DISCOURSE STRATEGIES MODEL: AN INITIAL PHASE FOR DISCOVERY OF THE FACT-BASED STATEMENTS FROM DESCRIPTIVE TEXT.

Bruce A. Calway and Ross Smith
Swinburne at Lilydale
Swinburne University of Technology
Email: bcalway@swin.edu.au

ABSTRACT

Fact-based conceptual modelling approaches seek, as one goal, to express detail about a universe of discourse (UoD) as elementary declarative statements, generalised from collections of data. A further resource available to the systems analyst, as the fact-modeller, is to discover fact-based statements from descriptive natural language information systems specifications. However, such extant formalised approaches as exist, that can assist the fact-based analyst to process textual resources, are incomplete. This paper proposes a discourse strategies model, for use as an initial process for the fact-based analyst processing descriptive text. The model is synthesised from extant linguistic and fact-based modelling approaches represented in literature. The model proposed is exercised using an example descriptive text.

KEYWORDS

Fact-based modelling; Discourse analysis.

OVERVIEW

Each textual resource is inherently compressed through linguistic referencing and ellipsis techniques. The application of these techniques by an author is aimed at making the resource readable, and within a particular universe of discourse (UoD), communicable. The author of a text uses many linguistic methods for this process, and it might be argued that for the fact-based analysis of a textual resource to be more thorough the reversal of these processes should be of primary importance. The discourse strategies model (DSM), proposed as an approach in this paper, calls for all implicit lexical material to be made explicit within the textual resource and that the simple sentence, as a single clause, forms the largest sentential element.

This approach typifies that of a data oriented representation of textual content. In that each clause is taken to represent a proposition and therefore a fact without regard for the clause's relative position to any other clause which exists within the text, nor regard for how the elements of texture function relative to each other. To achieve a clause based narrative of a textual resource involves the simplification or unpacking of discourse as written text into clause based descriptions of either simple (i.e. one clause sentences) or cosub-ordinate clause sentences.

This work was first reported in (Calway and Sykes 1996) and is further detailed as a series of strategies in this paper. This study takes as its premise the concept that a simple clause is expressed through one proposition (itself being a set of atomic propositions) and expressing one fact as a declarative statement (cf. Wintraecken 1990; Sykes 1994; van Dijk and Kintsch 1983; Clark and Clark 1977).

Having unpacked and simplified a text then allows the analyst to approach the text to isolate and report propositions to be found in the clause based statements. Several theories apply and are discussed as optional approaches (i.e. syntactic and semantic strategies (cf. Clark and Clark 1977; van Dijk and Kintsch 1983), and lexico-grammatical processes (cf. Eggins 1994; Winter 1982; Halliday 1994).

In combination with fact-based processes, syntactic and semantic processes could be incorporated as a production system and therefore be developed using an approach similar to augmented transition networks. The significant theoretical variance to previous research, which uses transformational approaches (cf. Dunn 1992; Meziane 1994; Su 1988), is the use of psycholinguistic theories for proposition comprehension (van Dijk Kintsch 1983); functional linguistics for clause complex theories (cf. Halliday 1994; Winter 1982); and strategies theory as a processing model (cf. Bever 1970; Clark and Clark 1977; van Dijk and Kintsch 1983).

Further, underlying the discourse strategies model is that "Understanding sentences as part of a discourse is a different process from understanding sentences in isolation". Also..... as yet there is no system able to parse arbitrary English text reliably". (van Dijk and Kintsch 1983:32)

The linguistic processes developed in the following sections are a synthesis of strategies taken principally from van Dijk and Kintsch (1983), Clark and Clark (1977), Eggins (1994), Halliday et al. (1976, 1978, 1985, 1994), and Martin (1992).

In the next section a background to the research is established and a formal research question is proposed. Arising from this the structure of the remainder of the paper emerges, as a discourse strategies model (DSM) is synthesised and subsequently executed.

NATURAL LANGUAGE MODELS

There are many examples of models of natural language and language use which take account of linguistic objects syntactically, semantically and pragmatically. A major assumption underpinning the model of discourse processing proposed in this paper is that information from each level interacts in and with each other. Semantic knowledge does not explicitly follow or equate to a syntactic analysis of sentential surface structure. However, both semantic and syntactic structure would seem to be required for conceptual modelling.

Calway and Sykes (1996) have shown that information can be lost when concentrating on text from a context free syntactic position. Also, van Dijk and Kintsch (1983) among others have empirically shown that we do not break discourse into elementary pieces and parse them independently. Chunks of discourse are processed relative to each other using a combination of structural and semantic strategies. Conceptual modelling not only seeks structure but also meaning or conceptual intent.

It is important that textual cohesion be retained as a step in the model because, unlike day to day discourse in which we might engage for the purpose of communication and comprehension, descriptive discourse places significant emphasis on the entities within the discourse and the role relationships which are extant. That is, this model places greater emphasis on the objects of direct interest by placing them at the focal point of a sentence (thematic, psychological subject) rather than concentrating for example on the actor undergoer (informal, logical subject) or subject/verb/object (grammatical subject) (Eggins, 1994:141).

Derivational rule governed processes like those of transformational grammars will produce structural descriptions of a sentence by syntactic parsing rules (an example of this approach for fact modelling is Dunn (1992)). However the process may be complex, and guaranteed success is only available if all the rules are correct and applied correctly. Some linguists argue (cf. Quirk et al. 1985) that language is an organism and therefore subject to evolutionary change, having a direct impact on the veracity of a derivational rule based approach. This point is further highlighted by the fact that no universal grammar is available for parsing natural language, as yet. Quirk, et al. (1985) would not claim to have solved all issues of grammar, even though they have documented in excess of 1500 pages of detail (including a 110 page index). This highlights the extremes which must be considered if technology is employed in a prescriptive and deterministic way.

The model proposed in the present work is rather like a hypothesis testing scenario, where no unique representation of text is assumed. This means that an initial working hypothesis may be understood but that this may be re-interpreted by further processing. Strategies theory (cf. Bever 1970) also takes account of the textual characteristics and the language user in terms of goals and world knowledge (a theme acknowledged by interaction of a systems analyst and domain expert in fact-based methods). The strategies detailed in this paper are an expression of the procedural knowledge about understanding discourse as information systems descriptions. However the work in no way forms the totality of possible processing strategies. It is plausible that strategies can be learnt and formalised, and therefore available to be automated as a production system shell.

A Textbase (as an outcome of the application of DSM) is generated as a collected declarative representation formed from the input discourse. The Textbase will be defined in terms of elementary statements expressed as declarative clauses and cohesive relations between entities/entity-types. This approach is similar to those proposed by Fillmore (1968), Wintraecken (1990), and Vendler (1967) for example.

Therefore the underpinning question which is investigated, in this paper, through the use of a literature survey and exercising of the proposed model using and annotated example is:

> Can a discourse strategies model be synthesised from extant fact-based data modelling and linguistics literature which enables the development of elementary fact-based statements from descriptive textual information systems specifications?

The balance of this paper addresses this question as follows:

- Review of the fact-based processes (Section 3);
- Discussion of the foundations upon which the discourse strategies model is built (Section 4, 5 & 6);
- Detailing of the discourse strategies model (Section 7 and 8); and
- Exercising the proposed model as an annotated example of the application of the model (Section 9 and 10).

FACT DISCOVERY USING FACT-BASED PROCESSES

Hirschheim et al. (1995:177-186) outline the fundamentals of conceptual modelling as typically represented in fact-based approaches. Fact-based approaches argue that the meaning of language is determined by the way in

which the elements of language correspond to entities and facts, as elements acting in some world. This accords with the goal of context free grammatical parsing through generative structures. Currently generative grammar, as algorithmic rule-governed processes, is used for parsing natural language declarative sentences, however, this procedure depends on automation and a holistic context free grammar.

For the proposed discourse strategies model it is important to note that the linguistic processes proposed are manually applicable, non-deterministic, open ended and highly context sensitive. As van Dijk and Kintsch (1983:31) suggest this does not preclude the rigour and objectivity required for a scientific theory or formalism. They remain open to automated strategies parsing by suggesting that both the rule based (context free) and strategy based (context sensitive) models can be implemented as an augmented transition network, currently a preferred parsing approach for context free grammars (cf. Wood 1970; Wanner and Maratsos 1977; Dunn 1992; Kaplan 1972; Meziane 1994).

The issue of context free versus context sensitive parsing offers a significant theoretical dichotomy that remains unresolved in conceptual modelling schools of thought (Hirschheim et al. 1995:198ff). It is not resolved or advanced further in this paper, which draws on existing theoretical assumptions of the fact-based school.

When considering sentence parsing approaches, understanding sentences as part of a textual discourse is different from understanding sentences in isolation from one another. While studying sentences in isolation may tell us something, it may also mislead us. With textual discourse there may be an elementary clause that is part of a complex sentence which itself forms part of a sentence sequence. This observation requires contextual considerations to be observed in terms of UoD coherence, texture and reference normalcy.

The proposed discourse strategies model does not seek a 100 per cent grammatical model of sentence analysis, but rather a normative approach which the fact-based analyst might plausibly apply to assist fact discovery when dealing with information systems specifications as descriptive text. Further, the model proposed does not deliver elementary facts or fact-type sentences sufficient for the immediate application of data base design approaches. The proposed model could best be described as providing an induced proposition set and consequent declarative sentence narrative. These statements could be used as input to conceptual modelling methods, where such generalised facts would be expressed or verified using specific UoD examples, sufficient that a data base designer would be sure of the accuracy of a deduced design framework.

Wintraecken (1990:41-104), when describing the NIAM information analysis method, offers the theoretical foundation for using natural language on which this dissertation also draws. This is substantially the same as that of Clark and Clark (1977) and van Dijk and Kintsch (1983), where propositions are offered as the representation of semantic detail expressed as a limited sentence type and syntax.

Once propositions as declarative statements are catalogued as a Textbase then the application of a conceptual modelling formalism can follow. For example, one could apply conceptual schema design process (CSDP) steps 1 and 2 from the NIAM conceptual schema design procedures of Nijssen and Halpin (1989) to the Textbase generated from the application of the discourse strategies approach.

A fact-based information model is stated as expressing entities and their relationships, which occur within a specified UoD. This being the case, there are three indicators to start the process strategies framework for fact discovery and declarative statements development:

- entities, entity-types;
- relationships between entities, fact-types; and
- a specified UoD for the context of the previous elements.

There are a number of general processes which are defined for a fact-based analysis of discourse. Burg and van de Riet (1996) have stated that the starting process is to:

1. Transform the user sentences into elementary sentences;
2. Find common nouns in the elementary sentences (as entity-types);
3. Find transitive verbs in the elementary sentences (as actions);
4. Identify adjectives (as attributes); and
5. Identify adverbs (as attribute relationships).

Similarly Edmond (1992:220) suggests that the fact-based analysis stages should be:

1. Uncover the relevant entity-types and the fact-types that join them;
2. Look for any uniqueness constraints involved in each fact-type; and
3. Construct record types by merging fact-types, where appropriate.

However these approaches rely upon a sentence-by-sentence analysis with no regard in the first instant to inter textual detail.

STRATEGIES THEORY

In addressing alternative theoretical foundations Psycholinguistic approaches as proposition comprehension strategies form a further theoretical focus for the present research, for comprehension of discourse. They were introduced by Bever (1970), extensively developed and experimentally tested by Clark and Clark (1977) and form the basis of strategies for discourse comprehension of van Dijk and Kintsch (1983). Reasons for advancing a strategies based model were that rule based (a term used by van Dijk which is different in intent from earlier references to the term rule-based used by Hirschheim et al.) systems which linguists use to parse sentences were implausible as psychological process models. Semantic proposition comprehension strategies best emulate the heuristics of the intuitive fact-based analysis approach (the subject of a further paper by Calway, yet to be published).

The present research suggests that fact-based analysts using intuitive heuristics to discover facts in textual discourse actually use psychological processes. Secondly, the calculation resource demanded for a rule-governed grammatical parsing simply exceeds human processing limitations, as demonstrated by Noble (1988). Moreover, as stated earlier, a definitive grammar of English does not exist and therefore context free grammar parsers remain incomplete.

Dunn (1992), and Meziane (1994), et al., have shown the complexity of attempting to implement context free grammars through automated augmented transition networks. Van Dijk and Kintsch (1983:28) suggest that "Strategies were simpler". As they say, strategies do not guarantee the right results, but it is plausible that people can, and in fact do, parse sentences on the basis of strategies.

What then is the nub of the psycholinguistic approach as a strategy? It is using the proposition as the basis for parsing, as demonstrated by Bever (1970), Clark and Clark (1977) and confirmed in van Dijk and Kintsch (1983) where they concluded that evidence for the psychological reality of prepositional units is overwhelming.

Howard (1983:304) commented that there is a failure in the derivational theory, due to complexity, which indicates that such approaches have not worked. Howard continues, "... psycholinguists have tried a different approach. They argue that perhaps language comprehension is more accurately viewed as involving a set of heuristics, i.e. a set of rules of thumb that provide guidelines for uncovering the propositions encoded in the string of words."

Therefore processing strategy theory assumes that people use heuristics rather than algorithms during language comprehension. The derivational theory of complexity posits syntactic processes (e.g. transformational rules) that are largely independent of semantic strategies. Whereas the processing strategy approach assumes that both syntactic cues (such as word order) and semantic information (such as word meaning) are used throughout the processing of a text (Howard 1983:306,307)

TEXT AND TEXTURE

Using proposition comprehension strategies only provides a partial approach to text processing for conceptual modelling. Within descriptive text there could be several referent and cohesive elements/items that are not explicit in terms of lexical labels. To emphasise these items from a textual rather than a sentential perspective, the proposed model also draws on strategies from functional linguistics disciplines (Halliday 1994; Eggins 1994; Martin 1992)

Referencing elements contained in sentences act as textual elements to provide cohesion. (Halliday and Hasan 1976:1) state that text is "any passage (of language), spoken or written, of whatever length, that does form a unified whole". For any research to move from a single clause to a clause complex as a sentential structure, and further to sentence sequences, requires recognising or differentiating between text and non-text. To analyse discourse requires examination of the nature of texture. Texture results from contextual coherence (coherence of register and genre) and cohesion (internal semantic ties through which parts of the text depend for their interpretation or comprehension on other parts (Eggins 1994:95)).

Eggins suggests (1994:95-109) that text can exist as a single clause or multiples of clauses and depends on texture created by:

- Referencing - Endophoric (retrieved from within the text) and Exophoric (retrieved from shared immediate context);
- Lexical relations - how lexical items (nouns, verbs, adjectives) and event sequences are used to relate the text consistently to its focus; and
- Conjunctive relations - how the author creates and expresses logical relationships between the parts of text (elaboration, extension, enhancement).

Elementary facts, and formal facts (as rules or fact-type sentences) are to be expressed as simple declarative clauses and constraints which act to modify the relationships expressed in and between entities. The strategies envisaged for expressing these items as a fact oriented resource include sentence unpacking and simplification by splitting clause complexes and recording statements expressing the main, property, and constraint propositions.

STRATEGIES SELECTION - PROPOSITION DEVELOPMENT

The strategies framework provides for specific selection and application of processes which underpin the development of the discourse strategies model for use by a fact-based analyst desiring a more methodical approach to declarative statement discovery from a textual resource. Central to this position is that the intuitive and documented processes for fact discovery displayed, by fact-based analysts when using a textual information systems description, were in fact reflected in psycholinguistic discourse comprehension processes for discourse analysis.
The starting point for developing strategies is the definition of the relationships between clausal or sentential structures and consequently prepositional structures. As noted, a schema is given by van Dijk and Kintsch (1983:120) and Wintraecken (1990:42-43) where one simple clause corresponds to one proposition, that proposition representing one fact which is the referent of the clause in some possible world. This is the elementary level; above this level exists the clause complex that will reflect a composite schema (of clauses, propositions and facts), i.e. one schema may be coordinated or subordinated with respect to another schema. Van Dijk and Kintsch (1983) gives the example:

> The professor hired an assistant who had written a dissertation on discourse comprehension.

This sentence is said to denote one fact, i.e. the action of hiring somebody who has certain properties. This fact is complex due to the property of one of the arguments being described in terms of another (previous) fact. The sentence structure suggests the embedded fact has no independent function other than to specify a (main) fact. To illustrate this, a further example might be:

> A car is of a particular model and is given a serial number by its manufacturer that is unique among the cars made by that manufacturer.

Here the main clause and therefore main proposition and main fact is that the manufacturer gives a serial number to a car. There also exists one coordinated clause and therefore fact (1 below) and one subordinated clause and therefore fact (2 below), as follows:

1. a car is of a particular model made by a manufacturer,
2. a serial number is unique among the cars made by the manufacturer.

These three facts (i.e. one main and two relative facts) are sufficient for a fact modeller to begin the conceptual modelling process as outlined by authors such as Wintraecken, Halpin, and Edmond, et al.

This small exercise points to the strategies required above the clause or proposition level, i.e.:

- how to discover the proposition hierarchy in the clause complex; and
- the importance of knowing how semantic information is placed in, or distributed across, several clause complexes.

Before stating the processing strategies for discovery of the main and relative clauses there is a question of clause independence, where the sentence as a clause complex is re-phrased into individual clauses. This re-phrasing process creates a situation where each clause identified contains a main proposition and therefore denotes an independent fact. The proposition schemata for such a structure could then be linked textually by degrees of closeness (van Dijk and Kintsch 1983:122-123).
A sentence or clause complex is the highest level of grammatical and therefore syntactic structure. However in text, propositions as semantic structure extend beyond those bounds. The semantic relation can be achieved through text forming resources such as conjunctive relations, i.e. relations between messages or between larger complexes. There are two conjunctives, external (ideational) which are seen in a series of events or as a narrative, and internal (interpersonal), setting semantic relations between steps in an argument.
There are three levels of cohesion that will require analysis:

- cohesion between entities within a clause;
- cohesion between entities within a clause complex; and
- cohesion between entities within a sentence complex (text).

The grammar of a clause accounts for how various parts participate or function within a clause. This can be extended to include how clauses function together. Once the textual resource moves beyond a single sentence then traditional grammars offer little support.

A fourth level of cohesion, that is not dealt with in this work but which forms the basis of considerable debate in information systems development research, is that of context free and context sensitive text best expressed by Hirschheim et al. (1995) in their book. This traces the various philosophical and practical developments in Data Modelling.

A distinction must be drawn between what is the actual outcome from a textual simplification or reconstruction of a textual resource and that of a data driven intuitive approach. If, as is the case with the present research, the input to the process is descriptive text, then consideration must be maintained as to the information content through cohesion in the text, as distinct from the data structures which neglect the cohesion that may exist between clause and sentence complexes. This is exemplified when sentences are split on a coordinating conjunction boundary, as suggested by Falkenberg (1986:6.1).

Hirschheim et al. (1995:26,27) make little distinction between information, conceptual and data modelling, preferring their definition of the later to be inclusive of the former. They do state in passing that it is the information or conceptual level that deals only with linguistic modelling. It is this which forms the basis of the output from the discourse strategies model approach. Therefore it is the information model as a Textbase that forms the input to the modelling formalism of fact-based approaches.

A DISCOURSE STRATEGIES MODEL

There have been several strategy groupings identified, in this paper, for the fact discovery process of a fact-based analyst:

1. Clause and therefore declarative statement discovery strategies sufficient to identify inter and intra textual cohesion elements within the text;
2. Sentence clarification and simplification strategies which remove all implicit referencing and ellipsis by substitution; and
3. Declarative statement development strategies for identification and collection of expressed elementary statements.

The expression of an elementary fact is described in terms of a simple declarative sentence as a single clause. The availability of descriptive discourse for expressing information systems structure is common, however as indicated earlier, documented heuristics used by the fact-based analyst for comprehension and modelling of fact statements are lacking.

Therefore, this section begins with strategies, which it is believed will help disclose the clausal and prepositional content of a descriptive textual discourse. It is plausible that discourse simplification should be considered as the starting point for expressing propositions within and across the sentential content of text. Each of the strategies discussed below, individually or when combined, allow the fact-based analyst a plausible model for analysing textual material for the development of a collection of statements which express the UoD information system specification.

The previous sections have indicated that the starting point for fact discovery from a descriptive text is to make explicit how the text is implicitly structured as constituent parts. An important component of this approach is to understand that although comprehension, and to a degree interpretation, are important, these are only a means to an end, which is a detailed set of declarative statements made available for the conceptual modeller.

As with any text it is shown that considerable content is sacrificed through methods such as ellipsis and substitution, endophoric and exophoric referencing, etc. For a more complete treatment of propositions, and therefore facts, as much content as possible is retrieved prior to the declarative statement Textbase development.

The desired sentence structure for the proposed Textbase is as:

- indicative affirmative clauses (i.e. declarative); and
- indicative conditional clauses.

Each clause will be asserted as true (cf. Nijssen and Halpin 1989:16).

Also, all property facts (i.e. facts which have only an attribute relationship expressed (e.g. a car is of a particular model)) should be tested for an associated activity fact to which the property ascribes. The identified detail may not be relevant to the conceptual modeller but does give opportunity to discover how a state of affairs came about. Further, proper names or data examples used in sentential material are to be associated with their entity-type and entity-label detail, as prepositional statements, as this forms an integral part of developing fact sentences of the kind expressed in a fact-based formalism (e.g. NIAM):

e.g. John is the First-name of a Person *or*
Person with First name John.

Within descriptive text it is unlikely for the analyst to observe sufficient (or any) examples sufficient for a full understanding of all data constraints, unless they are explicitly mentioned within the text, as is often the case:

e.g. The maximum number of employees per department is 10.

Textual resources, when considered conceptually, can be handled in one of two ways to deliver the prepositional structure as a semantic representation of the text. One is to take the text and represent the semantic detail as a narrative of elementary clauses and therefore elementary propositions which express a one-to-one fact sentence relationship. The other is to deal with the text as a set of sentences (clause complexes) which will contain one main proposition per sentence and therefore one main fact as a proposition hierarchy.
The first requires that each clause be identified and made independent of others within the text, whereas the second, would suggest one main clause which is supported by statements which relate to, or qualify, the main clause. It is the first option which is most like the outcome proposition statement set of the fact-based approach. The proposition comprehension strategies discussed in Calway (1998:116-146) were shown to take content words and create a series of elementary propositions without regard to any implicit context or hierarchy within the text. Also, if an analyst were to look for one main clause, and therefore one main fact, that clause would be the one which expresses the main finite predicate structure. All other clauses would be considered relative, in that they are qualifying the main clause. This is exemplified in the functional approach to mood analysis as an understanding of sentences Eggins (1994).
As a proof-of-concept it is necessary to express the discourse processing findings of the present research as a synthesised theoretical strategies model which can be aligned to existing fact-based analysts heuristics (to be reported in Calway, yet to be published). The discourse strategies model will be exercised subsequently in this paper, using the textual examples given by Wintraecken (1990:41). As a future work, the model could be automated and tested in a longitudinal study as to the process and product of the theorised approach to analysing descriptive text.

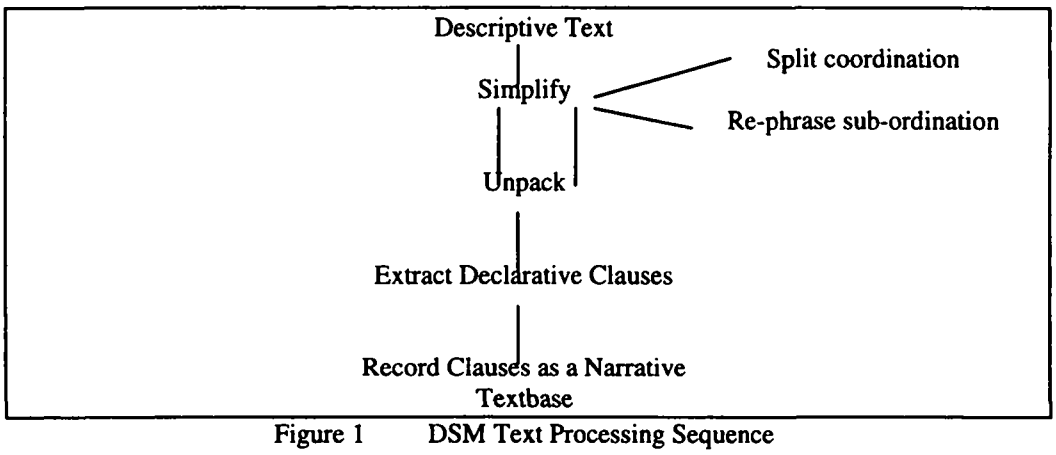

Figure 1 DSM Text Processing Sequence

The discourse strategies model described in the following sections has a series of steps, with each step having a set of discourse strategies to apply for various textual problems as they may be encountered by the fact-based analyst. The text processing sequence as shown in Figure 1 includes:

- Splitting - where compound sentences (i.e. co-ordinated clauses) are split at points where conjunctions are operative and clause boundaries are identified;

- Unpacking - replacement of linguistic items of referent, substitution and ellipsis for each simplified sentence (clause complex);
- Re-phrasing - taking each clause and phrase complex and identifying clause/phrase boundaries. The contents are re-phrased into independent clauses as a narrative where the above steps can be re-applied; and
- Extraction - extracting the content words, identifying the main predicate and developing an elementary proposition as a declarative simple sentence.

A flow diagram (Figure 2) provides a step by step indication of the strategies that might be applied, in what order, and what decisions might be taken during the analysis process. The process steps need not be applied in a linear form; it is more important to choose the appropriate processing strategies than to attempt a prescriptive hierarchy of actions. The DSM strategies are summarised in the following sections (a detailed analysis of the DSM strategies development can be found in Calway (1998).

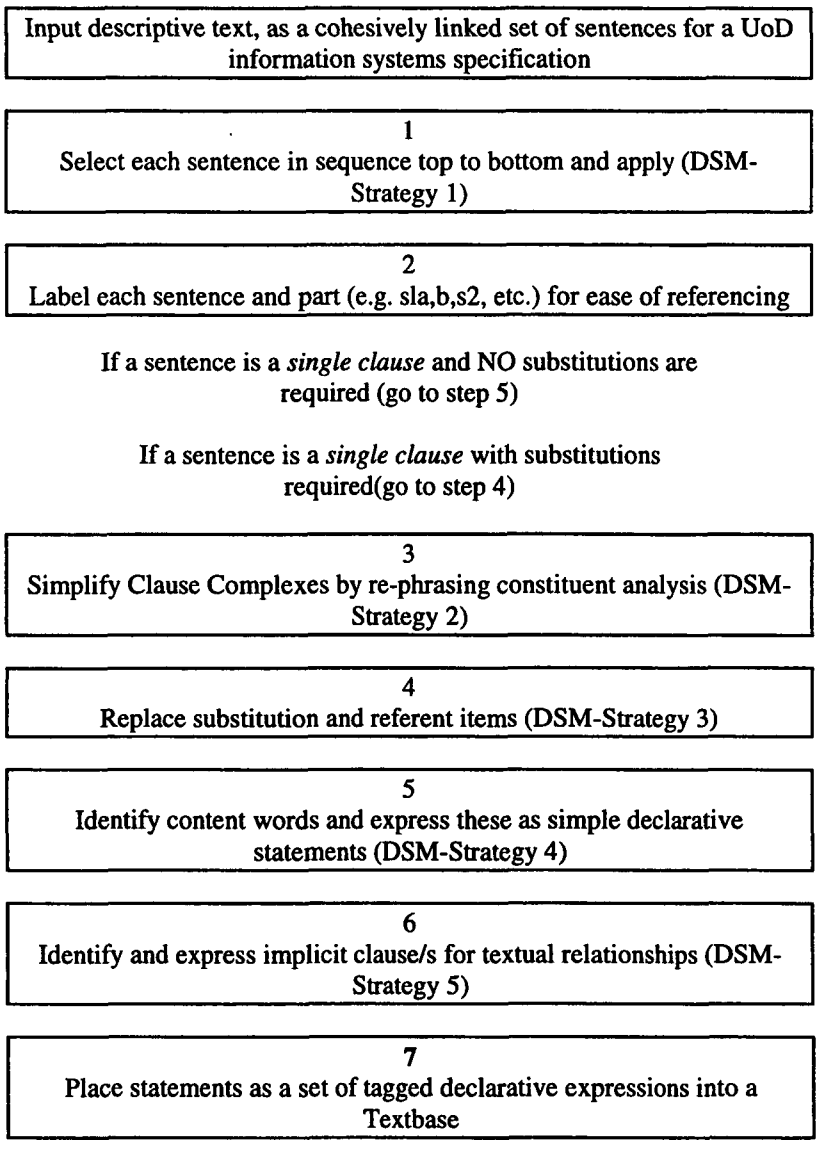

Figure 2 Model Flow Diagram

(Note: Strategy numbers shown i.e. Strategy 1, 2, etc. refer to specific, related strategies discussed later in this paper)

ELEMENTARY STATEMENT DEVELOPMENT

There are at least two approaches, which can be considered when developing elementary statements from unconstrained textual discourse. Of the two approaches it is the first which is taken as constituting the discourse strategies model. The second approach is more likely to replicate the current fact-based text processing heuristics of the fact-based analyst. This second approach could form the basis of a further research project that could reveal in more detail specific text processing heuristics or conditional approaches.

Approach One

- Sentential Splitting and Unpacking

The text is analysed for opportunities to split compound sentences into less complex structures. This is achieved by observing conjunctive items which join clauses together (e.g. and, but, or, that). Divide the sentence into two and discard the conjunction, making each new part into an independent clause or clause complex.

Unpacking of text means replacing significant contextual items in each new or existing clause complex so that each new structure can make an independent sense. This is achieved by replacing non lexical items, substituted items and ellipsis items within the structure. Each unpacked clause or clause complex may offer new opportunity for further splitting before moving to declarative statement development, although, it is not mandatory to split clause complexes into smaller parts.

- Extracting Propositions, Facts, Rules, Constraints

Once the text is developed so that clause complexes are simplified, then elementary declarative statement can be extracted. It will be noticed that each clause or clause complex will have one main proposition and possibly one or more relative propositions as attributes or constraints on the entity/entities in the main fact. As one approach the content words (i.e. nouns and predicate) are selected and written as a declarative simple clause based sentence. Although declarative statements will be extracted for every clause in the text initially, this does not mean that they must be used in any conceptual schema modelling process.

Approach Two

A second approach is to identify each main clause at the surface level of a sentence and dissect each clause identified. This approach accepts the main clause in-situ, with the development of propositions emulating the semantic strategies approach expressed earlier. Such an approach relies on the ability of the fact-based analyst to identify the main proposition through the content words. This approach is more likely to replicate the current text processing heuristics of the fact-based analyst (research reported in the PhD dissertation (Calway, 1998)).

Approach One - Sentence Splitting and Unpacking Strategies

Before any attempt is made to process textual material it is necessary to determine a starting point. Within the literature reviewed there is no explicit statement as to whether a top down approach should be taken, when processing text. The use of a top down approach is implicit in the work of Eggins 1994; Wintraecken 1990; and van Dijk and Kintsch 1983); and therefore, for the present research, a top down, sentence by sentence sequential approach has been taken as a basis for the initial research.
The goal of the following strategies is to identify how the various parts of a sentence are functioning. The text, and therefore a sentence, can be split into components and have all implicit detail made explicit. Two main conjunctive relationships exist which can be identified and allow sentence simplification: namely co-ordination and subordination. Coordination is expressed as follows:

DSM-Strategy 1 Whenever co-ordinated clauses are detected joined by a coordinating conjunction, then begin a new clause and remove the conjunction.

Co-ordination is where co-equal clauses are joined by a co-ordinating conjunction (and, or, but, nor) in order to create a single sentence.
This strategy reflects the observation that conjunctions are used to connect ideas or to remove repetition between sentences in a textual discourse.
Sub-ordination is expressed as follows:

DSM-Strategy 2 Whenever a sub-ordinate clause or clause component is detected;
1) expand the word or phrase and add a co-ordinating conjunction to produce a connected clause sentence which can then be split at the co-ordinating conjunction OR
(2) move the sub-ordinate clause to the head of the clause complex and separate the clauses as if co-ordinated (refer DSM-Strategy I.)

It is possible to simplify a sentence structure by taking account of relative, *subordinating and complement* elements of a sentence. Often these elements can be made independent through re-expressing the clauses or clause constituents as co-ordinated clauses which can then be split further, as follows:

DSM-Strategy 2a Use the first word (or major constituent) of a clause to identify the function of that clause in the current sentence.

DSM-Strategy 2b Assume the first clause to be a main clause unless it is marked at, or prior to, the main verb as something other than a main clause.

This strategy also distinguishes main from qualifying clauses by noting the clause markers.

Additional optional syntactic strategies include:

Strategy 2c Whenever a function word is used, begin a new constituent larger than one word.

Function words (e.g. pronouns, determiners, qualifiers, propositions) are easy to detect and very reliable cues to constituent structure. Second, function words signal the type of constituent.

Strategy 2d After identifying the beginning of a constituent, look for content words appropriate to that type of constituent.

As with Strategy 2c, there is a correlation between the variety of function words with those of content words (e.g. nouns, verbs, adjectives, adverbs). Most content words cannot be identified unambiguously and are therefore dependent on the use of function words for disambiguation.
Although strategies 2c and 2d can clarify the constituents of a clause there is a requirement for additional strategies to identify the clause and how it functions. (Halliday 1994; Eggins 1994; and Martin 1992), when developing functional linguistic concepts, suggest that the mood constituent of a clause identifies what the function of a clause complement will be, as follows:

DSM-Strategy 3 Replace all implicit lexical details for the text. (ie. Identify all ellipsis, substitutions and referent co-ordination and subordination (endophoric/exophoric referencing).

The application of this strategy allows the analyst to unpack the implicit entity and role relationship substitutions encapsulated textually within the discourse.

Sub-strategies include:

DSM-Strategy 3a When a pronoun is identified look for the referent to which it is associated and replace the pronoun with the referent.

DSM-Strategy 3b Endophoric and exophoric referencing - analyse the text for the referent and replace the reference with the referent lexical items.

DSM-Strategy 3c Ellipsis analysis and Substitutions - analyse each sentence for ellipsis clauses or phrases expanding the sentence with the ellipses clause or phrase.

Approach One - Strategies for Extracting Propositions, Facts, Rules, Constraints

The goal of the following strategies is to determine how each sentence was meant to be utilised. This is achieved by following two working principles (van Dijk and Kintsch 1983; Clark and Clark 1977: 72-85):

- reality - concerned with the substance; and
- co-operative - concerned with the way ideas are expressed.

As stated earlier, for fact-based approaches, there is the assumption that the writer of the information systems description means what is stated and that sufficient UoD knowledge is available to interpret sentences.
The following strategy is suggested:

DSM-Strategy 4 Using content words alone, build declarative statements that make sense using the subject/finite/predicated/complement/adjunct structure.

Within the sentence which has been subjected to Strategies 1 to 3, there are contentive (content words) items which constitute the substance of the sentence and therefore the substance of the propositions. It cannot be stated that a declarative statement at this point constitutes an elementary fact until further conceptual schema design processes have been undertaken.
One should apply the functional analysis processes to determine the mood constituents and therefore assist in determining the Subject, Finite, Predicate, Complement and Adjunct structure. This allows consequent parsing of these elements into elementary (declarative) statements. Sub-strategies include:

DSM-Strategy 4a **Identify the Subject (psychological) and Finite Mood items for a sentence or clause.**

DSM-Strategy 4b Identify the Predicate, Complementary and Adjunct Mood items for a sentence or clause.

These processes (strategies 1-4) in large part satisfy the stages of fact-based analysis suggested by Burg and van de Riet (1996), and Edmond (1992), detailed earlier in this paper.
To identify and express, implicitly, clauses for textual relationships the following strategy is suggested:

DSM-Strategy 5 Develop elementary statements to represent the indicative relationship which exists cohesively between clauses/entities(types) within the text.

Sub-strategies include:

DSM-Strategy 5a Whenever a proper name or literal (i.e. a name or literal given to represent one real or imagined person, place, or thing) is identified also discover the referent label and entity type for that item(entity).

There are two processes required as part of this sub-strategy:

- identify the lexical noun, as entity type and label type, for the proper name or variable of the lexical noun. (e.g. person with last-name, Johnson)
- record a statement which expresses this proposition.

DSM-Strategy 5b Whenever a thematic connection is implicit within a discourse create a linking proposition and tag the proposition as implicit.

The most common text construction indicates some relationship at the type level for entities and then leaves the connection between specific examples within the text as implicit thematic references.

DSM-Strategy 5c Review the Adjunct items for constraint and qualification phrases and create a proposition statement to express the detail.

Resultant Textbase

The outcome from the analysis is a collection of elementary statements as declarative representations. These statements represent both elementary and formal fact options, with referent constraints or qualifying propositions:

- Elementary facts - in the Textbase contain a declarative statement with one or more of the identified entities being a literal or proper noun as an instance;
- Formal facts - in the Textbase represents declarative statements identifying non instance entities as label or type entities;

- Fact constraints - as qualifier or conditional statements operating on an entity and expressed as an indicative conditional statement; and
- Implicit facts - being those statements derived when checking for entity label and type detail expressed for an identified entity as a literal or proper noun.

All detail is expressed as declarative clause based simple statements that can be asserted as true for a given UoD. However, in no way should they be taken as other than indicative of the possible statements for a conceptual model until such time as a full set of specific examples and formalisms have been applied as a part of any fact-based conceptual modelling approach. Equally, as stated earlier, the elementary statements expressed within the Textbase should not be equated to elementary facts as expressed by any fact-based formalism (e.g. NIAM).

Approach Two Strategies

Although the use of a psychological model, where propositions are discovered and recorded as a hierarchy, is a plausible model there have been no strategies synthesised or tested in this dissertation. However, strategies could well be incorporated into further research which tests the veracity of one model synthesis over another (i.e. narrative vis-Avis hierarchy).

ANNOTATED EXAMPLE

The following is a step by step rendering of declarative detail using the discourse strategies model as described above. As such, this example should facilitate understanding of the DSM synthesised in earlier sections.

The example text is taken from Wintraecken (1990:4 1) which states:

> "In this company, there are employees and departments. Employees have last names and departments have department names. Employees can only work for one department. Johnson and Peterson work for the Purchasing Department and King works for the Personnel Department. A department may not consist of more than 10 employees."

A step by step rendering of declarative detail using the discourse strategies model could proceed as follows:

- Apply Strategy 1 to the body of text as far as possible, splitting sentences at the co-ordinating conjunction.
  In this company, there are employees *and*

  ... departments.

  Employees have last names *and*

  departments have department names.

  Employees can only work for one department.

  Johnson *and*
  Peterson work for the Purchasing Department *and*

  King works for the Personnel Department.

  A department may not consist of more than 10 employees.

- Strategy 2 allows for splitting of clause complexes, however there are no obvious candidate sentences in the analysis.

- Apply Strategy 3 as the approach to filling the void left by splitting the sentence and number the sentences as a reference back to the originating text. Also it is practical to sequence the new text structure for further reference and connection with the Textbase.

S1a In this company, there are employees.
S1b In this company, there are departments.
S2a Employees have last names.
S2b Departments have department names.
S3 Employees can only work for one department.
S4a Johnson works for the Purchasing Department.
S4b Peterson works for the Purchasing Department.
S4c King works for the Personnel Department.
S5 A department may not consist of more than 10 employees.

- Apply Strategy 4 to the body of text in order to discover content words (**Subject,** *Finite,* **Predicate,** complement, Adjunct).

S1a In **this company,** *there are* employees.
S1b In **this company,** *there are* departments.
S2a **Employees** *have* last-names.
S2b **Departments** *have* department-names.
S3 **Employees work for** one department.
S4a **Johnson works for** Purchasing Department.
S4b **Peterson works for** Purchasing Department.
S4c **King works for** Personnel Department.
S5 **department** *consists of* <= 10 employees.

The first and main proposition can be determined and expressed as an indicative (declarative) clause structure. These clauses represent the main elementary propositions and therefore facts in a Textbase. For the purpose of comparison with the set of propositions offered by Wintraecken (1990:43) some re-phrasing and article elimination have been activated on the above declarative statements:

p1 *company* has *employees.*
p2 *company* has *departments.*
p3 *Employee* has *last-name.*
p4 *Department* has *department*
p5 *Employee* works for (one) *department.*
  p5a *Employee* works for *department.*
  p5b maximum 1 *department per employee.*
p6 *Johnson* works for *Purchasing Department.*
p7 *Peterson* works for *Purchasing Department.*
p8 *King* works for *Personnel Department.*
p9 *department* consists of <= *10 employees.*
  p9a *department* consists of *employees.*
  p9b Maximum 10 *employees per department.*

- Using Strategy 5c produces a representative set of elementary statements for sub-ordinate clause/s. These propositions, when expressed as indicative clauses, represent property and constraint facts and are represented in the Textbase (refer to the propositions 5 and 9 above for an indication of what could exist).

In some statements data is present as proper names, etc. Where these are identified they also should be represented by a fact statement which expresses the example, its entity-label, and entity-type where available by induction of plausible propositions.

p10 Peterson is the last-name of Employee
p11 Johnson is the last-name of Employee
p12 King is the last-name of Employee
p13 Purchasing Department is the department-name of Department
p14 Personnel Department is the department-name of Department

Having completed the discovery process from the textual resource it is important to pass the outcome Textbase on to the conceptual modelling processes to further be developed using example data which operate within the UoD and which form the fundamental input into fact-based conceptual schema formalism.

ANNOTATED EXAMPLE COMPARATIVE ANALYSIS

Table 1 is a comparative analysis of the results for the proposition set offered by Wintraecken (1990:43) and the demonstration study argued above:

(Wintraecken 1990) Proposition Set	Demonstration Study Equivalent
W1 There are employees.	P1
W2 There are departments.	p2
W3 Employees work for departments.	p5a
W4 An employee may only work for one department.	p5b
W5 There are last names. (this would be considered an atomic proposition and therefore part of (W7)	
W6There are department names. (as for W5 but part of W11)	
W7Employees are denoted by their last names.	p3
W8Johnson is the last name of a certain employee.	P11
W9Peterson is the last name of a certain employee.	P10
W10 King is the last name of a certain employee.	p12
W11 Departments are referred to by department names.	p4
W12 Purchasing is the department name of a certain department.	p13
W13 Personnel is the department name of a certain department.	p14
W14 Johnson works for the Purchasing Department.	p6
W15 Peterson works for the Purchasing Department.	p7
W16 King works for the Purchasing Department.	p8
W17 A department may not contain more than 10 employees.	p9a,b

Table 1 Demonstration Study Statement Comparison

It can be seen from this example that a simple textual example can be easily developed into an independent set of declarative statements each expressing a single fact. However, it can also be seen that Wintraecken has not identified fully the content or constraints operating. This is observed by the lack of reference to 'This company' in W1 and W2; and the inclusion of atomic propositions W5 and W6. As a fact-based conceptual modelling approach, Wintraecken may be eliminating non relevant data items (e.g. where only one value is valid for a data item). However, as has been argued earlier, design issues should not be considered until after a full analysis of the text has taken place. It can only be assumed that W3 is meant to represent the relationship of employee and department (p5a) and the inverse (p9a), and that W17 expresses a constraint only.

SUMMARY

In this paper, the fact-based processes have been reviewed, and the foundations upon which one might build a discourse strategies model have been summarised. Building upon this base, a discourse strategies model has been synthesised and illustrated with an annotated example. As a consequence, the question posed at the commencement of this paper (i.e. can a discourse strategies model be synthesised from extant knowledge which enables this development of elementary statements from descriptive textual information systems specification?) is answered in the affirmative.

This paper has been limited to reporting the discourse strategy model synthesis. Further aspects of the wider research project, of which this is part, form the content of a PhD dissertation (Calway 1998), and further investigation of the model, and its application, is the subject of continuing research.

REFERENCES

ABBOTT, R. (1983) "Program Design by Informal English Descriptions", **Communications of the ACM**, Vol. 26 No. 11, pp. 882-894

ALLEN, J. (1987) **Natural Language Understanding**, Benjamin Cummings Publishing Company.

BEVER, T. G. (1970) "The Cognitive Basis for Linguistic Structures", in Hayes, J. R. (ed.), **Cognition and the Development of Language**, John Wiley and Sons.

BEVERAGE, R. (1981) **A Structured Methodology to Ascertain Requirements for Computer Based Decision Support Systems**, PhD Thesis, George Washington University, USA.

BOLINGER, D., and SEARS, D. A. (1981) **Aspects of Language**, 3rd edition, Harcourt Brace Jovanovich Publishers.

BURG, J.F.M., and van de RIET, R.P. (1996) "Analyzing Informal Requirements Specifications: A First Step Towards Conceptual Modelling", in van de RIET, R.P. et al. (Eds.), **Application of Natural Language to Information Systems**, IOS Press.

CALWAY, B. A. (1998) **A Discourse Analysis and Strategies Based Model for Discovering Facts from Natural Language Information Systems Specifications**, PhD Thesis Swinburne University of Technology. (Submitted)

CALWAY, B. A., and SYKES, J. A. (1996) "An Application of Discourse Analysis in Conceptual Modelling", **Australian Journal of Information Systems**, Vol. 3 No. 2, pp. 10- 19.

CHEN, P. P. (1983) "English Sentence Structure and Entity Relationship Diagrams", **Information Sciences**, No. 29, pp. 127-149.

CHOMSKY, N. (1956) "Three Models for the Description of Language", **IRE Transactions**, No. 2, pp. 113-124.

CHOMSKY, N. (1965) **Aspects of the Theory of Syntax**, MIT Press.

CLARK, H. H., and CLARK, E. V. (1977) **Psychology and Language - An Introduction to Psycholinguistics**, Harcourt Brace Jovanovich Publishers.

DARKE, P., and SHANKS, G. (1994) **Defining Systems Requirements - A Critical Assessment of the NIAM Conceptual Schema Design Procedure**, working paper 8/94, Department of Information Systems, Monash University, Australia.

van DIJK, T. A., and KINTSCH, W. (1983) **Strategies of Discourse Comprehension**, Academic Press.

DUNN, L. J. (1992) **Relational Database Engineering - a Natural Language Driven Approach**, PhD Thesis, University of Queensland, Australia.

EDMOND, D. (1992) **Information Modelling - Specification and Implementation**, Prentice Hall.

EGGINS, S. (1994) **An Introduction to Systemic Functional Linguistics**, Pinter Publishing. FALKENBERG, E. D. (1976) "Concepts for Information Modelling", in NIJSSEN, G. M. (ed.), **Modelling in Data Base Management Systems, North-Holland Publishing Company**.

FALKENBERG, E. D. (1986) **Data Base and Information Systems**, Lecture Notes. FILLMORE, C. J. (1968) "The Case for Case", in BACH, E., and HARMS, R. T. (eds.), **Universals in Linguistic Theory**, Holt Rinehart and Winston Inc., pp. 1-88.

FLAVIN, M. (1981) **Fundamental Concepts of Information Modelling**, Yourdon Press.

FODOR, J. A., BEVER, T. G. and GARPETT, M.F. (1974) **The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar**, McGraw Hill.

van GRIETHAYSEN, J. J. (ed.) (1982) **Concepts and Terminology for the Conceptual Schema and the Information Base**, ISO/TC97/SC5-N695, ANSI.

HALLIDAY, M. A. K. (1978) **Language as a Social Semiotic - The Social Interpretation of Language and Meaning**, Edward Arnold.

HALLIDAY, M. A. K. (1994) **Introduction to Functional Grammar**, 2nd edition, Edward Arnold.

HALLIDAY, M. A. K., and HASAN, R. (1976) **Cohesion in English**, Longman Publishing.

HALLIDAY, M. A. K., and HASAN, R. (1993)**Language, Context and Text - Aspects of Language in a Social-Semiotic Perspective**, Deakin University, Australia, (Reprint from 1985).

HALPIN, T. A. (1995)**Conceptual Schema and Relational Database Design**, 2nd Edition, Prentice Hall.

HIRSCHHEIM, R., KLEIN H. K., and LYYTINEN, K. (1995)**Information Systems Development and Data Modelling**, Cambridge University Press.

HOFFMAN, R.R. (1987) "The Problem of Extracting the Knowledge of Experts From the Perspective of Experimental Psychology", **AI Magazine**, Vol. 8 No. 2, pp. 53-67.

HOFFMAN, R.R., SHADBOLT, N.R., BURTON, A.M., and KLIEN, G.A. (1995) "Eliciting Knowledge from Experts: A Methodological Analysis", **Organization Behaviour and Human Decision Processes**, Vol. 62 No. 2, pp. 129-158.

HOWARD, D. V. (1983) **Cognitive Psychology - Memory, Language and Thought**, MacMillan.

ISO/TR 9007 (1987) **Information Processing Systems - Concepts and Terminology for the Conceptual Schema and the Information Base**, International Organization for Standardization. Technical Report UDC 681-3-02, Published 1987-07-01.

KAPLAN, R. W. (1972) "Augmented Transition Networks as Psychological Models of Sentence Comprehension", **Artificial Intelligence**, Vol. 3, pp. 77-100.

KERL, S. A (1985) **Common-School Grammar of the English Language (1878)**, Scholars' Facsimiles and Reprints, Delmar, NY.

KIMBALL, J. (1973) "Seven Principles of Surface Structure Parsing in Natural Language", **Cognition**, Vol. 2, pp. 15-47.

MARTIN, J. (1992) **English Text - System and Structure**, John Benjamin Publishers.

MEZIANE, F. (1994) **From English to Formal Specifications**, PhD Thesis, University of Salford, UK.

MOULIN, B., and CREASY, P. (1992) "Extending the Conceptual Graph Approach for Data Conceptual Modelling", **Data and Knowledge Engineering**, Vol. 8, pp. 223-248.

NESFIELD, J. C. (1936) **Outline of English Grammar**, MacMillan and Co.

NIJSSEN, G. M., and HALPIN, T. A. (1989) **Conceptual Schema and Relational Database Design - A Fact Oriented Approach**, Prentice Hall.

NOBLE, H. M. (1988) **Natural Language Processing**, Blackwell Scientific Publications.

QUIRK, R., GREENBAUM, S., LEECH, G., and SVARTVIK, J. A (1985) **Comprehensive Grammar of the English Language**, Longman.

REED, A., and K-ELLOGG, B. (1886) **Higher Lessons in English 1886**, Scholars' Facsimiles and Reprints, Delmar, NY.

van de RIET, R. P., and Meersman, R. A. (eds.) (1992) **Linguistic Instruments in Knowledge Engineering**, Elsevier Science Publishers.

RILOFF, E., and LEHNERT, W. (1994) "Information Extraction as a Base for High-precision Text Classification", **ACM Transactions on Information Systems**, Vol. 12 No. 3, pp. 296-333.

ROTHE, H. J., TIMPE, K. P., and WARNING, J. (1989) "Psychological Methods of Knowledge Elicitation within the Domain of Mechanical Engineering", in KLIX, F., STREITZ, N. A., and WAERN, Y. (eds.) **Man - Computer Interaction Research** MACINTER II, North Holland Publishing Company.

STAMPER, R. K. (1992) "Language and Computing in Organised Behaviour", in van de RIET, R. P., and MEERSMAN, R. A. (eds.), **Linguistic Instruments in Knowledge Engineering**, Elsevier Science Publishers.

SU, M. (1988) **The Design of a Natural Language Modelling Environment for Manufacturing System Simulation**, PhD Thesis, The University of Iowa, USA.

SYKES, J. A. (1994) **English Grammar as a Sentence Model for Conceptual Modelling using NIAM** (working paper 10/94), Centre for Information Systems Research, Swinburne University of Technology, Australia.

VADERA, S., and MEZIANE, F. (1994) "From English to Formal Specification", **The Computer Journal**, Vol. 37 No. 9, pp. 753-763.

VENDLER, Z. (1967) **Linguistics in Philosophy**, Ithaca - Cornell University Press.

WANNER, E., and MARATSOS, M. (1977) "An ATN Approach to Comprehension" in BRESNAN, J., and HALE, M. (eds.), **Linguistic Theory and Psychological Reality**, MIT Press.

WINTER, E. 0. (1982) **Towards a Contextual Grammar of English**, Allen and Unwin. WINTRAECKEN, J. J. V. R. (1990) **The NIAM Information Analysis Method - Theory and Practice**, Kluwer Academic Publishers.

WOOD, D.D. (1993) "Process-Tracing methods for the study of cognition outside the experimental psychology laboratory", in KLIEN, G., ORASANU, J., COLDERWOOD, R. and ZSAMBOK, E. (eds.), **Decision Making in Action: Models and Methods**. Norwood, pp. 228-251.

WOOD, W. A. (1970) "Transition Network Grammars for Natural Language Analysis" **Communications of the ACM**, Vol. 13, pp. 591-606.