[redacted]

Conjunctions and Knowledge Acquisition from Text

by

Laura Jane Proctor
B.Sc., University of Guelph, 1975

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Arts

in the Department of Linguistics

We accept this thesis as conforming
to the required standard

[redacted]

Dr. J.E. Kess, Supervisor (Linguistics)

[redacted]

Dr. T.F. Hukari, Department Member (Linguistics)

[redacted]

Dr. A.C. Brett, Outside Member (Computing User Services)

[redacted]

Dr. B.A. Schaefer, Additional Member (Acquired Intelligence, Inc.)

[redacted]

Dr. C.K. Leong, External Examiner (Psychological Foundations)

UNIVERSITY OF VICTORIA

Supervisor: Dr. J.F. Kess

ABSTRACT

A simple model for interpreting subordinating conjunctions is presented and applied to the task of automatic knowledge acquisition from written documents. Knowledge acquisition involves identifying features of the written document which correspond to the basic units of the underlying representation. The relationships between basic units then establishes an organizational component of interpretation. Subordinating conjunctions are not only indicators of structural (syntactic) form but also organize the text's interpretation. The organizational component of a text's interpretation is complex and may involve a number of different levels such as temporal sequence, physical consequence or cause, physical location, and logical contingency. In the application presented here, only the logical contingency level is addressed. However, the basic model may be applied to any other level of organization.

The basic units of the representation used are called objects and correspond to clauses or clause complexes. The internal organization of these basic units is not important to the model as it is presented here. What is important, however, are the relationships among objects. These relationships are represented by directed links which capture the ordering of objects in terms of logical contingency among them. Such links can be indicated by subordinating conjunctions which can be classified into three groups, based on the ordering they signal between their

subordinate and main clauses. The three groups are defined as pre-ordered, post-ordered, and parallel-ordered, corresponding to the position of the subordinate clause relative to the main clause in the representation.

The initial research reported here has successfully generated a knowledge base from a sample Bylaw of the City of Victoria. Such a preliminary knowledge base is intended to provide a basic knowledge schema for further use by Expert Systems developers. This knowledge schema efficiently extracts the important relationships between objects, and at the same time provides the basis for a hypertext interface to the on-line document.

Examiners:

[redacted]

Dr. J.F. Kess, Supervisor (Linguistics)

[redacted]

Dr. T.E. Hukari, Department Member (Linguistics)

[redacted]

Dr. A.C. Brett, Outside Member (Computing User Services)

[redacted]

Dr. B.A. Schaefer, Additional Member (Acquired Intelligence, Inc.)

[redacted]

Dr. C.K. Leong, External Examiner (Psychological Foundations)

CONTENTS

LIST OF FIGURES

LIST OF EXAMPLES

ACKNOWLEDGEMENTS

I would like to express my appreciation to all of those who helped me over the last three years in the research which lead to successful completion of this thesis. First, many thanks to all of the members of my committee who had the confidence in me to give me plenty of rope, but were kind enough to make sure that I did not hang myself. Thanks also to all my fellow graduate students for many stimulating discussions and lots of moral support.

This thesis would not have been possible without the support of Acquired Intelligence Inc. who provided equipment, software and advice throughout the project. I would particularly like to thank Brian Schaefer and Paul Sihota for their interest and often crucial comments that helped me immeasurably. Thanks also to Bev, Ian, Rob, Rick, Omer, and Mike for making me feel so much a part of the group at AI.

Financial support for this project was provided by the B.C. Science Council and the Advanced Systems Institute.

And finally, special thanks to my friends Owen Griffiths and Chris & Doris Main and my parents who all stood by me and provided the encouragement I needed to get through the times of uncertainty.

Chapter I

INTRODUCTION

Written texts hold a wealth of information about our knowledge of the world. Writers use language to encode this knowledge to communicate with others. Readers gain knowledge by decoding the message contained in these written texts. Text linguistics (discourse analysis), psycholinguistics, and artificial intelligence (natural language processing) are specifically concerned with how these two processes are accomplished. Traditionally; these fields have operated quite independently of each other. However, driven by the recent demand for practical results in research and an increasing interest in computational models in linguistic theory, experts in these fields have started to work together. This change has resulted from the realization that many of the issues that were addressed separately are, in fact, common to all three disciplines. This investigation will bring together previous work in several areas of each of these disciplines by focussing on two inter-related issues: the role of conjunctions in discourse and automatic acquisition of knowledge from text.

Any definition of text or discourse as a linguistic unit involves the notion of connectedness. A sequence of sentences intuitively understood to be a text has cohesion of both form and meaning. Although a precise technical definition of this concept is still developing, there is general agreement that a text model must incorporate this concept of connectedness in order to provide an explanation for

both comprehension and production of text. In the first case, linguistic (and other) constructs must provide sufficient information to allow a reader to understand a text; in the second, the same constructs are used by the writer to encode the intended message. The study of text cohesion involves investigating both how text elements are connected and what those connections contribute to the reader's understanding of textual meaning.

Conjunction of both clauses and sentences is one linguistic feature that contributes to cohesion in text. The clauses and sentences of a text occur in a linear sequence. In any grammar of English, syntactic rules assign structural relations like subordination and coordination between clauses. The syntactic structure does not uniquely differentiate among the possible semantic inter-clause relations. For example,

I went to the store because I was hungry.
I was hungry therefore I went to the store.

or even

I was hungry and went to the store.

In each case, the same semantic relation of causal connection between the two clauses is possible, although their syntactic roles are reversed. In order to capture the semantic relation between the two clauses, other information must be considered. The explicit conjunctions, *because* and *therefore*, indicate clearly the intended relation in the preceding examples. As the following sentence demonstrates, no explicit conjunction nor syntactic structure is necessary to convey the intended meaning because readers can infer the relationship from their existing knowledge about *hungry* and *store*.

I was hungry. I went to the store.

Syntactic structure, semantic relations between clause or sentence meanings, and explicit connectors like conjunctions (in the examples above) and prepositions, are all manifestations of conjunction which contribute to the cohesion of a text.

The few studies of conjunctions that have been undertaken (Halliday & Hasan 1976, Martin 1983) have each proposed similar typologies covering the function of these words in discourse. However, much less has been done towards utilizing these insights in a computational framework. This is a result of the fact that models of discourse structure, which these words affect, have only recently been developed sufficiently to permit some understanding of the kinds of operations they represent. Markers of conjunction signal the dependency between propositions conveyed by the clauses (and other units) of a text, and therefore their function must be defined as operations on the representation of the clausal material. Recent work on applied systems for machine translation and natural language generation have highlighted a number of additional factors which affect the use and function of conjunctions in textual meaning. The writer's style, the type of text (expository, narrative, etc.), and the writer's intentions all affect the use of explicit conjunctions.

Clearly the connections explicitly marked by conjunctions are not sufficient for complete comprehension of the relationships between portions of a text. Reference, substitution, ellipsis, lexical cohesion and the inferences implicit in general "knowledge" about the domain all contribute significantly to building a complete representation of the discourse. Systems which have been developed to "understand" text input have usually incorporated a pre-existing knowledge base

of the restricted domain to build the discourse structure, identifying the flow of topic. However, if no such knowledge base exists, and this is often the case, then this approach is not possible.

Particularly in expository texts, conjunctions are frequently used to explicitly mark connections between portions of the text in an effort by the writer to clarify the important logical connections within the propositional content. Therefore, the structure signalled by these markers provides a rough approximation to the overall organization of the domain knowledge in terms of complex propositions.

The practical goal of this research project, undertaken in co-operation with colleagues at Acquired Intelligence, Inc., is to develop a methodology to automatically generate a partial structure representing textual relations to be used as a starting point in the development of a knowledge base.[1] Acquired Intelligence, Inc. has developed software, called ACQUIRE[2], to facilitate the creation and development of knowledge based expert systems. This project will enhance the current system a) by automatically producing a skeleton knowledge base which can be completed by a domain specialist and b) by investigating how links between the knowledge base and on-line text can be used to provide improved "explanatory" capability in an expert system.

Both the type of text and the knowledge representation used in the study were selected because they are directly relevant to research on knowledge acquisition techniques at Acquired Intelligence, Inc. The "Parking Lot Bylaw" of the City of Victoria (1987) is used as a sample text. It is representative of the

---

[1] This research has been supported by Science Council of British Columbia G.R.E.A.T. fellowship award and a Graduate Student Scholarship from the B.C. Advanced Systems Institute.

[2] ACQUIRE is a registered trademark of Acquired Intelligence, Inc.

formal, expository style characteristic of official regulations and procedures. Such documents exist in many fields for which expert systems could be extremely useful. This type of document usually has a highly structured format and style that has been developed with considerable effort towards clarity of meaning. The information contained in these documents forms a part of the domain's knowledge which must be identified to build a knowledge base. The characteristics of these formal documents make them useful in examining the role of conjunctions in reflecting and identifying the structural representation of the domain knowledge.

The automated analysis of textual material, in order to derive the knowledge represented by the text and encode it in a knowledge base, has not been addressed in any depth. Shaw and Gaines (1987a) have used a statistical procedure to identify important terms and some of their relations. This analysis examines the co-occurrence of content words in sentences of a text. By using the most frequently occurring content words as representative of important concepts in the domain, an initial schema of the area is generated in terms of elements and constructs. The representation derived by this method works from the "bottom-up", in the sense that the words identify the objects, characteristics, or actions which are used in the area. In computational linguistics, sublanguage analysis of legal and medical documents is also primarily concerned with content words and their co-occurrence patterns. Although the purpose of most such analyses is to develop specialized dictionaries for language processing systems, the information gathered is very similar to that required in a domain knowledge base.

There are two other aspects of documents which are used to encode the organization of their content. First, the logical structure of the printed document

in terms of components like "headings", "paragraphs", "sections", and the like, provide structure within printed texts. Often these structural components will correspond with at least one level of organization of the content of the document. In addition, many of the non-content words, in particular, conjunctions and prepositions, provide clues as to the type of logical connection between or within text components. Both of these structural characteristics of texts help to indicate a higher-level organization of the text content. Particularly in the case of formal or technical documents such as bylaws, regulations or operations manuals, these features are used extensively to provide a clear and useful organization for the benefit of those using such documents.

In the four chapters that follow, the research undertaken to investigate the use of connectives in discourse and in signalling the logical structure of the discourse content is presented. Chapter 2 reviews some recent literature in the area of knowledge acquisition (Artificial Intelligence), including a description of the ACQUIRE software developed by Acquired Intelligence Inc. Then, Chapter 3 presents a survey of work in discourse analysis (text linguistics) to demonstrate how it is that these two areas share many of the same questions, but approach the questions from different perspectives. Chapter 4 presents the analysis of a sample bylaw of the City of Victoria, describing the analysis of conjunctions and how this analysis was applied in a knowledge acquisition system. Finally, a summary of the project and the conclusions drawn from the experience is set out in chapter 5.

Chapter II

BACKGROUND IN EXPERT SYSTEMS AND KNOWLEDGE ACQUISITION

2.1 Expert Systems

The term Expert System is used to describe a computer program whose purpose is to fulfill the role of a human expert in a restricted task domain. The expert roles that have been most amenable to automation are classification or diagnostic tasks and decision-making. Providing relatively unrestricted access to the specialized knowledge of one or more experts is a practical goal of the growing field of Expert System development.

Some of the areas in which expert systems have been developed are medicine (MYCIN), organic chemistry (DENDRAL) and Law (LDS), to mention only a few. For example, MYCIN (Davis et al. 1977) is a program designed to consult with a physician in the diagnosis and therapy of infectious diseases. The physician works interactively with the program by first providing needed case details. The program then provides possible diagnoses and recommendations for therapy. In addition, the path of reasoning used by the program in coming to conclusions is also accessible to the physician. DENDRAL (Barr & Feigenbaum 1981b) uses the chemist's input to provide structural analysis of organic compounds. The LDS (Legal Decision-making System) (Waterman, Paul, Peterson 1986) system addresses product liability cases, computing defendant liability, case worth, and settlement amount.

Traditionally, software development is carried out in a procedurally oriented environment. "Procedural" means that a program must specify exactly what operations are to be applied to incoming data and in what order alternative options are selected. Each new program requires a new procedural description, tailored to the characteristics of the specific application. The relationships among data elements are captured implicitly in the program procedures. The knowledge of how to use the data is, however, explicit in the procedures.

Knowledge-based, or "declarative", programming is a paradigm which has become nearly synonymous with expert system development. It provides a programming environment which supports very complex interaction of information sources. Knowledge-based systems make a clear distinction between "knowledge", or what is known about the application area, and how that knowledge is applied to solve specific problems. This is often called declarative programming because the programmer defines only the domain-specific information. The procedures which access and apply this knowledge to specific input values are built-in to the programming environment. This provides a metaphor where programming can be viewed as specifying an appropriate description of specific knowledge that is operated on by generalized reasoning or inference procedures. From the point of view of one using a piece of software, it is not usually obvious how a system has been implemented (procedurally or declaratively, or both). The difference is crucial, however, to the software developers and maintainers (the programmers and analysts). Though it is yet to be proved (by experience), promoters of the declarative technique suggest that it is easier to develop and maintain programs in this way.

There are usually three major components in a knowledge-based programming system: a knowledge base, a reasoning module and a knowledge acquisition module. The knowledge base consists of structured data representing general, domain knowledge. The knowledge acquisition module provides an interface through which data in the knowledge base can be augmented or changed. And, the reasoning module, often called an "inference engine", comprises procedures for using a knowledge representation to derive a diagnosis, classification or decision, based on specific case input (Savory 1988). A number of different expert system development packages have been produced, both commercially and experimentally, each with this same basic design. They provide the basic framework for an expert system; however, no "knowledge" of any particular domain is provided. It is the job of the system developer to appropriately formalize the domain knowledge as "data" on which the reasoning mechanism can operate. To anthropomorphize the description, one could say that the system comes with the ability to think, but has nothing to think about (Savory 1988).

Expert systems have been developed in many diverse domains. The application to public regulatory documents, such as the municipal Bylaw used here as a sample, is the target of this research. Successful expert systems have been created from such documents in a number of areas. The British Nationality Act was manually translated into PROLOG, a programming language for knowledge based programming (Sergot et al. 1986). A draft fire safety standard provided the basis for creation of an advisory expert system in New Zealand (Buis et al. 1987). An advisory expert system on shipping regulations has been

developed based, in part, on the Canada Shipping Act (Lockwood 1988). All of these projects involved human, not machine, interpretation of the source documents.

All of these projects resulted in successful expert systems. It is interesting to note that researchers in each of these developments reported that ambiguity, incompleteness and inconsistency in the text, which was not easily recognized, was revealed by the operation of the resulting expert systems. This suggests that expert system development might be a valuable tool for drafting new regulations, helping authors avoid these kinds of communication problems.

2.2 Knowledge Representation

Research into knowledge representation formalisms and reasoning procedures to operate on them is a major area of Artificial Intelligence (Davis & Lenat 1982, Brachman & Levesque 1985, Barr & Feigenbaum 1981a). Semantic networks were adopted from psychology (Anderson & Bower 1973) and introduced to Artificial Intelligence by Quillian in his 1966 dissertation (Brachman 1979). So also were scripts (Schank 1980), frames (Minsky 1981), and production rules (Davis et al. 1977), and these are all forms of representation that have been used in expert systems. Although this area is extremely important, the relative merits of these forms of representation will not be discussed here.

Most commercial Expert System software (including the software used in the project described in Chapter 4) uses a production rule format to encode "knowledge" (Quinlan 1987). Production rules are used extensively because they are easy to create and edit incrementally since each rule is content-wise

independent of all other rules (Savory 1988). Although, it should be added that the interaction of the rules will affect the content and quality of the overall system's advice. The rules also provide a fairly clear path through the reasoning procedures. This feature has been used to provide explanation capabilities with some success (MYCIN).

Production rules have a left-land side (LHS) which defines a context and a right-hand side (RHS) which expresses a result or conclusion. Presenting this schematically:

LHS --> RHS

which can be interpreted as:

IF the LHS condition holds true
THEN the RHS condition holds

In many cases, the conditions may be quite complex, depending upon the particular system. Following are two example rules from XCON (Winston 1987: 9), a system used to configure computer-system components:

if	the context is layout and assigning a power supply an sbi module of any type has been put in a cabinet the position it occupies in the cabinet is known there is space available in the cabinet for a power supply there is no available power supply the voltage and frequency of the components is known
then	add an appropriate power supply

if	the context is layout and assigning a power supply an sbi module of any type has been put in a cabinet the position it occupies in the cabinet is known there is space available in the cabinet for a power supply there is an available power supply
then	put the power supply in the cabinet in the available space

Another example (Clancey 1984: 60) is a rule taken from MYCIN, a medical expert system.

RULE535

if 1) The infection which requires therapy is meningitis,
2) Only circumstantial evidence is available for this case,
3) The type of the infection is bacterial,
4) The age of the patient is greater than 17 years,
5) The patient is an alcoholic

then There is evidence that the organisms which might be causing the infection are diplococcus-pneumoniae (.3) and e.coli (.2).

The basic units of the knowledge representation used in this project are called objects. Objects represent concepts relevant to reasoning about a domain. An object may be a physical object (eg. a tool), a characteristic (eg. patient's age), a measurement (eg. body temperature) or a more complex concept such as a hypothesis (eg. "the infection which requires therapy is meningitis").

A set of possible values is associated with each object. The LHS of a rule is a conjunction of clauses (possibly only one), each specifying a value or range of values; a relation, such as "equals"; and an object. When a rule is used, the actual value of each object (entered by the user or derived through the application of other rules) is compared under the relation to the value specified in the LHS clause. If all clauses of the premise (LHS) are met, the RHS action is performed, usually setting values for one or more objects.

In a very simple example from a medical expert system, a rule might be as follows.

if Patient's-Temperature is (equals) ABOVE-NORMAL

then Likelihood-of-disease-A is (equals) HIGH

When this rule is used, the value of the object, **Patient's-Temperature**, will be compared to the value **ABOVE-NORMAL.** If the values are equal, then the value of the object **Likelihood-of-disease-A**, will be set to **HIGH.**

The reasoning module in this type of system must select which rules should be applied or "fired". Particularly in large systems with many hundreds of rules, the choice of the appropriate rule or rules to use is complex. A number of general strategies are commonly used. Forward-chaining describes a strategy often called data-driven or bottom-up reasoning. That is, when the values on the LHS of the rule are all available, then the rule will be applied, usually setting values referenced in the RHS. At the next step, with new values assigned, new rules will become available for application.

In some applications, a top-down or goal-driven strategy, called backward-chaining, is more appropriate. Using this strategy, one RHS is selected as the "goal" and then the system attempts to find an appropriate value for it by establishing values for the LHS. If the LHS objects do not have values assigned, rules which can be used to set their values (that is, the same objects are on the RHS of a rule) will be examined. Eventually this "backward" search will end by finding a rule in which all the LHS values are set. The system then retraces its path through the sequence of rules, setting RHS values on the way. (This is essentially equivalent to forward-chaining).

Whichever of these two strategies is used (perhaps both), there will often be more than one rule which could be applied. Therefore, some strategy must be used either to resolve which to use, or in which order to apply them all. Some strategies that have been used are:

- the first rule in terms of the linear arrangement in the knowledge-base.
- the rule with the highest "priority" as defined by the programmer.
- the rule with the most specific context (the rules from XCON are designed to work with this strategy)
- use all of the rules (Feigenbaum & Barr 1981a).

These rule selection strategies represent the system's methods of reasoning about the knowledge encoded in rules. Some control in the reasoning processes is also implicit in the content of the rules themselves. Thus, in addition to representing basic entities and relations in a domain, the knowledge base must also include some information about how basic rules should be applied. In the case of production rule systems, this knowledge is, in part, captured by the strategy used for rule-selection, as well as by the content of the rules themselves. In a system developed to serve a single function or role, this situation has provided satisfactory results, but to adapt the system to a new function can entail restructuring of the entire knowledge base. This difficulty was encountered in the development of NEOMYCIN, a teaching system, from MYCIN which was designed as a consultation system (Clancey 1984).

An example (Clancey 1984: 60) of implicitly coded knowledge is demonstrated in the following rule from MYCIN:

RULE535

if
1) The infection which requires therapy is meningitis,
2) Only circumstantial evidence is available for this case,
3) The type of the infection is bacterial,
4) The age of the patient is greater than 17 years,
5) The patient is an alcoholic

then
There is evidence that the organisms which might be causing the infection are diplococcus-pneumoniae (.3) and e.coli (.2).

The first three clauses specify conditions for applying this rule, and so implicitly order the rule relative to others used to establish the conditions. Implicit general knowledge, that children are not usually alcoholic, is embodied in clause 4) and the rationale for linking alcoholism with the suggested bacteria is not explicitly coded.

While this rule was adequate for MYCIN, designed as an expert system for consultation, it was not adequate to fulfill the instructional role required in the NEOMYCIN system. The MYCIN knowledge base required significant restructuring for this new use. For example, knowledge about taxonomic relations between diseases encoded implicitly in the ordering of LHS goals in rules had to be made explicit in order for the system to function effectively in a teaching role. In general, how experts organize and remember their knowledge, and what problem solving strategies they use, are not explicitly coded in MYCIN (Clancey 1984: 56).

The role, or function, the expert system is expected to fulfill has influenced the choice of representation and reasoning strategy for particular knowledge domains. However, problems like the one described above have led to investigations of human reasoning and initial steps towards a theoretical basis for discriminating types of knowledge. This work is also pursued in the area of cognitive psychology, where the objective is to provide adequate explanations of human reasoning. The distinction between procedural and declarative programming is essentially the same as that made between procedural and declarative knowledge in human memory (Anderson 1982). Much of this work has focussed on developing models of learning or skill acquisition (Anderson 1987) as a way to understand just what it is that people know.

All of these issues are important in the design of systems for acquiring knowledge for expert systems. In this research, a production system representation was used, in keeping with the ACQUIRE knowledge acquisition software made available by Acquired Intelligence, Inc. In the next section, some current methods in knowledge acquisition will be reviewed.

2.3 Knowledge Acquisition

This section will provide an overview of some issues in knowledge acquisition research. Special attention will be directed towards the way textual materials have been used to date since the research reported in this thesis is applied to improve the effectiveness of using written documents as a knowledge source. Knowledge acquisition is the process of eliciting, analysing and interpreting knowledge in a specific domain. This first, crucial stage of knowledge base development continues to be a "bottleneck" in commercial development of expert systems (Shaw & Gaines 1987a, Gaines 1987, Kidd 1987, Savory 1988, Davis & Lenat 1982).

The knowledge acquisition problem has been addressed by developing techniques for interviewing experts about their knowledge and for observing experts in the performance of their specialized skills. Methods for automatic learning by example (Michie 1987, Hart 1986) and for learning by "being told" (Haas & Hendrix, 1983) are being developed to augment the tools available to system developers. In addition, more attention is being directed towards creating software to aid domain experts to formalize their knowledge and to enable them to interact directly with the acquisition system. Overall, there has not yet emerged any overall framework that can capture the nature of the knowledge acquisition process. What is certain, is that many different factors are involved, and simple answers are unlikely.

Identifying the relevant concepts and their relationships, and encoding them as objects and rules is often difficult for those not familiar with the process. In many cases, an expert's ability to articulate the methods they use in performing a

task is very poor. Therefore, a new area of expertise emerged, called "knowledge engineering". Most typical expert system development projects involve one or more domain experts (a person who has a "complete" knowledge of the subject area) and a "knowledge engineer", a person who is not necessarily knowledgeable in the subject area, but is skilled in identifying and organizing data for a knowledge base (a computer analyst). This process is not only expensive in terms of the time required, but also extremely difficult. Very accurate communication between these people is crucial for the success of a development project, and it has been the general experience that it is necessary to involve the area expert in very detailed levels of design. "Rapid prototyping" is an experimental trial and error approach which has been typical of expert system development (Breuker & Wielinga 1987). Neither the characteristics of potential applications, nor of available implementation methods (that is, methods of representation and reasoning) are yet understood sufficiently to have provided a foundation for knowledge acquisition methodology. This problem is well recognized (Clancey 1985, Breuker & Wielinga 1987, Buchanan & Smith 1988) and numerous efforts are being made to develop well-founded knowledge acquisition methodologies (Kidd 1987, Hart 1986).

The interviewing of domain experts has been the primary method of acquiring knowledge for new applications (Shaw Gaines 1987a, Kidd 1987). The analysis of verbal protocols (Ericsson & Simon 1985) in "thinking aloud" experiments have provided insight into expert reasoning strategies in a number of different task domains. The difficulty with this type of analysis lies in finding appropriate units in the protocol for analysis (Johnson 1985). Written documents have been

used to identify the basic structure of the knowledge domain (Breuker & Wielinga 1987). Once key terms used in the documents are organized, a structured approach to the interpretation of verbal protocols can be taken (Lesgold et al. 1988).

In this way, written documentation is seen to play a role in the initial stages of knowledge acquisition when the basic units or concepts of the domain are collected and organized. This analysis focuses on lexical identification of content words, like nouns and verbs, taking them to represent objects and actions basic to the domain. Sublanguage analysis, a domain-specific approach to natural language processing of recent interest in computational linguistics, also involves analysis of lexical items and patterns. This work will be discussed in chapter 3.

Gaines (1987) views knowledge transfer as a social process operating under evolutionary survival constraints. In this multi-level model, he suggests that basic distinctions and rules are part of the social environment, acquired by individuals through mimicry and reinforcement of experience. A level called "computational", equated with rational explanation, links these lower level with "higher levels" of organization in terms of alternative and abstract models. More than one model might be relevant. For example, the same set of distinctions might be part of both a causal and a temporal model. It is at the level of rational explanation that language (in particular texts) is seen to be a source of information.

A text base is one of the central information sources in the automated knowledge acquisition environment envisioned by Shaw & Gaines (1987a). This environment is based on Personal Construct Psychology (Shaw & Gaines 1987b)

from which the repertory grid is taken. The repertory grid defines relations between elements, or basic concepts on a number of binary dimensions, or constructs. The text base is used to seed the repertory grid to give the developer a starting point. Frequently occurring "non-noise" words in the text are identified as elements. Associations between them are computed on the basis of their co-occurrence in sentences of the text. Clustering the high-frequency associations provides an indication of how key concepts should be grouped together (Smith 1976).

While this analysis of word associations provides valuable information about the relative importance and groupings of key concepts reflected by the words used in a text, it does not make use of the information conveyed through connectives. Connectives are among the "noise words" ignored. In this system, no connection between the identified concepts and the original text is maintained. As a result, the developer cannot directly refer to the actual text to see exactly how the words are used and what kind of connections are made between them.

2.3.1 Summary

Recent research and development in the area of knowledge acquisition have focussed on developing a structured methodology, drawing past experiences of systems developed in an experimental, "rapid prototyping" environment. A few of the directions in the field have been reviewed here to provide some background to the knowledge acquisition problem.

Documents are used as a source of valuable information for expert system developers. However, only limited attempts have been made to integrate automatic text analysis methods. Content word analysis provides a direct way of

identifying basic concepts and a general indication of the relations between them. Little mention is made in the literature of function words (connectives) or document format as a source of information about the structure of concepts in the domain. Undoubtedly, knowledge engineers use this type of information in analyzing documents. However, they are often not explicitly aware of doing so. The diversity of relations that can be expressed through these features and the lack of a clear theoretic basis for deciding which relations are relevant for a particular application are obstacles to understanding exactly how this kind of information can be utilized.

The project described in Chapter 4 has undertaken to use both document format and connectives to produce a "seed" knowledge base intended to aid the system developer in the initial stages of knowledge acquisition. Acquired Intelligence Inc. have made their knowledge acquisition software, ACQUIRE, available for use in the project. The following section briefly describes this system and the knowledge representation it uses.

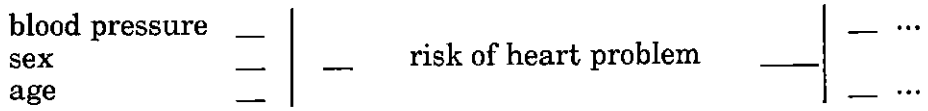
2.4 The ACQUIRE Knowledge Acquisition Software

The knowledge representation used here is comprised of two different forms of information: objects and rules. The use of the term object for the basic units in the knowledge base is chosen to be as neutral as possible about what kind of things they represent. Objects in the system may represent states of affairs, values of observable qualities or states, physical objects, or any other type of information that is necessary in the application domain. Each object must be given a name, a description, a set of possible values and a list of the other objects

it supports or that support it. A schematic description of this representation is presented in Chapter 4 (beginning on page 53) and the implementation details may be found in Appendix B.

Rules are then created to express the actual relationship among objects in terms of their values. The LHS, or context, of each rule contains one or more expressions relating an object to one of its possible values. The RHS, or action of the rule, causes one or more objects to be assigned one of their possible values. Although in practice both LHS and RHS may have more effects than testing relations and assigning values, this is not of concern here. Including these extra operations is seen to remain the developer's responsibility. This research attempts to give the developer a reasonable starting point from which to construct the expert system.

The rule set gives the most detailed level of description of the knowledge base. Another, complementary level of description which summarizes the knowledge base at the object level is also available in ACQUIRE. At this level, there is a network in which each node represents an object and the connecting lines indicate a support relation. A graphic display is available to the developer, similar to the following diagram.



In this example, **blood pressure**, **sex**, and **age** are all objects which appear on the LHS of a rule with **risk of heart problems** as the RHS object. These three objects all support the object, **risk of heart problems.** We could also say that **risk of heart problems** is supported by the other three objects. At this level of description, the

actual values the objects may take on is not important, nor is it important whether there are one, two or many rules with a particular object on its RHS, we see here only a summary of all supporting relationships between the objects of the system. This representation is the desired output from the automated text analysis attempted in this project.

Chapter III

LINGUISTIC RESEARCH ON DISCOURSE

Conjunction or juncture of clauses is a feature of language that is not restricted to the domain of a sentence, the traditional maximum unit of linguistic enquiry. Conjunction or joining also operates between sentences and is one of the features that helps connects ideas together both within and between sentences. Thus, the study of conjunction has been undertaken as a topic in the field of Discourse Analysis and/or Text Linguistics since it is concerned with language units larger than single sentences or utterances.

3.1 Cohesion in Discourse

Discourse is a unit of language in use (Halliday and Hasan 1976:1) and so, has a purpose and a focal topic. It is realized as a sequence of one or more sentences. The message communicated by a discourse is coherent in the sense that its component parts are understood to be connected. It is this appeal to "use", "purpose" and "connectedness" that distinguishes a discourse as a linguistic unit and at the same time makes its investigation so difficult. This is because the overt, or surface form of a discourse can be so varied that the most basic units familiar to linguists (words, phrases, clauses and sentences) do not seem to provide building blocks that explain discourse structure. Rather it is necessary to appeal to constructs like "topic", "purpose" and "intention", all of

which are abstract features, more connected with the question of mental representations than with the words on a printed page.

Halliday & Hasan describe discourse (or text) as follows:

> "A text is best regarded as a SEMANTIC unit: a unit not of form but of meaning. ... A text does not CONSIST of sentences; it is REALIZED by, or encoded in, sentences." (Halliday & Hasan 1976: 2)

Although text or discourse is intuitively easy to understand, a clear definition is very difficult. A sequence of unrelated statements or questions such as the following is not considered a discourse.

> The weather has improved today. Regarding the matter of fees, it is important that every member ensure their account is up-to-date. It does so deliberately and on the basis of considerable thought. And so, trailing his coat behind, he wandered off.

It does not have the connectedness that characterizes our concept of a realistic unit of language. Cohesion and coherence are terms that have been used to describe the features of connectedness in text. Cohesion refers to the linguistic devices used to signal connections and coherence to the structure of the resulting conceptual understanding derived from the surface text.

Without a reasonable context in which to place this discourse, the topic of the first sentence (today's weather) has little in common with *the matter of fees*. The reference to *every member* cannot be understood clearly without further context. In the third sentence, the pronoun *It* comes as a surprise because there is nothing in the preceding text to suggest the reference that we are expecting. A similar, though less striking, example occurs in the last sentence (*he*). *And so* which introduces the final sentence of the sequence implies that the following statement provides a result to the events described by the preceding text.

Each of these features of topic continuity, reference/anaphora by pronouns, and conjunction, are examples of cohesion which are not functioning properly in the example. The lack of coherence in the ideas conveyed by the discourse is a result of these features not fulfilling their discourse function.

The study of discourse must involve the communicative role of language and has indeed grown out of the realization that interpretation of individual sentences is not independent of context, either within the text or relative to non-textual elements like the shared knowledge of the speaker and hearer, intentions of the speaker, and many other factors. Modern studies in text linguistics and discourse analysis have all had to grapple with the problem of relating the surface form of language use with the "message" or information perceived to underlie it.

The information conveyed through language can be affected by the reader's existing store of information (knowledge) through various processes. These processes are not well understood, but have been described by various concepts, such as inference, feature inheritance, and spreading activation. The study of discourse thus raises questions whose answers involve understanding how readers/hearers use their knowledge in discourse understanding.

Linguistic enquiry takes the approach of identifying sets of distinctions that may be made and associated forms of expression for those distinctions. Linguistic studies of discourse are directed towards seeing HOW meaning and reference of sentences depend on other sentences in a text (van Dijk 1977, Halliday & Hasan 1976). The basic assumption that component clauses of a text each incrementally contribute to the overall meaning of a text is shared, at least implicitly, by most theorists. Among the problems that must be addressed are: how the components

may be effectively represented; how they can be linked together to form a composite representation; and, lastly, what form that representation should take.

The interaction between cohesion and coherence indicates that discourse processing must provide a bridge between the surface form of a text and representation of the resulting knowledge of the reader. Forms and systems of language have been subjected to extensive study; however, a definitive set of rules or systems governing sentence forms has not emerged. The study of knowledge, whether in the field of philosophy, psychology or artificial intelligence, is also indeterminate at this point. Many insights and observations have been made, but a comprehensive explanation of "knowledge" is still a distant goal.

Therefore, the problem we are facing in discussing discourse is indeterminate from both points of approach. The language data can be analyzed on a number of complementary levels: lexical, syntactic, semantic or pragmatic. These levels of analysis are not, however, independent of each other. Models of the underlying "knowledge" or cognitive representation which results from language understanding have only begun to emerge and are far from complete. It does seem clear, however, that an interaction between the two lies at the heart of the problem of discourse understanding and production.

Halliday & Hasan (1976) investigated the linguistic forms that may be used to establish connectedness, or cohesion, in texts. Based on their examination of these linguistic structures, they present a taxonomy of textual relationships, including amongst others, conjunction. In contrast, others have started with a set of "notional" or logical relationships (Longacre 1983, Givon 1983, de Beaugrande and Dressler 1981) and looked for correlations between these and

surface forms. The distinction between these approaches is one of perspective, rather than substance. In both cases, the basic intuition is that a text communicates not only isolated facts or ideas, but an organized collection of these.

From a cognitive processing perspective, the analysis of discourse can be viewed from two perspectives: that of the hearer/reader or that of the speaker/writer. In the first case, surface structure is the input to the cognitive system from which some representation or understanding is derived. In the second case, the opposite relation holds. (See Figure 1)

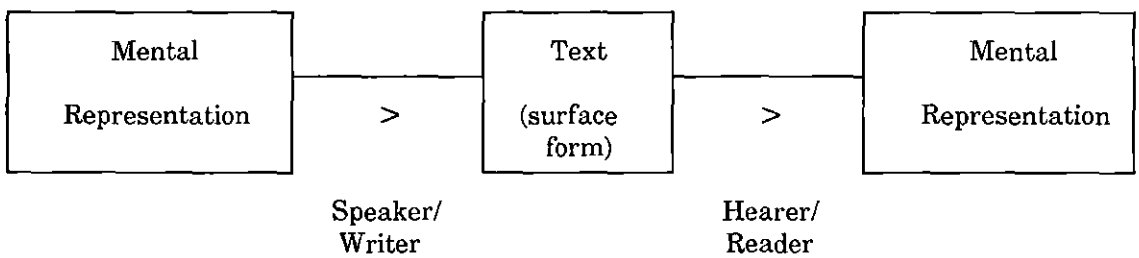


Figure 1: Model of Discourse Comprehension Processing

The representation, which is a theoretical construct, attempts to capture the "meaning" of a text in its broadest sense. If we want to make claims about the psychological reality of this construct, then we could view the model as a snap-shot of that portion of an individual's memory that participates in the understanding of the text. The representation should capture the knowledge that results from "understanding" the text.

The basic building blocks of knowledge are concepts and relations (de Beaugrande and Dressler 1981). One form of representation is that of predicates and arguments (Longacre 1983:77), a formalism that is based on a distinction between things or objects and the relationships between them. Objects represent

possible values for the arguments to predicates (roughly akin to the notions of nominal and verbal concepts, respectively). Clauses are represented by predications which form the "atomic particles of discourse" (Longacre 1983:77). Longacre (1983) proposes a form of statement calculus in which a set of operators is defined to represent the relations between predications. He includes among these relations not only conjoining, alternation, temporal sequence, and implication, but also paraphrase, illustration, deixis, attribution and frustration. The result is to provide a common representation for different surface forms conveying the same relationship. For example, compare the following sentences.

> John went downtown and then bought a hamburger
> John's buying a hamburger followed his going downtown.

In the first sentence the conjunction *and then* conveys the same relation between predications as the verb *followed* in the second. His analysis does not attempt to provide the description of what surface forms correspond (or might) to each of his operators but rather tries to find a good set of operators to express the necessary relations.

De Beaugrande and Dressler (1981) use a conceptual associative network to represent the outcome of discourse processing. Nodes in the network represent concepts which are labelled by the "content words" in the examples. The lines connecting the nodes are labelled with a set of labels intended to be a typology of language relations, largely drawn from case grammars (Fillmore 1968). The two notions of coherence and cohesion in discourse are captured procedurally in this type of representation. Coherence arises from the creation of a network comprised of "knowledge spaces" centred around "topics". Some cohesive devices, like pronominal anaphora, establish links between portions of the network generated

by separate clauses or sentences. The contribution of existing knowledge is described as the addition of nodes to the network, as it is built, through the techniques of spreading activation, inheritance and inference, all of which are triggered by the textual content. In terms of Kintsch (1988), word meanings drive the building of a discourse model, the final stage of which is the integration of concepts into that model. The meaning of conjunctions and other function words is to guide the integration of following components with the model under construction.

The notion of "knowledge spaces" centred around "topics" is common to an number of different studies in text linguistics. Topic continuity in discourse has been studied (Givon 1983 :1-42) by examining how it is expressed through grammatical marking. Discourse is seen to be structured into thematic paragraphs (Longacre 1983), which by definition are continuous with respect to "theme" or "aboutness".

Thematic paragraphs comprise a group of connected propositions which are realized as clauses. The clause, or micro, level is related to the macro level (thematic paragraphs) through three levels of organization: THEME > ACTION > TOPICS/PARTICIPANTS. These levels are related in an "implicational hierarchy" (or 'inclusion set'). Topic/participant continuity is the most "concrete" and, therefore, most strongly coded in the grammatical form of low level clauses. Zero anaphora, pronouns, and definite NPs are among the grammatical devices used to code this level. Action, which refers to temporal sequence and adjacency, is less strongly coded in the grammar, usually realized through verbal tense-aspect subsystems. Theme is the most abstract level and is only intuitively

defined by Givon. It is only weakly coded through grammatical devices. Conjunction is the only grammatical coding of theme identified in English. Givon's studies are concerned with finding how languages compare in terms of the grammatical systems used to mark continuity in narratives and spoken conversations. Once again the inter-relationship between coherence and cohesion underlies these investigations.

Built-in structures like frames, schemata, scripts and plans (de Beaugrande and Dressler 1981) or macrostructures (van Dijk 1980) have been suggested to explain how a reader can identify the higher-level or conceptual organization within text. These structures are not, however, part of the text, but rather in the mind of the reader. In the context of processing a text with a computer program, these kinds of structures can be seen as providing constraints on the possible options for interpretation. However, in the absence of a clear definition of precisely what these structures are and how they are correlated with the text, they do not provide us with immediately useable tools. These higher levels of organization are part of the information which knowledge acquisition techniques try to elicit from experts to provide additional structure to the lower levels of representation. The levels suggested in Gaines'(1987) model of knowledge transfer are very similar to the levels suggested by these authors.

Common to all of these views of discourse is the notion that clauses (or simple sentences) are basic units of discourse representation. Clause or sentence sequences must be connected in the discourse representation to capture the relationships between individual units. The purpose of this investigation is to examine the role of function words in conveying relationships between phrasal

and clausal elements. The question of what clausal representation would be most adequate will not be addressed. But I am assuming that such units are identifiable and function as building blocks in discourse understanding. The question that will be pursued is what connections can be derived from written texts from the cues provided by functors like prepositions and conjunctions.

Halliday & Hasan (1976) identified five different kinds of cohesive "ties": reference, substitution, ellipsis, conjunction and lexical cohesion. Each of these linguistic devices contribute to connecting the clauses in discourse. In this research, the "ties" categorized as conjunction have been of prime interest, providing the basis for identifying objects in regulatory texts. The type of text most often studied in text linguistics/discourse analysis described above have been narratives. The actors, their relationships and actions are the salient features of narratives. Usually narratives rely heavily on shared knowledge of how the world is to provide the background or context for understanding. In expository texts, however, the purpose is generally to convey new knowledge. Although clearly shared basic knowledge of the world is also a crucial element, expository text is intended to convey new or reinforce existing connections. Therefore, the use of explicit connections is both more common and more essential in this type of text.

3.2 Studies of Conjunction

Conjunctions and prepositions are all explicit indicators which contribute to the cohesion of a text. Textual cohesion expressed in the surface structure both rests on and is an indicator of the underlying coherence in the domain of the discourse. Thus, in the absence of predefined knowledge about the textual

domain, cohesive devices provide guidance in building or learning relationships between objects, events, and situations.

This functional role of connectors, a term used here to collectively refer to conjunctions and prepositions, is suggested in the work cited above. Morrow (1986) draws together other similar work to support his position that grammatical morphemes convey not only grammatical distinctions but content distinctions as well. Grammatical morphemes of the function word variety are characterized as guiding the process of discourse comprehension in organizing textual content. Rudolf (1988) presents a similar view of connective expressions as "instructions for cognitive operations" (Rudolf 1988:109). The content of these instructions aids the reader to perceive both information about the factual content of a text, as well as the writer's view of the relative important of events and situations. Halliday and Hasan (1976:227) had earlier described connective expressions as "... a specification of the way in which what is to follow is systematically connected to what has gone before."

The use of connectors to mark cohesive relations is neither necessary nor sufficient to independently guarantee that a sequence of sentences or clauses is interpretable (van Dijk 1977, Rudolf 1988). For example,

1. a) John was sick. He went to the doctor.
   b) John went to the doctor because he was sick.

2. a) The wind was blowing from the west. John went to the doctor.
   b) John went to the doctor because the wind was blowing from the west.

In 1(a) the two sentences are connected, but not by any explicit marking. Rather because we, as readers, know that going to the doctor is a common

reaction to being sick, so we can recognize the connection of cause (being sick) and result (going to the doctor) between the two sentences. In 1(b) the connection is made explicit by the use of the conjunction *because*. The purpose behind each of these expressions on the part of the speaker is likely quite different. In 1(a) a sequence of events is related, but the purpose is not clear without further knowledge about the context of the segment. The order of the sentences is as suggestive of a temporal ordering of the two events as it is of a causal relationship. However, 1(b) in which the causal relation connection is emphasized by the use of the conjunction, the possibilities are narrowed (though the same temporal relation can still be inferred).

The second pair of examples show how important the relatedness of underlying propositions (or facts about the world) are to establishing relations like causation or temporal ordering. The connection between *went to the doctor* and *the wind was blowing from the west* is not supported by general knowledge of the world, so 2(a) does not suggest any causal relation. The temporal ordering assumptions are different from 1(a), again appealing to our knowledge of the world that wind blowing is not a discrete action but rather a continuous one. And 2(b) seems incorrect, although structurally there is no problem. The explicit statement of cause leads to an unacceptable conclusion on the basis of our knowledge (beliefs) about how the world is. (That is, the wind blowing would not normally have anything to do with visiting the doctor).

The principle of relevance is assumed to underlie the intentions of the speaker or writer of a discourse. Although we can construct examples of structurally anomalous or incorrect sentence sequences, we do not expect to find this kind of

sentence intentionally placed in a discourse, particularly not in the type of written documents of interest in this study (manuals, regulations, etc.).

We can take a new perspective towards the role of conjunction in discourse by leaving aside the question of how to identify incorrect connections. Instead we begin with the assumption that the connections expressed by a text are correct and proceed to examine how many of the connections can be extracted by analysis of the explicitly marked conjunctions. In essence this approach asks the question, to what extent can we derive a representation of the organization of propositions from written text. This approach is particularly relevant to illuminating the relationship between text meaning and that elusive notion "world knowledge".

Halliday and Hasan's (1976) study Cohesion in English presents a detailed examination of linguistic features in English texts that contribute to creating their "texture" or cohesion. Cohesion is seen as arising from meaning relations (which they call "ties") between linguistic items of the discourse. For example, there is a cohesive tie formed between *John* and *He* in the following passage:

John returned yesterday. He had a good trip.

These authors were working towards an "explanation of why and how it [a text] means what it does." (Halliday & Hasan 1976:328) Their focus was on identifying the linguistic features observable in texts that contribute to cohesion. They proposed a classification of conjunctive relations that has been incorporated into broader treatments of overall discourse structure (van Dijk 1977). A similar and more detailed study of conjunctions (Martin 1983) resulted in a more finely grained system of distinctions that can be expressed by connectors.

Halliday and Hasan's work has been valuable because they looked at texts to see what kinds of links actually occur and so contribute to the idea that relations between parts of a text are crucial to its meaning. In the context of this study, this is a contribution to identifying what features of a text can provide the clues or information required to interpret the text. Their work is descriptive, and relies on the analyst's intuitions to recognize where the cohesive links occur. It has, however, provided a framework in which to explore what rules might be used to guide an automatic text understanding system.

Several different classifications of conjunctions that have been proposed share, at least at the highest level, common intuitions about the type of connective relation expressed by these words. For example, from Halliday and Hasan (1976), inter-sentential connective relations, realized as conjuncts, are classified into four major categories:

Additive
Adversative
Causal
Temporal

Essentially the same categories are presented in Rudolf (1988:107-8) with very similar terminology:

Connection of Addition
Connection of Contrast
Connection of Causation
Temporal Connection

In Martin (1983), in addition to the inter-sentential relations (or non-subordinating conjunctions in his terms) which were the main focus of Halliday and Hasan, relations between clauses in a single sentence are all treated in the same systemic classification. The four relations:

Temporal
Consequential

Comparative
Additive

are cross-classified with the features explicit/implicit and subordinating/non-subordinating, integrating linguistic form into the network of distinctions. The level of detail in the networks he describes is beyond the scope of the present enquiry, but the information captured in the networks would clearly be of particular importance to dialogue processing.

The concern of these researchers is in identifying distinctions made in language. They have not been concerned with explicitly formulating how this relates to conceptual representations. However, the distinctions they have made among possible types of conjunction are of a semantic nature and are based on the differences among these types. Each of the categories express an underlying organization of concepts according to criteria which are significant to representation of knowledge captured by expert systems.

Causal and temporal relations between situations or events are crucial to organizing knowledge representations. Ordering of rules in a knowledge representation determine the path of reasoning (as described in Chapter 2), and both causal and temporal orderings are important to the type of knowledge contained in regulatory documents. Contingency of facts, whether based on the perception of temporal ordering like the sequence of steps in a task or conditional/causal ordering which occur in regulations is an essential aspect of a knowledge representation. With respect to using documents as a knowledge source, the presence of explicit markers of conjunction not only supply hints about what kind of ordering should be imposed, but also indicate what concepts should be represented. For if such a relationship is marked between two clauses, then it is also likely that the two clauses should represent separate concepts.

The following section looks at some of the work done in the area of computational models of discourse. Some of the important aspects of discourse structure are closely connected with markers of conjunction.

3.3 Computational Model of Discourse Processing

Computational linguists assume this connexity in text or discourse, and attempt to formulate appropriate forms of representation and processing methods to build systems that can "understand" (or "generate") different genres of text.

Analysis of discourse must include many different forms of linguistic information. Lexical meaning, morphology, syntactic structure and sentence semantics are all contributing factors to the overall communicative meaning of the text. Within a computational framework, an appropriate representation for each of these types of information forms the input to a processing system. A procedure (or set of procedures) that can operate on this representation to produce a representation of the discourse meaning forms the comprehension ability of the system. The current research in this area is still focussed on determining what information needs to be represented to capture discourse meaning.

The models will be considered from the perspective of discourse understanding, since the application of this work is to create from text, a representation of the structure of the discourse in an attempt to approximate the structure of the knowledge it embodies.

The model proposed by Grosz and Sidner (1986) includes three components: the linguistic structure, the attentional structure, and the intentional structure. These three interact in the processing of the utterances in a discourse. The model

attempts to provide a processing oriented framework in which referring expressions, cue phrases, and discourse purpose can be explained.

The linguistic structure is a partitioning of the discourse into Discourse Segments (DS). A crucial characteristic of a DS is a unique "topic", very like the idea of thematic paragraph. Grosz and Sidner view this as a reflection of "natural" segmenting based on their observations of general agreement between readers as to where a discourse can be partitioned. Although no specific claims are made about the size of these units, in the example texts they use, a sentence is typically the minimum size of a DS.

The intentional structure is made up of the "intentions" associated with the overall discourse and each particular segment. This roughly corresponds to the "writer's" purpose for conveying the particular information contained in each section. The whole discourse is assumed to have at least one Discourse Purpose (DP) and each DS is associated with a Discourse Segment Purpose (DSP). Each DSP is related to the DP and some other DSPs in a hierarchical structure. This structure indicates how satisfaction of DSPs contribute to the satisfaction of others in a "dominance" relation.

Another relation called "satisfaction-precedence" is proposed to capture ordering between DSPs. In Task Oriented Dialogues, one of the examples they use, two DSPs might be

1. "INTEND speaker (hearer (do something))" and
2. "INTEND speaker (hearer (know how to do something))"

It is intuitively clear that 2 must precede 1 (from our knowledge of "how things are"). Therefore, DSP2 "satisfaction precedes" DSP1.

Attentional structure, the third component of this model, is a dynamic structure represented as a push-down stack (Aho, Hopcroft & Ullman 1983). A Focus Space associated with each Discourse Segment records the objects, properties, relations, and purpose (DSP) of the segment. As a new Discourse Segment begins, its associated Focus Space is created and added to the top of the stack. If the new DSP is dominated by the DSP already on the stack, the new Focus Space will be added on top of those already on the stack. However, if the new DSP is not dominated by the DSP on the top of the stack, then the Focus Space on the stack will be removed before the new one is added. In this way, the intentional structure controls the state of the focus stack during processing. The attentional structure models the participants' focus of attention and is essential to the treatment of referring expressions.

The "dominance" relation in the intentional structure not only reflects the structure of the discourse, but also reflects the perceived connections between the propositional content of the component utterances. Two relations between propositions are given as examples: "supports" and "generates". The first is used in describing a logical argument and the second a task-oriented dialogue. Both the supports and generates relations reflect real-world connections perceived between propositions conveyed by each utterance.

In an argument, Discourse Segment Purposes are based on the author's intention that believing one proposition will help convince the reader of another. This is the relation called "supports" and it is the real-world relation on which the "dominance" hierarchy of the intentional structure is based. In the case of the task dialogue, the "generates" relation reflects the relation between steps in a

complex task. The individual steps which contribute to accomplishing a task are related by the relation called "generates". For example, "finding a screwdriver" is a step which contributes to accomplishing the task of "removing a set screw". For a task dialogue, therefore, the "generates" relation forms the basis for the "dominance" hierarchy.

The particular relation varies with the type of discourse, but the hierarchical nature of the "dominates" relation is the same. The parallelism between the "dominance" relation and propositional relations like "supports" and "generates" suggests that if the intentional structure, often marked by explicit linguistic cues, can be recognized, then we can infer the structure of the propositional content. This is then one step towards discovering the structure of the knowledge extracted from a discourse.

The knowledge structure identified in this way should be thought of as a high level description of the relation between complex propositions represented linguistically by clauses of the discourse. Whether the representation of knowledge is characterized as an associative network or production rules or some other formalism, there are necessarily connections between each of the elements in the representation. Through a discourse writers try to convey the relationships between objects, actions or states of affairs, usually depending on the reader to have basic knowledge about the topic. In this way, a discourse adds to the reader's knowledge by adding (or reinforcing) high-level connections. The structure among these propositions is part of the meaning derived by the reader and must be seen as part of what the reader can "know" from understanding a text. The relevant observation here is that the logical relations between propositions and the intentional structure of DSPs are essentially the same.

It is important to remember that a discourse structure does not exist a priori but is a dynamic structure that is created and changes as a discourse unfolds. Basic levels of linguistic processing (word identification, morphology, syntactic analysis) are not explicitly described in the Grosz-Sidner model; however, the results of these processes both contribute to and are supported by the discourse structure.

Cue words or phrases are an important factor in communicating to the reader the structure of the discourse and changes in it. These linguistic clues can signal changes in linguistic structure, attentional state, intentional state, and relations between DSPs such as Dominance and Satisfaction-precedes. Not all of the discourse structure is marked by cue words, or linguistic elements whose function is to indicate structure (*I must interrupt, in the first place, furthermore*). The clausal content of the linguistic elements augmented by the reader's knowledge about the objects and relations talked about implicitly indicates how the discourse is structured.

The model proposed by Grosz and Sidner (1986) provides a general framework for approaching the analysis of discourse. Separating the linguistic structure from intentional structure with the mediation of a focussing mechanism such as attentional structure allows the sources of information in a discourse to be treated independently. It also highlights the interaction between these sources of information and their symbiotic relationship. From their discussion of example discourses, the role of function words (which they have generally grouped under the name "cue words") as indicators of structural relationships is given a processing oriented description.

Work in the area of text generation (Matthiessen 1987, Mann & Thompson 1987 & 1986) has addressed the issue of representing connectivity in discourse. Matthiessen (1987) makes a distinction, originally due to Halliday & Hasan (1976: 26), between logical and experiential types within the ideational component of the linguistic system. The ideational component is "that part of the linguistic system which is concerned with the expression of 'content', ..." (26). Further, Matthiessen correlates this functional distinction with two modes of organization in the representation of discourse. The logical type correlates with a "sequential" organization and the experiential type with a functor/argument representation. Sequential organization is also a general organizing principle in text. Conjunction, explicitly marked or not, is the part of the cohesive system that reflects this organization. Rhetorical Structure Theory (Mann & Thompson 1987 & 1986) introduces the idea of rhetorical relations which connect spans of text. In this account, fifteen predicates are proposed to cover the possible types of relations between parts of a text. These predicates are often not signalled in text, but conjunctions are one of the explicit signals that can be used. Matthiessen suggests that this theory could support the role of conjunction in discourse.

3.4 Sublanguage Analysis

A sublanguage is defined as the language used by specialists in a restricted domain (Grishman and Kittredge 1986). The basic premise for examining sublanguages is that they are a restricted form of natural language in which selectional restrictions (determined by distributional analysis) reflect semantic classes of the domain. This methodology involves two crucial processes: that of

regularizing natural language statements through transformations (giving a common format to multiple linguistic forms that convey the same meaning); and, identifying subject-verb-object patterns of these regularized forms. Words in the sublanguage are grouped into classes, based on their occurrence in specific environments. For example, the words "X-ray", "film" and "scan" can all serve as subject in "_____ revealed a tumour" (Grishman et al. 1986:206) and so form a semantic class. This analysis has been applied in a number of different fields, including immunology (Harris et al. 1989) and medical reports (Sager et al. 1987). In addition to identifying word classes, characteristic statements (or formulae) of a sublanguage can be identified by examining the patterns of word classes which occur in "fact statements". That is, what are the characteristic patterns of subject-verb-object in terms of the sublanguage word classes associated with each element (subject, verb, and object).

The treatment of conjunctions in this framework has been varied. In some cases, conjunctions are collapsed as features of predications, other times they are treated as functors taking domain "fact-statements" (to use Harris' terminology) as arguments. Recalling the rules shown previously, we can see that the "fact-statements" identified in these analyses will likely play a role as objects in a knowledge base. The operators representing conjunctions can be seen to capture some of the "model" level descriptions (time, cause, etc.) by providing links between pairs (or groups) of objects.

Although little reference to this work is found in the literature on knowledge acquisition, the results of these analyses would appear to capture the same kind of information about a domain that is required in the development of an expert

system knowledge base. The word classes are based on semantic distinctions and correspond to basic concepts in the domain. Sager et al. (1987) describe the use of sublanguage analysis to derive a structured "frame-like" representation of narrative medical reports that can be a database component of an enquiry/information retrieval system. The word classes and statement formulae could also be utilized as "objects" in a knowledge base format as described in chapter 2.

Connectives have not been treated in depth, in such discussions, although connective relations are identified as an important element of the final representation. The essential observation, reflective of other linguistic studies, is that the connectives establish relationships between situations or events expressed by the connected clauses. Grammatical type distinctions are identified in relation to medical reports (Friedman 1987) and a classification of temporal connectives is used to generate a "time graph", along with other linguistic clues such as tense and aspect (Johnson 1987). The "time graph" provides the information necessary to respond to time-related queries. The "time relation classes" mentioned by Johnson (1987:184) suggest a general approach to temporal connectives that is similar to the approach taken in the following work.

3.5 Summary

This chapter has reviewed recent work in several areas of linguistics, all related to various issues in the processing and representation of connections in discourse. In formal linguistics, cohesion and cohesive relations have been studied with a view to identifying the semantic distinctions made by subordinating conjunctions

as well as other linguistic devices. Descriptive grammars like that of Quirk et al. (1972) also make this same type of distinction. Morrow (1986) and Rudolf (1988) approach connective relations from a cognitive perspective, characterizing connectives as signals that serve to guide comprehension processing. de Beaugrande and Dressler (1981) and van Dijk (1980), working in the field of discourse analysis, both address issues of what form the representation of a discourse might take. In their work, correspondence between surface elements and underlying units of representation is examined. The relationships between units are described as "links". Computational linguistics in both discourse processing (Grosz and Sidner 1986) and sublanguage analysis (Grishman and Kittredge 1986) also focus on representation issues drawing on work from other areas of linguistics.

The basic form of representation assumed as the basis for the analysis of conjunctions presented in the next chapter is drawn from the work cited above. It is assumed that clauses correspond to distinct units in the discourse representation. These units will be called objects. Connections between objects are described as links. The overall discourse representation, thus, takes the form of a network of objects.

The view taken in this thesis which forms the basis for the analysis described in the following chapter is that the discourse network is, in fact, comprised of a number of different sets of links. Each set of links represents a different semantic relationship, like time or cause. These sets of links are a different level of organization in the discourse. What is common to all levels is that the relationships are all represented as orderings on the included objects.

This ordering relation is also a major concern addressed in the area of knowledge representation. Rules in a knowledge base implicitly link LHS conditions (or objects) to the RHS conditions. The strategies used for rule selection in the reasoning procedures effectively impose an ordering among the RHS conditions, and so, implicitly among the LHS conditions as well. The kind of relationship implied between LHS and RHS objects is not explicitly represented. That is, a LHS condition which specifies a temporal contingency is not distinguished from one that specifies a conditional contingency. It is only by examining the descriptive label that a person can determine, based on their own general knowledge, what contingency is implied. Grosz and Sidner (1986) seem to imply a similar view that relations between propositional content in different types of discourse are basically of the same form, distinguished in name only (i.e. "supports" and "generated"). The particular name associated with the relation is dependent on the subject matter of the particular discourse.

Combining these various observations, it is proposed that the connecting function of conjunctions can be interpreted as defining an ordering relation between objects, specifying how a new object is linked to the discourse representation. In itself, the conjunction does not provide enough information to uniquely determine what type of ordering should be represented. Each conjunction does, however, entail a specific order between objects representing the main and subordinate clause. A classification of the conjunctions occurring in the sample document is presented in the next chapter.

Chapter IV

ANALYSIS OF THE SAMPLE DOCUMENT

Collectively, the observations and conclusions about the roles of connectives such as conjunctions and prepositions suggest that they serve to signal relations of cause, contingency, temporal order, and comparison among the clauses of a text. Similarly, structural markings provided by headings and numbering systems provide an additional means by which a writer can express the hierarchical organization of the document's content. Thus, both document layout and grammatical connectives are used to encode some aspects of a document's logical organization and provide sources of "information" about the structure of the text's interpretation.

In this chapter, a method for interpreting and representing conjunctions will be presented. A processing oriented approach is taken in the analysis. A discourse representation is seen as a dynamic structure which is built through comprehension processes following Grosz and Sidner (1986). It is assumed that individual clauses correspond to distinct units in the discourse representation, an idea common to many researchers in the area of discourse analysis, among them Kintsch (1988) and van Dijk (1980). Conjunctions are seen as signaling relationships between the units of representation, and thus, their interpretation is crucial to discourse comprehension.

Bylaw No. 87-248 of the City of Victoria, British Columbia (1987) is the sample document analysed in this chapter. This Bylaw sets out conditions which must be met by the operators and users of parking lots in the City of Victoria. These conditions and the relationships between them must encoded in the discourse representation. Examination of the bylaw suggests that the text can be segmented into sentence, clause and phrase size units corresponding to conditions that must be represented. In the discourse representation, these units will be called objects. The relations among these conditions may be causal, contingent, and/or temporal and these are frequently marked in the text by explicit connectives and/or layout distinctions. Each of the relationships will also have to be included in the discourse representation as connections between objects.

The knowledge base for an expert system based, in whole or in part, on this document will also include this same information. Using the terminology of the ACQUIRE system, the conditions will be objects in the knowledge base. The connections between them are encoded in the support links of each object. Thus, the final discourse representation can be used to generate a knowledge base. And indeed, the data structures of the knowledge base have been used as a model for representing the discourse structure.

All of the relationships between objects indicate an ordering among the objects that must be captured and encoded. No attempt has been made to encode the type of relationship; only the direction of the connection is addressed, for this is the function which is common to all of the connectives considered here. The ordering among objects provides the structural form of the discourse interpretation. In a knowledge base, this ordering among objects represents the

order in which they must be considered when the knowledge base is used by inference procedures. The proposed analysis of conjunctions is applied to automatically derive these links between objects.

In the following sections, the analysis of conjunctions will be presented first. Then, an overview of the processing method implemented using this analysis is provided. The last sections provide examples from the sample Bylaw to illustrate each of the stages of processing. A description of the programs used to implement a prototype system is included in Appendix B.

4.1 Analysis of Conjunctions

It is proposed that conjunctions can be split into three groups, based on the ordering they indicate between subordinate and main clause. Figure 2 lists all the conjunctions and prepositions used in the analysis of the Bylaw and the ordering relation they signal.

Pre-Ordered	Post-Ordered	Parallel-Ordered
after	before	
where	until	
unless	upon	
except	notwithstanding	
if		and
as		or
without		

Figure 2: Function Words Classified by Direction of Contingency

In Figure 2, the headings "Pre-Ordered", "Post-ordered" and "Parallel-Ordered" indicate the ordering between subordinate and main clause that is entailed by each conjunction.

Those conjunctions listed under "Pre-Ordered" are those which specify that the content of the subordinate clause precedes, or must be considered before, that of the main clause. For example, in the following sentence, taken from the sample Bylaw, *where* marks the subordinate clause.

Example 1: Subsection 10. (2)

> "(2) [Where any parking space on a licensed parking lot is equipped with a parking meter], [no person shall park a vehicle within such parking space] [without having deposited the appropriate fee for parking in the manner and at the rate prescribed or measured by the meter]."

The condition expressed in this clause must be evaluated to determine whether or not the main clause need be considered. Therefore, this conjunction is placed in the "pre-ordered" category. In the same way, *without* indicates that the subordinate clause expresses a pre-condition for its main clause.

Each of the conjunctions in this category will generate the same structural relation between objects in the discourse representation. Regardless of the basis for the ordering (i.e. time, cause, location) of objects which correspond to each clause, the direction of the links between them will be the same. The subordinate clause will precede the object representing the main clause. Graphically, this can be illustrated by connecting the subordinate clause object below that representing the main clause. In terms of the knowledge base, this means that the subordinate clause supports the main clause.

The particular ordering related to each lexical form, independent of its semantic category is illustrated by a number of conjunctions which belong to more than one such category. The conjunction *where* can indicate either a locational relationship or a conditional relationship depending on the content of its clause.

When a conditional relationship is indicated, *where* takes on the meaning *in cases where* ... (Quirk et al. 1972:745). However, whichever meaning is appropriate, the ordering relation between the clauses will be the same. The *where* clause expresses a condition which must be met before the main clause should be considered. In this example, the relationship is clearly conditional. An example that shows the same ordering based on a locational relationship might be:

> "A protective shield must be installed where the intake valve is connected."

"Post-Ordered" conjunctions are those which specify that the content of the subordinate clause follows that of the main clause in the logical sequence. The following example from the sample Bylaw illustrates this relationship.

Example 2: Subsection 4. (2)

> "4. (1) .....
>
> (2) [Notwithstanding the provisions of subsection (1)], [no certificate as to screening is necessary in respect of any side of a parking lot constituting a boundary with an adjoining lot] [where the elevation of such parking lot is at least 2 m lower at such boundary than the finished elevation of the adjoining parking lot]."

In this case, the main clause provides an exception to the requirements specified in the prepositional phrase. Therefore, reasoning must proceed from the main clause, *no certificate as to....*, first, and only then the content of the phrase *the provisions of subsection (1)* should be evaluated. Therefore, this preposition or conjunction is placed in the "post-ordered" category. In the discourse representation, the object for the *notwithstanding* phrase will follow the main clause object and this will be illustrated by placing the former object above the latter. The phrase marked by *notwithstanding* will thus be supported by the main clause in the knowledge base.

This example also shows the type of prepositional phrase that has been treated as equivalent to a subordinate clause in this analysis. These phrases are often equivalent in meaning with subordinate clauses through insertion of a verb (Quirk et al. 1972: 733). In this case, the phrase could be replaced by *Notwithstanding the provisions the have been specified in subsection (1).* A number of other conjunctions also function as prepositions in this way. Some examples are *because (of)*, *before*, and *after*.

The third category, "parallel-ordered", includes the coordinating conjunctions *and* and *or*. This category of conjunction will generate a structure in which neither of the clauses is superior to the other. Rather the relationship between them exists by virtue of their relationship to the object representing the sentence (in this case a subsection as well) as a whole. Thus, the objects in the discourse representation are not directly linked and neither object in the knowledge base supports the other.

The semantic classifications suggested by Halliday and Hasan (1976), Rudolf (1988), Martin (1983) or Quirk et al. (1972) have not been considered in this analysis. It is recognized that a complete representation of any discourse must involve the information conveyed by the kinds of distinctions that these classifications attempt to capture. However, in this work, the common role of all connectives as imposing an abstract ordering of concepts has been the major concern. The semantic distinctions such as time, cause, or location can be seen as information which would be used to include each link in the appropriate set of links or model within the representation (Gaines 1987, Johnson 1987). The connective itself does not, however, completely determine in which model(s) the

link should be included. The semantic category of the connective will interact with the content of the linked clauses to make this determination.

The application of the above analysis is the subject of the rest of this chapter. The next section provides an overview of the proposed processing method which is then explained in more detail in the last sections.

4.2 Overview

Knowledge acquisition for expert systems has been described in chapter 2 as the process of identifying key concepts in a particular domain and the relationships that hold between them. Specifically, in the ACQUIRE knowledge acquisition system, the key concepts are represented by objects. The relationships between objects are expressed as rules. Each object description includes link fields which specify the object's place in a support network. This network summarizes the interconnection among objects expressed in all of the rules. The first step in the knowledge acquisition process is to define the objects, including their support links, that represent the domain knowledge.

The knowledge representation used by Acquired Intelligence, Inc. is a production rule system. Production rules are IF-THEN statements, where the values of symbolic "variables" in the condition (IF) part are evaluated and values conditionally are assigned to other symbolic "variables" in the action (THEN) part. The symbolic "variables" are called objects in this system. Each object has a set of possible values and represents an entity, action or state of affairs in the knowledge domain. The rules represent decisions made in reasoning about the domain. Collectively the rules in a knowledge base define a decision network.

This thesis will focus on identifying segments of a text which will likely embody "concepts" that must be represented as symbolic variables in the knowledge base, and where possible, determine the form of rules involving these variables.

Some concepts, or objects as they are called in the terminology of ACQUIRE, can be identified by structural features of a document and will be taken to represent "high-level" objects in the support network. The smallest units of text considered are clauses and a restricted number of prepositional phrases. The objective is to proceed top-down in creating a support network amongst objects whose meaning is reflected in segments of the text. The support network summarizes all support relationships amongst objects. The rules necessary to complete the knowledge base specify the relationships between the actual values of the objects. Thus, support links between several objects may lead to several different rules, depending on domain specific information (see Section 2.4). However, the support network constitutes a skeleton knowledge base in which basic objects and their relationships are already specified, suitable for further refinement by a domain expert or knowledge engineer. In addition, by making the connections between clauses explicit, this representation might prove useful for automatic procedures involving distributional analysis of terms.

At the same time, links inserted in the text provide on-line access to the text of the document for the developer and for the end-users of the system. In the first case, access to the text is a valuable aid to refining the automatically generated structure. End-users of the system will have access to the document for their own reference or as an "explanation" facility. The wording of the official document from which the expert system has been derived can provide a familiar framework to assist system users understand their interaction with the system.

The aim of the project described below has been to apply the analysis of connective relations described above in a procedure to automatically extract a set of object descriptions from an on-line document. To do so, we will identify salient text segments and use the relationships among them to build a network of objects. It is hypothesized that in the formal, regulatory documents that are the specific type of text addressed, the identified segments will correspond to concepts that must be part of the domain knowledge base. In the ACQUIRE system, by mapping the text segments, or concepts, to objects and the relationships between them to support links, an **intermediate text representation** can be created. This representation will be a first approximation of the knowledge base.

This process should not be viewed as "transforming" a text into a knowledge base, but rather as creating a structured text representation which could be implemented in a hypertext system (Conklin 1987). This independent representation may then be linked to a separate and distinct knowledge representation or knowledge base. This is shown schematically in Figure 3.

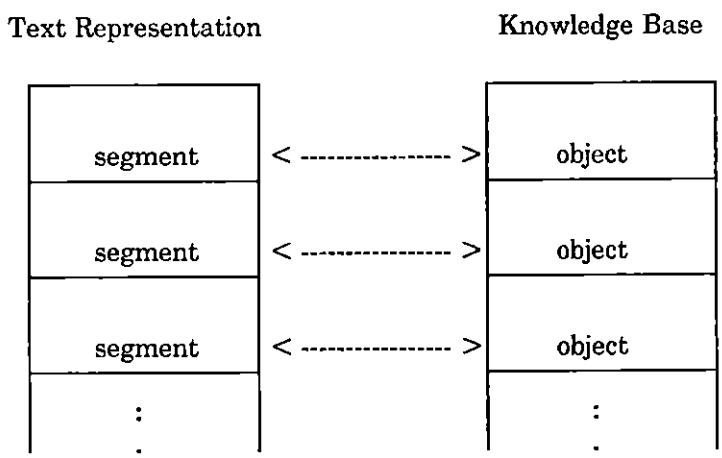


Figure 3: Relation between Intermediate Text Representation and Knowledge Base.

Both of these structures will initially have essentially the "same" structure. However, the knowledge base created in this way will clearly be neither complete nor entirely accurate at this stage. Other information that would be necessary in a complete expert system knowledge base would be: how strictly conditions are enforced, who is responsible for enforcement, and what paperwork is required. This information must be elicited from the people who actually handle bylaw enforcement, that is, the domain experts. The intermediate knowledge base will undergo considerable revision by developers and/or domain experts as changes and additions are made to this intermediate structure. Having a separate text representation leaves open the possibility that links between objects and text segments can be maintained when either the document or knowledge base is edited (although this topic is not discussed here).

In the following discussion, the characteristics of the document layout are addressed first along with a discussion of how they contribute to structuring the document's content. Then, the actual language used in the bylaw is addressed. This second part of the discussion focuses on those linguistic features which are immediately useful in identifying relevant concepts or objects without recourse to a pre-existing representation of a domain lexicon or "world knowledge". For this reason, our analysis has focussed on function words like conjunctions and prepositions which are commonly used in formal documents and have a reasonably consistent meaning across many domains. The relatively frequent use of connectives in this discourse style provides a rich source of information that can be used to establish the direction of connections between the concepts represented by the clauses or phrases.

These two types of characteristics, document layout and linguistic structure, of the sample Bylaw are discussed separately because of their different nature. Document layout characteristics are visual cues to human comprehension imposed on the linguistic content of the document. Many types of text, like most narratives, lack the wealth of document layout features that are exhibited in our sample document. However, this research is specifically concerned with official, regulatory document which are characteristically highly structured. Therefore, we have taken advantage of the information provided by these visual features.

In this processing model, the document format characteristics are used to provide the basis for linguistic interpretation. That is, the segmentation indicated by the document layout is done first and then serves to guide the interpretation of the linguistic structure.

4.3 Document Layout

Examples from the sample Bylaw are used in the following discussion of the structural description derivable from typographic layout of a document. The structure derived from the document layout features will be called the **document structure representation**, or more simply, the **document structure**. This representation is one "view" of the input text which captures the logical segmentation of the document. The additional information derived from the linguistic features (Section 4.4) will be added to this **document structure** to create what will be called the **intermediate text representation**.

4.3.1 Typographical Features

The following is an excerpt from the bylaw. The complete text is included in Appendix A.

1. This bylaw may be cited as the "PARKING LOT BYLAW".

2. In this bylaw

   "vehicle" has the meaning assigned to it in the Motor Vehicle Act;

   "parking lot" means a place, on one parcel of land, which is used or set aside for use for the parking of one or more vehicles in consideration of the payment of money.

3. No person shall operate a parking lot unless he holds a valid and subsisting licence for it, issued under the provisions of this bylaw and of the Business Licence Bylaw.

4. (1) No licence for a parking lot shall be issued unless and until the City Engineer certifies:

   (a) That the surface area of the parking lot has been completely paved and is adequately drained;

   (b) where the parking lot is in or adjoining an area zoned by bylaw or lawfully used for residential use, that it is screened from adjoining parcels of land either by evergreen hedges or by view obscuring fences or both and that such hedges or fences are of a height of not less than 1.3 m and, for fences, not more than 2 m, along the common boundaries of such adjoining properties and of the parking lot;

   (c) where the parking lot abuts on a street, that it is screened along its entire street boundary, except for necessary vehicular access points, either by an evergreen hedge or shrubs or by permanent masonry planters with plants growing in them, or by both methods, in such a manner as to provide an effective screen of the parking lot along all street boundaries and of a height of at least 1.3 m above ground level;

(d) that all lighting used to illuminate the parking lot is deflected from adjoining lots and streets; and

(e) that there is only one sign, not exceeding 2 m^2 in area, at each entrance and at each exit, and that such sign does not contain any words or signs other than to designate entrances, exits, conditions of use of the parking lot, the name of the parking lot and conditions relating to the towing away of vehicles.

(2) Notwithstanding the provisions of subsection (1), no certificate as to screening is necessary in respect of any side of a parking lot constituting a boundary with an adjoining lot where the elevation of such parking lot is at least 2 m lower at such boundary than the finished elevation of the adjoining parking lot.

(3) Where the provisions of subsection (2) apply the City Engineer may stipulate any modifications of the screening requirements as may be necessary to conform to zoning bylaws and traffic bylaws in respect to safety.

5. . . . " (Victoria 1987)

Figure 4: Excerpt from Bylaw 87-248, City of Victoria

The typographical layout used in this document provides many visual cues which help readers in identifying the organization of its content. Drafters have used numbering or labelling, in conjunction with punctuation, indentation and spacing to indicate logical segmentation of the document. For example, labels which are Arabic numbers followed by a period, like 1.,2.,3., etc., indicate the beginning of a section of the bylaw. These sections are further marked by extra spacing, both before and after the section's text. The text of the section is aligned at the leftmost indentation point. Each of these layout features provides a visually prominent indication of the extent of the segment.

Each section in the Bylaw addresses a specific topic relevant to the operation or use of parking lots. It is possible to distinguish different functions for some of the sections. For example, section 1. simply provides the "name" of the Bylaw. Section 2. lists the definition of important terms used in the rest of the sections. The remaining sections of the Bylaw, like 3. and 4. shown above, stipulate conditions on specific aspects of parking lot operation or use. In this project, no attempt has been made to identify or make use of these functional distinctions. However, because these distinctions are conventionally used in the presentation of regulatory documents, they could be profitably utilized. For example, recognition of the name of the Bylaw would be extremely important if an attempt were to be made to incorporate all, or even a few, Bylaws in a single representation. Also, any lexical analysis would be aided by having a list of important terms and their definitions available.

Each of these sections will be represented as a node in the **document structure** representation. These nodes will be directly linked to a node representing the whole document in a hierarchical relation. The nodes in the **document structure** represent segments of the document. Since no typographical features indicate any further grouping, the document structure derived for these segments can be represented by the tree diagram shown in Figure 5.

In section 3.,there is no further segmentation indicated by the typographical layout. Section 4., however, is divided into a number of subsections. The beginning of each subsection is labelled by an Arabic number enclosed in parentheses. In this case, the labels are (1), (2) and (3). The change in style of labelling indicates the beginning of a new segment in the text and a new

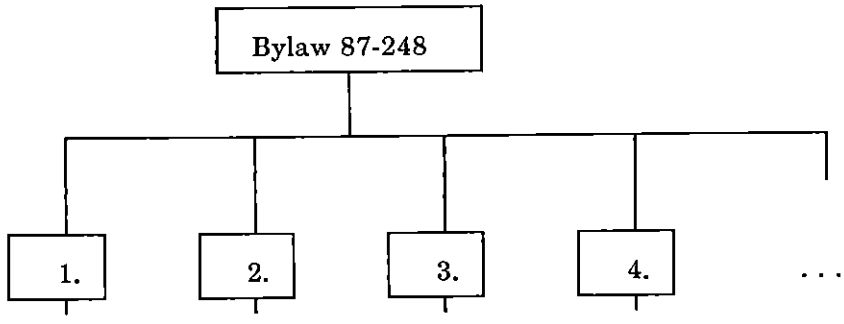


Figure 5: Structure of Bylaw Sections

grouping of segments. The numbers themselves explicitly suggest (to the human reader who is familiar with the order relation between the symbols "1"; "2", etc.) an ordered sequence among these units. Subsections labels begin again at the start of the numeric sequence and, thereby, indicate an interruption in the ordering between segments.

The hierarchical, or subset, relation of these new sections is visually emphasized by indentation. The subsection label is indented relative to the section labels. The text of the subsection is indented further to the right than the text wholly contained in a section (as in 3.). The first level structure of this section is graphically illustrated in Figure 6

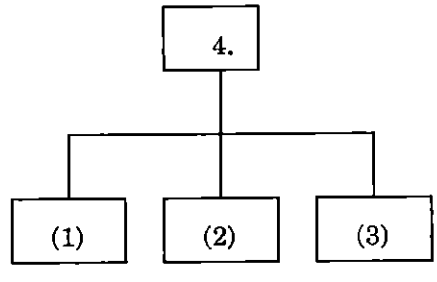


Figure 6: Structure of Bylaw Section 4

The third level of segmentation is labelled by lower case alphabetic characters enclosed in parentheses (for example (a),(b), etc.). The same indentation and spacing used to distinguish subsections from sections are used in this case to distinguish clauses (or "list" items) from subsections. In addition, punctuation between the clauses reinforces, even more, the subordinate nature of these segments. Unlike sections and subsections which are terminated by periods, the clauses (except the last) are all terminated by semi-colons.

These observations will seem "obvious" because, as skilled readers, we have all learned the conventions used in printed publications and are not usually aware of using this source of information. However, if all section numbering, indentation and spacing were removed from the document, the result would be far less easily understood. In this project, these typographical features are used to automatically build the document structure representation which will serve as the basis for the balance of the analysis.

The initial data is in the form of an ASCII file containing a print image of the Bylaw. The clause markers discussed below are included in the text. The first program in the prototype system removes all blank lines, leading blanks and segment labels (1.,a), etc.). In their place, Standard Generalized Markup Language (SGML) style tags are inserted in place of each segment label. The details of this procedure are outlined in Appendix B.

Many documents created on-line are already marked with codes equivalent to the SGML tags used here. However, documents which are not on-line can be captured by the use of an Optical Character Reader. In this case, or where the document creation language does not provide sufficient marking of document

segments, the suggested procedure would be a necessary step in the document analysis.

The structure of the first four sections of the sample Bylaw can be graphically represented as in Figure 7.

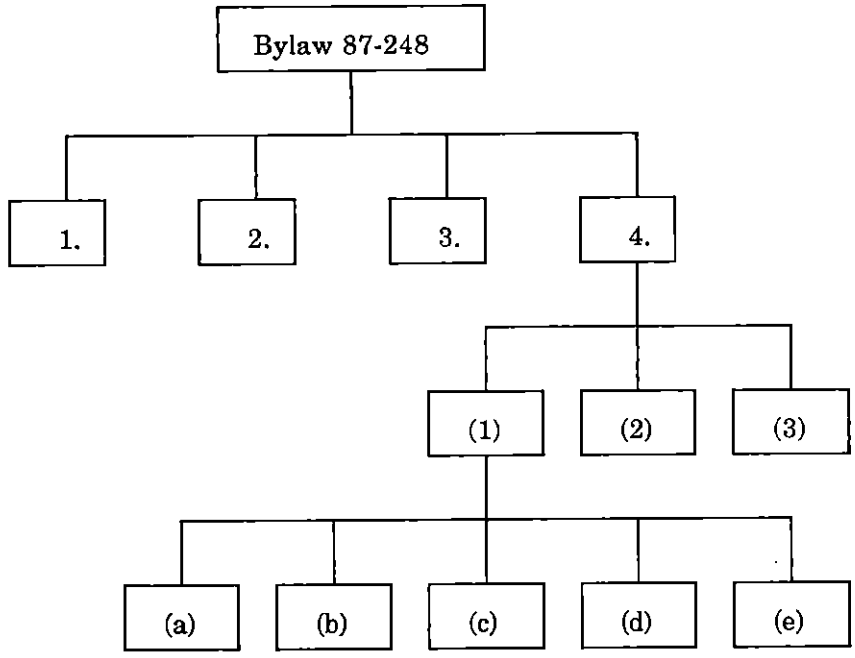


Figure 7: Bylaw Document Structure

The physical form of the document imposes this strict hierarchy which can be viewed as a tree structure. Terminal nodes, or leaves, of the tree represent document segments which are not further subdivided and are directly associated with continuous portions of the actual text. Internal nodes represent groupings of the segments. These nodes are associated with portions of text through links with the nodes they contain. The **document structure** is important for both further analysis and for the maintenance of links between the text and the knowledge base.

4.3.2 Interpretation of the Document Structure

The strictly hierarchical structure of the document components is a reflection of the strict sequential ordering imposed by the presentation medium in the original document. This structure can be graphically represented as a tree. The graphic representation embodies a composed-of relation between a node and its subordinate nodes. For example, take the following excerpt from the Victoria Parking Bylaw.

Example 3: Section 10.

> "10. (1) Where parking spaces on a licensed parking lot are clearly delineated by painted lines or barriers, no person shall park a vehicle on such parking lot, except in such parking spaces, and no person shall park a vehicle in such a manner as to straddle the line between two parking spaces."
>
> (2) Where any parking space on a licensed parking lot is equipped with a parking meter, no person shall park a vehicle within such parking space without having deposited the appropriate fee for parking in the manner and at the rate prescribed or measured by the meter."

The document structure will represent the section (10.) and its two subsections as distinct components with the two subsections contained in the section as shown in Figure 8. Section 10. is composed of subsections (1) and (2). Equally, both subsection (1) and (2) are in an element-of relation with Section 10.

In order to use the **document structure** to create a knowledge base, the physically defined structure must be interpreted in terms of objects and support. The interpretation used here equates each document component, or node in the tree, with an object in the knowledge base. The composed-of and element-of relationships, represented by branches in the tree, are then equivalent to the support links. The physically defined composed-of relation will be interpreted as

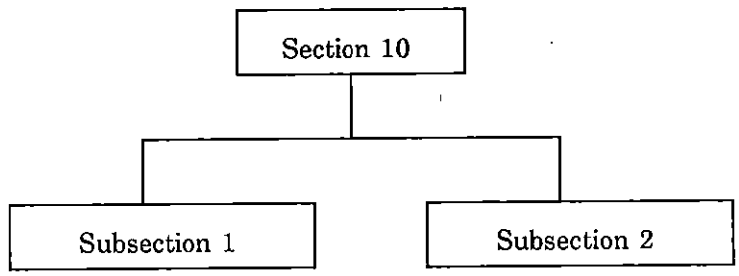


Figure 8: Document Structure - Section 10.

indicating that the dominating object is supported by the subordinate object(s). So, in Figure 8, the object, "Section 10." is supported by both "subsection 1" and "subsection 2" objects

The relationship between a document component and those subordinate to it often, though not necessarily, reflects logical relations which should be included in the text representation. Therefore, we can directly map the hierarchical relations of the document structure into relations between corresponding nodes in the text representation. That is, the composed-of relation in the document representation will become the support-from relation in the text representation. Similarly, the element-of relation will become the support to relation. The links in the document representation thus provide information about the probable structure of the **text representation**. This will not always yield an accurate description of the logical connections between document components; however, in a significant number of cases it does.

Each document component described above has a distinct format, sequential labelling, indentation, and spacing. These format distinctions are used by document writers to help readers organize their understanding of the document's content. Therefore, where the format indicates a division of the document into

subcomponents, we will assume that a corresponding component in the text representation is justified.

In this document, each section component comprises exactly one sentence, unless it contains subsections. Subsections all contain exactly one sentence. Whatever the status of the "sentence" as a linguistic unit, in written discourse the boundaries of sentences are explicitly and unambiguously marked by punctuation. Grouping ideas into complex sentences demonstrates the author's intention that those ideas are closely connected. We assume that authors of public documents intend to express correct and accurate information. Therefore, we will take this characteristic of the sections and subsections as additional justification for identifying each as a node in the text representation network.

Initially, this hierarchical structure will constitute the **intermediate text representation.** Each document component will map directly to a node in the text network and the document structure links will correspond to the support links between them. In this case, the nodes representing Subsections 1 and 2 will both have a support-to link with the Section 10 node and Section 10 will have support-from link with both Subsections 1 and 2. The next section describes how the **intermediate text representation** is further refined.

4.4 Intermediate Text Representation

The default text representation that is derived from the document structure can be both extended and revised by utilizing signals that are contained in the linguistic realization of each component. Explicitly marked adverbial prepositional phrases and subordinate clauses, can be used to further divide the

lowest level document components (leaves on the tree) into separate text components and establish appropriate links between them. Explicit references to document components can also be used to prevent the duplication in text nodes and correctly link potentially non-adjacent document components.

The **intermediate text representation** is a network identifying salient textual components as nodes and the relationships between these components as bi-directional links. Textual components are defined as contiguous portions of a text whose interpretations represent decision points in reasoning about the text's knowledge domain. Unlike the document structure, the text representation is not necessarily hierarchical and cannot be modeled as a tree structure. Instead, a network provides a more accurate description of this **intermediate text representation.**

The hierarchical organization of a tree means that a node may be linked to only one node higher in the tree, although it may link to several nodes below itself. This restriction is reflected in the terminology often used to describe directly linked nodes as mother and daughter, where the mother node is higher in the tree than the daughter. A mother may have several daughters but only one mother.

The **text representation** will not have this restriction on the links between objects or nodes. It has been pointed out previously that there may be many sets of links between objects, each representing a different model or view of the discourse. Thus any object can be linked to any number of other objects either higher or lower in the structure. This kind of organization is described as a network.

This representation attempts to identify segments of the text which can be easily interpreted by people as decision points in a reasoning network. The analysis does not attempt to establish the "meaning" of each segment, but only derives the ordering imposed by the logical contingency between them. Thus, the network represents only the ordering among the identified decision points, not the specific content. The developer or experts who will use this representation are active participants in the system and they will be responsible for attributing the "meaning" to each segment.

4.4.1 Clause Structure

Complex sentences provide a structural mechanism for expressing the connection between related concepts. The complexity of a sentence is dependent on the stylistic choices of the writer, but the reason for the choice is not of concern here. The relevant observation is simply that complex sentences are used extensively in formal documents such as that addressed in this study. Therefore, the structural characteristics of these sentences can be exploited to derive a representation of the logical ordering of concepts related to the structural components.

For example, Section 3 of the Bylaw, shown below, is one of the document components that can be further subdivided on the basis of clause structure.

Example 4: Section 3.

> "3. [No person shall operate a parking lot] [unless he holds a valid and subsisting licence for it, issued under the provisions of this bylaw and of the Business Licence Bylaw]."

In this example, the square brackets indicate the major clause breaks in the sentence. The two clauses both express concepts that are crucial to the knowledge structure for this domain. *No person shall operate a parking lot* clearly includes the concept of operating a parking lot which is one of the top level concepts that the target knowledge base must include. The subordinate clause, *unless he holds a valid and subsisting licence ...*, also includes reference to an important concept, that of holding a licence. These two concepts are directly related in terms of reasoning about this domain of parking lot operation. That is, in order to establish whether *a person can operate a parking lot* it is necessary to determine if *he holds a valid licence.* This relationship is represented in a knowledge base through support links between objects. These links must indicate that the object, *he holds a valid licence*, supports the object, *a (this) person can operate a parking lot.*

It is not necessary to consider the meaning of the two clauses to establish this relationship as long as we assume that the writer is presenting the content in a truthful and accurate way. It is sufficient to recognize the clausal divisions in the sentence to identify new objects.

In the construction process, a new object will be generated for each marked clause. Thus, structural form of the text is interpreted as marking units of the text that correspond to units of the discourse representation. The direction of the link between these two objects will be determined by the particular conjunction introducing the subordinate clause.

Although no automatic syntactic analysis is attempted in this project, one can see how the syntactic structures act as discourse signals to indicate

connections between clauses. Since we need to recognize phrasal boundaries, these crucial divisions have been inserted by hand. The clause boundaries that were marked, and thus used in further analysis, are as follows:

- Subordinate adverbial clauses explicitly marked by a conjunction,
- Verbal constituents conjoined by *and* and *or*,
- Preposed prepositional phrases.

The next section describes how the conjunctions are used to establish the support links between objects.

4.4.2 Conjunction

The subordinate clause in Section 3., introduced by *unless*, expresses a condition for determining the status of the proposition expressed in the main clause. (See Example 4 on page 68.) That is, holding an appropriate licence is a condition for operating a parking lot. If we consider how these two clauses are used in reasoning about this domain, it is clear that the value of the *unless* clause, *he holds a valid and subsisting licence ...*, supports whatever conclusion can be made about the main clause, *no person shall operate a parking lot.* That is, it is necessary to make some conclusion about *holding a licence* before *operating a parking lot* can be determined. Thus, *unless* is a member of the category called "pre-ordered" as described in Section 4.1.

In this item, the syntactic realization divides the sentence into two clauses. The subordinating conjunction *unless* explicitly marks the subordinate clause functioning as an adverbial clause of condition (Quirk et al., 1972). *Unless* expresses a conditional relation in which the subordinate clause states a condition which must be considered in establishing the meaning (or consequence)

of the main clause. In this case, if we are reasoning about parking lot operation (content of the main clause), then the situation represented by the subordinate clause must be considered before or, in order that, the "value" of the main clause can be determined.

In the text network, this relation can be captured by establishing a support to link from the node representing the subordinate clause to the node representing the main clause. The inverse relation is captured with a support from link from the main node to the subordinate node. This will result in the configuration shown in Figure 9. Since these links are always bi-directional, only a single line will be used to indicate the links between nodes in the diagrams. The physical placement on the page in which one object appears above another will serve to indicate the direction of links. That is, support-to links are always pointing upwards and support-from links point towards the bottom of the page.

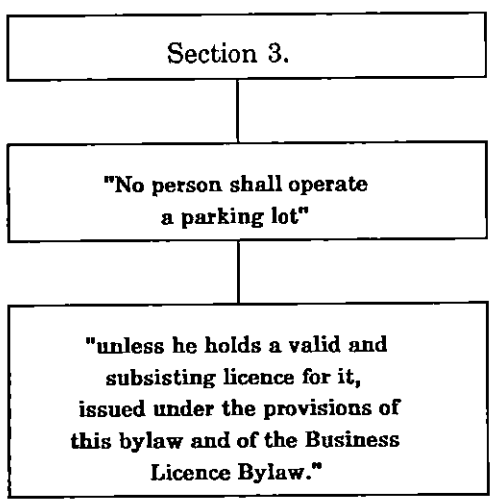


Figure 9: Structure of Section 3.

The actual interpretation of each clause that is suggested above is only implicit in this representation. The nodes themselves are simply symbolic entities. An interpretation is attributed to a node only by the system's users: developers, experts, or others. Therefore, the clauses themselves will be used as descriptive labels for the nodes, so that they can be readily interpreted. The significance of the links themselves is represented in part through their use by the reasoning procedures. These procedures do not directly consider what kind of link is represented: only the sequence of connections is important. However, the conjunctions themselves remain as part of the descriptive labels so that this information will be available to the system developers.

Other conjunctions which have the semantic force of temporal sequence, cause, or condition impose the same kind of abstract ordering on the situations described by clauses. Two such conjunctions are *where* and *without.* Each of these conjunctions is a member of the "pre-ordered" category and indicates that the associated phrase or clause is in a supporting relation to the clause it modifies. For example, both of these conjunctions appear in the following subsection (10.(2)) of the Bylaw.

Example 5: Subsection 10. (2)

> "(2) [Where any parking space on a licenced parking lot is equipped with a parking meter], [no person shall park a vehicle within such parking space] [without having deposited the appropriate fee for parking in the manner and at the rate prescribed or measured by the meter]."

The *where* clause expresses a condition which must be met before the main clause should be considered. *Without* imposes the same ordering between its clause and the main clause. Therefore, the structure shown in Figure 10 is derived from the text of subsection 10.(2).

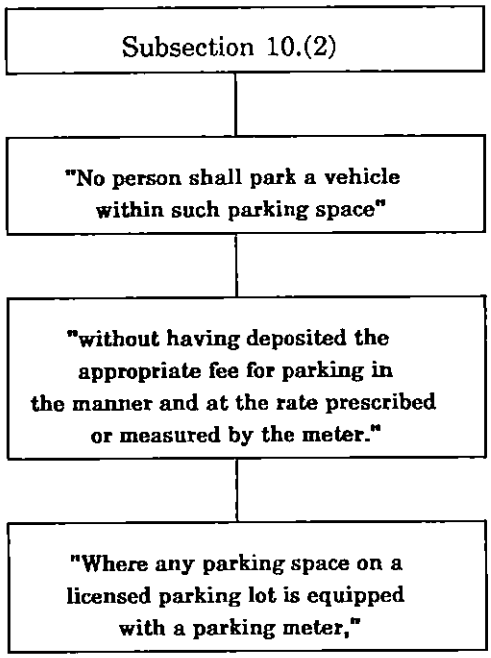


Figure 10: Structure of Subsection 10.(2)

Notwithstanding is a connective that also signals that a further division in the textual content should be made. This is an example of the type of prepositional phrase that has been treated as equivalent to subordinate clauses.

Unlike the preceding examples, the opposite ordering of clauses is indicated by *notwithstanding* since it is a member of the "post-ordered" category. The *notwithstanding* phrase or clause is supported by the main clause, rather than supporting it. Thus, it is an example of the category of conjunctions called "post-ordered". For example, subsection 4.(2).

In this case, the main clause provides an exception to the requirements specified in the prepositional phrase. Therefore, reasoning must proceed from the *no certificate* .... clause first, and then to *the provisions of subsection (1)*. The structure generated from this section is shown in Figure 11.

Example 6: Section 4.

"4. (1) .....

(2) [Notwithstanding the provisions of subsection (1)], [no certificate as to screening is necessary in respect of any side of a parking lot constituting a boundary with an adjoining lot] [where the elevation of such parking lot is at least 2 m lower at such boundary than the finished elevation of the adjoining parking lot]."

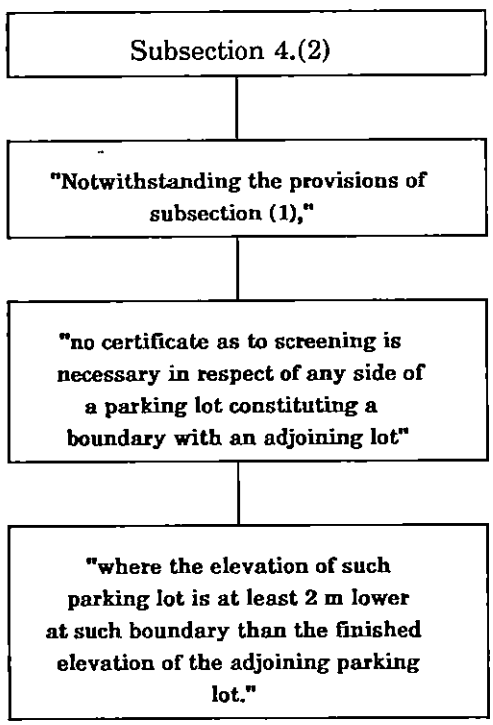


Figure 11: Structure of Subsection 4.(2)

So far, how the links between nodes representing clauses are inserted has been described. However, within a document segment, once the links between the generated objects (if any) are determined, a link must be established to connect these new objects with the one from which they were both derived. All of the

derived objects will at least indirectly give support to the objects representing the document segment.

If there are no generated objects, that is, the text contained in the document segment cannot be further subdivided, the new object will be linked into the network supporting the document segment node. When objects are generated and links inserted by reference to the connectives marking the subordinate clause, at least one object will not have had a support-to link added to it. That is, in the context of this document segment, one object will not give support to any of the other objects. Any such object will be connected to the document segment node with a support-to link.

Thus, for example, in 4.(1)(a) two new objects will be generated.

Example 7: Clause 4. (1)(a)

> (a) [that the surface area of the parking lot has been completely paved] [and is adequately drained;]

Since the conjunction "and", of the category "parallel-ordered", occurs at the beginning of one of the clauses, no support links will be established between them. They were derived from the object representing 4.(1)(a), and since neither is supporting any other object, both will support object 4.(1)(a) in the text network as shown in Figure 12.

Examples used to illustrate connections made for clauses introduced by "pre-ordered" and "post-ordered" categories of conjunctions have all been illustrated with a link to the document segment. (See Figure 11 on page 74 and Figure 10 on page 73). From these illustrations it should be clear that the object representing the main clause will be the one which does not support any other

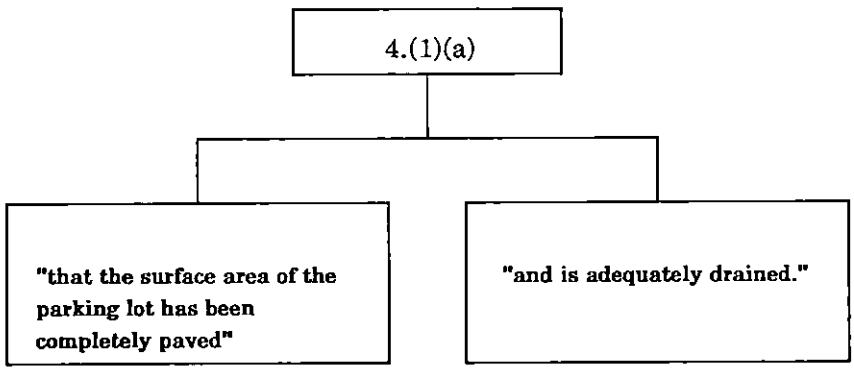


Figure 12: Structure of 4.(1)(a)

object locally. Thus, it will be directly linked to the document segment node with a support-to link. In the case of clauses introduced by conjunctions of the category "post-ordered", it will be the object representing the subordinate clause that will be linked in support of the document segment node.

4.4.3 Internal References

There are a number of internal references to particular sections or subsections of the Bylaw which should prevent the creation of a new node in the text representation. One example is found in subsection 4.(3).

Example 8: Subsection 4. (2)

> (3) Where the provisions of subsection (2) apply the City Engineer may stipulate any modifications of the screening requirements as may be necessary to conform to zoning bylaws and traffic bylaws in respect to safety.

If the clause "Where the provisions of subsection (2) apply" were treated in that same way as described above, a new node would be generated to represent this clause. However, this would effectively duplicate the node representing subsection (2) that was added from the document structure. This type of internal reference occurs a number of times in this Bylaw.

The problem is addressed by searching each clause for such references, using both the style of number used in the reference and the occurrence of the words 'section' and 'subsection' to identify them. The consistent style used for such internal references permits this very simple scheme to provide a reasonably reliable result. When such a reference is encountered, no new node is generated for the clause. Instead, the node already in the text structure representing the referenced document component will be used. Links will be added to connect the old node into the text representation in a new location. An example will help to make this clear.

One new node will be generated from subsection 4.(3) to represent the main clause *the City Engineer may ....* The first clause, introduced by *Where* will not generate a new node. Rather, the node representing subsection 4.(2) (generated from the document structure) is treated as if it were the node representing the *Where* clause. Therefore, on the basis of the conjunction *where*, the node 4.(2) will be linked to support the node representing the clause *the City Engineer ....* That is, it will take the place of the node which would otherwise have been generated. The representation derived for this subsection is shown in Figure 11 on page 74.

The supporting link between subsection 4.(2) and section 4., taken from the document representation is not affected by this new processing. The new link is simply added to the representation. In this way, the text network begins to capture some of the interconnection between the conditions laid out in the Bylaw.

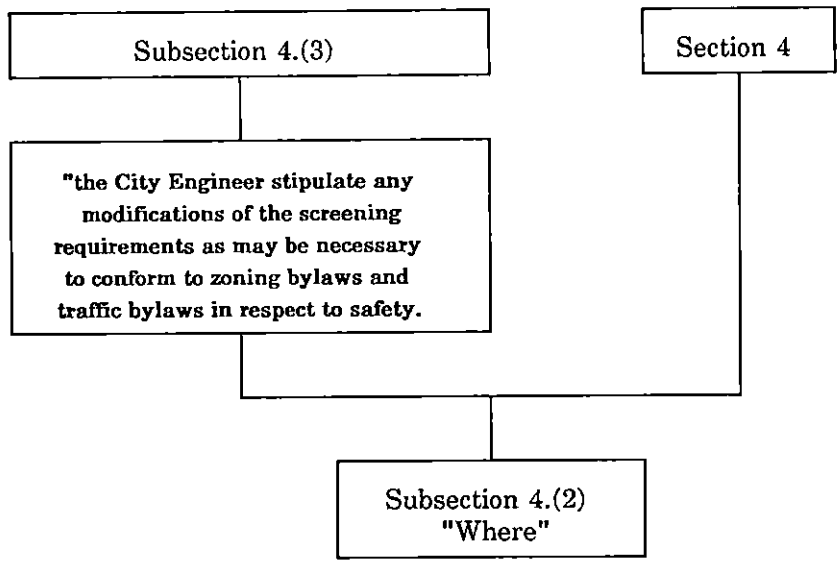


Figure 11: Structure of Subsection 4.(2)

4.5 Summary

In general, it appears that each of the function words addressed above has the effect of imposing a logical ordering between the node representing the clause or phrase it introduces and the node which is its associated main clause. So, not only do these words provide cues as to syntactic structure, but they also provide cues to the structure of the knowledge represented. This is the important structural characteristic which is the motivation for the processing method outlined in this chapter.

The **support network** of ACQUIRE, the knowledge acquisition software used in this research, defines an ordering relation between **objects** in a knowledge base. That is, the **support network** must link an **object** to all other **objects** that it supports and that support it.

Conjunctions have been treated as signals of the logical ordering between clauses in the text without addressing exactly what type of ordering is implied. Depending on the topic of the document, support could be one of the following types: temporal or causal dependence between events, actions, or propositions; elaboration of detail; or contrastive relationships. In spite of these distinctions, all of these kinds of "support" imply an ordering between pairs of nodes. This ordering is that part of the target knowledge representation with which this project has been concerned.

A similar approach to structuring discourse representation is taken by Grosz & Sidner (1986) in their analysis of two types of discourse, an essay and a task-oriented dialogue. They use two different relations, "supports" and "generates", which connect propositions in the essay and actions in the task dialogue, respectively. Although these two relations are intuitively quite different, both have the effect of ordering the components of discourse content. Grosz and Sidner also observe that hierarchical relations of the attentional structure that are explicitly marked by linguistic cues can be used to infer relations of the intentional structure. This is precisely what we are attempting to do here, but in the context of the sample Bylaw chosen for analysis.

Chapter V

SUMMARY AND CONCLUSIONS

In this thesis, the basic notion is that one of the functions of "structure" in language is to communicate, in another type of symbolic system, the concepts conveyed by a text's content and the relationships among these concepts. Specifically, the organizing role that subordinate clauses and the conjunctions which frequently introduce them is investigated. The purpose has been to make explicit how these functors and their associated structures can be interpreted and represented.

It has been assumed that clauses in a text can be represented as distinct units in the representation of the interpretation or meaning of the text. The details of this representation have not been addressed in depth, assuming only that each concept may be represented by a distinct unit, which will be called an object. The operations or processes involved in deriving a representation of the text operate on these units. The text representation will include an object representing each of these clause-level concepts. In addition, relationships among these basic objects are represented by links between them. On the basis of structural relationships entailed by sentence membership and document format, additional objects representing more complex concepts are added to the representation. The links between basic and complex objects capture some of the higher levels of organization in the text representation.

It is generally assumed that subordinating conjunctions serve to signal the relationship that holds between a subordinate clause and its associated main clause. In addition, they signal a number of different types of relationships; as for example, causal, temporal, or comparative relationships. Regardless of the subtlety of distinctions that may be made in terms of the kind of relationship signalled by a particular conjunction, all of these relationships may be seen as imposing an ordering on objects in the text representation. Thus, each subordinating conjunction may be categorized according to one of the three possible ordering relations called pre-ordered, post-ordered and parallel-ordered.

All relationships between objects are represented by links in the text representation. The different types of relationships (causal, temporal, comparative, etc.) can be viewed as distinct models, or levels of organization, which combine to yield the overall representation. In a particular text, there may be more than one model depending upon the subject matter and the writer's purpose. Each model can be seen as an independent set of links among objects in the representation. The conjunctions in the text do not themselves uniquely indicate which relationship is intended, and thus do not provide sufficient information to distinguish possible models that should be represented. Identifying what kind of models are necessary to represent a text's interpretation would require an analysis of the text's topic structure and thus detailed consideration of the internal representation of objects.

Therefore, the analysis of the sample Bylaw presented here has assumed that a single model is sufficient and no attempt has been made to differentiate types of links according to a particular model. In the context of the project undertaken to

test the view of conjunctions described above, such a single model approach has been successful. This is a result of the restricted purpose of the formal, regulatory document chosen for analysis. In this document, the content can be viewed via a single model based on logical contingency between basic objects (clauses).

In the review of linguistic literature, both theoretical and psycholinguistic research was used to motivate the use of subordinating conjunctions to organize objects, or units, of the text representation. The types of connections signalled by conjunctions has been investigated by Halliday and Hasan (1976), Martin (1983) and Rudolph (1988); each of these studies established similar typologies of connections. From a cognitive perspective, both Rudolph (1988) and Morrow (1986) have suggested viewing conjunctions as cues which guide the construction and organization of text comprehension. The view of function words as organizing the interpretation of language presented in this thesis is based on these previous studies. That is, function words are traditionally treated as "empty" words, without significant meaning in themselves; their meaning has been said to be realized in context, where they serve to connect the "content" words of an utterance or discourse. The same can be said of syntactic phrase structure which reflects the compositional nature of phrases and clauses which are intuitively recognizable units of "meaning". Clauses are connected by syntactic form or explicit connectives or both.

Function words do more than indicate syntactic structure, they also make a significant contribution to communication of meaning. The categorization of conjunctions according to an ordering relation proposed here is an explicit

expression of meaning of these words from a processing viewpoint. Although the different types of ordering, or models, that conjunctions may suggest is another important aspect of their meaning, the present analysis does not address this issue. This is because, as previously mentioned, the conjunctions do not specify the type of relation independently. Rather, there is an interaction between these structural words and the content of the text.

This view has been suggested by other workers in the area of computational linguistics. Grosz and Sidner (1987) suggest that the organization of a discourse is based on interaction between form and content. The functions which connect elements in the intentional structure differ according to the topic of the discourse, but are parallel in form. The structural parallelism can be captured through a general ordering relation which is independent of the particular domain. In addition, this intentional structure can be inferred from attentional structure, which is in turn built from linguistic structure. That is, features of the linguistic structure or form are reflected in the structure of the text's interpretation.

This model of conjunctions as imposing ordering relations between objects or concepts in a text representation has been applied to the problem of knowledge acquisition for expert systems. Ordering among elements is an important feature in all schemes for knowledge representation. Whether the representation is a set of production rules, a network of objects and values, or a combination, some form of ordering is imposed to relate the individual components. The general ordering relations in the organization of knowledge representation have been described by Gaines (1987) as linking lower levels with "higher levels" of organization in terms of alternative and abstract models and by Breuker & Wielinga (1987) as

dependencies between objects captured in a model (or view of) the object organization. The models suggested in both cases range from causal, conditional, and spatial to empirical models based on experiences, perhaps incorporating temporal ordering. Thus, the ordering relations entailed by conjunctions are an essential part of the information required in a knowledge base.

The process of knowledge acquisition involves the integration of information from many sources. Written texts are used extensively by knowledge engineers, but only limited attempts have been made to incorporate automatic analysis of texts into knowledge acquisition systems. Therefore, this project was undertaken to apply the proposed analysis of conjunctions to automatically generate a knowledge base.

In the prototype processing system developed, syntactic structure inserted in the text serves to segment the original sequence of linguistic units into concepts or objects in the representation. The linear order of syntactic units in the text imposes a basic, default organization among these components. Conjunctions are used to identify where links should be inserted between objects. In this way, conjunctions function in cooperation with the patterns of syntactic structure, to organize the representation.

While the data chosen are restricted to formal, regulatory documents, the basic interpretation should remain valid across discourse types. There is a difference in how strongly these features contribute to understanding different types of discourse. In formal documents, the contribution is significant, and this is one reason for choosing this type of document for analysis. The contribution of these functors to the interpretation of other types of discourse, narratives for example, may not be as great, but the meaning they symbolize remains the same.

In addition, the access to textual materials during the knowledge acquisition process would be enhanced by the establishment of links between objects or rules in the knowledge base and the source document. To fulfill this criterion, it is not enough to have extracted an abstract representation of text components. It is also necessary to allow the developer access to the document to help in the interpretation of the object network. Here again it is well to keep in mind that arbitrary labels on objects in the knowledge base do not necessarily provide enough information for a human user to interpret the meaning of an object. But seeing the source of that object, in its full written form, can allow the developer to assign the correct meaning to each object by seeing it in context.

The organization of information in written texts is conveyed by many other attributes as well. The document layout itself has been used here to effect the initial segmentation of the document. Other forms of cohesive devices, such as lexical reference, would likely enhance the completeness and accuracy of the knowledge representation extracted from a text. In the current project, no concerted effort has been made to address these aspects of language. However, explicit references to the components of the document itself were addressed because the processing was straight-forward, and eliminating duplicate objects significantly improved the usefulness of the resulting representation.

The prototype system successfully generated a set of objects definitions for the sample document. These definitions were used to produce an object network in the ACQUIRE system. The resulting knowledge base was not as complete as that prepared manually; however, those parts of the network that were generated were accurate. The main source of incompleteness was in the topical or thematic

organization among the document components. This is certainly to be expected since no lexical analysis was done. The methodology used by Shaw & Gaines (1987) for lexical analysis might yield another set of links among the objects on the database, imposing yet another ordering, this time based on topical relations.

The usefulness of the resulting knowledge base is limited by the technology available to fully implement the interface between the on-line text and the object definitions. Currently, the object definitions are simply labelled with the portions of the text to which they correspond. The facility to implement dynamic links between the knowledge base and the on-line text, a type of hypertext system, is necessary to make this type of system truly useful. The text associated with objects in the knowledge base does not necessarily provide enough information for a human user to interpret the object's meaning. The segments of text, out of context, are not always helpful. However, if these labels were augmented with links to the location of the segment in the document, users would be able to see the segment in its context and so allow them to correctly interpret each object.

The study has demonstrated that one part of the meaning of these conjunctions is to impose an ordering on components of semantic representation. The sequential or ordering nature of the relations signalled by all conjunctions is presented. This principle, then, has been used as the basis of a strategy for automatically extracting a knowledge representation from written texts. In addition to presenting a processing oriented analysis of conjunctions, this thesis has applied linguistic analysis to the problem of knowledge acquisition. In doing so, it is hoped that the common questions of knowledge representation and acquisition addressed by discourse analysts in linguistics and computer scientists

have been further illuminated and the often suggested potential for cooperation between these fields demonstrated.

BIBLIOGRAPHY

Aho, Alfred V., John E. Hopcroft and Jeffrey D. Ullman (1983) *Data Structures and Algorithms*. Addison-Wesley: Reading, Mass. Anderson, J.R. (1982) "Acquisition of cognitive skill." *Psychological Review* 89: 369-406.

Anderson, J.R. (1987) "Skill Acquisition: Compilation of Weak-Method Problem Solutions." *Psychological Review* 94: 192-210.

Anderson, J.R. and G.H. Bower (1973) *Human Associative Memory*. V.H. Winston and Sons: Washington, D.C.

Barr, Avron and Edward A. Feigenbaum, Eds. (1981a) *The Handbook of Artificial Intelligence Vol 1*. Addison-Wesley: Reading, MASS.

Barr, Avron and Edward A. Feigenbaum, Eds. (1981b) *The Handbook of Artificial Intelligence Vol 2*. Addison-Wesley: Reading, MASS.

de Beaugrande, Robert-Alain and Wolfgang Ulrich Dressler (1981) *Introduction to Text Linguistics*. Longman: London.

Brachman, Ronald J. (1979) "On the Epistemological Status of Semantic Networks," in *Associative Networks: Representation and Use of Knowledge by Computers*, ed. N.V. Findler. Academic Press: New York, 3-50. Reprinted in Brachman and Levesque (1985): 191-215.

Brachman, Ronald J. and Hector J. Levesque, Eds. (1985) *Readings in Knowledge Representation*. Morgan Kaufmann: Los Altos.

Breuker, Joost and Bob Wielinga (1987) "Use of Models in the Interpretation of Verbal Data," in Kidd (1987): 17-44.

Buchanan, Bruce G. (1988) "Fundamentals of Expert Systems." *Annual Review of Computer Science* 3: 23-58.

Buis, M., J. Hamer, J.G. Hosking, and W.B. Mugridge. (1987) "An Expert Advisory System for a Fire Safety Code," in Quinlan (1987): 85-101.

Clancey, William J. (1984) "Methodology for Building an Intelligent Tutoring System," in *Methods and Tactics in Cognitive Science*, Ed. W. Kintsch, J. R. Miller and P.G. Polson. Erlbaum: Hillsdale, NJ., 51-83.

Clancey, William J. (1985) *Heuristic Classification.* Technical Report STAN-CS-85-1066. Department of Computer Science, Stanford University: Stanford, CA.

Conklin, Jeff (1987) "Hypertext: An Introduction and Survey." *Computer* 20: 17-41.

Davis, R. (1977) "Production Rules as a Representation for a Knowledge-Based Consultation System." *Artificial Intelligence* 8: 15-45. Reprinted in Brachman and Levesque (1985): 371-387.

Davis, Randall and Douglas B. Lenat (1982) *Knowledge-Based Systems in Artificial Intelligence.* McGraw-Hill: New York.

Davis, Randall, Bruce Buchanan, and Edward Shortliffe (1977) "Production Rules as a Representation for a Knowledge-Based Consultation Program." *Artificial Intelligence* 8: 15-45.

van Dijk, Tuen A. (1977) *Text and Context Explorations in the semantics and pragmatics of discourse.* Longman: London.

van Dijk, Tuen A. (1980) *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition.* Lawrence Erlbaum: Hillsdale, NJ.

Ericsson, K. Anders and Herbert A. Simon (1985) "Protocol Analysis," in *Handbook of Discourse Analysis, Vol. 2,* Ed. T. van Dijk. Academic Press: London, 259-268.

Fillmore, Charles J. (1968) "The Case for Case," in *Universals in Linguistic Theory.* ed. Emmon Bach and Robert T. Harms. Holt, Rinehart and Winston: New York, 1-88.

Friedman, Carol (1987) "A Sublanguage Narrative Processor," in Sager, Friedman and Lyman (1987), 81-112.

Gaines, Brian R. (1987) "An overview of knowledge-acquisition and transfer." *International Journal of Man-Machine Studies* 26: 453-472.

Gardner, Anne von der Lieth (1987) *An Artificial Intelligence Approach to Legal Reasoning.* The MIT Press: Cambridge, Mass.

Givon, T. (1983) "Topic Continuity in Discourse: An Introduction," in *Topic Continuity in Discourse: A Quantitative Cross-Language Study.* ed. T. Givon. John Benjamins: Amsterdam, 1-42.

Goldfarb, Charles F. (1986) "Introduction to Generalized Markup," in *Information Processing - Text and office systems - Standard Generalized Markup Language (SGML) ISO 8879-1986(E).* ed. Technical Committee ISO/TC 97. International Standards Organization: Switzerland, 59-65.

Grishman, Ralph, Lynette Hirschman, and Ngo Thanh Nhan (1986) "Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments," *Computational Linguistics* 12: 205-215.

Grishman, R. and R. Kittredge, Eds. (1986) *Analyzing Language in Restricted Domains: Sublanguage Description and Processing.* Lawrence Erlbaum: Hillsdale, N.J.

Grosz, B. (1982) "Discourse Analysis," in *Sublanguage: Studies of Language in Restricted Semantic Domains*, Ed. R. Kittredge and J. Lehrberger. de Gruyter: Berlin.: 138-174.

Grosz, Barbara J. and Candace L. Sidner (1986) "Attention, Intentions, and the Structure of Discourse." *Computational Linguistics* 12: 175-204.

Haas, Norman and Gary G. Hendrix (1983) "Learning By Being Told: Acquiring Knowledge For Information Management," in *Machine Learning An Artificial Intelligence Approach*, ed. Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell. Tioga: Palo Alto, 405-427.

Hafner, Carol. (1981) *An Information Retrieval System Based on a Computer Model of Legal Knowledge.* UMI Research Press: Ann Arbor.

Hahn, U. and U. Reimer (1988a) "Knowledge-Based Text Analysis in Office Environments: The Text Condensation System TOPIC," in *OFFICE KNOWLEDGE: Representation, Management, and Utilization*, Ed. W. Lamersdorf. North-Holland: Amsterdam, 197-215.

Hahn, Udo and Ulrich Reimer (1988b) "Automatic Generation of Hypertext Knowledge Bases," in *Conference on Office Information Systems*, ed. Robert B. Allen. ACM: New York, 182-188.

Hajicova, Eva (1987) "Focussing - A Meeting Point of Linguistics and Artificial Intelligence," in *ARTIFICIAL INTELLIGENCE II: Methodology, Systems, Applications*, ed. Ph. Jorrand and V. Sgurev. Elsevier (North-Holland): Amsterdam, 311-321.

Halliday, M.A.K. and Ruqaiya Hasan (1976) *Cohesion in English.* Longman: London.

Harris, Zellig, Michael Gottfried, Thomas Ryckman, Paul Mattick Jr., Anne Daladier, T.N. Harris, and S. Harris (1989) *The Form of Information in Science: Analysis of an Immunology Sublanguage.* Kluwer Academic: Dordrecht.

Hart, Anna (1986) *Knowledge Acquisition for Expert Systems.* Kogan Page: London.

Hayes-Roth, Frederick, Donald A. Waterman, and Douglas B. Lenat, Eds. (1983) *Building Expert Systems*. Addison-Wesley: Reading, MASS.

Hirst, Graeme (1988) "Semantic Interpretation and Ambiguity." *Artificial Intelligence* 34: 131-177.

Ingold, Rolf (1989) "Text Structure Recognition in Optical Reading," in *Structured Documents*. ed. J. Andre, R. Furuta, and V. Quint. Cambridge University Press: Cambridge, 133-142.

International Standards Organization (1986) *Information Processing - Text and office systems - Standard Generalized Markup Language (SGML)*. ISO 8879:1986. Technical Committee ISO/TC 97. International Standards Organization: Switzerland.

Jackendoff, Ray (1983) *Semantics and Cognition*. Paperback edition 1985. MIT Press: Cambridge, MASS.

Johnson, Nancy S. (1985) "Extracting the Proof from the Pudding: Coding and Analyzing Experimental Protocols," in *Handbook of Discourse Analysis, Vol. 2*, ed. T. van Dijk. Academic Press: London, 245-257.

Johnson, Steven (1987) "Temporal Information in Medical Narratives," in Sager, Friedman and Lyman (1987), 175-194.

Kempen, Gerard, Ed. (1987) *Natural Language Generation*. NATO ASI Series. Martinus Nijhoff: Dordrecht.

Kidd, Alison L., Ed. (1987) *Knowledge Acquisition for Expert Systems: A Practical Handbook*. Plenum Press: New York.

Kintsch, Walter (1988) "The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model." *Psychological Review* 95: 163-182.

Kittredge, R. and J. Lehrberger, Eds. (1982) *Sublanguage: Studies of Language in Restricted Semantic Domains*. de Gruyter: Berlin.

Kuipers, Benjamin and Jerome P. Kassirer (1987) "Knowledge Acquisition by Analysis of Verbatim Protocols," in Kidd (1987): 45-72.

Lesgold, Alan, H. Rubinson, P. Feltovich, R. Glaser, D. Klopfer, and Y. Wang (1988) "Expertise in a complex skill: Diagnosing X-ray pictures," in *The nature of expertise*, ed. M.R.H. Chi, R. Glaser, and M.J. Farr. Erlbaum: Hillsdale, NJ., 311-342.

Lockwood, Glen (1988) Letter to Dr. B.A. Schafer, June 14, 1988. Acquired Intelligence Inc. Victoria, B.C.

Longacre, Robert E. (1983) *The Grammar of Discourse.* Plenum Press: New York.

Mann, William C. and Sandra A. Thompson (1986) "Relational Propositions in Discourse." *Discourse Processes* 9: 57-90.

Mann, William C. and Sandra A. Thompson (1987) "Rhetorical Structure Theory: description and construction of text structures," in Kempen (1987): 85-96.

Martin, Jim R. (1983) "Conjunction: The Logic of English Text," in *Micro and macro connexity of texts*, ed. J. S. Petöfi and E. Sözer. Helmut Buske: Hamburg, 1-71.

Matthiessen, Christian (1987) "Notes on the Organization of the Environment of a Text Generation Grammar," in Kempen (1987): 253-278.

Michie, Donald (1987) "Current Developments in Expert Systems," in Quinlan (1987): 137-156.

Minksy, Marvin (1981) "A Framework for Representing Knowledge," in *Mind Design*, ed. J. Haugeland. MIT Press: Cambridge, MA, 95-128.

Morrow, Daniel G. (1986) "Grammatical Morphemes and Conceptual Structure in Discourse Processing." *Cognitive Science* 10: 423-455.

Municipal Council of the Corporation of the City of Victoria (1987) Bylaw No. 87-248. The "Parking Lot Bylaw".

Nishida, Fujio, Shinobu Takamatsu, and Tadaaki Tani (1986) "Text Analysis and Knowledge Extraction," in *Proceedings of 11th International Conference on Computational Linguistics*. International Committee on Computational Linguistics: Bonn, 241-243.

Quinlan, J. Ross Ed. (1987) *Applications of Expert Systems.* Addison-Wesley: Sydney.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1972) *A Grammar of Contemporary English.* Longman: London.

Rudolf, Elisabeth (1988) "Connective Relations - Connective Expressions - Connective Structures," in *Text and Discourse Constitution - Empirical Aspects, Theoretical Approaches*, ed. Janos S. Petöfi. de Gruyter: Berlin, 97-133.

Sager, Naomi, Carol Friedman, and Margaret S. Lyman, M.D., Eds. (1987) *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley: Reading, Mass.

Savory, Stuart, Ed. (1988) *Artificial Intelligence and Expert Systems*. M.W. Rogers (trans.) Ellis Horwood Limited: Chichester.

Schank, Roger C. (1980) "Language and Memory." *Cognitive Science* 4: 243-284.

Sergot, M.J., F. Sadri, R.A. Kowalski, F. Kriwaczek, P. Hammond, and H.T. Cory (1986) "The British Nationality Act as a Logic Program." *Communications of the ACM* 29: 370-386.

Shaw, Mildred L.G. and Brian R. Gaines (1987a) "KITTEN: Knowledge initiation and transfer tools for experts and novices." *International Journal of Man-Machine Studies* 27: 251-280.

Shaw, Mildred L.G. and Brian R. Gaines (1987b) "An Interactive Knowledge-Elicitation Technique Using Personal Construct Technology," in Kidd (1987): 109-136.

Smith, R.A. (1976) "Computer-based structural analysis in the development and administration of educational materials." *International Journal of Man-Machine Studies* 8: 439-463.

Talmy, Leonard (1988) "Force Dynamics in Language and Cognition." *Cognitive Science* 12: 49-100.

Tucker, Allen B., Sergei Nirenburgu, and Victor Raskin (1986) "Discourse and Cohesion in Expository Text," in *Proceedings of 11th International Conference on Computational Linguistics*. International Committee on Computational Linguistics: Bonn, 181-183.

Waterman, Donald A., Jody Paul, and Mark Peterson (1987) "Expert Systems for Legal Decision Making," in Quinlan (1987): 23-47.

Winston, Patrick H. (1987) "The Commercial Debut of Artificial Intelligence," in J. Ross Quinlan, Ed. *Applications of Expert Systems*. Addison-Wesley : Sydney.: 3-20.

APPENDIX A

SAMPLE BYLAW DOCUMENT

NO. 87-248

A BYLAW OF THE CITY OF VICTORIA

By virtue of the powers conferred upon it by Section 18 of the Victoria City Act, 1919, and other enabling powers, the Municipal Council of the Corporation of the City of Victoria enacts as follows:

1. [This bylaw may be cited as the "PARKING LOT BYLAW"].

2. [In this bylaw

["vehicle" has the meaning assigned to it in the Motor Vehicle Act][;

"parking lot"
means a place, on one parcel of land, which is used or set aside for use for the parking of one or more vehicles in consideration of the payment of money].

3. [No person shall operate a parking lot] [unless he holds a valid and subsisting licence for it, issued under the provisions of this bylaw and.of the Business Licence Bylaw].

4. (1) [No licence for a parking lot shall be issued] [unless and until the City Engineer certifies:

(a) [That the surface area of the parking lot has been completely paved] [and is adequately drained];

(b) [where the parking lot is in or adjoining an area zoned by bylaw or lawfully used for residential use], [that it is screened from adjoining parcels of land either by evergreen hedges or by view obscuring fences or both]

[and [that such hedges or fences are of a height of not less than 1.3 m and, for fences, not more than 2 m, along the common boundaries of such adjoining properties and of the parking lot]];

(c) [where the parking lot abuts on a street], [that it is screened along its entire street boundary, except for necessary vehicular access points, either by an evergreen hedge or shrubs or by permanent masonry planters with plants growing in them, or by both methods, in such a manner as to provide an effective screen of the parking lot along all street boundaries and of a height of at least 1.3 m above ground level];

(d) [that all lighting used to illuminate the parking lot is deflected from adjoining lots and streets]; and

(e) [that there is only one sign, not exceeding 2 m_2 in area, at each entrance and at each exit], [and [that such sign does not contain any words or signs other than to designate entrances, exits, conditions of use of the parking lot, the name of the parking lot and conditions relating to the towing away of vehicles]].

(2) [Notwithstanding the provisions of subsection (1)], [no certificate as to screening is necessary in respect of any side of a parking lot constituting a boundary with an adjoining lot] [where the elevation of such parking lot is at least 2 m lower at such boundary than the finished elevation of the adjoining parking lot].

(3) [Where the provisions of subsection (2) apply] [the City Engineer may stipulate any modifications of the screening requirements as may be necessary to conform to zoning bylaws and traffic bylaws in respect to safety].

5. [[Upon it appearing that the holder of a licence issued for a parking lot is not watering the trees, shrubs or other plants on the lot adequately], [or is allowing litter to accumulate on the lot] [or is otherwise failing to maintain the parking lot up to the standards required in terms of the preceding section]], [[the matter shall be reported to the Council, who, after affording the applicant an opportunity to be heard, shall, if satisfied that the report is correct, either cancel the licence or suspend its operation until such time as the matters complained of are corrected], [and may, upon cancelling the licence, direct that it shall not be renewed]].

6. [Before renewing any licence and without assigning any reason], [the Business Licence Inspector may demand that the applicant obtain and

produce a fresh certificate from the City Engineer, complying with the provisions of Section 4], [and if the applicant fails to do so, shall refuse to renew his licence].

7. [Notwithstanding the provisions of Sections 4 and 6], [the Business Licence Inspector shall waive the requirements of a certificate under Section 4] [upon production by the applicant of a certificate from the City Engineer to the effect that:

(a) [the proposed use of the premises for parking lot purposes does not appear to constitute a hazard to traffic in the area or to the users of the parking lot];

(b) [it appears that the proposed use is only temporary, following demolition of a building and pending new construction];

(c) [no parking lot licence under this bylaw has previously been issued in respect of the same premises]; and

(d) [the premises are in that part of the City included within the boundaries of one of the areas at that time zoned as

"M-1 Zone, Limited Light Industrial District,"
"M-2 Zone, Light Industrial District,"
"M-3 Zone, Heavy Industrial District," or
"M-3S Zone, Special Heavy Industrial District."
"M-3T Zone, Heavy Industry Tank Farm District."]

8. [A licence issued pursuant to the preceding section may not be renewed] [except upon production of a certificate under Section 4].

9. (1) [Where a parking lot which complies with the provisions of this bylaw and in respect of which a valid and subsisting business licence is held under the Business Licence Bylaw is equipped with a ticket dispensing machine in good working order, into which coins must be deposited as advance payment for parking on such parking lot], [[no person shall park or leave a vehicle on such parking lot] [without immediately depositing the required amount of money into the ticket dispensing machine and leaving the appropriate and valid ticket on the front dashboard of vehicle in such a manner that the number printed on the ticket and the date and time if printed on the ticket are clearly visible from outside the vehicle]].

(2) [No person shall leave a vehicle which has been lawfully parked pursuant to subsection (1) for a time longer than that authorized as indicated by a notice at or on the ticket dispensing machine].

10. (1) [Where parking spaces on a licensed parking lot are clearly delineated by painted lines or barriers],[[no person shall park a vehicle on such parking lot], [except in such parking spaces], [and no person shall park a vehicle in such a manner as to straddle the line between two parking spaces]].

(2) [Where any parking space on a licensed parking lot is equipped with a parking meter], [[no person shall park a vehicle within such parking space] [without having deposited the appropriate fee for parking in the manner and at the rate prescribed or measured by the meter]].

11. (1) [Where any parking space on a licensed parking lot is clearly marked as reserved], [no person other than the person for whom it is reserved shall park a vehicle in such space].

(2) [Where any area on a licensed parking lot is designated as a Truck Loading Zone pursuant to the Streets and Traffic Bylaw], [the regulations and prohibitions applicable in the Streets and Traffic Bylaw apply to the Zone].

12. [Subject to the provisions of Section 19], [[no person shall tow away or cause a vehicle to be towed away from a parking lot, whether or not such parking lot is licensed under the Business Licence Bylaw, by reason of such vehicle having been left there without the consent of the occupier of the parking lot if such lack of consent consists of either the failure of the drive of the vehicle to purchase a parking ticket or the expiry of a parking ticket], [unless such towing occurs between the hours of 2 o'clock in the morning and 7 o'clock in the morning]].

13. [No person shall cause or authorize a vehicle to be towed away for failure to obtain a parking ticket from a ticket dispenser on a parking lot] [unless such ticket dispenser

(a) [at that time is in sound working order] [and supplied with parking tickets], and

(b) [dispenses parking tickets in duplicate]].

14. [No person shall cause to authorize a vehicle to be towed away to a destination outside the boundaries of the City of Victoria].

15. [The provisions of Sections 4 to 8, inclusive, do not apply to a parking lot

(a) [that is wholly in, under or on a building]; or

(b) [that is not operated at any time between 7 o'clock in the morning and 5 o'clock in the afternoon]].

16. [No person shall erect, maintain, operate, control, keep or manage a ticket dispenser] [unless:

(a) [the owner or occupier of the premises holds a valid and subsisting licence for a parking lot, applicable to the location of the ticket dispenser]; and

(b) [the prices charged by the operator of the parking lot for the parking of vehicles are clearly stated in a conspicuous notice on or at the ticket dispenser]].

17. [No person shall charge or attempt to charge a fee for the parking of a vehicle in respect of which there is not a valid or unexpired ticket issued by a ticket dispenser that is more than twice the fee that would have been payable if such ticket had been purchased or had not expired].

18. [No person shall put or keep a sign on a parking lot or deliver a document on a parking lot] [[if such sign or document states or implies that the operator has the right to cause a vehicle to be towed away] [without the restrictions imposed by the bylaw]].

19. [Notwithstanding the provisions of Section 12], [where more than 10 violation issued pursuant to Section 21 have been put on any vehicle], [[the owner or operator of a parking lot may cause or authorize the vehicle to be towed away from the parking lot] [where the driver of the vehicle has failed to purchase a parking ticket or the parking ticket purchased has expired]].

20. [[Any person contravening any provision of this bylaw is guilty of an offence] [and liable on conviction to the fines prescribed by the Offence Act which shall be not less than $25.00]].

21. [Where any person authorized by or pursuant to a resolution of the Council puts a violation notice on a vehicle whose owner is alleged to have committed an offence against Section 9, 10 or 11], [[no prosecution shall be commenced against the alleged offender] [if there is paid to the City a voluntary penalty of

(a) [$10.00] [if paid within 7 days from the date of the violation notice];

(b) [$15.00] [if paid after 7 days but within 45 days from the date of the violation notice];

(c) [$25.00] [if paid after 45 days from the date of the violation notice]]].

22. [Bylaw No. 82-85, the "Parking Lot Bylaw," and all its amendments are repealed].

Passed and received third reading by the Municipal Council the 26th day of November, 1987.

Reconsidered and adopted by the Municipal Council the 17th day of December, 1987.

APPENDIX B

AUTOMATIC DOCUMENT ANALYSIS

The analysis presented in Chapter 4 has been implemented in computer programs. Three prototype programs have been written, each one accomplishing one aspect of the analysis.

The three programs accomplish the following tasks:

1. Translate the document format implicit in the physical layout into explicit SGML style tags marking the logical document structure.

2. Generate a representation of the document structure in the form of a tree structure in which each logical component is a node.

3. Generate a network of objects from the document structure, creating new objects and links in accordance with the clause markings and occurrence of conjunctions.

B.1 Step 1 - Generalized Document Markup

The analysis of the document format carried out here is specific to the sample Bylaw. In this document, the major components are identified sufficiently by the numeric or alphabetic labels in the text, without taking into consideration the horizontal or vertical spacing on the page or any other possible attributes. A general method for identifying document structure would include consideration of the placement of characters, as well as the font and size of type (Ingold 1989). However, for the current project these factors have not been considered.

Initial processing of the raw text replaces the numeric/alphabetic labels (1., (2), (c)) with SGML style tags (International Standards Organization 1986). The actual reference numbers are maintained as values of tag attributes so that they can be used in later processing to resolve internal references. The tags make explicit the hierarchical document structure implicitly coded in the labels and the indentation of the text associated with each type of label.

Each logical component of the document structure is marked by the same type of label. That is, each segment of the document marked by an Arabic numeral followed by a period is a "section" of this document. "Sections" are the major logical divisions in the document structure. Arabic numerals enclosed in parentheses label "subsections" of the document. Subsections are strictly embedded in a section of the document, so can be considered part of the section. Similarly, the labels made up of a lowercase letter enclosed in parentheses mark the beginning of each item in a "list". List items may be embedded in section or subsection components of the document.

Take for example the excerpt from the Bylaw document shown in Example 9.

Example 9: Section 11.

> 11. (1) [Where any parking space on a licensed parking lot is clearly marked as reserved], [no person other than the person for whom it is reserved shall park a vehicle in such space].
>
> (2) [Where any area on a licensed parking lot is designated as a Truck Loading Zone pursuant to the Streets and Traffic Bylaw], [the regulations and prohibitions applicable in the Streets and Traffic Bylaw apply to the Zone].

The label "11." each will be replaced with the tag, :H1. An attribute, called REF, is associated with each :H1. tag and will have the actual section number, in

this case 11, as its value. The tag marking the section shown above would be: :H1. REF=11.

The two labels (1) and (2) will both be replace by :H2 tags. The REF attribute associated with each tag will be assigned the numeric value of the text label. The output from this stage of processing corresponding to section 11 will be as shown in Figure 13.

```
:H1 REF = 11.
:H2 REF = 1.
[Where any parking space on a licensed parking lot is
clearly marked as reserved], [no person other than the
person for whom it is reserved shall park a vehicle in
such space].
:EH2 REF = 1.
:H2 REF = 2
[Where any area on a licensed parking lot is designated
as a Truck Loading Zone pursuant to the Streets and
Traffic Bylaw], [the regulations and prohibitions
applicable in the Streets and Traffic Bylaw apply to
the Zone].
:EH2 REF = 2.
:EH1 REF = 11.
```

Figure 13: Example of Text Markup

The beginning of the document, including the title and preamble were marked with the tag :BEGIN, and the final section which includes relevant dates were marked by :END. The content of these two portions is not considered further here.

Figure 14 lists the tags used in the markup of this document and the attributes defined for the purpose of the analysis undertaken in this study. For each tag listed a matching ending tag is also defined. Ending tags are all named by appending the letter *E* to the beginning of a tag name. The REF attribute of the main tag is also defined on each ending tag.

Tag Name:	BEGIN
Attribute Name:	n/a
Attribute Value:	n/a
Description:	Marks the preliminary section of the text, including the title and preamble.
Tag Name:	END
Attribute Name:	n/a
Attribute Value:	n/a
Description:	Marks the final portion of the text, which includes the date and signatures.
Tag Name:	H1
Attribute Name:	REF
Attribute Value:	A number represented as an Arabic numeral.
Description:	Marks the beginning of a major section of the text. Identified by an Arabic number terminated by a period. The name 'section' is the reference word used for this type of element.
Tag Name:	H2
Attribute Name:	REF
Attribute Value:	A number represented as an Arabic numeral.
Description:	Marks the beginning of a major section of the text. Identified by Arabic numerals enclosed in parentheses. The name 'subsection' is the reference word used for this type of element.
Tag Name:	OL
Attribute Name:	n/a
Attribute Value:	n/a
Description:	Marks the beginning of a list. Identified by the occurrence of the first list item. This composite element is supported by the punctuation used between list items, either commas or semi-colons rather than periods.
Tag Name:	LI
Attribute Name:	REF
Attribute Value:	The label on the list item.
Description:	Marks the beginning of a list item. Identified by a lower case letter enclosed in parentheses, occurring in a sentence.

Figure 14: Document Tags

The names and punctuation used in this markup process were chosen for consistency with the formatting program Waterloo SCRIPT which was available to process the SGML tags to produce physical output. There is no reason why

different names or punctuation could not be used. The important point is that there is no specification of formatting requirement in the document itself. Rather the logical components of the document are marked, or tagged, with format independent tags. A set of output processing instructions may be associated with each of the tags to produce any desired output format (Goldfarb 1986).

B.2 Step 2 - Creating the Document Representation

The output from the first stage of processing, that is, the text which has be marked with generalized tags indicating the logical structure of the document (ISO 1986), can now easily be processed to create the document representation. The hierarchical relationship between the logical document components is captured by representing each component as a complex data record. The format of the data record is shown in Figure 15.

NAME	:	A generated name of the form "OBJECTnn", where "nn" is an integer value.
ALIAS	:	A string made up of the markup tag and its parameters.
SUPPORT-TO	:	The integer value associated with the component directly above this one in the tree. Only one value may appear in this field.
SUPPORT-FROM	:	A list of integer values associated with the components directly below this one in the tree. There is no restriction on the number of links that can appear in this field.
COMMENT	:	The actual text covered by the component.

Figure 15: Document Structure Nodes - Field Definitions

A stack is used to keep track of embedded structures. Each new component encountered in the text will be recorded on the stack. Each component tag is assigned a level number which is used to determine whether components on the stack should be closed before the new component is processed. The level numbers are as follows:

TAG	Level Number
BEGIN	1
H1	1
H2	2
OL	3
LI	3
END	1

The processing proceeds as follows:

1. On encountering an H1,H2,OL,LI,BEGIN or END tag in the input, a new object is generated.
   - The NAME field is filled in by appending the current object number to the string OBJECT.
   - The current object number is incremented by one.
   - The tag and its attributes are stored in the ALIAS field.
   - Place the new object on top of a stack.
2. Lines of text will be appended to the COMMENT field of each tag that is currently on the stack.
3. On encountering an 'end' tag, the object at the top of the stack will be removed (popped from the stack), filling in the appropriate support links.
   - The number of the popped object is appended to the "SUPPORT-FROM" list of the new top of stack object.
   - The number of the new top of stack object is appended to the "SUPPORT-TO" list of the popped object.

The resulting data structure, for section 11 of the Bylaw is as shown in Figure 16.

NAME: OBJECT31
ALIAS: H1 REF = 11.
SUPPORT_TO: None
SUPPORT_FROM: 32,33
COMMENTS: [Where any parking space on a licensed parking lot is clearly marked as reserved], [no person other than the person for whom it is reserved shall park a vehicle in such space]. [Where any area on a licensed parking lot is designated as a Truck Loading Zone pursuant to the Streets and Traffic Bylaw], [the regulations and prohibitions applicable in the Streets and Traffic Bylaw apply to the Zone].

NAME: OBJECT32
ALIAS: H2 REF = 1.
SUPPORT_TO: 31
SUPPORT_FROM: None
COMMENTS: [Where any parking space on a licensed parking lot is clearly marked as reserved], [no person other than the person for whom it is reserved shall park a vehicle in such space].

NAME: OBJECT33
ALIAS: H2 REF = 2.
SUPPORT_TO: 31
SUPPORT_FROM: None
COMMENTS: [Where any area on a licensed parking lot is designated as a Truck Loading Zone pursuant to the Streets and Traffic Bylaw], [the regulations and prohibitions applicable in the Streets and Traffic Bylaw apply to the Zone].

Figure 16: Sample Document Structure Data Records

This stage of processing only changes the format of information which is implicit in the marked-up document. The hierarchical relationship of the logical documents components is made explicit through the links in the SUPPORT_TO and SUPPORT_FROM fields. The text itself has been copied into the COMMENT field because at this time there is no facility in the ACQUIRE system

for accomplishing an indirect linkage with the document file itself. However, it would be possible, using the information in the ALIAS field and the SUPPORT_FROM links to accomplish this in the future.

The output from this processing comprises the document representation. The format of the data elements was chosen to match that of objects in the ACQUIRE knowledge base. As a result, a simple transformation can be performed to import this structure into the ACQUIRE system. Although the document structure captured in this representation can be viewed as a naive approximation to the "knowledge structure" of the document, it is certainly far from complete.

Additional information in the actual text of each object can be used to a) further subdivide some of the document components and b) identify internal references which should not generate new objects, but rather substitute already existing objects in the network. The processing suggested to utilize this information is discussed in the next section.

B.3 Creating the Text Representation

The document representation described above is the starting point for creating what is called the **text representation**. It is this representation that can be directly translated into the knowledge base format. The data record used to represent objects is that same as that used in Step 2, except the restriction that SUPPORT_TO field hold only a single link, imposed by the hierarchical nature of the document structure, is not imposed on the text representation. In fact, there are many examples in the the Bylaw document where an object (now seen as an interpretation of the actual text) is used in support of several other objects.

NAME	:	A generated name of the form "OBJECTnn", where "nn" is an integer value.
ALIAS	:	The markup tag of the document component containing this text and it's parameters.
SUPPORT-TO	:	A list of integer values associated with the components directly above this one in the network.
SUPPORT-FROM	:	A list of integer values associated with the components directly below this one in the network.
COMMENT	:	The actual text covered by the component.

Figure 17: Text Structure Nodes - Field Definitions

The data records of the document structure are processed in sequence. Each of these data records is copied to the text structure without change. The data records which have at least one value in the SUPPORT-FROM field are not processed further. These records are all further subdivided by the document structure. Therefore, the text associated with them is only indirectly referenced by these nodes.

Data records of the document structure which do not have any values in their SUPPORT-FROM field represent leaves in the document structure tree. The text of these nodes is processed further to generate new elements in the text structure. The clause markers, '[]', are used to identify sequences in the text that will be represented by distinct elements. A data record will be generated for each pair of clause markers.

If only a single new element is generated, no further analysis is required. Support links are inserted into this data record and that of the original data record so that the new record supports the old one. Otherwise, based on the classification shown in Figure 2 on page 49, the link fields of the newly generated data records are completed.

The data record whose text is not marked by a connective is taken to be the "main" element. For a data record whose text is marked by a "pre-ordered" conjunction, a link (the OBJECT number) to the main element is inserted in its SUPPORT-TO field. Its OBJECT number is inserted in the SUPPORT-FROM field of the main element's data record.

The processing for "post-ordered" conjunctions causes the opposite linkage between data records to be established. For a data record whose text is marked by a "post-ordered" conjunction, a link to the main element is inserted in its SUPPORT-FROM field. Its OBJECT number is inserted in the SUPPORT-TO field of the main element's data record. No links are added at this stage if the text associated with the data record is marked by a conjunction of the category "parallel-ordered".

Finally, each new data record is considered in turn to determine whether or not it should be linked to the data record representing the document component. If no link has been added to a data record's SUPPORT-TO field in the processing of the current document structure element, then the OBJECT number of the document structure element will be added to its SUPPORT-TO field. The OBJECT number of this data record will be added to the SUPPORT-FROM field of the document structure element's data record.

The data records generated from OBJECT32 (see Figure 16 on page 106) is as follows:

NAME:	OBJECT32
ALIAS:	H2 REF = 1.
SUPPORT_TO:	31
SUPPORT_FROM:	96
COMMENTS:	[Where any parking space on a licensed parking lot is clearly marked as reserved], [no person other than the person for whom it is reserved shall park a vehicle in such space].

NAME:	OBJECT95
ALIAS:	H2 REF = 1.
SUPPORT_TO:	95
SUPPORT_FROM:	None
COMMENTS:	[Where any area on a licensed parking lot is designated as a Truck Loading Zone pursuant to the Streets and Traffic Bylaw]

NAME:	OBJECT96
ALIAS:	H2 REF = 1.
SUPPORT_TO:	32
SUPPORT_FROM:	95
COMMENTS:	[no person other than the person for whom it is reserved shall park a vehicle in such space].

Figure 18: Sample Text Structure Data Records

VITA

Surname: Proctor *Given Names:* Laura Jane

Place of Birth: East York, Ontario *Date of Birth:* 7 September 1952

Educational Institutions Attended:

University of Guelph	1971-1975
University of Victoria	1979-1990
McGill University - Summer Program	1987 (held at Zhejiang University, Hangzhou and Qinghua University, Beijing)

Degrees Awarded:

Bachelor of Science, University of Guelph	1975

Honours and Awards:

B.C. Science Council Graduate Research in Engineering And Technology Scholarship	1988 - 1990
Advanced Systems Institute Graduate Award	1988 - 1990
Dean's Scholarship, University of Victoria	1988

Publications:

Review of Computer Translation of Natural Language by W. Goshawke, I.D.K. Kelly, J.D. Wigg (1987) Language, 65, December 1989.

Review of Natural Language Generation Ed. Gerard Kempen (1987) forthcoming in Language, 67, June 1990.

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis: Conjunctions and Knowledge Acquisition From Text

Author [redacted]

LAURA PROCTOR
(Name in Block Letters)

24 July 1990
(Date)