3 Getting the Message Across in RST-based Text Generation

Donia R. Scott and Clarisse Sieckenius de Souza

Abstract

This chapter examines the problem of generating texts that achieve their communicative goals in an effective way. We discuss an approach to producing effective text that is geared towards ensuring that the rhetorical aspects of a message are not only preserved but enhanced in the text. This approach is strongly influenced by research in psycholinguistics and the psychology of memory, and is based on a view of stylistics as a matter having rather more to do with cognition than aesthetics.

3.1 Introduction

Text generation can be characterised as a process of transforming a message into a text. This process is successful if, and only if, the reader of the text is able to derive its intended message. The ultimate criterion of what it means for a text to be good is thus a cognitive rather than a strictly linguistic one: the easier it is for the reader to decode the intended message from the text, the better the text will be. Given this, it is therefore important to be able to specify just what are the characteristics of a text whose message is easily retrievable. Grice [1975] addresses this issue in his Manner maxim:

> Be perspicuous: avoid obscurity of expression, avoid ambiguity, be brief and be orderly.

Clearly, if it is our aim to generate text that conforms to this directive then it is crucial for us to be able to determine just what are the defining features of the characteristics that Grice so strongly recommends. When, for example, is a text brief and when is it not? Unfortunately, Grice provides us with no more than a quick glimpse of what these directives amount to in practice. Neither do other potential sources of information provide us with much more to go on; essays and textbooks on good writing tell us what we should and should not do, but give us next to no indication about how we could or should go about doing it. For example, Mark Twain, in his critical essay on the art of good writing (*Fenimore Cooper's Literary Offences*), simply tells us that the writer must:

- say what he is proposing to say, not merely come near to it;
- use the right word, not its second cousin;
- employ a simple and straightforward style;
- not omit necessary details;
- avoid slovenliness of form;
- eschew surplusage; and
- use good grammar.

Similarly, Strunk and White [1979], in what is perhaps the most popular book of the genre of writing textbooks, define conciseness in the following terms:

> A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, for the same reason that a drawing should have no unnecessary lines and a machine no unnecessary parts. This requires not that the writer make all his sentences short, or that he avoid all detail and treat his subjects only in outline, but that every word tell.

In the absence of clearer guidelines, we are thus left with the task of specifying these directives by hypothesising their meaning, implementing our hypotheses in working systems, and judging the validity of the hypotheses by their effect on the goodness of the resulting text. Given the cognitive nature of our stated criterion for good text, it makes obvious sense to look to cognitive models of reading to provide the foundation for our hypotheses. That is, our hypotheses should be based on psycholinguistic models of language comprehension. In the work presented here, we discuss some of the hypotheses we have developed in this way for planning the generation of good text. The hypotheses we discuss are based on a general model of text understanding suggested by Clark and Clark [1977]. They are aimed at achieving clarity and conciseness in the textual expression of the message at the rhetorical level.

In our approach, the message of a text is comprised of a set of propositions which form the leaves of a hierarchical rhetorical structure that expresses the writer's intentions behind the inclusion of each proposition. If the message is coherent, then each of its constituent propositions contributes to the overall intention—that is, to the writer's communicative goal. We have chosen to represent messages within the framework of Rhetorical Structure Theory (RST) [Mann and Thompson 1985, 1987a, 1987b]. RST provides a number of extremely useful features for text generation, many of which are discussed at length in Hovy [this volume]. Most important among them for the present discussion is that RST can be used to represent both the message and the text plan and that it provides a means for capturing the notions of relevance and coherence within the representation of messages. The advantage of these is that, taken together, they provide the basis for maintaining the control that is required during the generation process to ensure that the necessary mappings from message to discourse and syntax are meaning-preserving. Like many other text planning systems (e.g. McKeown [1985]; Paris [1987]; Hovy [1988a], [1988b]) the basic elements of our messages are verb-based, clause-sized propositions, each of which can be expressed as a single sentence.[1]

Our particular concern here is that of ensuring that the texts we generate convey the rhetorical structure of the message in an effective manner.[2] This means that our texts must conform to the following three basic requirements. First, they must be sensitive to the communicative context in which they are set, i.e., one where the writer is an artificial interlocutor, with few resources for predicting or judging the impact of the text on the reader. Second, the chosen expression of the message must be a valid and unambiguous textual rendition of its rhetorical structure (i.e. the rhetorical structure of the message must be derivable from the text). Third, the chosen expression must be the most easily processable member of the set of all valid and unambiguous expressions of the message.

3.2 Making the Text Sensitive to the Communicative Setting

Clearly an important factor in effective writing is the tailoring of the expression of the message to suit one's intended audience. The need to provide tailored communication is now well recognised among designers of cooperative human-computer interfaces, and there are a growing number of dialogue systems which attempt to provide just this sort of interaction by incorporating a model of the user (see Kobsa and Wahlster [1989] for a detailed discussion of these systems). Although user models provide a useful basis for tailoring a system's contribution to a dialogue, they cannot be expected to be reliable representations of the user. Since artificial interlocutors clearly have fewer possibilities to make reliable assessments of their audience's ability to 'get the message' than do their human equivalents, their expressions of the message often need to be more explicit than would be ideal. This is particularly the case with respect to the rhetorical aspects of messages, whose understanding generally relies heavily on common-sense knowledge. Also of relevance is the

[1] It is worth noting that this also accords with the prevailing view in psycholinguistics. See Clark and Clark [1977] and van Dijk [1977] for a discussion of this, and Johnson Laird [1983] and Garnham [1985] for an alternative view.

[2] Meteer [1988a, 1988b, 1989] presents a rather different approach for planning text that is clear and concise at the propositional level.

fact that although most user models attempt to represent the overlap between the system's knowledge base the user's beliefs, few attempt to represent the user's beliefs about the relationship between the facts in the knowledge base.

It follows from the above that since the present generation of computer systems cannot reliably determine whether their users will be able to correctly infer the rhetorical structure of a message from its constituent facts, they should therefore make this structure explicit to the user. Our first hypothesis is thus:

> Hypothesis: Readers are unlikely to retrieve the rhetorical structure of a message unless it is stated explicitly.

To account for this factor, one of our heuristics for guiding generation is therefore:

> Heuristic 1: Always generate accurate and unambiguous textual markers of the rhetorical relations that hold between the propositions of the message.

Our examination of two unrelated languages, British English and Brazilian Portuguese (see Souza, Scott and Nunes [1989]), shows that at least in these languages all rhetorical relations of the set proposed by Mann and Thompson [1987b] can be marked textually. These rhetorical markers may be lexical, phrasal or purely syntactic, and their roles in the language are strictly pragmatic. For example, the ANTITHESIS relation can be signalled in a number of ways: *rather than, instead of, however,* and *yet.* Similarly, the EVIDENCE relation can be signalled by *since*, *because*, and *therefore*, RESTATEMENT by *in other words* and PURPOSE by *in order to.* Some relations, in particular ELABORATION, can only marked by syntax.

A requirement of text generators, then, is that they include information about the appropriate textual markers of each rhetorical relation.

3.3 Generating Valid and Unambiguous Textual Markers of Rhetorical Relations

In order to ensure that only valid and unambiguous markers are generated, it is important for the generator to know not only what the markers of each relation are, but also how they can be used in the target language. It must know, for example, that as a marker of EVIDENCE, *since* can only be attached to the satellite of the relation, and can only be used intrasententially, but with any ordering of the satellite and nucleus, whereas *therefore* can only be attached to the nucleus, can be used intersententially and can only be used with the satellite presented before the nucleus. This type of information permits the

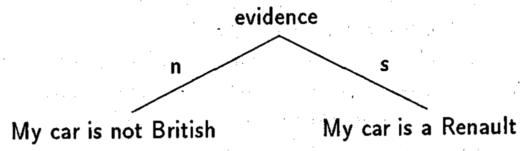


Figure 3.1: An instance of the EVIDENCE relation

generation of a message, such as that shown in Figure 3.1, as sentences (1)–(3), all of which are valid expressions of the message and therefore ones from which the message is retrievable. It also prevents the generation of sentences such as (4)–(8), none of which are valid expressions of the message and some of which (sentences (6)–(8)) are valid expressions of some other message and therefore likely to 'give the wrong message'.

(1) Since my car is a Renault, it's not British.

(2) My car is not British since it's a Renault.

(3) My car is a Renault, therefore it's not British.

(4) *Therefore my car is not British, it's a Renault.

(5) *My car is not British. Since it's a Renault.

(6) *My car is not British, although it's a Renault.

(7) *Since my car is not British, it's a Renault.

(8) *My car is not British, therefore it's a Renault.

In addition to the above constraints on the use of rhetorical markers, there are also situations where strong constraints are placed on the tense that can be used to express elements of a relation in conjunction with particular markers. For example, in Brazilian Portuguese, the satellite of a CONCESSION relation marked by *embora* can only be expressed in the subjunctive.

Avoiding the generation of ambiguous markers can often prove difficult, since there are a number of markers that apply to more than one relation.

When this happens, however, it is generally the case that the set of applicable relations form a superclass. One such superclass involves the abovementioned markers of EVIDENCE, which also happen to be markers of what Mann and Thompson [1987a] call the **cause cluster**: the VOLITIONAL CAUSE, VOLITIONAL RESULT, NON-VOLITIONAL CAUSE, NON-VOLITIONAL RESULT and PURPOSE relations. That the marker *because* (i.e., *be* + *cause*) should validly apply to them all is hardly surprising, since they are all organised around the concept of causation. Although not considered to be part of the cause cluster, it can be argued that EVIDENCE properly belongs to it, since it too involves the notion of causation. Given all this, the applicability of *because* to all of six rhetorical relations does not prove problematic, since it narrows down the choice of applicable relations to a well-defined set whose differences are not so great as to lead the reader off-track.

There are other markers, however, which are so ambiguous as to be almost meaningless; *and* is a case in point. *And* can be used to link the elements of most, if not all, rhetorical relations. It is a strong marker of only a few of these relations (to be discussed below) and an extremely weak marker of the rest, where it tends to mark not a rhetorical relation between the elements that it is linking, but merely the fact that they are part of the same piece of discourse [Gleitman 1965; Lakoff 1971]. Its weak use is rather more prevalent in speech than in text, and this may well be related to the fact that the transient and immediate nature of speech, as compared to text, means that it is rather more difficult to undo errors of omission without disrupting the comprehension of the message. But the real point to be made here is that rhetorical markers are better thought of as strong clues to the presence of a specific relation than as proof of its presence, and that although some degree of ambiguity will have to be tolerated, ambiguities that arise from the generation of a very weak marker that also happens to be a stronger marker of another rhetorical relation should not count among them. There are good reasons why this should be the case. Firstly, there is a wealth of psycholinguistic evidence from studies performed in the 1970s which suggest that ambiguities can produce quite severe disruptions to the comprehension process (e.g. [Lackner and Garrett 1972; Foss and Jenkins 1973]). Secondly, there is also strong evidence to suggest that rhetorical markers have such a powerful influence on language comprehension that people will try to make sense of what they read purely on the basis of the marker, even when what they are reading makes no sense at all [Fillenbaum 1971, 1974a, 1974b].

Having established the means by which we can provide the reader with strong clues to the identity of rhetorical relations, we are well on our way to being able to generate texts whose message is *retrievable.* But it must be remembered that we are aiming for more than this—for texts whose message is *easily* retrievable. So, the obvious question here is: given the range of possibilities for signalling a rhetorical relation, what is the most effective way to do so? Clearly, the answer to this will be related to the psychology of language comprehension.

The task of comprehending a text involves transforming a linear string (the text) into its underlying, hierarchical structure (the message). This holds for texts (and messages) of all levels of complexity—from those consisting of a single clause (one proposition, no rhetorical relations) to those of paragraph or even book length (many propositions, any number of rhetorical relations). Psycholinguistic studies suggest that readers interpret text by reconstructing its propositions and using them to continually build onto a hierarchical representation of propositions.[3] The ease with which this construction process occurs is heavily dependent on the effect it has on the consumption of working memory resources.

Working memory is the mental work space we use during comprehension [Newell and Simon 1972]. During reading, it is where we reconstruct the message in order to interpret its meaning. Comprehension proceeds by processing the basic constituents of the message and then using them to build coherent units with previously-processed constituents that are being stored in working memory. Once a coherent unit is formed, it is added to the main structure (the interpreted message so far) in episodic memory. There is known to be a rather direct relationship between the processing and storage resources of working memory: the more storage that is required at any particular moment, the fewer the resources available for processing the incoming data [Baddely 1986; Daneman and Carpenter 1980]. Because of this trade-off, the human comprehension process is organised so as to minimise the number of partial products that need to be held in working memory. This means that the more structured the input is, the easier it will be for the reader to derive its underlying message. Syntax plays a major role in helping structure the incoming information so that it can be retained in working memory until the parts needed to complete it have been processed. Textual features such as long distance dependencies, digressions, or constituents that involve a lot of processing, will place heavy demands on working memory and thus slow down the comprehension process. Similarly, undoing previously built structures will be costly. Related to this is the suggestion that readers typically purge working memory, retaining only the gist of what was stored, at sentence boundaries [Jarvella 1970, 1971]. This view of language comprehension presents at least two immediate tips for us. The first is:

> Hypothesis: The greater the amount of intervening text between the propositions of a relation, the more difficult it will be to reconstruct its message.

[3] See Clark and Clark [1977] for a review of the literature on this topic.

which leads us to include the heuristic:

> Heuristic 2: Keep the propositions of a rhetorical relation together in the text.

Failure to keep the propositions of a rhetorical relation together in the text will result in the introduction of long-distance dependencies. As mentioned above, long-distance dependencies hinder comprehension. One reason for this is that the closer the propositions of a relation are in the text, the less time they will need to be stored in working memory before their rhetorical link can be recovered. Another reason is that there is a natural tendency for readers to attach each new constituent to the one that came immediately before it [Kimball 1973]. This explains why the ordering of unlinked sentences in a text has such a strong effect on its overall interpretation. For example, by producing Hovy's example message about the ship Knox (Hovy, this volume, Figure 2.3) as:

(9) Knox is heading SSW. It is of readiness C4. It is at 79N 18E. It will arrive on 4/24. It will load for 4 days. It is en route to Sasebo.

instead of the suggested

(10) Knox is en route to Sasebo. It is of readiness C4. It is at 79N 18E. It is heading SSW. It will arrive on 4/24. It will load for 4 days.

we end up conveying the wrong message, since the text incorrectly implies that the place where Knox is intended to arrive on 4/24 for 4 days loading is not Sasebo, but some place on Knox's route between 79N 18E and Sasebo.

The second tip we get from the cognitive model is:

> Hypothesis: Rhetorical relations that are expressed within a single sentence are more easily understood than those expressed in more than one sentence.

This hypothesis is suggested by the finding that readers tend to purge working memory at or soon after the end of a sentence. If this is the case, then it makes obvious sense to include the heuristic:

> Heuristic 3: Make a single sentence out of every rhetorical relation.

The question of how to distribute the propositions of a message as sentences in the text is one which Hovy [this volume] poses as one of the unresolved issues in paragraph planning. It has a major bearing on the style of the final text, especially in approaches (such as ours, and those mentioned above) where the input units to the generator are clause-sized propositions. In such cases, the text can, in principle, contain any number of sentences: from one to as many as there are propositions.

The above heuristic addresses this issue rather directly. It proposes that parts of a rhetorical relation should not be realised as individual sentences; neither should they be combined as sentences with parts of other rhetorical relations. Rather, they should all together form a sentence.

Application of this heuristic will increase the efficacy of the generated text in the following ways. First, the text will be more concise, since it will contain fewer sentences. Second, its message will be clearer since (a) there will be more opportunities for generating the textual markers of its rhetorical relations (most rhetorical markers can only be used intrasententially) and (b) sentence scoping will be guaranteed not to distort the hierarchical structure of the message since sentence boundaries will be conterminous with boundaries of rhetorical relations. Finally, the text will be easier to understand since the presentation of rhetorical units as sentence units will require less storage and processing in working memory, thus making its underlying rhetorical structure easier to rebuild.

In summary, this heuristic not only provides an effective approach to the problem of determining sentence scope, but it also provides one that is theoretically motivated. Decisions about where to place sentence boundaries are not based on *ad hoc* aesthetic criteria to do with how good the text will look, but rather on criteria which ensure that sentence allocations enhance rather than disturb the accessibility of the message, and on psycholinguistic factors that are known to facilitate the processing of a piece of text. This is not to say, however, that the issue of sentence scoping is now resolved. Although we have a fairly clear idea of what sort of text we should not generate in this respect, we have little idea of what we should generate. In particular, two outstanding problems remain: that of when to stop adding propositions to sentences, and that of how best to combine our clause-sized propositions to form complex sentences.

The first problem arises from the fact that the rhetorical relations of a message may be complex structures comprised of a number of other relations. It goes without saying that a coherent text can only be produced from a coherent message, which, in terms of RST, is a message that is spanned by a single rhetorical relation. Given Heuristic 3, this means that any message could, in principle, be packed in its entirety into a single sentence. Such complete freedom would be undesirable, since there is clearly a point where a sentence becomes 'too long'. Although as writers we seem to be able to recognise, and thus avoid producing, overly lengthy sentences, it is difficult to specify what the criterion of 'too long' actually is. Again, it is easier to say what it is not. It is clearly not the number of words *per se*. For example, (11a) below is more acceptable than (11b), even though it has more than double the number of words.

(11) a Mary's son Lawrence, the difficult one that everyone always said would come to a bad end, was fatally attacked by piranhas in the Pantanal last month, despite having been warned repeatedly by the local fishermen that it was dangerous to swim in the Cuiaba river.

b Lawrence, who married the very elegant young Austrian woman who used to run a boarding school for the illegitimate children of aspiring back-benchers, is a solicitor.

Neither does it seem to have much to do with number of rhetorical relations or number of propositions *per se*, since (11a) also has more of both of these than (11b). Rather, the answer seems to lie in some complex combination of factors which include number of words, number of relations, number of propositions, and syntax; factors such as the 'balance' of the text also seem to play a role. Just what the magic algorithm is, is unclear to us, and we do not know of any empirical studies on this topic.[4]

The second problem arises from the fact that there is more than one way to make a complex sentence from a set of clauses: through embedding, paratactic coordination or hypotactic coordination. Not surprisingly, there is a strong correlation between the syntactic specification of a complex sentence and its perceived rhetorical structure. This means that certain types of complex sentences are likely to be better expressions of a given rhetorical relation than others, and that the wrong choice of sentence type may lead to the wrong interpretation of the underlying rhetorical relations. So in addition to knowing *when* we must combine propositions to form complex sentences, we also need to know *how* we should combine them.

The remainder of this chapter presents an approach to the problem of producing only the most appropriate choice of complex sentence for a given rhetorical configuration. In what follows, we will be discussing only two of the three types of clause combining: embedding and paratactic coordination. Our approach to the generation of hypotactically coordinated sentences is discussed in greater detail in Scott [in preparation].

3.3.1 Embedding

Although embedding is considered by some linguists[5] to be a separate activity from clause combining, we do not adhere to that distinction here, on the purely practical grounds that our propositions are always clausal units.

[4] There are, however, a number of studies which examine the individual effect of some of these factors.

[5] See Matthiessen and Thompson [1987] for a discussion of this.

Following from Heuristic 3, that only valid and unambiguous markers should be generated, our investigations have led us to apply the following heuristic for embedding:

> Heuristic 4: Embedding can only be applied to the ELABORATION relation.

We restrict embedding to the ELABORATION relation since this relation appears to us to be the only one of the set of existing relations for which it is appropriate. It is also significant that embedding provides the only valid means by which the propositions of an ELABORATION relation can be combined to form a complex sentence. It is also the only available textual marker of ELABORATION.[6]

Embedding provides an extremely reliable syntactic cue to the semantic subordination of the embedded material to that of its matrix. It marks the embedded material as being less relevant to the message. The implication of this for RST is clear:

> Heuristic 5: When embedding, the nucleus of the relation must form the matrix of the sentence, and the satellite the embedded clause.

This heuristic guarantees that embedding preserves the hierarchical relationship of the propositions to which it is applied. So, for example, an ELABORATION relation with (12) as nucleus and (13) as satellite could result in (14) or (15) but not (16) or (17).

(12) The substance is fatal.

(13) The substance is illegal.

(14) The illegal substance is fatal.

(15) The substance, which is illegal, is fatal.

(16) The fatal substance is illegal.

(17) The substance, which is fatal, is illegal.

If, on the other hand, (13) were the nucleus and (12) the satellite, then (16) or (17) could be produced, but never (14) or (15).

In cases where the nucleus of the ELABORATION relation is complex, then there may be more than one candidate matrix proposition. Some direction for choosing among them is therefore required. This is provided by:

[6] That is, with the possible exception of phrases like *by the way* or *to be specific*. These, however, are more likely to be repair markers, for introducing information that has been erroneously left out, than rhetorical ones.

> Heuristic 6: When embedding, the matrix proposition must be the earliest occurring candidate in the immediate nucleus of the to-be-embedded proposition.

This means that given the RST structures in Figure 3.2 with both (a) and (b) as candidate matrix clauses for (c), the chosen proposition will be (a) in structures (i)–(iv) and (b) in all others.

The impact of this heuristic on the prevention of stylistic blunders is considerable. For example, suppose the following instantiations were made to the elements in Figure 3.2:

(18) a My car is French
b My car is a Renault
c My car is new
R1 EVIDENCE
R2 ELABORATION

Then Heuristic 6 would ensure that only sentences like (19) and (20) could result from embedding, and never ones like (21) or (22):

(19) Since my new car is a Renault, it's French.

(20) My new car is French, since it's a Renault.

(21) Since my car is a Renault, it, which is new, is French.

(22) My car is French since it, which is new, is a Renault.

Not only does the complexity of the nucleus provide opportunities for producing stylistic blunders when embedding, but so too does the complexity of the satellite. This occurs when embedding has the side effect of destroying the integrity of another relation. An example of this arises in cases where the embedded proposition is an element of a LIST relation and the result of embedding it is a LIST containing only one element. Heuristic 7 acts to prevent such an occurrence.

> Heuristic 7: Propositions of a LIST relation should not be embedded if doing so would make the number of remaining propositions in the relation equal to 1.

This heuristic not only preserves the integrity of the message (since the LIST relation requires more than one proposition), but it prevents the production of dangling sentences.

Dangling sentences occur when information that is only weakly relevant to the message is produced as a separate sentence. This is always the case

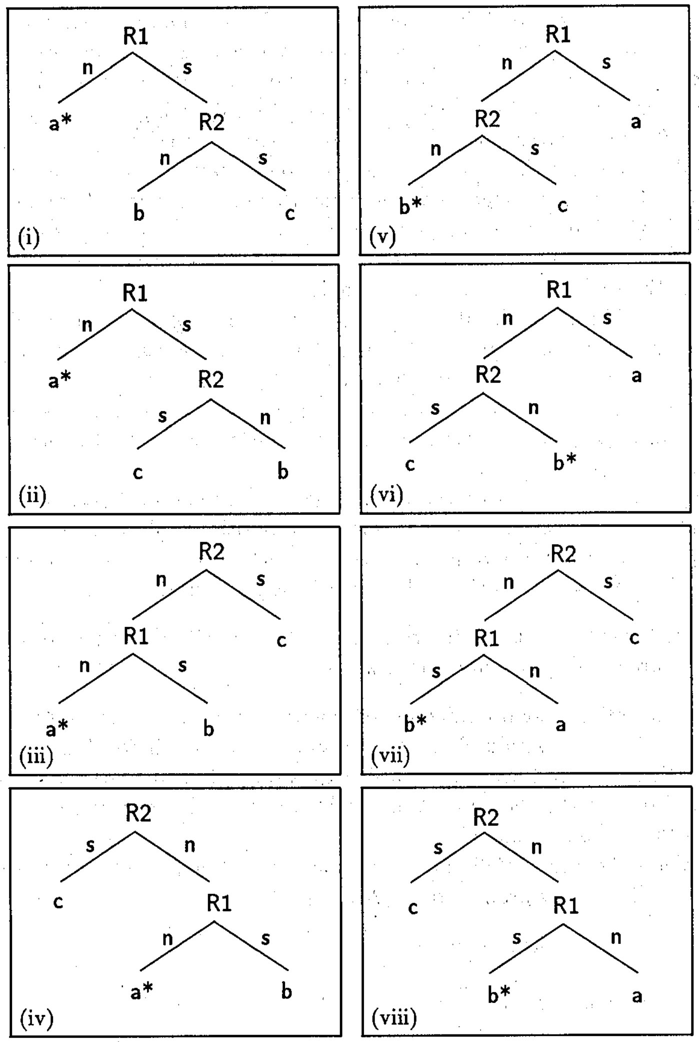


Figure 3.2: Structures involving embedding

when satellites of an ELABORATION relation are not subject to embedding. ELABORATION is the weakest of all rhetorical relations in that its semantic role is simply one of providing 'more detail'. The information contained in its satellite is thus only weakly relevant to the message. Since embedding is the only textual marker of ELABORATION, the only alternative to not applying it is to generate the satellite as a separate, and thus dangling, sentence.

Dangling sentences can have a severely disruptive effect on the comprehensibility of a text. They give the impression of having been included as an afterthought, or of introducing a new topic which is then abruptly abandoned. Integrating the content of such sentences with the preceding text is made difficult by the fact that their content is made more perceptually prominent in the text than it actually is in the message.

Stylistic blunders can also arise from the choice of syntactic realisation for the embedded satellite. Embedded clauses can be realised as nominals, adjectivals or adverbials. Although the choice between these realisation classes will be determined by strictly semantic aspects of the propositions, there is still a choice to be made regarding the most appropriate syntactic form within the chosen class.

Adjectivals can be expressed as an adjective, a relative clause or a prepositional phrase, adverbials as an adverb or prepositional phrase, and nominals as a noun or an appositive phrase. Within each class, some expressions will lead to better text than others. This can be expressed as:

> Heuristic 8: Syntactically simple expressions of embedding are to be preferred over more complex ones.

We use the notion of syntactic complexity here for want of a better term to refer to the move from lexicalised to phrasal and clausal modifiers. This heuristic biases the generation process towards expressing the embedded clause as an adjective or adverb. The impact of this heuristic on the resulting text can be seen in the following examples.

When embedding (24) in (23), preference would be given to a rendition as (25) over the equally grammatical (26) or (27):

(23) A man bought the picture.

(24) The man had blond hair.

(25) A blond man bought the picture.

(26) A man with blond hair bought the picture.

(27) A man who had blond hair bought the picture.

Similarly, preference would be given to the production of (28) over (29):

(28) Paula danced with Peter willingly.

(29) Paula danced with Peter with willingness.

Heuristic 8 enhances the readability of the resulting text in two important ways. Firstly, it reduces the possibility of generating ambiguities, since relative clauses can be restrictive or non-restrictive, and prepositional phrases can be adjectival or adverbial. Secondly, it necessarily makes the text more concise, since lexicalised modifiers involve fewer words than phrasal or clausal ones: (25) is clearly more concise than (26) and (27), and (28) is more concise than (29).

It should be noted, however, that there are exceptional cases where the application of Heuristic 8 may lead to stylistic blunders. Notable among them are those resulting from the generation of low-frequency adjectives or adverbs over their more commonplace, wordy equivalents: for example, *rancourously* instead of *with rancour*.

There are strong similarities between our use of Heuristic 8 and Meteer's [1988a, 1988b, 1989] treatment of the expression of verbal predicates in SPOKESMAN, which has the effect of preferring simple verbs over their corresponding complex ones or verb phrases (e.g. *decided* over *made a decision*, or *fed* over *gave food to*).

Our final heuristic for embedding controls the types of multiple embeddings that are allowed.

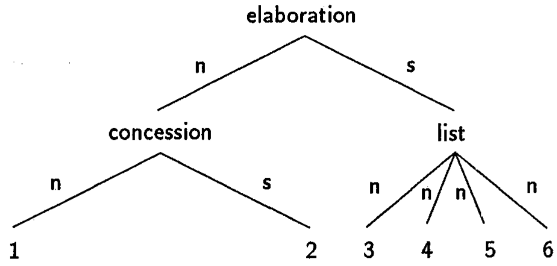
> Heuristic 9: Self-embedding is only allowed in cases where the proposition that is the deeper of the two embeddings is expressed as an adjective or adverb.

This heuristic ensures that self-embeddings do not lead to comprehension difficulties. So, for example, it allows sentences like (30) to be generated, but not ones like (31) with the same number of self-embeddings.

(30) The dog [that likes the [black] cat] is sad.

(31) *The dog [that likes the cat [that disappeared]] is sad.

Even more important, it guarantees that double centre embedded sentences are never generated. Double centre embedded sentences (e.g. *The dog that the cat that the rat saw chased died*) are definitely to be avoided since they are notoriously difficult to process [Miller and Isard 1963, 1964; Schlesinger 1968; Freedle and Craun 1970; Carpenter and Just 1989], and are known to



[1] George received a letter from Peter.

[2] George had told Peter never to contact him.

[3] George and Peter are brothers.

[4] George and Peter are estranged.

[5] The letter was long.

[6] George is my friend.

Figure 3.3: A message to which the embedding heuristics can be applied

slow down the comprehension process by as much as 58% [Larkin and Burns 1977].

The global impact of the heuristics controlling embedding can be demonstrated by their effect on the process of transforming the message shown in Figure 3.3. Taken together, their application would result in embedding that provides for the possibility of generating:

(32) My friend George received a long letter from his estranged brother Peter, even though he had told Peter never to contact him.

which is a stylistically good rendition of the message. The heuristics prevent the generation of alternative, equally grammatical but less easily understood renditions such as (33)–(37):

(33) *My friend George received a long letter from his estranged brother Peter, who he had told never to contact him.

(34) *George, who received a long letter from his estranged brother Peter even though he had told Peter never to contact him, is my friend.

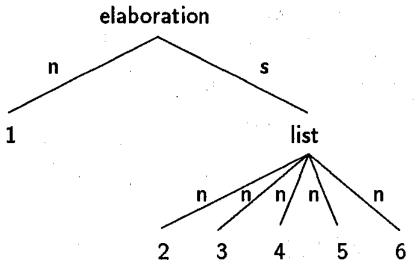


Figure 3.4: The message conveyed by example (33)

(35) *My friend George received a long letter from his brother Peter, even though he had told Peter, from whom he is estranged, never to contact him.

(36) My friend George received a letter from his estranged brother, Peter even though he had told Peter never to contact him. The letter was long.

(37) My friend George received a letter from Peter, who is his brother and from whom he is estranged, even though he had told Peter never to contact him.

In (33), the embedding of proposition [2] in proposition [1] results in the loss of the CONCESSION relation to which they belong. As a result, the text does not convey the message shown in Figure 3.3. Rather, it conveys the message shown in Figure 3.4. The possibility of generating this text from the given message is prevented by Heuristic 4.

Again, the text in (34) does not convey the message in Figure 3.3. By having [6] as the main clause, the text expresses instead the message in Figure 3.5. Such a possibility is prevented by Heuristic 5.

Although the message in Figure 3.3 is, in fact, derivable from the text in (35), the stylistic blunder that is created by the embedding of [4] in [2] instead of [1] makes message retrieval rather more difficult than it need be. The possibility of producing this stylistic blunder is prevented by Heuristic 6.

Like (35), (36) is a valid expression of the desired message. However, extracting this message is made difficult by the presentation of [5] as a separate, dangling sentence. This is prevented by Heuristic 7.

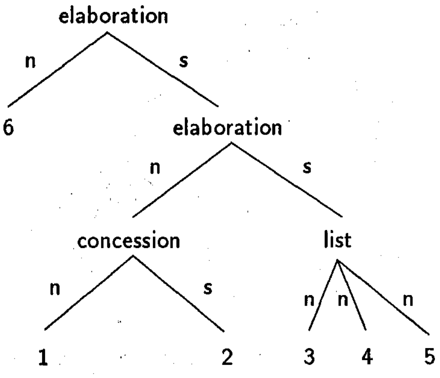


Figure 3.5: The message conveyed by example (34)

Sentence (37) is another valid expression of the desired message but one that is made unnecessarily difficult to process. This difficulty is caused by presenting [3] and [4] as relative clauses instead of adjectives. The production of this type of stylistic blunder is prevented by Heuristic 8.

The possibility of providing an example message to demonstrate all 6 embedding heuristics has only been prevented by our lack of imagination in constructing examples.

3.4 Paratactic Coordination

Paratactic constructions are complex sentences involving the *coordinate* conjoining of one or more sentential units linked by a coordinating conjunction (*and*, *or*, *but*).[7] The use of the coordinate conjunctions as rhetorical markers is not, however, restricted to paratactic conjunctions. They are also often used as weak rhetorical markers in hypotactic constructions, as can be seen in (38)–(40), where the (a) versions involve a coordinate conjunction and the (b) versions a subordinate one.

[7] Our use of the term **parataxis** is thus wider than that of Quirk *et al.* [1985].

(38) a The printer is broken and the chapter is due tomorrow.
b The printer is broken and I haven't been able to print out the chapter.

(39) a The laser printer is broken but the line printer is working.
b The laser printer appears to be broken but it does work.

(40) a Turn off the printer or unplug it at the wall.
b Turn off the printer or it will overheat.

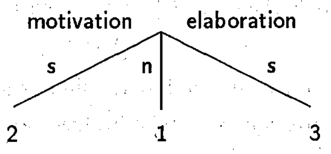
This overuse of coordinating conjunctions means that it is not always easy, or even possible, to identify the rhetorical relationship that they are intended to be signalling. For example, our understanding of the (b) versions above is heavily reliant on extralinguistic information; it is only our knowledge of the possible consequences of a broken printer that allows us to recognise the subordinate role of the second part of (38b) to the first, and thus to recover the underlying NON-VOLITIONAL RESULT relation. Similarly, it is only our knowledge of the possible consequences of not turning off an electrical appliance in certain circumstances that allows us to recover the underlying OTHERWISE relation in (40b).

Given our generation goal to convey the rhetorical role of all propositions of a message, and the previously discussed constraints imposed by the communicative setting in which generation is performed, it is therefore important for us to ensure that the syntactic operation of paratactic coordination is only ever used in genuine cases of the conjoining of rhetorically coordinate propositions. By being precise in our usage, we increase the chances of the reader recovering the intended rhetorical relation between the coordinated propositions. Our first heuristic for paratactic coordination identifies the criterion for determining which rhetorical relations it can be applied to:

> Heuristic 10: Paratactic Coordination can only be applied to multi-nuclear relations.

Mann and his colleagues identify three multi-nuclear relations: SEQUENCE, CONTRAST and LIST [Mann and Thompson 1987b; Matthiessen and Thompson 1987]. We have added a fourth, ALTERNATIVE, to this set. ALTERNATIVE is one of the two relations that Grimes [1975] considers to be 'purely paratactic'. It is closely related to Mann and Thompson's OTHERWISE relation, the crucial difference being that it does not involve a dependency relationship between its elements. This difference is shown in (40) above, where (40a) involves the ALTERNATIVE relation and (40b) OTHERWISE.

Heuristic 10 ensures the unambiguous mapping between propositions that are coordinate at the rhetorical level, and coordinate structures at the syntactic



[1] John wants to be a diplomat.

[2] John likes travelling.

[3] John will take the Foreign Office exams tomorrow.

Figure 3.6: A message to which Heuristic 10 can be applied

level. It guarantees the possibility of producing all the (a) versions of (38)–(40) but none of the (b) versions as expressions of the same message, even though they are grammatical and semantically non-anomalous.

By restricting paratactic coordination to information units that are not only coordinate but nuclear, this heuristic also prevents the undesirable coordination of some multi-satellite structures. Consider, for example, the message in Figure 3.6.

Although propositions [2] and [3] are coordinate with respect to [1]), the fact that they belong to different rhetorical relations makes them unsuitable candidates for joint membership of the multi-nuclear LIST relation. Heuristic 10 thus prevents them from being generated as a paratactically coordinated complex sentence like (41) or (42).

(41) *John wants to be a diplomat because *he likes travelling and will take the Foreign Service exams next week.*

(42) *John wants to be a diplomat so *he will take the Foreign Service exams next week and he likes travelling.*

If, however, [3] were replaced in the message by *John likes going to diplomatic parties*, also in a MOTIVATION relation with [1], then (43), which is clearly appropriate, could be generated.

(43) John wants to be a diplomat because *he likes travelling and going to diplomatic parties.*

Having determined which rhetorical links can be expressed through paratactic coordination, we now need to stipulate which members of the set of coordinating conjunctions can be applied to which rhetorical relations.

> Heuristic 11: The paratactic marker *and* must only be applied to SEQUENCE and LIST, *but* to CONTRAST, and *or* to ALTERNATIVE.

This heuristic guarantees that the paratactic coordination of propositions does not result in the generation of invalid or rhetorically ambiguous text. For example, it will ensure that only the unstarred sentences could be generated as expressions of the following:

SEQUENCE:

(44) Put the loose tea in the teapot and pour in the boiling water.

(45) *Put the loose tea in the teapot but pour in the boiling water.

(46) *Put the loose tea in the teapot or pour in the boiling water.

LIST:

(47) John likes apples and bananas.

(48) *John likes apples but bananas.

(49) *John likes apples or bananas.

CONTRAST:

(50) The meal looked good but tasted like poached cardboard.

(51) *The meal looked good and tasted like poached cardboard.

(52) *The meal looked good or tasted like poached cardboard.

ALTERNATIVE:

(53) John wants to go to Sussex or Essex.

(54) *John wants to go to Sussex and Essex.

(55) *John wants to go to Sussex but Essex.

Clearly, (45) and (46) are not synonymous with (44). Neither are (48) and (49) with (47), (51) and (52) with (50), or (54) and (55) with (53).

A characteristic feature of multi-nuclear rhetorical relations is that the order of appearance of their elements in the text tends not to affect the message. The only exception to this is SEQUENCE, which involves the notion of temporal priority.[8] Heuristic 12 allows for a different ordering of the propositions of multi-nuclear relations in the text than in the message in situations where this is appropriate.

> Heuristic 12: Propositions of all relations except SEQUENCE can be reordered during paratactic coordination.

This heuristic provides the flexibility for generating (56), (57) and (58) as synonyms of (47), (50) and (52) respectively, and prevents the generation of (59) as a synonym of (44).

(56) John likes bananas and apples.

(57) The meal tasted like poached cardboard but looked good.

(58) John wants to go to Essex or Sussex.

(59) *Pour the boiling water in the teapot and put in the loose tea.

This flexibility is often useful, especially in cases where the order of presentation of the propositions affects the thematic flow of the text. For example, (38a) would be more appropriate than its alternative (60) if the preceding sentence were (61), and *vice versa* if the preceding sentence were (62).

(60) The chapter is due tomorrow and the printer is broken.

(61) The printer always fails when I most need it.

(62) I doubt that I'll be able to finish this chapter on time.

It should be noted, however, that there are cases where the linguistic realisation of the individual propositions will rule out certain otherwise permissible orderings. For example, the alternative ordering of propositions [2] and [3] in (43), which would lead to:

(63) *John wants to be a diplomat because *he likes going to diplomatic parties and travelling.*

[8] See Lakoff [1971] and Schmerling [1975] for a more detailed discussion of this.

would clearly not be desirable. Situations like these do not become apparent until quite late in the generation process and thus cannot be taken into account when the message itself is being planned. It is therefore important to have the flexibility for reordering that is provided by Heuristic 12.

It is also important that this flexibility should *not* be extended to the SEQUENCE relation, since the resulting text would violate Grice's directive that the text be 'orderly', and lead to an erroneous interpretation of the message (see also Schmerling [1975]). Reorderings of the propositions of SEQUENCE must be marked by hypotactic coordination.

It follows from Heuristic 12 that the number of orderings of the relevant propositions that are possible during the construction of paratactically coordinated complex sentences is the factorial of the number of propositions of the relation. As we have just seen, some of these will lead to better text than others. This is especially true in cases where there are a fair number of propositions to be considered, and thus often more than one sentence to be formed. In such situations there will be a need to bias the generation process towards the production of the best combination of coordinated propositions. Heuristic 13 provides one way of doing this.

> Heuristic 13: The greater the number of shared elements between propositions, the more desirable it is to coordinate them.

This heuristic biases paratactic coordination towards complex sentences which, to use Lakoff's [1971] terms, share a common topic. By promoting the generation of paratactically coordinated complex sentences with similar propositions, this heuristic has a direct bearing on the conciseness of the resulting text, since it encourages coordinations which provide the greatest opportunities for ellipsis.

A broad view of the operation of the heuristics for paratactic coordination can be seen with reference to the message shown in Figure 3.7. Taken together, the above heuristics will provide the possibility for expressing this message as (64).

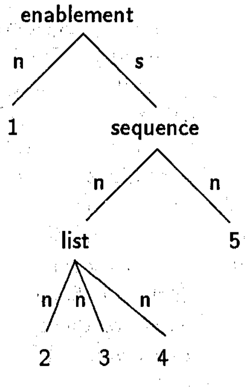
(64) In order to change the oil in the tank, one must drain the tank and sump of oil, replace the oil filter, and refill the tank with oil.

The possibility of applying coordination to propositions [1] and [2], thereby generating something like (65) is prevented by Heuristic 10.

(65) *Change and drain the oil in the tank ...

Heuristic 11 blocks the possibility of generating an incorrect paratactic marker, thereby conveying the wrong message, as in (66):

(66) *...drain the tank or sump ...



[1] Change the oil in the tank.

[2] Drain the oil in the tank.

[3] Replace the oil filter.

[4] Drain the oil in the sump.

[5] Refill the oil in the tank.

Figure 3.7: A message to which the heuristics for paratactic coordination can be applied

Changing the order of the elements of the SEQUENCE relation in the text would result in something like (67). Although clearly grammatical, (67) would be undesirable since, like (66), it conveys the wrong message. This is prevented by Heuristic 12.

(67) *... refill and drain the oil in the tank ...

Finally, by biasing the generation process towards the paratactic coordination of propositions [2] and [4], Heuristic 13 prevents the production of the stylistic blunder that would occur if [3] were chosen over [2], thereby giving (68):

(68) *... drain the oil from the tank and replace the oil filter, drain the oil from the sump ...

3.5 Summary

There is little need to argue for the importance of stylistic factors in the readability of a text. Until now, the problem has always been that of determining just how we should go about giving our texts good style. One approach to this problem is to allow the process of text production to be guided by what is known about the way in which readers understand texts. We have shown here that this approach is effective, at least with regard to the rhetorical aspects of text generation. It allows us to maximise the possibility that the message will be retrieved from the text by ensuring that the rhetorical structure of the message is enhanced by the choice of discourse structure for the text, which is in turn reflected in the choice of syntactic structures. An added advantage of this approach is that it provides us with a sound theoretical basis for dealing with some aspects of the issue of sentence content and organisation (see Hovy [this volume]).

Acknowledgements

We are extremely grateful to Graça Volpe Nunes, Patrizia Tabossi and Richard Cole for their helpful discussion of many of the ideas discussed in this chapter.

References

Baddely, A D [1986] *Working Memory.* Oxford: Oxford University Press.

Carpenter, P A and Just, M A [1989] The role of working memory in language comprehension. In D Klahr and K Kotovsky (eds) *Complex Information Processing,* pp31–68. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, H H and Clark, E V [1977] *Psychology and Language.* New York: Harcourt Brace Jovanovich.

Daneman, M and Carpenter, P A [1980] Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behaviour,* **19**, 450–466.

Fillenbaum, S [1971] On coping with ordered and unordered conjunctive sentences. *Journal of Experimental Psychology,* **87**, 93–98.

Fillenbaum, S [1974a] Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology,* **102**, 574–578.

Fillenbaum, S [1974b] Or: Some uses. *Journal of Experimental Psychology,* **103**, 913–921.

Foss, D J and Jenkins, C M [1973] Some effects of context on the comprehension of ambiguous sentences. *Journal of Verbal Learning and Verbal Behaviour,* **12**, 577–589.

Freedle, R O and Craun, M [1970] Observations with self-embedded sentences using written aids. *Perception and Psychophysics,* **7**, 247–249.

Garnham, A [1985] *Psycholinguistics: central topics.* London: Methuen.

Gleitman, L R [1965] Coordinating conjunctions in English. *Language*, **41**, 260–293.

Grice, H P [1975] Logic and conversation. In P Cole and J L Morgan (eds) *Syntax and Semantics*, Volume 3: *Speech Acts*, pp41–58. New York: Academic Press.

Grimes, J E [1975] *The Thread of Discourse.* The Hague: Mouton.

Hovy, E H [1988] Planning coherent multisentential text. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, State University of New York at Buffalo, Buffalo, NY, 7–10 June 1988, pp163–169.

Hovy, E H [1988] Approaches to the planning of coherent text. Presented at the *4th International Workshop on Text Generation*, Los Angeles, 1988. Also in C L Paris, W R Swartout, and W C Mann (eds), *Natural Language in Artificial Intelligence and Computational Linguistics*, to appear.

Hovy, E H [1990] Unresolved Issues in Paragraph Planning. This volume.

Jarvella, R J [1970] Effects of syntax on running memory span for connected discourse. *Psychonomic Science*, **19**, 235–236.

Jarvella, R J [1971] Syntactic processing and connected speech. *Journal of Verbal Learning and Verbal Behavior*, **10**, 409–416.

Johnson-Laird, P N [1983] *Mental Models.* Cambridge: Cambridge University Press.

Kimball, J P [1973] Seven principles of surface structure parsing in natural language. *Cognition*, **2**, 15–47.

Kobsa, A and Wahlster, W (eds) [1989] *User Models in Dialog Systems.* Berlin: Springer-Verlag.

Lackner, J R and Garrett, M F [1972] Resolving ambiguity: Effects of biasing context in the unattended ear. *Cognition*, **1**, 359–372.

Lakoff, R [1971] If's, And's, and But's about conjunction. In C J Fillmore and D T Langendoen (eds.), *Studies in Linguistic Semantics*, pp114–149. New York: Holt, Rinehart and Winston.

Larkin, W and Burns, D [1977] Sentence comprehension and memory for embedded structure. *Memory and Cognition*, **5**, 17–22.

Mann, W and Thompson, S A [1985] Assertions from discourse structure. *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, Berkeley, California, 16–18 February 1985.

Mann, W C and Thompson, S A [1987a] Rhetorical Structure Theory: A Framework for the Analysis of Texts. USC/Information Sciences Institute Research Report RS-87-185.

Mann, W and Thompson, S [1987b] Rhetorical Structure Theory: A theory of text organisation. In L Polanyi (ed.) *The Structure of Discourse.* Norwood, NJ: Ablex.

Matthiessen, C M I M and Thompson, S A [1987] The structure of discourse and 'subordination'. In J Halman and S A Thompson (eds), *Clause Combining in Discourse and 'Subordination'.* Amsterdam: John Benjamins Publishing Company.

McKeown, K R [1985] *Text Generation.* Cambridge: Cambridge University Press.

Meteer, M W [1988a] Defining a vocabulary for text planning. *Proceedings of the AAAI-88 Workshop on Text Planning and Generation*, St. Paul, Minnesota, 25 August 1988.

Meteer, M W [1988b] The implication of revisions for natural language generation. *Proceedings of the Fourth International Workshop on Natural Language Generation*, Catalina Island, California, 17–21 July 1988.

Meteer, M W [1989] The SPOKESMAN Natural Language Generation System. Bolt, Beranek and Newman Technical Report No. 7090.

Miller, G A and Isard, S D [1963] Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, **2**, 217–228.

Miller, G A and Isard, S D [1964] Free recall of self-embedded English sentences, *Information and Control*, **7**, 292–303.

Newell, A and Simon, H A [1972] *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

Paris, C L [1987] The Use of Explicit User Models in Text Generation: Tailoring to a User's Level of Expertise. PhD Thesis, Department of Computer Science, Columbia University.

Quirk, R, Greenbaum, S, Leech, G, and Svartvik, J [1985] *A Comprehensive Grammar of the English Language.* London: Longman.

Schlesinger, I M [1968] *Sentence Structure and the Reading Process.* The Hague: Mouton.

Schmerling, S F [1975] Asymmetric conjunction and rules of conversation. In P Cole and J L Morgan (eds) *Syntax and Semantics*, Volume 3: *Speech Acts*, pp211-231. New York: Academic Press.

Scott, D R [In preparation] A cognitive approach to the generation of hypotaxis.

Souza, C S, Scott, D R and Nunes, M G V [1989] Enhancing text quality in a question-answering system. In J Siekmann (ed) *Lecture Notes in Artificial Intelligence.* Berlin: Springer-Verlag.

Strunk, W and White, E B [1979] *The Elements of Style*, 3rd Edition. New York: The Macmillan Co.

van Dijk, T A [1977] *Text and Context.* London: Longman.