GRAMMATICAL CONVERSION OF DESCRIPTIVE NARRATIVE - AN APPLICATION OF DISCOURSE ANALYSIS IN CONCEPTUAL MODELLING

Bruce A. Calway & James A. Sykes
Centre for Information Systems Research
School of Information Systems
Swinburne University of Technology
BCalway@gpo.swin.edu.au
P.O. Box 218, Hawthorn Vic. Australia, 3122
Phn +61-3-214-8451 Fax +61-3-819-1240

ABSTRACT

Fact-oriented conceptual modelling begins with the search for facts about a universe of discourse (UoD). These facts may be obtained from many sources, including information systems reports, tables, manuals and descriptive narrative both verbal and written. This paper presents some initial findings that support the use of discourse analysis techniques as an approach to developing elementary fact based sentences for information systems conceptual schema development from written text.

Although this discussion paper only considers the NIAM (fact-oriented) conceptual schema modelling method, the ISO87 report from which the research case study is taken describes other conceptual methods for which the research contained in this paper could be applicable (e.g. Entity Relationship analysis). The case study could be modelled exactly in the form in which the text is initially found, but grammatical analysis focuses consideration on alternative, potentially better, expressions of a sentence, a theme which is described and demonstrated. As a result of having applied grammatical sentence simplification with co-ordinate clause splitting, each sentence could be expressed as a complete, finite, independent collection of declarative simple statements.

The outcome from the application of the techniques described provides at a minimum a discourse analysis of descriptive narrative which will have retained its meaning and contextual integrity while at the same time providing a simplified and independent clause representation for input to the fact-oriented conceptual schema modelling procedure.

PROBLEM DESCRIPTION AND RESEARCH OUTLINE

Context

Fact-oriented conceptual modelling of information systems begins with the search for facts about the universe of discourse (UoD), expressed as elementary declarative sentences in natural language. These facts may be obtained from many sources, including reports, tables, manuals and descriptive narrative both verbal and written. This paper presents some initial findings that support the use of discourse analysis techniques as a way of extracting elementary facts (i.e. simple relationships which exist between people, places and things) from written text.

Nijssen and Halpin state (Nijssen 1989:31) the fundamental approach to building a conceptual schema as taking "specific examples" then populating the resultant schema with instances for validation purposes. The specific examples may appear in many forms, or combination of forms including graphical, tabular, textual, etc. It is often a role of the systems analyst to take such disparate material and UoD discourse as the input to the conceptual schema development process. One step which the analyst may undertake is to collect discourse to describe in detail the UoD from the perspective of the various information systems agents (e.g. experts, users, developers). Discourse both written and verbalised (spoken) can provide a wealth of knowledge about the organisation's UoD and the application of information systems within that UoD

Falkenberg (Nijssen 1989:31) is credited with proposing the idea to base conceptual schema development on elementary natural language sentences, a proposal which is said to have been influenced by the grammatical work of C.J. Fillmore. Despite such work there persists a general perception that natural language lacks structure or is too complex for analysis purposes. This is not necessarily supported if only on the basis that if this were true no individual could understand communicative discourse. Sykes (Sykes 1994), Andersen (Andersen 1991), Stamper (Stamper 1992b), Halliday (Halliday 1994), have shown that natural language, although very complex in structure, contains considerable detail available for analysis.

The International Organization for Standardization released a Technical Report 9007 (ISO87) which outlined the concepts and terminology for developing conceptual schema models for Information Systems specification. Within that report was a detailed theoretical foundation for describing a universe of discourse (UoD) which was defined as "the collection of all objects that ever have been, are, or ever will be in a selected portion of a real world or postulated world of interest that is being described" (ISO87:12). An integral part of that description
was the formal system of logic to describe the UoD (ISO87:47-48) expressed at two levels ie. syntactic and semantic.

An analyst conducting a system analysis of an organisation is likely to seek out at least three forms of discourse, namely:

- prescriptive discourse; which expresses an organisation's work practices formally as manuals, process specifications, etc.,
- descriptive discourse; where an individual or group provides a description, in prose form, of part or all of an organisation's UoD,
- interview discourse; written records (transcripts, notes, etc.) of an interactive dialogue of questions and statements about an organisations UoD, between some individual/s or group/s from within the UoD and an individual or group investigating the UoD in order to describe aspects of that organisation's UoD.

All would seem possible sources relevant to the application of discourse analysis techniques. A collective descriptive discourse can be developed for the UoD by restatement as prose from prescriptive, interview and descriptive sources, for communicating structures and meanings through a written textual medium. Given that natural language discourse, as written text, could be valuable in the analysis and design process when developing information systems using information technology, the question then arises as to an appropriate approach for investigating and restating, as simple facts for conceptual model development, the content and contextual expression of the discourse.

Problem Description and Research Method

The success of the analysis and development of conceptual schema specifications for a UoD depends upon completeness and accuracy of the UoD fact discovery. As stated earlier fact-oriented conceptual schema modelling begins with the search for facts about the universe of discourse (UoD). When the analyst must search written text, fact-oriented methods like NIAM offer no guidelines for discovering these facts (other than that the analyst knows what they should look like, and how to verify them when discovered by getting agreement from the UoD expert (Nijssen 1989)). Prior work by Sykes (Sykes 1994) has shown that the study of the grammar of natural language sentences has value for the early stages of schema development. It was also shown that a syntactic grammatical model of English sentences is applicable. The present work can be seen as seeking to extend that line of study, which only considered a single sentence/clause, to dealing with groups of complex sentences in a descriptive narrative expressed as written text.

The larger research project, of which this paper forms a part, proposes:

that techniques from the field of discourse analysis, being the relevant aspect of linguistics, might offer a repeatable methodical way for a systems analyst to deal with UoD descriptions when found as written textual discourse.

For this paper it is proposed to survey the discipline of discourse analysis initially to find individual techniques that appear suitable, and not to be concerned at this point that a method necessarily be the best available. Each method selected will individually be applied to a selected case study (below) and evaluated for results obtained, specifically for the ability to:

- allow the separation of clauses within compound sentences,
- identify and provide for inter-clausal referential identification without loss of meaning,
- deliver independent clauses, while retaining the narrative sequence of the clauses in the text.

Discourse Analysis

Discourse analysis is concerned with the understanding of bodies of written or spoken text. Some of its techniques are aimed at simplifying complex sentences, the subject of this current study. The use of such techniques can yield simple declarative clauses (i.e. simplest form of sentence) without loss of meaning. Therefore discourse analysis can be a useful addition to fact-oriented conceptual schema development processes. Subsequently, if the methods can be easily understood and applied, they can be taught to systems modelling analysts.
A case study based approach is taken for this work with the discourse analysis being limited in scope to a single discourse case, i.e. the ISO87 technical report example UoD. The selected case study is well documented in the technical report and provides an established framework of definition and analysis for conceptual schema development using various conceptual schema methods.

Case Study

"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. The manufacturer is registered as the owner of the car as soon as practicable. At this time it is given one registration number, unique for all cars and for all time. The year of production is also recorded. During the month of January only, a car may be declared to have been produced in the previous year. Eventually a car is destroyed and the date of destruction is registered. The history of a car must be kept until the end of the second calendar year following its destruction." (ISO87:77)

DISCOURSE ANALYSIS FRAMEWORK

Discourse Analysis and Simple Terms

Discourse analysis is generally considered as "the analysis of language in use" and as such is dependent upon the description of language usage within the UoD (e.g. work language, etc.). The methods for studying the diversity of discourse are extensive and therefore for this discussion have necessarily been constrained to specific domains of language analysis. The subsequent discussion is limited to written text as declarative narrative with the discourse analysis techniques arbitrarily drawn from the syntactic and semantic disciplines as described by Quirk et al. (Quirk 1985), Halliday (Halliday 1994), Nesfield (Nesfield 1936), Eggins (Eggins 1994), Winter (Winter 1982).

Choosing a path which allows the discourse analysis techniques to be selected is first clouded by the issues surrounding the definition of "sentence". The problem arises when groups of sentences are to be considered, because of the need to have a workable grammar of English besides the linguistic grammar of phonics, morphemes, words, phrases, etc. As Halliday (Halliday 1994) points out it is often assumed that discourse analysis can be undertaken without grammar and states that discourse analysis that is not based on grammar is not an analysis at all. A contextual grammar for describing sentences is necessary in order for real-world (day to day) communication which will use clauses and sentences in unfamiliar ways as part of written or spoken discourse.

Grammarians (Nesfield 1936, Winter 1982, Quirk 1985, et al.) traditionally have considered sentences in terms of form or class labels ranked as word (i.e. parts of speech as noun, article, etc.), phrase/group (i.e prepositional phrase, adverbial phrase, etc.) and clause (finite, non-finite, dependent clause, etc.). The independent clause (single subject with single predicate) is often referred to as the minimum sentence (or simple sentence). This approach is adopted by Sykes (Sykes 1994) when considering the NIAM conceptual modelling method, displaying that the use of English grammar expressed by Quirk et al. was of value. Quirk et al. (Quirk 1985), as an alternative grammatical classification, subdivide a clause into a series of functional parts (i.e subject, verb, object, complement, adverbial (SVOCA)).

Nijssen and Halpin (Nijssen 1989) state, as an objective, the development of facts about a UoD. These facts are expressed in a very structured format which reflects two known grammatical representations of a sentence (i.e. subject/predicate, and subject/verb/object, ref. below). Nijssen and Halpin describe these facts as elementary sentences which therefore could reasonably be expected to respond to discourse analysis in terms of these grammatical representations.

The two English sentence representations under consideration are:

- Subject - Predicate; where it is implied that a single subject is expressed or implied and that the predicate expresses something about the subject referred to in the sentence. One can regard the subject/predicate form as focussing on the logic of the sentence.

- Subject, verb, object, complement, adverbial (SVOCA); where the sentence is seen to operate as functioning parts, the simplest being subject and verb e.g. "Dogs bark". The SVOCA analysis emphasises the grammatical structure.
During this study what is being sought from the application of discourse analysis are elementary facts expressed as independent descriptive or declarative clauses. A brief study of grammar basics will assist in clarification of the use of these terms and also point to the reasons for their use within this paper. This is due to the explicit limitations in meaning for each term used. The following definitions represent a synthesis of those offered by Kerl (Kerl 1985), Kellogg (Reed 1886), Nesfield (Nesfield 1936), Quirk (Quirk 1985) et al.

- **Discourse** - any series of thoughts embodied in language which may be:
  - **Descriptive** - an account of persons, places and things,
  - **Narrative** - chiefly temporal as a series of events and states,
  - **Illustrative** - generally rhetorical and used to make the meaning more intelligible or impressive.

(NB. any or all forms may exist in a single piece of discourse)

Discourse can be taken as a sentence or a series of sentences (e.g. paragraphs) (series here is used advisedly as the removal or restructuring of sentences within a series may alter the contextual meaning, a point discussed later in this paper when considering sentence analysis and conversion). The division of discourse can also be by volumes, books, parts, cantos, verse, chapters, sections, paragraphs, sentences, and is generally left to the writer's own tastes. Of interest to this study are:

- **Sentence** - being a thought expressed by at least one subject/predicate clause, or a union of clauses, followed by a full pause (e.g. .!? etc.).
- **Clause** - a sentence as a whole, or that makes part of a larger sentence.

A sentence or clause expresses at least three forms:

- **Declarative** - an expression about something,
- **Interrogative** - expressing a question,
- **Imperative** - expressing a command, permission, etc.

Kerl (Kerl 1985:229) suggests a further option ie. Exclamatory which could be seen as the co-operation of Interrogative and Imperative. Also many other compounded forms may exist e.g. declarative and interrogative, etc.

Senteuces may also be divided into classes:

- **Simple** - having only one clause expressed or implied,
- **Complex** - being a sentence containing one independent or principal clause with one or more dependent clauses,
- **Compound** - being a sentence that has two or more independent clauses. It should be noted that the compound sentence may consist of two or more complex sentences or a mixture of all classes.

**Fact-oriented Sentence**

Therefore definitionally what would seem to be required for fact-oriented conceptual schema development using discourse as simple sentences is:

*single clause sentences to express a proposition which is declarative in form and descriptive and/or narrative by nature.*

The recommended practice for interpretation of source material into fact-oriented elementary sentences (simple sentences) (Nijssen 1989:32-48) for example, is that elementary facts should emerge from the analysis of natural language utterances about a UoD that can be or have been validated by a UoD domain expert. The word "fact" indicates that the system is to treat the assertion contained in the sentence as being true of the UoD whether or not this really is so. A "simple sentence" in English as with fact-orientation is one which has only one finite verb expressed or understood, although there may exist many verbs within the sentence. The sentence following is an example of a simple sentence in English with a single subject/predicate structure. It is the intent of NIAM conceptual modelling (Nijssen 1989) to have a subject/predicate structure, but only for a single clause (ie.
no subordinate or coordinate clauses either existing or implied in the sentence) of the SVOCA form as Sykes points out (Sykes 1994).

- Simple sentence with single subject/predicate

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>The manufacturer, having a large volume of cars to sell</td>
<td>caused all of its stock to be conveyed by train, there being no market in that area.</td>
</tr>
</tbody>
</table>

In this sentence there are five verbs; having, to sell, caused, to be conveyed, being. Of these "caused" is to be taken as the finite verb, as it is the only verb out of the five which has or could have a subject attached to it, but there exist several phrases which could be converted to (dependent) clauses or even made independent. One objective in creating sentences for input to the fact-oriented conceptual modelling method would be to create sentences which as a worst case contain co-ordinate clauses with no subordinate clauses existent. A co-ordinate clause is one that can be separated from other clauses to make an independent sentence which gives an independent sense. It is possible to convert simple, compound and complex sentences into purely co-ordinate clause sentences (Nesfield 1936, Kerl 1985, et al.) as shown later in this paper.

Sentence Analysis and Conversion

The discourse analysis approach which will be applied to the case study and provide the balance of the discussion in this paper is:

- Grammatical Conversion - where sentences are grammatically modified to remove subordinate clauses and provide either simple (one clause) or co-ordinate clauses (self contained but contextually connected).

Winter (Winter 1982) has laid a foundation which has been followed within this discussion where he investigates and redefines "sentence" in terms of the clause. That definitional work is founded upon work by such significant grammarians as Fries, Halliday, Kerl, Kellogg, Jespersen, Quirk, et al. in attempts to bridge the dichotomy between clause and sentence structure within discourse. The definition of sentence offered by Winter (Winter 1982) and Halliday (Halliday 1994) argues that the term "clause complex", which refers to the clause relationship as a sentence, is taken as the starting point for discussion. Therefore, the clause becomes the significant first level decomposition of a sentence and the SVOCA structure (ie. subject, verb, etc.) offered by Quirk et al. becomes the relevant grammatical structure of a clause. The current study is limited to descriptive discourse because the discourse considerations for analysis of information systems are normally descriptive representation of the UoD and declarative in nature. The structure of declarative sentences can be very complex with subordinate and co-ordinate components producing implied meaning through grammatical structure as is the scenario with the first sentence of the case study where the implication is that there is a relationship between a car, the model and the manufacturer.

Winter, Eggins, Kerl and Nesfield (Winter 1982, Eggins 1994 Kerl 1985, Nesfield 1936) each offer similar principles for grammatical conversion of complex declarative sentences into simple sentences; approaches which may prove beneficial to sentence structuring for elementary sentence development and conversion. In the first sentence from the case study "A car...... manufacturer" there are four verbs or verb groups. Of these only one has or could have a subject attached to it. Therefore since there is only one finite verb the sentence could be considered simple but in reality is not because of the clause structure. A simple sentence (clause) could be "A car is of a particular model". In order to convert from simple sentences to multiple co-ordinate clauses Nesfield and Kerl among other authors suggest (Nesfield 1936:143,152/3, Kerl 1985) the following substitutions which can occur:

- Simple sentences can be converted into multiple co-ordinate clauses, by expanding words or phrases into clauses joined with co-ordination conjunctions or connectives (e.g. and, before, which).

- Cumulative connective (adding)

Simple  *Besides making* a promise, he kept it.
Multiple He not only made a promise, but also kept it.

- Alternative connective (offering a choice)
  Simple He must confess his fault to escape being fined.
  Multiple He must confess his fault, or he will be fined.

- Adversative connective (expressing a contrast or a difference)
  Simple Notwithstanding his sorrow, he is hopeful.
  Multiple He is sorrowful, but yet hopeful.

- Illative connective (drawing an inference)
  Simple Owing to bad health, he could not work.
  Multiple He was in bad health, and therefore he could not work.

Simple sentences as seen can be converted to that of double or multiple co-ordinate clauses (should more than one clause be observed in the sentence), by expanding words or phrases to form the co-ordinate clauses. Similarly, to convert complex sentences to that of double or multiple clauses, the main clause must be placed last, and the subordinate clause/s must be placed first (Nesfield 1936 pp 152-153, 162).

- Complex to Multiple (main clause last).
  Complex I am certain that he will not recover.
  Multiple He will not recover, and of this I am certain.

A further grammatical simplification (Eggins 1994:134) which can be considered is to divide a clause into its constituent parts and decide when a particular component (phrase or group, as part of a clause) is operating at clause level. Three tests are available:

- movability - where the component is independently moveable within a sentence,
- substitution - where a clause or phrase can be replaced with a single substitute verb or pronoun. This is achieved by asking what the substitution item is standing for.
- probe questions - which allow testing for verb groups, prepositional phrases, adverbial elements, etc.

There are three kinds of subordinate relationships recognised (Nesfield 1936:155) and therefore available for expansion as co-ordinate clauses:

- noun clauses - acting as a noun in relation to some word in some other clause,
- adjective clauses - acting as an adjective in relation to some word in some other clause,
- adverb clauses - acting as an adverb in relation to some word in some other clause.

Even after dealing with issues of grammatical conversion, problems of referential packing still exist (ie. packing being the avoidance of needless repetition of the words within a clause or sentence), bearing in mind the targeted output from any analysis is a set of sentences of no more than one clause expressed or implied, therefore:

- all subordinate clauses within a sentence must be removed.
- each clause within a sentence must be made independent of all other clauses.

When analysing sentences in consideration of this latter point, for the purpose of creating independent clause sentences, the following referential unpacking substitutions suggested by Nesfield et al. must be completed (Nesfield 1936, Kerl 1985):

- When there are two finite verbs to the same nominative, the nominative is not usually mentioned more than once, but it must be repeated in the analysis.
  e.g. The sun rose and (the sun) filled the sky with light.
• When there are two nominatives to the same finite verb, the finite verb is not usually mentioned more than once, but it must usually be repeated in the analysis.
e.g. Either this man sinned or his parents (sinned).
However with some sentences the verb must not be repeated, because nominatives are seen as a collective.
e.g. Truth and honesty is the best policy.

• The omission of the conjunction "and" can often be seen when the author is stating a single fact with a string of related objects.
e.g. A steam engine can propel, elevate, pump, (and more thing than can be enumerated.)
A steam engine can propel and a steam engine can elevate, etc.

Case Study Examined using Grammatical Conversion

(a) Here the original case study is reproduced with sentence sequencing, ie. numbering for ease of identification in subsequent analysis, as grammatical conversion can alter the structure of the sentence:

S1 - 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.
S2 - The manufacturer is registered as the owner of the car as soon as practicable.
S3 - At this time it is given one registration number, unique for all cars and for all time.
S4 - The year of production is also recorded.
S5 - During the month of January only, a car may be declared to have been produced in the previous year.
S6 - Eventually a car is destroyed and the date of destruction is registered.
S7 - The history of a car must be kept until the end of the second calendar year following its destruction.

(b) One possible solution using grammatical conversion of the first sentence S1 of the above (a) is therefore:

S1 - A car is of a particular model
and (a car) is given a serial number by its manufacturer
that (serial number) is unique among the cars made by that manufacturer.
• This sentence has two co-ordinate clauses and one relative (dependent) clause, therefore this is a complex sentence.
• There are a number of finite verbs to choose from, but if one were to consider one which was pivotal it is "is given".

(c) One possible solution using grammatical conversion of sentences S2-S7 of the above (a) is therefore:

S2 - The manufacturer is registered as the owner of the car as soon as practicable.
S3 - One registration number is given (to a car) at this time
and (the registration number) is unique for all cars
and (the registration number) is for all time.
S4 - The year of production (of a car) is also recorded.
S5 - If the month of production (of a car) is January only,
a car may be declared to have been produced in the previous year.
S6 - Eventually a car is destroyed
and the date of destruction (of a car) is registered.
S7 - The history of a car must be kept.
[Until the end of the second calendar year following (a cars) destruction.]
By applying grammatical conversion, a complex and/or compound sentence structure can be changed to create a double or multiple co-ordinate clause structure, but co-ordinate clause sentences still do not comply with the structure of elementary facts where an elementary fact has a subject and single predicate as a single independent clause. Second, clauses and phrases contained within a discourse have pronouns and inferential or thematic referential connections and while this allows ease of reading it hinders elementary fact development when sentences are grammatically converted and then considered independently, i.e. outside of the discourse context, as would be the case for sentence S4. Subordinate clauses and sentences are thematically and inferentially connected as is the case for sentences S1 - S2, S2 - S3 - S4, S2 - S5, etc. Further, clausal items may refer to each other but with differing lexical items as in S1 "car is made" and S4, S5 where the reference is to "production". Therefore, although clauses may be separated and made independent this does not provide the referential identification which is required between clauses within a sentence or descriptive narrative.

- The first finding is that sentences which are analysed and converted from complex or compound into co-ordinate clauses require a degree of observer (analytical) intuition to derive independent elementary facts based upon the observable co-ordinate clauses structure contained in the sentence.

- The second finding is that no clause or sentence can be made independent of the discourse in which it is found by the application of grammatical sentence conversion alone. This finding identifies the need for referential identification to occur prior to, or at a minimum in parallel with, grammatical clause analysis.

The resulting list (above) of clauses is a declarative narrative expression of the independent clause/phrase structure to be found within the case study text. Each clause contains an independent meaning (albeit intuitively derived for referential identification) in that a clause can now be taken out of sequence without loss of referential meaning and modelled as an elementary fact. If a sentence is individually considered it can be observed that contextual meaning is still required to be interpreted as referential constraints, e.g. (Eventually a car is destroyed) where, as a sentence it syntactically correct and semantically plausible, but out of context is functionally meaningless.

SUMMARY AND CONCLUSIONS

The research project, of which this discussion paper forms a part, proposed:

*that techniques from the field of discourse analysis, being the relevant aspect of linguistics, might offer a repeatable methodical way for a systems analyst to deal with UoD description when found as written textual discourse.*

It has been shown that the grammatical approaches taken do:

- allow the separation of clauses within compound sentences to be delivered as a series of independent co-ordinate clauses for a single discourse as written text,
- provide a methodical approach of discourse analysis for a system analyst to deal with UoD descriptions,
- highlight the information packing and shortcomings inherent within the complex sentences.

Grammatical reconstruction of discourse into elementary sentences could prove difficult on its own for analysts who are not familiar with the rules of English Grammar, but it is believed that the methods when documented fully could be taught to systems analysts faced with using descriptive narrative for derivation of conceptual schemas or schema verification.

In summary, it can be concluded that methods of analysis are available within the linguistics discipline which when applied to descriptive narrative allow the derivation of simple sentences of the SVOCA form. Although sentences in the case study narrative could be modelled from the descriptive narrative form in which they were initially found, grammatical analysis methods focus consideration on alternative, potentially better, expressions of a sentence, a theme which has been described and to some extent demonstrated through the analysis processes described within this paper.
REFERENCES


Nord C.;(1991) *Text Analysis in Translation* Rodopi Amsterdam (90-5183-311-3)


Palmer M.S.;(1990) *Semantic processing for finite domains* Cambridge University Press (0-521-36226-1)


Reed A, Kellogg B; (1886) *Higher Lessons in English 1886*. Scholars' Facsimiles & Reprints, Delmar NY.


van Dijk T.A.; (1986) Text and Context - Explorations in the semantics and pragmatics of discourse. Longman (0-582-29105-4)