

Capturing the Interaction between Aggregation and Text Planning in Two Generation Systems

Hua Cheng and Chris Mellish

Division of Informatics, University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, UK
huac,chrism@dai.ed.ac.uk

Abstract

In natural language generation, different generation tasks often interact with each other in a complex way. We think that how to resolve the complex interactions inside and between tasks is more important to the generation of a coherent text than how to model each individual factor. This paper focuses on the interaction between aggregation and text planning, and tries to explore what preferences exist among the features considered by the two tasks. The preferences are implemented in two generation systems, namely ILEX-TS and a text planner using a Genetic Algorithm. The evaluation emphasises the second implementation and shows that capturing these preferences properly can lead to coherent text.

1 Discourse coherence and aggregation

In NLG, theories based on domain-independent rhetorical relations, in particular, Rhetorical Structure Theory (Mann and Thompson, 1987), are often used in text planning, whose task is to select the relevant information to be expressed and organise it into a hierarchical structure which captures certain discourse preferences such as preferences for the use of rhetorical relations.

In the theory of discourse structure developed by Grosz and Sidner (1986), each discourse segment exhibits two types of coherence: local coherence among utterances inside the segment, and global coherence between this segment and other discourse segments. Discourse segments are connected by either a *dominance* relation or a *satisfaction-precedence* relation.

There has been an effort to synthesise the two accounts of discourse structure. Moser and Moore (1996) argue that the two theories have

considerable common ground, which lies in the correspondence between the notion of dominance and nuclearity. It is possible to map between Grosz and Sidner's linguistic structure and RST text structure, and relation-based coherence and global coherence capture similar discourse properties.

Oberlander et al. (1999) propose a distinction between two types of discourse coherence: *proposition-based coherence*, which exists between text spans connected by RST relations except for *object-attribute elaboration*, and *entity-based coherence*, which exists between spans of text in virtue of shared entities. *entity-based coherence* captures the coherence among adjacent propositions, which resembles *local coherence* in Grosz and Sidner's theory.

To generate a coherent text, the text planning process must try to achieve both local (entity-based) and global (relation-based) coherence. Since the task of aggregation is to combine simple representations together to form a complex one, which in the mean time leads to a shorter text as a whole, aggregation could affect the ordering of text plans and the length of the whole text. Therefore, it is closely related to the task of maintaining both types of coherence. Here we treat embedding as a type of aggregation.

There is no consensus as to where aggregation should happen or how it is related to other generation processes (Wilkinson, 1995; Reape and Mellish, 1999). In many NLG systems, aggregation is a post planning process whose preferences are only partially taken into account by the text planner.

1.1 Aggregation and local coherence

In a structured text plan produced by the text planner, local coherence is normally maintained through the ordering of the selected facts, where

certain types of center transition (e.g. center continuation) are preferred over others (e.g. center shifting) (Centering Theory (Grosz et al., 1995)). Aggregation may affect text planning by taking away facts from a sequence featuring preferred center movements for embedding or subordination. As a result, the preferred center transitions in the original sequences could be cut off. For example, comparing the first two descriptions of a necklace in Figure 1, 2 is less coherent than 1 because of the shifting from the description of the necklace to that of the designer, which is a side effect of embedding.

Since the centers of sentences are normally NPs and embedding adds non-restrictive components into an NP, it could affect the way a *Cb* is realised (e.g. preventing it from being a pronoun). As pointed out in (Grosz et al., 1995), different realisations (e.g. pronoun vs. definite description) are not equivalent with respect to their effect on coherence. Therefore, embedding could influence local coherence by forcing a different realisation from that preferred by Centering Theory. There is an obvious need to balance the consideration for local coherence and stylistic preferences.

1.2 Aggregation and global coherence

Different types of aggregation need to be compatible among themselves, in particular, embedding and semantic parataxis and hypotaxis. Using the abstraction of RST, semantic parataxis concerns facts related by explicit multi-nuclear semantic relations (e.g. *sequence* and *contrast*) or by implicit connections like parallel common parts. If two facts have at least two identical parallel components, we say that a *conjunct* or *disjunct* relation exists between them, and these relations are multi-nuclear relations. Semantic hypotaxis concerns facts connected by nucleus-satellite relations (e.g. *cause*). Semantic parataxis and hypotaxis feature in relation-based coherence and they depend on the text planner to put the related facts next to each other in order to perform a combination.

(Cheng, 1998) describes interactions that need to be taken into account in aggregation. Firstly, complex embedded components like non-restrictive clauses may interrupt the semantic connection or syntactic similarity between a set of clauses. Secondly, the possibilities of other types of aggregation should be consid-

ered for both the main fact and the fact to be embedded during embedding decision making. And thirdly, performing parataxis inside a hypotaxis could convey wrong information.

We argue that the effect of aggregation is not limited to the particular NP or sentence where aggregation happens, but to the coherence of the text as a whole. The complex interactions demand the features of aggregation to be evaluated together with other coherence features and aggregation to be planned as a part of text structuring. This requires better coordination between aggregation and other generation tasks as well as among different types of aggregation than is present in current NLG systems.

In this paper, we describe how to capture the above interactions as preferences among related features, and the implementation of the preferences in two very different generation architectures to produce descriptions of museum objects on display.

2 Preferences among coherence features

We claim that it is the relative preferences among features rather than the absolute magnitude of each individual one that play the crucial role in the production of a coherent text. In this section we discuss the preferences among features related to text planning, based on which those for embedding can be introduced.

2.1 Preferences for global coherence

A semantic relation other than *conjunct* or *disjunct* is preferred to be used whenever possible because it usually conveys interesting information about domain objects and leads to a coherent text span. If a *conjunct* relation shares a fact with a semantic relation, the *conjunct* should be suppressed. For example, in 3 of Figure 1, apart from other relations, there is an *amplification* relation signalled by *indeed* and a *conjunct* between the last two propositions. Compared with 3, 4 is less preferred because it misses the *amplification* and the center transition from *the necklace* to *an Arts and Crafts style jewel* is not so smooth, whereas 3 expresses the *amplification* explicitly and the *conjunct* implicitly.

However, a semantic relation can only be used if the knowledge assumed to be shared by the hearer is introduced in the previous discourse (Mellish et al., 1998a). We assume the strategy

1. *This necklace is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance, this necklace has floral motifs. It was designed by Jessie King. King was Scottish. She once lived in London.*
2. *This necklace, which was designed by Jessie King, is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance, this necklace has floral motifs. King was Scottish. She once lived in London.*
3. *The necklace is in the Arts and Crafts style. It is set with jewels in that it features cabuchon stones. Indeed, an Arts and Crafts style jewel usually uses cabuchon stones. It usually uses oval stones.*
4. *The necklace is in the Arts and Crafts style. It is set with jewels in that it features cabuchon stones. An Arts and Crafts style jewel usually uses cabuchon stones and oval stones.*

Figure 1: Aggregation examples

of (Mellish et al., 1998a) which uses a *joint* relation to connect every two text spans that do not have a semantic relation other than *object-attribute elaboration* and *conjunct/disjunct* in between. Although *joint* is not preferred when other relations are present, it is better than missing presuppositions or embedding a *conjunct* relation inside a semantic relation. Therefore, we have the following heuristics, where “A>B” means that A is preferred over B.

Heuristic 1 *Preferences among features for global coherence:*

*a semantic relation > Conjunct/Disjunct > Joint > presuppositions not met
Joint > Conjunct inside a semantic relation*

2.2 Preferences for local coherence

One way to achieve local coherence is to control center transitions among utterances. In Centering Theory, Rule 2 specifies preferences among center movement in a locally coherent discourse segment: sequences of *continuation* are preferred over sequences of *retaining*, which are then preferred over sequences of *shifting*.

Brennan et al. (1987) also describe typical discourse topic movements in terms of center transitions between pairs of utterances. They argue that the order of coherence among the transitions is *continuing > retaining > smooth shifting > abrupt shifting*. Instead of claiming that these are the best models, we use them simply as an example of linguistic models being used for evaluating features of text planning.

A type of center transition that appears frequently in descriptive text is that the description starts with an object, but shifts to associated objects or perspectives of that object. This

is a type of abrupt shifting, but it is appropriate as long as the objects are highly associated to the original object (Schank, 1977). This phenomenon is handled in the system of (Grosz, 1977), where subparts of an object are included into a focus space as the implicit foci when the object itself is to be included.

We call this center movement an *associate shifting*, where the center moves from a trigger entity to a closely associated entity. Our informal observation from museum descriptions shows that *associate shifting* is preferred by human writers to all other types of center movements except for *continuation*. There are two types of *associate shifting*: where the trigger is in the previous utterance or two entities in two adjacent utterances have the same trigger. There is no preference between them.

Heuristic 2 summarises the above preferences. We admit that these are strict heuristics and that human texts are sometimes more flexible.

Heuristic 2 *Preferences among center transitions:*

Continuation > Associate shifting > Retaining > Smooth shifting > Abrupt shifting

2.3 Preferences for both types of coherence

Two propositions can be connected in different ways, e.g. through a semantic relation or a smooth center transition only. Since a semantic relation is always preferred, we have the following heuristic:

Heuristic 3 *Preferences among semantic relations and center transitions:*

a semantic relation > Joint + Continuation

2.4 Preferences for embedding

We distinguish between a *good*, *normal* and *bad embedding* based on the features it bears. We do not claim that the set of features is complete. In a different context, more criteria might have to be considered.

A *good embedding* is one satisfying all the following conditions:

1. The referring part is an indefinite, a demonstrative or a bridging description (as defined in (Poesio et al., 1997)).
2. The embedded part can be realised as an adjective or a prepositional phrase (Scott and de Souza, 1990).
3. In the resulting text, the embedded part does not lie between text spans connected by semantic parataxis and hypotaxis.
4. There is an available syntactic slot to hold the embedded part.

A *good embedding* is highly preferred and should be performed whenever possible. A *normal embedding* is one satisfying condition 1, 3 and 4 and the embedded part is a relative clause which provides additional information about the referent. *Bad embeddings* are all those left, for example, if there is no available syntactic slot for the embedded part.

Since semantic parataxis has a higher priority than embedding (Cheng, 1998), a *good embedding* should be less preferred than using a *conjunct* relation, but it should be preferred over a center continuation for it to happen.

To decide the interaction between an embedding and a center transition, we use the first two examples in Figure 1 again. The only difference between 1 and 2 is the position of the sentence “*This necklace was designed by Jessie King*”, which can be represented in terms of features of local coherence and embedding as follows:

the last three sentences in 1: *Joint + Continuation + Joint + Smooth shifting*

the last two sentences plus embedding in 2: *Joint + Abrupt shifting + Normal embedding*

1 is preferred over 2 because the center moves more smoothly in 1. The heuristics derived from the above discussions are summarised below:

Heuristic 4 *Preferences among features for embedding and center transition:*

Good embedding > Normal embedding > Joint > Bad embedding

Continuation + Smooth shifting + Joint > Abrupt shifting + Normal embedding

Good embedding > Continuation + Joint

Conjunct > Good embedding

The ‘+’ symbol can be interpreted in different ways, depending on how the features are used in an NLG system. In a traditional system, it means the coexistence of two features. In a system using numbers for planning, it can have the same meaning as the arithmetic symbol.

3 Capturing the preferences in ILEX

The architecture of text planning has a great effect on aggregation possibilities. In object descriptive text generation, there lacks a central overriding communicative goal which could be decomposed in a structured way into subgoals. The main goal is to provide interesting information about the target object. There are generally only a small number of relations, mainly *object-attribute elaboration* and *joint*. For such a genre, a domain-dependent bottom-up planner (Marcu, 1997) or opportunistic planner (Mellish et al., 1998b) suits better than a domain-independent top-down planner. In these architectures, aggregation is important to text planning because it changes the order in which information is expressed. The first implementation we will describe is based on ILEX (Oberlander et al., 1998).

ILEX is an adaptive hypertext generation system, providing natural language descriptions for museum objects. The bottom-up text planning is fulfilled in two steps: a content selection procedure, where a set of fact nodes with high relevance is selected from the Content Potential (following a search algorithm), and a content structuring procedure, where selected facts are reorganised to form entity-chains (based on the theory of entity-based coherence), which represent a coherent text arrangement.

To make it possible for the ILEX planner to take into account aggregation, we use a revised version of Meteer’s Text Structure (Meteer, 1992; Panaget, 1997) as the intermediate level of representation between text planning and sentence realisation to provide abstract syntactic constraints to the planning. We call this system ILEX-TS (ILEX based on Text Structure).

In ILEX-TS, abstract referring expression determination and aggregation are performed during text structuring. For each fact whose Text Structure is being built, if an NP in the fact can take modifiers, the embedding process will find a list of elaboration facts to the referent and make embedding decisions based on the constraints imposed by the NP form. The decisions include what to embed and what syntactic form the embedded part should use.

Heuristic 1, 2 and 3 are followed naturally by the ILEX text planner, which calculates the best RS tree and puts facts connected by the imaginary *conjunct* relation next to each other. It tries to feature center continuations as often as possible. When it needs to shift topic, it uses a smooth shifting.

ILEX-TS has a set of embedding rules, where those rules featuring good embedding are always used first, then a rule featuring a normal embedding. Bad embedding is not allowed at all. To coordinate different types of aggregation, the algorithm checks parataxis and hypotaxis possibilities for each nucleus fact and the fact to be embedded before it applies an embedding rule. These realise most of Heuristic 4 (except for the second set). However, because the various factors are optimised in order (with no backtracking), there is no guarantee that the best overall text will be found. In addition, complex interactions between aggregation and center transition cannot be easily captured.

4 Text planning using a GA

Although most heuristics can be followed in ILEX-TS, some interactions are missing, for example, 2 of Figure 1 will probably be generated. For better coordination, we adopt the text planner based on a Genetic Algorithm (GA) as described in (Mellish et al., 1998a). The task is, given a set of facts and a set of relations between facts, to produce a legal rhetorical structure tree using all the facts and some relations.

A fact is represented in terms of a subject, a verb and a complement (as well as a unique identifier). A relation is represented in terms of the relation name, the two facts that are connected by the relation and a list of precondition facts which need to have been mentioned before the relation can be used¹.

¹As this is an experimental system, the ability of the

A genetic algorithm is suitable for such a problem because the number of possible combinations is huge and the search space is not perfectly smooth and unimodal (there can be many good combinations). Also the generation task does not require a global optimum to be found. What we need is a combination that is coherent enough for people to understand.

(Mellish et al., 1998a) summarises the genetic algorithm roughly as follows:

1. Enumerate a set of random initial sequences by loosely following sequences of facts where consecutive facts mention the same entity.
2. Evaluate sequences by evaluating the rhetorical structure trees they give rise to.
3. Perform mutation and crossover on the sequences.
4. Stop after a given number of iterations, and return the tree for the “best” sequence.

The advantage of this approach is that it provides a mechanism to integrate planning factors in the evaluation function and search for the best combinations of them. So it is an excellent framework for experimenting with the interaction between aggregation and text planning.

In the algorithm, the RS trees are right-branching and are almost deterministically built from sequences of facts. Given two sequences, crossover inserts a random segment from one sequence in a random position in the other to produce two new sequences. Mutation selects a random segment of a sequence and moves it into a random position in the same sequence.

To explore the whole space of aggregation, we decide not to perform aggregation on structured facts or on adjacent facts in a linear sequence because they might restrict the possibilities and even miss out good candidates. Instead, we define a third operator called *embedding mutation*. Suppose we have a sequence $[U_1, U_2, \dots, U_i, \dots, U_n]$, where we call each element of the sequence a *unit*, which can be either a fact or a list of facts or units with no depth limit. For a list, we call its very first fact the *main fact*,

system is limited in all aspects. It does not have a real realisation component, so the parts we are less interested in are realised by canned phrases for readability.

Features/Factors	Values (raters)	
	1	2
Semantic relations		
a joint	-20	-46
a conjunct or disjunct	10	11
a relation other than joint, conjunct or disjunct	21	69
a conjunct inside other semantic relations	-50	-63
a precondition not satisfied	-30	-61
Focus moves		
a continuing	20	7
an associate shifting	16	1
a smooth shifting	14	-3
resuming a previous focus	6	-43
Embedding		
a good embedding	6	3
a normal embedding	3	0
a bad embedding	-30	-64
Others		
topic not mentioned in the first sentence	-10	-12

Table 1: Two different raters satisfying the same constraints

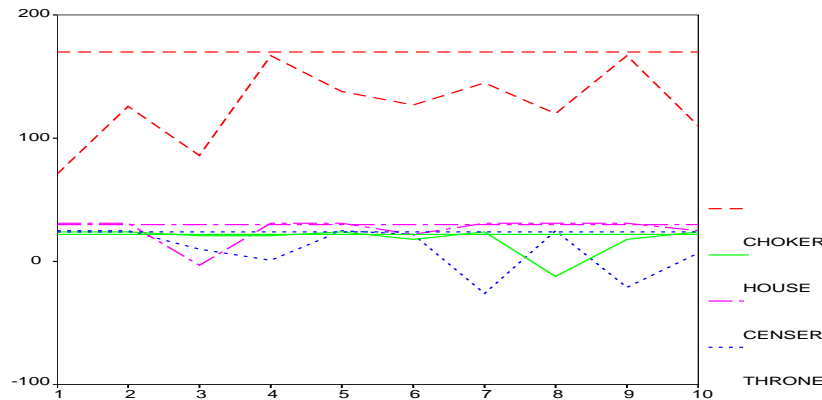


Figure 2: Scores for four museum descriptive texts

into which the remaining facts in the list are to be embedded. The embedding mutation randomly selects a unit U_i from the sequence and an entity in its main fact. It then collects all the units mentioning this entity and randomly chooses one U_k . The list containing these two units $[U_i, U_k]$ represents a random embedding and will be treated as a single unit in later operations. It takes the position of U_i to produce a new sequence $[U_1, U_2, \dots, [U_i, U_k], \dots, U_n]$ and all repetitions outside $[U_i, U_k]$ are removed. This sequence is then evaluated and ordered in the population.

The probabilities of applying the three operators are: 65% for crossover, 30% for embedding

mutation and 5% for normal mutation. This is because the first two are more likely to produce sequences bearing desired properties by either combining the good bits of two sequences or performing a reasonable amount of embedding, whereas normal mutation is entirely random².

5 Justifying the GA evaluation function

The linguistic theories discussed in Section 2 only give evidence in qualitative terms. For a GA-based planner to work, we have to come up with actual numbers that can be used to evalu-

²The values for crossover and mutation rate used in our algorithm are fairly standard.

The **small portable throne from the time of the Qianlong Emperor 1736-95** is made of **lacquered wood with decoration in gold and red**. It was used in the private apartments of the Imperial Palaces. The cover **from the reign of Jiaqing, 1796-1820** is woven in yellow silk, **which is the imperial colour of the Qing Dynasty, 1644-1911**. It would have covered the throne when not in use.

The design on the seat is a **imperial five clawed dragon in a circular medallion**. On the inside of the arm pieces are small shelves. Precious possessions can be placed in small shelves and can be studied as an aid to contemplation.

Figure 3: A generated text scored the highest, with the embedded parts highlighted

	Score2	Score3	Score4	Score5	Score6
Score1	.9567	.9337	.9631	.9419	.9515
Score2		.9435	.8819	.9280	.9185
Score3			.8650	.8462	.9574
Score4				.9503	.8940
Score5					.8486

Table 2: Correlations between six raters

ate an RS tree. Mellish et al. (1998a) present some scores for evaluating the basic features of a tree, but they make it clear that the scores are there for descriptive purpose, not for making any serious claim about the best way of evaluating RS trees.

The methodology we adopted was that we took the existing evaluation function and extended it to take into account features for local coherence, embedding and semantic parataxis. This resulted in rater 1 in Table 1, which satisfied all the heuristics mentioned in Section 2. We manually broke down four human written museum descriptions into individual facts and relations and reconstructed sequences of facts with the same orderings and aggregations as in the original texts. We then used our evaluation function to score the RS trees built from these sequences. In the mean time, we ran the GA algorithm for 5000 iterations on the facts and relations for 10 times. The results are shown in Figure 2, where the four line styles correspond to the four texts. The jagged lines represent the scores of the machine generated texts and the straight lines represent the scores for the corresponding human texts.

All human texts were scored among the highest and machine generated texts can get scores very close to human ones sometimes. Since the human texts were written and revised by museum experts, they can be treated as “nearly

best texts”. The figure shows that the evaluation function based on our heuristics can find good and correct combinations. The reason for a relatively bad text being generated sometimes might be that really bad sequences were produced at the beginning. This could be improved by using certain heuristics to get better initial sequences. Also when the number of facts becomes larger, more iterations are needed to get readable texts. Figure 3 gives a text generated using rater 1.

To justify our claim that it is the preferences among generation factors that decide the coherence of a text, we fed the preferences into a constraint based program. If a feature can take a range of values, the program randomly selects a number in that range. A number of raters compatible with the constraints were generated and one of them is given in Table 1 as rater 2. We then generated all possible combinations, including embedding, of seven facts from a human text and used six randomly produced raters to score each of them.

The qualities of the generated texts are normal distributed according to all raters. The raters distinguish between good and bad texts and they classify the majority of texts as of moderate quality and only very small percentages as very good or very bad texts. The behaviours of the raters are very similar as the histograms are of roughly the same shape.

To see to what extent the six raters agree with each other, we calculated the Pearson correlation coefficient between them, which is shown in Table 2. We can claim from the table that for this data, the scores from the six raters correlate, and we have a fairly good chance to believe that the six raters, randomly produced in a sense, agree with each other on evaluating the text and they measure basically the same thing.

6 Conclusions and future work

This paper describes an experiment with the preferences among features concerning aggregation and text planning, in particular, we present an mechanism for how relevant features can be scored to contribute together to the planning of a coherent text. The statistical results partially justify our claim that it is the preferences among generation features that decide the coherence of a text.

Our experiment could be extended in many ways, for example, validating the evaluation function through empirical analysis of human assessments of the generated texts, and using more texts to test the correlation between raters. The architecture based on the Genetic Algorithm can also be used for testing interactions between or within other text generation modules.

References

- Susan Brennan, Marilyn Walker Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, Stanford, CA.
- Hua Cheng. 1998. Embedding new information into referring expressions. In *Proceedings of COLING-ACL'98*, pages 1478–1480, Montreal, Canada.
- Barbara Grosz and Candace Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.
- Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.
- Barbara Grosz. 1977. The representation and use of focus in dialogue understanding. Technical report 151, SRI International.
- William Mann and Sandra Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RR-87-190, Information Sciences Institute, University of Southern California.
- Daniel Marcu. 1997. From local to global coherence: A bottom-up approach to text planning. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 629–635, Providence, Rhode Island.
- Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998a. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, Ontario, Canada.
- Chris Mellish, Mick O'Donnell, Jon Oberlander, and Alistair Knott. 1998b. An architecture for opportunistic text generation. In *Proceedings of the 9th International Workshop on Natural Language Generation*, Ontario, Canada.
- Marie Meteer. 1992. *Expressibility and The Problem of Efficient Text Planning*. Communication in Artificial Intelligence. Pinter Publishers Limited, London.
- Megan Moser and Johanna Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.
- Jon Oberlander, Mick O'Donnell, Ali Knott, and Chris Mellish. 1998. Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia*, 4:11–32.
- Jon Oberlander, Alistair Knott, Mick O'Donnell, and Chris Mellish. 1999. Beyond elaboration: Generating descriptive texts containing it-clefts. In T Sanders, J Schilperoord, and W Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*. Benjamins, Amsterdam.
- Franck Panaget. 1997. Micro-planning: A unified representation of lexical and grammatical resources. In *Proceeding of the 6th European Workshop on Natural Language Generation*, pages 97–106.
- Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. Research paper hcrc-rp87, Centre for Cognitive Science, University of Edinburgh.
- Michael Reape and Chris Mellish. 1999. Just what is aggregation anyway? In *Proceedings of the 7th European Workshop on Natural Language Generation*, pages 20–29, Toulouse, France.
- Roger Schank. 1977. Rules and topics in conversation. *Cognitive Science*, 1(1):421–441.
- Donia Scott and Clarisse Sieckenius de Souza. 1990. Getting the message across in rst-based text generation. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*, pages 47–73. Academic Press.
- John Wilkinson. 1995. Aggregation in Natural Language Generation: Another Look. Technical report, Computer Science Department, University of Waterloo.