CONJUNCTIONS AND KNOWLEDGE ACQUISITION

Laura Proctor

Department of Linguistic
University of Victoria

1. 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 brings 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.

2. CONJUNCTIONS 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.

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 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".

That is, function words traditionally treated as "empty" words, without significant meaning in themselves, do contribute to text meaning by providing some explicit connections among the meanings of 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, thus, 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. Although the different types of ordering (temporal, causal, etc.), or models, that conjunctions may suggest is another important aspect of their meaning, the present analysis does not address this issue. This is because, 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 (1986) 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.

3. KNOWLEDGE ACQUISITION

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.

4. DOCUMENT ANALYSIS

In this section, a method for interpreting and representing conjunctions in a discourse representation is presented in relation to the process of knowledge acquisition. 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 including 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. 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 be 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.

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 1 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 1: Function Words Classified by Direction of Contingency

In Figure 1, 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.

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.

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 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.

4.2 Overview of Application

Knowledge acquisition for expert systems is 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 project focussed 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. 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.

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 2.

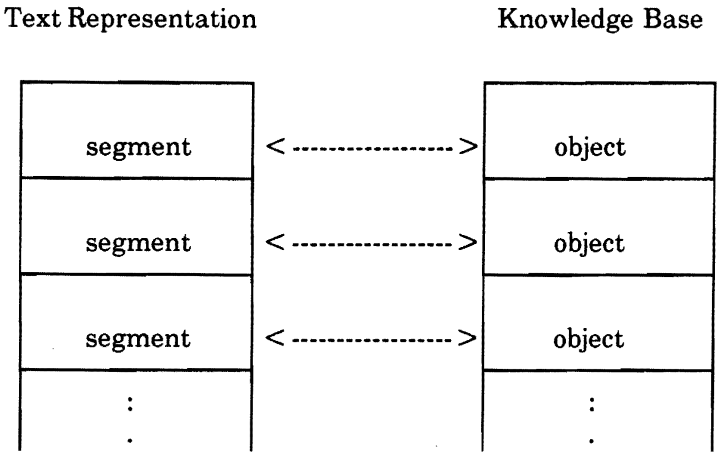

Figure 2:
Relation between Intermediate Text Representation and Knowledge Base.

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.**

The following is an excerpt from the bylaw.

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;

"parkinglot" 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 3:
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.

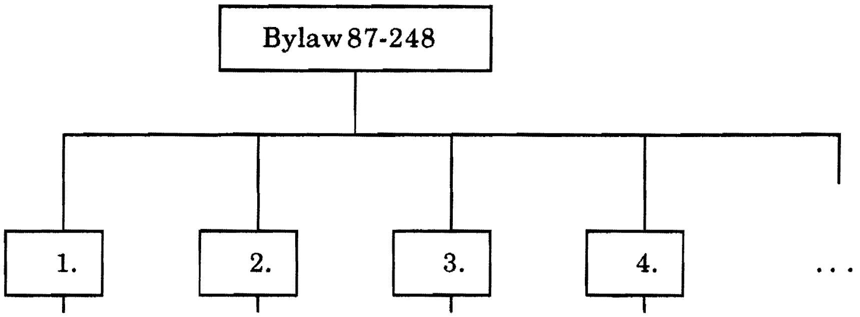

Figure 5: Structure of Bylaw Sections

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 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.

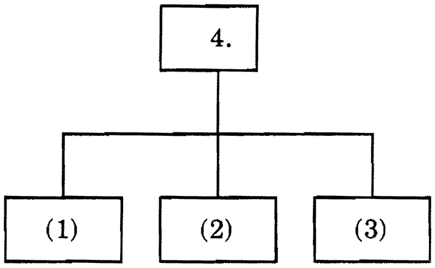

Figure 6: Structure of Bylaw Section 4

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.

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.

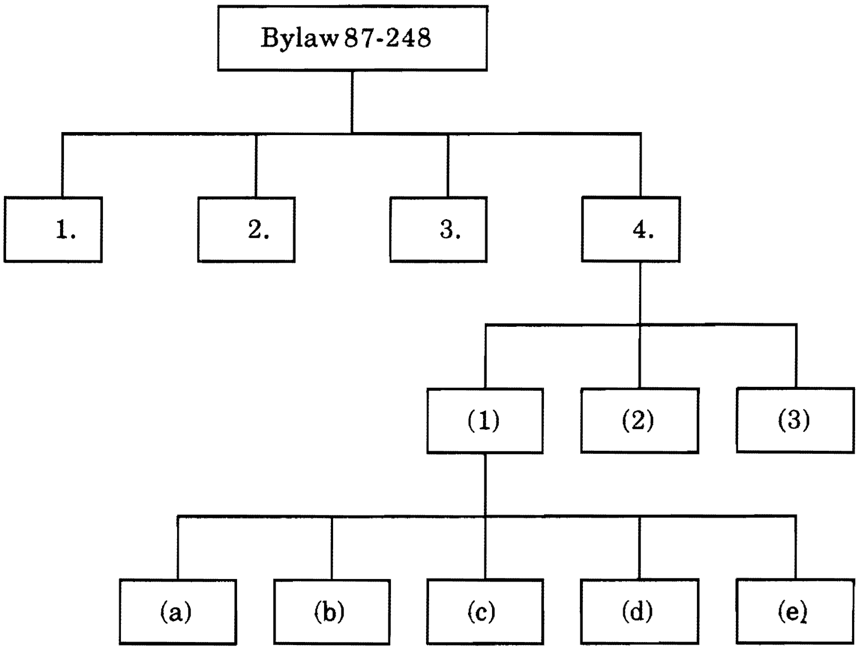

Figure 7: Bylaw Document Structure

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.

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.

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.

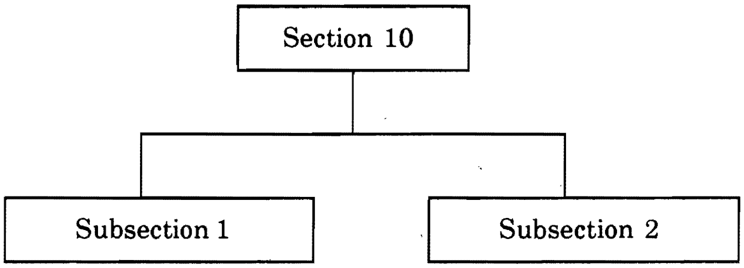

Figure 8: Document Structure - Section 10.

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 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 the author 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.

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.

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 conjunctions in the text are used to establish the support links between objects. The subordinate clause in Section 3., introduced by *unless*, expresses a condition for determining the status of the proposition expressed in the main clause. 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 the value of *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.

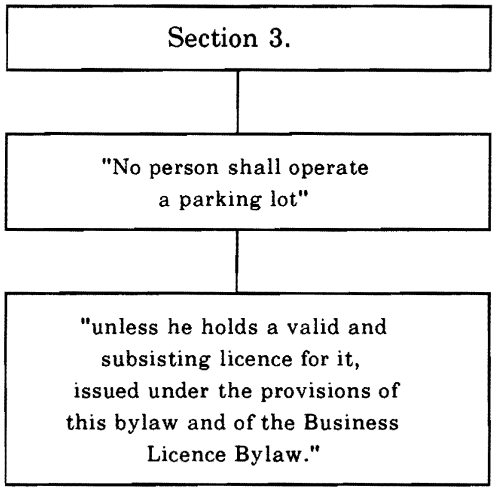
Section 3.
"No person shall operate a parking lot"
"unless he holds a valid and subsisting licence for it, issued under the provisions of this bylaw and of the Business Licence Bylaw."

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.

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).

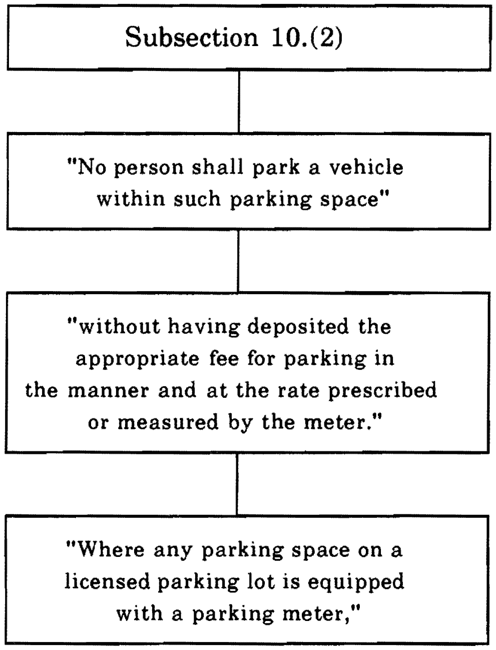
Subsection 10.(2)
"No person shall park a vehicle within such parking space"
"without having deposited the appropriate fee for parking in the manner and at the rate prescribed or measured by the meter."
"Where any parking space on a licensed parking lot is equipped with a parking meter,"

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).

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]."

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.

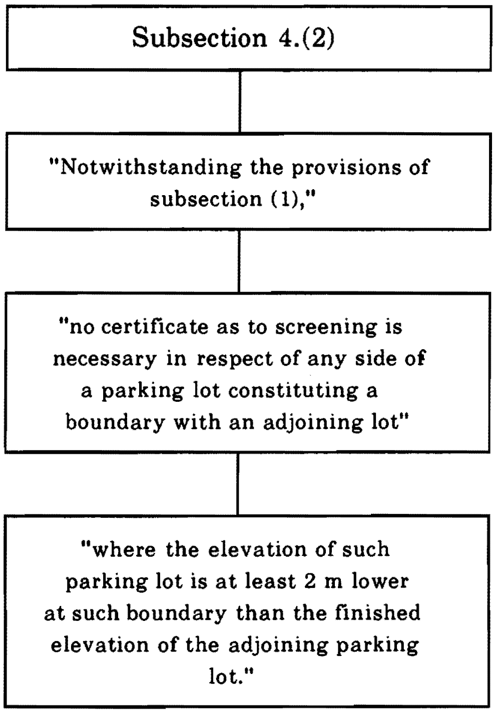
Subsection 4.(2)
"Notwithstanding the provisions of subsection (1),"
"no certificate as to screening is necessary in respect of any side of a parking lot constituting a boundary with an adjoining lot"
"where the elevation of such parking lot is at least 2 m lower at such boundary than the finished elevation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

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.

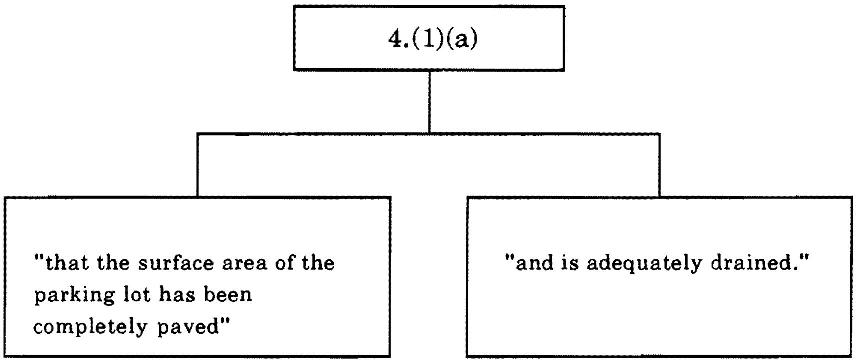
4.(1)(a)
"that the surface area of the parking lot has been completely paved"
"and is adequately drained."

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

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 Figures 10 and 11). From these illustrations it should be clear that the object representing the main clause will be the one which does not support any other 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.

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 here.

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.

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 an analysis of conjunctions, linguistic research and perspective has been applied to 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.

REFERENCES

Breuker, J., & Wielinga, B. (1987). Use of models in the interpretation of verbal data. In A. L. Kidd (Ed.), Knowledge Acquisition for Expert Systems: A Practical Handbook (pp. 17-44). New York: Plenum Press.

Conklin, J. (1987). Hypertext: an introduction and survey. Computer, 20, 17-41.

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

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

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. Computational Linguistics, 12, 175-204.

Halliday, M.A.K., & Hasan, R. (1976). Cohesion in English. London: Longman.

Johnson, S. (1987). Temporal information in medical narratives. In Sager, N., Friedman, C., & Lyman, M.S. (Eds.), Medical Language Processing: Computer Management of Narrative Data. (pp. 175-194). Reading, Mass.: Addison-Wesley.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. Psychological Review, 95, 163-182.

Martin, J. R. (1983). Conjunction: the logic of English text. In Petòfi, J. S., & Sòzer, E. (Eds.), Micro and macro connexity of texts (pp. 1-71), Hamburg: Helmut Buske.

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

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). A Grammar of Contemporary English. London: Longman.

Rudolf, E. (1988). Connective relations - connective expressions - connective structures. In Petòfi, J. S. (Ed.), Text and discourse constitution - empirical aspects, theoretical approaches (pp. 97-133), Berlin: de Gruyter.

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