and 'frighten'. Although both statements in the pair "He frightened me", "I feared him" are active, they refer to one act, not two⁵⁹. The factors operating here seem in some way to be akin to the two of the types of speech act, locutionary ("I feared him") and perlocutionary ("he frightened me"), generalized from cases of deliberate action to all acts.

* this linguistic phenomenon prompted Chomsky to develop the notion of deep structure and the theory of transformational grammar (Winograd, 1983).

§ it is unusual for the active and passive voices of a statement to express the same sense.

19. Homomorphy (*similar syntax, different semantics*) "Falling planes can be dangerous", "Flying planes can be dangerous", "Buying planes can be dangerous"; "Warren is eager to please", "Warren is easy to please".

The first set of examples use ergative verbs to illustrate homomorphy and the second set show that it is something other than ergation. (1) and (3) can be paraphrased without ambiguity as "Planes which are falling can be dangerous" and "<To be> buying planes can be dangerous". The second sentence can be read both ways: "Planes which are flying can be dangerous" and "<to be> flying planes can be dangerous". It seems to lack a constituent which would distinguish between the two senses. The paraphrases suggest that the confusion arises from the different senses of ergative verbs; "Planes which are being flown can be dangerous" is a near paraphrase of "<To be> flying planes can be dangerous".

The same distinction appears in the second example without the presence of ambiguity. A paraphrase of (4) makes 'Warren' the direct object in "<it> is easy to please Warren." while in (5) 'Warren' is a complemented subject because 'easy to please' is an idiom meaning 'easily pleased'*.

20. Long Range Reference "The man left the house. The sun was shining. It was warm and windless. A bird sang. He felt good", "He is the man who in the fullness of time and after talking to all of us who take an interest in these matters when the issue becomes important did it".

In the first example the correct reference for 'he' must be found across three intervening sentences. Sentences frequently make references to neighboring or more remote statements and these references can easily be misunderstood. In the second example

* are there any other constructions which employ complete transformations like from adj+inf to adv+part?

long range reference occurs within a sentence across a 24-word interjection between 'who' and 'did it'.

References can also occur between noun and noun; we are expected to recognize entities which have been previously introduced into a narrative. Common noun references are harder to handle than a pronoun one because the latter obviously requires a referent whereas the former may be the introduction of a new actor. The problem is compounded when the reference is not direct but through a synonym.

21. Comparison "Tide is better", "Tide is better than ABC", "Tide is better than any soap on sale in this country"; "Tide is the best soap on sale in this country".

All comparisons, whether comparative (1) or superlative (4), need to have their context specified. In the comparative case, this means identifying the particular individual (2) or set of individuals to which the subject is being compared (3). When the comparative evaluation is made against a set, the resulting statement is easily paraphrased into a superlative comparison. Thus (3) is a paraphrase of (4).

Semantic Cases

Lexical

22. Literal vs Idiomatic Use "I turn on the road", "I turn on the tap"; "The men agreed to play ball", "The men agreed to play ball Tuesday nights", "Acknowledging their other differences, the men agreed to play ball on this issue".

Each word in the first sentence executes a traditional syntactic function; 'on' heads a prepositional phrase that identifies where the rotation takes place. In the second sentence, 'turn on' is a idiom in which the particle 'on' indicates the direction of rotation. (4) and (5) show how the same sentence can be modified to select between its literal and idiomatic senses 58.

There is a dilemma here. A parser which treats (2) as a verb+prep will misunderstand it, while always taking the longest possible string from input (ie idiomatic use) means that (1) will never be correctly recognized.

23. Intensive vs Extensive Usage "The president is at Camp David", "The president is elected every four years".

Sometimes the name of a role is intended to identify the individual who currently fill it, and sometimes the role itself. This parallels the distinction between a term's intension and its extension.

24. .Universal vs .Specific Quantification "I am looking for a man", "The golden tamarin¹ is smaller than a cat but twice as cute", "I am holding a man", "The man who sees me left the room", "The man who saw me left the room".

The first example illustrates the problem; it is uncertain whether the sought man is a particular individual or simply any man at all. The second and third examples demonstrate that both definite and indefinite quantifiers can experience this kind of ambiguity. 'A cat' is .universal and 'the tamarind' may be; 'a man' however must be interpreted specifically because no one cannot hold all men. The fourth and fifth sentences show that quantificational ambiguity can arise in tensed verbs. Given that I am within a room, a man's seeing me must precede his leaving the room. Therefore 'sees' can only be understood as the generic sees-on-a-regular-basis rather than the specific single-act-of-sight. Because both verbs in the fifth sentence are past tense, no temporal order is established and 'saw' can refer to either quantification.

No general method of resolving quantificational ambiguity has been discovered*. However if a quantified phrase is highly specified by definition or through context, it is reasonable to assume that reference is to a particular entity. Thus in the absence of contrary

* 'the situations ... are not well understood at present' (Alshawi, 1987, p. 39).

context (such as reading the classified ads) "I am looking for a red 1977 Mustang" suggests that the speaker seeks a particular car known to himself. Similarly "The man who saw me for three weeks at 9 o'clock left the room" suggests that the seeing is an act unrelated to the leaving.

25. Ambiguity

inherent (*polysemy and homonymy*) 'right'.

Words can be both polysemous and homonymous. 'Right' has homonymous senses of -righteous- and -right-handed-. Each of these senses is polysemous*. When employed in "the right thing to do" and "do right", -righteous- has homologs of correct-under-the-circumstances and -absolutely-correct-. The homologs of -right-handed- are demonstrated in 'on the right' and 'my right hand'. They are somewhere-to-the-right and on-the-right-side-of-my-body.

in quantifier sense "List the trains that service every city", "Not everyone came", "Everyone doesn't do that".

These sentences introduce ambiguity because 'every' means both the universal quantifier 'all' and the existential quantifier 'some'. As a result the first example leaves it unclear whether the trains have to visit all cities or just one of them to be listed. The next examples show that negation can interact with quantifier ambiguity. 'Not everyone' is existentially 'not some one' and therefore someone, and universally 'not all ones' and therefore no one.

Syntactic

26. Grammatical Subordination (*to aid communication*) "Everyone loves his wife",

* 'right' also occupies different syntactic categories: "Might makes right" (noun), "I right the bottle" (verb), 'the right stuff' (adjective) and "Do right" (adverb).

"Although it was predicted Goldwater didn't win", "The man who shows he wants it will get the prize he deserves".

In each of these examples structure has been subordinated to communication; each is readily understood although it infringes on the rules of grammar. The first example would be more correctly stated as "Everyone loves their wives", which would introduce a syntactic ambiguity about the number of wives each person has.

The reference for 'it' in (2) is an elided 'his win' phrase or 'that he would win' clause. The presence of either makes an awkward construction: "Although it was predicted that he would win, Goldwater didn't (win)"; "Although his win was predicted, Goldwater didn't (win)". The third case is similar to the second. It can be paraphrased as "If a man deserves a prize and shows that he wants it, he will receive the prize" but this fails to express the need to show desire emphasized in (3).

27. Passivization Effect "Few people read many books", "Many books are read by few people"; "Some books are read by few people", "Few people read some books"; "No books are read by all people", "All people read no books".

In the first example (1) emphasizes the scarcity of voracious readers while its passive transformation (2) emphasizes the plenitude of poorly-read books. A sentence's first phrase seems to enjoy a position of importance. (1) and (2) make different statements because the focus of emphasis shifts from 'books' to 'readers' when voice is transformed from passive to active. Although the problem disappears when quantification is shifted to the neutral 'some' in the second example (3, 4), 'some' in this case appears to be a qualifier of identity rather than a quantifier of measure because the problem persists when the quantifiers are universal (5, 6).

28. Referential Opacity/Obliquity "Ross believes that Nadia's dog is lost", "Ross believes that Nadia's spaniel is lost".

(2) cannot be inferred from (1) because the first sentence does not explicitly say that

Ross knows the breed of Nadia's dog. An individual's personal knowledge is necessarily circumscribed and it cannot be assumed that objective facts are subjectively known by everyone. Reference to subjective knowledge occurs in words like 'believe', 'say', 'think' that report indirect discourse and mental activity.

29. Conventional Ambiguity "The chickens are ready to eat", "the Boston train is on track nine", 'human offspring', 'human race', 'human dog', human consumption', 'human experimentation'; 'dog meat'.

Both examples contain ambiguous elements. 'Ready to eat' can be interpreted as an idiom meaning -fully-prepared- or as a construction meaning -hungry-. In the second example, modifying 'train' with 'Boston' leaves it unclear whether the train is going to or has come from that city. These ambiguities arise from conventions which we routinely assume when we read either of these statements. However the first example should not really be ambiguous at all. Its -fully-prepared- sense can only be understood by misreading it; to make that statement in correct grammar would be "The chickens are ready to be eaten". Advertising has made 'ready to eat' mean ready-to-be-eaten.

The -to- or -from- confusion in the second example is based on a similar assumption, but one that has not been formally taught through the media. 'Boston' in this case could also mean -made-in-Boston-, -named-after-Boston- or even -owned-by-the-Boston-subway-corporation-. The directional senses are selected as the ones to be confused about because movement and travel is the first thing most people associate with trains. Disambiguation comes (if it does) through knowledge of the timetable.

The five 'human' examples all use the adjective to communicate different relationships between its object and human beings: offspring-of-humans, race-composed-of-humans, consumption-by-humans, experimentation-on-humans, dog-with-human-traits,. Similarly, although both words in the last example are individually

unambiguous, when collocated it is unclear whether they mean meat-to-feed-to-the-dog or flesh-of-the-dog.

30. Complementary or Copulative Verbs "John is quiet".

The twenty copula* are a special case of relationship. They are used in a statement to allow an entity to be qualified in potentially elaborate manner and not to express a formal relationship. Notwithstanding this, these verbs do qualify their complement's existence.

31. Uninstantiated Reference "They're always watching", "It rained yesterday".

The referents for 'they' and 'it' are supplied by convention and not context; none will be found in the narrative.

32. Confounded Senses (*meaning shift through extension*) "The oscilloscope Tom used to fix", "The oscilloscope Tom used to fix the radio".

This example has elements of garden pathing and idiom, but the construction it is based on may be unique. In the first sentence, 'oscilloscope' is the direct object; in the second, it is an instrument and 'radio' is the direct object. Winograd (1983) believes different senses of 'used to' are employed. In the first sentence the phrase, which he represents as 'usetā', expresses past tense idiomatically; in the second, instrumentality.

33. Reference through Synonyms "My sister's kids are at the store again. Those three tykes sure like candy", "The man left the house. The sun was shining. It was warm and windless. A bird sang. He felt good".

We understand that 'those three tykes' refers to 'kids' because the two noun phrases are highly synonymous (though not exactly so; little girls can be tykes too). Less synonymous pairs and reference through negated antonyms are harder to understand.

* subjective complements: 'be', 'become', 'grow', 'get', 'feel', 'seem', 'appear', 'taste', 'sound', 'smell', 'look', 'remain', 'act', 'go', 'turn', 'make'; objective complements: 'make', 'prove', 'imagine', 'consider'.

Synonyms and long range together are a recipe for expensive processing in cases of reference. If 'the man' or 'Mr. Smith' is substituted for 'he' in the last sentence of the second example, the previous four sentences or 15 words would need to be tested to catch the intended reference.

Discourse

34. Diexis in Dialogue " 'This is the man. He put it there' ", "John pointed to his damaged car and said, 'look at that! Do you think I can see through that? It'll run, but that tire'll be torn to shreds in a mile if I do.' "

The first reported statement cannot be understood without context. In the second we infer that the car's windshield has been made opaque by cracks and stars, and that although the tires are not flat, part of the fender or frame is pressing against one of them.

Spoken words can identify things in the scene which are not mentioned in the utterance. The same effect can be accomplished in a written narrative only if the identified item is explicitly or implicitly introduced in the vicinity of the dialogue.

35. Dialogue Attribution "I did it." "No, you didn't." "Yes, he did!"

The content of these three sentences imply three speakers rather than two individuals taking turns which would otherwise have been assumed.

Written English does not identify speakers of directly-quoted dialogue once they have been introduced. Quotation marks are used to delimit utterances and speakers are assumed to take turns. This convention becomes problematic when a conversation involves more than two people; we then must rely on our models of the participants' interests, motivations and perspectives to determine attributions. There is a distinct risk of misinterpretation in a dialogue among three or more people.

36. Speech Acts "Could you pass the salt shaker?", "She turned to her two suitors who stood there with their gifts in their hands and, 'I'll take the flowers'."

Speech acts can be divided into those in which the locutionary sense is incongruous with the context and those in which it is congruous. The first case, illustrated by (1) covers what is commonly known by Austin's (1962) term. The locutor's interest in the auditor's physical capacity is irrelevant to his current activity of eating.

The second case is more contextual. In (2) all parties know that the woman is expressing a preference for more than a gift. Although the second utterance is superficially similar to a metonymy*, it differs in that its sense is meant literally; no doubt the woman really expects to and does receive the flowers. Her utterance is therefore not solely figurative.

37. *de re* vs *de dicto* Scope "He confidently walked down the garden path", "He apparently walked down the garden path."

'Confidently' applies to the act of walking within the statement and elaborates its sense; 'apparently' applies to the entire statement and qualifies the proposition it communicates. The former is a *de re* usage and the latter *de dicto*. A frame to distinguish between these two cases translates candidate words to their cognate noun form: "It is apparent that ...". Any statement that can be sensibly paraphrased in this way is *de dicto*.

37. Modal and Hypothetical Existence "It might snow", "he ought to come", "Were he here, he would disagree".

Modal verbs are a closed set which qualify the possibility, capability and necessity of a specific act or class of acts occurring (1). In addition, modals qualify the volition, intention, prediction, permission and obligation of an animate agent to perform an act (2).

* in that the speech act consists of the flowers being understood to stand for the man who is offering them.

The subjunctive mood is used to express statements which have an imaginary or hypothetical existence (3).

38. Necessary vs Accidental Consequence "Tom entered the room. The apple fell off the table", "Tom slammed his fist on the table. The apple fell off of it", "Tom looked at the river. The apple fell off the table", "Tom is at work. The apple fell off the table".

In the absence of explicit connectors like 'because', 'however' and 'since' (colloquial) we must employ commonsense knowledge to infer that causal consequence accompanies temporal sequence. The first example is open to either a necessary or an accidental interpretation. In my view it leans slightly towards causality because it goes to the trouble of mentioning the events. The second example is clearly to be understood causally. The third and fourth examples involves events, and a state and an event respectively. Readers are not likely to think that the apple's fall in either case proceeded from Tom's action or state.

39. Indeterminate Set Membership "The family went to a play. It was so poorly acted that some of the audience booed. Others left and some people began to talk among themselves".

After reading the above sentences, it is impossible to know whether any of the family members booed, left, began to talk among themselves or continued to watch the play. Language utterances routinely define groups and perform operations* on them which result in their memberships being unknown. Language users trust that they will be told the composition of a group when it is important.

* conjunction ('the gang and I'), disjunction ("It's them or us"), member and subset identification and selection ("the whole family was among the audience ...", 'one of the staff').

A.4 Issues Surrounding the Selection and Description of a Computational Lexicon

The original design of this project called for a general purpose dictionary capable of supporting sentences of unrestricted content and large enough to allow those sentences to exhibit limited but definite expressive power. This goal arose from my objective of exploring semantics in a nonspecific and unpredetermined way. Many previous research efforts in NLU have limited themselves to the semantics of some restricted or specialized area of natural language, ie discourse involved in database operations, military logistics, interaction with the UNIX operating system etc.

The first task to tackle was the choice of words to be included in the lexicon. A criterion for word selection which initially seemed reasonable was most frequen. use. Three compilations made on this basis are readily available: 1) the Leuven English Teaching Vocabulary-List (LET) (Catholic University of Leuven, 1981), 2) the American Heritage Word Frequency Book (WFB) (Carroll, Davies & Richman, 1971), 3) Word Frequencies of Spoken American English (WFS) (Dahl, 1979). These lists differ on the size of the corpus from which they were selected, its nature (spoken or written medium; educational, psychoanalytic or general factual and fictional content) and whether or not inflected forms, colloquialisms and particles are counted as distinct words. As a result of these differences no two of the word lists duplicate more than 66% of each other's most common selections and certain pairs match in less than half*. Similarity diminishes further when order in the list is considered. Pairwise correlations of frequency rank range between 0.41 for WFB and LET, the lists based on written corpora, and -0.37 for WFB and WFS.

Although the problem of poor agreement between lists could be solved by including in the dictionary all words which appear in any list, a more serious problem immediately

* when the first 200 words from the WFB and the WFS are compared with the first 149 from the LET, the total number of unique words is 267 and the lowest rate of matches was 47%.

becomes evident—none of the lists were very serviceable; they do not permit descriptive statements to be composed. In other words, a vocabulary based on word frequency provides poor semantic and syntactic coverage. It seems that people most frequently use a small set of articles, adjectives, and inflected and contracted modal and auxiliary verbs in their linguistic expressions. These forms occupy the top levels of any list based on frequency and displace the more varied nouns and main verb forms which give natural language its richness and descriptive power.

More work at the library revealed an alternate source of word lists. Language educators have been assembling basic vocabulary lists for instructional purposes for most of this century (Horn, 1926; Thorndyke & Lorge, 1944). A more current and substantially smaller* set of such core words appears in the two Pacemaker Core Vocabularies (PCV1, PCV2) (1980). These two disjoint lists of 1,021 and 730 words have been identified as core words for children reading at the third- and fourth-grade levels respectively. Although PCV1 lists 172 of the first 267 words in a consolidated frequency list, the words it contains have been chosen to facilitate communication and the remaining 800-odd appear to be the nouns and verbs missing from the first series of word lists.

Once the contents of the dictionary were selected, the next task was to develop a semantics comprehensive enough to permit construction of meaningful descriptions for each commonly-recognized sense of the 1,021 words. Although this was only partially completed, semantic features and relations were identified sufficiently well to outline a comprehensive semantics for natural language—the semantic feature system.

The last task was to use this system to describe senses for the words in the dictionary. This goal was largely unmet. With few exceptions, only a few words from each of the various syntactic classes were described, entries for these words are schematic rather than complete, and word selection was not directed by an analytic principle but

* Horn catalogued 10,000 words, Thorndyke 30,000.

rather by the practical need to compose grammatically correct examples to be submitted to the demonstration program.

The reason that performance fell off in the second stage and was so far short of initial expectation in the third is simply that this situation marks the point at which it became clear that the scale of the task was beyond what could be accomplished in the time and resource frame of a master's thesis. Notwithstanding this, the first stages of semantic analysis were carried out for most of the thousand-odd words in PCV1. This involved categorizing and recategorizing words organized by part of speech into smaller and smaller sets*. An example is provided below.

A relatively complete set of word sense descriptions were developed for English prepositions. Prepositions are a good candidate for a scoped development since they constitute a small and fairly closed syntactic class[§]. They are also interesting from a semantic point of view; they are ubiquitous, express a wide variety of meanings, and while not central in the sentence, play important roles in it. Prepositions express relationships and their senses call into play the senses of nouns and verbs; but they do this indirectly and do not need the exhaustive analysis that a direct approach to the two larger grammatic categories would require. 42 prepositions appear in PCV1 and 11 more in PCV2; 46 of these were included in the 86 defined in the course of this project.

The First Pacemaker Core Vocabulary (PCV1) ©

This vocabulary list is reproduced with permission in order to make it available to other interested parties. Its compilers take a relatively relaxed view of word identity; contractions, semi-acronyms like 'TV' and irregular verb tenses are all considered to be separate words. Notwithstanding the PCV1's superiority to pure frequency lists for

* this procedure is discussed in detail in the body of the thesis.

§ but not a completely closed one. Constructions such as the compound 'up on' and the idiom 'with the exception of' demonstrate how the set gets expanded.

semantic coverage, I find a bias in it towards a schoolchild's perspective in words like 'mom', 'dad', 'ring', 'rang', 'bell'; and 'king', 'queen' etc. To condense it, the list is presented in a run-on format here; no entries begin with X or Z.

a able about above across act add afraid after afternoon again against age ago air airplane all almost along already also always am an and angry animal another answer any anyone anything are arm army around art as ask at ate aunt awake away

baby back bad bag ball band bang bank bark basket be bear beat beautiful because become bed been before beg began begin behind believe bell belong below bend beside best better between big bird birthday bit black blank block blood blow blue board boat body book born both bottle bottom bow box boy brave bread break breakfast bright bring broke broken brother brought brown build built bump burn bus business busy but buy by

cake call came camp can candy cannot can't car card care careful carry cat catch caught cause cent center chair change check chest chicken child children church circle city class clean clear climb clock close clothes coat cold color come coming company complete control cook cool corner cost could couldn't country course cover cow crack crash cross crowd cry cup cut

dad dance dark day dead dear decide deep desk did didn't die different dig dinner dirt dish do doctor does doesn't dog doll dollar done don't door dot down drank draw dream dress drink drive drop drove drum dry duck dug during dust

each ear early earn earth east easy eat egg eight either empty end engine enjoy enough escape even ever every everyone everything explain eye

face fact fair fall family famous far farm fast fat father favorite feed feel feet fell felt fence few field fight fill finally find fine finger finish fire first fish fit five fix flag flash flat flew float floor flower fly fold follow food foot for force forest forget forgot forward found four free fresh friend frighten from front fruit full fun funny

game garden gas gate gave get giant girl give given glad glass go gold gone good good-bye got grab grand grandfather grandmother grass gray great green grew ground group grow guess gun

had hair half hall hand handle hang happen happy hard has hat hate have haven't he head hear heard heart heat heavy hello help her here herself hid high highest hill him himself his history hit hold hole home hope horn horse hot hour house how however human hundred hung hungry hunt husband

I ice I'd idea if I'll I'm important in inch inside instead interest interesting into iron is island isn't it its it's
itself I've

jail jam jar jet job join joke joy jump just

keep kept key kick kid kill kind king kitchen knew knock know

lake lamp land large last late later laugh law lay lazy lead learn leave led left leg less let let's letter lie life
lift light like line lion list listen little live lock log long longer look lost lot loud love low luck lunch

machine mad made mail make making man many map mark married matter may maybe me mean meat
meet member men mess met might mile milk mind mine minute miss mix mom money monkey moon
more morning most mother motor mountain mouse mouth move movie Mr. Mrs. much mud music must
myself

nail name near neat necessary neck need neighbor net never new next nice night nine no noise noon nor
north nose not note nothing now number nurse

ocean o'clock of off office often oh oil OK old on once once one only open or orange order other our out
outside over own

pack page paid paint pair pan paper parents park part party pass past paw pay peace pen penny people
perhaps person pet pick picnic picture pie piece pig pile pin pink pipe place plan plane plant play please
pocket point pole police policeman pond pool poor pop possible post pot pound power present press pretty
print probably problem pull push put

queen question quick

race radio rag rain rang rat rather reach read ready real really reason record red remember rent rest return reward
rich ride right ring river road rob rock rocket rode roll roof room rope round row rub rule run rush

sad safe said sail sale same sand sang sat save saw say school science scream sea season seat second see
seem seen sell send sent sentence set seven several shake shall shape sharp she sheep sheet shell shine ship
shirt shoe shook shoot shop short should shout show sick side sight sign silly simple since sing sink sister
sit six size skin sky sleep slide slip slow slowly small smart smash small smile smoke snow so soap soft
sold some someone something sometime son song soon sorry sound soup south space spark speak special
speed spend spent spill spin splash spoil spoke sport spot spring square stairs stamp stand star start state
station stay steal steel step stick stiff still stone stood stop store storm story stove strange stream street

strike string strong student study stuff such sudden suddenly sugar summer sun support sure surprise sweet
swim swing

table tail take taken talk tall tap taste teach teacher team tear teeth telephone television tell ten test than
thank that that's the their them themselves then there these they thick thin thing think third this those three
threw throw ticket tie time tiny tire to today toe together told tomorrow tonight too took tool tooth top
torn touch toward town toy track trade trail train trap tree trick trip trouble truck true try turn turtle TV
twice two

ugly uncle under understand until up upon us use

vacation vegetable very visit voice

wagon wait wake walk wall want war warm was wasn't waste watch water wave way war weak wear weather
week weight welcome well went west wet what wheel when where which while whisper white who whole
why wide wife wild will win wind window wing winter wipe wire wish with without woke woman won
wonder won't wood word work worker world worn worry worse worst worth would wouldn't write written
wrong wrote

yard year yellow yes yesterday yet you you'll young your you're

An Example of Semantic Categorical Analysis of Verbs in PCV1

607 words, almost two-thirds of the in the first PCV list—607 words—have verb senses. 73 of these words are contracted and tensed forms which can immediately be eliminated from consideration because their senses will be expressed through their root forms. A cursory inspection of the remaining 533 highlights a problem common to all vocabulary lists which do not fully analyze their subject matter: a child may or should recognize one sense of a word for which an adult immediately understands additional senses. This is a particularly frequent occurrence when the word list contains the simple and often homonymous words taught in the first grades of school. However so long as the sense or senses that qualify such words to be in the list are not identified, no problem exists*—within the part of speech. If a word has additional senses when taken as another part of speech, compilers of a vocabulary list must decide whether or not to include it under the other part of speech. The authors of the PCV lists appear to have included such senses. Homonymy eliminates another 31 words from the verb list. I believe that the verb senses of words like 'book', 'ice', 'neck' and 'season' are not commonly understood by children reading at the third-grade level. Their noun rather than verb senses justify these words' inclusion in the list. 503 words remain of interest as verbs. It would be tedious to list them here, particularly since they have been reproduced above (although in an unorganized form). A first set of verb categories was developed by grouping these words on what I judge to be their most striking aspects. The process of grouping on salience produces the minimum amount of semantic output because word senses are analyzed only to the point where they can be distinguished from one another.

The resulting categories are unsystematic, arising directly from raw material as they do. They are not guaranteed to cover 'acts' area of semantic space fully, nor do they agree

* and yet that is exactly the problem; if the senses of interest were specified, there would be no ambiguity.

with the current feature system in many respects. Parenthetical features show how senses in each of nine primary categories can be extended:

1)	possession	(direction)
2)	action	(purpose)(class)(causation)(orientation)
3)	process	(domain)
4)	composition	(union, division, combination)
5)	construction	(orientation)(number)
6)	change	(domain)(aspect)(purpose)
7)	emotion	(orientation)(perspective)
8)	condition	(time)(extent)
9)	motion	(medium)(instrument)(case)(proximity)(direction)

FIGURE 14: An Initial Classification of Verbs

The last category, verbs of motion, offers a particularly good opportunity for further categorization. 116 verbs emphasize a thing's movement in space. A number of criteria can be used to divide them into smaller subsets.

Two points must be stressed. This categorization is not particularly coherent or comprehensive. This is the stage during which it became clear that defining senses for a thousand words was too large a task. The effort was discontinued after most verbs were provisionally classified but before feature selections were subjected to the repeated scrutiny of editing. There is no question that many important aspects of the verb classes below are not expressed by the half dozen or so features that describe them. It is also quite likely that the choice of feature descriptors may not be the best possible.

This classification is also not consistent with the feature system which was actually implemented. The analysis presented here was superseded by the simpler but more coherent feature system presented in the body of the text and used in the demonstration program. This categorization must therefore be understood in its own terms. To do this one must set aside previously presented details of features and consider each word in the

following classes only in terms of the features describing the class. A quick inspection will show just how complex a set of features are needed just to distinguish among the sense of word classes which most third-grade students know.

In the categories appearing below, defining features begin each entry and start with capital letters. They are followed by verb tokens which are not set in the usual single quotes so that they may be read more easily. A word followed by a parenthesized number is a homonym; the number labels the particular sense intended. A set of features enclosed in parentheses should be considered as a unit which extends the sense expressed by the first feature in the set. Words in angle brackets identify aspects of a verb sense which other words in a usage context must specify.

- a) Medium Gas, Proximity Interior, Agent Vehicle: jet, fly(1)
- b) Medium Gas, Proximity Interior, Agent Animate: fly(2), wing
- c) Medium Liquid, Proximity Contact, Agent Vehicle: sail
- d) Medium Liquid, Proximity Contact, Object Vehicle: row
- e) Medium Liquid, Proximity Contact, Agent Animate: swim
- f) Medium Solid, Proximity Contact, Agent Vehicle: bus, drive(1)
- g) Medium Solid, Proximity Contact, Object Vehicle: drive(2)
- h) Medium Solid, Proximity Contact, Agent Animate, Perspective Extensive, Plane Vertical: jump, spring
- i) Medium Solid, Proximity Contact, Agent Animate, Perspective Extensive, Plane Horizontal: walk, step, run, dance, crawl
- j) Agent Animate, Perspective Intensive, Plane Vertical: duck, bow
- k) Agent Animate, Direction Linear, Instrument Hand: pat, slap
- l) Agent Animate, Direction Linear, Instrument Finger: tickle
- m) Agent Animate, Instrument Hand: handle, paw
- n) Agent Animate, Direction Linear, Instrument Hand, Object Hand: clap
- o) Agent Animate, Direction Circular, Instrument Hand: stir
- p) Agent Animate, Force Compressive, Instrument Hand: choke
- q) Agent Animate, Force Compressive, Instrument Arm: hug
- r) Direction Linear, Number Multiple, Energy Great: drum, beat, pound
- s) Direction Linear, Energy Great: strike, knock, hit, bang, prod
- t) Direction Linear, Energy Little: tap
- u) Direction Linear, Instrument Agent*: bump
- v) Direction Circular, Perspective Extensive: roll
- w) Direction Circular, Perspective Intensive: wheel, turn, spin, face
- x) Direction Oscillatory, Number Multiple: shake, rock
- y) Direction Oscillatory, Number Multiple, Object Agent: shudder, shiver, nod
- z) Agent Animate, Direction Along, Instrument Hand: stroke, pet, rub, smooth, wipe
- aa) Direction Oscillatory: bounce
- ab) Direction Oscillatory, Perspective Extensive: wave, swing
- ac) Agent Animate, Direction Linear, Instrument Foot, Energy Little: toe
- ad) Agent Animate, Direction Linear, Instrument Foot, Energy Great: kick, stamp
- ae) (Object Agent Substance Fluid): flow, stream
- af) (Object Substance Fluid): pour, spill, splash
- ag) Direction To, Proximity Contact, Medium Liquid: land
- ah) Direction To, Proximity Contact, Number Single: come, arrive, reach

- ai) Direction To, Proximity Contact, Number Plural, (Object Mobile): return(1)
- aj) Direction To, Proximity Vicinity, Number Single: approach, near
- ak) Direction From, Proximity Contact, Number Single: go, depart, leave
- al) Direction Linear, Plane Vertical Greater: rise, climb
- am) Direction Linear, Plane Horizontal: slide, slip
- an) Direction Linear, Plane Vertical Lesser: fall, sink, drop, ground
- ao) Direction Serial Greater[§]: lead
- ap) Direction Serial Lesser: follow, trail, track
- aq) Instrument Agent: carry, bear
- ar) Force Compressive, Instrument Agent: push, press
- as) Force Expansive, Instrument Agent: drag, haul, pull
- at) Direction To, Instrument Agent: bring
- au) Direction From, Instrument Agent: remove, from
- av) Direction To, Instrument Hand: hand
- aw) Direction From, Perspective Extensive, Medium Gas: toss, throw
- ax) Direction To, Perspective Extensive, Medium Gas: catch, field, grasp, seize, grab
- ay) Plane Vertical Greater, Instrument Agent: lift
- az) Direction From, (Object Mobile): leave(2), abandon(2), discard
- ba) Direction From, (Agent Mobile): leave(1), abandon(1)
- bb) Direction To, Proximity Contact, Number Plural, (Object Stationery): return(2)
- bc) Direction To, (Object Mobile), Position <unspecified>: put, place, set, station, seat, lay, stand(1)
- bd) Direction To, (Agent Mobile), Position <unspecified>: rest, lie, sit, stand(2)
- be) Medium Liquid, (Agent Mobile), Position <unspecified>: float
- bf) Medium Gas, (Agent Mobile), Position <unspecified>: hover

FIGURE 15: A Classification of Verbs of Motion

* 'agent' used as a value indicates a reflexive relationship; a thing uses its own body to bump another thing.

§ each feature extends the sense of the one preceding it. 'Serial' specifies the direction while 'greater' specifies the particular position in the series.

A.5 Classification of Semantic Theories Used in Computational Linguistic Systems

In a recent survey Graeme Hirst (1987) identifies model-theoretic semantics and compositional semantics as the two major schools of semantic theory used in computational natural language systems. Moyne (1985) agrees with this classification and adds speech act theory which Hirst deliberately ignores (I will do likewise since the feature system does not treat pragmatics on an organized basis). Hirst also groups computational processing practices for semantics into three categories: procedural, knowledge and object-oriented. Other classification schemes ⁶⁰ can be reconciled with this organization.

Compositional Semantics or Semantic Primitive Systems

Compositional semantic theories 'attempt to represent the meaning of each word by breaking it down into semantic primitives' (Hirst, 1987). In his view these systems fail to account for the fact that we can recognize objects which lack criterial features, many words are difficult to decompose, and the issue of representation is ignored. He also believes that frame-based systems are essentially structured compositional systems which must also eventually face the problem of selecting primitives.

Moyne (1985) groups conceptual dependency theory (Schank, 1972) with compositional semantic theories. In that connection, CD theory has been criticized for failing to address the complexity of real-world syntax (ie anaphora) and for inflexible processing (Klappholtz & Lockman, 1977 in Moyne, 1985).

Preference semantics is another major compositional theory. Wilks (1973) identifies 80 primitives in five classes (entity, action, case, quality, type). Processing emphasizes mutual word sense disambiguation (preference); if an initial CD-like attempt to identify a word sense is unsuccessful, various readings of the current template (equivalent

to the CD ACT) are evaluated in terms of previously parsed ones. Wilks' templates do not contain frames' predictive values, though he believes these could be usefully added (Wilks, 1981 in Moyne, 1983).

A connectionist model of lexical semantics defines words as sets of nodes, each of which is a semantic microfeature describing a dimension of meaning (Zeidenberg, 1987).

Truth-Conditional and Model-Theoretic Semantics

Drawing on Tarski's ideas and expressed in the systematizations of Montague and Davidson, these compositional theories evaluate the necessary and sufficient conditions for a statement to be true in a (mathematically) formal model of the real or other possible worlds. In model-theoretic theory processing uses a context-free grammar to translate a sentence into a metalinguistic form, frequently intensional logic, and then interprets the form in terms of the model.

Although control can be diffuse and processing flexible, problems with model-theoretic semantics are its computational intractability and the irrelevance of truth conditions to NL understanding. Another weakness is that it is not an interpretive theory (Foster, 1976 in Putnam, 1985); it gives us no understanding of the utterances considered. Production systems based on the Newell and Simon prototype (Newell, 1973 in Moyne, 1985) parallel this approach.

Procedural Semantics

According to procedural semantic theory, a sentence is compiled into a procedure in the mind or in the machine. This procedure is considered to be the sentence's meaning (intension) and the results of its execution to be the sentence's extension (Hirst, 1987). Meaning resides solely in the text and any knowledge (data) base is not used in interpretation. In systems guided by this view, meaning is encoded in a set of rules which

if matched call what are essentially database procedures to return relevant information. An example is Woods' airline reservation system. Weaknesses are the embeddedness of the semantics in rules and procedures, and the specificity and rigidity (non-compositionality) of the rules. Although procedural semantics is a dominant approach today it remains controversial (Hirst, 1987). Winograd's special format, text-based or limited logic categories (1972 in Moyné, 1985) are all procedural semantic theories.

Knowledge Semantics

Knowledge semantics could be considered a refinement of procedural semantics. Proposing that a sentence's meaning is contingent on both its context and its own content, implementations interpret the results of syntactic analysis in terms of the existing knowledge base as well as on the basis of its semantic composition. This may cause a statement in some form of representation language such as higher-order predicate logic to be added to the knowledge base resulting from some form of transformational grammar.

Since representation is integrated with processing, systems of this sort are no longer structurally tied to an area of discourse; this is a strength. One weakness is a failure to interpret the elements composing the propositions during processing (Hirst, 1987). Knowledge semantics is well suited to the definition of general knowledge representation languages which users can employ to build applications for particular domains. Winograd calls these are general deduction systems and his SHRDLU falls amongst them.

Object-Oriented Semantics

A number of systems use the semantic content of individual words or parse elements in their operations; Hirst (1987) calls them 'object-oriented'. These systems do not appear to be based on any particular theory or even any explicit theory at all, and have only their use of knowledge structures in common. The name object is applied to these

systems because they continue the general trend found in procedural and knowledge-based semantics towards large and more complex data structures, some of which are in fact objects in the message-passing paradigm. Most recent research systems fall in this group.

3 OF/DE 3

A.6

A Theoretic Context for the Feature System

These models of the world, the human mind and language express a conceptual framework consistent with the feature system. The first section presents a simple ontological classification of the world and the next considers how people deal with it. The third section discusses psychological aspects of the nature of linguistic media and draws a conclusion which has implications for decompositional semantics. The final section discusses the general character of language.

Presentations are couched in very general terms to ensure that they represent common or uncontroversial opinion in their respective disciplines and also because the feature system does not need highly detailed models. These models are optional; computational linguistic systems are judged on their performance and neither require nor need to propose theoretical expressions in philosophy, psychology or linguistics.

World

A model produced by generalizing ontologies presented in the literature ⁶¹ views the fundamental being is the thing or entity. Things have a coherent and unitary structure by which distinguishes them from one another. They persist over time and can be recognized as the same thing at different times. Natural kinds are an important class of thing which have a physical existence and are known through direct perception. Other things are mental constructs that have no objective reality outside of the human mind: qualities (greenness, thickness), states and actions (running, hunger), and abstractions (idea, cause). These immaterial things are derived from direct perception by one or more stages of mental processing.

Qualities (color, shape) are the particular sensory data by which we know natural kind things. They are directly given and differentiable by perceptual mode. The notions

that convey immaterial things (speed, number) are derived from qualities*. Although derived notions may express things of any kind, qualities can only be applied to natural kinds meaningfully: *colored cause, *thick running, three ideas.

A thing does not exist in isolation. Things which occur together regularly have a relationship or significant connection. Relationships are epiphenomena, contingent on and sharing the nature of the things that hold them. Thus natural kind things give rise to natural kind relationships. Inverting this dependency creates a view in which relationships exist separately from things and impose requirements on the sorts which hold them.

Mind

"The tendency to divide the world into categories is a pervasive aspect of human thought. Categorization is a fundamental cognitive process because every experience is in some sense unique. ... However, if each experience were given a unique mental representation, we would be quickly overwhelmed by the sheer complexity, and we could not apply what we had already learned to deal with new situations. By encoding information into an organized system of categories, we are able to recognize significant commonalities in different experiences." (Glass & Holyoak, 1986, p. 149)

As Glass and Holyoak note, people make sense out of the sensory experiences constantly thrust upon them by organizing them categorically. This presupposes that the sensory flow has been structured into a set of concepts. Conceptualization involves routinely compartmentalizing sensory data into individual instances (discrimination) which are uniquely identified (recognition), and repeatedly associating new instances with previously perceived ones of the same type (recall).

Sensory phenomena convey themselves to us in a direct and peremptory way.

* in Meinong's terms, qualia subsist in things (Lacey, 1986).

Although we can focus to some degree on particular inputs within a field, our visual and aural senses are both noisy channels from which other sensations cannot be entirely blocked out. We can close our eyes and cover our ears but there is little we can do to veil our senses of touch and taste. So long as we are alive we experience our surroundings whether we wish to or not.

Concepts are our permanent mental representations⁶² of temporary sensory experiences. A concept based on a sensory experience is different from the bundle of distinguished sensation that is the experience itself. When we perceive a man walking into a store, the walking is fleeting and embodied in the particular man and store. Notwithstanding this, the relationship understood by seeing a man walking into a store does qualify as an concept: it can be distinguished from other relationships and is associated with a particular set of perceptions (visual, aural or sensorimotor). Nominalizations of 'walk' like 'walking' or '(a) walk' are separate concepts again; they are things rather than relationships.

The first things we perceive after birth are the physical objects called natural kind things, and perceptions of such things remain the most obvious ones we receive for the rest of our lives. As we encounter different instances of natural kind things, we place them in categories⁶³. Similar things go into the same category and different ones into different categories. When we create a new category, we simultaneously abstract concepts of the particular criteria which distinguish it from other categories*. These concepts embody particular aspects of our perceptions of the categorized natural kind things. Thus when we see different leaves which are thin and green, thin and red, and green and irregularly-shaped, we gain concepts not only of what leaves are, but also of -green-,

* an alternative approach would be for concepts to give rise to primitive elements directly. To analogize from the alphabet, 'material' might be A, 'living' B, 'human' C and 'artifact' D. If abstraction (in Locke's sense) were employed, 'material' might be A, 'living' BA, 'human' BAC and 'artifact' CAD. If the alphabetic concatenations are treated as names it can be seen that the two relations have equivalent results.

-red-, -thin-, and irregularity-of-shape, qualities have no existence outside of the things they convey*. Green ink or paint might initially lead us to think otherwise, but it is impossible to have a material thing which is solely green; green paint has a paint character above and beyond its greenness. Conversely, a typically green quality may be part of a person's concept of 'leaf', but that does not mean that the two concepts are synonyms or share more than the intersection set of perceptual experiences from which both are derived.

The categorical process can be applied repeatedly to the same objects, with each new category and criterial concept becoming farther and farther removed from the original perceptual basis. Thus from things we learn concepts of -red- and -green-, -heavy- and -light-, -round- and -straight-; and from these concepts we learn ones of -color- and -weight- and -shape-; and in turn from these concepts we learn -sensory-mode-. Notwithstanding this, many concepts contain sensory components or images which, as the name might suggest, are primarily visual⁶⁴. The images associated with concepts have only a fraction of the intensity of direct perception; if we want to visualize something, we are usually forced to close our eyes in order to 'see' it.

Meaning and concepts are linked in this way: we can conceive of many more concepts than we find meaningful⁶⁵. We retain meaningful concepts because they have some connection with the world and eventually discard those which do not. In practice we can assume that meanings and concepts are isomorphic.

* aspects, which may or may not be the totality of a concept's being, are either on the same level of the 'attentional-judgmental abstractions' introduced by Miller and Johnson-Laird (1985) or on the level immediately superior to them. Thus 'green' is equivalent to $\text{Color}(x)$ and 'speed' is induced from $\text{Place}(x,y1) \ \& \ \text{Time}(e1,t1)$ plus $\text{Place}(x,y2) \ \& \ \text{Time}(e2,t2)$ and probably $\text{Travel}(x)$. These sorts of similarities are called 'qualia' in philosophy (Lacey, 1986) and fit some of the traditional definitions of 'universal' (and of 'idea', and of 'form') used in metaphysics.

Psycholanguage

The involuntary nature of most perceptual experience contrasts with the deliberate way in which we must pay attention to understand speech we hear or text we see. If we fail to focus on a linguistic message we do not grasp it. After overhearing a conversation to which we did not pay attention we may be able to describe the tones of the speakers' voices or who took longer turns speaking without being able to recollect with certainty a single word uttered. Thus one difference between hearing sounds and listening to speech, seeing letters and reading words, is the voluntariness of the experience.

Words also differ from other sensory experiences in the way in which they convey their content. Vision* operates in parallel to give us all elements of an image simultaneously; we see an entire picture or a whole page of words in an instant. To read a page however we must spend a perceptible amount of time focusing on its individual words one after the other, recognizing them and using their meanings to incrementally expand our understanding of the text's complete message.

Both of these differences are accounted for by the fact that words are semiotic. A word is understood only after we recognize its form and recollect its sense(s) from our memory. Thus words require interpretation once they themselves have been perceived, something which is not true of most direct sensory experiences[§]. On the other hand because a word has no inherent meaning, it can act as a token for whatever meaning or meanings language users choose.

The objective or dictionary meanings attached to words are necessarily more general

* hearing differs from sight in that we cannot perceive the entirety of a sound in one instant. Notwithstanding this, listening is a process almost entirely akin to reading in its sequential nature; if we can jump around the page with our eyes we defeat the organization and sense of the text as a whole.

§ there are other symbols with arbitrary meanings established by convention. There is no inherent reason why a white flag raised on a field of battle should express the surrender of an army. Words lack the tangibility of many non-verbal signs; one can literally wrap oneself in a flag.

than the concepts that arise from directly received sensory experiences. There is a reason for this. An individual bases his concepts on sensory experiences of particular instances which contribute a great number of accidental details in addition to those included in the concept. I never see a dog; I always see a black or brown or white one. As a result my idea of 'dog' is somewhat different from yours.

However for a word to be generally intelligible across a population, its meaning must be limited to the intersection of the various understandings different users have of the concept it labels. Confusion occurs when a person does not observe this restriction; we have all been misled by someone's use of a word in an idiosyncratic way. Dictionaries adjudicate the common understanding. Thus every person develops a set of concepts that arise from his own experiences. By trial and error we learn to associate words with the meanings which are generally understood for them. Once this has been done we are able to share new concepts with other language users via utterances which combine words' dictionary meanings in new ways.

Significant differences exist between the two modes through which we receive language. Cohen, Fertig and Starr (1982) found that speech is wordier than text and has a different structure. It is more fine-grained and is dominated by speech act requests to identify objects in the field of view. The prosody of speech conveys substantially more information than that of writing; speech has an intonation and speakers have aspect* (this is why we look at people to whom we are listening). Finally, sensations in the aural field are less localized than those in the field of view making it is harder to focus with the ear than with the eye.

These differences deserve discussion because they bear on the presumptions on which compositional semantics are based. Before the phonograph and the wireless made it portable, speech was tied to the particular times and places in which a speaker spoke and

* aspect is facial expression, carriage and demeanor; hearing may or may not involve seeing but reading never involves listening. Television communicates aspect and radio does not.

an audience listened. Aware of the context in which his message was received and understood, a speaker could use shared knowledge and ostensive references. To a limited degree, the medium was bidirectional^{*}; a speaker could tailor his message to the identities of his listeners and their reactions.

Although some texts do make use of their context[§], writing has the potential to be acontextual or autonomous of its physical context. As soon as words are written on a portable medium like paper, authors lose control over who will receive their message and the context in which it will be interpreted. Lacking a known physical context, texts require a pragmatic one, ie a set of assumptions about texts that writers and readers share. For example, as I write this there is a plaid armchair five feet ahead and to the right of me. I know that I cannot count on the reader having a similar chair in his field of vision, or even make a reference to my chair until I have introduced it (as I just did).

Some pragmatic assumptions are quite basic and span all texts. A reader can expect a text's literal content to be chosen systematically rather than at random[‡]. He also reads expecting to encounter sufficient information to communicate a coherent message. A reader can therefore make routine assumptions about issues to which no reference is made, confident that these assumptions are either appropriate or irrelevant and in any case do not defeat the purpose of the communication. These fundamental and pervasive assumptions are called Charity, Cogency, Cooperation and Communication etc in psycholinguistic literature 66.

The context shared by reader and writer may contain more specific pragmatic information. An author may know his intended audience and be able to make additional

* truly bidirectional spoken utterance where both participants take turns is dialogue.

§ a placard posted at a scenic lookout which is oriented and fixed so as to be sure to describe what lies in the reader's field of view.

‡ however a primary purpose of language is simply to communicate information for purposes to be determined by the recipient rather than the utterer. And even if a text is intended to serve a particular end, there is no guarantee that a reader will not use it for some other purposes. As a result, apprehended narratives may never be subsequently used.

presumptions in light of this knowledge; and a reader may be able to discern information about a text above and beyond the sense of its words, information such as who wrote it, when it was written, and the author's circumstances at the time of writing. However the amount of specific pragmatic information that can be gleaned from a text is variable and may be non-existent. Specific pragmatics cannot therefore play a necessary role in the semantic theory of written language. Therefore text has an irreducible and minimal sense exclusive of any particular context which is based on the constituent words' dictionary meanings.

To recapitulate: because a text can express its meaning only through the words that compose it and without recourse to its surroundings, that meaning is at least minimally comprehensible from the literal senses of its words even if there are additional, perhaps overriding, pragmatic senses. These observations are language-invariant and they support the concept of a feature system.

Language

A natural languages is composed of a lexicon and a grammar. A lexicon is a list of words recognized in the language and a grammar is a set of rules for combining words into well-formed or correct expressions. Grammars operate on all four linguistic levels (morphological, syntactic, semantic and pragmatic) and these levels are not distinct from their neighbors. For instance, it is only because number and gender* affect syntactic correctness that they are considered the province of syntax while color or shape are not. I speculate that the originally 'noun' and 'verb' were semantic concepts meaning what 'entity' and 'formal relationship' do in the feature system. The descriptive senses of these were lost when the particulars and inconsistencies of usage were codified in grammars.

* the question mark and exclamation mark have semantic senses in addition to their syntactic uses.

Seeking to develop computation models of the world described by language, we now return to the original senses with new terms.

Meanings are distinct from the words that convey them. Every word has at least one meaning and can have more than one (homonymy). Conversely some meanings are expressed by more than one word (synonymy) while others are not embodied in a word (lexical gap). Words are objective and known; meanings are subjective and in some way ineffable. Although a word can have a variety of different meanings⁵, its literal one is most integral (Leech, 1981) and can therefore be called its canonical sense.

It should be noted that some words or symbols have almost no meaning. There is little sense to grasp in a telephone number's arbitrary sequence of digits, and the need to cue certain recollections by reciting their items suggests that rote memorization is an almost 'sense'-less apprehension. Nonsense words and babies' babbling (unlike grunts or groans, a baby's enunciations are phonetically well-formed and could be words) are other words with almost no intrinsic sense.

People use natural language to communicate concepts more complicated than the sense of any individual word. For this reason individual words rarely occur in utterances by themselves; instead, they are combined into clauses and sentences according to a language's rules of grammar. The sentence is the unit of syntactic structure but the clause is the unit of conceptual structure*. This suggests that many sentences and sequences of sentences could be ordered differently without changing their meanings and with only minor periphraxis on the clausal level. People also usually want to communicate conceptual sets too elaborate to be communicated by a single sentence. As a result single sentences rarely occur in natural language; conceptual packages are conveyed by groups of related sentences called episodes. Because utterance is linear, this imposes an organization on a message's presentation which may not be particularly suitable to its contents.

* language structure is similarly arbitrary on the word level. Any number of meanings can be realized syntactically (by phrases) rather than lexically (by single words).

Clauses and sentences are organized around their verbs, and in the absence of explicit information to the contrary, verbs are understood to occur in a temporal order which follows the order of their occurrence in an utterance.

An initial model of linguistic truth holds that unqualified language contains propositions which are true and consistent with each other and occur in an objective universal context. When the expression of language is qualified, readers understand that its propositions occur in limited subjective contexts and are not necessarily true or consistent with the objective context. A typical subjective context is an person's set of facts (propositions he holds to be true but which need not be objectively true) and beliefs (conjectures which the person knows to be not necessarily true). Beliefs vary in type* and likelihood.

The first extension of this model accounts for linguistic communications passing from a speaker to a hearer. Each individual has distinct subjective contexts[§] for the set of statements he has made and the set of utterances he has heard. Since people can lie, statements made are not necessarily true but have truth values known to the maker; and since we know that people can lie, utterances we hear have unknown truth values which are not absolute. The second extension of the model accounts for the plurality and combination of speakers and utterers in the objective context. Individuals participate in groups which speak on behalf of their members, and instances of communication occur where there are multiple speakers or hearers. An individual may therefore be a co-hearer or co-speaker[‡].

* operational premises, imaginative speculations and esthetic conventions are a few of the different types of conjectural propositions. The question mark and exclamation mark have semantic senses in addition to their syntactic uses.

§ these contexts are differentiated. Utterances heard and statements made are distinguished on their speaker or hearer, and are usually more consistent with one another than with the utterances of other speakers or statements to other hearers. We try to be consistent in our statements.

‡ the idea of co-speakers is a simplification suitable to a prototype model. Strictly speaking, only individuals can make utterances. The degree to which group communications express their beliefs of their members ranges from jointly-authored reports to government legislation.

A.7

Semantic Features and Words Defined in the System

The system contains 307 semantic features.

.*	.+	.-	/
.ABSTRACT	.ACROSS	.ACT	.ACTIVE
.ACTIVITY	.ADJECTIVE	.ADVERB	.AGE
.ALONG	.AMORPHOUS	.AMUSEMENT	.ANGER
.ANIMAL	.ANIMATE	.ANXIETY	.APPETITE
.APPROXIMATE	.ARITHMETIC	.AROUND	.ARTICLE
.ARTIFICIAL	.ASSERTION	.ATTENTION	.AUXILIARY
.AXIS	.BANG	.BIRD	.BRIGHTNESS
.CAPABILITY	.CASE	.CAUSATION	.CAUSE
.CHANGE	.CHARACTER	.CIRCUMSTANTIAL	.CLASS
.CLOSEQUOTE	.CLOSEQUOTES	.CLOTHING	.COLON
.COLOR	.COMMA	.COMMON	.COMMUNICATION
.COMPARATIVE	.COMPARISON	.COMPONENT	.COMPLEX
.COMPREHENSION	.CONCORDANCE	.CONDITION	.CONJUNCT
.CONJUNCTION	.CONSCIOUSNESS	.CONSTITUENT	.CONSTRAINT
.CONSTRUCTION	.CONTACT	.CONTINUOUS	.COUNT
.CREATION	.DEFINITE	.DEGREE	.DEMONSTRATION
.DEMONSTRATIVE	.DENOTATION	.DESTRUCTION	.DIRECTION
.DISCONTINUOUS	.DISCRETE	.DISJUNCT	.DRINK
.DURATION	.ECONOMIC	.EFFECT	.EFFORT
.ELASTICITY	.EMOTION	.ENERGY	.ENTITY
.EQUAL	.ESTHETIC	.EXACT	.EXACTNESS
.EXCLAMATION	.EXISTENCE	.EXTENDER	.EXTENSIVE
.EXTERIOR	.FAR	.FEAR	.FEMALE
.FISH	.FLAVOR	.FOOD	.FORCE
.FREQUENCY	.FROM	.FRUIT	.FURNISHING
.FUTURE	.GAS	.GENDER	.GLOBAL
.GOAL	.GREAT	.GREATERQUAL	.GREATER
.GREEN	.HABITATION	.HAPPINESS	.HARDNESS
.HAS	.HEALTH	.HEARING	.HEAT

.HORIZONTAL	.HUMAN	.HUMIDITY	.IDENTITY
.IMMATERIAL	.IMMOBILE	.IMMOBILITY	.IMPORTANCE
.INANIMATE	.INDEFINITE	.INDOBJ	.INFINITIVE
.INGREDIENT	.INHIBITION	.INSECT	.INSTRUMENT
.INTEGRITY	.INTELLIGENCE	.INTENSIVE	.INTERIOR
.INTERROGATION	.INTERROGATIVE	.INVOLUNTARY	.IS
.LENGTH	.LESSEQUAL	.LESSER	.LIFE
.LIQUID	.LITTLE	.LOVE	.MACHINE
.MAIN	.MALE	.MAMMAL	.MATERIAL
.MEASURE	.MECHANISM	.MEMORY	.METRIC
.MIND	.MINIMAL	.MOBILE	.MOBILITY
.MODERATE	.MODIFIER	.MOOD	.MOTION
.MUCH	.MULTIPLE	.NAMING	.NARRATIVE
.NATURAL	.NATURE	.NEAR	.NECESSITY
.NEGATION	.NEGATIVE	.NEGJUNCT	.NONE
.NORMALCY	.NOUN	.NOURISHMENT	.NUMBER
.NUMERAL	.OBJECT	.OBLIGATION	.ODOR
.OPACITY	.OPENQUOTE	.OPENQUOTES	.ORDER
.ORIENTATION	.PAPART	.PART	.PARTICIPLE
.PASSIVE	.PAST	.PATENSE	.PERFECTION
.PERIOD	.PERSON	.PERSONAL	.PERVASIVE
.PHRASE	.PLANE	.PLANT	.PLASTICITY
.PLURAL	.POSITION	.POSITIVE	.POSSESSION
.POSSESSIVE	.POSSIBILITY	.POTENCY	.PRECEDING
.PREPOSITION	.PRESENT	.PRIDE	.PRIMACY
.PROCESS	.PRODUCT	.PRONOUN	.PROPER
.PROPULSIVE	.PROXIMITY	.PRPART	.PRTENSE
.PUNCTUATION	.PURPOSE	.QUALIFIER	.QUALITY
.QUESTION	.QUOTATIONAL	.RADIATION	.RATIONALITY
.READINESS	.REALITY	.REASONING	.RECOGNITION
.REFLEXIVE	.RELATION	.RELATIVE	.RELATOR
.REPTILE	.REPUTATION	.RISK	.ROLE
.SCOPE	.SEE	.SEMICOLON	.SENSE
.SEQUENCE	.SET	.SIGHT	.SIMPLE
.SINGLE	.SMELL	.SOCIETY	.SOLID

.SOME	.SORT	.SOUND	.SPATIAL
.SPECIFIC	.SPECIFICATION	.SPEED	.STATE
.STATEMENT	.STRUCTURAL	.STYLE	.SUBJECT
.SUBSTANCE	.SUBSTANTIAL	.SUCCEEDING	.SUFFICIENCY
.SWEET	.SYNTAX	.TASTE	.TEMPORAL
.TENSE	.TERMINAL	.TEXTURE	.THING
.TIME	.TO	.TOOL	.TOPIC
.TOTAL	.TOUCH	.TRANSFER	.TRANSFORMATION
.TRANSPORTATION	.TWO	.TYPE	.TYPICALITY
.UNIT	.UNIVERSAL	.UNIVERSALITY	.VEHICLE
.VERB	.VERTICAL	.VICINITY	.VISCERAL
.VITALITY	.VOICE	.VOLUNTARY	.WEIGHT
.WIDTH	.WILL		

The system contains 203 defined words.

Ottawa	a	about	above
across	across	from	after
against	ahead	ahead of	all
all over	all through	along	alongside
also	altogether	amid	amidst
among	amongst	an	and
anger	apart	from	are
around	arrival	arrive	as
at	away	away from	ball
be	because	because of	been
before	behind	being	below
beneath	beside	besides	between
beyond	book	boy	but
but for	by	can	certainly
city	concrete	could	despite
did	do	does	dog
doing	doubtless	down	during
enough	even	except	except for

for	from	gave	girl
give	given	green	had
hard	hardly	has	have
having	he	her	hers
him	his	hour	in
in front of	in spite of	inside	inside of
instead	instead of	into	is
it	just	leg	less
like	likewise	love	man
may	maybe	men	might
mile	minus	morning	move
must	near	nor	not
notwithstanding	of	off	off of
on	on top of	only	onto
opposite	or	out	out of
outside	outside of	over	over to
page	past	perhaps	plus
pretty	probably	quick	quite
rather	read	really	reason
shall	she	should	since
slightly	somewhat	speed	sufficient
surely	that	the	them
they	this	through	throughout
till	times	to	too
towards	under	undemeath	unlike
until	up	up to	very
walk	was	well	were
will	with	within	with the exception of
without	work	would	young

A.8 Thematic Evaluations and Remarks on the Feature and Linguistic Systems

This appendix is an attempt to evaluate the systems proposed here in terms of traditional topics and issues. In addition to these general matters, certain details of the implementation seems interesting enough to bear repeating. The following remarks apply variously to feature systems of the same general class as this one, to the particular feature system specified here, and to the system implemented in the program. Theoretical and practical issues are discussed separately and problems encountered in any stage of development appear in a section of their own.

Theoretical

Change. Things change over time both in accidentals and fundamentals and a representation must be able to incorporate alterations to its contents based on new information ³⁵. It has already been noted that certain of a word sense's features are essential and that others are accidental, and that enough essential ones must be represented in a word meaning to allow it to be distinguished from all other words. If we look behind words at the concepts they embody, we see that a concept becomes fundamentally different when one of its essential features changes. An individual is no longer a husband if a medical operation changes it from man to woman. Changes in accidentals are different. We consider a fat man who becomes thin to be the same man and see no inconsistency between his present and past states.

However necessary and accidental features are not sharply divided. In the definition of 'man', having-a-head may be necessary, -hirsute- is expected but not necessary, -blond- is possible but not expected, and -green-haired- is unlikely. Someone is human without having hair or even without having limbs. However as expected accidental

features accumulate, they tend *en masse* to achieve a state of necessity. Alternatively, we can start with an individual who would generally be regarded by onlookers as human and, by subtracting more and more of its typical features*, move through a position in which its humanity is controversial to one where most observers would not see the entity as manlike. In fact in the features of natural kind things, necessity may simply be a shorthand way of expressing 'an accumulation of expected features sufficient to identify'.

Inconsistency—fat to thin to fat again—is sometimes not merely tolerable but expected. A hungry man who does not change his state will not be a man for long. Every young thing becomes old. Such facts are the essence of pragmatics.

Non-monotonicity can be accommodated in a number of ways: if links and symbols created under conditions of uncertainty are marked they could be reversed; discarded senses could be kept and reinstalled if they gain credibility; or the whole utterance could be reparsed in the light of new information. The system has the right to reevaluate and reverse controversial elements if necessary. When this perspective is generalized from particular symbols to entire contexts it could result in copies of previous environments or continuations being kept for the last few previous sentences.

Ambiguity. In the simplest possible terms, disambiguation is a process by which all words or phrases (idioms) in an utterance start with all their dictionary senses and these are systematically eliminated on the various grounds provided by the utterance until each word or phrase is left with a single sense. The first stage in this activity is determining which part of speech each word plays in the sentence; this usually eliminates a large number of senses.

Once that has been done, disambiguating a word's alternative senses is organized around utterance network nodes. In order to disambiguate a word's meaning its senses

* senses, form, a brain, life.

must be compared with those of other words in the same node and a consistent subset of the senses of each selected. Disambiguation can be undertaken on each node in isolation immediately after it has been constructed by the parser if the node has been modified by relations which might serve to identify each other's and the node's own correct sense. The process can be reinitiated when an ambiguous symbol is linked to others in the utterance network and again when that structure is integrated with the episode network if either of these two events adds additional constraints on a symbol's meaning.

In this system, words naming things have few restrictions on the contexts in which they can occur by their nature. Relations restrict the classes of things which can hold them and thereby permit the appropriate sense of a word to be distinguished. For this reason entity nodes have no provision for associated restrictions. Relation restrictions can be explicitly attached to .slot pairs in senses themselves or occur implicitly through relationships between features. There are a number of different sorts of constraining relationships between features: different features may belong to the same domain and hold ancestry (consistent) or sibling (inconsistent) positions towards one another, they may belong to different domains which hold 'affect'/'accept' or 'has'/'qualifies' to one another, or they can be explicitly linked in a unidirectional 'implies' relationship.

Some practical conclusions about processing are in order. Consistency is transitive, so disambiguation processing can be restricted to testing the consistency of the senses in a symbol's .sense category's single entry with the senses of each of entry in its .relation category and does not need to be performed between pairs of .relation entries. On the other hand, disambiguation interacts between roles in a relation, so whenever an ambiguous* relation has a sense eliminated through disambiguation with one role it should be tested against its other roles to determine further advances are forthcoming.

* which remains ambiguous after internal disambiguation and external efforts.

Cases. Given the task of identifying various sorts of relations between things, one can either 1) attach the identification to the relation and have a small number of cases appearing in many instances; 2) identify a large number of cases and thereby decrease the variety of kinds of relations. Sparck Jones and Boguraev (1987) take the latter approach; the former is used here ⁶⁷.

Cases express relationships and most relationships are directed. This directional character must be expressed in some way in any representation. Three alternatives suggest themselves; 1) using paired names for each relation; 2) adding a directional specifier to a single relation name; 3) employing conventions so as to avoid direct representation.

The third option offers an advantage (a economical representation is possible if a relation's direction does not need to be represented) so let's examine it first. The issue to resolve is whether a purely conventional representation gives rise to situations that can be interpreted equivocally. Modifier relationships can readily be understood because they are monadic and not links between different individuals; their sense is simply inserted in the appropriate place in the representation of the object they apply to and interpretation is not in question. Acts can also be understood easily when their cases are held by nouns. In "The boy runs", 'boy' can be seen as the subject of 'runs' because no other interpretation is intelligible. However confusion arises when a case slot is filled by a nominalization; 'winning pleases' can be confused with 'pleasing wins' if the two concepts are linked through an undirected subject case. The problem is even more common with prepositional phrases and conjuncted clauses such as 'the boy on the bench' and "How you win counts". A purely conventional approach is out.

The other two approaches have enough power to express directed character. The first does so at the cost of verbosity and obscurity; so long as one understands what .subject and .object mean, .part .subject and part .object distinguish between 'finger' and 'hand' in 'the fingers of my hand'. However this understanding is arbitrary and must

therefore be explicitly articulated. The second approach, using unique terms like *.part* and *.compound* for the same example, has much to recommend it. Its expressions are shorter and clearer. This approach runs into difficulty when one tries to find terms to express the two directions of more elusive relations such as the *.style* relationship in 'he spoke with great dignity and surprising candor'*. If exactly descriptive terms cannot be found, new ones can be coined either from suggestive roots (*.fugal* for 'from', *.transjacent* for 'on the other side') or from phrases which express the correct sense (ie, 'is-a'), or by compounding the relation name and either '-subject' and '-object' (or a derivative). In the first two cases a convention of interpretation is again required, while the third invites giving the entire relation a different, descriptive name that suggest its two directional components.

The approach used here is the first, with *.is* replacing *.object* and *.has* replacing *.subject* for greater expressiveness and so that the latter terms can be used to identify the positional cases of *.act* relations. *.Is* and *.has* are interpreted in this way: *.is* identifies the thing which is affected by the relation-as-named and *.has* identifies the thing experiencing the relation from the opposite direction. (*.Ispart |thing-1|*) appearing in *|thing-2|* identifies *|thing-2|* as a part of *|thing-1|* and in *|thing-1| (.haspart |thing-2|)* identifies *|thing-2|* as one of its parts. Formal relationships are expressed differently than informal relationships in the representation; the sense of an *.act* is stored in a separate symbol and a pointer to that symbol is placed in each of its case-holders' lists of relations. As a result the sense expressed by *.part* etc is only accessible at one remove. Accordingly, (*.is .subject |act-1|*) appears in *|thing-1|*'s symbol while (*.has .subject |thing-1|*) appears in *|act-1|*.

Processing Strategy. Frequently inadequate information is available to allow semantic interpretation of an utterance to be made with certainty at the end of syntactic processing.

* Although some *.style* relations can be rendered as adverbs ('he came with speed' -> 'he came speedily'), this is not true in all cases and it is not feasible to eliminate these obscure relations by converting them to other representational forms.

And even when the necessary character of constituent relations permits an interpretation to be determined with complete certainty, there is no guarantee that an utterance is self-consistent and that subsequent input will not contradict the current construction placed on it. In many cases it is unnecessary to solve problems of linguistic interpretation because the problematic construction will not figure in future processing. These points, as well as the inherent economy or 'lazy evaluation', argue for the postponement of secondary semantic processing and problem resolution until it is actually needed.

Operations on the representation should therefore be conservative, require fairly conclusive evidence, and be corrigible if possible. As a practical demonstration of this attitude, where Schank attaches a maximal set of conceptual case predictions to the actions represented by verbs (Schank, 1972 in Moyne, 1985) the linguistic system makes no assumptions. If the text supplies a tenant for a case role, the role is added to the utterance network (and fills the existing empty slot in Schank's representation).

Practical

Inheritance. The classic example of an inheritance problem is the observation that although penguins are birds they do not fly. While it should be noted in the linguistic system that word sense descriptions contain only the minimum features necessary to distinguish among senses (at least initially) and that they almost always employ structural rather than behavioral criteria, the problem still needs to be addressed. Some birdy could lack wings and wings could be criterial to the description of 'bird'.

Inheritance is intended to increase information density and unfortunately both of the fixes that suggest themselves add structure to the system and lower that density. The first fix would be to qualify the .existence of the .wing pair in the definition of 'bird' so that it would be understood as .measure .great rather than total .assertion. This is veridical; If birdy does not have wings than not all birds do. Alternatively, the definition of 'birdy'

could have a negated .wing pair added to its own definition. As is usual in such matters, the local values would override the inherited one. b) Defaults. Certain values are assumed implicitly and others are explicitly used to establish initial domain values which may be modified during subsequent processing. Default domain values are: .verb .syntax is .prtense; .number is .single; .existence is .assertion (100% confidence); .mood is declarative and .person is .third.

Role Tenancy. The selection restrictions ²¹ that relations impose on the entities that hold them can be reasonably complex. A relation might require that a thing have particular features from more than one domain or accept more than one combination of features. The first case requires that a series of domain lists be delimited: ((nn ...) ...) while the second requires that these delimited lists be themselves delimited: (((nn ...) ...)(oo ...) ...)). .Act relations accept three cases; to distinguish these, a key can be added to the head of the restriction list: (.agent ((nn ...) ...)(oo ...) ...)).

Restrictions are interpreted in this way: a candidate thing is accepted if it has concurring domain values and rejected if it has conflicting domain values. If it lacks specification in a domain it is accepted if no other candidate provides a better match. Thus a pair with a key but no value will match any member of the associated domain; while a key is not the domain head, will obtain the same relationship with the appropriate subdomain.

A second complexity arises in the interaction between different cases in a relation's sense. These can co-occur in three ways: tenants must satisfy one, two or more case criteria independently; tenants must satisfy one of two or more case criteria independently; two tenants must satisfy the criteria of different cases at the same time.

Although the first two cases can be handled by the standard or-and representation, it is difficult to represent linkage between different cases. An alternative approach sees the alternation level deleted from the representation and the convention established that two different domains must be satisfied at the same time for consistency to be achieved.

Alternation would be expressed by duplicating the sense definition in the word with the alternate conditions. Since the implied restrictions or features are a part of the sense's meaning, this is a theoretically satisfactory treatment; two different ways of satisfying the restrictions do suggest two different meanings.

Organization. In any choice among a set of alternatives the most specific satisfactory one should be selected first. Alternatives should therefore be organized so the most specific cases are tested before more general cases which might also satisfy the test criteria. Because the most specific cases are often also the most infrequent, least- to most-specific organizations are slow to process. To minimize this performance cost, an alternative's most restrictive requirements should head its set of conditions. An 'else' case or its equivalent is required unless failure of all cases is acceptable.

In the interests of parsimony, pairs in a word sense contain only the domain head and the value of interest. In the interests of efficiency, intervening values in the linkage connecting these two features appear in 'is' (hyponymic) relations because of this relation is so frequently looked up. .Human is defined as .class .material .animate .animal .mammal. If it only .class .mammal in its 'is' category (as it would were this another category), a check for consistency with .material would require accessing each of the three intervening values.

Condensation of other domain pairs proceeds according to these rules: the domain head must appear as a key to identify the pair; if the value of interest does not lie in the keyed domain but in a domain which affects one of its values, the affected value must appear in the pair as a pointer to the value of interest. Given that all domains can affect the .abstract subdomain of .class, the construction (.class .abstract .time) is necessary to indicate that we wish to represent time on a metalevel, ie to talk about time rather than to give a value of time.

Subordinate Clauses. Subordinate clauses cannot be positioned in sentences arbitrarily. Consider a sentence such as "He will talk to us when we come". The subordinate clause "When we come" could only be located in three of the six interword positions in the superordinate clause "He will talk to us": "When we come he will talk to us", "He will talk when we come to us" and the original "He will talk to us when we come". Other examples are more restricted. In "the man who I saw is my father's friend" the subordinate clause can only appear in one other place, at the end of the sentence.

Processing would proceed in this way. When a subordinate clause is encountered, the current parse contents would be stacked and a new parse originated with any relative pronoun introducer of the clause being treated as a legitimate tenant of one of its verb's cases. Processing would end with the end of the sentence or when grammatic or semantic inconsistency is encountered*. In terms of the first example's alternative forms, this occurs at "When we come he will" in "When we come he will talk to us" (two verbs); at "When we come to us" in "He will talk when we come to us" ('to us' is semantically inconsistent with "We come"; contrast with "He will talk when we come to the office") and at the sentence end in "He will talk to us when we come".

From the point of view of actual processing, pro-forms would be required in both sub- and superordinate clause parses to link them even if one did not appear in the utterance. This form would play different roles in each clause. Typically it would be a qualifier in the parent clause and .subject or .object in the child. When a subordinate clause was completely parsed the results would be stacked and not be processed further until the superordinate clause was also completely parsed.

I cannot determine whether there are sufficient grounds to choose between

* processing should include copying unmatched relations in correlative clauses to produce the full sets in each clause. In "On Sunday the boy sang and the girl danced at the show", a .spatial relation is added to 'sing' and a .temporal relation to 'dance'.

bottom-up or top-down processing of individual subordinate clause parses*. A bottom-up approach is more localized, and subordinate clauses do not depend on their superordinates for interpretation but rather vice-versa; on the other hand, most other processing in this system proceeds in a left-to-right or forward direction.

In any case the superordinate parse and the stacked subordinate parses would be stacked again after being processed into individual conceptual network fragments. Final consolidation of these individual networks with the existing episode network would proceed from super- to subordinate network, with pro-form links between clauses being replaced by the structures they represent. In the first example 'the man' would replace 'who' in the "who we saw" clause, and a pointer to the entire clause (ie to the 'saw' network) would replace 'who' as the object of the qualification.

Problems

Banality. As the number of features in the system increases the uniqueness of each diminishes and it becomes increasingly difficult to tell similar features apart. The difference between .local and .distant is clear but that between .local and .vicinity is less so. There is little in the literal sense of either of these two words to suggest that the first refers to a thing's frame of reference and the second to the spatial proximity of two things. This supports the idea of 'semantic distance' as a veridical concept, at least for words. It also highlights the inadequacy of words as media of communication; human beings have no difficulty distinguishing mentally between -reference-frame- and -spatial-proximity-, but there is no simple and concise way to capture that difference linguistically.

Idiosyncrasy. The system is based on categorical criteria peculiar to each domain rather

* or the parse at the bottom of the stack if there is more than one.

than ones applicable across many domains. As a result criterial aspects may proliferate and introduce an element of incommensurability to the representation; a man could be .propulsive in the context of mobility, .oriented in all three spatial axes etc. It is not clear whether the process of adding features would eventually exhaust this idiosyncrasy.

Pair Structure. It is difficult to know on what basis to group features in word senses into pairs. On one hand, all features in a sense could conceivably be placed in a single flat list with sequence determining their application. This could cause a problem when two features from the same domain appear in a sense (ie .measure). Distinguishing between them would require looking at their context; because that context has not established structurally, it would need to be identified by code—a great deal of code.

Alternatively a series of features in a word sense which constitutes a whole could be separated from other features by being embedded in a list. Defining these pair structures causes the aforementioned problem. If 'boy' is not going to be (.class .human .age .young .gender .male), which of the possible permutations and groupings of (.class .human)(.age .young)(.gender .male) should it be? While I cannot enunciate a general rule, I can make one observation: if a sense's interpretation would be changed by moving feature(s) from one pair to another, then the list structure is significant; if it is not significant, in the interests of simplicity do not use it. While this might seem so obvious as to not require mention, it does suggest a methodology.

Alternative Criteria. Different criteria can describe the same semantic space. For example tridimensional space can viewed both intensively or extensively. An .intensive or thing-focused view partitions it into .length, .width and .height; seen .extensively or in an external framework, the same space is .vertical, .horizontal (.breadth, .depth). Similarly, verbs can be classified syntactically either on their five conjugate forms or on the main/auxiliary distinction.

No single and straightforward solution to this problem has been found. One approach is to define a contingent domain on one of the two criteria and then apply that domain to the other in cases where it serves to describe instances. Thus 'do' is both .prtense and .infinitive, and its .prtense sense can be either .main ('I do it') or .auxiliary ('I do wed thee.'). Another technique is to see whether one of the criteria is actually meronymic, in which case the classifications can coexist without conflict. The .slot domain interacts with .case in this way.

A.9

Feature Descriptions of Nine Prepositions

164 senses have been identified for 86 prepositions and for three conjunctions which express set relationships not otherwise defined. 14 of the prepositions are synonyms and four more occur only in compound phrases.

Descriptions for the nine prepositions used in testing appear below. The clause or phrase following each sense description illustrates its meaning.

([about| ((.syntax (.preposition
((.type .position .spatial .interior .pervasive)
(.subject ((.class .material)))(.object ((.class .material)))) *he looked about the room*
((.type .position .spatial .exterior .around)
(.subject ((.class .material)))'.object ((.class .material)))) *smoke swirled about the house*
((.type .position .temporal .vicinity)
(.subject ((.time)))(.object ((.time)))) *it happened about noon*
((.type .position .metric .vicinity)
(.object ((.unit))((.measure)))) *the fish weighed about a pound*
((.type .specification .topic)
(.subject ((.communication)))))) *a book about politics*

(([around| ((.syntax (.preposition
((.type .position .spatial .interior .pervasive)
(.subject ((.class .material)))(.object ((.class .material)))) *he looked around the house*
((.type .position .spatial .exterior .around)
(.subject ((.class .material)))(.object ((.class .material)))) *he walked around the car*
((.type .position .metric .vicinity)
(.object ((.unit))((.measure)))))) *a book costs around five dollars*

([despite| ((.syntax (.preposition
((.type .causation .inhibition .negation)))))) *he acted despite my request*

(for ((.syntax (.preposition		
((.type .position .spatial .to)		
(.subject ((.class .material)))(.object ((.class .material))))		<i>the bus for town</i>
((.type .position .temporal .interior .pervasive)		
(.subject ((.time)))(.object ((.time))))		<i>he worked for an hour</i>
((.type .position .metric .equal)		
(.object ((.measure)))(.unit))))		<i>he sold it for five dollars</i>
((.type .causation .goal)		
(.object ((.type .naming)))(.orientation .positive))))		<i>he did it for love</i>
((.type .causation .effect)(.subject ((.type .thing))		
(.object ((.type .naming)))(.type .quality)))(.emotion))))))		<i>food for thought</i>
(from ((.syntax (.preposition		
((.type .position .spatial .from)		
(.subject ((.class .material)))(.object ((.class .material))))		<i>they came from town</i>
((.type .position .temporal .from)		
(.subject ((.time)))(.object ((.time))))		<i>work from nine till five</i>
((.type .causation .cause)		
(.subject ((.type .act)))(.type .naming))))		
(.object ((.type .naming)))(.orientation))))))		<i>weak from lack of food</i>
(in ((.phrase "front of" "spite of")(.(syntax (.preposition		
((.type .position .spatial .interior)		
(.subject ((.class .material)))(.object ((.class .material))))		<i>we jumped in the pool</i>
((.type .position .temporal .interior)		
(.subject ((.time)))(.object ((.time))))		<i>it happened in the morning</i>
((.type .position .temporal .interior .pervasive)		
(.subject ((.time)))(.object ((.time))))		<i>he slept in the night</i>
((.type .position .state)(.object ((.class .immaterial))))		
((.type .specification .style)(.object ((.type .quality))))))		<i>it went down in an awkward way</i>
(like ((.syntax (.preposition		
((.type .comparison)(.subject ((.type .act))))))		
		<i>he did it like an expert</i>

([of] ((.syntax (.preposition	
((.type .specification .identity)	
(.subject ((.type .sort)))(.object ((.denotation .proper))))	<i>the city of Ottawa</i>
((.type .specification .substance)(.subject ((.type .sort)((.type .part)))	
(.object ((.type .ingredient))))	
	<i>boxes of cardboard</i>
((.type .component)(.subject ((.type .part))))	
	<i>the fingers of my hand</i>
((.type .possession)(.subject ((.type .sort))))))	
	<i>that car of John's is broken again</i>
 ([with] ((.phrase "the exception of")(syntax (.preposition	
((.type .specification .style)	
(.object ((.class .immaterial)((.type .quality))))	<i>he did it with panache</i>
((.type .set))	
	<i>they came with their friends</i>
((.type .component)(.object ((.type .part))))	
	<i>hands with fingers</i>
((.type .causation .instrument)	
(.subject ((.type .act)((.type .naming)))	
(.object ((.type .naming)((.class .artificial)((.type .part))))	<i>stir it with the spoon</i>
((.type .possession))))))	
	<i>he came with all his tools</i>

Notes

1. To date neither an integrated nor a sequential approach to processing syntax and semantics has been recognized as unequivocally superior. In recent work Hirst (1987) separates, Lytinen (1984) compromises (in Hirst's view), and Riesbeck and Schank (Riesbeck & Schank, 1978) integrate, giving primacy to semantic processing. Hirst notes that intuition tells us the processes proceed in tandem and cites a generally accepted psycholinguistic finding: do it as early as possible.

2. I can identify three influences on research into semantics: previous work in the field, the researcher's own view of the world, and his ideas about human cognitive function. With regard to my own understandings in these areas, references to NLU research appear in the body of the discussion, Appendix 5 and particularly in these notes; here are a few of my other opinions about this field.

Cognitive science is a pure science and artificial intelligence is its applied counterpart. Work in an applied discipline is primarily validated by results; theory plays a distinctly subordinate role and can be dispensed with in favor of a comprehensive development methodology so long as it yields satisfactory results. My work is in artificial intelligence; speculations appear here only to increase the cogency and intelligibility of the feature system. They are unsystematic and eclectic, and will not be defended against charges of internal or external inconsistencies.

The nature of the subject also argues for practical approaches. Natural language is too complex to be readily reduced to a formalism. For the foreseeable future computational systems will be based more on heuristics than on comprehensive analysis. For that reason no attempt has been made to formalize the feature system used here. If it is useful it can eventually be axiomatized; in the meantime, better that it be freely changeable without need to satisfy and be expressed in an theoretical superstructure.

Much computational linguistic research tends to focus on narrow issues and be undertaken in relative isolation, a trait rather caustically described by Jensen & Binot (1987). Hirst (1987) demonstrates a similar attitude when he speaks of artificial difficulties. These traits should be deliberately countered by trying to develop systems that emphasize breadth rather than depth of coverage and substitute serviceability for rigor. The field is also not sufficiently mature to permit a system's failure to meet human cognitive performance to be a legitimate criticism.

Techniques of human cognitive performance revealed by introspection and intuition are compelling to this degree—they show us one way of doing a task without proving that it is the only possible way. Because we have no other examples of higher intelligence to draw on this evidence is sometimes enough to

be conclusive. It also means that we cannot distinguish what is particular and what is general in our own intellectual circumstances, and that cognitive science and artificial intelligence will continue to be permeated by anthropomorphic terms and images.

3. Although their observations are directed towards theories based on formal logic, some of the methodological principles Hobbs *et al* (1987) suggest for the development of a semantic theory are pertinent here: 1) try to accommodate competing ontologies⁶³; 2) pay only minimal attention to philosophical black holes (notions like 'relevant' and 'normative'); 3) aim for concept characterization rather than definition. "One cannot generally expect to find necessary and sufficient conditions for a concept. The most we can hope for is to find a number of necessary conditions and a number of sufficient conditions. This amounts to saying that a great many predicates are primitives, but they are primitives that are highly interrelated with the rest of the knowledge base." (p.242).

4. Hirst (1987) remarks that much NLU research has proceeded without firm theoretical basis and has therefore been rather narrowly conceived. Putnam (1975 in Moyne, 1985) suggests that the reason semantic theory lags behind syntactic theory is because the prescientific concept on which semantics is based is less clear. Moyne (1985) extends Hirst's criticism to the entire field of artificial intelligence, citing supporting evidence from a variety of sources (Ringle, Sayre, Wilks and Schank). Notwithstanding this (or perhaps because of it), Moyne sees a general movement (in terms of Schank's (1979) three stages of AI programs) towards a 'deepening and broadening' of the knowledge representation and a simultaneous positioning of AI within the larger multidiscipline of cognitive science.

The atheoretical basis on which previous semantic studies have been undertaken has resulted a set of demonstration systems and research findings which cover the field in a haphazard and piecemeal manner. In an opening epigram Jensen and Binot's (1987) paper characterize a customary posture for NLU papers thus: "1) Here's a difficult language situation. 2) Here's the semantic (pragmatic/discourse/whatever) information necessary to interpret the situation as we humans interpret it. 3) Here's how I put this information into my system. 4) Therefore, my system can handle the situation."

5. Leach (1974) identifies seven different types of word meanings. In addition to its literal or conceptual denotation a word may have connotative, social, affective, reflected, collocative or thematic meanings. These different sorts are listed in their decreasing order of inherence in the word itself*. Connotative meaning has its usual sense. Social meaning is the social circumstance implied by a word; affective meaning is what a word communicates about the utterer's feeling and attitudes; reflected meaning is the sum of a word's senses other than the one in use; collocative meaning is the meaning a word picks up

from others it is habitually used with. Thematic meaning does not inhere in a word; it is the order and emphasis attached to a word by virtue of the syntactic structure of the sentence.

• the first five sorts of meaning are attached to the word itself and grouped under the heading 'associative'.

6. Bobrow and Winograd (1977 in Moyne, 1985) believe that concepts cannot be decomposed but must be represented by multiple (contextual) views; depending on the view chosen, kissing may be seen as seen-physical-act- or expression-of-sentiment. Charniak (1977, p.39 in Moyne, 1985) thinks that knowledge is 'represented by concepts at all levels of generality' rather than on a single level.

Schank on the other hand believes that there is 'available to the mind an interlingual, i.e., language-free, representation of meaning' (Schank, 1980). Katz (1972 in Moyne, 1985) posits compositionality and decompositionality as two of the four criteria of any semantic theory must meet (the other two are formality and a capacity to account for semantic properties and relations such as synonymy, antonymy etc.).

7. If a semantics based on primitives is troublesome, the original words of the utterance (or any satisfactory set of tokens) can be used in place of feature descriptions of word senses in the network, thereby guaranteeing that the original meaning in the text is maintained. A substitution of this kind will have no effect whatsoever on the network's internodal structure, and changes within nodes to fit them to atomic symbols would be straightforward. Disambiguation could still take place during parsing if word symbols were indexed on their meanings, ie 'BOW-3' would be the third sense of 'bow'. Since all knowledge in the system is represented explicitly by symbols, the sense of a token can be the same as that of a complex expression, one lookup and some processing further removed (that processing would be non-trivial, since the issue of symbolically representing the sense of 'BOW-3' remains to be resolved).

8. Wilks (1977 in Barr & Fiegenbaum, 1981) believes that there is no single correct set of semantic primitives and that particular selections are justified operationally. Five criteria to use in evaluating primitive sets are 1) finitude: the number of primitives should be finite and smaller than the number of words encoded; 2) comprehensiveness (coverage): each word sense should be uniquely identified and adequately described with essential rather than relational information, although this cannot be exhaustive; 3) independence: no primitive should be definable in terms of other primitives; 4) noncircularity: no two primitives should be definable in terms of each other; 5) primitiveness (minimality): no subset of primitives should be replaceable by a smaller subset. Criteria from other sources are 6) salience (expressiveness, suggestiveness): primitives should express significant aspects of meaning^{*}; 7) regularity (consistency): primitives should be interpreted similarly in different cases[§].

Primitive sets can express meanings in different ways. Sensory impressions are by nature positive assertions of particular data; primitives reflecting this characteristic identify particular semantic aspects by directly asserting them. However aspects can also be identified by elimination of others from a parent set. The difference is between -non-metallic- and -plastic-, -wood-, -stone- etc. Primitive systems which incorporate both inclusion and exclusion are more powerful than purely inclusive ones, but exclusive primitives require interpretation before they can be used (primarily to identify the correct parent set). Is flesh non-metallic? Is air or water?

• "These features express this word sense to my satisfaction".

§ "This feature captures the difference between these two word senses" or "This feature expresses the appropriate dimension of this word sense, given its use in that sense and the semantic analogy between the two".

9. Computer scientists use heuristics that offer suboptimal results in acceptable time and space when forced to deal with NP-hard and NP-complete problems. Algorithms employed in heuristically-based programs are either demonstrably incorrect or undemonstrably correct. Computational semantics is of this order of complexity, so designs to deal with it must operate in the uncharted area between well-understood working systems too small to be very useful, and formally rigorous theoretical systems too large to be practicable.

10. Because word definitions contain an inexhaustible fund of semantic detail which no conceivable representation could express explicitly, compositional systems can only describe and not exhaust word meanings. As Korzybski put it, "the map is not the territory" (in Chapnick, 1988b). Moreover there is some evidence that a residuum of concrete semantic content remains in a word sense after analysis is exhausted⁶⁷. A representer must therefore select which aspects of knowledge are important for his purposes and ignore the rest. The feature system circumvents this restriction by viewing sense descriptions as initial semantic pegs for words which will receive most of their (open-ended) meaning as a result of use in subsequently processed texts.

11. The analysis used here identifies a minimal set of features that allow word senses to be distinguished from one another. Thus if -non-flying- and white-and-black-coloured distinguish penguins from other birds, living-in-cold-climate will not be added to their description. However senses may be able to be distinguished from one another without the use of all their salient features, and the analytic process may need to be pushed further to some indeterminate point. Systems which do this will become more and more comprehensive, but the idea that word definition or exhaustive sense description can be achieved is a chimera¹⁰. Prove this to yourself by trying to completely define a word verbally.

12. Examples of historical taxonomies are the classical Greek 'Tree of Porphyry' (Rapaport, 1985), the categorization used in Chinese encyclopedia titled 'Celestial Emporium of Benevolent Knowledge' circa 200 BC (Borges, 1966 in Rosch, 1978). More recently the 1933 American edition of Roget's Thesaurus (1933/1979) classified a large English lexicon into exactly 1000 categories.

13. Examples of other relations between features are: 1) equivalence (of measure), the equality relationship between an hour and sixty minutes, a pound and sixteen ounces; 2) scale, the relationship between things and their order of magnitude, or the unit most appropriate to measure them; 3) typicality*, the relationship between a set and its most frequently-occurring member; 4) order[§], the relationship between siblings which are ordered within their class; 5) synonymy, the undirected relationship between semantic neighbors.

Particular cases of antonymy, a similar relationship, are expressed by the .correlate relation.

* typicality indicates the value to assume in the absence of other information.

§ the paradigm order is numeric—one, two, three; first, second, third—but things are also ordered by size (hamlet, village, town, city, metropolis) and by sequence (spring, summer, winter, fall).

14. Productive relationships are semantically salient. Meronymy is an example; when we identify a feature's parts we generate a mental inventory of other features, some of which we might not have thought of otherwise and some of which may not even bear names. In addition, the fact that a feature is a part is an important component of its sense.

On the other hand, relationships like scale and equivalence link already defined features which have a more important sense as units of measure in another domain. Many domains do not have typical values (what is the most typical color?) and selections for those that do should be based on experience rather than assumption (suggesting that processing texts might lead to modification of the feature system's structure).

15. Rigorous categorization can proceed on the basis of only one criterion at a time so a choice must be made between alternative categorical principles. An example may clarify the need for this. Imagine three species of birds exhibiting two different characteristics; one species has a hooked beak and webbed feet, the second only a hooked beak and the third only webbed feet. Any individual classification of individual birds into species and species into a hierarchy of sets must ignore one of the birds' characteristics; if it fails to do so, it may be ignoring potential categories or is at least capable of decomposition into simpler classes.

16. The relation of parts to wholes is a many-to-many relationship in the real world. Human beings have fingers; but so do apes, frogs and most other mammals. Yet it would be inefficient to create heterogeneously populated classes solely on the basis of their members' common possession of a part. Since word descriptions are intended to convey only enough of the most salient aspects of a word's meaning

to allow it to be distinguished from other words, a component thing will be considered to be a part of only its most obvious whole. Thus fingers will be considered a part of human beings, and their occurrence in other animals and artifacts will be recorded not in the feature system but through the representations of utterances that state the fact.

It is difficult to determine the most obvious meronymic relationship; a bolt must be a canonical part of a great many different things. Two ways to classify meronymic relationships seem obvious; one is to generalize from particular sorts that incorporate the part to some superordinate which includes all the sorts as members, a sort of 'lowest-common-whole'. In the bolt example, one might initially generalize from engines, tools and vehicles to 'machine'. After noting that bolts are used as fasteners in building construction and other non-mechanical situations, one might generalize further to 'artifact'. This is very broad but it does encompass all bolts. Alternatively, putting a feature in a contextless frame ("What is a bolt?"), may help identify its most obvious or default sense (a bolt is a fastener) and thereby the appropriate relationship.

A problem arises when the two techniques identify different meronymic relationships. Consider 'leg'. The most general sort of thing that has legs is 'animal'*. However if I were asked to identify what a 'leg' is a part of, I would say a human being.

- * although machines, furniture and other things can have legs, the feature is not part of their canonical definitions. Such uses originate as metaphors and gradually become alternative word senses.

17. Some linear qualities have a central value of unique importance relative to which other values are located by symmetric positive or negative measurement (honor). In others the polar or terminal values are paramount, interior values are measured relative to the nearer pole, and the middle value is either nonexistent or unimportant (heat).

In fact this whole characterization of qualities' structural nature is only an approximation to facilitate machine reasoning. When binary qualities are inspected closely, one value can frequently be seen to head a subdomain with its own (different) character. Consider mobility. 'Immobile' things are all equally immobile but 'mobile' things move to a greater or lesser degrees; immobility has a static character, mobility a dynamic one. Consider color. In the feature system this domain has an enumerated top level and shallow hierarchic lower levels (blue-green). However its enumerated values can be located on the line based on the visible portion of the electromagnetic spectrum*. If this line is inspected closely discontinuities between neighboring values are filled by values from lower levels in the hierarchy and the spectrum becomes continuous.

Complex hierarchies cannot be represented in tridimensional space. Because the top level organization of such domains is a graph rather than a line, features are not restricted to association with

their immediate neighbors. Any two (or more) top-level features can combine to create new values on a lower level.

- however not all color values can be located on the spectrum and this representation ignores colors' essential incommensurability and hierarchic nature (ie prototypical color values are most readily identifiable).

18. Quirk & Greenbaum (1973) identify two classes of modifiers which act as qualifiers of measure: intensifiers heighten the degree of the following sense either absolutely (emphasizers: 'certain', 'real') or in terms of a norm (amplifiers: 'complete', 'perfect'); limiters particularize the following sense ('main', 'precise', 'only', 'same'). Intensification seems veridical but the members of the class of limiters seem too heterogeneous to constitute a single class.

19. Because the linguistic system cannot readily express unmediated many-to-many interrelationships an effort has been made to avoid needing to do this. In the system the paradigm case of multiplicity is the presence of a set of functionally equivalent alternative senses in an word's meaning description making it ambiguous. The situation to avoid is directly juxtaposing two ambiguous word meanings; facts* about the meanings' interrelationships would be lost.

In practice this case does not arise. Qualifying modifiers are directly concatenated onto the feature lists that compose other word's senses, but words like 'very' and 'possibly' are not ambiguous. Extending features are each given their own entry in a node's .relation category; in 'bright green apple', both 'bright' and 'green' are represented as separate relations. Recognition that the two adjectives interdepend to specify a color would occur in context of the noun§. All other relationships—entity and entity, entity and .act, .act and .act—are mediated through .cases which serve to constrain the possible interactions.

- ie, that the first meaning's first sense is consistent with the second meaning's second and third senses etc.

§ contrast with 'bright green student'. Its modifiers, -intelligent- and -inexperienced-, are independent rather than interdependent. Although these readings could only be applied to a .human object, it is theoretically possible that a human being could be colored green, and commonsense knowledge is required for disambiguation.

20. The subject is positional in active sentences, prepositional in passive ones. The indirect object (an ambiguous case—to write a letter to someone is not the same as writing a letter for someone) can be either positional or prepositional, but precedes the direct object if positional. The direct object is always positional and precedes the verb in a passive sentence.

21. The notion of selection restriction is almost but not quite correct in this circumstance. The term accentuates the function rather than the character of the features attached to a role; in the linguistic system, features attached to a role are an integral part of the relationship's own sense and not simply conditions to be satisfied before the role is occupied. In fact their function of ensuring that the correct type of .thing in

installed in a role is subordinate to the way in which they extend the relationship's sense. All features in a relationship's sense must be consistent with its role tenants, those inherent in the relationship's own character as well as those attached to its peripheral roles.

The term is also less than fully descriptive. Although the selection restriction attached to the agent of 'kick' could be completely satisfied by specifying an .animal .subject, doing so would completely ignore the unarguably significant sense of .instrument .leg. Moreover a requirement that a .leg be the .instrument forces the .subject to be or to have a .leg: because a .leg is a .part, the relevant relationship is meronymy, and things that have .legs are .animals. Therefore the .subject of 'kick' must be an .animal. Tracing processing in this way shows how far we must move from the peripheral sense to reach the actual restriction on the tenant's .type.

22. Specification, qualification, modification—all characterize aspects of this sort of relation. The name .specification has been chosen instead of qualification because this particular system of semantic features has been designed for language applications rather than for general cognitive purposes. There are differences between these two environments. Perceptual processes find individuals' attributes to be self-evident; however most of an individual's attributes remain unexpressed in language. When one is mentioned, the important thing is not its relationship of qualification but rather the fact that its value has been specified.

23. It is difficult to unambiguously identify a particular individual. Names label classes rather than individuals, and although many proper names label classes with a single known member, there is always a possibility that other members may exist. So naming an ostrich Clyde identifies the class of ostriches-named-Clyde and does not unarguably specify a particular individual. In cases where more than one individual has a particular name, context may aid in selecting the proper party.

A more difficult case arises with artifacts which are mass produced and therefore both numerous and almost identical. Speakers usually end up by designating ownership —'John's silver 1988 Mazda RX7'—or accident—'the stainless steel steak knife with the scratch on its handle'—to unambiguously designate an individual.

24. The first context is straightforward. It includes cases where something is inside or outside another thing, near or far from it, or moving towards or away from it. Although we usually assume horizontal orientation, nothing in the sense of this relationship requires it.

The second and third contexts are more complex. Consider the statement "the man stood in front of the car". With no additional information we could draw a cartoon showing the relative positions of the man and the car, and the car's orientation relative to the man (but not the man's orientation). We would draw the

same cartoon regardless of where we stood. However the relationship loses its universality if we substitute 'tree' for 'car'. Imagine the trunk of the tree is marked in some way so that we can distinguish its different sides. If we stand opposite the first side and draw a cartoon, it will show the man between us and the tree; that is, nearest to side A. Should we walk 90 degrees clockwise around the tree to face side B and draw again, our new drawing will place the man nearest side B.

We can draw some conclusions from this example. Many natural kind things have an shape attribute, call it 'orientation' (unimplemented), by which we identify their front, sides and rear, and their top and bottom. A car has an orientation in both horizontal and vertical planes while a tree has only horizontal orientation. A thing's orientation is based on its axes (generally axes of symmetry). A cube has axes which we can use to distinguish radical from orthogonal position while a sphere has neither orientation nor axes.

"The man is in front of the tree" does not provide enough information to permit us determine its spatial relations, and by introducing ourselves into the picture we can provide the additional specification required. It seems that if we cannot understand statements in their own terms we try to understand them in terms of ourselves. If this fails, we postpone further attempts until we have more information or the issue is no longer of interest (ie we change the focus of our attention).

25. It is unclear where to place human beings in the hierarchy. Considered objectively man is just another member of the large class of mammals. Considered subjectively human consciousness alone gives existence to immaterial things—all the ideas, thoughts and notions that are imaginable and that compose our mental life. This categorical oddity is resolved by entering human beings in the classification twice, once as a kind of .mammal and a second time as the unidentified subject of the .immaterial subdomain named .person*. Adding this subdomain has the practical effect of broadening and shortening the tree.

* the alternative would be to consider mental abstractions such as time and space to be minor phenomena of less importance than .fish.

26. The inadequacy of celestial terms in the language of earthbound people provides an opportunity to show how names are selected for features. The question is what to name individual non-stellar spatial objects. Properly speaking, a body circling a star is a satellite, but that word has come to be more commonly understood to mean a machine circling Earth for communication or spying. 'Planetoid' is too obscure. 'Planet' and 'world' mean much the same thing; in a choice between them, 'planet' was selected to be the superordinate because of its more extensive celestial connotations, while 'world' suggests a more intensive habitability.

27. Roget's Thesaurus (1933/1979) classified affections as personal (passive, discriminative, prospective,

contemplative, extrinsic), sympathetic (social, diffusive, special, retrospective), moral (obligations, sentiments, conditions, practice) or religious (sentiment).

28. The serviceability of this view can be seen in its capacity to explain the moment in the course of understanding a sentence when a word is recognized as denoting a semantic relationship but before the things to perform that relationship (occupants of its roles) have been identified. An example is "While walking down the street he saw a hat he wanted to buy". There is a point in our left-to-right reading when we have read 'walking' but not its pronoun subject.

29. Exactly what cases are is unclear. Fillmore (1968) originally defined them as 'semantically relevant syntactic relations'. Bruce (1975) emphasized their semantic side when he called a case as 'a relation which is "important" for an event in the context in which it is described'. Case grammar has also been interpreted as a way of treating a sentence as an assertion whose predicate is denoted by the verb and whose arguments are denoted by the noun phrases (Hirst, 1987). Some researchers see it as simply a notational convenience (Charniak, 1975 in Feigenbaum, 1981) or as a suspect or even inimical concept (Mellema, 1974 in Kunze, 1987).

Lacking a clear criterion for what constitutes case, case grammars seem to have gradually grown in size and scope to usurp some of the functions of other parts of syntax*. Traditional grammar identified only a few cases[§] and Fillmore's 1968 theory postulated a small number of them. However as time passed Fillmore himself identified additional cases[‡]. As more cases proliferated the issue of organizing them arose. Conceptual Dependency theory groups 16 cases into three classes on the basis of the type of tenant they accept; it also recognizes pairings like 'source' and 'destination' (Schank, 1975 in Winograd, 1983). Grimes (1975 in Winograd, 1983) organized 13 cases into 4 groups in an attempt to discover a smaller and more abstract defining set. To date however no classification scheme has been generally accepted; when Sparck Jones and Boguraev (1987) recently presented 28 cases realizable in prepositional phrases[¶], they simply listed them. The net result of these tendencies appears in the index of *Language as a Cognitive Process* (Winograd, 1983) which has entries for each uniquely-named case mentioned in the systems it discusses; there are 61 entries and many are almost identical.

Certain facts about case are unequivocal. Surface case, the purely positional 'subject' and 'direct object' of traditional grammar, is not the same as deep case, the composition of semantic elements and roles. Surface cases are only one of the ways deep cases can be realized in syntax; prepositional phrases are more common. For example the deep case structure { AGENT we } { park } { OBJECT truck } { MANNER easily } can be expressed in different syntax by the sentences "We parked the truck easily", "The truck was parked by us easily" and "The truck parked easily".

Prepositional phrases have a great impact on case assignment. Their presence or absence can change the way surface cases are construed, ie "He wrote his friend", "He wrote to his friend", "He wrote about his friend". Sometimes verbs require prepositions to invoke particular senses, ie 'chance on', 'see off' and 'account for', making these phrases subidioms (Appendix 3).

- thus postulating a case of Accompaniment to express one sense of 'with' creates the problem of determining how to parse syncategoric paraphrases like "She accompanied him" from "She went with him". Should the verb's action be understood to be a case relation?
- § Quirk and Greenbaum (1973) suggest four cases: agentive, affected, recipient and attribute.
- ‡ eight or nine at different times in 1971.
- ¶ part of the reason there are so many cases is because the authors consider that the same role defines different cases if it is used in different spheres.

30. 'Across' is a subjective relationship (what's across from you may be onside to me) based on the point of view of an implied case of observer held by the utterance writer or reader. It also has cases for an object which is an intermediate thing, and for a subject which is a thing beyond the object. In 'the man across the street', 'man' is subject, 'street' is object, and the implied observer from whose point of view this spatial relationship is true is the participant.

'Between' is an objective relationship (what lies between two things is between them from any perspective barring occlusion) which has a subject that is an intermediate thing and an object which must be binary either by nature ('us') or by explicit conjunction ('John and Mary').

31. Gender provides us with an opportunity to determine how to handle inapplicable or unspecified domains in word descriptions. 'Man' and 'woman', 'he' and 'she' are the straightforward cases; their descriptions simply specify the appropriate gender feature. 'It' is less obvious. The word has little sense other than its gender and that gender is neither male or female. Although adding .neuter to the domain is one solution to the problem, a better one is to acknowledge the aspect of gender in 'it' by using the domain name in the description but not follow it with a value.

This problem does not arise with words that have senses other than the aspect which is unspecified in its domain; unless it is salient they do not use the domain in their descriptions.

32. An argument could be made for subordinating .degree to the .single value of .count. A thing whose degree is being measured is considered to be a single individual for the purposes of measurement even if it is plural in other contexts. "He ate most of the eight pieces of pizza" suggests that he ate perhaps six pieces (interpreting 'most' as 75%)*. A similar idea underlies the concept of percent. Percent increases our capacity to exactly describe proportions of things that are difficult to number by consolidating them into unities which are then divided into 100 equal parts.

- an alternate and nonintuitive interpretation would be that he ate some part of all eight pieces.

33. Flanagan (1988) describes a five-valued logic or measure that allows rigorous evaluation of certain proportional statements. If none, some, half, most and all match the 0%, 50% and 100% points and the intervening ranges, the intersection of most and most is non-empty. As this and the domain defined in the feature system suggest, an odd-numbered set of measures seems superior to an even-numbered set because of its capability of representing the middle.

34. Because periphrastic utterances reduce to a single representation in the linguistic system (*a la* Chomskian deep structure) and also because in practice the representation's grain must be finite, the set of possible meaning representations is smaller than the set of possible natural language utterances. The unboundedness of feature composition ensures that the set of meaning representations remains infinite, however.

35. It can happen that a new statement is inconsistent with all of a preexisting symbol's senses because some of them have already been eliminated by disambiguation. If routine processing went ahead the symbol would be left without any sense whatsoever. This can never be permitted if discourse is understood to be meaningful. Such a situation could cause words previously subsumed in the symbol to be reprocessed in the context of the new word, with the result that a previously eliminated reading would now be preserved because of its consistency with the new senses. Alternatively, routine processing might be suspended and both sets of readings (the current and the new) might be stored in the symbol for future processes to resolve if necessary.

36. "The man I saw" is an example of a sentence in which a subordinate clause begins without an introducing word; it is as acceptable as "the man whom I saw". Sometimes an absent introducing word can be inferred from the structure of the subordinate clause. Only modifying relative clauses can be constructed with implicit links; 'when' cannot be dropped from "when we come he will talk to us".

37. If subsequent processing consolidates similar statements into sets, it is important that the exact sense of the separate expressions be retained; when formal relationships are involved this is not straightforward. Although "John saw Bill and John saw Jane" is equivalent to "John saw Bill and Jane" in some circumstances, "John wrote Bill a letter and John wrote Jane a letter" is never equivalent to "John wrote Bill and Jane a letter".

38. Series introduced by 'either' and 'whether' are syntactically disjoint. 'Either' asserts that one set actually member exists ("I think either John or Mary came"); 'whether' is equivalent to 'if either' and

indicates that a set member's existence is uncertain or irrelevant ("Find out whether John or Mary came"; "I'm going whether he goes or not").

The only complication I am aware of arises from unspecified temporal relations. In "The man who had seen me left the room. He went out of the building", 'left' would point to 'go' as its successor. 'Seen', known to be temporally prior to 'left' in their indirect relations through 'man' because of its past perfective sense, would not appear in the History list. Changing the tense to 'has seen' or 'sees' makes this priority unclear*, and there may be a need to add a general rule to the grammar about tense interpretation in clauses.

- * ie, "The man who has seen me left the room." and "The man who sees me left the room." The latter invokes a different, nonspecific sense of 'see', ie see-on-a-regular-basis.

39. 'S' poses special problems because it can stand for either an 'is'-contraction or the possessive case. The program looks in SENTENCE for a verb to precede a noun; if this occurs, contraction ("John's very young to go camping") is more probable than possession ("John's very young brother is ill"). Handling the possessive case immediately above would require 'of John' replace 'John's' and the phrase be moved nonlocally to produce 'the very young brother of John'. This should occur in the parser and not in the morphological translator and has not been implemented.

40. Although they will not be directly addressed here, Winograd (1983) listed issues about which a parser-designer must make decisions. Some of these are general: the degree of completeness of the parse, the content and structure of the output, the comprehensiveness of grammatic coverage and the scope of the target domain of discourse. Once the size and shape of the parser 'black box' is fixed, the designer must specify how he will fill it: pick a theoretical framework within which to work, make decisions about processing tactics*, and determine the particulars and functionalities of the actual implementation.

- * such as searching (parallel processing, backtracking, determinism) and handling ambiguity (forbid it, use the first sense found, allow many senses), the direction of processing (top-down, bottom-up) and its method of advance (serial, largest chunk).

41. Given grammatic input, an analytic grammar is theoretically syntactically simpler than a generative one. Analysis involves recognizing which of a large but finite number of syntactic templates matches a particular word phrase. Generation must select one of the theoretically unbounded and practically very large set of alternative realizations for each suitable template as the most appropriate expression of a semantic structure. This is part of the reason that the pause for the right word is usually longer than any hesitation encountered in apprehending text. Alternatively, because grammatic input is already well-formed, the issue in analysis reduces to identifying which good form is in use. Concepts which we wish to utter are not

equivalently constrained to good formation so this must occur before the counterpart task of picking the most suitable form is begun. For instance, I write these words and then edit them so that they say exactly what I mean; this is generation and it has no analytic counterpart.

42. At present only three heuristic rules have been coded. These require: that a sentence have at least one noun phrase and one verb phrase before it is ended; suggest that it is unlikely that the next phrase is a noun phrase if one has just been processed and no verb phrase has been processed yet; suggest that the next phrase is likely to be a verb phrase if none has been processed yet. Only the sentence-ending rule is implemented in the parser; the other two existing rules have not yet been used because defining two priority levels for state processing—an optimal one and a regular one—would require rewriting the code for almost all states. It is envisioned that Suitable? would be used to control the order of execution of a state's transition blocks in a dynamic way. If the most suitable transition failed and no other was distinguishable, processing would proceed as present in the order fixed by the sequence of transition blocks within states.

43. Consider the phrase 'hard ball'. Both words are ambiguous; 'hard' means both -rigid-surfaced- and -difficult- and 'ball' means both -spherical-object- and -social-event-. This phrase can be properly understood not only in the obvious -rigid-surfaced- -spherical-object- way but also as -difficult- -social-event-. These associations can be made correctly because both senses in each word differ in .class; a -spherical-object- is .material and a -social-event- is .immaterial, and -rigid-surfaced- implies a .material object and -difficult- implies an .act (nominalized) which must be .immaterial.

44. Because modifiers are routinely attached to the next following noun when possible, autonomous modifier phrases are only likely to occur just before the verb or at the end of the sentence. Semantically equivalent noun phrase complements could be represented in the same way as modifier phrases are (ie "the boy is a youth" converted to "the boy is young") Modifier phrases are not expected to contain implied senses for their subjects.

45. Attaching a prepositional phrase to its .subject enriches its semantic context with the result that previously indeterminable senses in its .object may now be able to be disambiguated. This also suggests (and the pseudocode shows) that consistency between .subject and .object across the preposition is an integral part of making an attachment determination.

46. Word sense descriptions can be generated for contracted, inflected and synonymous forms of the words entered in the dictionary. Synonyms can be exact or near. Word production shows how differences in a word's morphology can cause adjustments in its meaning.

47. **Dell** destructively eliminates any element from a list (Delete won't remove the first item); **Bound?** returns its symbol argument if it is bound in an environment; **Push**, **Pop** and **Peek** implement stack operations; **Wr** writes its arguments to output without an terminal linefeed and returns them (making it transparent in code); **String->Integer** performs a missing type coercion; **Set!** evaluates its arguments; **Twr** and **Tset!** return guaranteed true results (**Write** always returns nil and **Set!** does so when an environment is specified); **Call/CC** converts **Call/CC** to MIT or lambdaless style.

48. **Map** applies an function to a list. The first stage of refinement allows the single list to be replaced by an indefinite number of lists used cross-serially, for nil results to be expunged from the result list and for that list to be flat rather than nested list.

The second stage of refinement (implemented here) tailors mapping functions to be used on particular data structures. Options can be illustrated by decomposing function names from front to back. The digit following **'Map-'** specifies on what level of its argument list(s) the mapper will operate; an optional **'-can-'** followed by a digit indicates that the results will have that many less levels than otherwise; **'-*'** means that null results are eliminated; and a final digit shows how many lists the mapped function expects. Restricting list levels (ie digits) to four identifies a possible 256 unique functions.

49. The text could have any particular content, popular or specialist, so long as it is used literally and all of its words are defined in the dictionary. The resulting representation is completely deterministic; the frequencies with which particular word senses occurred absolutely and in particular conjunctions would be established by the texts chosen. Thus no two dictionary representations would be identical unless they were developed by presenting the same texts to the program in the same order (order is important because selection between alternative senses is based on past experience and in many cases the selection of a particular sense will never be contradicted by any subsequent information). This situation makes possible the creation of linguistic knowledge bases that have known points of view and predispositions to interpretation, ie biases.

50. Assume that a fully-developed apprehender program has been implemented and run on a large volume of text ⁴⁹ in an environment unconstrained by resource limitations. Millions of words have been input, perhaps as many as most people read during their entire elementary schooling. The senses of each of these words have been retrieved from the dictionary, disambiguated as necessary, combined with the senses of related words and stored in one network episode after another. These episode records constitute the substance of the meaning of that volume of text. Leaving aside the details of how the required gigabytes of secondary

storage would be organized and accessed, assume that no episode is no more than a few dozen microseconds away from recollection to main memory.

51. Although the representation could be accessed through its episodic rather than its lexical structure, no access path is likely to be more serviceable than the dictionary. The representation stores semantic information. Its contents are not sought on sequential, random or arbitrary keys but on meanings and overt meanings are structured into words. Demonstrate the human counterpart of this to yourself by trying to recall all the words that end with 't' and have 'm' as the third last letter, or the word which follows 'attempt' in your personal mental dictionary. Answers are not forthcoming; the search criteria, especially the latter, are nonsensical.

52. Because prepositions, conjunctions, articles, adjectives and adverbs are all types of relationships, their senses are combined with those of the things or acts which they relate and are not represented by symbols in the knowledge base. Pronouns' symbolic existence ends when they are replaced by their referents.

As a result there are no symbol lists attached to the dictionary descriptions of these parts of speech. This means that instances of the adjective 'green' could not be recalled (although the noun 'green' could be). To find instances of green things, one would have to inspect symbols of things using some search strategy (recency, frequency, locality). This seems congruent with human psychology.

53. Different aspects of the abstract nature of relations have been identified by people working independently in various fields and seven have some degree of pertinence to semantics. Philosophy supplies notions of transitivity, reflexivity, direction and connection (Lacey, 1986). Set theory adds associativity and cardinality (Lipschutz, 1964). It is axiomatic to say that every relation has a unique character or identity, while relation strength is a final aspect which may be peculiar to semantics.

transitivity: a relation is transitive when its effect extends beyond its immediate tuple of tenants to include tenants in adjoining relations of the same character. Transitive relations are necessarily non-exclusive. The inverse is intransitivity or non-transitivity. -Greater-than- is a transitive relation while -father-of- is intransitive.

reflexivity: reflexive relations apply to their sole tenant. They are necessarily unary or monadic. The inverse form is non-reflexivity, relations which cannot apply to their tenant and therefore require more than one tenant. Irreflexive relations straddle the distinction and may operate in either fashion. -Identity- is a reflexive relation, -greater-than- is non-reflexive and -fond-of- is irreflexive.

direction: a directed relation has an orientation and is experienced differently depending on which

of its roles is held. A undirected or symmetric relation has no orientation and is experienced in the same way by any of its tenants. -Father-of- is a directed relation, -sibling-of- is undirected.

connection: a connected relation is pervasive and holds between any two individuals in the universe of discourse; all individuals are tenants of a connected relation. -Earlier-than- is connected in the domain of time.

associativity: associativity characterizes the independence of a relation's effect. Associating relations can be applied to their tenants in any order without affecting the final outcome, while non-associating relations are sensitive to the order of their application. -Size- and -shape- are associating relations when applied to material things while state-of-matter and -shape- are not.

cardinality: a relation's cardinality is the number of tenants it may require or be limited to. -Identity- is unary or monadic, -married-to- is binary or dyadic, and -between- is ternary or triadic.

strength: relations in formal logic are always strict and conclusions necessarily follow from their premises. Many real world relations are inherently probabilistic and therefore less certain; such weak forms of implication have been studied only recently (Zaden, 1983). A relation can be therefore be classified as necessary or accidental.

Necessity and impossibility are inverses which express strict contingency, the capacity of one individual to require or rule out the co-occurrence of another. A 'man' necessarily 'lives' but a 'thought' cannot possibly 'live'. Contrast this with the accidental way in which a 'man' may be 'hirsute' or 'thin'. The strength of an accidental relation can vary; more men have hair than are thin. An individual's most likely accidental relations are its typical values.

54. Dissimilarity is a correlative dimension (unimplemented); the degree to which an inconsistent pair of senses differ could also be compared or measured. Strength (unimplemented) is an absolute measure of agreement between two senses; agreement in 6 out of 8 features is stronger than agreement in 3 out of 4, even though the proportion and therefore the similarity are identical. Such precise measurement of an approximate system may be unnecessary however.

55. It is difficult to talk clearly and exactly about the human mind because our experience of it is subjective and non-verbal. For this reason it is useful to envision the qualitative contents of our minds as areas in tridimensional SEMANTIC SPACE. According to this metaphor each aspect occupies a small area of semantic space, and together all the aspects we know shape and tile our particular semantic space. Characters are assemblages of areas in semantic space. Features, which express whole aspects or parts of them, range downward in size from areas congruent with aspects to dimensionless points within an aspect's area.

To carry the spatial metaphor further, an aspect's unique nature could be shown by the color of its area. If further refinement of the system causes an aspect to be subdivided into two, each half would retain the basic color, tinted slightly to distinguish between it and its neighbors. The idea of semantic space is akin to that underlying Berlin & Kay's diagramming of color terms (1969 in Leech, 1981).

If mental constructs are envisioned to be volumes defined by the areas of their component aspects, then a 3-dimensional space cannot adequately express how such volumes would be formed. In 3-dimensional space the areas of component aspects would need to be translated to bring them into contact with each other and there is no guarantee that they would enclose a volume. In N-dimensional space, where N is the number of domains required to express the character of the construct and can vary, specifying each aspect would select an area in the associated dimension and would be sufficient to enclose a volume. (By convention dimensions are defined to have extent at right angles to one another, ie orthogonally, in terms of a global coordinate frame).

56. These two relationships are the only ones on which real-world hierarchies are commonly based. Although institutional staffs and document formats are exceptions, most other relationships are unsuited to representation by a hierarchy. They either lack transitivity, are too strongly or too weakly connected, or the hierarchical structure seems inappropriate to their nature.

Take causal relationships as an example. They are directed and transitive and it is easy to construct a graph of the forward flow of effects. However a single effect frequently has multiple causes and vice versa, so the representation must be a true graph rather than a tree. Temporal relationships are also directed and transitive, but the hierarchic structure is simply unsuited to represent priority and posteriority. Classing all events which precede an individual as it's members produces a linear structure rather than a hierarchic one (except for cases when events occurred simultaneously).

57. Terms in a system are often defined in terms of other terms within the system. This recursive character has two outcomes. The sequence of definitions may eventually become circular in which case meanings ultimately refer to themselves. Systems of this sort are called self-contained or hermetic, and natural language is an example of one; all words in a dictionary are defined in terms of other words. Alternatively, a term may appeal outside the system for meaning. Such appeals are made to the external system within which the given system exists. Because language is the preeminent external system that people share, objectively-understood systems from formal logic to baseball are usually based on axioms or terms which are defined linguistically.

This leads to a conundrum: language is an hermetic system but it is also shared. How is consistency between different individuals achieved? The external system within which language itself exists

is a person's sensory experience. The most noteworthy thing about sensory experience is that it is the ultimate frame of reference. A person can never escape from or exceed his sensations. However, sensory experience is also an hermetic system and that appears to magnify the problem; not only can a person never escape his sensations, he cannot share them.

Resolution lies in noting that language is initially acquired by diexis. Before we learn words' dictionary definitions, we are taught to associate them with—map them onto—certain sensory experiences shared by ourselves and our language teacher. In this way we are bootstrapped into an initial language capability which we use both to learn new words through language itself, and to correct our understanding of word meanings so that they are in harmony with objective definitions.

58. Not all emotions are expressed in paired verbs. Consider "He charmed me" and "I was charmed by him". The more intense role is the response to the effort^{*}; surprisingly it is cast in the passive. Contrast this with "He hated me" and "I was hated by him" where the more intense role is active. The main point here is that neither 'charm' nor 'hate' has a commonly-used counterpart to express its other side.

'Drip' is akin to 'charm'. An uninformed examination of "The faucet dripped" would suggest that the faucet did something when in fact the action portrayed is the falling of (unmentioned) drops of liquid from its lip.

* being charming can be effortless or inadvertent.

59. Idioms tend towards one of two poles on a scale of literal plausibility. "To play ball" describes something that happens every day, while "to throw one's hat in the ring" must be a relatively rare occurrence in real life. Although idioms are typically noun or verb clause constructions, a number are complete sentences: "It's not over till the fat lady sings". An idiom can always be paraphrased into a literal expression which says the same thing. Words which remain ambiguous after all attempts at resolution may be intentional, ie double entendres. It is difficult to know how to select between literal and idiomatic senses in a given case. Since there are fewer idiomatic constructions and because many idioms are rarely used literally, it is reasonable to initially attempt to construe input idiomatically. This requires that the tokenizing process try to match as many words of the input to dictionary entry labels as possible. Idioms can have embedded role variables^{*} which a pattern-matcher would have to recognize. Once identified, an idiom would be processed as if it were its literal synonym. In the event of failure to parse, it would be necessary to backtrack into the tokenizing function and pick a shorter string of words (processing would be time-consuming).

* a Canadian might 'throw his toque in the ring'.

60. The diversity of perspectives on NLU is reflected in the different ways that observers have classified

theories and systems when surveying the field. A variety of criteria have been used: historical development, processing strategy, type of representation, area of application etc. Some classifications are relatively encyclopedic and descriptive while others are more evaluative.

Although Winograd (1983) identified seven broad task categories into which he classified 52 different implemented systems, the primary criterion on which he categorized computation linguistic systems is their theoretical framework. Grosz, Sparck Jones and Webber (1986) provide a *de facto* classification by organizing 38 key readings in NLU into six sections, supplemented by prefatory remarks. The principle used seems to be effective partitioning, since the sections span different criteria*. Sparck Jones' earlier book co-edited with Wilks (Sparck Jones & Wilks, 1983) classifies theoretical frameworks for natural language understanding systems on the basis of their parsing techniques. Grishman (1986) organizes his introductory text in computational linguistics on a task basis, with three bifurcations each neatly producing four terminal classes. Hirst (1987) short classification of systems on their treatment of semantics is discussed elsewhere in this thesis.

* it should be noted that there is a companion volume covering knowledge representation readings. In it Brachman and Levesque (1985) classify 25 papers into six sections, also mixing their criteria.

61. Creary and Pollard (1985) posit the following real world categories: 1) individuals, which are things and situations, which in turn are states and events; 2) relations, which include properties, have associated roles of varying arity, and may be sortal; 3) concepts, which either subjective and objective and may be propositions. Communities of organisms communicate using a common subsystem called a grammar.

Cruse (1986) identifies hyponymy and meronymy as the two primary criteria for categorical organizations or taxonomies. In addition to the respectively corresponding membership and aggregation, Potter & Trueblood (1988) identify five other relations holding in hypersemantic data models composed of objects, operations and relationships: generalization, where similar object types are abstracted into a higher-level object type; classification, where object instances receive their type's characteristics; constraint, where objects, operations or relations are constrained in some fashion; heuristic, where inferences are derivable from existing instances; temporal, where object types are synchronous or asynchronous.

62. Concepts are primarily based on visual perceptions; visualization is almost a synonym for conceptualization. Although sight is more voluntary than our other senses (closing our eyes closes the visual channel), we receive by far the greatest amount of sensory information through this mode.

A practical way of showing the primacy of vision is to try to exhaustively describe* some simple object in words. Verbal description of a human hand can begin by listing its parts: fingers (little, second,

middle, index, thumb), palm, back, wrist. Repeated decomposition eventually reduces most⁶ of these items to simple solids. Each solid can then be indexed in some way to show its physical connections and typical measurements can be specified for each dimension. However this leaves the spatial relationships between parts unspecified; it does not show that the fingers of an open hand lie parallel to one another in a plane and in line with the longest axis of the palm, that the wrist lies at the opposite end of the palm from the fingers, and that the arc of motion of the thumb is at right angles to those of the fingers.

Even if the hand's description is somehow extended to include these configuration particulars, it has not yet touched on the color, texture and malleability of the skin or the distribution and details of the hair. Complete that, and a hand retains an indefinite number of undescribed local features: humps over knuckles and linear indentations, a pattern of wrinkles, junctures of fingernail and cuticle, surface deformations and discolorations around the tendons and veins etc.

The whole process is rendered useless as soon as one notices that words impose a granularity on a description that images do not. To achieve the effect of an image in words—if this can be done at all—would require a description of indefinite length. A picture appears to be worth considerably more than a thousand words.

- * describing something is a different task from the one which a dictionary sets itself. The first definitions for 'hand' in Webster's New Collegiate Dictionary refer to function rather than configuration or appearance. 'Description' itself has a strong visual component. The smell, taste or feel of a thing do not adequately convey it if its appearance is unspecified. The reverse is less true.
- § 'palm' and 'back' are not solids but surfaces, the front and back of the unnamed body part to which fingers are attached.

63. Recent research suggests that the traditional idea of a category's intension or set of criterial attributes unambiguously defining its extension or membership does not accord with psychological reality (Gardner, 1985). Attributes for most natural kind categories* appear to be universal rather than culturally arbitrary (Rosch, 1977), implying that categories themselves may be rooted in the natural world. Frequently attribute redundancy is a part of this universal nature, with feature occurrences correlating highly with one another rather than being orthogonal as criteriality generally stipulates. Attributes may inhere in a thing's relations (function, role) rather than itself and category intensions are also often mixed across relational and inherent attributes. Philosophers and anthropologists have proposed theories which parallel this view of categories (Gardner, 1985).

Rosch suggests that mental categories are organized about prototypes, are structured by levels of abstraction[§] and do not necessarily have rigid boundaries (ie unambiguous extension) (Rosch, 1978). Category formation proceeds in accordance with two principles: cognitive economy (most information for least effort) and veridicality (accordance with real world structure). Basic-level categories are the easiest to

form because they most completely mirror features in the world; abstract words[‡], both concretely based ones (Wittgenstein: 'game') and otherwise ('love', 'freedom'), are difficult to define even for competent language-users.

The analysis of cognitive psychologists is directed toward the ways in which individual categories are formed and not to the nature of the relations between categories in a taxonomy. Their discussions, especially the concept of category levels, seem to presume hyponymic organization.

- * some categories are truly arbitrary. These are learned either by rote memorization of the category's enumeration (the alphabet) or by mastering the generating formula (the number line) (Glass & Holyoak, 1986).
- ¶ basic ('chair'), superordinate ('furniture') and subordinate ('rocking chair'), with the basic level being the point of entry to the category.
- ‡ cognitive psychologists tend to treat a category and words naming it interchangeably in their discussions if not in their research.

64. There is some evidence that the psychological representations in which words are rooted are ultimately based on, and have accessible, concrete perceptual data. This accords with the hypothesis of a visuospatial code composed of subcodes for color, location and shape which integrates data from all senses (Glass & Holyoak, 1986). In his work on sentence verification, Clark (1976 in Moyne, 1985) assumes sufficient unity in the representations of knowledge from all sources that any two instances may be evaluated in each other's terms. Thus, 'fall' is meaningful because we have seen falling things and have on occasion fallen ourselves. Clark also recognizes that a word's perceptual basis may be indirect, as is the case with 'part' or 'thing'.

Communication is undertaken for some purpose, and the messages that compose it are organized to advance that purpose. This is Grice's Cooperative principle (Grice, 1975 in Moyne, 1985), and the Communicative and Literalness presumptions of Bach and Harnish (1975 in Moyne, 1985).

65. Most of the members of a power set on natural kind things have no semantic cogency. Some unnamed class might usefully superordain 'hand' and 'foot' but none is likely to be based on 'hand' and 'nose'. In a power set defined on features members without cogency are senses which do not directly or indirectly describe anything in the real world (an indirectly descriptive sense might be one that identifies a nonexistent thing worth imagining for some purpose).

66. Although they may differ substantially in particular cases and generally in detail, all users of a language share essentially the same set of words and meanings associated with them. This is the principle of Charity (Davidson, 1969 in Moyne, 1985). This is also Bach and Harnish's Linguistic presumption (Bach & Harnish, 1975 in Moyne, 1985). Bach and Harnish further posit that communicators share a set of mutual contextual beliefs which allows message initiators to determine what information must be

communicated to the recipient; ie both parties know what they are talking about.

67. Halliday's (1985) formulation provides a good starting point for a reductionist analysis of case; it is current and responds to the points raised by other authors. As the developer of the first systemic grammar, Halliday can also be considered authoritative; systemic and case grammar have a good deal in common, so much so that Winograd (1983) borrows Halliday's systemic grammar analysis and uses it as his treatment of case in an outline of English grammar.

Halliday proposes five disjoint classes of semantic process which account for 13 cases, and a sixth class of 8 more circumstantial cases which can co-occur with process cases and with each other. Phrases can hold more than one case role, ie reflexive actions. Table 6 lists the processes and the cases associated with them.

The process classes show a decided similarity. With the exception of the Equative class which expresses identity rather than action, the first case in each class is the active initiator of the process. Three process classes have only one additional case and in each that case plays the role of the object of the action, ie the thing with which the action is concerned. Verbalization in Verbal processes expresses the same relationship. Range and Medium are mutual exclusives which express the distinction between sentences

PROCESS	CENTRAL PARTICIPANTS (CASES)
Material	Agent, Medium, Range, Beneficiary
Mental	Cognizant, Phenomenon
Verbal	Sayer, Addressee, Verbalization
Attributive	Carrier, Attribute
Equative	Identified, Identifier
Circumstantial	Location, Extent, Motion, Cause, Manner
(all types)	Accompaniment, Purpose, Matter

TABLE 6. Winograd's Adaptation of Halliday's Case Roles

which can be reframed as "What he did to the X was Y it" and those which cannot. As the frame suggests, Range and Medium together express the object case in Material processes. Beneficiary in Material processes and Addressee in Verbal processes are the only cases remaining. These two cases are alike in being the subject's inverse in terms of the process operation. They are realized by the inaptly-named indirect object. I believe that all of these cases can be reduced to the deep equivalents of surface subject, direct object and

indirect object. Particular differences between case counterparts are attributable to the differences in the nature of the process and not inherent in the cases themselves. A reduction of Halliday's 13 process cases to the deep equivalents of surface subject, direct object and indirect object should not be entirely surprising. Those three cases are the only ones which can be realized purely by their position in the sentence; every other case requires a prepositional relator. This is not coincidental but a reflection of the special status these three cases have in sentence semantics.

References

- Allen, J.F. & H.A. Kautz. (1985). A model of naive temporal reasoning. In J.R. Hobbs & R.C. Moore (Eds.), *Formal theories of the commonsense world*, 251-268. Norwood NJ: Ablex.
- Alshawi, H. (1987). *Memory and context for language interpretation*. Cambridge UK: Cambridge University Press.
- Austin, J.L. (1962). *How to do things with words*. Cambridge MA: Harvard University Press.
- Bach, E. & R. Harnish. (1979). *Linguistic communication and speech acts*. Cambridge MA: MIT Press.
- Barr, A. & E.A. Feigenbaum. (1981). *The handbook of artificial intelligence*. Los Altos CA: William Kaufmann.
- Berlin, B. & P. Kay. (1969). *Basic color terms*. Berkeley CA: University of California Press.
- Bobrow, D.G. & T. Winograd. (1977). An overview of KRL. *Cognitive Science* 1(1), 3-46.
- Brachman, R.J. & H.J. Levesque. (1985). *Readings in knowledge representation*. Los Altos CA: William Kaufmann.
- Brittain, S. (1987). Understanding natural languages. *AI Expert*, 3(7), 31-38.
- Bruce, B. (1975). Case systems for natural language. *Artificial Intelligence* 6, 327-360. Catholic University of Leuven, Department of Linguistics, Section of Applied Linguistics.
- (1981). L.K. Engels, B. Van Beekhoven, Th. Leenders, I. Brasseur (Eds). *Leuven English teaching vocabulary list*. Leuven Belgium: Acco.
- Carroll, J.B, P. Davies & B. Richman. (1971). *Word frequency book*. New York: American Heritage.
- Chapnick, P. (1988). The concreteness of speech. *AI Expert*, 3(7), 5-6.
- Chapnick, P. (1988b). The map is not the territory. *AI Expert*, 3(8), 5-6.
- Charniak, E. (1977). A framed PAINTING: The representation of a common sense knowledge fragment. *Cognitive Science* 1, 355-394.
- Charniak, E. (1975). A brief on case (Report 22). Castagnola Switzerland: Institute for Semantic and Cognitive Studies.

- Cherniak, C. (1985). Undebuggability and cognitive science. *Communications of the Association for Computational Machinery*, 31(4), 402-412.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge MA: MIT Press.
- Chomsky, N. (1975). *Reflections on language*. New York: Pantheon Press.
- Clark, H.H. (1976). Inferences in comprehension. In D. Laberge & S. Samuels (Eds.), *Perception and comprehension* (pp. Erlbaum Associates, Hillsdale, N.J.).
- Cohen, P.R., S. Fertig & K. Starr. (1982) Dependencies of discourse structure on the modality of communication: Telephone vs. teletype. 20th Proceedings of the Association for Computational Linguistics, 28-35.
- Creary, L.G., C.J. Pollard. (1985). A computational semantics for natural language. 23rd Proceedings of the Association for Computational Linguistics. 172-179.
- Cruse, D.A. (1986). *Lexical semantics*. Cambridge UK: Cambridge University Press.
- Dahl, H. (1979). *Word frequencies of spoken American English*. Essex CT: Verbatim.
- Davidson, D. (1969). On saying that. In D. Davidson & J. Hintikka (Eds.), *Words and objects* (pp. 158-174). New York: Humanities Press.
- Dreyfus, H.L. (1985). From micro-worlds to knowledge representation: AI at an impasse. In R.J. Brachman & H.J. Levesque (Eds.), *Readings in Knowledge Representation*, (pp. 71-94). Los Altos CA: Morgan Kaufmann.
- Fillmore, C. (1968). The case for case. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 1-88). New York: Holt, Rinehard & Winston.
- Flanagan, T. (1988). Fiat lux: A reply to McDermott's A critique of pure reason. *Computational Intelligence*, 4(4), 404-414.
- Fodor, J. (1975). *The meaning of language*. New York: Crowell.
- Gardner, H. (1985). *The mind's new science*. New York: Basic Books.
- Glass, A.L., & K.J. Holyoak. (1986). *Cognition* (2nd ed). New York: Random House.
- Goebel, R. (1988). Exhuming the criticism of the logicist. *Computational Intelligence*, 4(4), 401-404.
- Goldfarb, W. (1955). Emotional and intellectual consequences of psychologic deprivation in infancy: A re-evaluation. In P.H. Hock & J. Zubin (Eds.), *Psychopathology of Childhood*, (pp. 105-119). New York: Grune & Stratton.
- Graesser, A.C., & S.M. Goodman. (1985). Implicit knowledge, question answering and the representation of expository text. In B.K. Britton & J.B. Black (Eds.), *Understanding Expository Text*, (pp. 109-172). Hillsdale NJ: Lawrence Erlbaum.

- Grice, H. (1975). *Logic and conversation*. In P. Cole & J. Morgan (Eds.), *Speech acts: Vol. 3. Syntax and semantics* (pp 41-58). New York: Seminar Press.
- Grimes, J. (1975). *The thread of discourse*. The Hague: Mouton.
- Grishman, R. (1986). *Computational linguistics*. Cambridge UK: Cambridge University Press.
- Grosz, B.J., K. Sparck Jones & B.L. Webber. (1986). *Readings in natural language processing*. Los Altos: Morgan Kaufmann.
- Halliday, M.A.K. (1985). *An introduction to functional grammar*. London: Edward Arnold.
- Hirst, G. (1987). *Semantic interpretation and the resolution of ambiguity*. Cambridge UK: Cambridge University Press.
- Hobbs, J., W. Croft, T. Davies, D. Edwards & K. Lewis. (1987). *Computational Linguistics*, 14(3-4), 241-250.
- Horn, E. (1926). *A basic writing vocabulary--10,000 words most commonly used in writing*. Ames IA: University of Iowa.
- Jensen, K., J.-L. Binot. (1987). Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics*, 13(3-4), 251-260.
- Johnson, G. (1986). *Machinery of the mind*. Redmond CA: Microsoft Press.
- Katz, J. (1972). *Semantic theory*. New York: Harper & Row.
- King, M. (1987). On the proper place of semantics in machine translation. In M. Nagao (Ed.), *Language and artificial intelligence*. Amsterdam: North-Holland.
- King, M. (1988 February). *Computational linguistics: Issues and solutions*. In A. Blaser (Ed.), *Natural language and the computer. Proceedings of the scientific symposium on syntax and semantics for text processing and man-machine-communication, Heidelberg*, (pp. 9-30). Berlin: Springer-Verlag.
- Klappholtz, A.D. & A.D. Lockman. (1977). *The use of dynamically extracted context for anaphoric reference resolution*. Unpublished manuscript. New York: Columbia University, Dept. of Electrical Engineering and Computer Science.
- Kunze, J. (1987). Some remarks on case. 3rd Conference of the European Chapter of the Association for Computational Linguistics, 302-305.
- Lacey, A.R. (1986). *A dictionary of philosophy*. London: Routledge & Keegan Paul.
- Laffal, J. (1973) *A concept dictionary of English*. New York: John Wiley & Sons.
- Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago IL: University of Chicago Press.

- Leech, G. (1981). *Semantics* (2nd ed.). New York: Penguin Books.
- Lehrer, A. (1974). North-Holland linguistic series: Vol 1. *Semantic fields and lexical structure*. Amsterdam: North-Holland.
- Lewis, D. (1972). *General semantics*. In D. Davidson and G. Harman (Eds.), *Semantics of Natural Language*, (pp. 169-218). Dordrecht: Reidel.
- Lipschutz, S. (1964). *Schaum's outline of theory and problems of set theory*. New York: McGraw-Hill.
- Lytinen, S.L. (1984). *The organization of knowledge in a multi-lingual, integrated parser* (Rsrch. Rep. No. 340). New Haven CT: Yale University, Department of Computer Science.
- Marcus, M.P. (1980). *A theory of syntactic recognition for natural language*. Cambridge MA: MIT Press.
- Mellema, P. (1974). A brief against case grammar. In *Foundations of Language* 11, 39-76.
- Miller, G.A., & P.N. Johnson-Laird. (1976). *Language and perception*. Cambridge MA: Harvard University Press.
- Milne, R. (1986). Resolving lexical ambiguity in a deterministic parser. *Computational Linguistics* 12(1), 1-12.
- Moyne, J.A. (1985). *Understanding language*. New York: Plenum.
- Newell, A. (1973). Production systems: Models of control systems. In W. Chase (Ed.), *Visual information processing*. Academic Press, New York.
- Obermeier, K.K. (1987). Natural-language processing. *BYTE*, 12(14), 225-232.
- Opdycke, J.B. (1965). *Harper's English grammar*. New York: Harper & Row.
- Quillian, M.R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge MA: MIT Press.
- Quick, R., & S. Greenbaum. (1973). *A University Grammar of English*. Harlow UK: Longman.
- Pacemaker Core Vocabularies. (1980). Belmont CA: Fearon Pitman Publishers.
- Pentland, A. (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence* 28, 293-331.
- Pereira, F.C. & D.H. Warren. (1980). Definite clause grammars for language analysis--a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence* 13(3), 231-278.

- Potter, W.D., & R.P. Trueblood. (1988) Traditional, semantic, and hyper-semantic approaches to data modelling. *IEEE Computer*, 21(6), 53-64.
- Putnam, H. (1975). The meaning of meaning. In K. Gunderson (ed.), *Minnesota Studies in the Philosophy of Science: Vol. 7*, (pp 131-193). Minneapolis: University of Minnesota Press.
- Rapaport, W.J. (1985). Meinongian semantics for propositional semantic networks. 23rd *Proceedings of the Association for Computational Linguistics*, 43-48.
- Rettig, M. & M. Bates. (1987). How to choose natural language software. *AI Expert*, 3(7), 41-49.
- Riesbeck, C.K. & R.C. Schank. (1978). Comprehension by computer: Expectation-based analysis of sentences in context. In W.J. Levelt & F. D'Arcais (Eds.), *Studies in the perception of language*. John Wiley, New York.
- Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Advances in cross-cultural psychology: Vol. 1*. London: Academic.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and categorization*, (pp. 28-49). Hillsdale NJ: Lawrence Erlbaum.
- Ross, J.R. (1967). Constraints on variables in syntax. Unpublished doctoral dissertation, Massachusetts Institute of Technology, New Haven CT.
- Roget, P.M. (1979). *Thesaurus of English words and phrases*. New York: Avenel. (Original work published 1933).
- Schank, R.C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3, 552-631.
- Schank, R.C. (Ed.). (1975). *Conceptual information processing*. Amsterdam: North-Holland.
- Schank, R.C. (1979). Natural language, philosophy, and intelligence. In M. Ringle (Ed.), *Philosophical perspectives in artificial intelligence*. Atlantic Highlands NJ: Humanities.
- Schank, R.C. (1980). Language and memory. *Cognitive Science*, 4, 243-284.
- Schank, R.C. & R.P. Abelson (1977). *Scripts, plans, goals, and understanding*. Hillsdale NJ: Lawrence Erlbaum.
- Searle, J.R. (1959). On determinables and resemblance. *Proceedings of the Aristotelian Society: Supp. Vol.*
- Searle, J.R. (1969). *Speech Acts*. Cambridge UK: Cambridge University Press.

- Sparck Jones, K. & B. Boguraev. (1987). A note on a study of cases. *Computational Linguistics*, 13(1-2), 65-68.
- Sparck Jones, K. & Y. Wilks. (1983). *Automatic natural language parsing*. Chichester UK: Ellis Horwood.
- Texas Instruments Data Systems Group. (1987). *PC Scheme. A simple modern Lisp user's guide* (rev. B). Austin: Author.
- Texas Instruments Data Systems Group. (1987b). *TI Scheme language reference manual* (rev. B). Austin: Author.
- Thorndyke, E.L., & I. Lorge. (1944). *The teacher's word book of 30,000 words*. New York: Columbia University Teachers' College, Publications Bureau.
- Wilks, Y.A. (1973). An artificial intelligence approach to machine translation. In R.C Schank & K.M. Colby (Eds.), *Computer models of thought and mind*, (pp. 114-151). San Francisco: Freeman.
- Wilks, Y.A. (1977). Methodological questions about artificial intelligence: Approaches to understanding natural language. *Journal of Pragmatics*, 1(1), 69-84.
- Wilks, Y.A. (1981). A position note on natural language understanding and artificial intelligence. *Cognition* 10, 53-74.
- Winograd, T. (1983). *Language as a cognitive process: Vol. 1. Syntax*. Reading MA: Addison-Wesley.
- Winograd, T. (1987). Natural language: The continuing challenge. *AI Expert*, 3(7), 7-8.
- Zadeh, L. (1983). Commonsense knowledge representation based on fuzzy logic. *IEEE Computer*, 16(10), 61-66.

END

17/09/90

FIN