Variable Length On-Line Document Generation

Abstract

This paper describes a system for *variable-length document presentation*: on-line documents whose length can be adjusted to the user's demands. The system depends on an initial marking-up of documents using an *RST Markup Tool* - a graphical interface for marking up the rhetorical structure of a text. During presentation, the rhetorical structure is used to prune the text down to the size requested by the user, allowing retention of the essentials of the text.

1 Introduction

As we move into the the use of the web, more and more documents are becoming available on-line. However, different users have different needs from these documents. Some users, in a hurry, may desire brief and succinct documents. Others may require more detail. Users may also vary as to the type of information they want from a document.

This paper describes an experiment with on-line text presentation – whereby the user specifies how long the document should be. The system then presents a coherent document fitting that space limitation. The user might choose to see the hundred-word version, or the thousand-word version, or somewhere between.

Figure 1 shows the web browser (Netscape) interface to the system, also showing (part of) the text before it is reduced. Figure 2 shows the same document, although with a 200 word limit set. The text is mostly coherent, with however some minor problems. One can see these as the cost of this sort of summarisation.

This technique, what we call *variable-length text presentation*, involves two steps:

1. **Document Preparation**: the document is marked-up according to its rhetorical structure. For this we use an RST Analysis Tool, which allows a user to graphically link segments of text into an RST-tree.

2. **Document Presentation**: a web-connected program is then used to present such documents. In response to a user's request, the program 'prunes' off less essential branches of the RST-tree until a text of the required size is produced.

The system was an attempt to see how far we could push a notion mentioned by Sparck Jones (1993), that RST can be used to summarise a text, shaving off less relevant satellites. Can we remove rhetorically dependent sub-sections of the text without markedly affecting the coherence of the text?

Our pruning method involves assigning a level of relevance between 0 and 1 to each RST relation. Using these values, we can work out the relevance of each node in the RST-tree: the top-node having 1.0 relevance, each of satellite in the tree having relevance proportional to the relevance of its nucleus times the relevance of the relation linking it. We then prune off text-nodes with lowest relevance until the required word-limit is reached. This process is described in section 3.

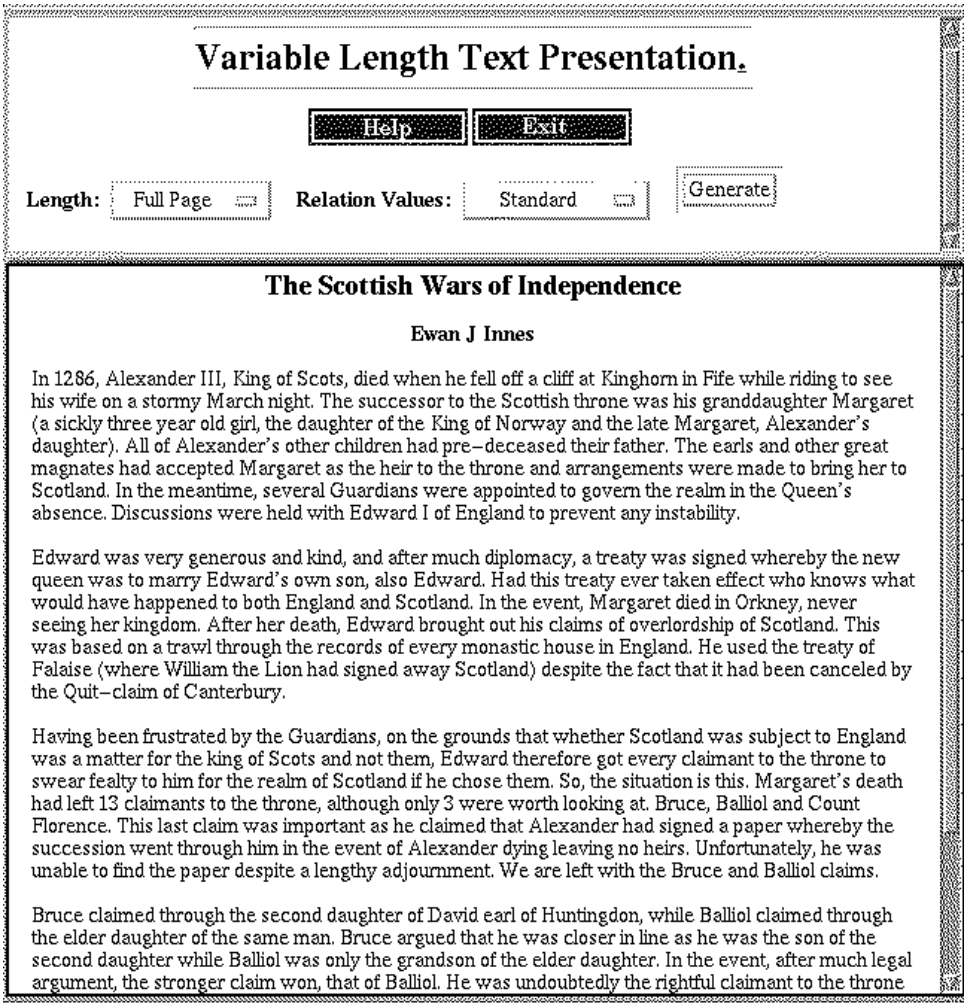

Figure 1: The VLTP interface

The Scottish Wars of Independence

Ewan J Innes

Alexander III, King of Scots, died. The successor to the Scottish throne was his granddaughter Margaret. The earls and other great magnates had accepted Margaret as the heir to the throne and arrangements were made to bring her to Scotland. In the meantime, several Guardians were appointed to govern the realm. Discussions were held with Edward I to prevent any instability. A treaty was signed whereby the new queen was to marry Edward's own son. Margaret died. Edward brought out his claims of overlordship of Scotland. Edward therefore got every claimant to the throne to swear fealty to him for the realm of Scotland if he chose them. Margaret's death had left 13 claimants.

Bruce claimed through the second daughter of David earl of Huntingdon, while Balliol claimed through the elder daughter of the same man. Bruce argued that he was closer in line. The stronger claim won, that of Balliol. Balliol was crowned and he was faced with constant pressure from Edward. He refused to do so. Edward gave the Scots an ultimatum. And the Scots instead signed a treaty of mutual aid with France. Edward invaded Scotland instead.

Figure 2: Scottish History text at 200 words

The system also allows a small degree of user-determination of the content. The RST-pruning uses information on the relative importance of each RST relation. If the user is given control of these importances, then they can tailor the kinds of information that is actually left in the document.

In section 4, we, we address various areas of incoherence introduced by the pruning (paragraphing, punctuation, reference and discourse markers), and our solutions to these problems.

In section 5, we describe the RST markup tool which makes it possible to conceive of doc-

ument presentation based on RST markup. RST-based document summarisation has been stopped in the past because of the present poor state of automatic discourse structure recognition. Hand-markup is an arduous task, but the tool we report here makes the task economical for some documents. However, keep in mind that because of the time-cost of document markup, this technique is only useful for documents with a longer shelf-life. We must weigh the cost of analysing the original document against the benefits of having a variable-length on-line document.

Finally, section 6 will attempt to assess the usefulness of this approach, detailing the quality of the presented documents, against the problems involved in the presentation. Some extensions of the work are also suggested.

1.1 Relevance to Generation

Given that this technique involves neither text-planning nor sentence-planning, one might ask how this paper is relevant to the Generation community. Firstly, the technique of *RST-pruning*, reported in section 2, is applicable to pruning of RST-structures generated by a full-blown text-planner. A text-planner could produce fully-elaborated rst-structures from an underlying knowledge-base, and then present pruned versions of the text depending on the users needs. We can thus apply the techniques reported here for variable-length document presentation to variable-length document *generation*.

Secondly, this work is also of interest to the Generation community because of the contained report of the RST Markup Tool. RST is used widely within the generation community, and this tool may prove useful to many, not only as an aide in their corpus studies, but also for preparing diagrams for publications.

1.2 Related Work

Summarisation via RST-pruning was suggested by Sparck Jones (1993), although the mechanism for determining which satellites to prune is unique here. Also, her work was limited by the lack of automated RST analysis, while I rely on semi-automated markup. The application of the technique to produce variable-length documents is also unique.

Rino & Scott (1996) offer a more detailed account of summarisation via pruning in a full generation environment. However, they prune the content structure rather than the discourse structure. The RST tree produced to express the pruned content structure is not itself pruned. On the other hand, their content structure is similar enough to RST that similarities to the present work are observed. They take intention structure into account to drive the pruning, which would be a valuable addition to the methods proposed here. While I believe they are right in that text summarisation needs to take both these areas (and others) into account, I am interested here to see how well rhetorical structure by itself can form the basis of summarisation.

Ed Hovy, in his involvement with the HealthDoc project, has suggested generation from a master document – a set of SPL (semantic specifications of sentences), each conditionalised by the user model (see DiMarco et al 1995). The text actually seen by the user is achieved by pruning out SPLs which are inappropriate for the user-type. The present system differs from this approach in that, while their master document is RST-structured, that structure is not used as the basis of the pruning, but only to restructure the pieces chosen. Also, the production of sub-documents is intended to produce user-tailored documents, not length-tailored ones.

I am aware of work by Veli J. Hakkoymaz (Hakkoymaz in-preparation; Hakkoymaz&Ozsoyoglu 1996) on Variable-Length Multimedia Presentations, whereby multimedia segments are added to or dropped from a presentation in order to meet the time constraints. That approach allows substitution of elements as well as deletion, which may be a useful technique.

2 Variable-Length Document Presentation

Any document marked up for RST can be used for variable-length document presentation. This section describes the process whereby the rst-structure is pruned to produce a suitable length document.

2.1 Assigning Relevance Scores to Text Nodes

As described in the introduction, the basic mechanism involves assigning each structural relation a relevance score between 0.0 and 1.0. For instance, ELABORATION may have a score of 0.40 (low relevance), while PURPOSE might be scored more highly.

By an RST-tree, I assume a tree with the top-nucleus as the root of the tree, and satellites hanging off this, and their satellites hanging off of them. Our task is then to prune branches off of this tree. The top-nucleus has a relevance value of 1.0 (maximum relevance).

Through a process of recursive descent, we assign each node in the tree the relevance level of its parent, multiplied by the relevance score of the relation which connects them to the parent. For instance, an ELABORATION of the top-nucleus would have relevance 0.4 (1.0 * 0.4), while an ELABORATION of that node would have relevance 0.16 (0.4 * 0.4). Nodes lower in the RST-tree (less nuclear) will thus have lower relevance than higher nodes (more nuclear), and will thus be the first to be pruned.

This is a simple mechanism, but it has shown good results in producing reasonable texts at whatever degree of verbosity. It is easy to see that an elaboration of an elaboration will in most cases be less essential to a text than the elaboration itself.

However, there are some cases where this method breaks down – nuclearity does not always reflect centrality of information. Sometimes an author introduces information in a rhetorically unimportant place, yet that information may be needed later to understand the argument. One example of this in the summary shown earlier is where the original text had said: *he was faced with constant pressure from Edward to sign. He refused to do so.* In the summary, "to sign" was pruned as, but it was actually a central concept, and the anaphoric "so" failed because of its pruning.

The text-nodes are then placed in a queue, position based on their relevance score.

2.2 Pruning the RST-tree

When a request is received to display the text at a particular length, the system needs to determine which text-nodes to display. Taking each node in turn from the relevance queue (starting with the most relevant), the program checks to see if including this text node will push the word-count over the limit. If not it adds the node to the nodes-to-be-expressed list, and increments the words-so-far count. When the word-limit is exceeded, the procedure then turns to expressing the selected nodes. The nodes are expressed in the order in which they appeared in the original full text.

Note that the satellites of a node will always have lower or equal relevance than the node itself, so we never include a satellite in the nodes-to-be-expressed list if its nucleus is not, which may produces incoherency.

2.3 Extensions on Basic RST

The RST Markup Tool, and consequently document presentation, allows markup of more than simple nuclear-satellite relations. This includes:

- **Multinuclear Relations**: such as JOINT and SEQUENCE.
- **Schemas**: what are sometimes called "story grammars" allowing a sequence of named elements of structure, e.g., INTRODUCTION, BODY, CONCLUSIONS, BIBLIOGRAPHY, etc.
- **Clause-Internal Structure**: for this summarisation work, I have been pushing RST analysis inside the sentence – not only in terms of analysing the relations between clauses in a sentence, but also analysing the relation between clausal adjuncts and the nuclear clause. For instance, (N: *Edward surrendered,*)(S: *in 1245*). Some of these adjuncts can be connected to the clause with standard RST relations, but many can not. A set of new relations, borrowed from the Systemic labelling of adjuncts (cf. Halliday 1985), has been added for this reason.

Allowing the intermixing of story grammars and RST greatly increases the representative power of the formalism, and subsequently helps in text pruning. For instance, if we provide the INTRODUCTION and CONCLUSIONS relations higher relevance values than BODY, then these sections will be more prominent in any summary.

All of these structures are handled in terms of the relation (role) linking the constituent to the whole, and this relation is handled identically to simple RST relations in text pruning.

2.4 User-Variation of Relation Weightings

The actual values associated with each relation are not fixed, but can be varied by the user. The user can select values which reflect their interests, highlighting some types of rhetorical relations, and ignoring others.

The system comes with three inbuilt 'user-models', representing different ranges of interest: (*standard*, (average values), *how&why* preferring cause, reason, purpose, conditionals, etc., and *when&where*, preferring spatial- and temporal-locations and extents. Figures 3 demonstrate the slight difference of information (bold font) included in the text when switching between the *when&where* set and the *how&why* set. We might also add such sets as *naive*, preferring definitions, clarifications, restatements, and elaborations, while an *expert* might value these less, but prefer generalisations, etc. Apart from these built-in values, the user can also assign values to each relation independently.

How&Why Summary: Alexander III, King of Scots, died. The successor to the Scottish throne was his granddaughter Margaret. The earls and other great magnates had accepted Margaret as the heir to the throne and arrangements were made to bring her to Scotland. Several Guardians were appointed **to govern the realm**. Discussions were held with Edward I **to prevent any instability**. A treaty was signed **whereby the new queen was to marry Edward's own son**. Margaret died. Edward brought out his claims of overlordship. **He used the treaty of Falaise**. ...

Where&When Summary: **In 1286,** Alexander III, King of Scots, died **at Kinghorn in Fife**. The successor to the Scottish throne was his granddaughter Margaret. The earls and other great magnates had accepted Margaret as the heir **to the throne** and arrangements were made to bring her to Scotland. **In the meantime,** several Guardians were appointed. Discussions were held with Edward I. A treaty was signed. Margaret died **in Orkney**. **After her death,** Edward brought out his claims of overlordship **of Scotland.** ...

Figure 3: Summaries with different weighting sets

---

3 Preserving Coherence in Dynamic Document Presentation

When summarising a document, we do damage to various aspects of the document's coherency. These aspects will be covered below under four topics: paragraphing, punctuation, referring expressions and discourse markers.

3.1 Paragraphing

Deleting sentences without changing paragraph boundaries would produce a text of many short paragraphs, reducing readability. Rather than attempt to repair document paragraphing, we have found it easier to throw away the original paragraphing, and re-determine paragraph boundaries as described below.

Paragraphing within a document is intended to make it easier to read. It segments the discourse into small chunks of sentences which are to some degree highly related. We found it plausible to use our RST structure to help in determining paragraph boundaries. From looking at texts, it is the usual case to see a paragraph representing a nucleus and its satellites (although some other of its satellites be in other paragraphs).

There is a useful notion used in speech synthesis and generation which claims that the spacing between spoken words can be predicted largely by the *syntactic distance* between them – the number of branches which have to be traversed in the parse tree to move from one word to the other. Thus, in *the Girl Guides fish*, we would expect little pause between noun *Guides* and its modifier *Girl*, while in the homophone *the girl guides fish* we would expect more pause between the verb *guides* and the subject *girl*.

We have applied this principle to paragraphing, arguing that two adjacent sentences which are more discoursally distant (more structurally separated in terms of the RST-tree) are more likely to be separated by a paragraph break.[1]

This is not the whole story however. Paragraphing is also constrained by the needs of *paragraphic rhythm*. Martinec (1995) argues that the division of texts into paragraphs is similar to the rhythmic structure of the sentence (divided into tonic feet of similar interval). Both are means of organising information into manageable chunks. The *rhythm* of a text requires that these chunks are of approximately the same size, not too long, not too short.

Our paragraphing algorithm combines these two notions – semantic distance and paragraphic rhythm – to determine paragraph boundaries in the presented texts. We assume there is an "ideal" paragraph length for the text, the paragraph rhythm (user configurable). Starting at the beginning of the text, we test each point between sentences for a possible paragraph-break. We evaluate two factors:

1. **Semantic Distance**: how many arcs of the RST-tree do we need to traverse to get from one sentence to the other. In a sense, we are looking for the weak-points in the text, textually adjacent sentences which are not semantically closely related.
2. **Projected Paragraph Size**: how much smaller or larger than our ideal would the paragraph be if we broke the paragraph at that point.

We use the following formula to evaluate each possible paragraph break, and select the point with the lowest value (I will leave fuller explanation to a paper dedicated to the topic):

$$Score(N_i, N_j) = (ideal\_length - actual\_length)^k + \frac{j}{sem\_dist(N_i, N_j)}$$

...where ideal_length, k and j are constants. I have found best results with values of 150, 1.2 and 75. Lower values of k allow more variation of paragraph size in seeking for better breaks on semantic distance grounds.

[1] An alternative approach might evaluate potential paragraph breaks on the basis of the *number* of nucleus-satellite links that boundary breaks compared to other possible breaks. This approach would reward paragraphs which are sub-trees of the RST. In addition, we might penalise what we might call *foster* sentences – sentences which have no direct relation to the other sentences in that paragraph.

Once a paragraph position is selected, we take that as our starting point and look for the next paragraph boundary after that, until the end of the text is reached. As you can see from figures 1 and 2 (both paragraphed using the above formula), the method produces quite plausible paragraphing.

3.2 Punctuation

As reported above, we have allowed the RST Tool to assign structure *within* the sentence as well as *between* sentences. This however creates a problem because, in deleting an intra-sentence nucleus, we may also delete the punctuation it carries. For instance, in (N: *Edward surrendered,*)(S: *in 1245*), deletion of the nucleus leaves us with a sentence terminated by a comma.

One module of the present system has been developed to correct such problems. It ensures all sentences start with a capital, and recovers the sentence-terminating punctuation from any pruned segments where necessary.

3.3 Referring Expressions

When deleting sections of a text, we may destroy the referential cohesion of a text in two ways. Firstly, we might delete the introduction of an entity, which provided the entities name, or other characteristics which allow the reader to identify the entity correctly. The remaining text may refer to this entity (e.g., "he"), but leave no clue as to who the entity is. The second, related, problem involves changing the referential environment of entities. References which are contextually unambiguous in the full text may be brought into close proximity to other entities which are potential confusers.

In the system as implemented so far, there has been no attempt to correct these problems. Cases of problems have been rare. However, for the next stage of implementation we are planning to introduce NP markup into the document preparation stage, allowing the document editor to indicate co-reference of NPs in the text. This would be a simple matter of allowing the editor to drag from each NP to a co-referring NP.

From this markup, we can deduce various things. We can identify the first-occurring reference for each entity, and with a reasonable level of certainty, use this as the first-mention of the entity in any pruned-text. We can analyse the remaining references to discover gender (from pronouns) or class (from definite or indefinite references). Where text-pruning places two entities of similar gender in proximity, the class-based or name-based reference form could be used if available. In this way, many of the reference problems can be repaired. An anaphora generation module being developed by Janet Hitzeman is a good candidate for use here.

The extra cost of NP markup needs to be weighed against the gain of coherency gained.

3.4 Discourse Markers

Markers of rhetorical relations are usually attached to satellites, and so there is no problem when the satellite is pruned. However, in some peoples analyses, some relations mark the nucleus, not the satellite. In others, both the nucleus and satellite are marked (e.g., if/then). When we delete the satellite, we should ensure that the discourse marker is removed also from the nucleus. However, due to the rarity of nucleus marking, this problem rarely occurs.[2]

For those cases where nucleus marking does occur, a future applications might avoid the problem by removing all discourse markers from the marked-up text, and generating these as appropriate. However, I envisage problems associated with this approach, including over-generation of discourse linkers (many are left implicit).

[2]In the case of if/then, I have the ELABORATION relation set to 100% relevance, since a clause without its condition has a totally different meaning.

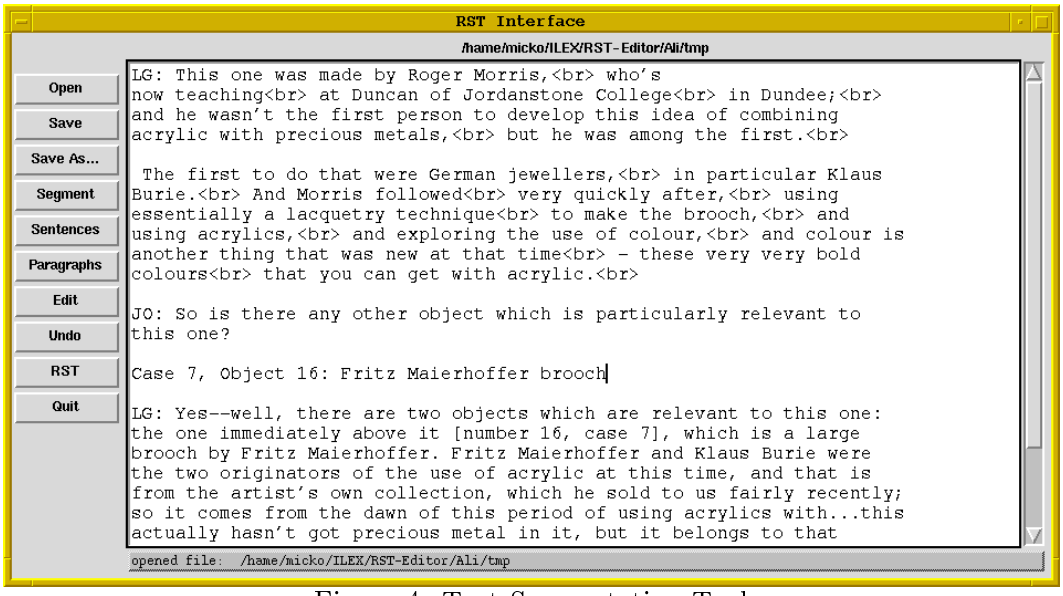

Figure 4: Text Segmentation Tool

4 Document Preparation

Before the text can be used for variable-length presentation, it needs to be marked-up in terms of RST structure. To facilitate this step, we have developed an RST Markup Tool, which allows a user to:

1. Segment the text.
2. Graphically link these segments together into an RST-tree.

4.1 Text Segmentation

Each of these tasks has a separate interface within the tool. The first is shown in figure 4. The buttons "Sentences" and "Paragraphs" result in automatic recognition of sentence and paragraph boundaries. If further segmentation is required, the user can switch into *segmentation* mode, during which they need only click at each segment boundary to introduce a segmentation marker. To edit the text (modifying the text, correcting spelling errors, etc.), switch to the *Edit* mode.

A problem occurs with *embedded elements* – cases where a rhetorically dependent stretch of text occurs within another node. For instance, we might wish to treat the embedded clause in the following as dependent on the main clause: *John, – I think you know him – is here for two weeks.* At present, the interface does not handle such cases. A simple solution is for the user to move the embedded text outside of the enclosing text.

4.2 Text Structuring

The second step of document preparation involves structuring the text. Another interface of the RST Markup Tool allows the user to connect the segments into a rhetorical structure tree, as shown in figure 5. We have followed the graphical style presented in Mann & Thompson (1987).

Initially, all segments are unconnected, ordered at the top of the window. The user can then drag the mouse from one segment (the nucleus) to another (the satellite). Upon releasing the mouse button, the system offers a menu of relations to choose from (the user can use the relation-sets provided with the system, or provide their own).

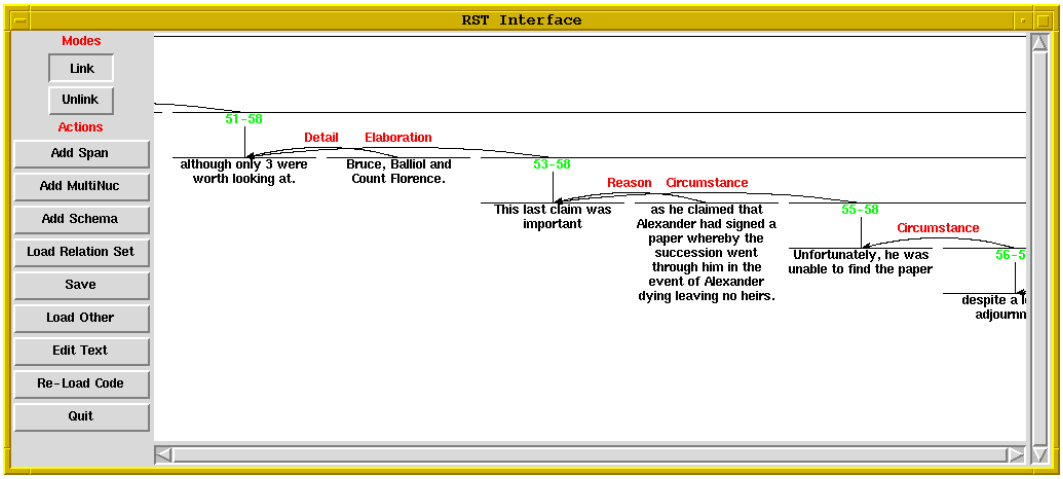

Figure 5: Text Structuring Tool

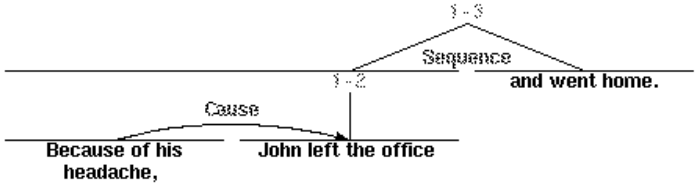

Figure 6: Scoping and multi-nuclear relations

The system allows both plain rst-relations and also multi-nuclear relations (e.g., joint, sequence, etc.). Scoping is also possible, whereby the user indicates that the nucleus of a relation is not a segment itself, but rather a segment and its satellites. See figure 6 for an example of both a multi-nuclear structure, and scoping. In addition, McKeon-style *schemas* (sometimes called *story-grammars*) can be used to represent constituency-type structures. See figure 7.

The user can switch freely between text segmentation and text structuring mode – to edit text, or to change segment boundaries. The system keeps track of the structure assigned so far. If the user, in editing the text, deletes a segment, the system forgets structuring information concerning that segment.

Because rst-structures can become very elaborate, the RST Tool allows the user to *collapse* sub-trees – hiding the substructure under a node, This makes it easier, for instance, to connect two nodes which normally would not appear on the same page of the editor.

The user can save the present state of the screen as postscript, for inclusion in Latex documents. Alternatively, a snapshot utility can be used to save selected parts of the structure in other formats. The structured text can be saved to a file, for later re-editing, or for use in variable-length document presentation.

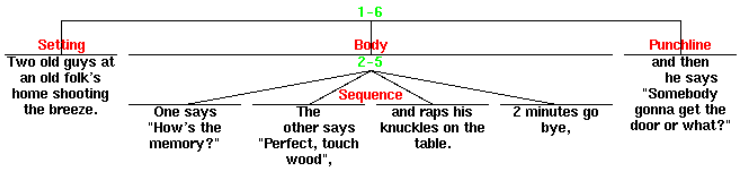

Figure 7: Constituent Structure

5 Summary

This paper has described a system for presenting variable-length on-line documentation, which allows the user to select the degree of verbosity of the text presented. The results so far on a small-scale have shown that reasonable-quality texts can be produced dynamically. The cost of document markup stops this approach being used on texts of short display-life, but makes it economical for documents of longer duration where length-variability has value.

Apart from text-length, Variable-Length documents allow the user a small degree of content-control, to the degree that they can determine the relevance of each RST relation (or of elements of a schema).

The major problem for the system involves restoring coherence after text-pruning, particularly in areas of reference, discourse markers, paragraphing and punctuation. The problems of paragraphing and punctuation have been solved, and solutions are suggested for the other two areas.

Another problem occurs when material important to the text is not included in nuclear positions in the RST-tree: nuclearity does not guarantee importance to discourse goals (although there is a strong correlation between nuclearity and importance). This is why, in the long term, approaches such as Rino&Scott (1996), which take intentional structure as well into account show some promise. While information about intention structure is not easy to mark up, it would be available in a system doing full text generation from intentions.

Regardless of the problems of this approach, the system is up and running on-line. New documents are being added as time allows, to test the generalisability of the approach.

Future development will include features such as allowing the user to zoom in on text by clicking on it. I will soon make sentence punctuation hyper-clickable, which would result in the pruned text under that sentence being provided.

The notion of variable length on-line documents has great value to information providers and information readers alike – imagine if this document had been provided variable-length, you could have read the two page version instead!

6 Bibliography

DiMarco, Chrysanne; Graeme Hirst, Leo Wanner & John Wilkinson 1995. "HealthDoc: Customizing patient information and health education by medical condition and personal characteristics". Workshop on Artificial Intelligence in Patient Education, Glasgow, August 1995.

Hakkoymaz, Veli J. (in prep) "Organizing Variable-Length Multimedia Presentations within a Given Deadline".

Hakkoymaz, Veli J. & Gultekin Ozsoyoglu 1996 "Automating the Organization of Presentations for Playout Management in Multimedia Databases". IEEE Int'l Workshop on Multi-Media Database Management Systems, Aug. 1996.

Halliday, M.A.K. 1985 *Introduction to Functional Grammar*. London: Edward Arnold.

Mann, William C. & Sandra Thompson, 1987. "Rhetorical Structure Theory: A Theory of Text Organization". Technical Report ISI/RS-87-190.

Martinec, Radan 1995 *Hierarchy of Rhythm in English Speech*, Ph.D. dissertation. Dept. of Semiotics, University of Sydney.

Rino, Lucia & Donia R. Scott 1996 "A Discourse Model for Gist Preservation". Lecture Notes issue, Special issue on the Proceedings of the XIIIth Brazilian Symposium on Artificial Intelligence.

Sparck Jones, Karen. 1993. "What might be in a summary?", Information Retrieval 93: Von der Modellierung zur Anwendung (Ed. Knorz, Krause and Womser-Hacker), Konstanz: Universitatsverlag Konstanz), 9-26.