

Number 290



UNIVERSITY OF  
CAMBRIDGE

Computer Laboratory

## Discourse modelling for automatic summarising

Karen Spärck Jones

February 1993

15 JJ Thomson Avenue  
Cambridge CB3 0FD  
United Kingdom  
phone +44 1223 763500  
<https://www.cl.cam.ac.uk/>

© 1993 Karen Spärck Jones

Technical reports published by the University of Cambridge  
Computer Laboratory are freely available via the Internet:

*<https://www.cl.cam.ac.uk/techreports/>*

ISSN 1476-2986

# Discourse modelling for automatic summarising

Karen Sparck Jones

Computer Laboratory, University of Cambridge  
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK  
sparckjones@uk.ac.cam.cl

February 23, 1993

## 1 Abstract

Automatic abstracting is a challenging task for natural language processing. It depends not only on sentence interpretation and the local context representation this requires, but also on the recognition and use of large-scale discourse structure. This paper describes research investigating the nature of different approaches to discourse representation and their value for summarising. This work is focused on comparative analysis, illustrated in the paper through the provision of different forms of representation, and different strategies for summary formation, for a short example text.

## 2 Introduction

From the start, research in natural language processing (NLP) had to address sentence processing. In the last decade in particular it has begun to address discourse processing. There have been two reasons for this. One is that sentence processing has required it, most obviously for anaphor resolution. The other is that the system's task has required it, for instance in order to provide contextually appropriate responses in interactive dialogue.

The requirements for processing going beyond the sentence have naturally stimulated work on discourse modelling. But though there are many views on what discourse models should be like (cf CL, 1988; Sparck Jones, 1991), computational modelling has in practice been primarily motivated by dialogue rather than 'single-source' discourse needs (e.g. Grosz and Sidner, 1986). Current work on text interpretation, for example within the framework of the Message Understanding Conferences (e.g. MUC-3, 1991) is indeed concerned with single-source discourse processing, but usually in very restricted ways. In particular there is little attempt to model the structure of a text as a whole other than as very weakly embodied in statistical

facts (cf Salton, 1989) or as very strongly embodied in instantiated a priori domain structures like frames (e.g. Young and Hayes, 1985). Research on text generation, on the other hand, has sought to apply relatively strong, but domain independent, rhetorical models of discourse (McKeown, 1985; Maybury, 1991). But the generated texts, even where their structure is not constrained by an interactive dialogue context, have not usually been very long. At the same time, general theories of discourse structure intended to apply to extended text (e.g. Kintsch and van Dijk, 1978; van Dijk and Kintsch, 1983; Mann and Thompson, 1987) have either not been applied computationally or have proved very difficult to apply, especially in interpretation rather than generation (cf Moore and Pollack, in press).

There is, however, one very important, generic, language-using task that appears to require thoroughgoing discourse modelling. This is summarising. Summarising is indeed viewed as a natural concomitant of discourse processing by Kintsch and van Dijk but, as just noted, in work which has not been computationally implemented. Research on automatic summarising, on the other hand, has either adopted shallow approaches most obviously illustrated by statistically or locationally motivated extraction (Salton, 1989; Paice, 1990), or has been based on very strong specifications of the kind of information to be derived from source texts, as in DeJong's FRUMP (DeJong, 1979).

It would clearly be very useful to be able to obtain summaries automatically, and it also seems self-evident that doing this properly requires deeper and more flexible strategies than those hitherto adopted. The rest of this paper presents the framework we have adopted for investigating summarising, and the initial work my colleagues John Beaven, Phil Gladwin and Steve Pulman, and I have done within this framework on a current Cambridge project. This research has naturally led us to consider the notion of discourse structure. Summarising directly addresses two key questions about discourse: what large-scale or global structure does discourse have, and how is this related to small-scale, local structure tying sentences into their immediate environment? Thus from the point of view of discourse, summarising can be seen as a lever for formulating and testing theories of discourse structure, and in particular of providing a context for relating the finer-grained functional perspective considered by Hajicova and her colleagues (Hajicova and Panevova, 1984; Sgall, Hajicova and Panevova, 1986) to more comprehensive text function; and the same holds for the treatment of cohesive devices identified by (e.g. Halliday and Hasan, 1976; Hobbs, 1990). Automatic summarising, like automatic question answering (e.g. Lambert and Carberry, 1992; Moore, 1989), naturally demands a rigorous account of whatever levels of discourse structure there are, the purposes they serve, and the ways they are recognised and exploited.

### 3 Framework

We are interested in general techniques for automatic summarising. Devising these is clearly a long-term enterprise, and in our project we are investigating foundations, and exploring options, through laboratory experiments in which we are deliberately,

in the interests of clarity and control, keeping everything simple. However as summarising is a complex task, this simplicity can only be informally achieved; and as many of the resources required for fully automatic experiments are not available and are not readily supplied, we have been forced to engage in simulation.

The reality with which we have to start is the variety of summaries that can be legitimately produced for the same text. Figure 1 shows an example text that will be used for illustration throughout this paper, and three summaries of it written by different people. Though these three summaries have something in common, they differ in detailed content and structure. The challenge of summarising is even more sharply illustrated by the contrast between the two summaries P3-1989 and P3-1988, written at a year's interval by the same person, and by the nature of the earlier of these, P3-1988.

### 3.1 Summary factors

It is not difficult to give a straightforward definition of a summary: a summary is a condensed derivative of a source, i.e. it is a content reduction achieved by selection and/or generalisation on what is important in the source. Then, taking the position that text content, or meaning, is a matter both of information and expression, condensing content involves both abridging information and shortening expression. The former has to be done without loss of coherence, the latter without loss of cohesion, in the summary text. From this point of view we can differentiate summaries from other kinds of reduced derivative from a source text in terms of the relative emphasis respectively placed on abridging and shortening, and on coherence and cohesion. Thus a summary can be distinguished from a precis which is designed to shorten expression, possibly with a loss of coherence, whereas an abstract concentrates on abridging information, with less regard for cohesion than coherence. An extract is further distinguished from all of these as a copy, with possible damage to both coherence and cohesion.

Inspection also shows that in practice summaries can be of very different types, or rather as differing along various dimensions. They can, for instance, be reflective or reorganising in relation to the way material is ordered in the summary as opposed to the source; direct or indirect, i.e. reproducing source content or merely pointing at it; neutral or evaluative in the summary stance.

Typological distinctions of this kind, which distinguish summaries from other forms of text condensation and distinguish different sorts of summary, are familiar in the abstracting literature: see for instance Rowley (1982). But they are typically not recognised in abstracting programs, which explicitly or implicitly assume a certain kind of relationship between source and summary, as in DeJong (1979). This is also the case where summarising is seen as an integral element in discourse processing, as in Kintsch and van Dijk (1978) or as intimately related to it, as in Rumelhart (1975, 1977).

This level of characterisation is, moreover, rather coarse. There are many more specific factors that affect, and may even determine, summarising. These fall into three classes: those associated with the input to summarising, i.e. the source text;

those associated with the purpose of summarising; and those associated with its output, the summary text. Like the previous notions, these factors require detailed analysis and definition, and for this paper they can only be briefly presented. Thus, for example, the input factors fall into two classes, that can be labelled those of input form and of input type respectively. The form factors include organisational structure, scale, and genre, so that if we have, for instance, a progress report as input text it may have a manifest organisational structure of introduction, tasks, performance, comments, plans etc, a scale of twenty pages, and fall under the genre heading of narrative. The subject type factors include source properties like whether this is ordinary or specialised so, for instance, dealing with a particular area of chemistry might be specialised, or is restricted, as the progress report could be in referring to local institutionally-defined tasks.

The second, and most important factors are those representing the purpose to be served by the summary. These are audience and function. To make a broad distinction, summary audiences may be untargetted, i.e. unknown or only very broadly characterised, or targetted, i.e. specifically known or at any rate narrowly specified. The summary's function, on the other hand, may be to impart information, or to alert to it; that is a summary may be taken as a substitute for the original text, or be designed to draw attention to it. The third class of factors affecting summarising, the output factors, refer to the material of the summary, i.e. broadly speaking whether the summary covers the entire source or only part of it, and to the format adopted, primarily either running text or some fielded or categorised style. Thus a summary might deal only with the results section of a paper reporting biological experiments, or use a standard headings sequence including "location" for archeological summaries.

These distinctions, crudely formulated though they are, emphasise the variety of summaries. Other language processing tasks may be similarly varied and, like summarising, produce the same kind of object for different purposes. For instance crude automatic translations may be used to decide what documents should be properly translated, manually, or as convenient indexes into source documents in other languages. However the range of summary possibilities is large, particularly since summaries may be used as end-products in themselves, for human readers, or as intermediates for other tasks, like routing.

Moreover, while it is evident that the nature of the input to summarising and of the summary purpose together impose constraints on the nature of the output, it is not clear they unequivocally determine it, certainly in the fine grain. The particular properties of the individual input text naturally also have their own significant consequences. However, even taking these into account, the essential point is that there cannot be a conditioning specification of a summary's purpose sufficient to define a correct summary, i.e. to tie the source text and summary text together in just one proper way. Moreover summaries are for human use, indeed for individual human uses on individual occasions. This has two consequences. One is that an analysis of source text properties and summary purposes cannot provide any very definite guidance for system design or operation. The other is that system evaluation can only be done on the basis of practical utility: it may be reasonable for experimen-

tal purposes to judge alternative summaries produced by different strategies on an intuitive basis, but summary system outputs can properly only be evaluated via extended practical testing with real users (cf Sparck Jones, 1990). The essential problem here is human unpredictability even if, in some circumstances, abstracts may be tested like indexes as indicators of relevant documents in retrieval, via recall and precision measures for instance.

Given the large space of summary possibilities, we have had to restrict our project investigations. We have therefore chosen to study what we have called basic summaries. A basic summary is one which imparts information rather than alerting to it; is comprehensive in covering all rather than only a predetermined part of a source text; is reflective in preserving the organisational structure of the source text; and is integrative in producing a single output text rather than, say, a decomposition as in a fielded form. However as mentioned earlier, these are broad characterising notions, and therefore ones which our project is intended to refine. Thus defining reflection in terms of surface organisation may be too restrictive, and it may have to be redefined as preserving whatever kind of structure is taken as primary for the source text. At the same time, in relation to audience and purpose, adopting a minimalist or neutral treatment suited to the idea of basic summary in fact assumes that the summary audience has the same properties as the source text audience, and that the communicative intention of the source is maintained, i.e. that the function of the summary does not call for more transformations of the source than those required by the reduction intrinsic to summarising. It clearly does not follow that what is learnt from summarising under these conditions will be applicable in others, but the exercise, approached with generality, should provide useful information.

### 3.2 Processing stages

It is also necessary to work within the framework of some view of the global summarising process. Unfortunately the kinds of instruction given in conventional abstracting manuals are not much assistance here. Rowley's recommended procedure (Rowley, 1982), for instance, is as follows:

1. read the document to grasp its content and scope
2. make notes of its main points
3. draft a rough abstract
4. check this for accuracy, omissions etc and edit
5. write the final abstract

For analytical purposes, in the context of seeking a general approach, it is also useful to treat the interpretation of a source text and the representation of its meaning as completely distinct from summary generation. Interpretation is viewed as delivering a complete source text representation to which summarising processes can be applied, first to construct the representation for the summary text, and then to

produce the actual text from this. The complete separation of source and summary processing is helpful because it makes it possible to consider the different types of information that source interpretation might provide and, independently, different ways in which these might be individually or collectively exploited for summaries. We do not want to restrict ourselves prematurely by assuming that interpretation is geared to summarising, as it is in DeJong (1979) and Young and Hayes (1985). Figure 2 thus shows the global process model we are assuming. In this we allow for two possible steps in source text interpretation, involving initial analysis, primarily at the sentence level, and subsequent integration for the complete source representation; and envisage two stages in summary generation. It is clearly the first, formation step in the latter, in which the source text representation undergoes a condensing transformation, that is the crucial one.

But it is clearly conditioned by the character of the source text representation. We have therefore to consider both the nature of this representation, as determined by source interpretation, and how adequate it is for summarising, from the point of view of the processing it needs and of further information resources that have to be brought to bear on it.

Thus what kinds of information might the source representation contain, and how might these be exploited?

### 3.3 Information types

In general, there are three types of information that source representations might give: linguistic information, world information, and communicative information.<sup>1</sup> For example, if we consider just the three sentences of the Biographies text shown in Figure 3, there is linguistic information in the sentence forms (declarative, 'generic') and lexical items used, and also in the parallels in forms and lexical items over the three. At the same time, the sentences convey information about the world, both about books of certain kinds (biographies etc) and about properties ascribed to them (valuable, etc). There is, moreover, communicative information in, or associated with, the sentences, namely that the writer wants the reader to adopt certain beliefs about books.

Even if these examples are only illustrative sketches, there is little doubt that there are different kinds of information embodied or conveyed by these three sentences. So while these are of course also intimately related, as signalled in a very obvious way in the Figure by the use of the word "valuable" in the linguistic characterisation, the descriptive property 'valuable' in the world model, and the assigned property 'valuable' for biographies in the communicative case, there is an important issue in whether these three types of information are all combined into a single source text representation.

---

<sup>1</sup>Note that I am using "information" here in a way which has no simple relationship to my informal earlier distinction between information and expression.



### 3.4 Large-scale structure

This issue becomes even clearer when large-scale text, and derivatively representation, structure is taken into account. There is no doubt that text has large scale structure, and therefore every reason to envisage a correlated text representation structure. It also seems likely that any such large-scale structure will play an important role in guiding summary formation. But again, it is possible to obtain quite different large-scale representation structures according to the different kinds of information on which they are based.

Thus for the Biographies example text as a whole, the three distinct structured representations shown in Figure 4 can be obtained. This is again a quite informal illustration, but is nonetheless one which is decently grounded in the actual source text. The important point about it is that even where, for this short text, the overall structure has a similar syntax, as for the linguistic and communication examples, the substance is quite different; and both are quite different in both syntax and semantic content from the representation organised in terms of facts about the world. Supposing then that these structures were used to constrain or guide summarising, they might each lead, in simple strategies suited to each type of representation, to the adoption of the material marked with S in the Figure for the summary. Thus for the linguistic case, looking for tightly repetitive or emphatic and proximate items would identify the substance of the last six sentences as summary content; for the world case picking out cause-effect-action relationships would identify biography giving as important; while operating on communicative representation might select 'strong' speech acts and their justifications. For the Biographies text the resulting summary representations would have much in common, but they would also differ significantly and would naturally lead to different output summary texts. Figure 4 shows some possible examples.

The other point that emerges, however, is that there are aspects of the original that are not necessarily captured in each one of the representations alone, for example in the communicative representation that a systematic comparison is being made, or that there is an emphasis on the contrast: biographies, NOT novels. It is further not clear whether any of the three representations captures the precise force of the real/not real, individual/not individual opposition as ground for biographies being best.

At the same time, these large-scale structures might be based on quite different constructive principles. These can be conveniently, if very crudely, categorised as bottom-up or top-down, though this is not intended to imply that during the interpretation of the source text there is a correspondingly simple processing line. What is meant here is how the structures are defined. Thus a large scale linguistic structure could be defined as bottom-up if it is derived by the progressive linking and consolidation of e.g. thematic units, using a few general rules; a world-model based representation could be analogously defined if found by combining the initial 'facts' given by individual sentences; and a communicative structure be obtained by the integration of bottom-level speech acts. Alternatively, the three forms of structure would be defined as top-down if they were instantiations of prior schemata, for

instance the linguistic one of standard patterns of sentence type (in a broad sense of this term), the world model of entity or activity frames, the communicative structure of standard speech act sequences. In both cases for any form, as Figure 5 shows, these could be compositional structures over levels of unit. The essential difference between the two views, bottom-up and top-down, is whether the structure or any particular source is wholly individual or just a manifestation of a type.

Thus with any of the bottom-up cases there are no expectations about the global structure in a text, where with the top-down ones there are expectations, which may be about the text's presentational organisation, or subject content, or communicative functions. However with both bottom-up and top-down alternatives the kinds of relation between the elements or components of the text are quite different: linguistic relations like 'repetition' link text segments; domain relations like 'temporal sequence' are between events in a world; communicative relations like 'counterclaim' hold between speech (or discourse) acts.

### 3.5 Summary implications

These construction alternatives perhaps carry with them types of strategy for forming summary representations. This could apply, too, whether or not summarising was treated as a matter of content selection or of content generalisation. For instance (and also for expository purposes) treating these construction options as strict alternatives, a bottom-up structure for source representations could naturally lead to similar bottom-up bases for determining summary content. Thus for example a linguistic representation, and the selection option, could pick up the most mentioned entity, while a world model selection could pick up the most novel or explanatory fact (e.g. as signalled by predication forms or representation structure), while for a communicative structure selection could pick out the most significant function (e.g. using speech act type or emphasis marking). Generalisation on the other hand, given a bottom-up type of structure, could emerge for a world model approach, say, with a merging of subcategory facts into supercategory ones. These approaches are illustrated in Figure 6.

On the other hand, with a top-down approach using instantiated schemata, selection would take key elements determined primarily by the nature of the global structure, as also shown in Figure 6, while generalisation would exploit the predetermined upper-level characterisation of the schema.

This discussion has been only an outline sketch, indicating the major features of the landscape we need to explore. Every concept mentioned needs to be properly analysed and defined. The area of large-scale discourse structure, in particular, is a daunting jungle. The problems involved in giving a better, actually or potentially computational account of large scale structure are well illustrated on the one hand by the ambiguity, with respect to information type, of terms I have used like "instruction" or "comparison" and of concepts like rhetorical predicate, and on the other by notions of discourse grammar, which appear to fall between the bottom up and top down forms of structure I have considered.

## 4 Investigative strategy

Thus the way I have considered discourse representation and summarising so far has been a very preliminary and rough staking out of the ground. All of the notions involved need much more investigation, and there are many questions I have not examined at all, for instance what other information resources are brought to bear during summary formation. The way we have chosen to proceed, given the framework I have presented, is by laboratory experiments constructing and using different types of source representation for the same set of test texts, to see both what the different approaches to discourse modelling can deliver and how appropriate each is for summarising. These experiments are therefore comparative ones, of a deliberately restricted kind, designed to throw light on the respective representation roles, and summary values, of different types of discourse information. Investigating each specific approach in this decompositional, analytic style should, however, not only be useful in its own right. It should also provide the foundation for consideration of the richer, multi-faceted representations, and of strategies combining different sorts of information, that are likely to be required for fully satisfactory summarising. It must however be emphasised that though we believe this is a sound overall approach, we are only making a first study which is unavoidably limited and simplified in relation to the real magnitude of the task, for instance in the test data we are using.

### 4.1 Data and procedures

Thus we are taking as our test texts ten paragraph-long pieces, three so-called 'noddy' texts originally composed for quite another purpose, and seven real weekly newspaper items. The texts are (to use the term informally) of different genres: narrative, description, argument, report. The noddy texts include the Biographies example already illustrated; the real texts deal, for instance, with acupuncture for pets, the kidnap of a sunbed salesman, and other unusual or striking topics.

We are not, in fact, processing these texts automatically to derive source representations, for two reasons. One is that we are not concerned primarily with what source text processing itself is like, only with what sort of thing it delivers. The other, perhaps more compelling reason, is that programs for completely interpreting extended text do not exist, and specifically ones for providing some of the kinds, and the range, of distinct representation types we are interested in. It is also the case that even where approaches have been implemented in the past, the specific resources, for instance characterising domain knowledge, needed to apply them to our test texts are not available or easily provided.

We are therefore obliged, as a working strategy, to engage in simulation. But we have attempted to anchor this reasonably fairly in reality by grounding our representations in the kinds of logical forms for sentences that current systems like the SRI (Cambridge) Core Language Engine (Alshawi, 1992) produce. We are taking these as our starting point for a 'baseline' source representation, consisting of sentence logical forms linked by resolved anaphors, so the baseline can be viewed as

a simple network of predicate argument forms. This baseline is then taken as the underpinning for elaborated full representations based on different principles of the kind sketched earlier.

The baseline thus has two consequences. One is that it provides primary information at the clause level of the kind that can actually be produced automatically for source text, though it must be accepted that this is generally true only in part even for our simple source texts, where the nature of the semantic representation to be delivered is not always understood. It is also of course the case that, in general, appropriate anaphor resolution can require inference on world knowledge, and the mechanisms for doing this are not to hand. For test purposes we have, however, kept fairly close to the the spirit of Sidner's strategy (Sidner, 1983), as a well-established one. The other advantage is that the baseline clause-level representation is a common underpinning for all the other full representations, so comparisons between these are better controlled. Figure 7 shows sample logical forms and resolved anaphors for part of Biographies: full details for all the test texts are given in Gladwin et al (1991).<sup>2</sup>

## 4.2 Initial experiments

In exploring alternative possibilities for full source text representations, and what information they might provide for summarising purposes, we have adopted a quite straightforward approach and have taken as type strategies either ones which are quite obvious or ones that have already been proposed by others. We make no claim to originality, and indeed have preferred to work with strategies with some independent standing because they have already been investigated by others. Our aim has been to try to cover the range of possibilities given by the combination of information type - linguistic, world, and communicative - and of derivation style - bottom-up or top-down, with at least one exemplar strategy per option; and to implement these in our simulations in as objective and general way as we can. However it must be emphasised that the illustrations we give here are heavily simplified, and also that in the more complex cases in particular, the details are necessarily only indicative.

We began by examining some extremely simple forms of elaboration, following naturally from the form of the baseline, but ones even so of an intuitively plausible kind. For example, we suppose that the source processing also delivers, in addition to the baseline, a focus history indicating what discourse entity was in focus for each clause, or rather at the end of processing each clause. This history follows from applying Sidner's algorithm, with results as illustrated in Figure 8. Having a focus history then suggests a natural strategy for summarising, namely pick out the discourse entity which has been most often in focus, with results as also shown in Figure 8. Of course, while there is something attractive about the idea that what a text is most focussed on is what the text is about, and what a text is about should be picked out as the material for a summary, simple counting is rather a crude strategy.

---

<sup>2</sup>The apparently elaborate logical forms in the figure have a wholly sound justification within the CLE framework.

More seriously, this strategy does not indicate what, for the purposes of summary, should be said, either in selective or generalising terms, about that entity.

The baseline representation also suggests another simple type of elaborated representation, barely more than a rearrangement, and an associated simple summarising strategy. This is to consider discourse entities from the point of view of the predications they figure in, i.e. to treat the elaborated representation as grouping predications, so a corresponding summary strategy is simply to select the entity which figures in most predications, as illustrated in Figure 8. But as with the previous strategy, it is not obvious what should be said about the selected entity, i.e. what predications to choose or construct. These strategies nevertheless do provide some handles on potential summary content, and might therefore be more properly viewed as starting procedures for summarising, and perhaps quite practical ones, rather than as complete summarising strategies.

At the same time, the kind of laboratory comparison we are interested in from the summarising point of view can be illustrated with these rather minimalist 'full' source representations and corresponding choices of discourse entities as elements for a summary representation, in particular the extent to which the strategies in fact identify the same key content in the source text as material for the summary. For our short texts the two strategies typically identify the same discourse entities as the central ones (see Gladwin et al 1991). Unfortunately an extremely crude technique constituting what might be described as a rock bottom strategy, namely just picking out the most frequently mentioned word, gives much the same result. But this is perhaps not surprising for such short texts.

More generally, of course, these are all rudimentary strategies: in particular, as summary strategies they are selective rather than generalising, and embody only very weak notions of discourse structure. There are obvious problems with them: thus as noted already, it is not clear what information from the source text should be given for the selected entities in the summary. The representations considered so far do not capture complex entities, especially events, or abstract over entities and events. So unless much of the work required has to be done by the summary formation component, these forms of representation do not offer much for either refined selection or serious generalisation as summarising strategies. It is clear that the baseline representation needs development for a full representation, for instance to obtain a more comprehensive picture of the event structure of a source text. It is of course also the case that our test texts are very short, and it is not clear how well simple approaches like those presented so far could work for larger ones. Correspondingly it is impossible, given such short sources, for summaries to be other than very brief, so the question not merely of large scale source structure but of its transformation into a large scale structure for the summary is not addressed.

The elaborated representations and summarising strategies using focus and predications can be viewed as concentrating respectively on linguistic and on world knowledge. We have not so far examined an analogous really simple communication-based approach, say considering speech acts as these might be associated with the baseline logical forms by an extended grammar.

At the same time, it would also be desirable to investigate an enriched baseline of

the kind used in Rieger (1975), i.e. one following the application of general inference rules to derive additional predications. These are inference rules on world knowledge, so while this enrichment might from one point of view be regarded as generating an adequate baseline, within our study framework these rules would by definition be taken as working on the baseline to obtain an elaborated full representation. With such a representation, the predication grouping and 'active' entity selection strategy would be natural for summarising.

However it is clearly essential to investigate some forms of source representation intended to capture any stronger or richer large-scale structure than those kinds considered so far. Given our working definition of bottom-up and top-down, and interest in applying prior art, we could take Kintsch and van Dijk's (1978) approach, delivering a propositional hierarchy based on discourse content, as an exemplar for bottom-up world knowledge, and Grosz and Sidner's (1986) theory as providing a bottom-up communication-based characterisation of large-scale structure. It is also possible, though somewhat problematic given its ambiguity, to treat Mann and Thompson's (1987) Rhetorical Structure Theory (RST) as dealing with bottom-up linguistic discourse notions, on the grounds that relations like 'Comparison' are marked by linguistic cues and embodied in linguistic patterns which are themselves independent of intended communicative effects. We have so far explored only the second and third of these possibilities.

Grosz and Sidner see their intentional structure as relevant primarily to the actual process of discourse production or interpretation but it can, like the focusing history, also be treated as recorded. This structure is formally defined only by two very abstract organisational relations, dominance and satisfaction-precedence, justified of course by the actual content of discourse segments. The intentional structure of a text is thus a tree, which may in processing be recognised decompositionally. However while it is possible to see intentional structure as grammar-based, this 'grammar' is so minimally constraining in terms of the discourse possibilities it allows that it is more appropriate to treat Grosz and Sidner's view of discourse, in our terms, as bottom-up. Following this approach for Biographies, to see what communicative intentions alone supply as source structure and what this offers for summarising, we get the representation shown in Figure 9, provided by Professor Grosz. A plausible approach to summarising would take the material for the top-most segment, from which the notional output summary text shown in Figure 8 might be produced. One obvious problem with this is that it is rather bulky. (It is also not clear whether the segment purpose should appear explicitly, or be left implicit.)

Our analysis for Biographies using RST is given in Figure 10. We had some difficulty in applying Mann and Thompson's definitions, and as Figure 10 shows, we could not preserve a strictly hierarchical composition. However RST source representations seem to lead very naturally to summaries, since the combination of a clear tree with the distinction between nucleus and satellite implies, as Mann and Thompson suggest, that the obvious material to be selected for a summary is that in the top-most nucleus, with the result shown in Figure 10. But while this is plausible for a short text like Biographies, because the top level relation normally spans an

entire text, its nucleus could often be far too large to be acceptable as a summary, and it would be preferable to recurse down or extract the lowest-level embedded clause.

To complement these bottom-up styles of source representation and naturally-associated approaches to summary formation we have been investigating alternative, top-down forms of representation, where individual texts are primarily instantiations of forms taken from a prior, and essentially limited, set of possibilities. (These forms need not, however, be strongly hierarchical.)

In considering potential source structure types for the three kinds of information, there is an obvious model for the use of world information, namely scripts and frames (collectively labelled scripts here). We have therefore simulated script application following the line of work developed by Cullingford (1978), DeJong (1979) and Tait (1983), but keeping our scripts simple so several may apply to one text. We have had to construct actual scripts, but have attempted to do this with some generality, so the scripts are not designed for the individual test texts: we have in part been able to do this by regarding it as normal for texts to instantiate a number of possibly not closely related scripts; we have not however followed Tait in allowing for non-scripty material in texts, as we wished to investigate a pure script strategy.

The results, illustrated for Biographies in Figure 11, which draws on 'Books' and 'Giving' scripts out of a total of around thirty for our test texts, show that it is possible to provide domain scripts covering most of the text material for each of the test texts. However they also show that important aspects of the text may not be captured in this way at all (as DeJong noted); and some texts in particular are very inadequately treated by characterisations motivated by the 'objective' structure of the world. Thus for Biographies, it is possible to construct a general-purpose frame called 'Comparison', as shown in Figure 11, but this is hardly an objective world fact, though it clearly exploits properties of the world to make comparison effective.

To obtain summaries, it is natural to follow the kind of strategy adopted in previous script-based summarising, and take header predications for instantiated scripts. We thus obtain the notional output of Figure 11.

As mentioned earlier, 'Comparison' falls into the class of so-called rhetorical discourse relations or structures; and as implied by the discussion of RST, rhetorical patterns can be viewed as linguistic types of structure, precompiled for communicative effects, rather than directly as communicative structures. Rhetorical schemata have been used for generation (e.g. by McKeown, 1985); However when we attempted to analyse our test texts as instantiations of such schemata, by analogy with our script/frame study, we found it difficult to identify candidate general-purpose schemata. We have therefore been driven to adopt an intermediate 'grammar' approach, of a more flexible kind than that used by Maybury (1991) but still defining large-scale structure much more strongly than the earlier bottom-up approaches. However, as there is not suitable rhetorical grammar to hand (and RST cannot readily be made into one), we have experimented with a rather different type of grammar for linguistic discourse structure stimulated by Rumelhart's story grammars (Rumelhart, 1975), which can perhaps be called a genre grammar. The kind of grammar, and the results of applying it to Biographies, are shown in Figure 12,

along with a resulting summary: this is obtained by applying the idea of summarising 'reduction' rules associated with grammar rules, in the way that semantic rules are associated with syntactic rules in a sentence grammar.

## 5 Conclusion

Our experiments are designed to fill the spaces in a grid of source representation possibilities, as defined by information types and derivatation styles. But as this grid is very broadly and informally characterised, and we have been taking only one or two exemplars for each space, what we are doing can only be regarded as a preliminary pass in examining discourse modelling in the context of summarising. We have also not yet completed the grid. We have, moreover, so far adopted a least-resistance approach to summarising, implying selection or generalisation as most naturally follows from each form of representation studied, and with virtually no explicit constructive effort for summary formation. We thus need to consider both versions of each generic attitude to summarising for each form of representation, even if working with short input texts means there is not much scope for constructive transformation of source into summary representation.

The conclusions we can draw now from what we have done are therefore very tentative.

First, reviewing the set of approaches, they all have some generic plausibility, even if the specific details of our exemplars are arbitrary. From the point of view of summarising in particular, they all pick up more important rather than less important material from the source. But this is not surprising, given that they are being applied in the very favourable experimental context provided by the highly-structured Biographies text. Nor is the fact that there is some overlap in the results for the different strategies. For even in the Biographies context, the strategies pick up different source content.

Thus it is already clear that while each approach to representation has its own strengths, it also has its limitations in failing to capture perceived, and hence potentially significant, generic properties and specific features of source texts. Communicative representations fail to capture text presentational patterns, for instance, while scripts miss communicative intentions. Particular key text content is thus lost, like the causal relation between the properties of books and their being recommended for children's reading. This lost source content (whether informative or expressive) thus cannot figure in summaries. At the same time, the summary strategies we have tried have their own problems. Thus they are all, as primarily selective strategies, liable to produce an arbitrary amount of poorly connected material: this makes the output using Grosz and Sidner's representation very uneven, for example.

It is also evident that the strengths and failing of the different approaches are to a considerable extent complementary, so no one is better than another. It is therefore natural to conclude that the correct line is to take account of all the information types, and to recognise that there are both a priori (top-down) and a posteriori (bottom-up) structures in discourse. Indeed the difficulties of categorising some



particular theories, like RST, might suggest that our distinctions between types and modes are really unviable. We nevertheless believe that while from one point of view 'Comparison', for instance, is a discourse structure which is at once a matter of form, substance, and intention, these can be separated out since each component individually could combine with alternatives for the others. At the same time, an intermediate grammar-based treatment of discourse organisation may still be too constraining, so a mixture of conventional and opportunistic structure may be more appropriate.

We thus see our basic framework, however crude, as a valuable heuristic device for exploring the nature of discourse structure, especially large-scale structure, and its role in supporting key discourse manipulation tasks like summarising. The obvious major question which follows from the idea that different sorts of structural information are all needed for an adequate source representation, is whether they are really compatible and can be genuinely combined in an additive way (as opposed to merely being placed in the same box); there are also the computational consequences of combination to consider. The presumption has to be that the different structures, as they can be shown to exist for texts, have distinctive and necessary roles in discourse representation; but they nevertheless, as they are partial, have to be combined to allow the required references from one to another. This appears to imply that a full discourse model may have to be defined as a complex, higher-level relational structure across lower types, as considered in Sparck Jones (1988) and, much more fully, in Grosz and Sparck Jones (1992).

We would like to investigate such 'covering' discourse models for our test texts, and their potential for summarising, but this will clearly be very difficult, even allowing for the fact that what we do is simulation not actual computation. However as DeJong's FRUMP, for example, showed, reasonable and appropriate summaries may still be obtained for particular contexts by type-limited strategies. Our first priority is therefore to complete the grid-based study in progress, and also to keep a check on the value of more ambitious treatments of representation and summarising by a comparison with a good shallow strategy, rather more sophisticated than those illustrated in Figure 8, but still text-bound.

From the point of view of nearer-term computational implementations these shallow strategies look the only viable approach, outside specific and well defined application contexts. The interest here is how modest bottom-up approaches building on sentences analyses might be combined with some sort of top-down schema resources of a simple and flexibly-applied kinds, for instance sketchy domain scripts or presentational rhetorical proformas to which cue words and even document layout and typography might provide access. We want to begin work in this area too.

Finally we need to look specifically, for both analytical and practical reasons, at the relation between local and global structure in the source representation and at its effects on, or use in, summarising. As we are not concerned with source text processing itself, we have not considered the way local and global factors interact during source interpretation, though this is clearly implied by e.g. the focussing record. However once we address summary formation in more detail than hitherto, and in particular multi-sentence summaries, we will have to study what form local

structure information about topic-focus articulation takes in source representations, and how it influences summary formation and figures in summary representation.

## 6 Acknowledgements

This work was carried out under Grant GR/F 24678, "General techniques for automatic summarising", funded by the UK Science and Engineering Research Council. The paper is derived from a lecture given at the University of Massachusetts at Amherst. I am grateful to John Beaven and Professor Barbara Grosz for discussions and examples, and to Steve Pulman for comments.

## References

- [1] Alshawi, H. (Ed) *The Core Language Engine*, Cambridge MA: MIT Press, 1992.
- [2] CL *Computational Linguistics*, Issue on User Modelling and Discourse Modelling, 14(3), 1988, Discussion section, 79-103.
- [3] Cullingford, R.E. *Script application: computer understanding of newspaper stories*, Thesis, Yale University, 1978.
- [4] DeJong, G.F. *Skimming stories in real time: an experiment in integrated understanding*, Thesis, Yale University, 1979.
- [5] Gladwin, P., Pulman, S. and Sparck Jones, K. *Shallow processing and automatic summarising: a first study*, Technical Report 223, Computer Laboratory, University of Cambridge, 1991.
- [6] Grosz, B.J. and Sidner, C.L. 'Attention, intention, and the structure of discourse', *Computational Linguistics* 12, 1986, 175-204.
- [7] Grosz, B. and Sparck Jones, K. 'Knitting discourse structure', Working paper, 1992.
- [8] Halliday, M.A.K. and Hasan, R. *Cohesion in English*, London: Longman, 1976.
- [9] Hajicova, E. and Panevova, J. 'Valency (case) frames of verbs' in *Contributions to functional syntax, semantics, and language comprehension*, (Ed Sgall), Prague: Academia, 1984.
- [10] Hobbs, J.R. *Literature and cognition*, Lecture Notes 21, Centre for the Study of Language and Information, Stanford University, 1990; especially Chapter 5.
- [11] Kintsch, W. and van Dijk, T.A. 'Toward a model of text comprehension and production', *Psychological Review* 85, 1978, 363-394.

- [12] Lambert, L. and Carberry, S. 'Modelling negotiation subdialogues', *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 1992, 193-200.
- [13] Mann, W.C. and Thompson, S.A. *Rhetorical structure theory: a theory of text organisation*, Report ISI/RS-87-190, Information Sciences Institute, University of Southern California, 1987.
- [14] Maybury, M.T. *Planning multisentential English text using communicative acts*, Thesis, University of Cambridge, 1991; Technical Report 239, 1991.
- [15] McKeown, K.R. *Text generation*, Cambridge: Cambridge University Press, 1985.
- [16] Moore, J.D. *A reactive approach to explanation in expert and advice-giving systems*, Thesis, University of California at Los Angeles, 1989.
- [17] Moore, J.D. and Pollack, M.E. 'A problem for RST: the need for multi-level discourse analysis', *Computational Linguistics*, in press.
- [18] MUC-3, *Proceedings, Third Message Understanding Conference (MUC-3)*, San Mateo, CA: Morgan Kaufmann, 1991.
- [19] Paice, C.D. 'Constructing literature abstracts by computer: techniques and prospects' *Information Processing and Management*, 26, 1990, 171-186.
- [20] Rieger, C.J. 'Conceptual memory and inference' in *Conceptual information processing* (Ed Schank), Amsterdam: North-Holland, 1975.
- [21] Rowley, J. *Abstracting and indexing*, London: Bingley, 1982.
- [22] Rumelhart, D.E. 'Notes on a schema for stories' in *Representation and understanding* (Ed Bobrow and Collins), New York: Academic, 1975.
- [23] Rumelhart, D.E. 'Understanding and summarising brief stories' in *Basic processes in reading*, (Ed Laberge and Samuels), Hillsdale NJ: Lawrence Erlbaum, 1977.
- [24] Salton, G. *Automatic text processing*, Reading MA: Addison-Wesley, 1989.
- [25] Sgall, P., Hajicova, E. and Panevova, J. *The meaning of the sentence in its semantic and pragmatic aspects*, Prague: Academia, 1986.
- [26] Sidner, C.L. 'Focusing in the comprehension of definite anaphora' in *Computational models of discourse*' (Ed Brady and Berwick), Cambridge MA: MIT Press, 1983.
- [27] Sparck Jones, K. 'User models, discourse models, and some others', *Computational Linguistics* 14, 1988, 98-100.

- [28] Sparck Jones, K. 'What sort of a thing is an AI experiment?' in *The foundations of artificial intelligence: a sourcebook* (Ed Partridge and Wilks), Cambridge: Cambridge University Press, 1990.
- [29] Sparck Jones, K. 'Discourse modelling: where are we now and where should we be going?' *Working Notes, AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, 1991, 142-145.
- [30] Tait, J.I. *Automatic summarising of English texts*, Thesis, University of Cambridge, 1983; Technical Report 47, Computer Laboratory, University of Cambridge, 1983.
- [31] van Dijk, T.A. and Kintsch, W. *Strategies of discourse comprehension*, New York: Academic, 1983.
- [32] Young, S.R. and Hayes, P.J. 'Automatic classification and summarisation of banking telexes', *Proceedings, Second Conference on Artificial Intelligence Applications*, 1985, 402-408.

FIGURE 1 : Example text, "Biographies", and human summaries

## TEXT

Biographies are the best books. They are about real things. They tell a true story. They are about particular people. They have lessons for their readers. Smiles wrote biographies. They are about successful workers. These biographies inspire their readers. Their readers imitate these workers. History books are about real things. They tell true stories. They are about states. They are not about particular people. They do not inspire their readers. Novels tell a story. They are about particular characters. They are not about real things. They are dangerous. Biographies are valuable books. Histories are useful. Novels are deceitful books. Give children biographies. Give them histories. Do not give them novels.

## SUMMARIES

### P 1 :

Biographies are about real individuals, while histories are not about individuals and novels are not about real people. So biographies are best for children.

### P 2 :

Biographies and histories are better for children than novels.

### P 3 : 1989

Give children biographies, maybe histories, but not novels.

### P 3 : 1988

Teach them facts.

FIGURE 2 : Global process model for summarising

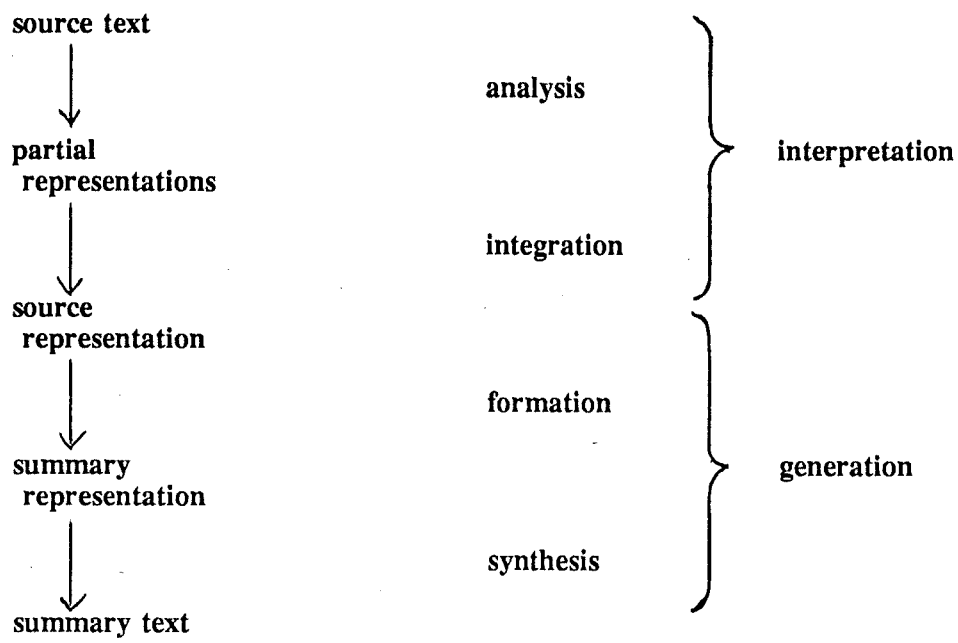


FIGURE 3 : Types of information in discourse

...  
Biographies are valuable books.  
Histories are useful books.  
Novels are deceitful books.  
...

**LINGUISTIC**

**sentence type** : declarative, generic ...  
**lexical item** : "valuable", "book" ...  
**discourse form** : parallel sentences, words

**WORLD**

books  
**type** : biography  
**property** : valuable  
**type** : history ...

**COMMUNICATIVE**

**speaker cause hearer believe** :  
biography valuable  
history ...

FIGURE 4 : Types of large-scale structure in discourse

**LARGE SCALE STRUCTURE**

1 = biography, 2 = history, 3 = novel, 4 = giving  
1,1a; 2,2a etc refer to the first and second chunks of material in  
the text about biographies, histories etc

**LINGUISTIC**

argument  
comparative description  
description 1 2 3  
description 1a 2a 3a  
instruction 4 } S ?

**WORLD**

books  
type [1 2 3]  
properties 1 1a 2 2a 3 3a  
possible effects 1 2 3 [items from]  
implied actions 4 } S ?

**COMMUNICATIVE**

S intend H act 4  
S motivate H  
S ensure H know (inform 1 2 3)  
S justify H act (inform 1a 2a 3a)  
S instruct H act } S ?

**Summaries**

**Linguistic**

Biographies are valuable, histories useful, novels deceitful. Give children biographies and histories, not novels.

**World**

Biographies are true and inspire, so give them to children. Histories are true, so give them. Novels are false, so do not give them.

**Communicative**

Give children biographies because they are valuable. Give histories because they are useful. Do not give novels as they deceive.



FIGURE 5 : Alternative derivation principles for large-scale structure

**BOTTOM-UP CONSTRUCTED**

**linguistic**

eg thematic units

**world**

eg linked facts

**communicative**

eg tied speech acts

**TOP-DOWN INSTANTIATED**

**linguistic**

eg sentence type pattern

**world**

eg object properties

**communicative**

eg speech act sequence

FIGURE 6 : Alternative styles of summary formation

### **BOTTOM-UP SELECTION**

#### **linguistic**

(most) mentioned entity  
biography

#### **world**

(most) novel fact  
valuable  
(most) explanatory fact

#### **communicative**

(most) significant function  
give

### **BOTTOM-UP GENERALISATION**

#### **world**

things + things → things  
books are about something

### **TOP-DOWN INSTANTIATION**

#### **linguistic schema ...**

#### **world frame**

reading matter  
select  
biography good  
generalise  
some books for reading

#### **communicative plan ...**

FIGURE 7 : Baseline source text representation

### LOGICAL FORM

Biographies are valuable books.

```
[dcl
  quant (forall A,
    [biography1, A]
    quant (exists B,
      [and [book1, B], [valuable1, B]]
      quant (exists C,
        [state C]
        [be C [eq A,B]] ))) ]
```

Give children biographies.

```
...
[biography1, A]
...
[child1, B]
...
[and [event C], [before now, C]]
[give1, C, hearer, A, B]
```

### ANAPHOR RESOLUTION

Biographies are the best books.

↑  
They are about real things.  
↑  
They ....  
↑  
They ....  
↑  
They ....

Smiles wrote biographies.

They are about successful workers.

↑  
These biographies ....

....

FIGURE 8 : Example text bottom-up representations and summaries - linguistic type

**FOCUS**

Biographies are ...	b1
...	b1
Smiles wrote biographies.	b2
...	b2
History books ...	b2
...	h1
Novels ...	h1
...	n1
Biographies are valuable.	b1
Histories ...	h1
Novels ...	n1
Give children biographies.	b1
Give them histories.	c1
Do not give them novels.	c1

<b>most focussed</b>	b1	7
	h1	6
	b2	5
	n1	4

**ACTIVITY (roles)**

be books biographies  
 have lessons readers biographies  
 tell stories history books  
 give biographies children

<b>most active</b>	
biographies	7
history books	7
novels	6

FIGURE 9 : Example text bottom-up representation and summary - communicative type

## INTENTIONS

**DS1 [Purpose: convince biographies best, histories alright for children, novels not alright**

biographies best ... real ... true ... particular

**DS2 [Purpose: provide illustrative example**

smiles wrote ... workers ... inspire ... imitate  
] DS2

**DS3 [Purpose: convince "real people" feature important**

histories real ... true ... states ... not people ... not inspire  
] DS3

**DS4 [Purpose: convince "real" is important**

novels ... story ... characters ... not real ... dangerous  
] DS4

**DS5 [Purpose: recapitulate important properties**

biographies valuable ... histories useful ... novels deceitful  
] DS5

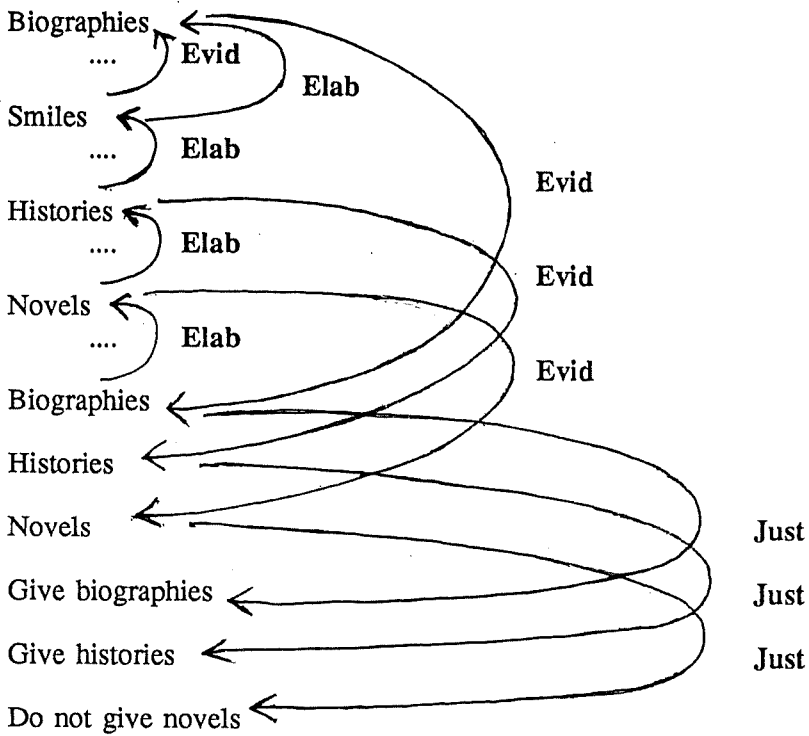
give children biographies ... histories ... not novels  
] DS1

**summary : topmost segment material**

Biographies are the best books. Biographies are about real things, true stories, particular people. Biographies have lessons for readers. Give children biographies, histories, not novels.

FIGURE 10 : Example text Rhetorical Structure Theory representation and summary

**Evid = Evidence; Elab = Elaboration; Just = Justification**



**summary : nucleus clause of topmost nucleus**

Give children biographies. Give children histories. Do not give children novels.

FIGURE 11 : Example text top-down representation and summary - world type

**SCRIPTS / FRAMES**

**BOOK**

**subset : BIOGRAPHY**  
    **subject : real things**  
        ...  
    **property : true story**  
        ...  
    **subset : (BIOGRAPHY)**  
        **author : smiles**  
        ...

**subset : HISTORY**

...

**COMPARISON**

**compared : BIOGRAPHY, HISTORY, NOVEL**  
    **parameter : ?**  
    **comparison : best BIOGRAPHY**

**GIVING**

**recipient : children**  
    **gift : BIOGRAPHY, HISTORY**

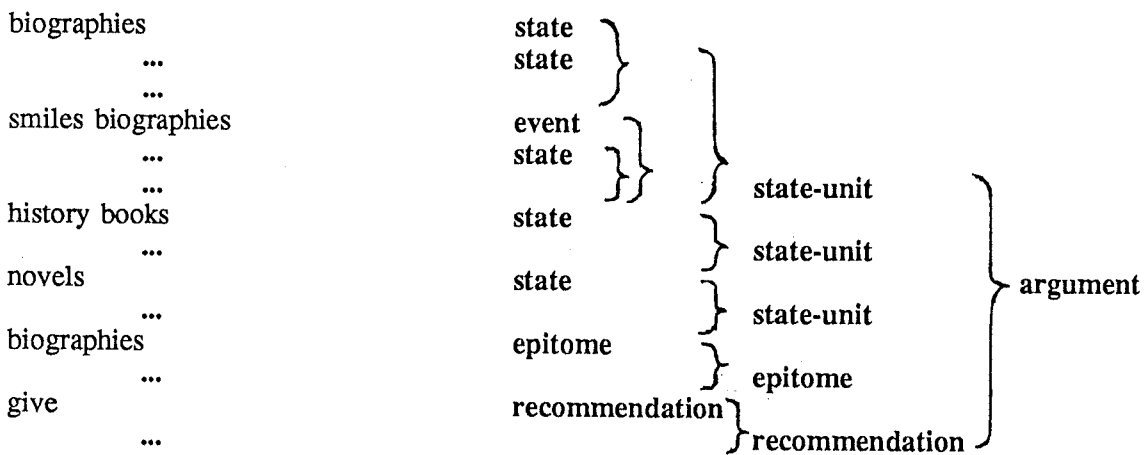
**summary : instantiated script header predication**

There are biographies, histories, novels with various properties. In comparison, biographies are best. Biographies, histories are gifts for children.

FIGURE 12 : Example text discourse grammar representation and summary

**TEXT GRAMMAR (part)**

text → narrative | ... | argument  
 narrative → { setting } ... commented-episode +  
 commented-episode → { comment } episode  
 comment → (event-unit | state-unit) +  
 episode → { ... } event-unit | state-unit +  
 event-unit → event<main>, event<subord> | state<subord> +  
 state-unit → state-unit, comment  
 state-unit → state +  
 argument → { epitome } event-unit | state-unit +, recommendation  
 epitome → epitome +  
 recommendation → recommendation +



summary : {epitome} recommendation

Biographies are valuable, histories useful, novels deceitful. Give children biographies, histories, not novels.